PEDOMETRIC MAPPING

bridging the gaps between
conventional and pedometric approaches

Tomislav Hengl

ITC Dissertation number 101
ITC, P.O. Box 6, 7500 AA Enschede, The Netherlands

ISBN: 90-5808-896-0

Thesis Wageningen University and ITC with summary in Dutch and Croatian

Pedometric mapping

bridging the gaps between conventional and pedometric approaches

Ph.D. THESIS

to fulfill the requirements for the degree of doctor
on the authority of the Rector Magnificus
of Wageningen University,
Prof. dr.ir L. Speelman,
to be publicly defended on
Wednesday 17 September 2003 at 15.00 hrs
in the auditorium of ITC, Enschede

PROMOTOR:
Prof.dr.ir. A. Stein,
(Wageningen University and ITC, Enschede)

CO-PROMOTORS:
Dr. D.G. Rossiter,
(ITC, Enschede)

Dr. S. Husnjak,
(Department of Soil Science, University of Zagreb, Zagreb)

EXAMINING COMMITTEE:
Prof.dr.ir. A. Bregt,
(Laboratory of Geo-Information Science and Remote Sensing, Wageningen University, Wageningen)

Prof.dr. A.B. McBratney,
(Faculty of Agriculture, Food & Natural Resources, University of Sydney, Sydney)

Dr.ir. G.B.M. Heuvelink,
(Laboratory of Soil Science and Geology, Wageningen University, Wageningen)

Dr.ir. P.A. Finke,
(Biometris, Wageningen University, Wageningen)

Foreword

This book is a small contribution to pedometric methodology, specifically to mapping methods. It could serve as a handbook or a user's guide for anyone seeking to collate soil geoinformation. Pedometrics is an emerging field of science that is bound to attract more and more attention from the soil scientists in the near future. It could also have an impact on other sciences dealing with natural resources and on biometrics in general. For example, vegetation mappers work with similar datasets and face similar problems when modelling the spatial variation of vegetation or when seeking the best classification system. For pedometricians, the true acknowledgement of their work will come when we read about continuous vegetation, geomorphological, geological or civil engineering maps, developed according to the concepts of continuous soil mapping.
Those readers who would require more technical details about specific methods or would even like to use some of my datasets, should note that there is a supplementary CD-ROM at the back of the book[1]. There you will find additional documents such as technical and lecture notes[2] (e.g. "Comparison of kriging with external drift and regression-kriging" and "Digital terrain analysis in ILWIS"), animations and full colour graphics that could not be fitted into this book. My homepage web address[3] can also be used to access these supplementary materials.

Before presenting the thesis, it might be of interest to outline the genesis of each chapter and how I found my collaborators. This is a small *curriculum* of major activities:
We first began discussing the usability of soil maps in Croatia in 1999 during fieldwork in Baranja. At that time, Dr. Rossiter (my supervisor) was introduced to Dr. Stjepan Husnjak and Prof. Bogunović from the Department of Soil Science in Zagreb. We have described two full profiles together and *ad hoc* discussed about

[1] The CD-ROM is only available in a limited number of copies. Part of this material is also available on-line.
[2] Caution — most of these materials are unpublish and contain possible errors and misstatements.
[3] See the full address in Curriculum Vitae, at the end of the book.

the differences and difficulties in existing local soil classification systems and soil maps. The Baranja dataset was used to test the applicability of terrain parameters for aerial photo-interpretation (chapter 4). In March 2000, I began my PhD study with the same supervisor. My predecessor on the Division of Soil Science, Dr. van Groenigen, just defended a thesis on spatial sampling optimisation. Hence, the most logical promotor was Prof. Stein who has been teaching geostatistics at ITC and with whom I latter on wrote two chapters of my thesis. In June of 2001, Dr. Rossiter came to Croatia for a second time, during which we produced additional eight full profile descriptions. This time the observations were drawn from the whole country. In Split, we visited the Institute for Adriatic crops and Karst reclamation, where we were introduced to the work of Dr. Miloš. In September 2001, I participated in my first Pedometrics conference with an oral presentation, in which I proposed a sampling methodology that considers point allocation in the feature space (chapter 2).

In February 2002, Stjepan came to ITC as a visiting scientist and we began work on the quality of soil data in Croatia (chapter 8). The major part of the data used in this report was collected and processed during June 2002, when we produced soil maps of control areas. In March of the same year, I went to the GSTAT meeting in Reading with some raw ideas for the generic framework for spatial prediction. The presentation was very useful as Dr. Papritz and Dr. Gerard Heuvelink offered important suggestions that helped me develop the generic framework (chapter 5). Subsequently, Gerard suggested that I also take over the web-administration of the pedometrics website, which I gladly accepted. Gerard also introduced me to the theory of universal kriging and generalized least square estimation, which then helped me to develop equations for sampling optimization in feature space and for generic framework based on regression-kriging. In July and August, I went for a long trip to Melbourne, Sydney and Bangkok (Accuracy conference and World Congress of Soil Science). At the Accuracy conference in Melbourne, I received interesting comments on my work on continuous maps from the Accuracy participants (chapter 6), which helped me to improve the paper. Prof. Fisher specifically suggested that I replace the saturation with brightness in coding the uncertainty in the colour mixture algorithm. Following my acquaintance with Dr. Odeh and Dr. Budiman from the Pedometrics 2001 conference in Ghent, I then moved to University of Sydney, Agricultural Chemistry and Soil Science Department. I followed the path of Dennis Walvoort, a former (regular) visiting scientist at this Department and my collaborator. The visit to Sydney was especially inspiring for my future work. In Bangkok, I had the privilege of presenting a paper on the use of auxiliary maps to improve the mapping of soil variables from existing datasets[4]. At that time, pedometrics has

[4]My final suggestion of how to achieve this is given in chapter 7

been promoted to a provisional commission under IUSS, which was a good sign for all of us.
During the last year of my study I was mostly in Enschede working on my publications. In February 2003, I went to Zürich, where I was collaborating with my fellow college Stephan Gruber (PhD student at the Department of Geography) on methodology for reduction of errors in terrain parameters (chapter 3). From this collaboration we also developed a lecture note — "Digital Terrain Analysis in ILWIS", which is available in the supplementary materials.

A PhD thesis is rarely a product of pure individual work. Therefore, I feel obliged to express acknowledgements to the following people and organizations. I would first like to thank my sponsor — the Croatian Ministry of Science and Technology for awarding me the scholarship and investing in the croatian 'brain-resource'. Likewise, I would like to thank ITC for giving me additional funds and hosting me in Enschede for all these years. Being a member of such international community was a unique experience, which I will never forget (*"Once an ITC student, always an ITC student!"*).
Secondly, I would like to thank my ITC supervisor David G. Rossiter for teaching me how to "Publish" and not "Vanish", how to achieve objectivity and be self-critical. His dedication to work and science has been a major inspiration for me in the last few years. David, in spite all the style-conflicts we had, I am sure that this thesis will make you fell proud in the coming years. Many thanks to Alfred Stein also for his instructions and suggestions on the diplomacy in academia.
Although I have spent last five years primarily at ITC, there are also a (large) number of people in Croatia that I need to thank to. None of this would have been possible without the full dedication and support from my employer, University in Osijek. I am especially grateful to Gordana Kralik, the Rector, and Dragica Steindl, the secretary, for their support over these years. I would also like to thank my dear college Mladen Jurišić, who helped me arrange facilities in Osijek. Many thanks also to Stjepan Husnjak and Matko Bogunović, from the Department of Soil Science in Zagreb, for supervising my work and participating in the field works and data analysis; to Boško Miloš, from Split, who passionately supported my work and assisted during the fieldwork and to Nikola Pernar, from the Faculty of Forestry in Zagreb, who provided us with the Croatian digital soil database. I must not forget our helping personnel, students and graduates (our slaves): Dario Mihin, Tomislav Krema and Božidar Žanko who helped during the fieldwork. Navigating to the points through dense vegetation during the field work was a heavy task in some terrains. Not to mention large spiders in karstic hills, countless mosquitoes and unfriendly bushes and grasslands. This environment has almost cost David a bone fracture, Stjepan an eye and Boško a head. Happily, we are still all in one peace!

At ITC, I need to acknowledge the expertise and cordiality from the following colleges. I am grateful to Wouther Siderius for the *vertaling* of the summary and Dhruba Shrestha and Alfred Zinck for suggestions on specific topics. Thanks to my dear college Arta Dilo I was able to grasp some of the 'undigestible' mathematical algebra. I owe to Rolf de By (the future rector of ITC) for teaching me first 'steps' in LaTeX (and first steps in *salsa*). I will miss my colleges, PhD students Martin Yemefack and Ivan Bacic and their wise advices on life and work. A big thanks to Jose Santos who also helped with LaTeX and taught me how to mix capiriña (*"The sweetest drink in the World"*). *Hartelijk bedankt* to Benno and Job for all the posters/notes/maps you printed for me, Marion Pierik for solving my financial problems, Jaap De Ruijter for organizing social activities and Hans and Roelof for making sure that I leave the building on time (*"Ladies and gentleman, it is almost closing time... you have 2 minutes to leave the building before I lock you in the dark"*). I am grateful to the research managers Loes Colenbrander and Martin Hale for their protective policy towards the PhD students. The ITC's facilities and these extensive professional personnel truly make a difference for a success of a PhD research. I was never in doubt that this is the best institute[5] for my PhD study.

About four years ago, I was first time introduced to the term *pedometrics*. From then on, I have evolved from a passive sympathizer to an active member, pedometric.org web-administrator and contributor to the new methods (I hope). I would like to thank the Pedometrics society for accepting me and giving me inspiration for this work. Special thanks to Alex McBratney, Inakwu Odeh and Gerard Heuvelink for supporting my 'brave but unpolished' ideas and for investing so much of their precious time. My next *spiritus movens* will be to organize a Pedometrics-related conference in Croatia (the 'angry young pedometrician' is still angry!).
Finally, I am grateful to my girlfriend Monija for supporting me in 'good and bad' (especially for being with me in the bad times). A big thanks to my family, relatives (from Australia and Europe), my friends in Enschede (Maura, Blanca, Lyande, Zoki, Darja, Dragče) and in the rest of the World (Stephan, Tommy, Lassie, Martina, Bojana).

by Tomislav Hengl

In Enschede, September 2003

[5] I was choosing between six international institutes/universities: University of Florida, Arizona, Cranfield, Reading, Wageningen and ITC.

To my mother and father,
Mirjana and Ivan Hengl

Eto, jel to bilo teško?

Definition of terms and abbreviations

A common problem in new scientific fields such as pedometrics is that the researchers use different terms for the same things and the same terms for (completely) different ones. Two good examples of this are: (1) the confusion between the universal kriging, regression-kriging and kriging with external drift and (2) confusion between the `CLORPT` techniques and environmental correlation. In fact, these terminological confusion inspired me to write a technical note with a detailed comparison (involving both mathematics and practical issues) of regression-kriging and kriging with external drift. To avoid similar confusion, I will first give (my) definition of the concepts and terms and their synonyms. These are then used consistently throughout the thesis.

Accuracy, precision, quality Accuracy is the degree of conformity with the reality and needs to be estimated using cross-checking, i.e. validation set. Precision is the measure of model uncertainty. In spatial prediction precision is expressed with prediction error, which is the measure of goodness of fit. Accuracy and precision in GIS are related with the concept of data quality — accuracy is the predictive power and precision is the efficiency of data fitting.

Auxiliary variables/maps In the case of soil mapping, these are non-soil data sources that are used to improve mapping efficiency. Also referred to as secondary data, ancillary maps or non-soil layers. Typical examples are terrain parameters, remote sensing and airborne images, geological, geomorphological and hydrological maps.

Choropleth map Thematic map showing number of classes using a set of colors, shading levels or hatching patterns. It commonly presents crisp transitions, which corresponds to a polygon map in a GIS. An example of the double-crisp soil map can be seen in chapter 7, page 158.

`CLORPT` or `CLORPAN` approach to spatial prediction This term was coined by McBratney *et al.* (2000, 2003) using an abbreviation of Jenny's soil-forming factors: climate, organisms, relief, parent material and time. It means that the spatial prediction is achieved by employing the correlation with auxiliary environmental information (and not some conventional geostatistical technique). A synonym, suggested by McKenzie & Austin (1993), is *environmental correlation* although environmental correlation does not imply that the method is used to map soil variables only.

Continuous soil map A map showing distribution of soil types as spatially and/or thematically continuous features. This term was first time used by Burrough *et al.* (1997) and de Gruijter *et al.* (1997). An example of the continuous soil map can be seen in chapter 7, page 158.

Conventional approach to soil mapping Conventional approach to soil mapping or conventional soil survey is a collection of the methods and systems applied by most of the soil survey teams in the World. It typically implies that aerial photo-interpretation is used to draw boundaries, double-crisp maps are used to present soil maps and key focus of the soil inventory are the soil classes.

Digital Elevation Model Map of elevations representing the Earth's surface (elevation is known at all locations of the study area). In this thesis, I only deal with gridded DEMs, i.e. raster maps.

Feature space Feature space also called state space, character space, property space or attribute space is a virtual space bounded by the range of a set of variables. Position of point measurements in feature space is related to the estimation of the prediction uncertainty and can be used to design sampling. A comparison between the feature and geographical space is available on page 16.

Mixed model of spatial variation A model that assumes both continuous and discrete changes (jumps or breaks) in the attribute values; it integrates both the continuous and discrete model of spatial variation.

Pedometric approach to soil mapping This is the new, (geo)statistical approach to the mapping of soils. Sort of a contradiction to soil taxonomies and conventional methods. For an overview of methods see McBratney *et al.* (2000, 2003).

Prediction error Prediction error is the estimate of the uncertainty of the prediction model. It is commonly expressed as the variance of the prediction error or the standard deviation of the prediction error, also referred to as the prediction error variance or prediction variance.

Regression-kriging Regression kriging is practically equivalent to Universal kriging (UK) or Kriging with External Drift (KED). All three methods should give the same predictions and prediction error. They differ, however, in the methodological steps used. See (Hengl *et al.*, 2003a) for a detailed comparison between the regression-kriging and kriging with external drift.

Soil Information System A thematic type of a GIS, specifically built to provide information on soils. A SIS includes: digital soil maps, soil databases with interpretations and manuals.

Soil variable This term is used as a generic name for all quantitative (measurable) and qualitative (descriptive) soil properties or characteristics. According to the WRB terminology (FAO, 1998, p. 13), one should make a clear distinction between the measured soil variables (soil characteristics) and inferred or compound soil properties.

Soil-landscape modeling A synonym for the use of terrain parameters to improve modelling (spatial prediction) of soil variables. Soil-landscape modelling is by some used as a synonym for the pedometric approach to soil mapping.

Spatial prediction or interpolation Spatial prediction is the process of estimating the target quantity (z) at a new, unvisited location (s_0), given its coordinates and interpolation data set. In GIS, we make predictions at all raster nodes or pixels in a new map. Hence, spatial prediction or interpolation is in fact a mapping process (Stein, 1991).

Terrain analysis or parameterization Terrain parameterization is a set of techniques used to derive terrain parameters from a DEM, i.e. a process of quantifying the morphology of a terrain. Terrain analysis is used as a general term for derivation of terrain parameters and their application (Hengl *et al.*, 2003b).

Terrain parameters Maps (or images) derived using some terrain analysis algorithm. Terrain parameters are commonly classified as geomorphological (e.g. curvature of terrain), hydrological (e.g. wetness index) or climatic (e.g. insolation).

Contents

Chapter 1

Pedometric mapping

"I am a pedomagician!!"

[by A. McBratney in Pedometron #14 "Pedometrics in a sentence", available via www.pedometrics.org]

1.1 Soil mapping

Soil mapping or soil survey is *a process of determining the spatial distribution of physical, chemical and descriptive soil properties and presenting it in an understandable and interpretable form to various users* (Beckett, 1976; Dent & Young, 1981). In general, it consists of the following steps:

1. Project planning;
2. Preparation for fieldwork;
3. Photo-interpretation and pre-processing of auxiliary data;
4. Field data collection and laboratory analysis;
5. Data input and organization;
6. Presentation and distribution of soil survey products.

Project planning is an especially important step for the success of a soil survey project as it includes the selection of sampling plan, inspection density, classification system and data organization system. The preparation for fieldwork typically includes literature study and reconnaissance surveys. The end product of a soil mapping project is a soil resource inventory, i.e. a map showing distribution of soils and its properties accompanied by a soil survey report (Avery, 1987; Rossiter, 2001).

In the age of information technologies, the soil resource inventory data is organized into a thematic type of a geoinformation system (GIS) called a **Soil Information System** (SIS), the major part of which is a **Soil Geographical Database** (SGDB) (Burrough, 1991). This is, in most cases, a combination of polygon and point map linked with attribute tables for profile observations, soil mapping units and soil classes. Often the soil mappers extend their expertise to land use planning and decision-making activities, so that a SIS not only offers information on soils but also on their potential (and actual) use, the environmental risks involved (e.g. erosion risk) and offers predictions of soil behavior on intended management.

Soil mapping projects differ in the inspection intensity levels, purpose and type of conceptual models used. Considering the intensity level, soil mapping projects typically range from small scale (1:100 K to 1:1 M) surveys to medium (1:50 K) and large scale surveys (1:25 K to 1:5 K or larger). With regards to the intended purpose, a soil mapping project can be classified as special purpose (commonly referred to as thematic) or general purpose. The first is completely demand-driven and focuses on a limited set of soil variables or a single soil variable, often ignoring soil boundaries and soil horizons. General purpose mapping is more holistic, but also more

complex, hence more expensive and often not affordable at large scales. The conceptual models of soils reflect the purpose of the mapping project: (i) special purpose mapping projects commonly follow the continuous model of spatial variation, thus geostatistical techniques are used to make predictions; (ii) general purpose mapping projects commonly rely on photo-interpretation and profile descriptions, following the discrete model of spatial variation.

Coping with soil variation has never been an easy task for soil surveyors. Soil variables vary not only horizontally but also with depth, not only continuously but also abruptly. In comparison with vegetation or land use mapping, soil mapping requires much denser field inspections. Moreover, soil horizons and soil types are fuzzy entities, often hard to distinguish or measure. The polygenetic nature of soils, in particular, has always been a main problem in description and classification of soils (White, 1997). In fact, many pioneer soil geographers have wondered whether we will ever be able to fully describe the patterns of soil cover (Jenny, 1941). The quality and usefulness of the polygon-type soil maps (area partitions) has for decades been subject of argument (Webster & Beckett, 1968). The technological and theoretical advances in the last 20 years, however, have lead to a number of new methodological improvements in the field of soil mapping. Most of these belong to the domain of the new emerging discipline — pedometrics.

1.2 What is Pedometrics?

Pedometrics, a term coined by Alex B. McBratney, is a neologism, derived from the Greek words $\pi\epsilon\delta o\varsigma$ [soil] and $\mu\epsilon\tau\rho o\nu$ [measurement]. It is formed and used analogously to other applied statistical fields such as biometrics, psychometrics or econometrics (Webster, 1994). The most recent definition of pedometrics, available via the website of the Pedometric society (www.pedometrics.org), is:

> **"the application of mathematical and statistical methods for the quantitative modelling of soils, with the purpose of analysing its distribution, properties and behaviors"**

Pedometrics gathers together many different scientific fields, ranging from geostatistics to soil microbiology. The domain of pedometrics has changed somewhat since its foundation. At the moment, it is best defined as an interdisciplinary field involving soil science, applied statistics/mathematics and geo-information science (Fig. 1.1). The domain of pedometrics, however, is not limited to only these three general sciences, as McBratney stated in his first communication: *"It can include numerical approaches to classification — ways of dealing with a supposed determin-*

istic variation... the definition is certainly incomplete but as the subject grows its core will become well defined" (preface of Geoderma, 1994: 62).

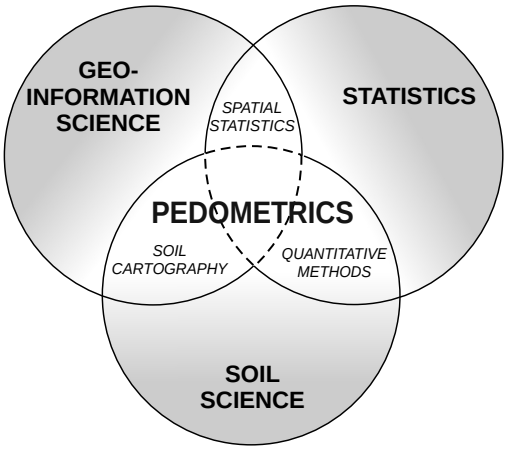

Figure 1.1: Pedometrics can be considered an interdisciplinary science where soil science, applied statistics and geoinformation science intersect.

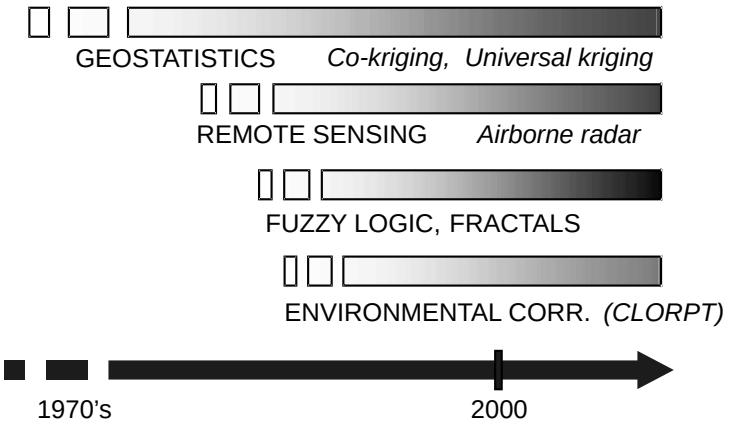

Figure 1.2: Some new emerging scientific fields that can be related to the development of pedometrics in the last decades.

Another way of looking at pedometrics is to see it as the implementation of newly emerging scientific theories, such as wavelets analysis and fuzzy set theory, in

soil data modelling applications (Fig. 1.2). The development of pedometrics is also a result of new technological discoveries and improvements, remote and close-range sensing techniques, GPS positioning and computers in general (Burrough *et al.*, 1994; McBratney *et al.*, 2003). The expansion of new applications in the early 90's has made pedometrics one of the leading sub-disciplines in the area of soil research (Hartemink *et al.*, 2002). Pedometrics is promoted and communicated via publications, conferences and workshops organized by the Pedometrics society, a working group under the International Union of Soil Sciences (IUSS). After a decade of existence and numerous conferences and workshops, this Working Group has been promoted, at the 17th World Congress of Soil Sciences, to become a Commission under the IUSS.

Recent topics covered by pedometrics include: multiscale data integration; the use of wavelets transforms to analyse complex variation; soil-landscape modelling using digital terrain analysis; quantification of uncertainty and fuzziness of information and evaluation criteria; soil genesis simulation; soil pattern analysis; design and evaluation of sampling schemes; incorporation of exhaustively sampled information (remote sensing) in spatial interpolation; precision agriculture applications and others. A major topic of pedometric research is the development of models and tools capable of dealing with the spatio-temporal variation of soils (McBratney *et al.*, 2000, 2003). These tools and methods can then be implemented to improve or replace conventional[1] soil mapping.

1.3 Pedometric mapping

Pedometric mapping is generally characterised as a quantitative, (geo)statistical production of soil geoinformation, also referred to as the *predictive soil mapping* (Scull *et al.*, 2003) or *digital soil mapping* (McBratney *et al.*, 2003), as it depends heavily on the use of information technologies. Pedometric mapping, however, specifically means that quantitative methods are used in the production of soil geoinformation.

In recent years, digital soil mapping had to encompass the rapid development of new and economic methods, mainly due to the increasing sources of auxiliary maps. Here, two main groups have played a key role: terrain parameters and remote sensing images (Dobos *et al.*, 2000). The terrain parameters are DEM-derived products that can be used to quantify the (geo)morphology of the terrain, i.e. accumulation and deposition potential, or to adjust the influence of climatic factors on the local terrain, while the remote sensing images reflect surface roughness, colour, moisture content and other surface characteristics of soils.

[1] See the definition of terms at the beginning of the book used consistently throughout the thesis.

Although it was originally expected that remote sensing would revolutionize soil mapping, as it had done for vegetation mapping, the direct derivation of soil properties from remote sensing data is still limited to areas of low vegetation cover, such as grasslands, semi-deserts or agricultural plots in fallow. Apart from some specific cases, such as using radar images to map soil moisture content (Hu *et al.*, 1997), it has not yet proved possible to use images of visible and infrared part of spectra directly to map soils in all parts of the study area. This is due to the complex illumination structure caused by terrain, cloud interference and atmospheric attenuation, or reflectance of vegetation (Skidmore *et al.*, 1997; Moran *et al.*, 2002). However, compound indices such as NDVI, which generally reflects biomass status, have been shown to correlate well with the distribution of the organic matter or epipedon thickness (McKenzie & Ryan, 1999). Even the coarse (1×1 km) AVHRR data have shown to be useful for mapping the clay content, CEC, EC or pH (Odeh & McBratney, 2000). A logical further development was to combine DEM-derived and remote sensing data to improve prediction models (Dobos *et al.*, 2000). The use of terrain data and remote sensing imagery has been especially interesting for medium scale-surveys (grid resolutions from 20–200 m), although there have also been an increasing number of field-site (precision agriculture) studies (Fig. 1.3).

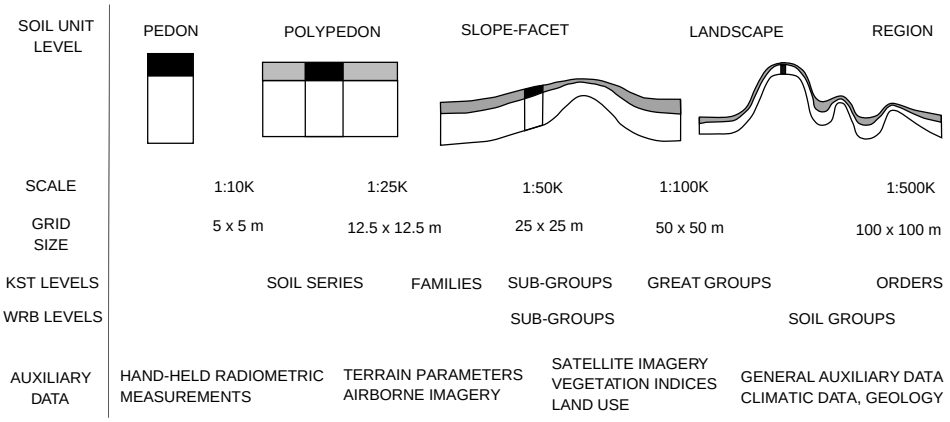

Figure 1.3: Relationship between the level of soil objects, scale, grid resolution and auxiliary maps used. Corresponding classification levels for Keys to Soil Taxonomy (KST) and World Reference Base (WRB) are also given.

There is a significant difference between the conventional and pedometric approaches to soil mapping. For a long time, the term *pedometrics* has been used as a challenge or contradiction of soil taxonomies, i.e. traditional systems. The key

differences between the two approaches are summarized in Table 1.1. The conventional soil survey relies on photo-interpretation and prediction of soil types, while pedometric techniques are (still) primarily focused on mapping individual soil variables at larger scales. The conventional survey typically leads to a polygon-based soil map whereas the products of pedometric techniques are fine-grained maps of soil properties.

Table 1.1: Comparison of pedometric and (analogue) conventional approach to soil survey.

	PEDOMETRIC APPROACH	CONVENTIONAL APPROACH
Preparation and project planning	Identification of key soil environmental variables (predictors)	Identification of key soil-forming factors (e.g. Catena concept)
Production of auxiliary data (pre-processing)	Remote sensing images; terrain parameters derived from a DEM; geological data etc.	Photo-interpretation; reconnaissance survey
Sampling design	Design-based (random sample, stratified random sample) or model-based (equal area stratification) sampling	Free survey
Field data collection and laboratory analysis	Navigation to points using a mobile-GIS (GPS receiver attached to a palm PC)	Navigation to points using aerial photos
Data input and organization	Data analysis and interpolation using some (geo)statistical technique	Designation of soil mapping units and theirs composition
Presentation and distribution of soil survey products	Fine-grained maps of soil variables with estimate of uncertainty (thematic mapping)	Polygon map with attributed soil properties (averaged)

1.4 Motives for the research

In recent years, there have been strong moves towards quantifying soil data: *"there has been corresponding increase in the demand for quantitative information at finer and finer resolutions"* (McBratney *et al.*, 2000). Even in the USA, surveyors anticipate a full transition to a quantitative (pedometric) survey in the 21st century (Indorante *et al.*, 1996). However, despite many appeals to abandon the conventional approach to mapping with mapping units, this approach remains more popular with most soil survey agencies. There are two probable reasons why pedometric techniques are still in the testing phase throughout most of the World. First, is the lack of

systematic knowledge about soil variability, as emphasized by Burrough (1993a): *"In spite of a huge research literature, knowledge about soil variability is still dispersed and not well organized. There is a need to organize and systematize our knowledge on soil variability in such a way that users of soil information unskilled in geostatistics and chaos theory can make the best possible decisions under conditions of uncertainty."* Secondly, pedometric techniques are still inappropriate to model such specific soil features as irregular soil stratigraphy, buried horizons, abrupt transitions between soils, fossil or karstic soils. These soil features and processes are still much easier to map (and generalize) using a mental model and photo-interpretation rather than geostatistics or auxiliary variables. Conventional soil mapping and classification have proven to be successful and popular, especially in the U.S. and Canada, where even the local farmers recognize different soil series. Experienced surveyors also find no need to change these systems. What is clearly needed is a compromise between the new methods (pedometric approach) and experienced soil survey teams (the conventional approach) that will satisfy both groups.

The integration of pedometric and conventional methods for operational surveys has not yet received widespread consideration. De Bruin (2000) emphasized the importance of combining different mapping techniques, noting that there are *"disciplinary gaps between the different techniques."* Even within the pedometric approach, there are rather isolated techniques that need to be combined. A good example is the gap that still exists between the `CLORPT` techniques and geostatistics. As de Gruijter stated in the preface of the Pedometrics '97 — International Conference held in Wisconsin, USA: *"... the second major theme of the Conference focused on spatial prediction methods. It was clear that there were two (somewhat) distinct approaches... The first is the geostatistical... the second is what Alex McBratny called 'clorp(t) approach, named from Jenny's equation or environmental regression... The synthesis of these two approaches was not really discussed. This will be an area for much further research in Pedometrics".* Both `CLORPT` and plain geostatistics have their advantages and disadvantages (Table 1.2). A disadvantage of ordinary kriging, for example, is that it ignores spatial variation of environmental factors, e.g. relief. Moreover, conventional geostatistical techniques have been shown to be inefficient at smaller scales (Yost *et al.*, 1982). A drawback of the plain `CLORPT` techniques, on the other hand, is that they ignore spatial location of points and spatial autocorrelation of residuals.

Another conceptual gap in soil mapping is that between the human perception of soil types and true nature of soils. One solution to the hidden and 'fuzzy' nature of soils is to use conceptual models that are more general: *"In order to bridge the gap, soil distribution modelling should be based on a new classification paradigm: that of a fuzzy set theory"* (de Gruijter *et al.*, 1997). How can such a system be operational-

Table 1.2: Comparison of some aspects of the conventional geostatistical and plain regression-based spatial prediction approaches.

GEOSTATISTICAL APPROACH	`CLORPT` APPROACH
Requires spatial dependence	Requires correlation with the auxiliary data
Higher sampling density desirable	Lower sampling density desirable
Data-driven	Knowledge-driven
Stratification desirable	One model over entire area
Deals with geographical space	Deals with feature space
Aims at spatially correlated random part of variation	Aims at structural part of variation (drift or trend)
Requires stationarity	Requires non-stationarity
Kriging variance reflects a geometry of the point locations while ignoring environmental patterns	Prediction error reflects the 'distance' of the point locations in the feature space while ignoring theirs spatial location
Numerous input parameters such as lag spacing, variogram function model, limiting distance, interpolation method, anisotropy model etc. are required; the predictions are non-unique for the same data set	For linear regression, in general, no input parameters are required; predictions are unique for the same data set; however, functional relationship between the auxiliary maps and soil variables is unknown and might differ for similar datasets

ized for routine survey? Indeed, is a universal method that can handle any type of soil data possible? McKenzie & Ryan (1999) think that *"the development of models for spatial prediction that are quantitative, mechanistic and mathematical is almost an impossible task in routine survey."*, considering the natural complexity of soils and soil properties. At this level of technology and knowledge, the development of hybrid or semi-automated, semi-subjective expert systems that integrate the empirical surveyor's knowledge of soils with GIS tools is the only feasible solution. This thesis is an attempt to bridge the gaps between the empirical and automated methods and improve the practice of soil mapping by designing an integrative pedometric methodology.

There are also practical motives for developing a flexible mapping methodology that can incorporate existing data sets. In Croatia, some 10 K profiles were described, analysed and classified during the 70's, 80's and 90's (National soil in-

ventory). These data are still not used spatially for soil prediction. In recent years, there has been considerable efforts to improve the effective scale of the Basic Soil Map of Croatia to a regional level, in this case the county level. There are 20 counties in Croatia and if the methodology proves successful, soil geoinformation could be improved in detail and brought to the 1:100 K effective scale or even less (i.e. a field resolution of 20–50 m). Similarly, there is a large amount of high quality soil field data worldwide that could be improved if the methodology proves to be successful.

1.5 Objectives

The main objective of this research was to develop a methodology for pedometric mapping that can be used to bridge the gaps between the pedometric and conventional techniques and that can be used for operational soil mapping at various scales. Specific objectives, addressed more closely in each chapter, are:

- To develop a methodology for optimal point allocation in both feature and geographical space and to recommend sampling strategies for the general purpose survey;
- To develop a systematic methodology to remove artefacts and inaccuracies in the terrain parameters used for soil-landscape modelling;
- To enhance the use of terrain analysis for photo-interpretation in soil survey;
- To develop and test generic interpolation algorithms that optimally employ both correlation with auxiliary maps and spatial dependence;
- To provide a basis for the integration of soil expertise (soil classification, photo-interpretation) and pedometric methods (regression-kriging, terrain analyis, pedo-transfer functions);
- To suggest methods to derive suitable grid resolution and investigate issues of combining multi-scale data sources;
- To develop a methodology to visualise fuzziness and uncertainty of soil information and enhance production of the continuous soil maps;
- To develop a methodology to assess the adequacy of soil maps and investigate the problems related to the usability of soil maps;

1.6 Outline of the thesis

The thesis was produced as a compilation of seven research papers, all written by myself as the principal author. These papers have been submitted to international peer-reviewed journals and have either been accepted for publication or are in the review process. Although the content of the thesis chapters and submitted papers does not differ in essence, I have made some minor changes in the thesis for the sake of coherence and textual harmony. I have also reduced some sections in the original papers to avoid a thematic overlap and repetition of phrases and statements. The list of the seven research topics can be seen in Fig. 1.4. It should be noted that these are all primarily methodological and do not depend on a specific study or scale. The research chapters are preceded by a definition of terms and concepts used and general introduction to soil mapping and pedometric techniques. To avoid terminological confusion, readers should first refer to the definition of terms at the beginning of the book.

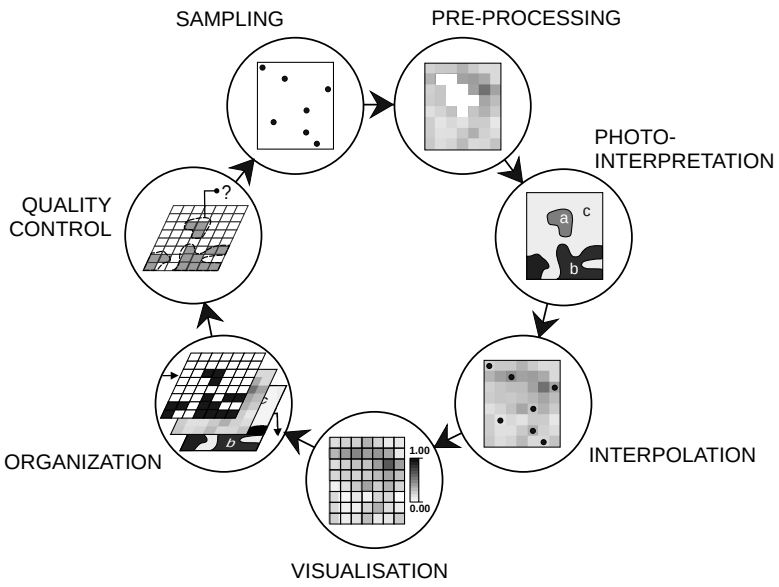

Figure 1.4: Schematic outline of the topics discussed in the thesis.

CHAPTER 2: SAMPLING This chapter gives a comparison of possible sampling strategies for the purpose of spatial prediction by correlation with aux-

iliary maps. This extends the existing sampling optimisation methodology to the issue of spreading in the feature space. The chapter demonstrates how allocation of points in the feature space influences the efficiency of prediction (overall prediction error). It suggests how to represent spatial multivariate soil forming environment; how to optimise sampling design for environmental correlation and which sampling strategies should be used for a general soil survey purposes. The concepts are illustrated using a 50×50 km study area in Central Croatia, four predictors (elevation, temperature, NDVI and CTI) and one target variable (organic matter in the top-soil).

CHAPTER 3: PRE-PROCESSING Because the pedometric mapping relies heavily on auxiliary maps, their quality plays an important role for the success of mapping. How do the inaccuracies and artefacts in auxiliary variables affect the prediction process and can these problems be reduced? In this chapter, systematic methods for reduction of errors (artefacts and outliers) in digital terrain parameters are suggested. These methods ensure more natural and more complete representation of the terrain morphology, which then also reflects on the success of spatial prediction. The Baranja Hill study area (3.8×3.8 km square) is used to demonstrate the effects of errors in the terrain parameters on mapping landform facets and predicting the thickness of the solum.

CHAPTER 4: PHOTO-INTERPRETATION Delineation of landform facets through the photo-interpretation is the key step in a conventional soil survey. It relies on subjective impressions of the terrain morphology and the mapper's experience of the specific study area. Can the subjective delineation of landform facets be improved with the help of terrain analysis? Moreover, should we aim at replacing photo-interpretation or search for a compromise solution? This chapter suggests a semi-automated method for extrapolating photo-interpretation from a limited number of study sub-areas to the whole area. The intention was to enhance and not to replace the mapper's knowledge and expertise. The map of landform facets was produced using nine terrain parameters for Baranja region (1062 km^2) in Eastern Croatia.

CHAPTER 5: INTERPOLATION This chapter considers the development of a flexible statistical framework for spatial prediction, that should be able to adopt both continuous and categorical soil variables. It suggests methods for dealing with non-normality of input data and multicollinearity of predictors. The logit transformation is suggested as a step to prevent predictions outside the physical limits. How well does this framework performs in real case studies and does it really improve the efficiency of prediction? The framework was

evaluated using the 135 profile observations of organic matter, pH and topsoil thickness from a 50×50 km study area in Central Croatia.

CHAPTER 6: VISUALISATION In conventional soil mapping, colours in the choropleth maps are typically selected following the human perception of soils. Continuous classification of soil classes, e.g. by using the fuzzy k-means, has been shown to have numerous advantages for mapping soil bodies. The result of continuous classification, however, is a set of membership maps that can be hard to visualise and manipulate at the same time. In this chapter, an algorithm is suggested to visualize multiple memberships and to analyse geographical and thematic confusion. Multiple memberships are visualized using the Hue-Saturation-Intensity model and GIS calculations on colours. This colour mixing was demonstrated using the landform classification of nine landform facets in the Baranja hill study area (3.8×3.8 km square).

CHAPTER 7: ORGANIZATION This chapter collates methods from chapters 2, 5 and 6. It answers the question of how to select a suitable grid size, how to aggregate and disaggregate soil information and what are the advantages and disadvantages of a grid-based SIS. Concepts, operations and organizational structure of a hybrid grid-based soil information system (SIS) are first described. The prediction maps are then made using both photo-interpretation and auxiliary maps, which ensures both continuous and crisp transitions. The grid-based SIS was produced using a soil survey data (59 profile observations) of Baranja hill and compared with a SIS produced using the conventional methodology.

CHAPTER 8: QUALITY CONTROL In this chapter, systematic steps are suggested to assess the effective scale, accuracy of soil boundaries, accuracy of map legends, thematic purity of mapped entities and overlap among the adjacent entities. This assessment was based on a number of control surveys including control profile observations and photo-interpretations. The adequacy and usability of soil resource inventories was assessed for the extensive National soil inventory in Croatia. This was done by:

- examining the average delineation area of six map sheets;
- comparing soil data from ten control profile observations with the original profile observations;
- examining the thematic overlap between the adjacent mapping units using the data from 2198 profile observations;
- evaluating the accuracy of soil boundaries and map legends using the three control survey sub-areas.

CHAPTER 9: CONCLUSIONS AND DISCUSSION In the last chapter general conclusions are given related to the questions posed above. This extends to a discussion on the limitations of this research, unexpected and conflicting findings. Finally, recommendations are offered for further research, emphasizing research problems in the area of pedometric mapping that still need to be tackled.

Chapter 2

Spatial prediction and feature space*

"I wish the surface of this soil wasn't so bloody rough... those fractals are a bit sore on the old legs... fat chance (probability) of these seeds being spread out evenly... I wonder what we could do about it... nothing, we're too interventionist anyway!"

[from a comical drawing in Pedometron #1, edited by A. McBratney, available via www.pedometrics.org]

*based on: Hengl, T., Rossiter, D.G. and Stein, A., 2003. Soil sampling strategies for spatial prediction by correlation with auxiliary maps. Australian Journal of Soil Research, Vol. 41(7), in press.

2.1 Introduction

A sampling design in soil survey specifies which points, transects, or areas will be visited for field measurements or observations. Sampling incorporates concepts of survey intensity, spatial variability and mapping scale, and is usually the most costly aspect of a survey (Webster & Olivier, 1990). Ideally, sampling should be as cheap as possible while consistent with the required level of accuracy and precision. In a conventional soil survey, sampling sites are selected subjectively by surveyors to support their mental predictive model of soil occurrence, a so-called free survey (White, 1997). Such designs are purposive and non-random, and do not provide statistical estimates. By contrast, a *pedometric* soil survey (McBratney *et al.*, 2000) aims at statistical modelling of the soil cover, including uncertainty about the predictions, using objective techniques.

In geostatistical applications, much attention has been given to optimisation techniques for sampling designs (Warrick & Myers, 1987; Odeh *et al.*, 1990; Brus & de Gruijter, 1997; van Groenigen *et al.*, 1999). This has not been the case for spatial prediction by environmental correlation. Here, several authors (Moore *et al.*, 1993; Bell *et al.*, 1994) have commonly applied the intuitively-appealing idea of placing samples at regular intervals along the steepest environmental gradients. An example is a toposequence with transects along the steepest slope, based on the concept of a hillside catena. Gessler *et al.* (1995) were among first to apply feature space stratification to sample evenly along the range of CTI (Compound Topographic Index). This principle can be extended to any environmental gradient, i.e. to multivariate gradients. McKenzie & Ryan (1999) used terrain parameters and geological and vegetation data to stratify an area and then randomly select samples inside the resulting patches. Lesch *et al.* (1995) developed an algorithm that combines model-based design with survey site spreading.

2.1.1 Feature and geographical spaces

Feature space (Lillesand & Kiefer, 2000) also called state space, character space, property space or attribute space, is not *'space'* in the geographic sense, but rather a virtual space bounded by the range of a set of variables. For multiple regression, the axes of the feature space are the soil-environmental variables or their transforms, which in the multivariate case form a hypercube. An important difference between the geographical and feature space is that the dimension of the feature space are on different scales. Points that are close in the geographical space can be far from each other in the feature space (and vice versa). Similarly, a study area has a different geometry when visualised in geographical and feature space. For example, a large but environmentally homogenous study area will occupy a small *'niche'* in

the feature space. If the predictors show normal distribution, the study area in the multivariate feature space forms a hypersphere.

2.1.2 Optimal point allocation for regression analysis

Let a set of observations of a soil variable z be denoted as $z(s_1)$, $z(s_2)$,..., $z(s_n)$, where $s_i = (x_i, y_i)$ is a location and x_i and y_i are the coordinates (primary locations) in geographical space and n is the number of observations. A discretized study area A, for example as represented in a grid-based ('raster') GIS, consists of m cells, which can be represented as nodes by their centres, such that $s_i \in A$. Let also the sampled auxiliary variables at primary locations be denoted as $q(s)$ and $Q(s)$ if considered at all nodes, with $\bar{q}$, s_q and $\bar{Q}$ and s_Q as the mean and standard deviation at primary locations and at all raster nodes respectively.

In the case of multiple regression, prediction at new, unvisited location (s_0) is made by the linear regression model (Odeh *et al.*, 1994):

$$\hat{z}(s_0) = \sum_{k=0}^{p} \beta_k \cdot q_k(s_0) \quad q_0(s_0) = 1 \tag{2.1}$$

where $\hat{z}(s_0)$ is the predicted or response variable, the β_k are model coefficients, the q_k's are auxiliary variables or predictors, i.e. their values at raster nodes or pixels of the map, and p is the number of predictors. The model coefficients are commonly solved using the ordinary least squares (OLS):

$$\hat{\beta} = \left(\mathbf{q^T} \cdot \mathbf{q}\right)^{-1} \cdot \mathbf{q^T} \cdot \mathbf{z} \tag{2.2}$$

where $\mathbf{q}$ is the matrix of predictors ($n \times p + 1$) and $\mathbf{z}$ is the vector of sampled observations. Prediction efficiency is quantified using the variance of the prediction error at s_0 is then (Neter *et al.*, 1996, p. 210):

$$\sigma^2(\mathrm{s_0}) = \sigma^2 \left\{\hat{z}(s_0) - z(s_0)\right\} = MSE \cdot \left[1 + \mathbf{q_0^T} \cdot \left(\mathbf{q^T} \cdot \mathbf{q}\right)^{-1} \cdot \mathbf{q_0}\right] \tag{2.3}$$

where MSE is the mean square (residual) error around the regression line:

$$MSE = \frac{\sum_{i=1}^{n} \left[z(s_i) - \hat{z}(s_i)\right]^2}{n-2} \tag{2.4}$$

and $\mathbf{q_0}$ is the vector of predictors at new, unvisited location. In the univariate case, the variance of the prediction error can also be derived using:

$$\sigma^2(\mathrm{s}_0) = MSE \cdot \left[1 + \frac{1}{n} + \frac{[q(s_0) - \bar{q}]^2}{\sum_{i=1}^{n} [q(s_i) - \bar{q}]^2}\right] = MSE \cdot [1 + v(s_0)] \qquad (2.5)$$

where v is the curvature of the confidence band around the regression line. This reflects the amount of extrapolation in the feature space (Ott & Longnecker, 2001, p. 570). It can be seen from Eq. (2.5) that the prediction error, for a given n (sampling intensity), depends on three factors:

1. Mean square residual error (MSE);
2. Spreading of points in the feature space $\sum [q(s_i) - \bar{q}]^2$;
3. 'Distance' of the new observation from the centre of the feature space $[q(s_0) - \bar{q}]$.

A common target of the sampling optimisation for spatial prediction is allocation of observations to minimise the prediction error (Lesch *et al.*, 1995; van Groenigen *et al.*, 1999). In this case, we are not only interested in minimising the prediction error at some new location, but in minimising the mean or overall prediction error, calculated at all raster nodes:

$$\bar{\sigma}^2 = \frac{\sum_{j=1}^{m} \sigma_j^2}{m} = MSE \cdot \left[1 + \frac{1}{n} + \frac{\sum_{j=1}^{m} (q_j - \bar{q})^2}{m \cdot \sum_{i=1}^{n} (q_i - \bar{q})^2}\right] = MSE\,[1 + \bar{v}] \qquad (2.6)$$

where $\bar{v}$ is the overall curvature and m is the total number of nodes. From Eq. (2.6), it can easily be shown that the overall curvature reduces to:

$$\begin{aligned} \bar{v} &= \frac{1}{n} + \frac{\sum_{j=1}^{m} q_j^2 - 2 \cdot \sum_{j=1}^{m} q_j \cdot \bar{q} + \sum_{j=1}^{m} \bar{q}^2}{m \cdot \sum_{i=1}^{n} (q_i - \bar{q})^2} = \frac{1}{n} + \frac{\frac{\sum_{j=1}^{m} q_j^2}{m} - 2 \cdot \bar{q} \cdot \bar{Q} + \bar{q}^2}{\sum_{i=1}^{n} (q_i - \bar{q})^2} \\ &= \frac{1}{n} + \frac{s_Q^2 + \bar{Q}^2 - 2 \cdot \bar{q} \cdot \bar{Q} + \bar{q}^2}{\sum_{i=1}^{n} (q_i - \bar{q})^2} = \frac{1}{n} + \frac{s_Q^2 + \left(\bar{Q} - \bar{q}\right)^2}{n \cdot s_q^2} \end{aligned} \qquad (2.7)$$

where $\bar{q}$ and s_q are the sampled mean and standard deviation of predictor and $\bar{Q}$ and s_Q are the mean and standard deviation of predictor at all raster nodes:

$$\begin{aligned} \bar{Q} &= \frac{\sum_{j=1}^{m} Q_j}{m} \\ s_Q^2 &= \frac{\sum_{j=1}^{m} \left(Q_j - \bar{Q}\right)^2}{m} \end{aligned} \tag{2.8}$$

Finally, it can be seen from Eq. (2.7) that, for a given data set (i.e. given n, m, $\bar{Q}$ and s_Q), the overall prediction error is minimised for $MSE \rightarrow \min$, $s_q \rightarrow \max$ and $\bar{Q} - \bar{q} = 0$. In other words, the prediction efficiency is controlled by success of fitting, the difference between the sample and population mean and between their variances as illustrated in Fig. 2.1a and b.

If the range of the feature space is $[-1, 1]$, Eq. (2.7) is minimised if half of the observations are taken at $q = -1$ and the other half at $q = 1$. This is the so-called *'minmax'* D-optimal design (Gaylor & Sweeny, 1978), here referred to as D1. It belongs to a group of experimental designs also known as response surface designs (Cochran & Cox, 1992, p. 335). If extended to a number of predictors (Fig. 2.1c shows a case with two predictors), it is also referred to as the first order central composite design. The D1 design is especially attractive for field surveys, as it will most likely reduce the number of samples and the spacing in-between them and therefore minimise the survey costs. However, it is optimal only if the model is linear. It is the worst possible design if the relation is quadratic. This is because it will give the worst estimates of the regression coefficients and therefore the lowest prediction accuracy at the validation points, as illustrated in Fig. 2.1d.

If the relationship is quadratic, the optimal response surface design is to allocate 25% of the observations each to the minimum and maximum and the remaining 50% to the central value, here called the D2 design (Atkinson & Donev, 1992). However, in the usual case when the functional relation between the predictor and target variable is unknown, designs such as D1 and D2 may perform poorly. Sampling is then designed to be resistant to the effects of an unknown model, even at the cost of inefficient estimation of model parameters. In the case the model is unknown, the most prudent design is sampling regularly along the feature space. This is often achieved by stratifying the area proportionally to the histogram of the predictor variable, also called *equal area stratification* (EA) design (Gessler *et al.*, 1995).

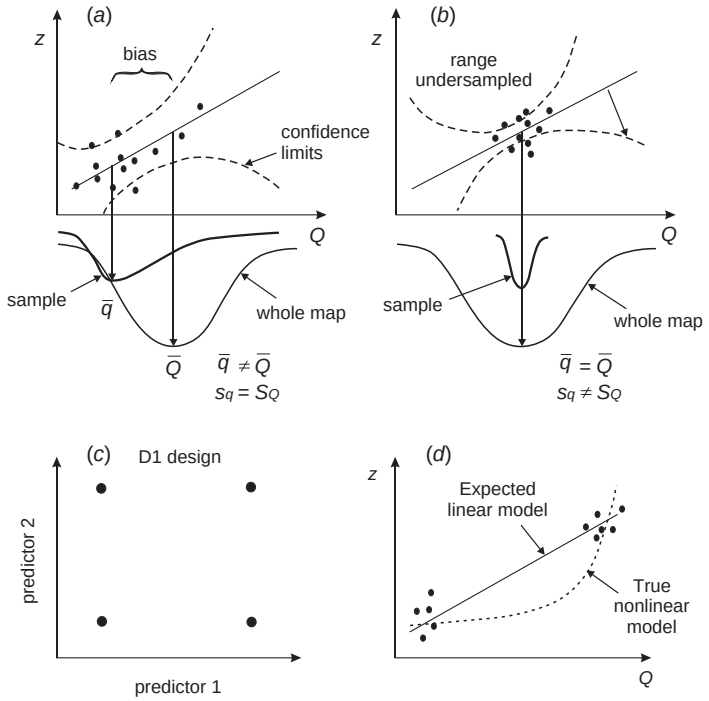

Figure 2.1: Sampling problems related to the feature space: (a) biased sample; (b) undersampling of total range causes extrapolation in most of the map; (c) D1 design in two-dimensional feature space and (d) D1 design can have a poor prediction for the whole map if the true model is actually non-linear.

2.1.3 Sampling optimisation and geographical space

The previous section showed that an optimal point allocation targets at increasing the spreading (variance) in the feature space and minimizing the difference between the sampled and population means. In the case of spatial prediction, however, the residuals may in addition show a strong spatial autocorrelation. Thus, estimation of regression parameters is over-optimistic and needs to be adjusted (Lark, 2000). General spatial prediction theory (Cressie, 1993, p. 166) states that trend model coefficients are optimally estimated using generalized least squares (GLS), i.e. by including the spatial correlation of residuals in estimation of coefficients (weighted regression):

$$\hat{\beta}_{gls} = \left(\mathbf{q^T} \cdot \mathbf{C^{-1}} \cdot \mathbf{q}\right)^{-1} \cdot \mathbf{q^T} \cdot \mathbf{C^{-1}} \cdot \mathbf{z} \tag{2.9}$$

where $\mathbf{C}$ is the covariance matrix of the residuals:

$$\mathbf{C} = \begin{bmatrix} C(s_1, s_1) & \cdots & C(s_1, s_n) \\ \vdots & \ddots & \vdots \\ C(s_n, s_1) & \cdots & C(s_n, s_n) \end{bmatrix} \tag{2.10}$$

$C(s_1, s_n)$ is the covariance between the ij'th point pair, estimated by modelling the variogram of regression residuals calculated by OLS estimation. Note that the variogram is first modelled using a semivariance function and then, for the reasons of computational efficiency, covariances are used. A flexible covariance function is, for example, the exponential:

$$C(\mathbf{h}) = \begin{cases} C_0 + C_1 & \text{if} \quad |\mathbf{h}| = 0 \\ C_1 \cdot e^{-\left(\frac{|\mathbf{h}|}{R}\right)} & \text{if} \quad |\mathbf{h}| > 0 \end{cases} \tag{2.11}$$

where $|\mathbf{h}|$ is the geographical distance between the point pairs and C_0, C_1, R are the covariance function parameters (Isaaks & Srivastava, 1989). The variance of the GLS prediction error is then:

$$\begin{aligned} \sigma^2_{gls}(\mathrm{s}_0) &= MSE + \mathbf{q_0^T} \cdot \left(\mathbf{q^T} \cdot \mathbf{C^{-1}} \cdot \mathbf{q}\right)^{\mathbf{-1}} \cdot \mathbf{q_0} \\ &= MSE + v_{gls}(s_0) \end{aligned} \tag{2.12}$$

and should be used instead of Eq. (2.3) to derive the mean prediction error. In the absence of spatial correlation, the covariance matrix ($\mathbf{C}$) reduces to the identity matrix:

$$\mathbf{C} = \begin{bmatrix} C_0 + C_1 & \cdots & 0 \\ \vdots & C_0 + C_1 & 0 \\ 0 & 0 & C_0 + C_1 \end{bmatrix} = (C_0 + C_1) \cdot \mathbf{I} \tag{2.13}$$

and Eq. (2.12) reduces to Eq. (2.3):

$$\begin{aligned} \sigma^2_{gls}(\mathrm{s}_0) &= MSE + \mathbf{q_0^T} \cdot \left(\mathbf{q^T} \cdot \frac{1}{(C_0 + C_1)} \cdot \mathbf{q}\right)^{\mathbf{-1}} \cdot \mathbf{q_0} \\ &= MSE \cdot \left[1 + \mathbf{q_0^T} \cdot \left(\mathbf{q^T} \cdot \mathbf{q}\right)^{\mathbf{-1}} \cdot \mathbf{q_0}\right] \end{aligned} \tag{2.14}$$

where $(C_0 + C_1) = C(0) = MSE$. $\mathbf{C}$ reduces to the identity matrix if the sampled are placed so that no pair is within the range of spatial dependence, in which case, OLS

estimation can be used instead of the GLS. McGwire *et al.* (1993) demonstrated that enforcing a minimum allowed distance between samples improves empirical models. Gessler *et al.* (1995) postulated that, in the absence of *a priori* information about soil attributes, the spatial dependence structure of the predictors can be used to derive the minimum distance at which the samples are spatially independent. However, this also assumes that the spatial dependence structure of a predictor is similar to the spatial dependence structure of the target variable, i.e. its residuals.

From Eq. (2.12), it follows that the D1 design might not give the minimum mean prediction error, even if the true regression model is linear, unless the sample points are spread outside the range of spatial dependence. This is difficult with a design that places all points at the extremes of the feature space range, since this will probably occupy a small portion of geographical space as well. If the covariance function of the residuals is unknown, the influence of $\mathbf{C}$ on the GLS prediction error is minimised if the samples are placed with a maximum geographical spreading, which leads to a grid sampling.

To evaluate the geographical spreading of the points, Mean of Shortest Distances (MSD) between point pairs can be used:

$$MSD = \frac{\sum_{i=1}^{n} \min_{j} (h_{ij})}{n} \tag{2.15}$$

where (h_{ij}) is the distance between two nearest point pairs. Note that in the geostatistical optimisation, the MSD to the equilateral triangular grid is more commonly used (van Groenigen *et al.*, 1999). Geographical spreading is then optimized if the MSD to the grid is minimized. The difference is that the maximisation of MSD to nearest point pairs will lead to outer points being pushed towards the borders of the region, while the minimisation of MSD to equilateral grid ensures that all points are spread equally within the study area. In this study we decided to use the MSD to the closest point pairs to emphasize the importance of spreading in the geographical space.

Finally, maximisation of MSD and feature space spreading may ask for adverse allocations, which means that the feature and geographical space criteria cannot be combined easily. In the absence of prior information on the spatial dependence structure of the residuals or knowledge of the nature of relationships, a sampling design that allows uniform spreading in both feature and geographical spaces is the safest strategy.

2.2 Methods

2.2.1 Study area and selected variables

A 50×50 km square in central Croatia (centred on 45°03'50" N, 15°17'39" E) was used as a case study. This is a relatively mountainous landscape, covered with coniferous and beech forests at the transition from the continental to Mediterranean Croatia (Fig. 2.2). The area is environmentally heterogeneous, which makes it especially attractive for spatial prediction by environmental correlation. The elevations range from 200 to 1400 m and the annual temperatures from 4 to 10° C. Four predictor variables following Jenny's (1980) conceptual equation of soil formation were selected: mean annual land surface temperature (LST), mean annual Normalised Difference Vegetation Index (NDVI), elevation (DEM) and Compound Topographic or wetness index (CTI), all at 1×1 km resolution. The LST map was calculated from the isotherm lines digitised from the climatic atlas of Croatia and adjusted up to ±1° C based on the aspect map. Mean annual NDVI was calculated for the year 1995 from a series of 36 NOAA AVHRR 1×1 km decadal images (USGS - NASA Distributed Active Archive Centre, 2001). The wetness index (CTI) was calculated based on the method of Quinn *et al.* (1991) using 60 iterations on a small-scale 1×1 km Digital Elevation Model (DEM). We used 100 measurements of topsoil organic matter expressed in % (OM), collected during the Croatian national soil survey in the 1980's (Bogunović *et al.*, 1998), as a target variable in regression modelling. The size of the dataset and variables used are typical for other similar environmental correlation applications (Moore *et al.*, 1993; Gessler *et al.*, 1995).

2.2.2 Uniform spreading in feature space - Equal range design

To achieve uniform spreading in feature space, stratification limits need to be set at equal distances in the feature space The weighting can be now done according to the histogram of predictor as in the EA design. We named this design **Equal range** (ER). The range of the predictor variable is divided into a number of equal-width clusters (also termed strata or histogram slices). The points are then randomly selected within each cluster following the given weights. For a normal distribution and five clusters, the stratification limits and weights can be calculated by dividing the standard statistical range of the normal distribution ($-3s_Q$ to $3s_Q$), which gives: 3.6%, 27.4%, 72.6%, 96.4% and 100% and the weights are: 0.036, 0.238, 0.452, 0.238 and 0.036 (Fig. 2.3a). The limits can be adjusted to any number of strata by calculating the percentage thresholds of the cumulative normal distribution.

Note that the ER design is in fact equivalent to the EA design in the sense that anywhere on the distribution a point has the same probability of being selected for

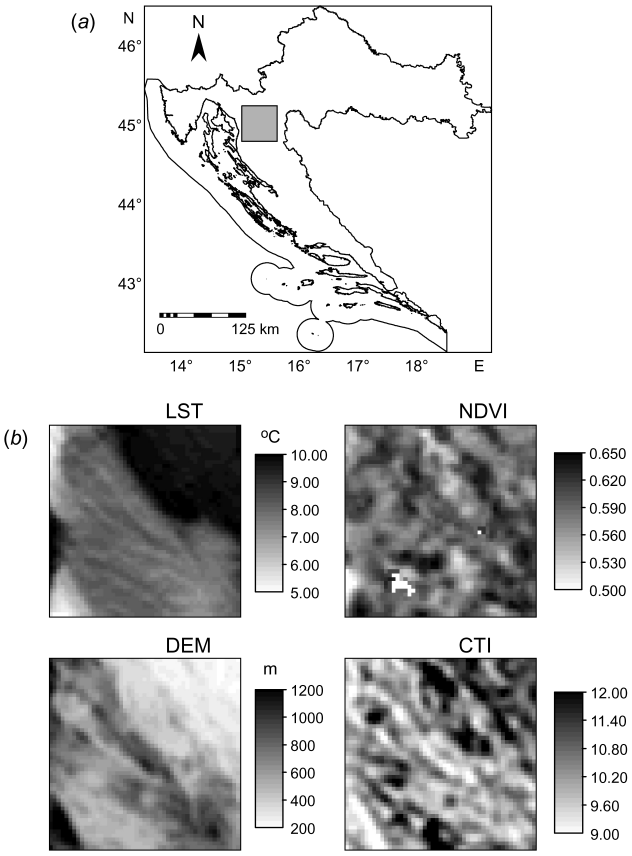

Figure 2.2: The 50×50 km study area: (a) location in Croatia and (b) selected predictors. LST - land surface temperature, NDVI - mean annual Normalised Difference Vegetation Index, DEM — elevation and CTI — Compound Topographic or wetness index. White patches in the NDVI map are lakes, i.e. water surfaces.

sampling. However, there is a key difference between the ER and EA designs (as proposed by Gessler *et al.* (1995)): the ER has different stratification limits and different weights. In particular, the tails of the distribution form strata, ensuring that some points are always selected from them, as in a D1 design. If the predictor shows a skewed distribution, the EA limits will largely shift towards one end of distribution and may by accident miss the tail. For the large number of sample points, however, the two designs will produce very similar results. If the predictors show a uniform distribution, the ER stratification limits are the same as in EA (e.g.

20%, 40%, 60%, 80% and 100%), i.e. the ER and EA designs are equivalent. We will hold to the term equal range, in further text, as it emphasizes the uniform spreading in the feature space.

An accurate method to determine stratification limits and weights, also used in this study, is to divide the range of predictor by the number of strata and then take the portion of the cumulative distribution between the limits as the weight. The ER design in the bivariate case (two predictors), in the case that all combinations of predictors are available (ideal conditions), is similar to grid sampling in geographical space (Fig. 2.3b).

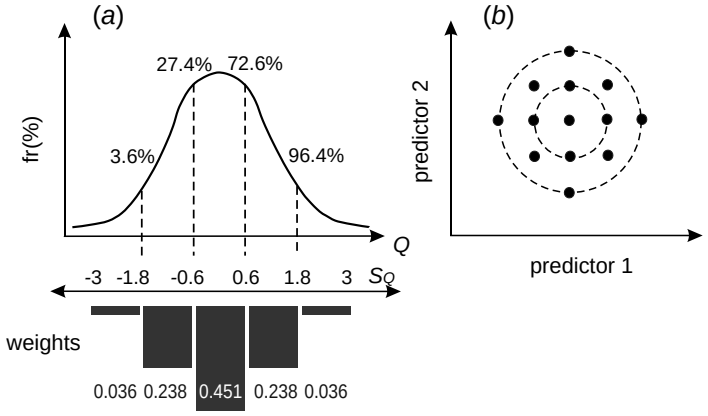

Figure 2.3: Equal range design (ER) with five strata: (a) histogram stratification with cumulative percentage limits and (b) 25 points allocated using two predictors (ideal case).

2.2.3 Comparison and evaluation of sampling schemes

We compared the ER and alternative designs by sub-sampling an existing set of 100 point observations from the original survey (ORIG). There were five set-ups in total: induced bias (here called Dx1), minimised spreading in the feature space (here called Dx2), D1, D2 and ER. We selected 25-point sub-samples of the 100 observation according to each design. For the Dx1 design we selected points at lower elevations only, and for the Dx2 design, around the mean elevation only. Because of the small sample size, some sub-samples did not correspond exactly to the theoretical designs. For example, in the case of D1 and D2, there were not enough points in the 5% tails or at the median. We then had to use the 25 points from the lowest and highest

elevations (approximation to D1), or around the median (approximation of D2), even though these occupy parts of the feature space outside the theoretical design. Therefore, this comparison of designs must be viewed as an approximation, with the advantage that it deals with a real dataset.

The mean overall GLS prediction error was calculated at all raster nodes (m=2500) for all designs (Eq. (2.12)). In this case, instead of using *MSE* for each subset, the residual error at the original 100 points (z_i^*) can be used to evaluate the true prediction error:

$$MSE^* = \frac{\sum_{i=1}^{n} (z_i^* - \hat{z}_i)^2}{n-2} \tag{2.16}$$

This means that the key evaluation issue is how close each design of a 25-point sample can come to estimating the original set. We used OM for response and DEM for predictor, as the reference model against which the various designs were evaluated.

The spatial dependence structure of the point data sets (soil properties and residuals) was modelled in VESPER using automated variogram fitting (Minasny *et al.*, 2002), in all cases with a lag spacing of 1 km, an exponential model, and a limiting distance of 25 km. Matrix calculations and regression analysis were done in the S-PLUS statistical package (MathSoft Inc., 1999).

2.2.4 Additional considerations

Since stratification of the feature space results in a large number of possible points in each cluster, there is an opportunity to run several randomisations and compare them for geographical spreading. We compared 10 simulations of the ER design with several alternative sampling strategies (random sampling, grid sampling, D1 and D2 designs) showing both the spreading in the feature space and geographical spreading (MSD) in a two-dimensional plot. In addition, we produced a transect ER design by allocating all points according to the ER feature space stratification on a single line in the direction of maximum contrast. The azimuth angle of maximum anisotropy was derived using the variogram surface function in ILWIS. The transect design has fairly small MSD and therefore is the most attractive realisation of the ER design considering the survey costs. Note that we did not make observations for the 10 simulations of the ER design, the transect design or the grid design. These were produced only to visualise the differences between the different sampling strategies in the context of feature and geographical space.

2.2.5 Multivariate case: Soil Predictive Components

When there are several predictors (the multivariate case), any of the stratifications proposed above must be adapted to multidimensional clusters. This is a highly relevant objective, since most realistic environmental correlations involve multiple predictors. This presents two problems. First, predictors are often significantly correlated, i.e. redundant in content, so that the dimensionality of the feature space is not as high as it first appears. Second, the final number of clusters obtained by crossing the several one-dimensional stratifications can easily be more than the intended number of sample points. We suggest the following sampling procedure for the multivariate case. To address multicollinearity, a principal component analysis can be used to produce uncorrelated Principal Components (PCs) (Neter *et al.*, 1996, p. 410). These are orthogonal and can be used instead of the original predictors to design sampling (Lesch *et al.*, 1995). In our case study, we first linearly stretched maps of continuous predictors in ILWIS to a dynamic range of 0–255 (8 bits). This set of maps was then transformed to PCs, yielding new synthetic 'bands' (Lillesand & Kiefer, 2000, p. 518), here named Soil Predictive Components (SPC). The SPCs were then stratified separately using the ER design as for the univariate case.

In addition, we partitioned the total number of sample points among the SPCs according to their proportion of the total variance calculated in the factor analysis. For example, since there were 100 new points to allocate, and SPC1 accounts for 64% of the variance, 64 points were assigned to SPC1. In each case, points were selected randomly within the each strata to ensure independence of sample measurements (Brus & de Gruijter, 1997). The randomly selected points for a single cluster may fall anywhere in the distribution of the other SPC. Because the SPCs are uncorrelated, the chance of overlap is minimised. We could not test this design for its predictive power, since we did not carry out actual soil observations and lab analysis. Thus we introduce these considerations to demonstrate how the ER design can be extended to the multivariate case.

2.3 Results

2.3.1 Regression models and spatial dependence structure

A plot of the relation between predictors and the target variable for all 100 observations showed the diffuse clouds of points, typical for soil-environmental relations (Fig. 2.4). OM showed a clear correlation with the selected predictors LST, DEM and CTI. All correlations, except those with NDVI as a predictor, are highly significant (p=0.01). Elevation (DEM) proved to be the most significant predictor of OM. Although we can infer a likely curve shape from the scatter plots, the true nature

of relationship between the predictors and soil variables is unknown. These noisy plots are typical for environmental data, where parsimonious models are suggested to avoid over-fitting the sample (Gauch & Hugh, 1993). As noted above, the correlation plot between OM and CTI shows a distinct curvature, which means that the prediction efficiency will be more sensitive to the sampling design (as illustrated in Fig. 2.1d).

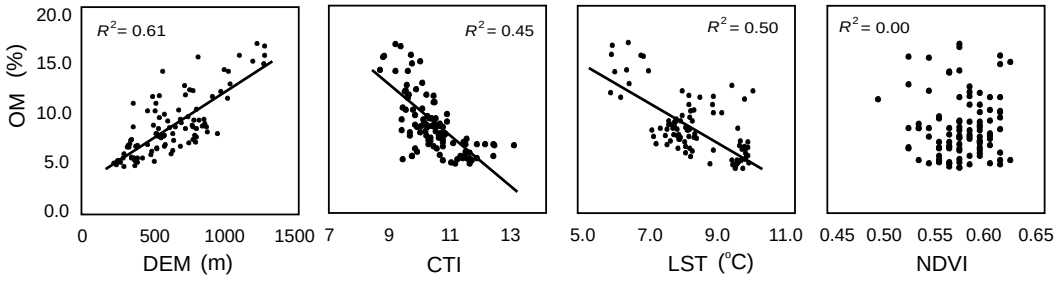

Figure 2.4: Observed ordinary least squares regression models for organic matter (OM) and their significance (R^2). DEM – elevation, CTI – Compound Topographic or wetness index, LST – land surface temperature and NDVI – mean annual Normalised Difference Vegetation Index.

In the case of predicting OM from DEM, OLS estimation gave:

$$OM = 2.96 + 9.10 \cdot 10^{-3} \cdot DEM \quad (R^2 = 0.61) \tag{2.17}$$

while GLS estimation gave a markedly higher intercept, lower slope, and a more realistic R^2:

$$OM = 4.96 + 5.89 \cdot 10^{-3} \cdot DEM \quad (R^2 = 0.53) \tag{2.18}$$

Automated variogram fitting for the DEM (used in the GLS estimation above) gave an unbounded variogram with a nearly linear shape in the radius of interest (Fig. 2.5a). The LST variable had a similar structure, while the variograms for NDVI and CTI showed much shorter range of spatial autocorrelation. The variogram surface of the DEM showed that the azimuth of highest anisotropy (shortest range) is 48.6°, i.e. northeast direction (Fig. 2.5b). Variograms of target variable (Fig. 2.5c) and residuals (Fig. 2.5d) from the reference model in Eq. (2.18) were both fitted by the exponential model, with a fivefold shorter range and threefold lower sill for the

residuals. The residuals showed spatial dependence to a distance of about 12.9 km (R=4.3 km), which implies that almost all points used in the regression modelling are spatially dependent. This confirms that geographical spreading has an effect on regression analysis and so is an important criterion for selection of the sampling design. Note also that the variogram model of a predictor might be quite different from the variogram model of the residuals, which means that the assumption made by Gessler et al. (1995) (see theoretical introduction) should be taken with care. The exponential variogram model (C_0=0, C_1=3.12, R=4.3 km) for the residuals from OLS was used as the reference model to calculate the overall prediction error for all sampling designs.

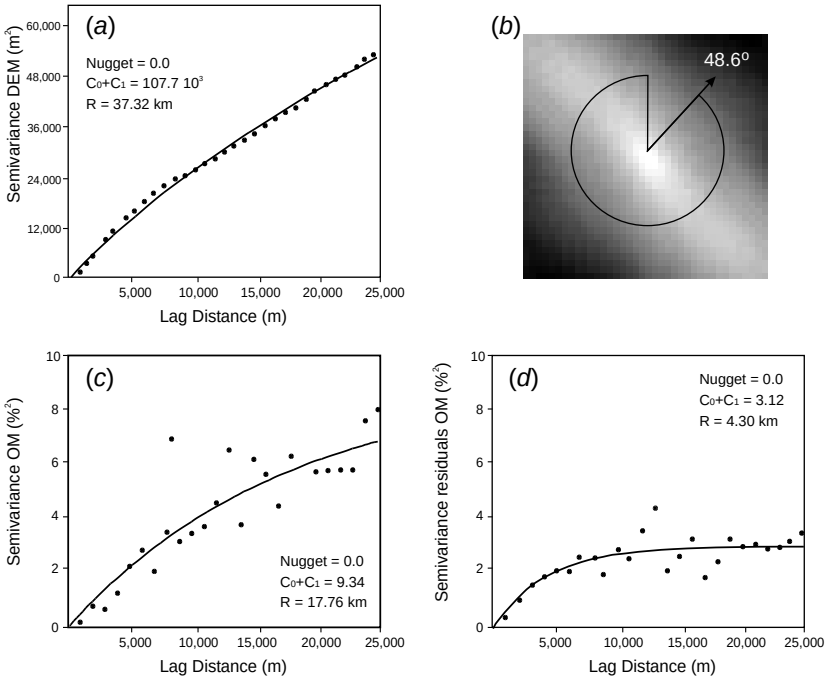

Figure 2.5: Geostatistical analysis: (a) variogram and (b) variogram surface of elevation map (DEM); (c) variogram of the target variable (organic matter, OM) and (d) regression residuals.

2.3.2 Comparison for prediction efficiency

A summary comparison of the designs is given in Table 2.1, and a visual comparison in Fig. 2.6. The Dx2 design, where the samples covered only the mean elevations gave, by far, the poorest overall prediction due to high spatial grouping and extrapolation in feature space. Similarly, the design with induced bias (Dx1) overestimated the β_1 coefficient and therefore the values in the areas of higher elevation (Fig. 2.6a). In general, maps of prediction error for D1, D2 and ER look fairly similar (Fig. 2.6b), although there are some differences. D1, the design with the highest spreading in the feature space, did not produce also the lowest overall GLS curvature of the confidence band: $\bar{v}_{gls}$=0.460 compared to 0.407 for ER design. This is due to the lowest MSD and the strong spatial correlation of residuals. Although the relationship is almost linear, which implies that the D1 design should be the optimal response surface design, the ER design gave a smaller overall curvature than the D1 design (Table 2.1). This agrees with our empirical assumption that ER is a good compromise between model estimation and geographical spreading. Also note (Fig. 2.6a) that ER came closest to estimating the reference model.

Table 2.1: Statistical comparison of internal properties and mapping efficiency for different sub-set designs and $OM = b_0 + b_1 \cdot DEM$ regression model.

Designs[a]	n	MSD	$\overline{q} - \overline{Q}$	s_q	b_0	b_1	R^2	$\overline{\mathbf{v}}_{gls}$	$RMSE^*$	$\overline{\sigma}_{gls}$
		(km)	(m)		-	(10^{-3})	-	-	(%)	
ORIG	100	2.52	24.8	266.8	4.96	5.89	0.53	0.196	1.95	2.18
Dx1	25	4.78	-191.7	113.5	2.78	10.1	0.33	1.42	2.02	2.35
Dx2	25	4.16	-13.2	52.0	8.98	-0.28	0.01	2.23	3.19	3.52
D1	25	3.41	84.6	447.6	3.01	8.85	0.90	0.460	1.96	2.07
D2	25	3.96	57.8	373.9	3.65	7.92	0.77	0.391	1.98	2.08
ER	25	5.43	-21.1	213.4	3.99	7.67	0.39	0.407	2.00	2.10

[a] n – number of observations; MSD – mean of shortest distances between the point pairs; $\overline{q} - \overline{Q}$ – bias between the sampled mean and mean for DEM calculated at all raster nodes; s_q – sampled standard deviation for DEM; b_0 and b_1 – GLS regression coefficients R^2 – coefficient of determination; $\overline{\mathbf{v}}_{gls}$ – overall curvature of the confidence bands; $RMSE^*$ – root mean square error (for estimating OM) between the predicted values and values at original locations; $\overline{\sigma}_{gls}$ – overall prediction error at all raster nodes; the summary statistics for DEM: $\overline{Q}$=599.3 and s_Q=232.8 m.

Because the D1 and D2 sub-sampled sets did not completely match the theoretical designs (Fig. 2.6a), differences between performances are less marked than

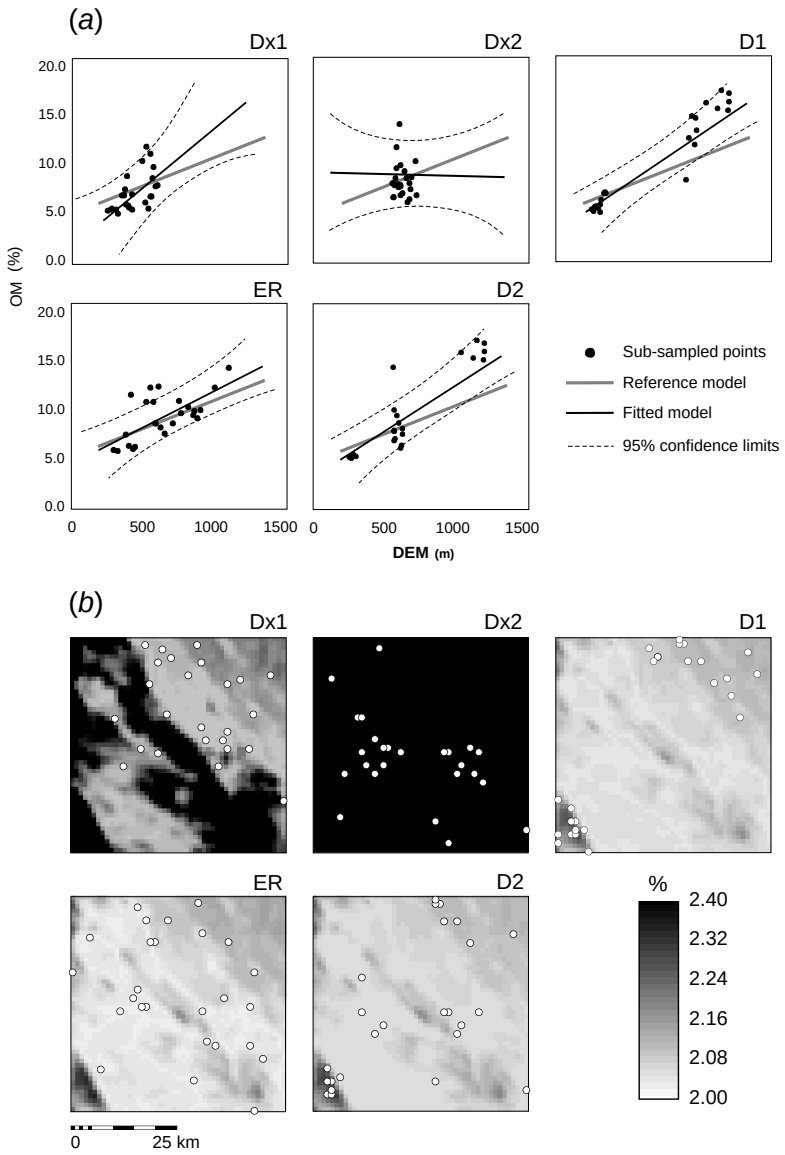

Figure 2.6: Comparison of different subsets based on different sampling designs - induced bias (Dx1), minimised spreading in feature space (Dx2), minmax design (D1), equal area stratification (EA), equal range (ER) and D2 design: (a) correlation plots showing fitted model of OM vs DEM with confidence limits and reference model estimated using all 100 points and GLS estimation; (b) standard deviation of the prediction error mapped at all raster nodes with geographical location of points.

expected by theory; in fact the D1 and D2 designs are quite close in performance to the ER design. Nevertheless, these sub-samples show some fundamental characteristics of the compared designs. For example, the D1 design achieves the best fit (R^2=0.90), but its points are clustered geographically (Fig. 2.6b), which finally raises the overall GLS prediction error so that it is somewhat inferior to the ER design.

2.3.3 Feature space and geographical space spreading

Summary comparison of different sampling strategies is shown in Fig. 2.7. We used a two dimensional plot with the feature space spreading (s_q) and the geographical spreading (MSD) as axes. As expected, designs D1, D2 and ER, show the highest spreading in feature space. Different realisations of the ER design will have different MSD values, although there is a limiting maximum spreading achievable within the ER strata. After few randomisations, an ER design with a higher spreading in both feature and geographical space than the existing soil survey was produced (Fig. 2.7a). An opposite strategy is a transect sampling along the steepest gradient, i.e. in the direction of the azimuth of highest anisotropy (Fig. 2.7b). On the other hand, the transect design, in this case (predicting organic matter from elevation), would give only sub-optimal estimation of model because of strong spatial autocorrelation between residuals.

2.3.4 Multivariate case

The predictor variables (LST, NDVI, DEM, CTI) were highly correlated, as shown by the large proportion of variance explained by first two SPCs (Table 2.2). The first component (SPC1) had approximately equal contributions from DEM, LST and CTI. The second component (SPC2) represented variation of biomass as estimated by NDVI. The third component (SPC3) reflected variation of CTI uncorrelated with DEM, whereas the fourth component (SPC4) reflected variation in DEM and LST. These SPCs form a orthogonal multivariate feature space of the study area. Fig. 2.8 shows the result of stratification of SPCs and the new points selected using the ER design according to the sampling plan (Table 2.3). Note that in some cases the number of points is not divisible by number of strata. For example, SPC4 receives four points, which need to be assigned to five strata. In such case we manually adjusted the number of points at the central class by adding or removing single observations to preserve the planned numbers (Table 2.3).

The points from the ER design with maximum spreading and the points from the existing survey are displayed in geographical and orthogonal feature space in Fig. 2.9. Both the new sample and the points from the existing survey had a similar spreading

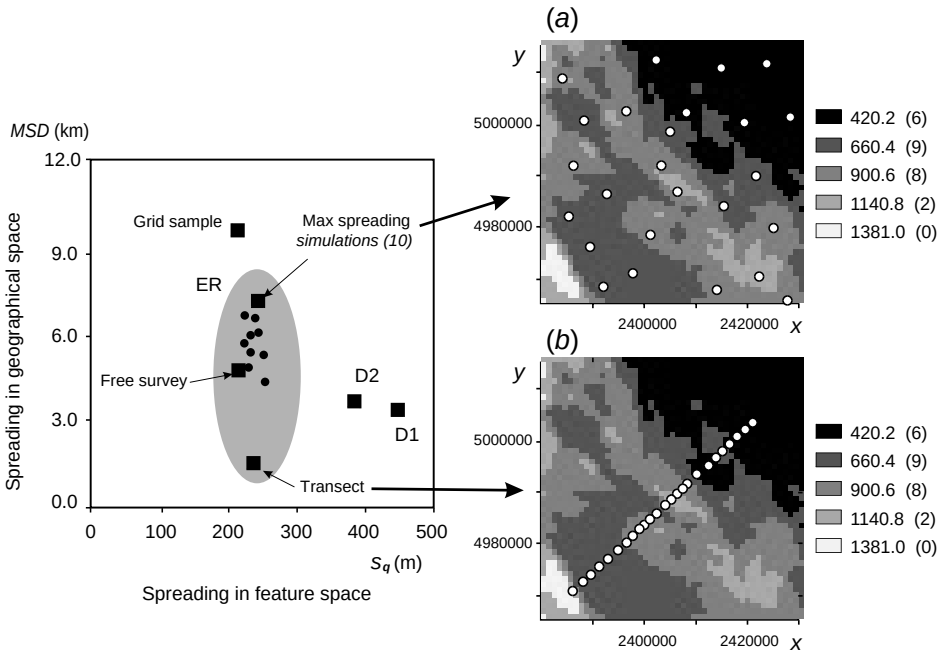

Figure 2.7: Comparison of different sampling strategies considering the spreading of the points in feature space (s_q) and geographical space (*MSD*): grid sample, free survey, D1, D2 and 10 simulations of the equal range (ER) design (left). Shaded area indicates assumed range of all possible realisations of ER design. Two different realisation of the ER design (25 points): (a) maximised spreading of points within strata and (b) minimised spreading by transect sampling consistent with ER feature space stratification (right).

in geographical space (Fig. 2.9a), with a MSD of 2.74 km and 2.52 km respectively, suggesting that the surveyors consciously spread their points to represent the whole area. Also when compared in the feature space using SPC1 and SPC2 as axes, ER and free survey (ORIG) show a higher spread towards the centre of the feature space cloud (Fig. 2.9b). Note that the total possible feature space, i.e. study area, spanned by SPC1 and SPC2, is limited due to a limited number of combinations between the predictors (Fig. 2.9b), which causes higher groupings in some areas. Consequently, the ER design does not look as similar to grid-sampling as we had expected.

Table 2.2: Loadings of the principal component analysis for four environmental bands and variance percentage explained.

	LST	NDVI	DEM	CTI	Variance explained per band (%)	
SPC1	0.567	-0.115	-0.637	0.509	64.1	64.1
SPC2	0.284	0.861	-0.199	-0.371	20.5	84.6
SPC3	-0.526	0.478	0.008	0.704	10.8	95.4
SPC4	-0.567	-0.127	-0.744	-0.329	4.6	100.0

Table 2.3: Allocation of points per strata: the distribution is based on the amount of variance explained by the factor analysis.

		Points per strata					
			Stratification limits (%)				
	Variance explained	n	L1	L2	L3	L4	L5
SPC1	64.1%	64	2	11	25	14	12
SPC2	20.5%	21	0	2	8	9	2
SPC3	10.8%	11	0	1	8	2	0
SPC4	4.6%	4	0	1	0	1	0
Total	100%	100	2	15	43	26	14

2.4 Discussion

2.4.1 The optimal design

In this chapter some theoretical concepts related to sampling optimisation by allocation in feature space were introduced and tested using a real data set. Our primary objective was to develop experimental sampling schemes that can be used prior to any data collection and extend the spatial prediction to the general case (GLS). We first used the model-based D-designs or response surface designs, which allow exclusion of most of the survey area. This may seem contradictory to field experience. For example, for the D1 design, it is not appealing to an experienced surveyor to put half of the points at the bottom of a slope and the remaining at the summit to determine the relation of organic matter to elevation. One reason is that we

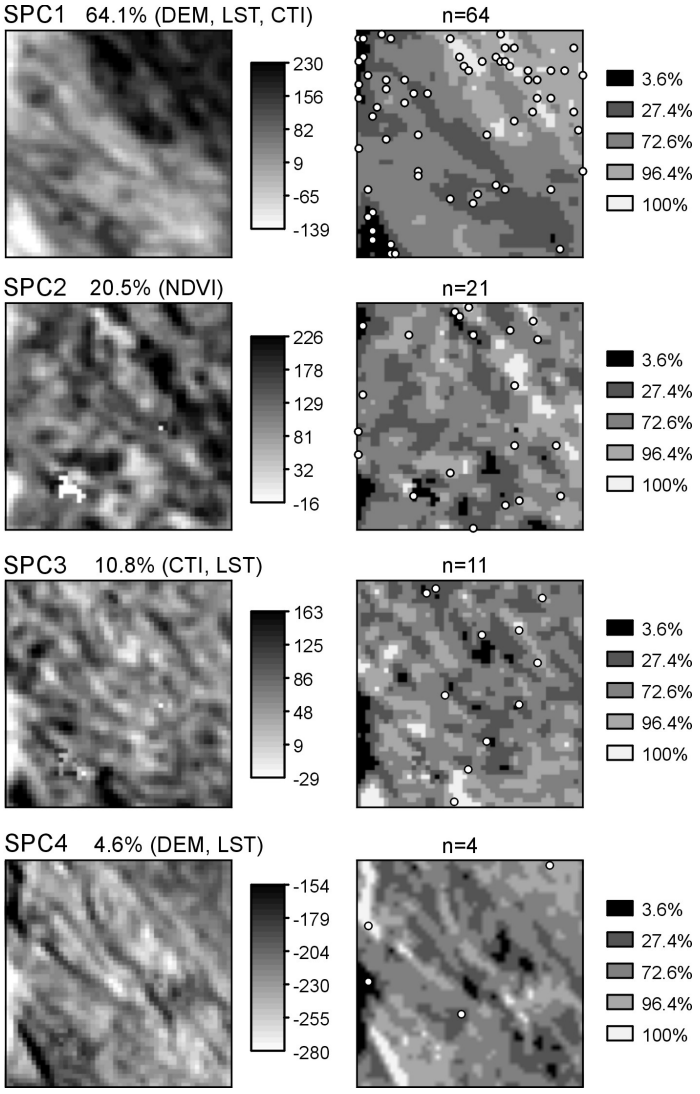

Figure 2.8: Soil Predictive Components (SPCs) from the principal component analysis, proportion of variance explained and most correlated environmental variables (left). Stratified SPCs based on the equal range design and 100 randomised point allocations spread along the four SPCs (right).

often expect that some other soil-forming factor, not correlated with the predictor variable, may vary between summit and footslope or that the relationship between the predictor and soil variable might be non-linear (as illustrated in Fig. 2.1d). This

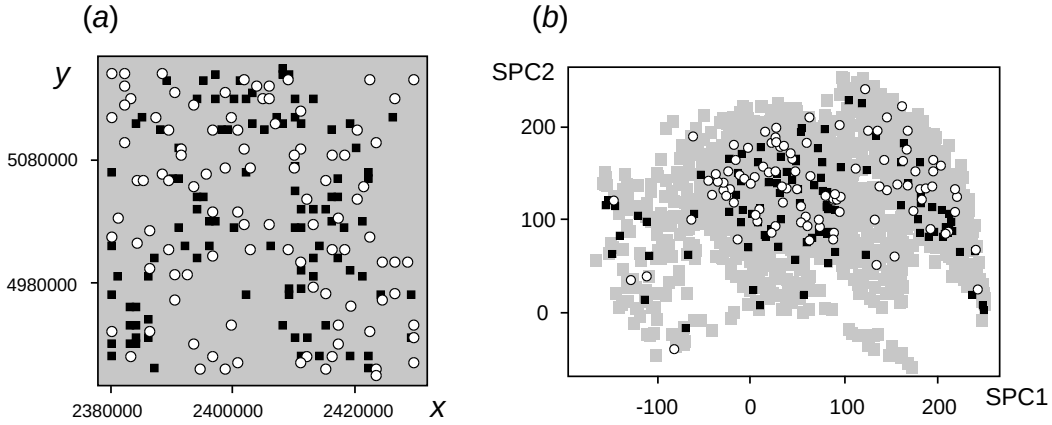

Figure 2.9: Profile locations from the existing survey (squares) and new simulated points using the equal range design (dots) displayed in geographical space (left) and in feature space spanned by SPC1 and SPC2 (right). The grey background indicates the study area.

means that even if a linear model is fit, the D1 design may result in a worse overall prediction.

Comparison of predictive power of designs for a univariate regression model in this study showed that D1 and D2 are indeed sub-optimal designs a strong spatial correlation between the residuals exists. This is because the points close to each other in feature space are often close in geographical space, since these narrow (in feature space) strata are also geographically small. On the contrary, the ER and EA designs achieved lower overall GLS curvature of the confidence band due to higher MSD between the points. In this case study, ER showed somewhat lower spreading in the feature space but the widest spreading in the geographical space.

This study also highlighted the importance of sampling optimisation in feature space, since serious bias or under-sampling of total variation (Dx1, Dx2) will result in poorer estimates of model coefficients, or higher overall curvature of the confidence band. Although one might argue that the Dx1 and Dx2 are obviously deficient designs, note that the Dx1 and Dx2 sub-samples did not have a poor geographical spreading at all (Fig. 2.6b). They could have be drawn out by simple random sampling for example.

Finally, the following four sampling principles, common for all compared designs, can be emphasized:

- Representation of feature space prior to sampling is important for overall prediction efficiency. Undersampling of the feature space may lead to poor esti-

mation of the model and high extrapolation in feature space (Dx1 example). A sample should cover the whole range of the feature space, so the extrapolation in feature space is minimised.

- Internal properties of predictors, histogram, range of spatial dependence and azimuth of maximum anisotropy can be used to design sampling prior to any knowledge on spatial variation of soil variables.

- The optimally placed sample is symmetrical around the central value of the feature space, i.e. shows a minimum bias between the sampled and the population mean of the predictors ($\bar{Q} = \bar{q}$).

- Maximisation of geographical spreading is important to represent areas (predictors) that are by accident overlooked and minimize spatial dependence between the observations. If the structure (range) of spatial dependence of residuals is known, we can be less strict. For example, if the residuals are spatially dependent at short ranges only, the samples can be placed at shorter distances and vice versa. In the case there is no prior information on soil variables and on functional relationship with predictors, one should aim at allocating the points in such way that they show both uniform spreading in (orthogonal) feature space and geographical space at the same time.

2.4.2 Equal range (area) or D-type designs?

Comparison of the ER (EA) and D-type designs showed that the ER design, when used for GLS estimation, is somewhat more appropriate for the spatial prediction of soil properties due to a more satisfactory spreading of points in geographical space and better estimate of the model coefficients. On the other hand, a drawback of ER design is that it gives more emphasis on the central values so even if the relation is close to linear, many points are not placed optimally in the feature space. Moreover, it appears that the EA or ER designs when used with a higher number of samples, does not have to necessarily differ much from a free survey or random design considering the spreading in the feature space. If the study area is stratified into equal areas and if equal weights are used, then the probability to select a location in the feature space is equal for the whole area (as with random sampling). It appears that the EA design, as used by Gessler *et al.* (1995, p. 426) and McKenzie & Ryan (1999, p. 78) could have been replaced with a simple random sampling.

When compared for the overall prediction error, however, the ER design proved to be justifiable. If the residuals are spatially correlated at shorter distances or if the plots do not show clear linear relationships, the safest design is to proportionally represent the feature space and allow higher geographical spreading. Open questions

are the relative performance of the ER designs with strongly non-normal predictors, and if higher-order polynomials are more parsimonious than the first-order linear models used in this paper.

The D-type designs seem to be attractive for cases where the relationship is linear or quadratic and where the geographical spreading can be satisfactory (e.g. in gilgai-type landscape samples can be placed at edges of the feature space but with satisfactory geographical spreading). A possible compromise between ER and D1 designs is to use a modification of the ER design with equal weights in all parts of range. Consequently, the sampled standard deviation (s_q) will be approximately 40% larger than the standard deviation of the whole map (s_Q), which is a desired property for minimisation of the prediction error (as explained in Eq. (2.7)).

2.4.3 Sampling along the multivariate gradient

In the multivariate case, we advocate a sampling procedure inspired by the intuitive idea of sampling in orthogonal multivariate feature space of predictors expected to represent soil-forming processes. The first step is the definition of key processes and variables according to concepts of soil formation in the study region. The second is data integration and stratification, which corresponds to the aerial photo-interpretation in a conventional soil survey. Here, we advocate the uniform spreading in feature space (ER design) and transformation of the predictive soil environmental maps to independent Soil Predictive Components (SPC). The samples can then be distributed using the proportion of variance explained by different SPC. The last step is the randomisation of points inside the clusters and selection of the randomisation with the maximum geographical spreading. In this case study we produced 10 randomisations and yielded a satisfying design with, simultaneously, a reasonable coverage of both feature and geographical space.

The development of a sampling scheme in multivariate feature space was more complex than in the univariate case, since the sampling points have to be selected simultaneously, i.e. represent a set of different predictors at the same time. Because of the large number of points in the initial set, there was no problem with randomly selecting points from 20 strata at the same time. Due to the high number of grid cells, the probability that the same point will be selected from the stratifications of two SPC at the same time was small. An alternative would be to make all possible combinations of clusters and then randomly select points inside these. The principle, used by McKenzie & Ryan (1999), was impractical in this case, since the number of combinations ($5^4 = 625$) would greatly exceed the total number of samples planned (100). This methodology can be adjusted for the general case where also the discrete predictors such as parent material are used. These, however, form strata a priori and cannot be processed together with continuous predictors.

Simultaneous analysis of feature and geographical space provides a basis for the development of sampling designs for hybrid interpolation techniques, such as kriging with external trend or regression-kriging (McBratney *et al.*, 2000). Moreover, it would be interesting to incorporate the proposed feature space criteria within the geostatistical optimisation algorithms such as simulated annealing.

Chapter 3

Reduction of errors in terrain parameters*

"Garbage in, garbage out!"

[a famous computer axiom meaning that if invalid data is entered into a system, the resulting output will also be invalid, available via www.webopedia.com]

*based on:

- Hengl, T., Gruber, S., and Shrestha, D.P., 2003. Reduction of errors in digital terrain parameters used in soil-landscape modelling. International Journal of Applied Earth Observation and Geoinformation (JAG), in review.
- Hengl, T. and Shrestha, D.P., 2003. Digital terrain analysis in ILWIS. Lecture notes, International Institute for Geo-Information Science & Earth Observation, pp. 56. available via: www.itc.nl/personal/shrestha/DTA/

3.1 Introduction

Digital terrain parameters, also known as topographic attributes (Wilson *et al.*, 2000) or morphometric variables (Shary *et al.*, 2002) are commonly derived from the digital elevation model (DEM) using some digital terrain analysis method. There has been an increasing interest in the use of relief data in the last decade accompanied by a growing availability of DEMs. The quality of terrain parameters is important as it directly affects the quality of spatial modeling. Several factors play an important role for the quality of DEM-derived products (Thompson *et al.*, 2001):

- terrain roughness and complexity;
- sampling density and DEM collection and interpolation method;
- grid spacing or pixel size;
- vertical resolution or precision and
- type and nature of algorithms used to derive terrain parameters.

Under slightly different input factors, e.g. coarser grid resolutions, vertical resolution or different filter algorithms, the terrain analysis can result in fundamentally different features (Wilson *et al.*, 2000). The importance of each factor, however, is usually driven by application-specific rules (Martinoni, 2002).

Terrain parameters are commonly used as auxiliary variables to improve spatial prediction of vegetation (Bolstad & Lillesand, 1992) or depositional/erosional processes (Mitasova *et al.*, 1996). A large group of terrain analysis applications is related to mapping and modelling of soil data. Terrain parameters are most commonly used as extensively mapped secondary or auxiliary variables to improve spatial prediction of soil-scapes and soil properties, such as thicknesses of horizons and other chemical (e.g. pH, organic matter) and physical (e.g. particle size fractions) properties (Moore *et al.*, 1993; Gessler *et al.*, 1995; McKenzie & Ryan, 1999). The application of statistical techniques for analysis of spatial distribution of soils using terrain and other environmental parameters is commonly referred to as *soil-lansdcape modelling*. McKenzie *et al.* (2000) gives an overview of applications for soil mapping. Dobos (2002) lists the most recent applications for regional scale soil mapping. In many of these applications, the errors in terrain parameters or terrain analysis algorithms are not considered as a quality control factor for successful soil-landscape modelling.

3.1.1 Errors in terrain parameters

For digital terrain analysis it is more important how well a DEM resembles actual terrain shapes and flow/deposition processes than what is the absolute accuracy of

the elevation values. This resemblance is often referred to as the *relative accuracy* of DEMs (Wise, 2000). Whereas absolute accuracy denotes the fit between the DEM and the real world, relative accuracy is a measure of the quality of DEM-derived products. The accuracy of terrain parameters is *"less a function of absolute accuracy of elevation values than of how well and how smoothly the landscape features are modeled"* (MacMillan *et al.*, 2000). Schneider (1998) introduced the term *"geomorphological plausibility"* to denote a compromise between the geomorphologic knowledge, sampled elevation data and interpolation techniques. In practice, field validation of accuracy of terrain parameters (e.g. hand measurements of slope, aspect and curvatures) has proven to be difficult due to the fractal nature of topography and abstract definition of many terrain parameters (Florinsky, 1998). The process of detecting and reducing errors is therefore somewhat different from detection of errors in remote sensing or other GIS data sources.

The errors in DEM and DEM-derived products can be roughly grouped in three types: (i) artefacts, blunders or gross errors, (ii) systematic errors and (iii) random errors or noise (Wise, 2000). Artefacts in terrain parameters are usually harder to detect than in the DEM, but they will certainly be visible in DEM-derived products. For example, interpolation of digitised contour lines using the linear interpolator will typically show artefacts in the slope and aspect maps (Burrough & McDonnell, 1998, Fig. 5.16). The most typical artefacts are so the called 'padi' or 'rice' terraces or cut-offs, which are absolutely flat. Although these are not visible in the DEM, the calculation of aspect or CTI fails due to division by zero, which finally results in part of the area being undefined. The padi terraces are somewhat similar to clouded pixels in remote sensing images, which suggests that similar geostatistical procedures (kriging or co-kriging) can be applied to remove them (Addink & Stein, 1999). Other common artefacts are 'ghost' lines or 'tiger stripes', which are obviously erratic features (Burrough & McDonnell, 1998). Systematic errors reflect the limitations of an algorithm and can be detected as local, unrealistic features or outliers.

Errors are especially common for terrain parameters derived using the higher order derivatives (curvatures), aspect map and hydrological parameters (CTI). Wise (2000) gives a comparison of different interpolation techniques and terrain analysis algorithms when applied in calculation of hydrological parameters. Thompson *et al.* (2001) evaluated the effect of the change in resolution on soil-landscape modelling and showed that with the increase of pixel size, spatial prediction of soil variables will be less discernible, while decreased vertical precision will typically show more erratic values. Wilson *et al.* (2000) emphasized the importance of the finer grid resolutions and flexible algorithms using a set of studies. Tang *et al.* (2002), showed that accuracy of DEM-derived hydrological data is directly related to DEM vertical resolution and terrain roughness. In the areas where the slope was less than

four degrees, the hydrological parameters were usually unreliable. Florinsky (1998) investigated the influence of different algorithms used to derive terrain parameters on the overall precision. Holmes *et al.* (2000) showed that local inaccuracies in the USGS 30 m DEM can be large and that the highest impact of the errors on terrain parameters is in valley bottoms.

In many cases, even simple smoothing of DEMs has proven to be beneficial in improving the quality of terrain parameters (Wise, 2000). Brown & Bara (1994) used low-pass filters in combination with analysis of spatial dependence to reduce outliers in elevation, slope and curvatures. MacMillan (2000) described a set filtering procedures to account for local noise and optimise terrain surfacing. Several other statistical image processing methods for reduction of errors have been proposed (Felicisimo, 1994; Lopez, 2000). In hydrological applications, quality of terrain parameters is usually improved by adjusting the interpolation to the existing network of streams and ridges or by removing the sinks. The automatic adjustment of DEMs has been implemented, for example, in the ANUDEM program (Hutchinson, 1989). However, ANUDEM and similar algorithms do not necessarily guarantee reduction of padi terraces, local outliers and other artefacts. There is still a need for flexible methods to improve plausibility of DEMs derived from contour lines. In addition, systematic methodology to quantify and reduce errors in number of morphometric and hydrological terrain parameters is lacking.

3.2 Methods

Let the elevation map be denoted as z or DEM, terrain parameters denoted as τ or TP and errors denoted as e, where z_i is the elevation value at ith grid location (z_1, z_2,...,z_n) and n is the number of pixels in a map. A realisation of elevation map is then denoted z^* and z^{*j} is the jth realisation of elevation map and filtered map is denoted z^+. Let also the derivation of terrain parameter from elevation be denoted as $\tau(z)$ or $TP(DEM)$ and local neighbourhood be denoted as z_{NB}. In a $k \times k$ window environment, $z_{NB\times}$ is the value of the central cell and z_{NBc} is the value at the cth neighbour of its k^2 neighbours. Commonly used window sizes are 3×3 and 5×5.

3.2.1 Detection and quantification of errors

Prior to the calculation of terrain parameters, it is important to first detect and reduce errors in the DEM (Wise, 2000). Padi terraces are areas where all surrounding pixels show the same value and can be defined as:

$$e \leftarrow \left[\ \forall c \quad z_{NBc} = z_{NB\times} \right] \tag{3.1}$$

Padi terraces are typical for closed contour lines and linear interpolators but can also appear when smoother interpolators are used. This happens because the hill tops, small ridges and valley bottoms are typically not recorded in the topo-map or no elevation value is attached to them. In a GIS, padi terraces can be detected using a neighbourhood operation.

The outliers can be defined as small, very unprobable features, which could have happen due to the gross error in the data collection method (very common for remote-sensing based instruments) or interpolation algorithm. They can be detected and quantified by using the statistical approach suggested by Felicisimo (1994). The probability to find a certain value within the neighbourhood is calculated by comparing the original elevation with the value estimated from the neighbours:

$$\delta_i = \hat{z}_i^{NB} - z_i \tag{3.2}$$

where δ_i is the difference between the original and estimated value and $\hat{z}_i^{NB}$ is the elevation (or terrain parameter) estimated from the neighbours. A statistically sound method to estimate the central value from the neighbouring pixels is to use the spatial dependence structure, i.e. predict the central value by kriging (Felicisimo, 1994). In a 3×3 window environment, there are only two types of distances (assuming the isotropic variation): in the cardinal (2,4,6,8) and diagonal directions (1,3,7,9) (Fig. 3.1a). An alternative is to use the 5×5 window size. Then there are 24 neighbours and five types of weights (Fig. 3.1b).

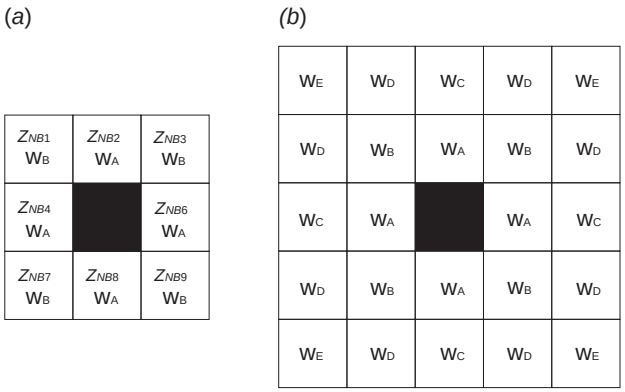

Figure 3.1: Schematic examples of filtering environments assuming isotropic model: (a) 3×3 window environment with common designation of neighbours and (b) 5×5 window environment with weights. The weights are used to predict the central pixel.

In a 3×3 window, the predictions are made by:

$$\begin{aligned}\hat{z}^{NB} &= w_{\mathrm{B}} \cdot [z_{NB1} + z_{NB3} + z_{NB7} + z_{NB9}] \\ &+ w_{\mathrm{A}} \cdot [z_{NB2} + z_{NB4} + z_{NB6} + z_{NB8}]\end{aligned} \tag{3.3}$$

where w_{A} is the weight in cardinal direction and w_{B} is the weight in diagonal direction. In general case ($k \times k$ window), the predictions are made by:

$$\begin{aligned}\hat{z}^{NB} &= \sum_{c=1}^{k^2} w_c \cdot z_{NBc} \\ \sum_{c=1}^{k^2} w_c &= 1\end{aligned} \tag{3.4}$$

where w_c is the weight at cth neighbour and $w_\times$ is the weight at the central pixel, so that $w_\times = 0$ and $\times = \frac{k^2+1}{2}$. Note that in the case of anisotropy, different weights can be used in different directions. The (kriging) weights are solved using the covariance function and relative distances between all pixels. Note that because we are only interested in the local spatial dependence, only first 10–15 surrounding pixels are considered for variogram modelling.

The difference between estimated and true value is calculated for each pixel to derive overall average and standard deviation ($\bar{\delta}$ and s_δ). Assuming a Gaussian distribution, Student's t test is used to standardise the differences by:

$$t_i = \frac{\delta_i - \bar{\delta}}{s_\delta}; \quad i = 1, ..., n \tag{3.5}$$

where n is the total number of pixels. Note that the overall average of differences should equal zero. The outliers (e) are then detected as:

$$e_i \leftarrow \left[|t_i| \geq t_{\alpha/2, n-2}\right] \tag{3.6}$$

For the two-tail 99.9% probability (a=0.01) t has value of 3.219.

3.2.2 Reduction of errors

Improving the plausibility of DEMs

Prior to actual filtering of terrain parameters, it is advisable to improving the plausibility of the DEM. First step in improving the DEMs derived from the contour data

is to account for features not shown by the contours such as break-lines indicating ridges or valley bottoms. This can be achieved by digitizing supplementary contour lines and spot heights indicating small channels, hilltops and ridges that are not indicated on the original topographic maps but can be inferred. The proportion of artefacts can be fairly high, especially in flat terrains, which means that the manual digitization can be a time consuming process. An alternative is the automated detection of medial axes between the closed contour lines. These are hypothetical ridges or valley bottoms, also called 'skeleton-lines' (Fig. 3.2a). First, the padi terraces need to be detected using Eq. (3.1). Then the medial axes can be detected using a distance operation from the bulk contour data (Pilouk, 1992). The new elevation is assigned to the medial axes between the closed contours by adding or subtracting some threshold elevation value, e.g. standard deviation of the elevation values (Hengl *et al.*, 2003b):

$$z_i^+ = \begin{cases} z_i + RMSE(z) & \text{if} \quad e = \text{terrace} \quad \text{and} \quad \tau = \text{convex} \\ z_i - RMSE(z) & \text{if} \quad e = \text{terrace} \quad \text{and} \quad \tau = \text{concave} \\ z_i & \text{otherwise} \end{cases} \tag{3.7}$$

where $RMSE(z)$ is the estimated accuracy of elevation measurements. For DEMs derived from the contour data, $RMSE(z)$ can be estimated from the contour interval h and local slope (Li, 1994):

$$RMSE(z) = B \cdot h + RMSE(xy) \cdot \tan\beta \tag{3.8}$$

where B is empirical number (commonly used is 0.16–0.33 range), $RMSE(xy)$ is the planimetric error and β is the local slope. In the case of padi terraces, the slope equals zero so the $RMSE(z)$ can be estimated directly from the contour interval. For example, if the contour interval is 10 m and B=0.25 then the $RMSE(z)$ is 2.5 m. Note that the adjustment of elevation is only done for hypothetical ridges (convex terrain) and valley bottoms (concave terrain). These additional lines are then added to the original contour data to re-interpolate the DEM.

The next step in improving the quality of the DEM is to reduce the outliers (Fig. 3.2b). These can be filtered using the parametric statistical method as explained in Eqs. (3.3) and (3.5). From the t value (Eq. (3.5)), we can derive the normal probability $p(t)$, which can be used as a weight function. The smoothed DEM can then be derived as a weighted average from the original DEM and estimated elevations:

$$z_i^+ = p(t_i) \cdot z_i + [1 - p(t_i)] \cdot \hat{z}_i^{NB}; \qquad p(t) \in [0, 1] \tag{3.9}$$

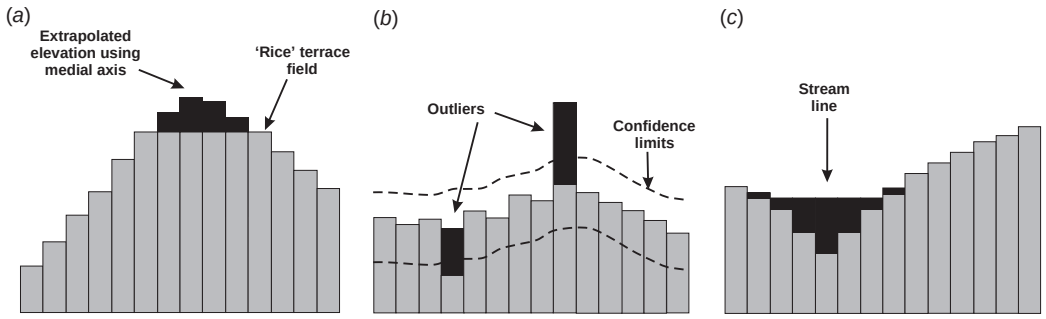

Figure 3.2: Schematic examples of DEM fitlering using cross-sections: (a) reduction of padi terrace fields; (b) reduction of outliers and (c) adjustment of the elevation using drainage lines. Black-coloured strips indicate the change in elevation values.

where z^+ is the filtered elevation map and $p(t)$ is the probability of exceeding a value estimated from the neighbours using the spatial dependence structure. The averaged elevation will be somewhat smoothed after the filtering for outliers. However, weak smoothing of elevation prior to terrain analysis is often recommended (Evens & Cox, 1999).

The last step in improving the geomorphic plausibility of a DEM is adjustment of elevations by incorporating the additional information, e.g. map of streams, water bodies and small channels. The streams (lines) and water bodies are first rasterized. A distance map (buffer) can then be used to calculate the DEM adjustment. We recommend the following formula:

$$\begin{aligned} \Delta z_i &= \left(\frac{p}{p+d_i}\right)^{\varphi} \cdot H \\ \Delta z_i &\in [0, H] \\ z_i^+ &= z_i - \Delta z_i \end{aligned} \tag{3.10}$$

where Δz_i is the adjustment of elevation, p is the pixel size, H is the maximum elevation difference, d_i is the distance from streams map and φ is the adjustment factor. This means that the original DEM will 'sink' proportionally to the distance from the streams (Fig. 3.2c). At the exact location of the streams or water bodies, the DEM will sink for the full value of H. The adjustment factor can be selected to reflect the field knowledge of relative local elevation differences. The maximum

elevation difference can be estimated from the field knowledge or an arbitrary small number can be used, e.g. half the contour interval h. This means that for $h = 10$ m the adjustment of elevation for < 5 m does not affect the original position of contours.

The suitable grid resolution can be estimated from the bulk contour data by using the total length of contours. As a rule of thumb, grid resolution (p) should be at least half the average spacing between the contours (Hengl *et al.*, 2003b):

$$p = \frac{A}{2 \cdot \sum L} \tag{3.11}$$

where A is the total size of the study area and $\sum L$ is the total cumulative length of digitised contours. Alternatively, the grid resolution can be estimated using cartographic standards. According to Tempfli (1999), the grid resolution should be optimally the maximum graphic resolution of lines shown on the maps, i.e. 0.4 mm at map scale. In the case of both estimating the pixel size and vertical resolution of the DEM, it is advisable to round down the numbers (the finer the grid size the better).

One can argue that the stream adjustment formula (Eq. (3.10)) can also modify elevations that are fairly far away from the streams. However, it can be shown that the adjustment of elevation affects only the local pixels. For example, for φ=1.5, p=25 m, H=5 m and vertical precision of 0.2 m, the stream-adjusted DEM differs for only first six neighbouring pixels from the streams (150 m in this case). The others pixels will stay practically unchanged (Fig. 3.3).

Finally, from the filtering steps explained above, four levels of DEM data can be distinguished:

1. (`DEM_L0`) the unfiltered DEM - derived from the contour data only;
2. (`DEM_L1`) terrace-free DEM - padi terraces are replaced by digitising ridges, peaks and sinks or by using automated extraction of medial axes;
3. (`DEM_L2`) smoothed DEM - filtered for the outliers and
4. (`DEM_L3`) streams-adjusted DEM - elevation adjusted for the streams and water bodies;

Note that these filtering steps can be applied regardless of the source, scale and quality of the input elevation data.

Reducing errors in terrain parameters

Even after the plausibility of the DEM has been improved, there can still be some remaining problematic features. Hence, filtering of errors in terrain parameters will

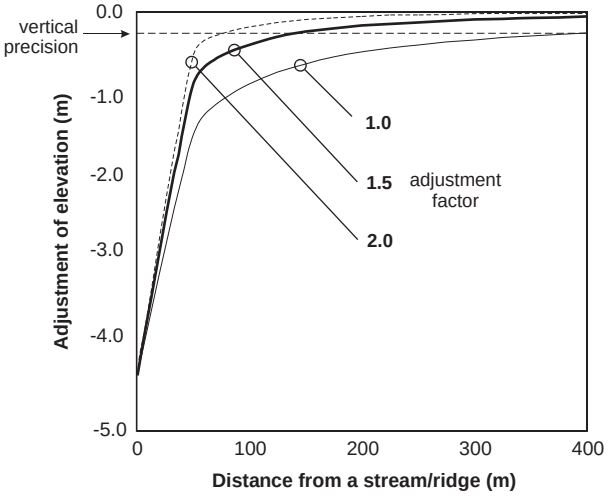

Figure 3.3: Elevation adjustment using different adjustment factors (φ).

also be necessary. The undefined pixels in terrain parameters can be filtered by iteratively replacing them using the predominant or average value from the neighbours:

$$\tau_i^+ = \begin{cases} \hat{\tau}_i^{NB} & \text{if} \quad \tau_i = ? \\ \tau_i & \text{otherwise} \end{cases} \tag{3.12}$$

where $\hat{\tau}$ is the terrain parameter estimated from the neighbours, e.g. using the kriging weights. A simpler solution is to take the average or the predominant value from the neighbours. For example, although CTI cannot be calculated in flat terrains due to division by zero, it can be estimated from surrounding pixels using filtering operation (extrapolation). One possibility is to approximate CTI by iteratively replacing the slope in the areas where it equals zero by averaging the neighbouring cells. This is done until all zero values are replaced with small values, which will have the effect of creating realistic pools of high CTI in the plain. The second possibility is to replace all small slopes with the threshold value, e.g. the smallest feasible slope. After all undefined pixels have been filled, terrain parameters can be filtered for outliers using the statistical approach as described in Eqs. (3.2) and (3.9). Note that each terrain parameter might show different structure of spatial variation from the elevation data, so that the modelling of a variogram is prerequisite.

In some cases, terrain parameters might not be undefined, but rather unrealistic. For example, aspect is extremely noise-sensitive in areas of very low relief. It

appears that it should be adjusted for the local slope. We recommend the following procedure. First, the aspect map needs to be converted to a linear scale (Beers *et al.*, 1966), e.g. the 'northness' map:

$$NORTH = |180 - ASPECT| \tag{3.13}$$

where $NORTH$ is the north-south aspect map where 0° means full northern orientation, 180° means full southern orientation and 90° means no orientation. The $NORTH$ map can now be adjusted for the slope using:

$$NORTH^{+} = 90 - (90 - NORTH) \cdot \left[1 - e^{-\frac{SLOPE}{RMSE_0(SLOPE)}}\right] \tag{3.14}$$

where $NORTH^{+}$ is the slope-adjusted northness map, $SLOPE$ is the slope map and $RMSE_0(SLOPE)$ is the estimated slope error in flat terrain. This is the precision of measuring slope in flat terrains. Note from the Eq. (3.14), in areas where slope tends to zero, the aspect exponentially tends to value of 90° (no-aspect). The slope error can be approximated from the $RMSE(z)$ and pixel size. For example, for the Evans and Young method (Florinsky, 1998):

$$RMSE(SLOPE) = \frac{0.41 \cdot RMSE(z)}{p \cdot (1 + G^2 + H^2)} \tag{3.15}$$

for $G^2 \to 0$ and $H^2 \to 0$, we get:

$$RMSE_0(SLOPE) = \frac{0.41 \cdot RMSE(z)}{p} \tag{3.16}$$

where G is the first derivative in x direction $\frac{\delta z}{\delta x}$ and H is the first derivative in y direction $\frac{\delta z}{\delta y}$. This means that, if $RMSE(z)$=2.5 m and p=25 m, the precision of measuring slope in flat terrains is 5% (3°).

Reducing errors by error propagation

Due to a high sensitivity of terrain analysis algorithms to local conditions, any single realisation represents only one view on terrain morphology. This is especially important for the calculation of hydrological parameters where we are more interested in the general picture of the processes. Even for the perfectly adjusted DEM, the location of the stream network can differ up to 3–4 cells from the true location (Burrough & McDonnell, 1998). A statistically robust approach to reduce the errors in terrain parameters is to average a set of possible realisations given the uncertainty in elevation values (Burrough *et al.*, 2000; Raaflaub & Collins, 2002). This is also referred to as the Monte Carlo method of error propagation (Heuvelink,

1998). The elevation values can be simulated using the inverse normal probability function (Banks, 1998):

$$\begin{aligned} z_i^* &= z_i + RMSE(z) \cdot \sqrt{-2 \cdot \ln(1-A)} \cdot \cos(2 \cdot \pi \cdot B); \quad i = 1, ..., n \\ A, B &\in [0, 1) \end{aligned} \tag{3.17}$$

where A and B are the independent random numbers within the $0 - 0.99\ldots$ range, z_i is the original value at ith location, is the simulated elevation with induced error and $RMSE(z)$ is the standard deviation of elevation values. The Eq. 3.17, however, will only induce noise in the original DEM and the spatial dependence structure of the simulated DEM will not be the same as the original.

In order to produce a realisation of DEM with similar spatial dependence structure (i.e. similar *smoothness*), point simulation needs to be used (Holmes *et al.*, 2000). It will produce a set of equiprobable realistic DEMs, each showing a similar histogram and variogram. Assuming gaussian spatial distribution of errors and for given $RMSE(z)$ and covariance function (C_0, C_1 and R), the realisation with same internal properties as the original DEM can be produced by simulating a point sample, inducing the error at point locations and then re-interpolating it over the whole area (Amstrong & Dowd, 1993). We suggest the following procedure for ILWIS:

(1) Randomly locate a set of points at locations α in the study area, so that the density of points corresponds to the original sampling density. In the case of contour data, average spacing between the contours can be used to estimate the original sampling density:

$$\upsilon = \left[\frac{p}{L}\right]^2; \qquad \upsilon \in [0, 1] \tag{3.18}$$

where p is the pixel size, and L is the average distance between the sampled points (contour data). Note that the sampling density is the key factor determining the smoothness of terrain. If the density of sampled points is high, it means that the terrain is more complex; if the density is low, the terrain is rather simple or smooth.

(2) At these locations, assign a random error using the inverse normal probability function and given $RSME(z)$ (Fig. 3.4a and b):

$$\Delta z^\alpha = RMSE(z) \cdot \sqrt{-2 \cdot \ln(1-A)} \cdot \cos(2 \cdot \pi \cdot B); \quad A, B \in [0, 1) \tag{3.19}$$

(3) Interpolate the error at all grid nodes using the same variogram function as for the original DEM (Fig. 3.4c and d):

$$\Delta z_i^* = \sum w_i^\alpha \cdot \Delta z^\alpha \tag{3.20}$$

(4) Add the error surface (Fig. 3.4e) to the original DEM:

$$z_i^* = z_i + \Delta z_i^*; \qquad i = 1, ..., n \tag{3.21}$$

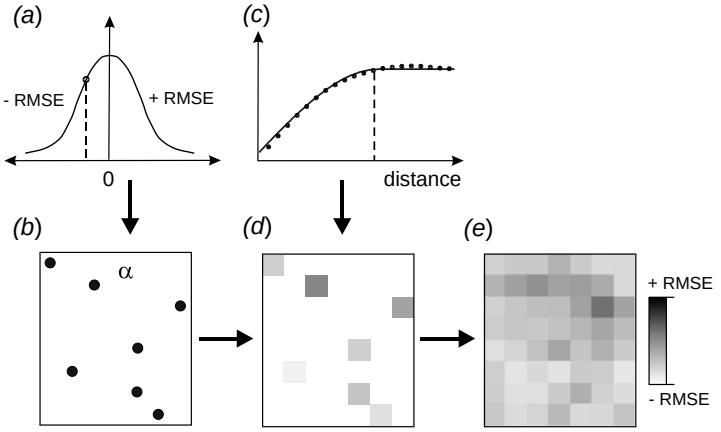

Figure 3.4: Simulation of an error surface: simulated error (a) is assigned to random locations (b) and then interpolated using the same variogram model (c,d) to produce a smooth error surface (e).

For each of the m simulated DEMs, terrain parameters are derived m times and then averaged per pixel:

$$\bar{\tau} = \frac{\sum_{j=1}^{m} \tau(z^{*j})}{m} \tag{3.22}$$

where $\bar{\tau}$ is the averaged map of a terrain parameter and $\tau(z^{*j})$ is the jth realisation of terrain parameter calculated from the simulated elevation map (z^*). The *RMSE* error of several simulations gives an estimate of the propagated uncertainty:

$$RMSE(\tau) = \sqrt{\frac{\sum_{j=1}^{m} \left[\tau(z^{*j}) - \bar{\tau}\right]^2}{m}} \tag{3.23}$$

Note that the map of propagated uncertainty can be used to depict problematic areas and digitize additional contours.

3.2.3 Study area

We used a small part of the Baranja hill located in Eastern Croatia (45°47'40" N, 18°41'27" E) to develop and test our methodology. This area is specifically suitable as it presents two contrasting landscapes: plains with terraces in the northwest direction and dissected hilland with small valleys. Contour lines were extracted from the 1:50 K topo-map, with the contour interval of 10 m and supplementary 5 m contours in areas of low relief. The total area is 13.69 km^2 and elevations range from 80 to 240 m. We digitised 127.6 km of contour lines (Fig. 3.5b), which means that the average spacing between the contours is 107 m and the pixel size should be at least 50 meters to present all mapped changes in relief. Considering the cartographic rule (the smallest distance of 0.4 m on the map), a grid spacing should be at least 20 m to satisfy this scale. Because the spacing between the contours is much narrower in the hill than in the plain, we finally decided to use a grid resolution of 25 m.

The contour lines were interpolated using the linear interpolator in ILWIS. The algorithm is described in more detail by Gorte & Koolhoven (1990). The contour interval was 10 m in hill and 5 m in the plain, hence we used the $RMSE(z)$ of 2.5 m. The precision has been setup to 0.2 m. From the interpolated DEM, we derived five terrain parameters: slope in % (SLOPE), profile curvature in rad/m (PROFC), plan curvature in rad/m (PLANC), Compound Topographic Index (CTI), and aspect, i.e. northness in degrees (NORTH). SLOPE, PROFC, PLANC and NORTH were calculated using the formulas by Shary *et al.* (2002), while the CTI was calculated based on the method of Quinn *et al.* (1991) using 20 iterations.

3.2.4 Evaluation and validation

We used two statistical measures to evaluate reduction of errors. The errors in DEM or DEM-derived data were quantified using the proportion (percentage) of errors in the total area:

$$A_e(\%) = \frac{A_e}{A} \cdot 100 \tag{3.24}$$

where A is the total area. This percentage was calculated for both padi terraces and outliers. In this case the outliers were detected as t-values exceeding the threshold value of 3.219 (Eq. (3.6)).

To validate the effect of reduction of errors on soil-landscape modelling, we used two applications. We first compared accuracy of classifying the landforms for unfiltered and filtered data. This was done by comparing an aerial photo-interpretation map with the results of supervised classifications (see chapter 4). Second, we used a data set of 59 soil observations of thickness of the solum (SOLUM). This is the

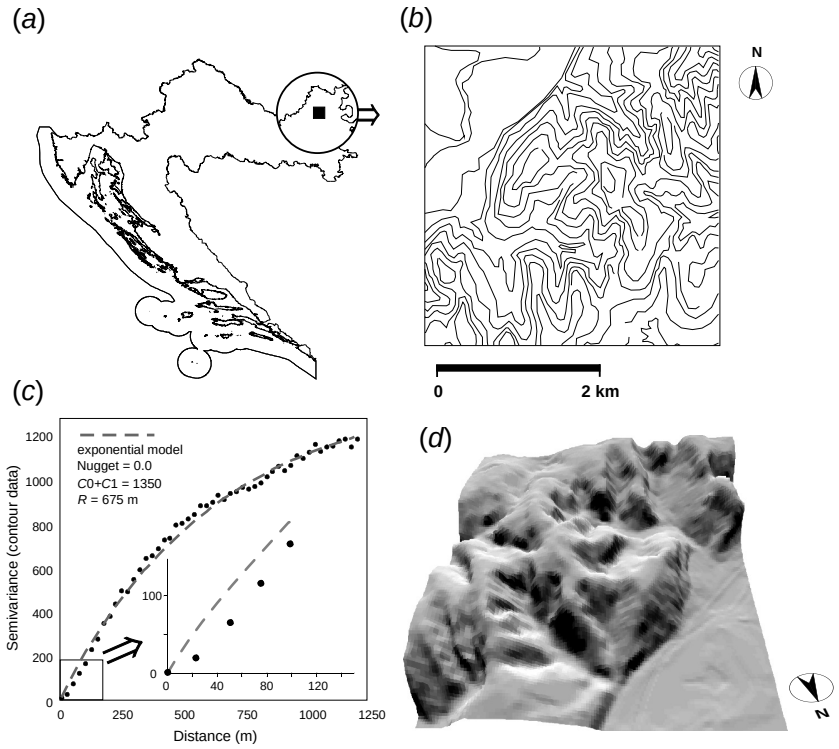

Figure 3.5: Study area: (a) part of Baranja hill, located in Eastern part of Croatia; (b) the contours lines; (c) variogram modelling of elevation with zoom on the local neighbourhood and (d) perspective view on interpolated DEM.

depth to parent material, in this case alluvial deposits and layers of loess. SOLUM was correlated with terrain parameters and mapped in the entire area. We observed the change in goodness of fit (R^2) for the unfiltered and filtered data.

3.3 Results

3.3.1 The plausibility of the DEM

First interpolation of bulk contour data resulted in 17.3% of the total area being represented with padi terraces (Fig. 3.6c), most of them located in the plain region (northwest corner). The automated extraction of medial axes detected hypothetical ridges and valley bottoms in 2.2% of the total area (Fig. 3.6a). After the second

interpolation using added medial axes, the proportion of padi terraces was reduced to 4.5% (Fig. 3.6c, `DEM_L1 - DEM_L0`). The biggest adjustment of elevation was in the plain region. The reduction of outliers (`DEM_L2 - DEM_L1`) did not contribute to the reduction of padi terraces. Finally, the proportion of the padi terraces was reduced to 2.2% (Fig. 3.6c).

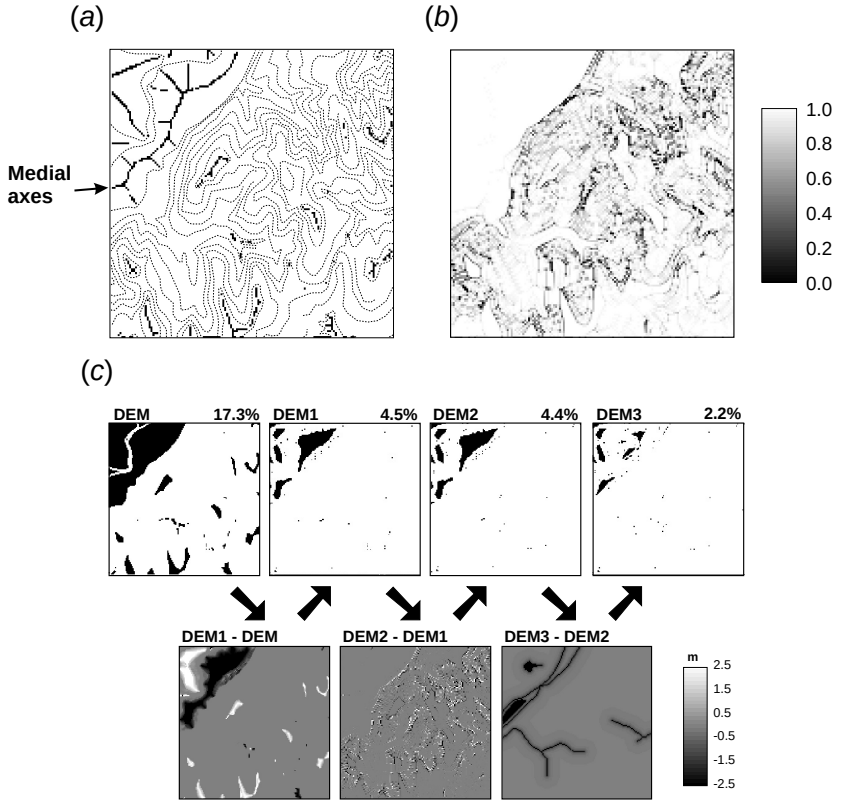

Figure 3.6: Semi-automated DEM filtering: (a) automated detection of medial axes; (b) normal probability of finding the elevation value within the given neighbourhood and (c) change in elevation values (above) and reduction of padi terraces in percentage (bellow). See text for more explanation.

The variogram analysis of the contour lines gave an isotropic exponential model with no nugget and fairly strong spatial dependence (Fig. 3.5c and Table 3.1). In the case of the 3×3 window environment, we calculated weights w_{A}=0.253 and w_{B}=-0.003. This means that this algorithm will give much higher importance to

the neighbours in the cardinal directions. For a comparison, the inverse distance interpolation would give weights w_A=0.146 and w_B=-0.104. For the 5×5 window environment, the prediction of elevation is still done mainly from the closest neighbours. The remaining weights (w_C, w_D, w_E) accounted for only 16.3% of the total cumulative weights (Table 3.1). This means that, although the elevation values are correlated at long distances, there is not a large difference between using the 3×3 and 5×5 filters. Comparison of the predicted and unfiltered values, showed that the difference is unbiased ($\bar{\delta}$=0), while the standard deviation (s_δ=1.28) was lower than the $RMSE(z)$. The values with the lowest probability occurred at locations where the density of contour lines was highest, e.g. at steep slopes (Fig. 3.6b). This confirms that the filtering of outliers has a smoothing effect. Note that the final change in elevation values, calculated as a difference between the old and adjusted elevation values does not exceed $RMSE(z)$ (Fig. 3.6c).

The changes in the DEM are in general relatively small. The finally adjusted (`DEM_L3`) does not differ significantly from the original DEM in the central tendency measures ($\bar{z}$=157.2, s_z=43.8 versus $\bar{z}$=156.4, s_z=43.1). The histogram comparison, however, showed smoother distribution of values for `DEM_L3`, while the unfiltered `DEM_L0` was characterized by the typical grouping around the contour values (z=90,100,...). This corresponds to the results of Brown & Bara (1994).

3.3.2 Errors in terrain parameters

The derivation of terrain parameters from unfiltered data (`DEM_L0`) resulted in large parts of area being undefined (Fig. 3.8a). The proportion of undefined pixels equals the proportion of padi terraces (17.3%), except for the CTI (22.6%). After the filtering of DEM, the number of undefined pixels was also reduced, although there are still some patches of undefined pixels (Fig. 3.7a,c and Fig. 3.8b). The proportion of undefined patches in the CTI is somewhat higher than the total proportion of padi terraces in the filtered DEM (5.3% compared to 1.9%). This is because the calculation of CTI is only possible if all neighbours are defined. Otherwise the error might propagate to other pixels.

The variogram analysis of each terrain parameter showed important differences between them (Table 3.1). The distinctly contiguous parameters were DEM and NORTH, while the curvatures proved to be locally variable features having bounded variograms and being correlated at relatively short distances. Also note that the estimated variogram models differently control the calculation of weights (Table 3.2). It appears that the key factors that determine the importance of neighbours are the nugget variation and distance at which covariance reaches 10% of the sill value. In the 3×3 window environment, the differences between the weights in cardinal and diagonal directions are mainly controlled by the nugget value. If the nugget value

Table 3.1: Variogram modelling of terrain parameters.

Terrain parameter[a]				Variogram modelling[b]					
	unit	AVG	STD	Model	Anisotropy	C_0	$C_0 + C_1$	R	$R(10\%)$
DEM	m	156.4	43.1	exponential	No	0	1350	675	1554
SLOPE	%	13.6	11.6	exponential	No	0	88	115	265
PROFC	rad m^{-1}	0.00	0.17	spherical	Yes	0.006	0.0235	156	104
PLANC	rad m^{-1}	-0.03	1.26	spherical	Yes	0.68	1.82	183	112
CTI	-	6.84	1.31	exponential	Yes	0.45	1.78	85	171
NORTH	-	90.0	47.0	exponential	Yes	0	2450	232	534

[a] AVG – mean value; STD – standard deviation

[b] C_0 – Nugget; $C_0 + C_1$ – Sill; R – range parameter; $R(10\%)$ – distance at which covariance reaches 10% of the sill.

is high, the weights are more or less equal and vice versa. In the 5×5 window environment, the importance of the outer neighbours ($w_{\mathrm{C,D,E}}$) is controlled by the distance at which covariance reaches 10% of the sill. For DEM, NORTH and SLOPE, the outer neighbours participate in approximately 15% of the total weights, while for the PLANC and CTI, this number is two times smaller. For prediction of PROFC the outer neighbours accounted for only 5% of the total weights. This confirms that the curvatures are more locally variable features. Note that the proportion of outliers in the study area was the highest for the PLANC and DEM and lowest for the CTI (Table 3.2). In all cases, the proportion of outliers did not exceed 2% of the total area.

After the filtering of undefined pixels and outliers, the terrain parameters appear more reliable (Fig. 3.7d and Fig. 3.8c). The same can be concluded for the results of error propagation (Fig. 3.7b and Fig. 3.8d). The latter technique showed to be especially suitable for reduction of errors in the aspect map and curvatures. On the other hand, it showed some difficulties in the areas of low relief. For example, in the case of calculating CTI, the previously induced adjustments in the elevation (streams, reduced pits and peaks) loose its importance in the error propagation. In the case of the lower number of iterations, it introduces a noisy pattern in the final derivative, which does not necessarily reflect the real case. There is certainly a difference in mapping CTI in plain using error propagation and filtering of slopes (Fig. 3.7c and Fig. 3.8d). The key reason for this difference is estimation of slope in the plain area. Error propagation in general increases slopes in the plain region

Table 3.2: Proportion of undefined pixels in terrain parameters derived from the unfiltered DEM and calculated weights for the 3×3 and 5×5 window environments.

	Kriging weights[a]								Outliers	
	3×3 window		5×5 window							
	w_A	w_B	w_A	w_B	w_C	w_D	w_E	$w_{C,D,E}$	s_δ	$A_e(t > t_a)$
DEM	0.253	-0.003	0.260	0.050	-0.025	-0.015	-0.006	16.3%	1.28	0.74%
SLOPE	0.249	0.001	0.258	0.050	-0.024	-0.014	-0.005	15.7%	2.19	0.41%
PROFC	0.183	0.067	0.183	0.070	0.006	-0.001	-0.007	5.2%	0.06	0.35%
PLANC	0.173	0.077	0.168	0.071	0.011	0.003	-0.006	8.9%	0.62	1.05%
CTI	0.176	0.074	0.174	0.073	0.010	0.001	-0.009	8.0%	0.95	0.12%
NORTH	0.251	-0.001	0.260	0.050	-0.025	-0.015	-0.006	16.2%	13.10	0.42%

[a] w_A – weights in cardinal direction; w_B – weights in diagonal directions; $w_{C,D,E}(\%)$ – percentage of outer neighbours in the total cumulative weights.

(under-estimation of CTI). The filtering of slope map, on the other hand, maintains fairly small values (over-estimation of CTI). In this case, the over-estimation of CTI appears to be more realistic as it portrays the watershed as being connected.

In other examples, error propagation seems to be the most robust way of producing smoother terrain parameters, with much less artefacts and more natural appearance. The Fig. 3.9 shows, for example, difference between PLANC calculated using a single, 20 and 50 realisations. The improvement is visible even after few realisations. After higher number of iterations, PLANC shows connected, smoother features; also note that the artefacts in the plain region disappeared from the map.

3.3.3 Effects on Soil-landscape modelling

Comparison of landform classification using unfiltered and filtered terrain parameters showed distinctive differences. The problem of artefacts and outliers propagates to the supervised classification of landform. This brings some new problems also: the classifier found valley bottoms on the hilltops, high terrace and floodplain could not be distinguished (Fig. 3.10b). After the filtering of DEM and terrain parameters, the overall classification accuracy increased from 51.3% to 72.0%. After the filtering of terrain parameters, the classified landform map (Fig. 3.10c) shows higher agreement with the reference aerial photo-interpretation map (Fig. 3.10a). This is because a mapper often tends to generalize and create smoother transitions during

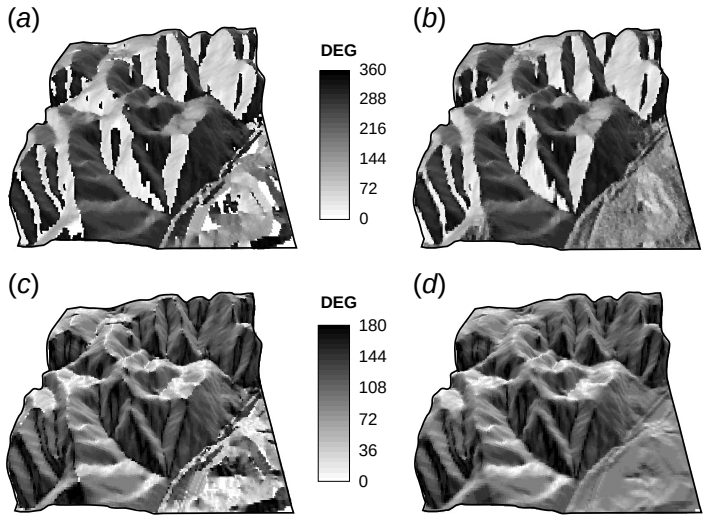

Figure 3.7: Aspect map (0-360°) derived from filtered DEM (a) and by using the error propagation (b). Northness map (0-180°) derived from filtered DEM (c) and slope-adjusted northness map (d).

the photo-interpretation, which is also a property of filtering. Hence, the two maps appear to be more similar visually.

The SOLUM was significantly correlated with DEM (r=-0.45), CTI (r=0.47), SLOPE (r=-0.32) and PLANC (r=-0.29). In soil survey terms, this means that the observed soils in the study area are in general shallower at higher elevations, steeper slopes and convex positions. On the other hand, deeper soil can be found in the areas of high potential accumulation. The step-wise regression analysis extracted DEM and CTI as the best predictors of the SOLUM with R^2=0.27. After the filtering of terrain parameters, the model improved to R^2=0.40, the best predictor being CTI. Comparison of the prediction maps is given in Fig. 3.11.

3.4 Discussion and conclusions

The objective of this work was to review and systematise methods to improve geomorphic plausibility of DEMs and minimise artefacts and outliers in terrain parameters. Three main approaches to the reduction of errors in DEM and DEM-derived products have been considered. The first is the empirical approach where the knowledge on features is used to reduce errors, primarily to improve the plausibility of the

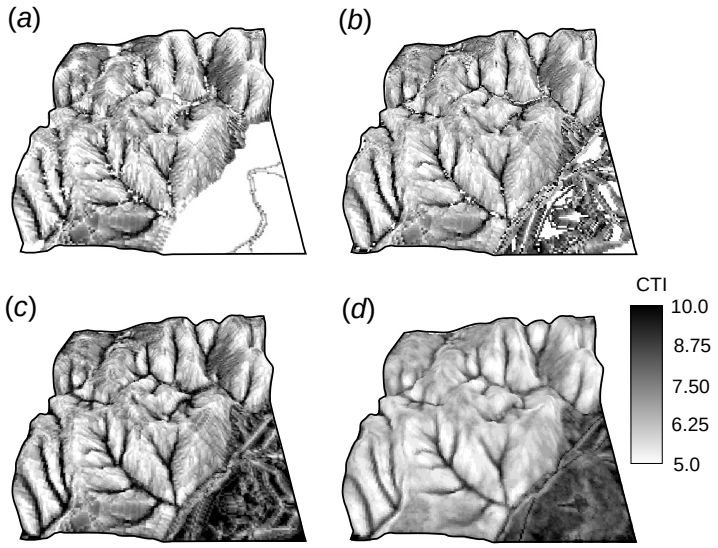

Figure 3.8: Compound Topographic Index (CTI): (a) derived from the original DEM; (b) from filtered DEM; (c) filtered for outliers and (d) averaged from 20 realisations. White patches in (a) and (b) are the undefined pixels.

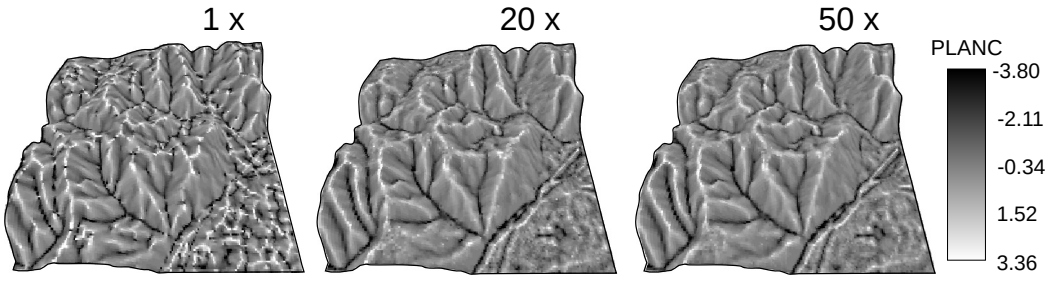

Figure 3.9: Comparison of PLANC calculated using a single, 20 and 50 realisations.

DEM. The examples are adjustment of the elevation using medial axes or stream network and modification of the northness map using the slope-adjustment formula. Limitation of the empirical approach is that it needs a good knowledge about terrain features. The automated methods, e.g. the automated extraction of medial axes, need to be taken with a care since the created ridges and valley bottoms might

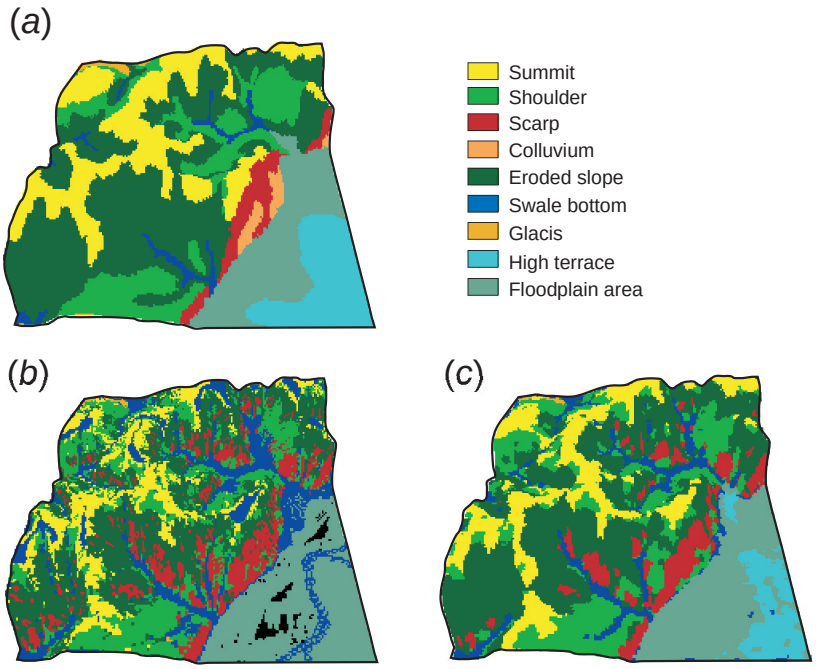

Figure 3.10: Landform classification: (a) the reference aerial photo-interpretation map with the legend; (b) results of classification using unfiltered terrain parameters and (c) after the filtering of terrain parameters. Black patches in (b) are the undefined pixels.

not reflect the reality. On the other hand, filtering of DEM does not guarantee that 100% of artefacts will be removed. It is advisable to check the percentage of artefacts and, if needed, digitise extra contours or use extra auxiliary information.

The second approach to reduction of errors is the filtering of values using the spatial dependence structure and probability of exceeding a value estimated from the neighbours. This approach is useful for filtering of outliers and, in general, gives somewhat smoother picture of the terrain. However, in rugged topography, this approach might smooth-out real features such as steep cliffs or sinkholes. Note that we could have simply applied median filter to reduce outliers. On the other hand, median filter does not take into account range of spatial dependence and can have unwanted effects. For example, if the elevation is strongly correlated spatially and at longer distances, then the confidence limits need to be much narrower. Similarly, if the elevations vary at small distances, than the definition of an outlier is not as strict. These aspects cannot be incorporated into a simple medial filtering. The problem with the geostatistical analysis is that the variogram models for terrain

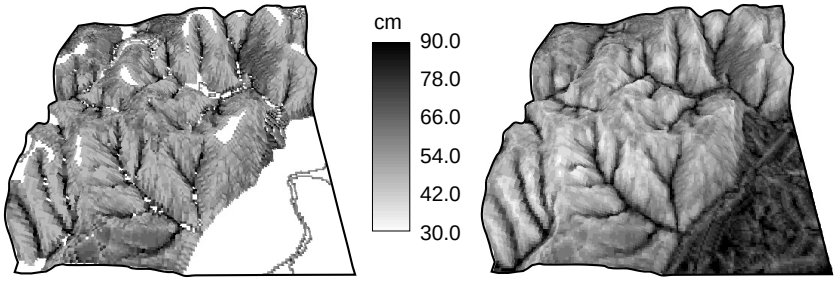

Figure 3.11: Depth to the parent material (SOLUM) predicted using: unfiltered terrain parameters (left) and filtered terrain parameters (right). White patches are the undefined pixels.

parameters and threshold distances are assumed correct, although they can differ greatly for different parts of the area as well as for different grid sizes. It may therefore be more reasonable to use empirical or field-validated models of spatial variation and threshold limits. For example, instead of calculating the standard deviation of the differences between the estimated and given elevation, we can use an empirical value (e.g. half $RMSE(z)$). An alternative is to estimate the true values of terrain parameters using transect studies, estimate variogram parameters and threshold values allowed and then use these for filtering. Another problem with this approach is the selection of the filtering window size. It appears that the 3×3 window environment is large enough for filtering of curvatures and CTI (i.e. locally varying features). Larger window sizes are computationally more demanding, but more accurate.

The third approach to the reduction of errors is the error propagation and is fully data-driven. The errors are reduced by calculating the average value of multiple realisations. This in general creates more natural and more contiguous picture of the geomorphology. The advantage of error propagation is that it does not need calculation of filtering weights or selection of the window size. The reduction of errors by error propagation is especially interesting as it can be fully automated. It is also attractive because it offers a (propagated) measure of the uncertainty of deriving a terrain parameter. The possible problems with error propagation, is that it can be time-consuming, as it often needs many realisations. It also needs a good estimate of the error in input values ($RMSE(z)$). For our sample area we have evaluated that the used input for the error propagation was too high in the plain region. This had an effect of increasing the slope (and CTI) in the plain.

One should keep in mind that elevation, i.e. topography is a non-stationary and

non-periodic feature. This means that it is probably more advisable to estimate its spatial variation model at local scales. Especially in karst and heavily dissected areas, it will be hard to estimate the global model of spatial variation, which typically means that filtering might over-smooth some untypical but important geomorphic features such as cockpits, cliffs, embankments or real padi terraces. It cannot be excluded, that even in our sample area we have incidentally corrected away some small number of real features such as real terraces and depressions that can occur naturally. One solution to this problem is to cut or mask out areas that make no sense for the terrain analysis, e.g. real rice terraces or escarpments.

The results of this case study have shown that the proportion of artefacts in the unfiltered DEM can be fairly high. In this case, the high proportion of padi terraces and spurious sinks and peaks (17.3%) was due to the limited interpolation technique and under-sampled features in the plain terrain. After the reduction of errors using filtering of DEM and terrain parameters, these were more successful for mapping of landform facets and prediction of solum thickness. Thus, the reduction of errors in DEM and DEM-derived data plays an important role for the success of soil-landscape analysis. Note that we did not evaluate effects of the grid size and vertical resolution, as it seems that these are not the real factors controlling the quality of terrain parameters but should be inferred from the scale of research and given data quality ($RMSE(z)$ or contour interval).

Chapter 4

Supervised landform classification*

"Is the replacement of experienced surveyors really what this paper is about?... the authors seem to criticize the field mappers for using excessive subjective methods and praise the computer programs... it appears that there would have been savings of total work: this work is being shifted to computer experts in the office rather than by mappers in the field."

[comments from an anonymous reviewer, referred to as experienced soil surveyor, on the paper referred to down-bellow]

*based on: Hengl T. and Rossiter D.G., 2003. Supervised landform classification to enhance and replace photo-interpretation in semi-detailed soil survey. Soil Science Society Journal of America, Vol. 67(5), in press.

4.1 Introduction

A product of the semi-detailed soil survey is an entity-class or polygon map of soil types at a typical scale of 1:50 K, with minimum legible delineations of 10 ha and optimal delineations of 40 ha (Forbes et al., 1982). This corresponds to "Order 3" to "Order 4" (Soil survey Division staff, 1993), semi-detailed or medium intensity soil surveys (Avery, 1987). These soil maps are intended for extensive land-use planning and to give a reasonably accurate picture of the distribution of soil types in an area at relatively low cost. The standard method of semi-detailed survey is to draw preliminary boundaries on aerial photos by means of stereoscopic landscape analysis, and then determine the soil types that occur in each map unit by field inspection of the soil at representative sites. A common inspection density is one observation per one to four map cm^2 (Western, 1978), which at this scale represents 25 to 100 ha. Often, the surveyors make sure that there is also at least one observation per each polygon. The observations are used to characterize the composition of photo-interpretation units, rather than to find or adjust every boundary.

One approach to semi-detailed survey is to study representative sample areas, typically covering about 10% of the survey area, more intensively in order to arrive at a better understanding of the soil-landscape relations and map unit composition. Field sampling is thus concentrated in comparison to the densities mentioned above, to an observation density of one per 2.5 to 10 ha in the sample area. This is at the cost of samples over the rest of the area, which is then mapped purely by photo-interpretation, extrapolating from the detailed understanding of the soil landscape built up in the sample areas. Because of the low inspection density, the only way that such maps can be reasonably accurate is if the surveyor is able to correctly understand the soil-landscape relations in the survey area, and then map these by surface features visible on the aerial photo (e.g. the landform as seen stereoscopically).

In many cases, however, including standard mapping procedures in the USA, surveyor's experience on soil-landscape relations is used without formalization (Soil survey Division staff, 1993, p. 219-231). Jenny's conceptual equation , i.e. soil property or class = f{climate, organism, relief, parent material, time} is thus used subjectively (and sometimes subconsciously) as a concept to guide photo-interpretation. Depending on the survey area, some aspects of the equation may be more important than others; for example, on a typical hillside, the catena or topo-sequence concept may be uppermost in the surveyor's mind, whereas in areas of recent deposition, parent material and time may be more important. In many areas, landform classification or segmentation, usually by photo-interpretation, is an important step in building up the soil map, since the landform delineations are often associated directly with natural soil bodies (Buringh, 1960).

For some time, there has been interest in replacing or supplementing the expert

judgment of the surveyor with reproducible procedures, particularly by the use of digital terrain analysis to predict the distribution of soil properties (Moore *et al.*, 1993; Gessler *et al.*, 1995; Bell *et al.*, 2000) and model depositional and erosional processes (Mitasova *et al.*, 1996). Irvin *et al.* (1997) were among first to use terrain parameters to derive soil-landscape elements and provide more objective basis for production of soil maps. They compared automated classification of landforms with the manual delineations by API using a small study area. Other authors have attempted to directly derive soil classes from terrain parameters (Thompson *et al.*, 1997; Thomas *et al.*, 1999). Recent developments include use of automated methods to detect landform facets using unsupervised fuzzy-set classifications (Burrough *et al.*, 2000). These are then applied even in the areas of lower relief to enhance crop production using site-specific management (MacMillan, 2000). A related effort by Zhu and collaborators 1996; 2001 attempts to infer soil classes identified by expert soil surveyors as being typical for each class directly from the terrain parameters, geological and remote sensing data.

4.2 Methods

4.2.1 Study area

The study area of 1062 km^2 corresponds to the Croatian portion of the historic region of Baranya. It is located in north-eastern Croatia, in the triangle formed by the Danube River to the east, the Drava River to the southwest, and the Hungarian border to the north (centered at N 45°42'14", E 18°40'35"). It is the half of Osijek-Baranja County lying between the Drava and Hungary, and is part of the large Panonian plain, which stretches through all of Hungary and ends in the northeastern part of Croatia and in the Vojvodina region of Yugoslavia. Soils are mostly formed in Pleistocene and Holocene sediments (Bognar, 1984). The major relief types can be seen in Fig. 4.1. The study area consists of number of different landforms and therefore was interesting how will the classification work for different types. The principal soil-forming factors differ between the essentially erosional hill land and depositional plain. The most extensive landscape is the fluvial plains of the Danube and Drava rivers, with temporary and permanent swamps and a series of terraces. The higher terraces are covered with 20–50 meters of Pleistocene loess overlying older fluvial sediments. The plains are fairly level, and cover about 85% of the total area. Elevations range from 80 to 250 m above sea level, and local relief is flat to gently undulating. About 20% of the area is in a separate landscape, Baranja hill. This is a dissected asymmetrical horst ridge of basalt and andesite, mostly blanketed by Pleistocene loess, from a few meters on the summit to 30 m at the bottom of the glacis. In some vales and on the glacis, there is gravely colluvium eroded from

bedrock.

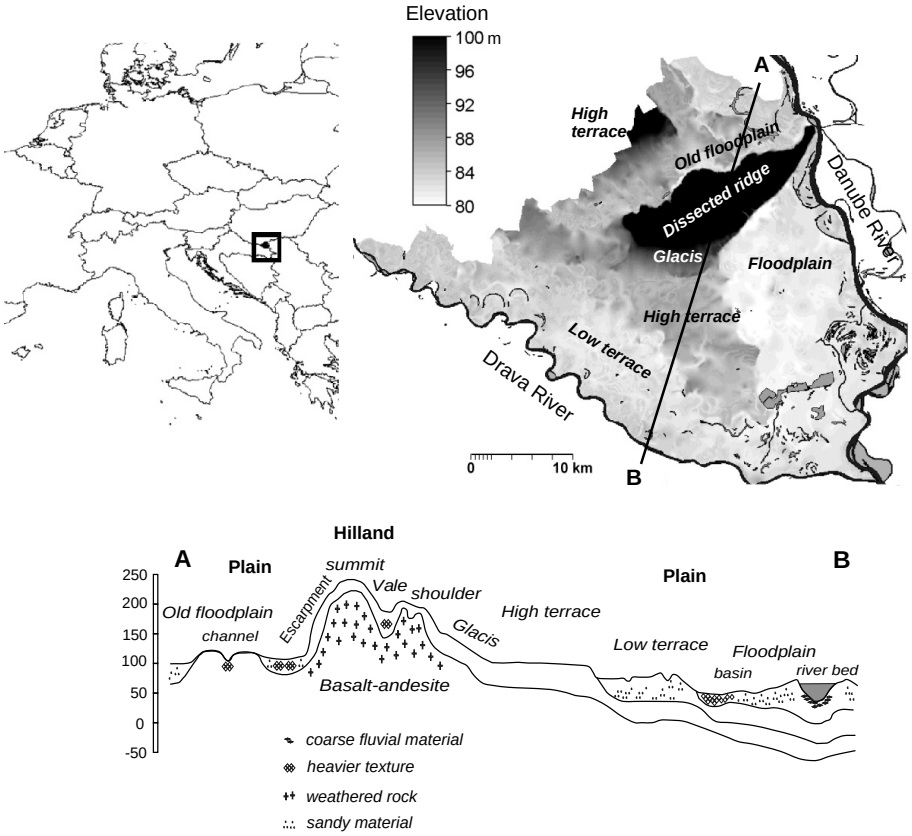

Figure 4.1: Location of the study area in Europe (upper left), main relief types as seen from DEM (upper right), and a cross section sketch indicated by line (bellow). Note that the DEM is stretched to the 80–100 m range to emphasize the main relief types.

The average monthly temperatures vary from 0° C in January to 21° C in August and with an overall annual average of 11° C. Annual precipitation varies from 630–750 mm. The ground water depth varies from above the surface in flooded swamps to more then 10 m on the higher terraces and is mainly determined by the two rivers and thus varies annually and seasonally. Water tables reach their maximum in April and minimum in October. During spring, the active floodplain and the lower terraces are occasionally flooded.

4.2.2 Data input and photo-interpretation

Topographic maps at 1:50 K covering the study area were scanned, imported to the GIS, geo-referenced to an accuracy of 2 to 8 m using overprinted grid intersections, and resampled to a 10 m cell size, all in ILWIS GIS (Unit Geo Software Development, 2001). The contour interval was 10 m, with supplementary 5 m contours and spot heights in areas of low relief. On-screen digitizing was used to create vector layers of contours and spot heights. We used the ANUDEM interpolation method with drainage enforcement (Hutchinson, 1989), as implemented in the TOPOGRID command of the ArcInfo 8 GIS (Environmental Systems Research Institute, 2001) to produce a digital elevation model (DEM). This algorithm required about 20 iterations in areas of low relief and 100 iterations elsewhere, to reach a steady state, i.e. when DEM does not change visually any more. Upon inspection of the results of the first efforts, artifacts such as slope breaks, cut-offs or spurious sinks, were still clearly visible, especially in areas of low relief and on hilltops. Thus, we decided to digitize supplementary contour lines and spot heights indicating small channels, hilltops and ridges that were not indicated on the original topographic maps. Their elevation (±2 m) was estimated from nearby contours and field knowledge of relative local elevation differences[2].

From total of 167 aerial photos covering the whole study area, we selected six training photos of 2116 ha (4.6×4.6 km) each, totaling 11,079 ha. These were selected subjectively to provide a representative sample of major soil landscapes. Training areas "A" and "F" covered sections of Baranja hill, the abandoned course of the Drava, and the edge of the low terrace, while the others ("B", "C", "D" and "E") covered the terraces and the floodplain (see Fig. 4.6). The middle photos from triplets of photogrammetric vertical 23×23 cm aerial photos at approximately 1:20 K scale were interpreted according to the geo-pedological method of Zinck (1988) and cross-checked in the field, resulting in a four-level hierarchical legend (Table 4.1). The minimum delineation size was 10 ha (0.4 cm^2 on the map), and the minimum delineation width was 150 m (3 mm on the map), since the objective was to make a 1:50 K soil map. Twenty-one soil-landscape units (seven in the hill land, fourteen in the plain) were identified, of which thirteen (six in the hill land, seven in the plain) accounted for 95% of the training area. Both the photos and the interpretation overlays were scanned, imported into ILWIS, geo-referenced with an ortho-correction to a horizontal precision of 3 to 15 m, using five to eight tie-points per photo (Rossiter & Hengl, 2002).

[2]See chapter 3 for more details about the reduction of errors in terrain parameters.

Table 4.1: Geo-pedologic legend showing hierarchical classification of landforms for Baranja region. Classes removed in reduced legend (3.5% of total area of training photo-interpretation) are set bold.

Landscape	Relief	Lithology	Landform	Code	(%)
Hill land	Dissected ridge(horst)	aeolian loess over basalt-andesite bedrock	Summit	Hi111	5.7%
		aeolian and deluvial loess over basalt-andesite bedrock	Shoulder/backslope	Hi112	3.4%
	Escarpment	aeolian and deluvial loess over basalt-andesite bedrock	Scarp	Hi211	1.4%
			Colluvium	Hi212	2.6%
	Vales	deluvial loess	Slope	Hi311	7.2%
		deluvial loess	Bottom	Hi312	0.9%
	Glacis	deluvial loess	Slope	Hi411	3.5%
Alluvial Plain	Recent floodplain	medium-textured fluvial sediments	Floodplain	Pl111	22.9%
			Levee	**Pl112**	1.3%
			Abandoned point bar complex	Pl113	3.5%
			Cut off channel	**Pl115**	0.8%
		coarse-textured fluvial sediments (sand, gravel)	Point bar complex	**Pl121**	0.8%
			Active channel banks	**Pl122**	0.2%
	Low terrace	medium textured fluvial sediments	Tread	Pl211	12.9%
			Overflow channel	Pl212	1.4%
			Elevations	**Pl213**	0.4%
		coarse fluvial sediments	Abandoned point bar complex	**Pl221**	0.1%
	High terrace	loess over fluvial sediments	Tread	Pl311	21.9%
			Abandoned channel	Pl312	1.1%
			Elevations	Pl313	0.5%
	Older flood-plain	fluvial sediments	Floodplain area	Pl411	7.8%

4.2.3 Extraction of terrain parameters

The DEM was used, directly or as a component, in calculating eight terrain parameters (maps): ground water depth (GWD), slope gradient (SLOPE), profile curvature (PROFC), plan curvature (TANGC), viewshed reflectance (VSHED), accumulation flow (FLOW), Compound Topographic Index or wetness index (CTI), and sediment transport index (STI), each at 30 m grid resolution (Fig. 4.2). SLOPE, PROFC, TANGC, and VSHED were calculated directly from the DEM with 5×5-pixel filters in ILWIS, which implements the Zevenbergen & Thorne (1987) formulas. Independently, the distance to nearest watercourse (DISTW) was computed from a map of

the drainage network.

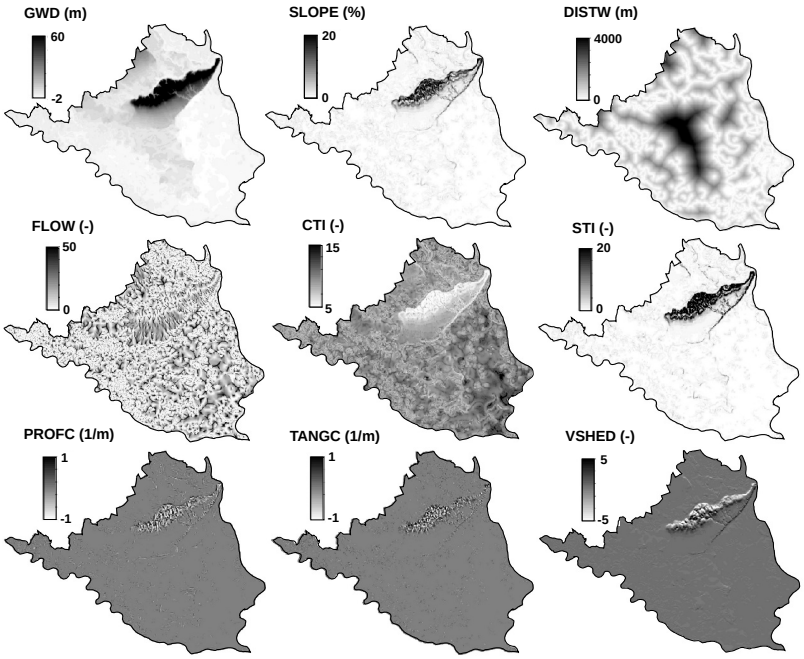

Figure 4.2: Terrain parameters of the study area derived from the DEM and related data.

GWD was calculated by using the additional information from the topographic map and hydrological stations. The base elevation of the water table was estimated from four benchmarks at Danube and Drava river level on the edges of the study area. The water table surface for the whole area was calculated using the mean annual water table height measurements and second-order trend function. The GWD was then calculated as the difference between the DEM and this surface. Thus, the GWD represents a slight adjustment of the DEM or relative elevation for the regional slope to the southeast. The CTI reflects the tendency of water to accumulate at any point in the landscape, while STI reflects the erosive power of the overland flow:

$$CTI = \ln\left(\frac{A_f}{\tan\beta}\right) \tag{4.1}$$

$$STI = \left(\frac{A_f}{22.13}\right)^{0.6} \cdot \left(\frac{\sin\beta}{0.0896}\right)^{1.3} \tag{4.2}$$

where A_f is the specific catchment or contributing area, which is the cumulative number of grid cells draining through the target cell (FLOW) and β is the local slope angle related to that cell (Burrough & McDonnell, 1998). Both CTI and STI were calculated in ILWIS using the multiple flow direction method of Quinn *et al.* (1991). Since the algorithm was developed using the neighborhood operation in GIS, it needs a number of iterations as an input. Here, we used 50 iterations, i.e. neighboring pixels to derive flow accumulation. This small number of iterations was sufficient, because remaining changes with further iteration were only in stream bottoms, which already had a high CTI relative to other landscape positions. A problem with the algorithm was that in pixels with zero slope, the CTI calculation fails due to division by zero. Also a zero accumulation flow is unrealistic and will produce an undefined pixel. In these positions, we approximated CTI by iteratively averaging the slope (β) and accumulation flow maps (A_f) from surrounding pixels until all zero values were replaced with small values. This has the effect of creating pools of high CTI in the plain, which in the study area is realistic due to lowest position of these areas[3].

The viewshed reflectance (VSHED) is a relative estimate of direct incoming radiation, i.e. an estimate of the solar energy reaching the surface. It was computed using the formula estimated by Horn (1981) and described in Burrough & McDonnell (1998). Here we assumed the sun to be at an elevation of 45° and azimuth of due South. This variable was selected to present different expositions and environmental conditions.

During the creation of the predictors, it became clear that there were major differences in central values and spread of terrain parameters (as seen on histograms). Some predictors (SLOPE, VSHED, CTI) showed asymmetrical, log-normal, while others (GWD, STI and DISTW) inverse distributions. This asymmetry in histograms reflects with a low contrast in images due to the domination of the plain landforms. In addition, there was a significant inter-predictor correlation between the two major landscapes (hill land and plain) (Table 4.2). Especially PROFC and TANGC are inversely related, as are CTI and SLOPE, while the strongest correlation show STI and SLOPE. Similarly, higher elevations are associated with steeper slopes and lower CTI. Because of these correlations and the difference in spread among predictors, the data reduction by factor analysis in a GIS (Eastman & Fulk, 1993) is interesting as an effective transformation to reduce multicollinearity and improve the contrast in predictors.

The principal component analysis (Table 4.3) shows that the first five components account for more than 80% of total variance. The most significant data reduction is in hill land, where first three components accounted for almost 75% of

[3]See explanation in chapter 3.

Table 4.2: Correlation between predictive variables for whole-area, hill land and plain only. The most significant values (> 0.5) are set bold.

Whole Hill Plain	FLOW	CTI	DISTW	GWD	PROFC	SLOPE	STI	TANGC	VSHED
	1.00	0.21	0.02	-0.02	0.25	0.00	-0.01	-0.27	0.00
FLOW[a]	1.00	0.25	0.01	-0.18	0.46	0.01	-0.05	**-0.66**	0.00
	1.00	0.27	0.02	-0.05	0.34	0.00	-0.09	-0.36	0.01
		1.00	-0.09	**-0.52**	0.13	**-0.59**	**-0.52**	-0.15	0.05
CTI[b]		1.00	**0.50**	**-0.51**	0.19	**-0.85**	**-0.83**	-0.24	0.22
		1.00	-0.08	-0.31	0.20	**-0.72**	**-0.64**	-0.18	-0.02
			1.00	0.17	-0.01	-0.01	-0.02	-0.01	0.08
DISTW[c]			1.00	-0.17	-0.01	-0.41	-0.38	-0.05	0.21
			1.00	0.49	-0.01	-0.02	-0.05	0.00	0.08
				1.00	-0.13	**0.70**	**0.67**	0.11	0.00
GWD[d]				1.00	-0.24	0.37	0.39	0.13	0.02
				1.00	-0.05	0.20	0.17	0.04	0.11
					1.00	-0.02	-0.12	**-0.55**	0.00
PROFC[e]					1.00	-0.03	-0.15	**-0.58**	0.01
					1.00	0.00	-0.19	-0.47	-0.03
						1.00	**0.98**	0.09	-0.25
SLOPE[f]						1.00	**0.98**	0.07	-0.34
						1.00	**0.89**	0.05	-0.04
							1.00	0.15	-0.28
STI[g]							1.00	0.16	-0.35
							1.00	0.17	-0.04
								1.00	-0.04
TANGC[h]								1.00	-0.04
								1.00	-0.00
									1.00
VSHED[i]									1.00
									1.00

[a]FLOW - accumulation flow;
[b]CTI – Compound Topographic Index;
[c]DISTW – distance to nearest watercourse;
[d]GWD – ground water depth;
[e]PROFC – profile curvature;
[f]SLOPE – slope gradient;
[g]STI – Sediment Transport Index;
[h]TANGC – plan curvature;
[i]VSHED – viewshed reflectance;

total variance. Still, the high proportion in higher components in whole area shows that the predictors had a fair degree of independence, which is often not the case with remote sensing images.

Table 4.3: Variance proportions explained by the principal component analysis for stretched principal components (PC1 to PC9).

	PC1 (%)	PC2 (%)	PC3 (%)	PC4 (%)	PC5 (%)	PC6 (%)	PC7 (%)	PC8 (%)	PC9 (%)
Whole area	35.86	18.37	12.81	9.79	9.22	5.36	4.98	3.49	0.11
Hill land only	39.25	22.65	11.36	8.48	6.90	6.20	3.64	1.45	0.08
Plain only	29.15	21.99	15.05	9.58	8.26	6.04	4.83	4.46	0.63

Since the predictors are expressed in different units with widely different ranges and distributions, all the maps of terrain parameters were rescaled to a dynamic range of 0–255, which is the one byte per pixel structure typical for satellite images. In this case, we decided to use a linear stretch with 0.5% truncation in each tail. The principal components were normalized to the same range but without truncation. The landform maps were then ready to be used in an image processing software (ILWIS) as *synthetic bands*, i.e. to classify the whole area as in the case of classification of remote sensing images (Janssen & Huurneman, 2001).

4.2.4 Training and classification stage

Two methods for selecting training samples were compared. In the first, the entire area of the interpreted photographs, i.e. API maps (in further text whole-API training set). In the second, training samples were created by manual selection on-screen of about 100 pixels within each photo-interpretation unit in the sample areas (in further text point-sample training set). Here by central concept we consider locations, which were in our mental model typical representatives of landform classes when observed stereoscopically. In addition, the photo-interpretation units were displayed as boundaries over false-color composites of synthetic bands and then point-samples checked to ensure that they fall in relatively homogenous facets (Fig. 4.5). The selection of variables for the three colors and their contrast was adjusted repeatedly to highlight the differences between API units. Thus the second method allows more precision, as the photo-interpretation units typically have inclusions within their minimum legible delineation that may confuse the classifier. This subjective

process is an extension of subjective photo-interpretation: the analyst is asked to find locations within each landform class where representative soils should be found, according to landform. In photo-interpretation, a 3D model is constructed in the analyst's visual perception by comparing adjacent photos of a stereo-pair, whereas in on-screen interpretation, a color composite is adjusted by the analyst until key geomorphic differences are evident by color alone.

In order to investigate whether different major landscapes should be classified separately, we also divided the two landscapes (hill land and plain) by a clearly visible master line. This line was manually delineated on-screen by visually interpreting a color composite of elevation and slope map. The two major landscapes were then classified separately and then merged to a single map.

The training samples were then used as input to maximum likelihood classifiers (Lillesand & Kiefer, 2000, § 7.9), with no distance thresholds, so that all pixels were classified. Automated (classification) and manual (photo) API maps were compared over the entire training area with a confusion matrix, with two test criteria: (1) the proportion of agreement between the two classifications, and (2) the kappa coefficient, which accounts for chance agreement (Congalton & Green, 1999). In addition, the nature and seriousness of the errors was evaluated subjectively, both from the confusion matrix and by a visual comparison of the maps. For the inclusive training set, the same samples were used for classification and accuracy assessment. Thus in this context *accuracy* is more properly termed *reproducibility*, i.e. the degree to which the automated classifier could reproduce the subjective photo-interpretation and sample point selection.

4.3 Results

Different terrain parameters, when examined visually, have shown stronger relationship with the delineations in hill land and in plain. In hill land, the CTI, GWD, SLOPE and PROFC showed strongest correspondence with the manual delineations (Fig. 4.3). When evaluated in the feature space (scatter plots), landform classes in the hill land area showed different clustering, while in the case of landforms in plain, the clouds of points were narrow and adjacent to each other (Fig. 4.4). For example, to distinguish between a channel (Pl312) and terrace (Pl311), a small difference in GWD matters. Mapping these classes is therefore much more dependent of how well are the training pixels selected. The overlay of the photo-interpretation boundaries on a false-color composites (Fig 4.5) clearly showed deficiencies in the photo-interpretation. First, some areas had not been correctly identified by photo-interpretation. For example, sharp bands of bright yellow indicate steep slopes at high elevations; these are either transitions (boundaries) between higher and lower

soil-landscape units (e.g. summit Hi111 and shoulder/backslope Hi112) or, if wide enough (>150 m at this scale), units of the scarp (Hi211). Second, some areas, while correctly identified, should have their boundaries adjusted in order to increase their homogeneity. These adjustments are easily achieved with on-screen digitizing. In this sense, the color composite provides an objective visualization of the geomorphology to supplement the stereovision of the photo-interpreter.

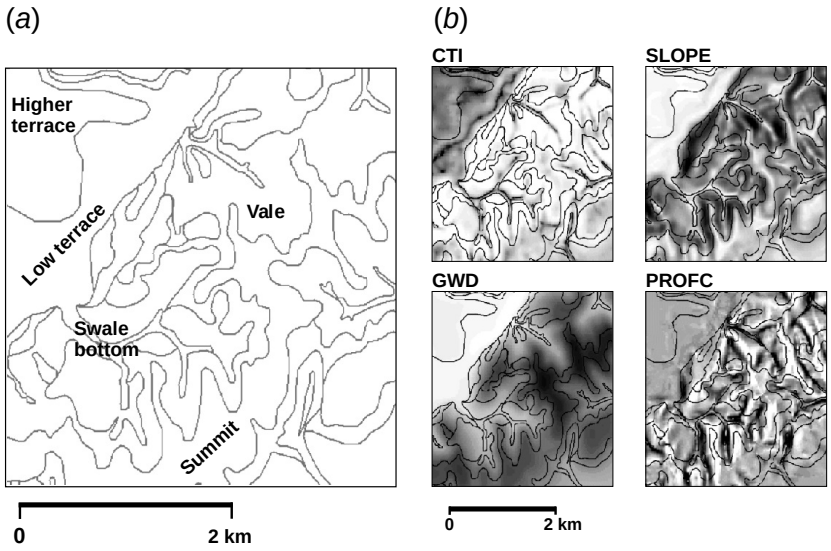

Figure 4.3: Visualization of relationship between terrain parameters and landform classes (training area A): (a) API delineations and main landforms and (b) boundaries overlaid over terrain parameters.

Classification with principal components showed that the first three components were sufficient in the hill land (47.5% of overall accuracy), but that in the plain, substantial improvement in the classification continued through the eighth component (Table 4.4). This means that information in higher level components is still useful for the classification of the landform classes and should not be discarded. Finally, because of the low data redundancy in this data set, there is little advantage in working further with principal components, so we concentrated on the original terrain parameters instead. In addition, the principal components are harder to interpret and therefore unpractical for visualization using false-colors or scatter plots.

Each class in the training set needed a non-zero estimate of the variance. Otherwise poorly conditioned or singular matrices will give unreliable results on inversion, and probability classifiers such as maximum likelihood will fail. This has happened

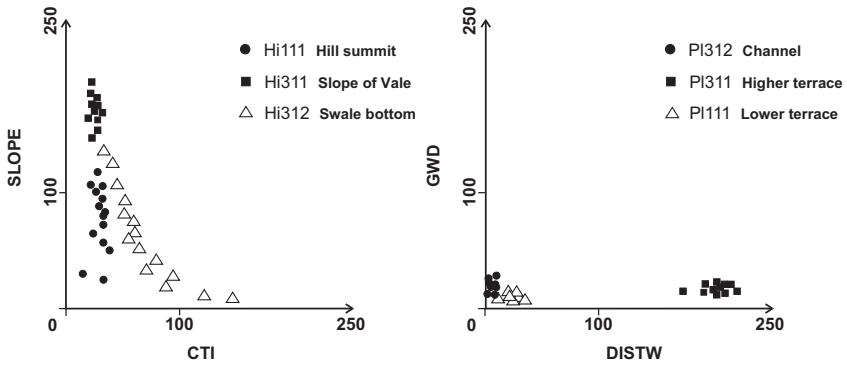

Figure 4.4: Scatter-plot of feature space formed using terrain parameters stretched to 0-255 range. CTI and SLOPE were shown to be good predictors in the hill land area (left), while DISTW and GWD show separation between the landform units in the plain (right).

Table 4.4: Overall accuracy of classification (%) for different number of principal components (PC).

	Number of principal components								
	1	2	3	4	5	6	7	8	9
Whole area (%)	16.0	20.5	27.5	31.8	32.2	32.6	33.9	40.7	41.5
Hill land only (%)	27.9	38.9	47.5	49.9	51.9	53.3	53.7	52.0	51.9
Plain only (%)	9.1	16.0	36.4	32.2	32.8	34.2	44.6	49.6	50.1

several times with our classification when we selected points from terraces where all pixels either had the same SLOPE or GWD. So it is not possible to only select central concepts, some variability must be also included. This can be achieved by iteratively inspecting the scatter-plots of all band combinations to ensure separability and variability of point clusters (training set).

4.3.1 Reproducibility

Initial attempts to classify the entire landscape, never achieved better than about 50% overall accuracy. The classifications using nine predictors, either as original predictors or their principal components, and all API legend classes showed clear differences between methods but similar overall results. The maximum-likelihood

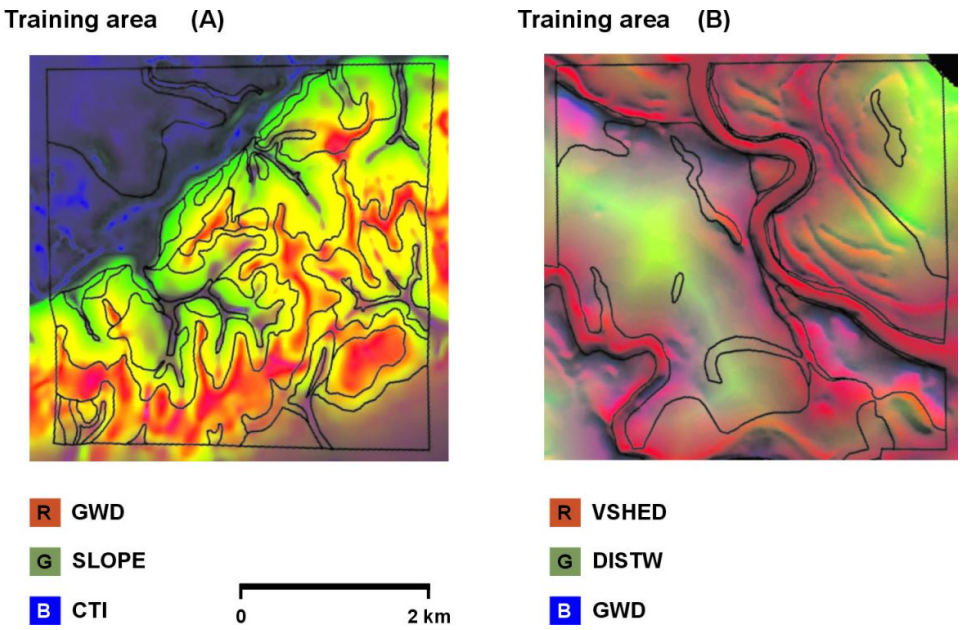

Figure 4.5: False color composite made from GWD, SLOPE and CTI (left) and VSHED, DISTW and GWD (right). Different band combinations are suitable for classification of landform elements in hill land and different in plain.

classification gave 45.3% (Kappa=42.6%) with whole-API training set and 36.8% with point-sample training set overall reproducibility. The corresponding figures for the classification of separate landscapes were 58.1% and 51.6% (hill land), and 39.1% and 34.4% (plain). The whole-API set was consistently superior to the point-sample set, and classification of separate landscapes was superior to a single classification, but in no case was the classification accuracy satisfactory (>80%).

Considering the major classification errors in the plain, both whole-API and point-samples gave similar results. When compared in the whole area, classes Pl115 (cut-off channels) and Pl121 (point bar complex in the floodplain) were grossly over-classified, mostly at the expense of class Pl111 (floodplain). Class Pl221 (abandoned point-bar complexes) was also grossly over-classified, at the expense of five other classes (Pl111, Pl113, Pl211, Pl311, and Pl411). This shows that the complex landform facets can not be easily distinguished by using the terrain parameters. The overall poor results can be attributed to poor separation in feature space, especially in plain region. Since these classes occupied a small proportion of the training areas, they were eliminated from the legend. Similar considerations applied to Pl112 (levee), Pl122 (active channel banks), and Pl213 (small elevations on the low ter-

race). Together these classes occupied only 3.5% of the training API in the study area (4.7% of the plain).

In the hill land, no single class contributed disproportionally to the poor reproducibility. Results are sensitive to sampling method. For the whole-API set, Hi112 (shoulder/backslope) was over-classified, mostly at the expense of Hi311 (vale slopes) and Hi212 (colluvial footslopes of escarpments). The latter may be due to uncertain placement of the photo-interpretation boundary between these two adjacent units. For point-sampling, Hi312 (vale bottoms) was over-classified, mostly at the expense of Hi212 and Hi411 (glacis). In both cases there was substantial confusion between most classes, resulting in moderate overall accuracy. In the case of whole-API set, this is attributed to the heterogeneous nature of the landform predictors, even in 'homogeneous' photo-interpretation units. This can not be corrected, so the 58.1% reproducibility in the hill land is the best possible with this set of predictors.

4.3.2 Improving reproducibility

In order to improve reproducibility, the original legend was reduced according to the results of the first classifications. This corresponded to *a priori* ideas about what differences might be difficult to detect by landform analysis alone. For example, recognition of the channel classes that are of high curvature and close to the water table is feasible. Distance to streams does not help, because some abandoned channels are quite close to active ones. A second group of features that were grouped in the reduced legend were morphologically-compound classes such as point bar complexes that consist of an array of smaller channels, levees and smaller elevations. These are inherently hard to classify, which is similar to the problem of automatic classification of urban areas (land cover classes), often consisting of mixed features.

For the whole-API training set, we assigned the photo-interpretation areas for the eliminated classes to the geomorphologically most similar class of the reduced legend, which turned out to always be adjacent to the merged class in geographical space. We did not consider the confusion in the first classification, but rather reduced the legend on these geomorphological criteria. The reclassifications were:

- Pl112 (levee) to Pl113 (abandoned point bar complex on floodplain);
- Pl115 (cut-off channels), Pl121 (point bar complex on floodplain), and Pl122 (active channel banks) to Pl111 (floodplain);
- Pl213 (elevations on low terrace) to Pl211 (tread of low terrace) and
- Pl221 (abandoned point bar complex on low terrace) to Pl211.

Thus, Pl113 in the reduced legend groups those units in the floodplain that have less active flooding than Pl111. Thus we gave up the attempt to map some details of the floodplain, namely levees, cut-off channels, active channel banks, and coarse-textured point bar complexes; and also for the low terrace, namely abandoned point bar complexes and small elevations. Some classes had to be merged with others of different lithology (here, dominant sediment size).

We then repeated the classification with the whole-API training set, resulting in overall accuracies relative to the API of 63.4% (whole area), 65.8% (plain), and 58.2% (hill land), i.e. an improvement of 26.7% in the plain and 18.1% overall; the results for the hill land were not affected. These results show that whole-API set is unlikely to produce satisfactory results, due to the unavoidable heterogeneity within an API unit and consequent overlap in feature space. However, they provide the basis for manual improvement. The maximum-likelihood classifier with the point-sampling method was quite sensitive to the training set, so that the classification could be improved considerably by iterative selection, classification, and evaluation of results. This effect was most pronounced in the plain, because of the clustering of classes in feature space, as illustrated by Fig. 4.4.

In both landscapes, the best results were achieved when the training sets were selected using the central concept and the landform classes used were defined as morphologically more or less homogenous units. After three iterations we were able to achieve high reproducibility for the point sample itself: 90.2% (Kappa=89.3%) (Table 4.5). Thus we were able to reproduce the classification of the central concept of each landform class. However, agreement with the whole-API set was only improved to 55.8% (hill land), 55.4% (plain), and 53.6% (entire area) using this iteratively-selected point sample. This shows that the API polygons are indeed heterogeneous, and their internal variability is best represented by all pixels in the map unit. On the other hand, to identify fine detail in the landscape, point-samples are preferred. The overclassification of some classes (e.g. Pl313) was probably because of under-interpretation in the original API. These are well-defined elevations but difficult to see stereoscopically, because of the low relative elevation difference. In this sense, the automatic classification is more in accordance with reality than the reference API.

An interesting question with a hierarchical legend is to what degree are the higher levels operational. In this case, to what degree are the misclassifications at detailed level within the same higher-level category. At the highest level (landscape), the automatic classifier using selected points and a reduced legend was quite good. Almost all pixels of the training sample (98.6%) were classified in the correct landscape. The only significant errors were areas of Pl311 (tread of high terrace) and Pl411 (old floodplain) misclassified as Hi312 (vale bottom). At the second

level (relief type), overall accuracy was 72.5%, which can be compared to 53.6% at the landform (detailed) level, as explained above. This shows that the hierarchical legend of landforms (Table 4.1) provides useful information as classes are grouped.

Table 4.5: Reproducibility of point training samples by the maximum-likelihood classification for the whole area, after legend simplification.

Class	User's Accuracy	Sample (pixels)	Proportion of sample	Proportion of API
Hi111	0.98	97	6.9%	5.8%
Hi112	0.97	67	4.8%	3.5%
Hi211	0.88	89	6.4%	1.4%
Hi212	0.95	38	2.7%	2.6%
Hi311	0.94	95	6.8%	7.3%
Hi312	1.00	59	4.2%	0.9%
Hi411	1.00	191	13.7%	3.6%
Pl111	0.85	126	9.0%	23.6%
Pl113	0.85	28	2.0%	4.8%
Pl121	0.95	165	11.8%	13.5%
Pl211	0.83	61	4.4%	1.5%
Pl212	0.71	138	9.9%	22.1%
Pl311	1.00	73	5.2%	1.2%
Pl312	0.94	90	6.4%	0.5%
Pl411	0.81	79	5.7%	7.8%
Overall accuracy	90.2%			
Kappa	89.3%			

4.3.3 Extrapolation to the entire study area

The final classification map produced by iteratively-selected point samples, reduced legend and maximum-likelihood classification is shown in Fig. 4.6. This gives a more mosaic-like pattern than the units produced manually through API, since these are already generalized and smooth because of cartographic considerations of scale and consequent minimum delineation size and width. Incongruous boundaries can be

explained by the lack of detailed contours both in areas of low relief and in areas with complex relief over a short distance in the hill land, both of which can be recognized on the aerial photos. In the plain, the automatic classification found details in small channels and ridges that the photo-interpreter had generalized or missed due to the low impression of relief.

In general, the visual agreement with a conventional soil-landscape map is strong. Especially for relatively homogeneous landforms that cover relatively large areas, such as terrace treads, channels, vale slope and hill summit, the correspondence is high. In some areas, features recognized through the API were not detected. The separate classification of the hill land was markedly superior to the separate classification of the plain and to the whole-area classification. This is because landscape elements derived from a DEM are more striking and easier to both photo-interpret and automatically classify when there is strong relief.

4.4 Conclusions and discussion

The results show that the supervised classification is an objective method to supplement photo-interpretation, especially in surveys where the funds are limited and only a few soil observations are made outside the training areas. To survey all of Baranja with the conventional method of semi-detailed soil survey based on landscape analysis would have required the manual interpretation of the center photos of 84 photo-triplets. In the present study, we used only 6 photos (6.25% of the total) to map the whole area and therefore largely decreased cost and effort. Overall accuracy of the supervised classification of landforms improved to 63.4% of the training API and 90.2% of the point-sample, once the legend was simplified to eliminate small classes that caused large relative misclassifications. A further step would be to do field sampling of the soils in these units to determine if they are distinct soil-landscape units.

Supervised classification can be applied over entire survey areas, or separately in major landscapes through easily-identified master lines which tend to follow slope breaks or abrupt changes of landscape type. In the current study, the results improved only slightly. However, stratification has the conceptual advantage that the predictive equations correspond better to conventional understanding of soil formation in different environments. Stratification of the area also enables selection of different predictor sets for each landscape. On the other hand, it is more practical to develop a single data-set and predictive map of the entire area at once.

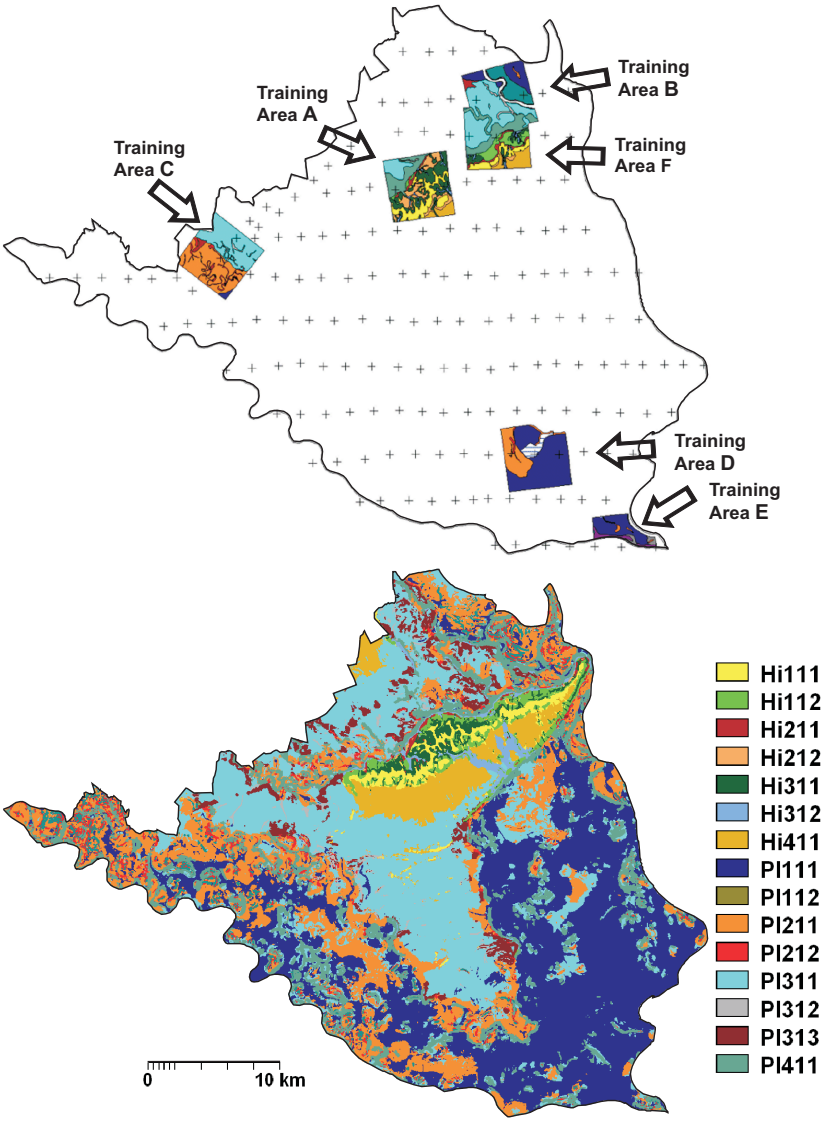

Figure 4.6: Location of the training areas and all aerial photos (indicated with crosses) taken to cover the whole area (above); final map produced through supervised classification of landforms (bellow). The legend was reduced to 15 classes.

4.4.1 Limitations and ways to overcome them

Some photo-interpretation classes were poorly identified by the supervised classification, especially in areas of low relief. The likely reasons for poor performance in the plain are:

- The limited vertical resolution of the DEM relative to the relief;
- Large distances between known elevations (contour lines), both of these leading to artifacts in the landform parameterization;
- The absence of predictor variables specifically adapted to the plain, other than relative elevation, such as distance to local drainage, and
- The presence of landform complexes which occupy too much of feature space.

We also discovered that many features in the plain differ on the topo-map (data collected in 1985) and aerial photo (1998). This is due to the fact that the fluvial processes such as flooding and building of dams and canals change the detailed geomorphology of the area much faster than in the hill land.

Considering the classification accuracy, the supervised landform classification provides, in general, poorer results than the typical accuracy of land-cover classification. The use of the whole-API training set was only moderately successful, even with a reduced legend; however, it does not require a further step of point-sample selection. On the other hand, point-samples can be reselected iteratively and accuracy improved. The following three steps refinements may be applied, by preference in the order given:

- refine the training set for classes that are misclassified, using scatter diagrams in feature space;
- simplify the legend or eliminate or merge classes;
- adjust the number of sample points for misclassified classes and
- consider addition of different predictors or improvement of the quality of existing ones.

This procedure is thus seen as a tool for the experienced soil mapper, not a replacement that could be applied by non-specialists. This is in contrarily to the unsupervised classification of landforms as described by Burrough *et al.* (2000), where the only input needed is the number of classes and fuzzy exponent. In the case of supervised classification, the analyst must still have a good knowledge of soil-landform relations, whether working with traditional or GIS-assisted methods.

The analyst must indeed intervene after the initial attempts to classify, in order to discover which landform units cannot reliably be identified. This causes a new collaboration between the geographers, GIS experts and mappers that seems beneficial to all.

Reproducibility could be further improved with the addition of extra and more detailed information directly related to soil-forming factors, e.g. on variation of the ground water table, satellite images showing flooding areas, vegetation indices, and soil moisture as estimated from radar data. For example, to differentiate between abandoned and actively-flooded channels, remote sensing images at times of flooding, or images which can be correlated to soil moisture should be used. Also the use of fuzzy classification algorithms will offer better insight into the spatial confusion among classes, i.e. uncertainty of classification.

4.4.2 Applicability of landform classification for Soil Survey

The supervised landform classification and visualization of terrain parameters as color-composites has two major advantages that can be used to enhance API for Soil Survey. First, the use of terrain parameters provides an objective and cost-effective basis for clustering of landscape facets. The classification results as shown in Fig. 4.6, prove that the procedure described can be used to extrapolate (generalized) photo-interpretation maps, with the only requirement, in addition to the aerial photos, being a good topographic map to show the contours, map of a drainage networks and geological survey map if possible. The classification also identified soil-landscape units smaller than the minimum legible delineation. If these are not just artifacts of the landform classification, this implies that semi-detailed survey could result in maps detailed enough for site-specific management. Second potential application of the proposed methods is to edit the soil boundaries of existing soil maps and improve their spatial accuracy within an existing GIS. This can be done by simply overlying the given boundaries over the false color-composites of terrain parameters or the results of a supervised classification and then modifying the boundaries to match changes in the terrain parameters.

In both cases, the mapper must take into account the usual considerations of legibility and delineation size, and also check the soil-landscape relations discovered by the classifier. This is faster than stereo-interpretation, since the automated classifier has already found the general location of the boundaries. The mapping project is thus the result of collaboration between an expert mapper and the GIS, able to find patterns with correct training. Far from eliminating soil surveyors, this process allows them to concentrate their efforts on their area of expertise, e.g. identification of soil-landscape relations, and map large areas as efficiently as possible.

Chapter 5

A generic interpolation based on regression-kriging*

"The soil is neither random nor stationary, but our models of it may be one or other or both. We should therefore ask whether our models are reasonable in the circumstances and whether they are profitable in leading to accurate predictions."

[R. Webster in "Is soil variation random?" Geoderma, vol. 97: 147-163]

*based on: Hengl, T., Heuvelink, G.B.M. and Stein, A., 2004. A generic framework for spatial prediction of soil variables based on regression-kriging. Geoderma, Vol. 123, in press.

5.1 Introduction

Spatial prediction is the process of estimating the values of a target quantity at unvisited locations. When applied to a whole study area, it is also referred to as spatial interpolation or mapping. Development of generic and robust spatial prediction techniques has been of interest for quite some time (Mitas & Mitasova, 1999). In land resource inventories, kriging and its variants have been widely recognised as primary spatial interpolation techniques from the 1970s. In the 1990s, with the emerging of GIS and remote sensing technologies, soil surveyors became interested to use exhaustively mapped secondary variables to directly map soil variables. The first applications were based on the use of simple linear regression models between terrain attribute maps and soil parameters (Gessler *et al.*, 1995; Moore *et al.*, 1993). In the next phase, the predictors were extended to a set of environmental variables and remote sensing images. This approach was termed *"environmental correlation"* by McKenzie & Ryan (1999), or spatial prediction by multiple regression with auxiliary variables (Odeh *et al.*, 1994, 1995). McBratney *et al.* (2000) coined the term `CLORPT` techniques. Geostatistics and the `CLORPT` techniques are two somewhat distinct approaches to spatial prediction and can both give satisfactory results independently one from another.

In the last decade, a number of 'hybrid' interpolation techniques, which combine kriging and use of auxiliary information, has been developed and tested. Here, two main paths can be recognised: co-kriging and kriging combined with regression (McBratney *et al.*, 2000). The latter path was shown to be more attractive for combination of kriging and `CLORPT` techniques, among others because fewer model parameters need to be estimated (Knotters *et al.*, 1995). In many cases, kriging combined with regression has proven to be superior to the plain geostatistical techniques yielding more detailed results and higher accuracy of prediction. Hudson & Wackernagel (1994) showed that kriging with use of elevation data improves mapping of temperature. Knotters *et al.* (1995) compared ordinary kriging with co-kriging and regression-kriging for soil mapping purposes, favouring the latter. Bourennane *et al.* (1996, 2000) showed that prediction of horizon thickness is more accurate with the use of a slope map as external drift. In several other studies (Odeh *et al.*, 1994, 1995; Goovaerts, 1999b; Bishop & McBratney, 2001), combination of kriging and correlation with auxiliary data outperformed ordinary kriging, co-kriging and plain regression. Although the hybrid interpolation techniques are becoming increasingly popular, there is still a need for a generic methodology that combines theory of generalized linear models (GLM) with universal kriging. Gotway & Stroup (1997) and Opsomer *et al.* (1999) give good starting points.

An (ideal) requirement for both linear regression analysis and ordinary kriging is that the target variable is normally distributed (Draper & Smith, 1998). In many

soil studies, however, the variables show skewed non-normal distributions, which then reflects on residuals also. To account for the normality requirement, transformations such as logarithmic and square root are often applied prior to the regression analysis (Gobin, 2000; Gobin *et al.*, 2001). Similarly, the log-transformation is often applied prior to kriging to account for positively skewed data. Here, the difficulties are the choice of transformation model and extreme sensitivity of errors for back-transformation (Isaaks & Srivastava, 1989). In the case of kriging combined with regression, a common problem is that the method might yield values outside the physical range (e.g. negative values) and these areas need to be manually masked or replaced (Goovaerts, 1997, p. 200). Another issue, in the case of large number of predictor maps, is the problem of multicollinearity (Neter *et al.*, 1996, p. 285). Moreover, it is not clear whether to use all available secondary variables in prediction or only the most correlated ones (Bourennane & King, 2003). These difficulties with data emphasize a need for a generic methodology that can be used with both continuous and categorical, both normal and non-normal data.

The objective of this study was to develop a methodological framework for spatial prediction based on the theory of universal kriging. This framework can then be used with most soil profile databases in a semi-automated or automated manner. We concentrate on the integration of different data processing steps, rather than on the development of new statistical techniques. In addition, we propose an image processing technique to simultaneously visualise predictions and uncertainty associated with prediction.

5.2 Methods

5.2.1 The generic framework

By a generic framework we consider a set of robust techniques that are used jointly to transform, fit, interpolate and visualise the data. Here, we primarily focus on the following aspects:

- reduction of multicollinearity among predictors;
- ensuring the normality of residuals;
- exploiting the 'best' of the data, i.e. correlation with auxiliary maps and spatial dependence at the same time and
- avoiding predictions outside the physical range.

A schematic diagram showing the generic framework is given in Fig. 5.1. The input variables are first transformed using logit transformation for target variables

and factor analysis for continuous predictor maps. The categorical predictors are transformed to indicator maps. The target variables are then fitted using step-wise regression and residuals interpolated using kriging. The final predictions are evaluated at control points. A generic visualisation method is used to simultaneously display both prediction and uncertainty of the prediction model. Technically speaking, the developed generic framework can be termed *step-wise principal component logistic regression-kriging*. For practical reasons, we simply refer to it as generic framework based on regression-kriging. The development of such a framework has been announced by McBratney & Walvoort (2001). We will now first introduce the theory of universal kriging and then extend its algebra using the above-described framework.

5.2.2 The spatial prediction technique: Regression-kriging

A spatial prediction technique, which jointly employs correlation with auxiliary maps and spatial correlation is universal kriging (UK), originally described by Matheron (1969). Many authors (Deutsch & Journel, 1992; Wackernagel, 1998), however, agree that the term *"Universal kriging"* should be reserved for the case where the drift (or trend) is modelled as a function of the coordinates only. The term *"Kriging with external drift"* (further referred to as KED) is more commonly used when the drift is defined 'externally' through some auxiliary variables (Chiles & Delfiner, 1999; Wackernagel, 1998). The drift and residuals can also be fitted separately and then summed afterwards. This was originally suggested by Odeh *et al.* (1994, 1995), who named it *"Regression-kriging"* (further referred to as RK), whereas Goovaerts (1999b) uses the term *"Kriging after detrending"*. UK, KED and RK are, in fact, equivalent methods and should, under the same assumptions, yield the same predictions (for more details, see Hengl *et al.* (2003a)). The advantage of KED is that the equations are solved at once, while the advantage of RK is that there is no danger of instability as with the KED system (Goovaerts, 1997, p. 195). Moreover, RK can be more easily combined with stratification, general additive modelling (GAM) and regression trees (McBratney *et al.*, 2000). Note that, although KED technique seems to be computationally more straightforward, it needs variogram parameters of the GLS regression residuals (which is often ignored), and therefore the GLS regression coefficients as with RK. Some authors make different assumptions and skip some computational step so that products of RK and KED might differ at the end. For example, Hudson & Wackernagel (1994); Bourennane & King (2003) make an assumption that the variogram of residuals (e) is equal to the variogram of target variable (z), which is a simplification. In this case, the KED prediction map will look more similar to the OK map. Other authors (Odeh *et al.*, 1994, 1995), use only ordinary least squares estimate of the drift, which is also sub-optimal but shorter

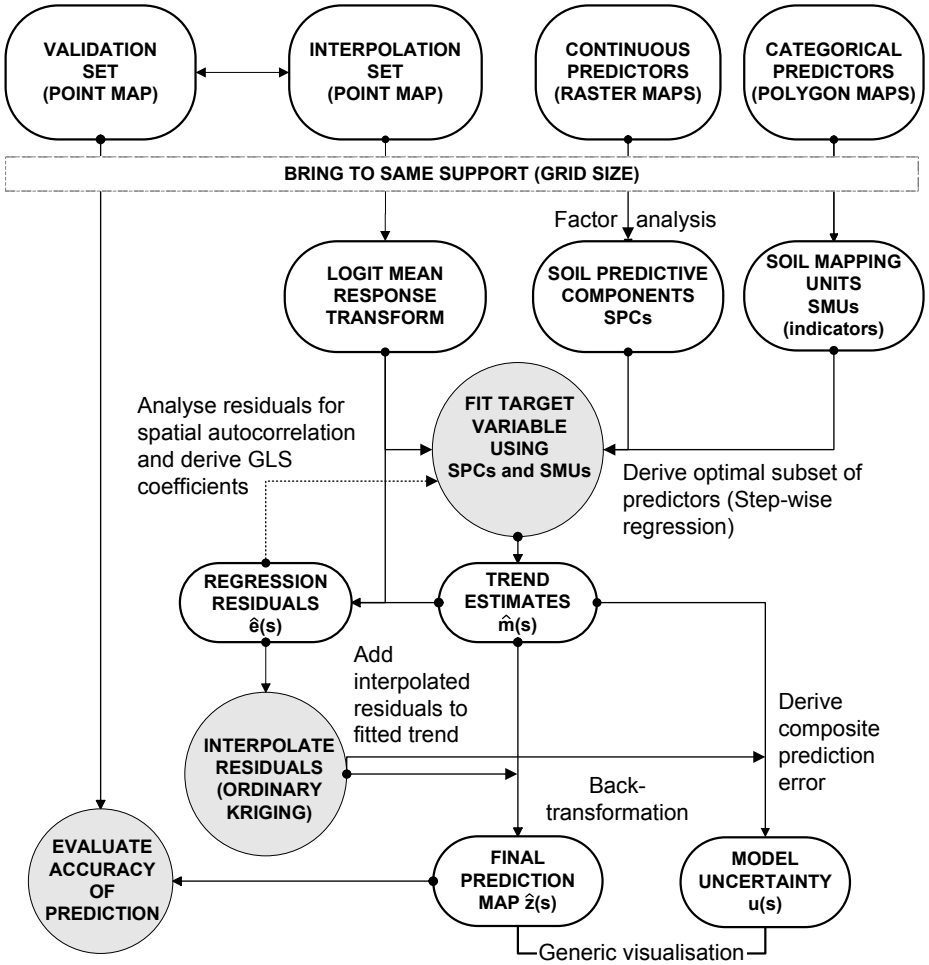

Figure 5.1: Flow diagram: generic framework for regression-kriging (in a GIS).

solution. These short-cuts might be more attractive for practical applications, but are sub-optimal statistically. In further text, we will hold to the term regression-kriging instead of kriging with external drift, as it specifically implies that regression is combined with kriging.

Let the observations of soil variables be denoted as $z(s_1), z(s_2), ..., z(s_n)$, where $s_i = (x_i, y_i)$ is a location and x_i and y_i are the coordinates and n is the number of observations. In the case of RK, a soil property at a new, unvisited location (s_0) is predicted by summing the predicted drift and residuals (Odeh *et al.*, 1994):

$$\hat{z}(s_0) = \hat{m}(s_0) + \hat{e}(s_0) \tag{5.1}$$

where the drift $\hat{m}$ is commonly fitted using linear regression analysis, and the residuals $\hat{e}$ are interpolated using ordinary kriging:

$$\hat{z}(s_0) = \sum_{k=0}^{p} \hat{\beta}_k \cdot q_k(s_0) + \sum_{i=1}^{n} w_i(s_0) \cdot e(s_i); \qquad q_0(s_0) = 1; \tag{5.2}$$

where $\hat{\beta}_k$ are the estimated drift model coefficients, $q_k(s_0)$ is the kth external explanatory variable or predictor at location s_0, p is the number of predictors, $w_i(s_0)$ are weights determined by the covariance function and $e(s_i)$ are the regression residuals. In matrix notation, the RK model is:

$$z = \mathbf{q^T} \cdot \beta + \varepsilon \tag{5.3}$$

where ε is the zero-mean regression residual. The predictions are made by:

$$\hat{z}(s_0) = \mathbf{q_0^T} \cdot \hat{\beta} + \lambda_\mathbf{0}^\mathbf{T} \cdot \mathbf{e} \tag{5.4}$$

where $\mathbf{q_0}$ is vector of $p+1$ predictors at s_0, $\hat{\beta}$ is vector of $p+1$ estimated drift model coefficients, $\lambda_\mathbf{0}$ is vector of n kriging weights and $\mathbf{e}$ is vector of n residuals. The drift model coefficients are preferably solved using the generalized least squares (GLS) estimation to account for spatial correlation of residuals (Cressie, 1993, p. 166):

$$\hat{\beta}_{gls} = \left(\mathbf{q^T} \cdot \mathbf{C^{-1}} \cdot \mathbf{q}\right)^{-1} \cdot \mathbf{q^T} \cdot \mathbf{C^{-1}} \cdot \mathbf{z} \tag{5.5}$$

where $\mathbf{q}$ is the matrix of predictors at all observed locations ($n \times p+1$), $\mathbf{z}$ is the vector of sampled observations and $\mathbf{C}$ is the $n \times n$ covariance matrix of residuals:

$$\mathbf{C} = \begin{bmatrix} C(s_1, s_1) & \cdots & C(s_1, s_n) \\ \vdots & \ddots & \vdots \\ C(s_n, s_1) & \cdots & C(s_n, s_n) \end{bmatrix} \tag{5.6}$$

The covariances between point pairs $C(s_i, s_j)$, under stationarity assumptions also written as $C(h)$, are typically estimated by modelling a variogram (Isaaks & Srivastava, 1989). A common variogram model is the exponential:

$$\gamma(\mathbf{h}) = \begin{cases} 0 & \text{if} \quad |\mathbf{h}| = 0 \\ C_0 + C_1 \cdot \left[1 - e^{-\left(\frac{\mathbf{h}}{R}\right)}\right] & \text{if} \quad |\mathbf{h}| > 0 \end{cases} \tag{5.7}$$

where $\gamma(\mathbf{h})$ is the semivariance function, which is related with the covariance function through $\gamma(h) = C_0 - C(h)$. C_0, C_1 and R are variogram parameters and $|\mathbf{h}|$ is

the Euclidean distance between the point pairs. Thus, RK in matrix notation is (Christensen, 1990):

$$\hat{z}(s_0) = \mathbf{q_0^T} \cdot \hat{\beta}_{gls} + \lambda_\mathbf{0}^\mathbf{T} \cdot (\mathbf{z} - \mathbf{q} \cdot \hat{\beta}_{gls}) \tag{5.8}$$

Note that estimation of GLS residuals is an iterative process: first the drift model coefficients are estimated using ordinary least squares (OLS), then the covariance function of the residuals is estimated and used to obtain the GLS coefficients. These can be used to re-compute residuals and so on. This is the major disadvantage of using KED or RK because both the regression model parameters and variogram parameters need to be estimated simultaneously. To estimate coefficients we need the covariance function of residuals, which can only be estimated after the coefficients. In practice, a single iteration can be used as a satisfactory solution (Kitanidis, 1994), although the optimal approach is to fit these components until convergence (Opsomer *et al.*, 1999).

The variance of the prediction error of RK is the UK variance (Cressie, 1993, p. 155):

$$\begin{aligned}\sigma^2_{\mathtt{E}}(s_0) = (C_0 + C_1) - \mathbf{c_0^T} \cdot \mathbf{C^{-1}} \cdot \mathbf{c_0} \\ + \left(\mathbf{q_0} - \mathbf{q^T} \cdot \mathbf{C^{-1}} \cdot \mathbf{c_0}\right)^\mathbf{T} \cdot \left(\mathbf{q^T} \cdot \mathbf{C^{-1}} \cdot \mathbf{q}\right)^{-1} \cdot \left(\mathbf{q_0} - \mathbf{q^T} \cdot \mathbf{C^{-1}} \cdot \mathbf{c_0}\right)\end{aligned} \tag{5.9}$$

where $\mathbf{c_0}$ is the vector of covariances between residuals at the unvisited and observation locations. The first part of Eq. (5.9) presents the kriging variance of residuals and the second part is associated with the error of estimating the drift. The latter, in statistical terms, is equivalent to the curvature of the confidence band around the regression line (Neter *et al.*, 1996, p. 210). Hence, the composite variance reflects the relative distance in geographical and feature space: the prediction uncertainty increases as the new predicted observation gets further away from observation points spatially and further away from the centre of the attribute or feature space.

5.2.3 Transformations of soil variables

In the case of `CLORPT` techniques, the functional relationship between environmental and soil variables is unknown and often very noisy (e.g. see the correlation plots by Moore *et al.* (1993, p. 448) and Gessler *et al.* (1995, p. 428)). Thus, simple linear regression modelling is most commonly used to model the data. It seems, however, that a general relationship between the soil and auxiliary variables is not necessarily linear. From empirical plots drawn by Buol & Hole (1980), Jenny (1980) and Birkeland (1999, p. 142), it can bee seen that a general relationship between soil

variables and the `CLORPT` factors is sigmoidal. This is often simply because many soil variables reach some physical minimum or maximum after a certain change of a `CLORPT` factor. In this situation, it is more advisable to adjust the model to the data by using some GLM transformation, rather than to adjust the data to the model (Lane, 2002). To approximate such a sigmoidal shape, we used a simple logistic response function (Neter *et al.*, 1996, p. 570):

$$z^{+} = \left[1 + \exp(-\beta^{\mathbf{T}} \cdot \mathbf{q})\right]^{-1} \tag{5.10}$$

The key property of the logit transformation is that it can be easily linearized by transforming the target variable to a logit variable:

$$z^{++} = \ln\left(\frac{z^{+}}{1 - z^{+}}\right); \qquad 0 < z^{+} < 1 \tag{5.11}$$

where z^{+} is the target variable standardised to the 0 to 1 range:

$$z^{+} = \frac{z - z_{\min}}{z_{\max} - z_{\min}}; \qquad z_{\min} < z < z_{\max} \tag{5.12}$$

and $z_{\min}$ and $z_{\max}$ are the physical minimum and maximum of z. This means that all new predicted values are constrained in-between these two limits.
Finally, we obtain the same RK linear equation (Eq. 5.8):

$$\hat{z}^{++}(s_0) = \mathbf{q_0^T} \cdot \hat{\beta}_{gls} + \lambda_{\mathbf{0}}^{\mathbf{T}} \cdot \left(\mathbf{z}^{++} - \mathbf{q} \cdot \hat{\beta}_{gls}\right) \tag{5.13}$$

For ratio-type or percentage-type variables (e.g. clay content, organic matter content etc.), $z_{\min}$ and $z_{\max}$ are given by definition. In other cases, the limits need to be defined using empirical or arbitrary numbers, such as expected or sampled minimum and maximum. For example, we know that a pH of soil, measured in water, in some area will never be below 4 or above 9. Therefore, including these limits will prevent predictions outside the given range. Note that all z values need to be different from the $z_{\min}$ and $z_{\max}$ to avoid $\ln(0)$ situations.

Another advantage of logit transformation is that it can adopt also the categorical data, which first needs to be converted to indicator variable (see later Eq. 5.18). The logit transformation has already been used prior to interpolation of soil data by Triantafilis *et al.* (2001). Gotway & Stroup (1997) used it as a link function prior to universal kriging of a binary target variable.

The predictions are back-transformed to original scale by:

$$\hat{z}(s_0) = \frac{e^{\hat{z}^{++}(s_0)}}{1 + e^{\hat{z}^{++}(s_0)}} \cdot (z_{\max} - z_{\min}) + z_{\min} \tag{5.14}$$

The variance of the prediction error calculated using Eq. (5.9), however, cannot be simply back-transformed as the error is not symmetrical around the regression plane. It can be used, though, to derive confidence limits:

$$\hat{z}_{\pm t}(s_0) = \frac{e^{\left[\hat{z}^{++}(s_0) \pm t \cdot \sigma_{\mathrm{E}}^{++}(s_0)\right]}}{1 + e^{\left[\hat{z}^{++}(s_0) \pm t \cdot \sigma_{\mathrm{E}}^{++}(s_0)\right]}} \cdot (z_{\max} - z_{\min}) + z_{\min} \tag{5.15}$$

where t is the threshold value of standard normal error and $\sigma_{\mathrm{E}}^{++}(s_0)$ is the standard deviation of the prediction error of transformed variable. From confidence limits, the probability density can be reconstructed to get an unbiased estimate of the mean and variance.

A simpler solution is to divide the prediction error of the transformed variable by the total standard deviation of observed samples. This is the normalized mean square error or relative prediction error (Park & Vlek, 2002):

$$\sigma_{\mathrm{E},r}(s_0) = \frac{\sigma_{\mathrm{E}}^{++}(s_0)}{s_{z^{++}}} \tag{5.16}$$

where $s_{z^{++}}$ is the standard deviation of the transformed observations:

$$s_{z^{++}} = \sqrt{\frac{\sum_{i=1}^{n} \left(z_i^{++} - \bar{z}^{++}\right)^2}{n-1}} \tag{5.17}$$

This estimate of the model uncertainty is scale-free and dimensionless. Hence, it will be further on used for visualisation purposes.

5.2.4 Transformation of predictors

To account for multicollinearity, we used a factor analysis prior to regression analysis to produce composite indices or standardised Principal Components (PCs). These are uncorrelated and standardised transforms, and can be then used instead of the original predictors in the regression analysis (Neter *et al.*, 1996, p. 410). Gobin (2000), for example, showed that use of standardized principal components instead of the original predictors improves the prediction for soil-landscape modelling. In addition, a stepwise regression is used as an automatic procedure to derive the 'best' subset of predictors and economize computational effort. Finally, a categorical map (soil map) was incorporated in the regression analysis by using indicator variables. Here, each class (c) in the categorical map resulted in an additional indicator variable:

$$q_c(s) = \begin{cases} 1 & \text{if} \quad q_c(s) = \text{class}(c) \\ 0 & \text{otherwise} \end{cases} \tag{5.18}$$

5.2.5 Evaluation

The performance of interpolation methods can be evaluated using interpolation and validation sets. The interpolation set is used to derive the sum of squares of residuals (SSE) and adjusted coefficient of multiple determination (R_a^2), which describe the goodness of fit:

$$R_a^2 = 1 - \left(\frac{n-1}{n-p}\right) \cdot \frac{SSE}{SSTO} = 1 - \left(\frac{n-1}{n-p}\right) \cdot \left(1 - R^2\right) \tag{5.19}$$

where $SSTO$ is the total sum of squares (Neter *et al.*, 1996), R^2 indicates amount of variance explained by model, whereas R_a^2 adjusts for the number of variables (p) used. In many cases, a $R_a^2 \geq 0.85$ is already a very stratificatory solution and higher values will typically only mean over-fitting of the data (Park & Vlek, 2002). Note that this number corresponds to the relative prediction error (Eq. 5.16) of $\geq 40\%$.

The true prediction accuracy can be evaluated by comparing estimated values ($\hat{z}(s_j)$) with actual observations at validation points ($z^*(s_j)$) in order to assess systematic error, calculated as mean prediction error (MPE):

$$MPE = \frac{1}{l} \cdot \sum_{j=1}^{l} \left[\hat{z}(s_j) - z^*(s_j)\right] \tag{5.20}$$

and accuracy of prediction, calculated as root mean square prediction error ($RMSPE$):

$$RMSPE = \sqrt{\frac{1}{l} \cdot \sum_{j=1}^{l} \left[\hat{z}(s_j) - z^*(s_j)\right]^2} \tag{5.21}$$

where l is the number of validation points. In order to compare accuracy of prediction between variables of different type, the $RMSPE$ can be normalized by the total variation, as in Eq. (5.16),:

$$RMSPE_r = \frac{RMSPE}{s_z} \tag{5.22}$$

As a rule of thumb, we can consider that a value of $RMSPE_r$ close to 40% means a fairly satisfactory accuracy of prediction. Otherwise, if the values get >71%, this means that the model accounted for less than 50% of variability at the validation points and the prediction is unsatisfactory.

5.2.6 Visualisation

A typical result of (geo)statistical interpolation is a map of predictions and prediction error, which is an estimate of prediction uncertainty. These two are commonly not visualised simultaneously. This can be achieved by using the pseudo colour scale and image calculations on colours, following the Hue-Saturation-Intensity (HSI) colour model (Hengl *et al.*, 2002). We suggest the following procedure. First the prediction values need to be transformed to the hue angle by:

$$\varphi_1 = -90 + z_r \cdot 300 \tag{5.23}$$

$$\varphi_2 = \left\{ \begin{array}{cc} \varphi_1 + 360 & \text{if} \quad \varphi_1 \leq -360 \\ \varphi_1 & \text{if} \quad \varphi_1 > -360 \end{array} \right\} \tag{5.24}$$

where φ is the hue angle in degrees measured clockwise and z_r are the predictions ($z_r \in [0, 1]$). The predictions and uncertainty (relative error) are then coded to HSI image by:

$$H = (\varphi_2 + 360) \cdot \frac{240}{360} \tag{5.25}$$

$$S = (1 - u_r) \cdot 240 \tag{5.26}$$

$$I = (1 + u_r) \cdot 120 \tag{5.27}$$

where φ is the hue angle in degrees measured clockwise, z_r are the predictions and u_r is the prediction uncertainty ($u_r \in [0, 1]$). Note that these values have to be stretched before coding by using:

$$z_r = \frac{\hat{z} - z_1}{z_2 - z_1} \tag{5.28}$$

$$u_r = \frac{\sigma_{\mathrm{E},r} - u_1}{u_2 - u_1} \tag{5.29}$$

where $\hat{z}$ is the prediction map derived using Eq. (5.13) and back-transformed using Eq. (5.14), $\sigma_{\mathrm{E},r}$ is the relative prediction error map derived using Eq. (5.9) and standardised using Eq. (5.16), z_1 and z_2, and u_1 and u_2 are the lower and upper inspection range limits for the predicted values and relative prediction error. From the HSI images, the RGB composite image can be derived in ILWIS using (Unit Geo Software Development, 2001):

$$z_{RGB} = colorhsi\,(H, S, I) \tag{5.30}$$

Note that, from Eqs (5.23) to (5.27), the lower values are coded bluish (hue angle from -90° to -150°) and highest values are coded reddish (hue angle from -330° to -30°). Consequently, the intermediate values are coded with cyan, green and yellowish (Fig. 5.2b). This model corresponds to the pseudo-colour scale used in many GIS packages for visualising continuous variables. Also note that a part of the hue circle representing magenta (-30° to -90°) is omitted to avoid confusion between high and low values. The second property of the HSI-coded image is that uncertainty is coded with whiteness. This has often proven to be the most suitable colour variable for visualisation of uncertainty (Jiang *et al.*, 1995). In this case, fully saturated colour indicates lowest uncertainty and white colour indicates full uncertainty within the given thresholds.

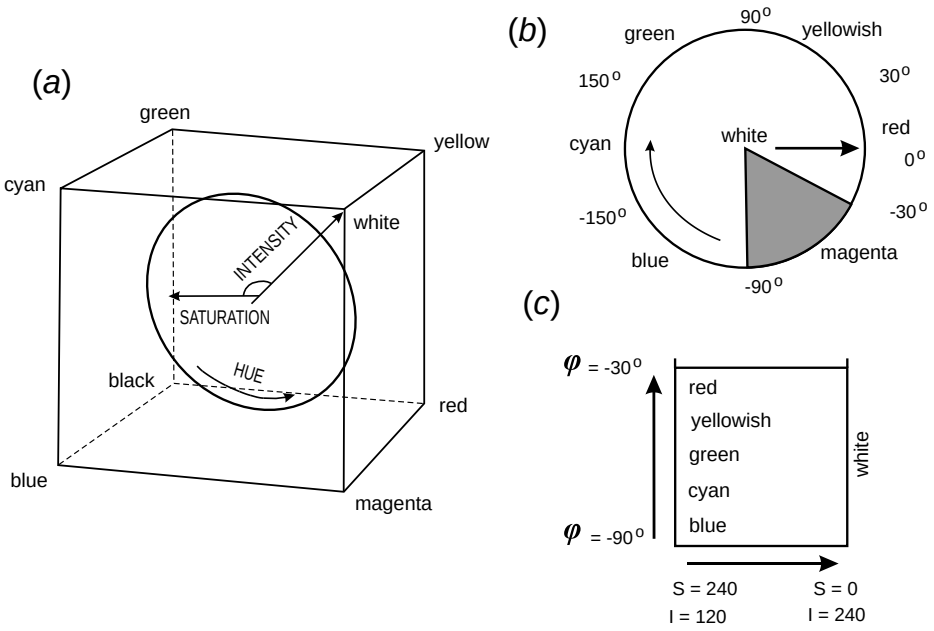

Figure 5.2: Hue-Saturation-Intensity colour model in Red-Green-Blue colour cube (a), main hue types used for the visualisation (b) and the same shown using a two-dimensional legend (c). φ is the hue angle in degrees measured clockwise. See text for explanation.

In addition to the colour map, we developed a special two-dimensional legend (see Fig. 5.2c and Fig. 5.8e) to accompany the HSI-coded image. The vertical direction indicates change of prediction values (from -90° to -30°), while the horizontal

direction indicates uncertainty and is coded with a linear increase of both intensity and saturation, i.e. whiteness. This visualisation algorithm can be applied in any image processing or general GIS package, which allows calculations on colours.

5.2.7 Case study

A set of 135 profile observations from the Croatian national Soil Geographical Database (Martinović & Vranković, 1997) was used as a case study. It was randomly divided into a interpolation (100 points) and validation set (35 points). The study area is a 50×50 km square located in the central part of Croatia. As target variables, the organic matter in the topsoil (OM), measured using a colorimetric wet oxidation method and expressed in %, the topsoil pH measured in H_2O (PH) and thickness of topsoil horizon expressed in cm (DEPTH) were used.

As predictors (auxiliary maps), we used five relief parameters derived from the 100×100 m resolution elevation data: elevation (DEM), slope (SLOPE), mean curvature (MEANC), Compound Topographic Index (CTI), Stream Power Index (SPI) and viewshed (VSHED) (Fig. 5.3), all derived in ILWIS (Hengl *et al.*, 2003b). These were first linearly stretched in an image processing software to a range of 0–255 to give each map equal contrast and then transformed to standardised principal components (further referred to as the Soil Predictive Components SPCs) using factor analysis in ILWIS. The 1:300 K soil map of Croatia was used as the categorical layer (Bogunović *et al.*, 1998). There were 28 soil mapping units (further referred to as SMU) in the study area. Due to a low number of points in the interpolation set, we first reduced the number of units to nine by merging some taxonomically adjacent units. Finally, there were six SPCs and nine SMUs making 15 predictors in total.

5.2.8 Data analysis

The OM, $DEPTH$ and PH were first transformed using Eq. (5.11):

$$\begin{aligned} OM^{++} &= \ln\left(\frac{OM^{+}}{1-OM^{+}}\right) \\ PH^{++} &= \ln\left(\frac{PH^{+}}{1-PH^{+}}\right) \\ DEPTH^{++} &= \ln\left(\frac{DEPTH^{+}}{1-DEPTH^{+}}\right) \end{aligned} \tag{5.31}$$

where the OM^{+}, PH^{+} and $DEPTH^{+}$ are values standardised to 0 to 1 scale:

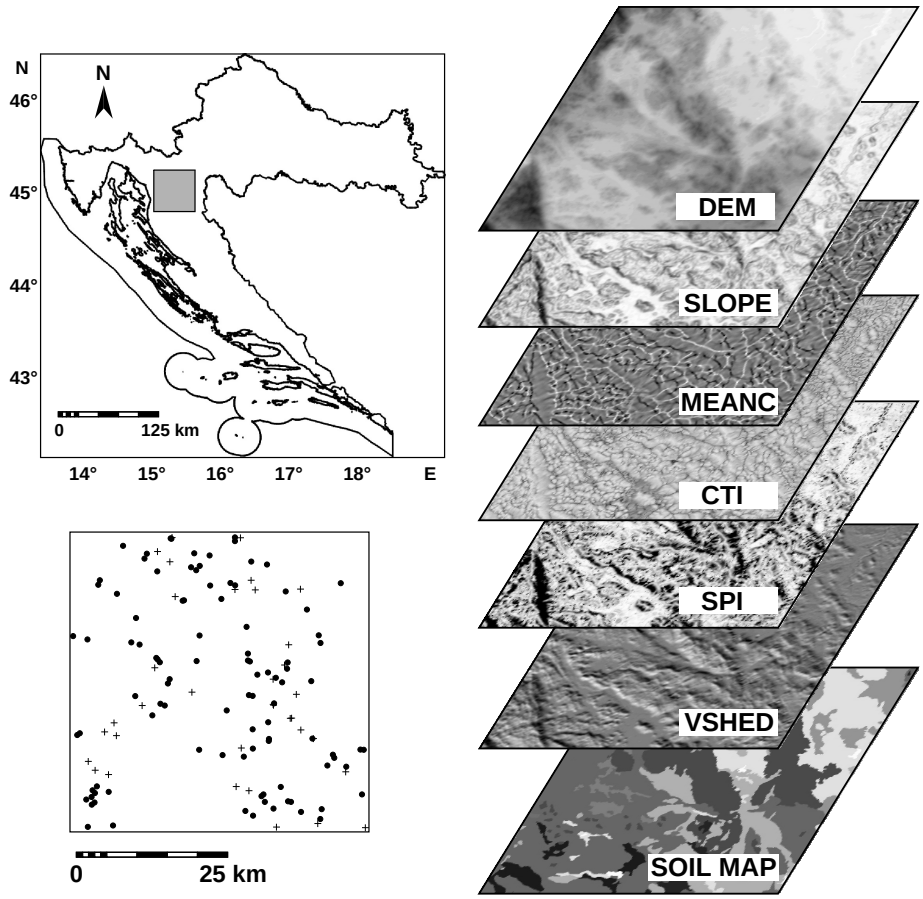

Figure 5.3: Location of the study area (upper-left), profiles used for interpolation (•) and validation (+) (lower-left) and maps of predictors (right).

$$
\begin{aligned}
OM^{+} &= \frac{OM - 0}{100 - 0} \\
PH^{+} &= \frac{PH - 4.1}{8.8 - 4.1} \\
DEPTH^{+} &= \frac{DEPTH - 0}{150 - 0}
\end{aligned}
\tag{5.32}
$$

In the case of *OM* and *DEPTH* we used the minimum and maximum values measured in the whole of Croatia. Also note that zero measurements need to be replaced with an arbitrary small number, e.g. the precision of measuring a variable in the laboratory or in the field.

First, prediction maps of *OM*, *PH* and *DEPTH* were made from the soil map by averaging profiles per SMU (Burrough, 1993a). Second, the variables were interpolated using ordinary kriging (OK) and OLS multiple regression (MR). These three prediction methods were then compared with the RK within the generic framework.

The spatial dependence structure of soil variables and residuals was modelled in VESPER using automated variogram fitting (Minasny *et al.*, 2002). We used an exponential model and a limiting distance of 25 km in all cases. In addition, the variogram modelling in VESPER gave the Akaike Information Criterion (AIC), which was used to compare different models for the goodness of fit (McBratney & Webster, 1986). The GLS coefficients were used to derive the drift maps using the map calculation in ILWIS. The re-estimated residuals were then interpolated and added to the fitted drift. The final estimates were then back-transformed to their original scale using the Eq. (5.14). Matrix calculations and fitting of the target variables using the stepwise regression was done in the S-PLUS statistical package (MathSoft Inc., 1999). Although most of the processing steps are feasible with a standard PC, the calculation of the variance of the prediction error (Eq. (5.9)) can be time-consuming, even for smaller size data sets.

5.3 Results

5.3.1 Regression modelling

Inspection of the distributions of target variables at primary locations showed that both *OM* and *DEPTH* have positively skewed distributions (Table 5.1). Also the predictors, especially SPI, CTI and MEANC show distinct asymmetry in their distributions. Similarly, the first univariate linear regression models showed that the residuals are skewed around the regression line and therefore do not satisfy the normality requirement for both regression analysis and kriging (Fig. 5.4b). In this case, the correlation test of normality for residuals (Neter *et al.*, 1996, p. 111) gave coefficient of correlation of 0.912 (OM) and 0.983 ($DEPTH$) between ordered residuals. Note that the critical value of coefficient of correlation between ordered residuals for $n = 100$ and 0.05 level of significance is 0.987 (Looney & Gulledge, 1985), which means that both variables significantly depart from the normal distribution. After the logit transformation of the target variables (OM^{++}, PH^{++}, $DEPTH^{++}$), the models became symmetrical around the fitted linear models, and the coefficients of correlation between ordered residuals were higher for OM^{++} (0.982) and simi-

lar for $DEPTH^{++}$ (0.981). This was also reflected in an increased R^2 (Fig. 5.4c). Note that for *DEPTH* under simple linear modelling, the predictions in areas of high slope (SLOPE> 60%) would yield negative estimates. The residuals for $DEPTH^{++}$, however, show skewness even after the transformation. In this case, this was a reflection of the log-normal distribution of SLOPE. Note that the observed relationships (Fig. 5.4d) correspond to the hypothetical plots described by Buol & Hole (1980) (Fig. 5.4a). This phenomenological correspondence is an extra guarantee to apply these models in spatial prediction.

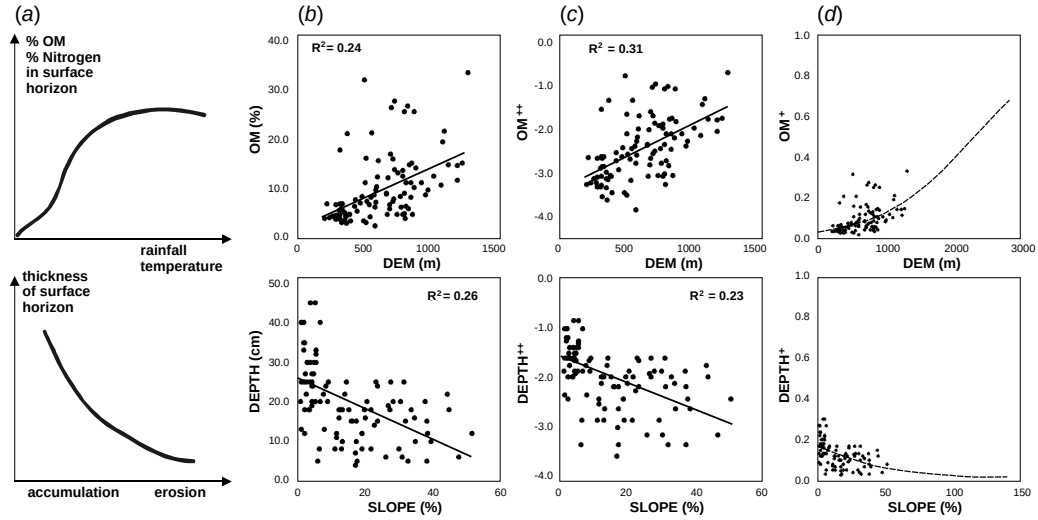

Figure 5.4: Comparison of empirical relationships (a), observed relationship (b), observed after the logit transformation (c) and back-transformed models (d).

The results of the factor analysis showed that there is an overlap in information and that the data can be reduced. The first four SPCs accounted for more than 90% of the total variation in the bands (42.6%, 21.3%, 14.5% and 12.0%). SPC1 as the main component was explained by variation in SLOPE, SPI, CTI and DEM. SPC2 accounted mainly for the variation in MEANC and SPI, while the third and fourth component accounted for DEM and VSHED. The fifth and sixth components showed some features already seen in the first four components and probably represent noise and artefacts in the relief parameters. Note that the SPCs show much lower skewness and kurtosis than the original predictors (Table 5.1).

The step-wise regression substantially reduced the number of predictors. In the case of OM^{++}, it selected SPC1, SPC4 and SMU8 as the optimal sub-set for

Table 5.1: Descriptive statistics for target variables, predictors and their transforms: MEAN – mean, STDEV – standard deviation, MED – median, MIN – minimum, MAX – maximum, SKEW – skewness, KURT – kurtosis.

	Target variables				Predictors				
	OM	PH	DEPTH	DEM	SLOPE	MEANC	CTI	SPI	VSHED
	%	-	cm	m	%	m^{-1}	-	-	-
MEAN	9.7	6.25	19.8	634	15.1	-1.19	9.1	90.3	0.63
STD	6.8	0.88	9.6	267	12.9	6.86	2.5	97.8	12.17
MED	7.3	6.20	18.5	604	12.3	0.23	8.5	56.8	1.00
MIN	2.1	4.50	4.0	207	0.9	-27.47	5.4	0.0	-30.60
MAX	33.4	7.70	45.0	1298	51.1	14.14	17.7	482.8	33.00
SKEW	1.51	-0.18	0.57	0.41	0.85	-1.12	1.14	1.60	0.01
KURT	2.00	-0.99	-0.01	-0.53	-0.23	1.97	1.01	2.71	0.98

	Transforms					
	SPC1	SPC2	SPC3	SPC4	SPC5	SPC6
MEAN	103	-48	86	79	-184	56
STD	105	55	55	53	31	28
MED	104	-38	95	74	-185	48
MIN	-84	-192	-61	-47	-244	-36
MAX	362	54	207	217	-96	138
SKEW	0.12	-0.55	-0.56	0.04	0.21	0.10
KURT	-0.99	-0.32	0.16	0.28	-0.28	0.78

prediction, while in the case of $DEPTH^{++}$, the algorithm selected SPC1, SPC3 and SPC4. In both cases the correlation was significant ($R_a^2 = 0.33$ for OM^{++} and $R_a^2 = 0.40$ for $DEPTH^{++}$). In the case of PH^{++}, the coefficient of multiple determination was small ($R_a^2 = 0.14$), but still significant at the 0.05 level, indicating weak correlation with the predictors. Here, the only significant predictors were

SMU5 and SMU6. The normality test for residuals showed that in all cases the residuals did not depart significantly from a normal distribution with coefficient of correlation between ordered residuals of 0.986 for OM^{++}, 0.981 for PH^{++} and 0.989 for $DEPTH^{++}$ (see also density histograms in Fig. 5.6b).

5.3.2 Geostatistical analysis

Both OM^{++} and $DEPTH^{++}$ showed a clear spatial dependence, whereas the variogram of PH^{++} was dominated by a pure nugget effect (Fig. 5.5a). For OM^{++} automated variogram modelling gave a small nugget and shorter range parameter (3 km) (Fig. 5.5b), whereas for $DEPTH^{++}$, the range parameter was fairly large (11.2 km). Analysis of spatial correlation of residuals reflected the success of regression fitting: the range of spatial dependence was much shorter and the sill was proportionally smaller to the variation accounted by regression modelling. Moreover, variograms of residuals tend to show a shorter range and bounded sill, which indicates that the drift has indeed been removed. This is especially distinct for $DEPTH^{++}$ where the target variable showed almost an unbounded variogram, whereas the residuals showed an almost five times shorter range of spatial dependence and 43% smaller sill (Fig. 5.5c). The AIC confirms that the variograms of residuals are somewhat easier to fit. Here, the best fit, i.e. the smallest AIC, was obtained for the residuals of $DEPTH^{++}$.

Regression coefficients using OLS and GLS estimation are given in Table 5.2. The differences between coefficients were in all cases relatively small, which indicates that there is no significant spatial clustering between the points.

5.3.3 Bias and accuracy of prediction

A problem with the logit transformation is that the back-transformation gives only an unbiased estimate of the median, as for example in the case of log-normal kriging. This is usually reflected in somewhat lower predictions, especially if values are grouped around zero. In this case, the mean of the interpolation set ($z(s_i)$) was somewhat higher than the mean of fitted values ($\hat{z}(s_i)$): 9.7% as compared to 9.1% for OM (or 6.7% lower in relative measures) and 19.8 cm compared to 19.5 cm for $DEPTH$ (or 1.5% lower). Comparison of histograms of the prediction maps for the same properties, also gave somewhat lower means of 8.9% for OM (or 8.2% lower in relative measures) and 18.8 cm for $DEPTH$ (or 5.1% lower) (Fig. 5.6c). The medians in the prediction maps, however, are somewhat higher than the medians at primary locations: 8.0% compared to 7.3% for OM and 19.0 cm compared to 18.5 cm (Fig. 5.6c). It seems, therefore, that there is no need for an unbiased back-transformation as the histograms, before and after the back-transformation, in

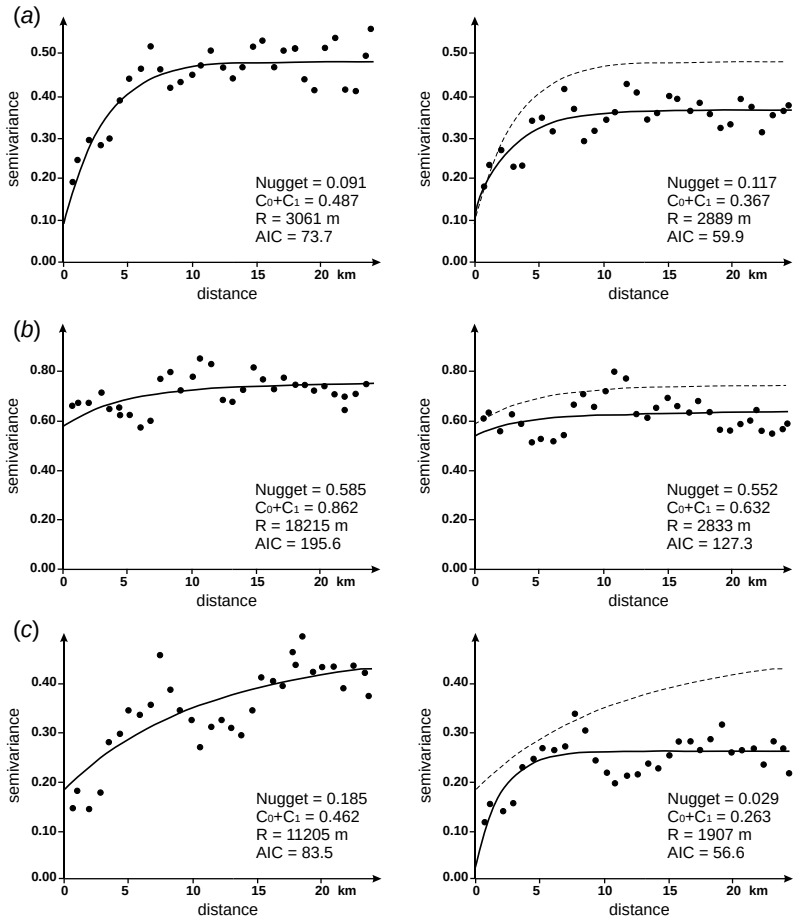

Figure 5.5: Semivariograms of target variables (left graphs and dotted line) and their residuals (right graphs): OM^{++} (a), PH^{++} (b) and $DEPTH^{++}$ (c). All fitted in VESPER using an exponential model.

general match. Note that the ranges in the prediction maps are somewhat narrower due to the smoothing effect of RK.

The summary comparison of prediction methods at the validation points is presented in Table 5.3. Note that the GLS coefficients result in somewhat higher *SSE*. For example, in the case of OM^{++}, the *SSTO* is 51.4; after the regression analysis the *SSE* has decreased to 34.7, meaning that 33% of the variation has been explained by the model. The GLS estimation, however, resulted in a somewhat higher *SSE* (36.7). Finally, the *SSE* after kriging the residuals has decreased to 5.3, indicating

Table 5.2: Summary results of the step-wise regression analysis for OM^{++}, PH^{++} and $DEPTH^{++}$ and improved coefficient estimates (n=100).

Target variable	Selected Predictors	Regression coefficients (OLS[a])	Regression coefficients (GLS[b])
OM^{++}	intercept	-3.124	-3.161
	SPC1	0.003228	0.003143
	SPC4	0.004843	0.005468
	SMU8	-0.7712	-0.4844
PH^{++}	intercept	-0.4258	-0.4194
	SMU5	0.6598	0.7010
	SMU6	0.9183	0.7008
$DEPTH^{++}$	intercept	-1.667	-1.664
	SPC1	-0.003212	-0.003131
	SPC3	0.002264	0.002189
	SPC4	-0.002726	-0.002823

[a]OLS - Ordinary least square estimation;
[b]GLS - General least square estimation based on the spatial covariance matrix of residuals.

that the RK model accounted for almost 90% of the total variation at the primary location grids (100×100 m). In the case of OM, RK achieved slightly better relative prediction accuracy than OK (53.3% versus 66.5%). In both cases the bias was small. Similarly, RK achieved a higher accuracy of prediction (66.5% versus 83.3%) and a smaller bias (0.15 versus 0.69 cm) for predicting $DEPTH$, when compared with OK. In general, the soil map was shown to be an inefficient predictor in all cases except for prediction of OM in topsoil. Relatively low bias for RK in all cases indicates that the logit transformation and use of SPCs served their purpose. The prediction of PH has proven to be difficult with a relative prediction error greater than 100%. This means that all compared methods can show almost any value within the given range and there is no justification for production of a PH map.

Comparison of different prediction methods for mapping $DEPTH$ is shown in

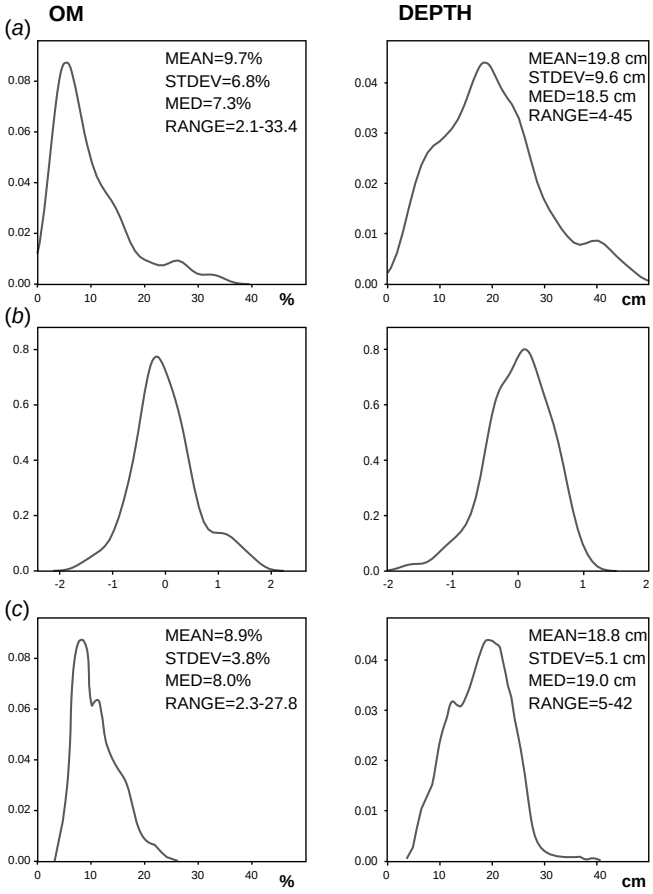

Figure 5.6: Density histograms and summary statistics for: target variables (*OM* and *DEPTH*) at primary locations (a), GLS residuals (b) and prediction maps (c). MEAN – mean, STDEV – standard deviation, MED – median and RANGE – range.

Fig. 5.7. The soil map in general over-smoothed the values, except for two SMUs (Fig. 5.7b). On the contrary, OK map (Fig. 5.7c) shows rather gradual transitions with fairly low level of detail, whereas the RK map (Fig. 5.7d) reflects change in elevation, slope and exposition. Finally the RK map (Fig. 5.7e) yields more detail than the OK map, at the same time showing the hot-spots not visible in the MR map.

Visualisation of the predicted *DEPTH* map together with the prediction er-

Table 5.3: Comparison of interpolation methods for goodness of fit (R_a^2), bias (*MPE*) and accuracy of the prediction at validation points (*RMSPE*).

		Interpolation set		Validation set		
	Method[a]	SSE[b]	R_a^2	*MPE*	*RMSPE*	$RMSPE_r$[c]
OM (%)	SOIL	-	-	-1.28	5.3	68.2%
	OK	3.2	-	0.01	5.2	66.5%
	MR	34.7	0.31	-0.10	3.4	44.1%
	RK	5.3	-	-0.04	4.2	53.3%
PH (-)	SOIL	-	-	0.11	1.024	128.1%
	OK	49.7	-	0.00	0.932	116.6%
	MR	64.5	0.13	0.06	0.892	111.5%
	RK	50.4	-	0.01	0.885	110.7%
DEPTH (cm)	SOIL	-	-	1.41	9.1	88.7%
	OK	10.4	-	0.69	8.5	83.3%
	MR	23.4	0.40	1.69	8.8	85.4%
	RK	0.7	-	0.15	6.8	66.5%

[a] SOIL – prediction from the soil map only; OK – ordinary kriging; MR – multiple regression; RK – regression kriging.

[b] Values for transformed variables (OM^{++}, PH^{++} and $DEPTH^{++}$).

[c] $RMSPE_r$ – relative prediction error (%).

ror is given in Fig. 5.8. The composite variance of RK reflects both the arrangement of points in geographical space (the kriging variance of residuals) and areas of extrapolation in attribute space. Note that the areas of higher slopes have been under-sampled (diagonal strips), which is also reflected in the prediction error map (Fig. 5.8b). This corresponds to previous results by Papritz & Stein (1999, p. 112), for example.

The combined visualization gives insight into the relationship between uncertainty and input data for the given thresholds. In this case, we visualised prediction of *DEPTH* using the following thresholds: z_1=5 and z_2=30 cm for the predictions and u_1=0.40 and u_2=0.80 and 1.00 for the errors (Fig. 5.8). The corrected brightness values are then: (a) equal to the original RGB for a relative uncertainty equal

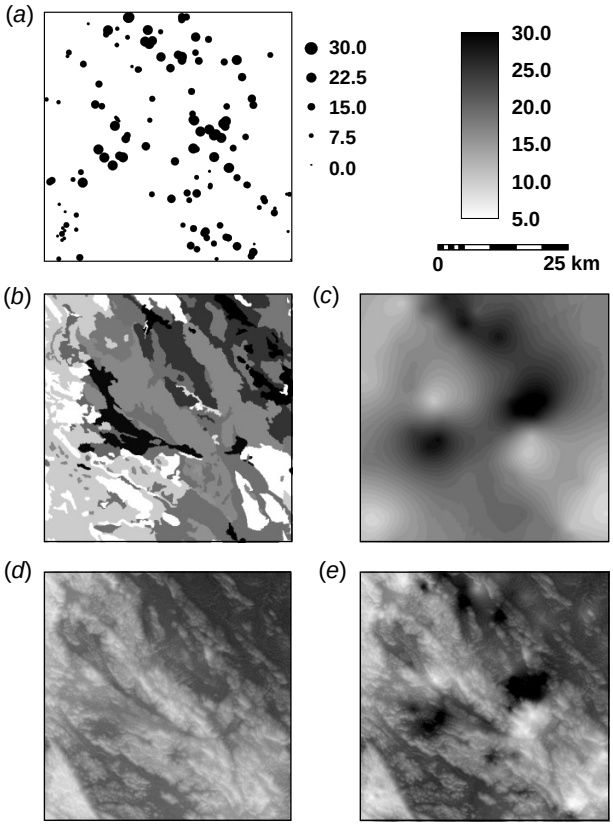

Figure 5.7: Topsoil thickness (*DEPTH*) measured at 135 locations (a), comparison of predictions made by using: soil map only (b), ordinary kriging (c), plain regression (d) and regression-kriging (e). Note that the hot spots re-appear in the RK prediction map.

or less than 0.40, and (b) completely white for relative uncertainty equal or higher than 0.80 or 1.00. In the first case (Fig. 5.8c), the visualisation resulted in most of the map distant from the points being pale, while in the case of a maximum feasible threshold (1.00), the HSI-coded image shows that prediction was efficient in most of the study area (Fig. 5.8d). A visual comparison between the HSI-coded RK and OK (Fig. 5.8f) maps shows that the OK predictions are somewhat less certain and, consequently, the colours are less distinct[2].

[2]See the supplementary materials for full-colour animation of prediction uncertainty.

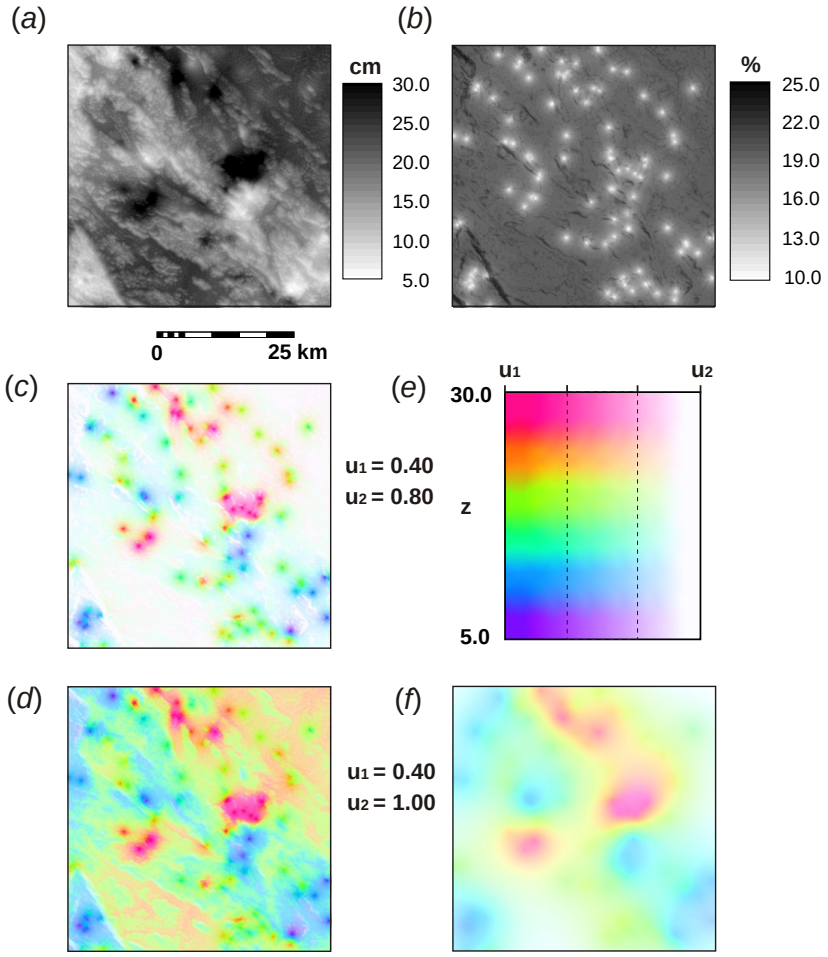

Figure 5.8: Generic visualisation of the RK prediction map for *DEPTH* in cm (a) and relative error (b), the HSI colour images with two inspection ranges (c and d), two-dimensional legend (e) and referent OK prediction map visualised using the same thresholds (f).

5.4 Conclusions and discussion

In this study we integrated several methodological steps to provide a framework for generic spatial prediction and visualisation of soil data. The key principle was to employ most of the available regression and kriging methods and let the system

exploit 'the best' of the data. The results show that the proposed methodology improves prediction efficiency, while ensuring a relative normality of residuals and predictors. Especially the logit transformation proved to be a useful step to model non-linear relationships and force prediction values to be within the physical limits. It is also attractive for a general case because it can be used to model both linear and curvilinear relationships with one or two inflection points and both quantitative and categorical target variables. The factor analysis on map sets was efficiently used to remove multicollinearity and reduce asymmetry in distributions. This helped the step-wise regression algorithm to come to an optimal subset of uncorrelated predictors. When the SPCs are at the same scale, then also the regression coefficients can be directly compared. Finally, visualisation of both predictions and prediction uncertainty offers a possibility to enhance visual exploration of the data uncertainty and make comparisons between different prediction methods. In several aspects of the developed framework, we advocate use of flexible statistical methods, such as factor analysis, step-wise selection of an optimal subset of predictors, logit transformation and automated variogram fitting. These flexible methods open a possibility to develop a user-friendly bundle algorithm that can be implemented in a GIS. Eventually, a user will be able to select a point map and maps of predictors, define some minimum needed criteria and then run the spatial prediction at once.

Recently, sources of auxiliary data are increasingly available from digital terrain modelling parameters to various air- and space-born remote sensing images. There is much auxiliary information at hand nowadays, even at farm level, i.e. for precision agriculture (McBratney *et al.*, 2003). The auxiliary variables in this study were cheaply obtained by digital terrain analysis, which makes the method inexpensive. Hence, the plain geostatistical methods are likely to be replaced with the regression-kriging techniques. One should keep in mind that both ordinary kriging and regression analysis are only special cases of one universal method of spatial prediction. In some cases, however, there will be no help from auxiliary maps and in other cases there will be no need to apply kriging (e.g. in the case of the pure nugget effect). Here, the key measures to decide on which method to use can be, for example, the correlation strength with auxiliary variables and distance at which semivariance reaches 90% of the sill (Fig. 5.9).

The limitations of RK are that it is more complex technique and, if misused, can give even worse estimates than straightforward ordinary kriging (Goovaerts, 1999a). Therefore, development of a fully automated generic method is still unrealistic. For example, we experienced problems with automatic fitting of the variogram functions in VESPER. Automatically fitted variogram parameters for *PH* did not show any physical meaning and needed to be adjusted by hand. This asks for a set of additional remedial measures. Similarly, we cannot guarantee that the sigmoidal shape is truly

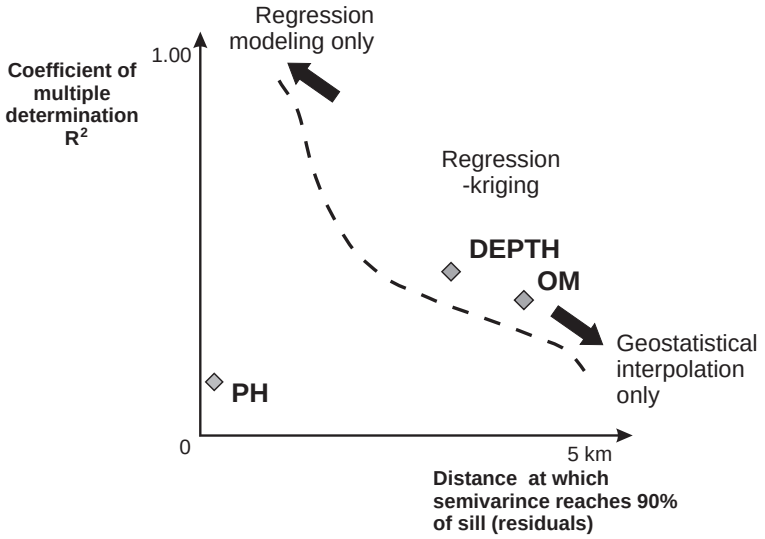

Figure 5.9: Regression-kriging, plain geostatistical and regression techniques in relation to the correlation coefficient (R^2) and distance at which semivariance reaches 90% of sill (spatial auto-correlation of residuals). Spatial prediction of pH was inefficient.

generic for all cases. There will be cases with more inflection points in the correlation plots, which will be more difficult for this framework to handle. Nevertheless, logit transformation has proven to be more beneficial for prediction than a simple linear regression.

Another constraint of RK is the number of point samples required to fit the regression model. Usually a large number of samples is needed to fit some 10–20 environmental variables. As a rule, Draper & Smith (1998) suggest at least 10 complete sets of observations for each potential variable to be included, while Ott & Longnecker (2001) show that the real minimum is $2p + 20$, where p is the number of predictors. In this case, due to the use of indicator variables, the size of the interpolation set was fairly close to the minimum required number (15 predictors to fit 100 points). It should be also emphasized that a point data set with a fairly equal spreading of points is more appropriate for regression-kriging, which is not a requirement for the plain `CLORPT` techniques.

In this study we have only dealt with the spatial (2D) aspect of soil variability. Note that there are three more aspects of soil variability that also play a role: temporal variability, depth (3D) and support size (Florinsky *et al.*, 2002). Spatial prediction of *PH* was probably limited due to measurement errors, high local vari-

ation and overseen factors. From the database description, we could not conclude at which part of the season the data were collected and what was the measurement error. If measurement errors are large and if data were collected during different seasons, not even the most optimal interpolator would make usable predictions. In such cases, a larger number and better quality of soil environmental variables should be used to improve predictions. Eventually, if not even these measures are useful, more intensive sampling strategies at fixed conditions (same season, same depths, same blocks of land) are required.

The next steps will be to integrate this statistical framework into a GIS package and provide a user-friendly procedure, which can be used to interpolate existing profile datasets. The methodology can then be extended so that it includes the temporal and internal (depth) variability of soil variables as well. Finally, the following three topics seems to be especially challenging for the future research:

- Development of soil genesis simulation models rather than just data fitting techniques (Heuvelink & Webster, 2001). Some rudimentary applications already exist (Minasny & McBratney, 2001).
- Development of robust automated variogram modelling algorithms that will account for serious outliers and over-parameterisation or over-fitting of the data.
- Further integration of geostatistical modules within GIS packages. Here, a promising development is the integration of geostatistical packages such as GSTAT with open-source statistical packages such as R or GIS packages like GRASS (Pebesma, 2003).

Chapter 6

Continuous categorical map*

"Traditional pedologists who use tacit knowledge and field morphological properties still have reservations towards such computerized quantitative applications. One major question emerges: can we reconcile these two different paradigms: quantitative/mathematical versus classification?"

[by S. Grunwald in the Pedometron #13, "The Dilemma of Pedometrics in the U.S.", available via www.pedometrics.org]

*based on: Hengl T., Walvoort D.J.J., Brown A. and Rossiter D.G., 2004. A double continuous approach to visualisation and analysis of categorical maps. International Journal of Geoinformation Science, accepted for publication, Vol. 18(1/2), in press.

6.1 Introduction

A GIS representation of a natural resource traditionally conforms to the *discrete model of spatial variation*, which is commonly modelled using object-based or polygon-based GIS (Goodchild *et al.*, 1993). In such system, the natural resource is divided into a set of discrete and mutually-exclusive classes, whose spatial distribution are traditionally presented using a sets of different colours in the so called *"choropleth"*, or more precisely *"chorochromatic"* map (Burrough & McDonnell, 1998). In GIS terms, this is a polygon map with fully defined adjacency, while the properties are recorded in attribute tables keyed to the polygon identifier. Since the geographical units are modelled as discrete objects and the taxonomic entities as discrete classes, this approach is referred to as the double crisp approach (Burrough *et al.*, 1997). For example, a soil surveyor uses sharp boundaries to delineate soil bodies (polypedons) on the landscape, and crisp classes of soil types to classify the typical soil individual found in the body. Variation of soil types within the crisp boundary may be mentioned in the linked database, however an impression of having crisp objects remains. Similar approaches are followed for other natural resources, including forest vegetation types, geomorphic classes, land cover classes, and geological units. In all these cases, the mapped objects are, in fact, usually not so crisp as the map indicates. On the contrary, these are often continuous in both their properties and distribution. In particular, soil units have always been considered to be poorly identifiable in both respects, being *"arbitrarily defined classes of mostly hidden objects"* (Burrough, 1993b). Similarly, classes of geoforms or landforms are hard to separate, as the geographers often emphasize (Fisher & Wood, 1998).

The alternative, *continuous model of spatial variation* and related field approach to GIS modelling (Heuvelink, 1998) was first utilised in geostatistical interpolations to produce maps of single continuous ratio or interval-scale variables (Burrough & McDonnell, 1998). With the emergence of the theory of fuzzy logic (Zadeh, 1965), it become interesting to model categorical variables, such as soil and landform classes. Here each class first has to be mapped separately as single class map, usually as the membership grade or membership value (μ), ranging from 0 (no membership in the class) to 1 (full membership). Consequently, a variable with n classes will result in several (n) single maps, i.e. multiple memberships maps. These are continuous categorical or fuzzy maps, with main difference from the conventional maps is that they can be analysed for ambiguity or indistinctness of a specific class and the overall confusion among all classes (Hootsmans, 1996).

In the past decade, a number of methods have been proposed to deal with fuzzily-defined natural objects. Lagacherie *et al.* (1996) drew the fundamental distinction between categorical and geographical fuzziness, and proposed a classification of boundaries between soil map units on this basis. Irvin *et al.* (1997) used five

landform parameters to produce membership maps of landform classes using fuzzy k-means classification. The classification of landforms was then further on discussed and improved by Burrough *et al.* (2000), who suggested an automated procedure to select the number of classes based on the classification entropy. Similar applications where continuous categorical maps are used instead of double-crisp maps can be found for vegetation (Brown, 1998), land cover (Hansen & Reed, 2000) and land evaluation studies (Hall *et al.*, 1992; de Gruijter & Boogaard, 2001). Woodcock & Gopal (2000) examined accuracy assessment for fuzzy maps, and concluded that they can give better estimates of areas covered by crisply-defined classes then the conventional double-crisp maps.

Single membership maps are easily visualised by grey or pseudo-colour scales that are related to strength of membership. However, each map is separate, which does not allow the user to visualise the membership maps as a whole and understand properties of the whole set of classes, e.g. overall fuzziness and relations between multiple classes. Visualisation of fuzziness and uncertainty is important as it allows users to explore it and investigate the effects of different decisions in the classification process (MacEachren & Kraak, 1997). This is not an easy task for maps of natural resources having tens of classes and resulting membership maps.

The simplest cartographic procedure commonly used to display the membership maps is *defuzzification* (Burrough *et al.*, 1997). Here a crisp colour map is produced by assigning the class with the highest membership in an individual pixel. Since defuzzification applies to individual pixels, it provides no means to visualise the nature of boundaries or the uncertainty of classification, although there have been some investigations in this direction (Fisher, 1993; Hootsmans, 1996, 7). de Gruijter *et al.* (1997) were among first to develop a cartographic technique called *"Pixel Mixture"* (PM) to visualise membership maps by including all membership values in the representation. This technique randomly assigns pixels to a sub-pixel grid, with a probability proportional to the membership grade in the class, thereby giving a visual impression of both the possible classes and their confusion (de Gruijter *et al.*, 1997). Similar concept of setting up a sub-pixel grid to visualise the transition zones between land cover classes was given by Atkinson (1997). de Bruin & Stein (1998) visualised multiple membership grades together with the uncertainty of classification, where the class map was displayed together with the confusion index (in further text CI), by combining hatching to visualise the classes and grey scale intensity to visualise the CI. Although in principle the latter method would work with any number of classes, more than about five classes leads to visual ambiguity between the grey-scale intensity and density of the hatching pattern.

Till now, a per-pixel colour-mixing approach was not pursued, assuming that digital technology was limited to three colours. Colour mixing is in fact possible,

using for example the standard red green and blue bands (RGB), but only if the number of classes is three. There is still a need for GIS methods to visually explore results of fuzzy classification and at the same time quantify abruptness of transitions (Goodchild *et al.*, 1994; Burrough & McDonnell, 1998, p. 284).

In this paper we extend the work of de Gruijter *et al.* (1997) and Burrough *et al.* (1997) by developing a new static visualisation technique based on calculations with colours. We propose a technique to simultaneously explore both the spatial and taxonomic confusion of our mapping system. For this purpose, we developed a so-called *"fuzzy-metric"* circular colour legend, suitable for quantified categorical data, and a method to analyse the abruptness of the derived boundaries.

6.2 Methods

6.2.1 Supervised fuzzy k-means classification

Membership maps can be derived by different algorithms. Most commonly, the membership maps are derived using a continuous classification algorithm such as fuzzy k-means (de Gruijter & McBratney, 1988). From the general theory of numerical taxonomy, a membership is calculated from the standardised distance in the attribute space (Sokal & Sneath, 1976):

$$\mu_c(i) = \frac{\left[d_c^2(i)\right]^{-\frac{1}{(q-1)}}}{\sum_{c=1}^{k}\left[d_c^2(i)\right]^{-\frac{1}{(q-1)}}} \qquad c = 1, 2, ...k \qquad i = 1, 2, ...n \tag{6.1}$$

$$\mu_c(i) \in [0, 1] \tag{6.2}$$

where $\mu_c(i)$ is a fuzzy membership value of the ith object in the cth cluster, d is the similarity distance, k is the number of clusters and q is the fuzzy exponent determining the amount of fuzziness. A commonly-suggested value for q, also used in this work, is 1.5 (Burrough *et al.*, 1997). The simplest similarity distance is the Euclidian distance, defined as the sum of squared differences from a given ith object to the class centres in attribute space (Gordon, 1981):

$$d_c(i) = \sqrt{\sum_{j=1}^{l}\left[x_j(i) - x_{c,j}\right]^2} \qquad c = 1, 2, ...k \qquad i = 1, 2, ...n \tag{6.3}$$

where $x_{c,j}$'s are the class centres of the jth attribute variable and k is the total number of classes. The fuzzy k-means classification was used to develop an unsupervised

classification algorithm where the memberships are calculated based on an objective function for the entire set of k classes (de Gruijter & McBratney, 1988). In this case, the only input is the number of classes and fuzzy exponent, while the algorithm finds the optimum class centres iteratively. The alternative to this approach is to predefine the class centres, so that the optimisation function is not necessary. This is a supervised fuzzy k-means classification or allocation, which is attractive for those cases where the surveyor has prior knowledge of the central concepts of the several classes (Hartigan, 1975).

In the case of supervised classification, the cluster centres for each attribute of each class are established prior to the classification by sampling. The similarity distances are first standardised using the sampled variance for each class and then used to derive the multiple memberships:

$$d_c(i) = \sqrt{\sum_{j=1}^{l} \left(\frac{x_j(i) - x_{c,j}}{s_{x_{c,j}}} \right)^2} \tag{6.4}$$

where $s_{x_{c,j}}$ is the sampled variance around the $x_{c,j}$'s and $d_c(i)$ is the *diagonal distance* (Hartigan, 1975).

6.2.2 The Colour mixture (CM)

Instead of mingling fixed-colour sub-pixels, as in case of PM, an impression of confusion can be achieved by mixing colours in each pixel (i) as an averaged intensity of RGB bands. The new derived R_i, G_i, B_i raster maps are first calculated separately for each of the three primary colours and then combined as a colour composite in an image processing software:

$$R_i = \frac{\sum_{c=1}^{k} (\mu_{i,c} \cdot R_c)}{\sum_{i=1}^{n} \mu_{i,c}} \tag{6.5}$$

$$G_i = \frac{\sum_{c=1}^{k} (\mu_{i,c} \cdot G_c)}{\sum_{i=1}^{n} \mu_{i,c}} \tag{6.6}$$

$$B_i = \frac{\sum_{c=1}^{k} (\mu_{i,c} \cdot B_c)}{\sum_{i=1}^{n} \mu_{i,c}} \tag{6.7}$$

where the R_i, G_i, B_i are the new derived mixed colours, R_c, G_c, B_c are the digital values (0-255) for selected class colours and k is the number of classes. We refer to this algorithm as the *Colour Mixture* (CM). For standardised fuzzy sets, the sum of memberships equals unity and the denominator in Eqs. (6.5), (6.6) and (6.7) can be discarded. A simplified example of how CM works, can be seen in Fig. 6.1.

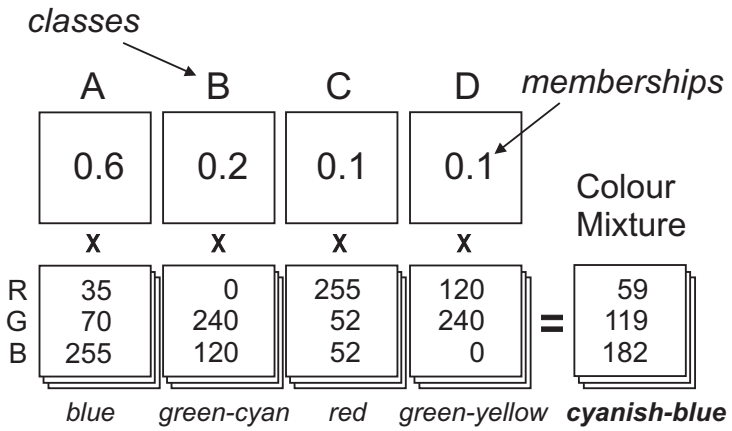

Figure 6.1: Schematic example of Colour mixture technique: the new RGB's are calculated as the weighted averages of class representations.

In the example above (Fig. 6.1), a new colour (cyanish-blue) is derived as a weighted average of four colours and reflects the highest memberships: blue and green-cyan colour. However, the CM technique is more complex than it appears on the first sight: how to interpret the new colours and how to create a legend showing all possible colours for all combinations of memberships? Moreover, if the legend class colours for each individual class are selected freely, a new-derived colour will not necessarily appear in the originally selected legend (accidental colour).

A more serious problem can be illustrated by the case of three classes (say `A`, `B`, `C`), where the representation of class `B` lies midway along the line connecting `A` and `C` in the RGB colour cube. A pixel that appears the same as the colour of class `B` in the legend could indeed indicate class `B`. However, it could also result from the maximum confusion between the `A` and `C` classes (i.e. $\mu_{\texttt{A}} = \mu_{\texttt{C}} = 0.5$). The possibility of such situations increases with the addition of more class colours, i.e. a higher density of points in the colour cube. To minimise such confusion, classes that are close together taxonomically should also be close in the colour cube.

To account for the above-listed problems, we decided to construct the legend following three principles:

- All legend colours should lie in the same plane in the RGB colour cube, in order to be able to produce a two-dimensional representation.
- It should not be possible to derive an exact colour in the legend from any combinations of colours representing other classes.
- A mixture in colour space should be the visual equivalent to a mixture in taxonomic space.

Following these three principles we designed a special legend for the purpose of CM. We named it "fuzzy-metric colour legend".

6.2.3 Fuzzy-metric colour legend

To keep all colours on one plane, we selected the cross-section (plane) of the RGB cube perpendicular to the diagonal, which connects black (0, 0, 0) and white (255, 255, 255). This is the *Hue, Saturation, Intensity* (HSI) colour model (Brown & Feringa, 1999). There are an infinite number of such planes along the diagonal, each corresponding to one intensity, also called the brightness. At any brightness we can produce a wheel-shaped palette, called a *"HSI colourwheel"*, which has the same brightness value and different hue and saturation values (Niblack, 1986, p. 61). Here, the hue (H) represents the visual sensation of the colour type, and is calculated as the number of degrees around the axis. The saturation (S) represents the degree to which the colour expresses its hue, and is calculated as the radial distance from the diagonal axis. The intensity (I) represents the visual sensation of brightness. The HSI colour model has already been recognised as promising for the visualisation of uncertainty by Jiang *et al.* (1995); Jiang (1996).

The formulas for RGB to HSI transformations often differ between software. In ILWIS, transformation is made using:

$$H = \frac{360}{2\pi} \cdot \arctan\left(\frac{\sqrt{3}}{2} \cdot [G - B], R - \frac{[G + B]}{2}\right) \cdot \frac{240}{360} \tag{6.8}$$

$$S = \sqrt{R^2 + G^2 + B^2 - R \cdot G - R \cdot B - G \cdot B} \cdot \frac{240}{255} \tag{6.9}$$

$$I = \frac{R + G + B}{3} \cdot \frac{240}{255} \tag{6.10}$$

where the input R, G and B are coded from 0–255 while H, S and I are coded to from 0–240 to comply with the colour scheme used in Microsoft Windows. The I is constant also for all new derived colours since:

$$I_i = \frac{R_i + G_i + B_i}{3} = \frac{\sum_{c=1}^{k} (\mu_{i,c} \cdot R_c) + \sum_{c=1}^{k} (\mu_{i,c} \cdot G_c) + \sum_{c=1}^{k} (\mu_{i,c} \cdot B_c)}{3 \cdot \sum_{i=1}^{n} \mu_{i,c}}$$
$$= \frac{\sum_{c=1}^{k} \mu_{i,c} \cdot (R_c + G_c + B_c)}{3 \cdot \sum_{i=1}^{n} \mu_{i,c}} = \frac{R_c + G_c + B_c}{3} \tag{6.11}$$

This way none of the mixed colours will appear too dominant visually. The class colour representations were located on the perimeter of the HSI colourwheel at a given brightness (here, arbitrarily set to 120 or half of the full brightness). In this way the class or taxonomic value is coded only with the hue, while the saturation is maximised (240) and brightness is kept constant.

We imagined that the relation between class centres in attribute space can be used to place them closer or further apart on the circumference of the HSI wheel. The class centres can be analysed in multivariate attribute space also referred to as taxonomic space. However, the centres rarely fall along a two-dimensional subspace (plane) within this space. Yet, we only have two dimensions that can be represented. Therefore, the dimensionality of the taxonomic space can be reduced to factor loadings by factor analysis (Tucker & MacCallum, 1997). The first (F_1) and the second (F_2) factor loading can then be used as the axes of a reduced attribute space[2]. The angular distances of the class centres can be then used to derive representation hues for each class. Fig. 6.2 shows an example of colour derivation for four classes, in which the angular distance between any two classes is proportional to the distance in taxonomic space spanned by the factor scores F_1 and F_2. Note that we lose information by using only first and second factor loadings, which is unavoidable, since we need to produce a two-dimensional legend. The Hue of each class is derived as the angular distance around the gravity centre, which, in the case of factor loadings, is the centre of the coordinate system:

$$H_c = \frac{360}{2\pi} \cdot \arctan(F_{1c}, F_{2c}) \cdot \frac{240}{360} \tag{6.12}$$

where H_c is the new derived class Hue and F_{1c} and F_{c2} are the factor loadings in the cth class. To display the class colours in the GIS, the H_c, S_c, I_c values (where S_c=240 and I_c=120) need to be transformed to RGB. In ILWIS, this is done by using the inverse HSI to RGB transformation. Towards the centre of the circle the colours

[2]This is the the so-called biplot display (Gabriel, 1971).

approach grey, that is, the hues become less distinct. This visualises the situation where we are less sure about the components of any mixture.

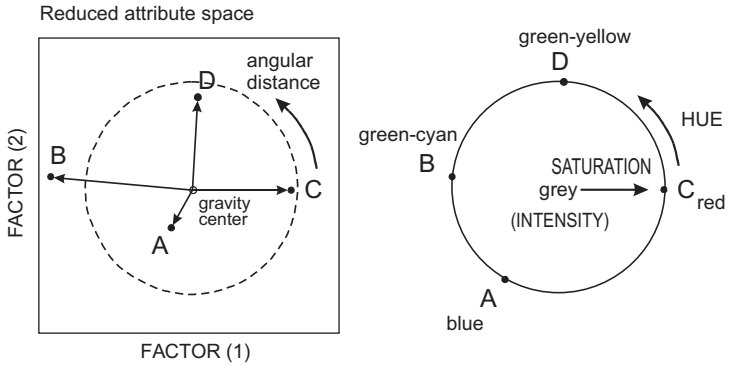

Figure 6.2: Schematic example of the fuzzy-metric colour legend construction for the CM technique: biplot display of factor scores for class centres in the reduced attribute space and angular coordinates (left), same transferred to the circumference of HSI colourwheel and the derived colour classes (right).

However, first maps produced showed that the coding of uncertainty of classification with saturation does not offer enough visual impression. This corresponds to results of perception tests conducted by cartographers, which clearly showed that brightness is the best variable to visualise uncertainty (Jiang, 1996, pp. 118-120). Therefore, we finally decided to visualise confusion using the whiteness, i.e. amount of white colour by replacing the constant brightness with the derived saturation:

$$[R_{i*}, G_{i*}, B_{i*}] = colorhsi\,[H_i, S_i, (120 + 0.5 \cdot S_i)] \tag{6.13}$$

where H_i, S_i are the hue and saturation maps derived from the CM, *colorhsi* is the ILWIS command to derive a RGB map from HSI maps and R_{i*}, G_{i*} and B_{i*} are the corrected RGB. Now, both saturation and brightness changes radially, i.e. brightness changes from 120 (circumference) to maximum brightness (255 or the centre of colourwheel).

From equation (6.12) we can back-transform the RGB map to estimate the input memberships. We first have to derive angular coordinates of each colour in the HSI colourwheel:

$$f_{1c} = S \cdot \cos(\varphi) \tag{6.14}$$
$$f_{2c} = S \cdot \sin(\varphi) \tag{6.15}$$
$$\varphi = H \cdot \frac{360}{240} \tag{6.16}$$

where the $f_{1,2c}$ are the angular coordinates of ith pixel, φ is the Hue angle in degrees and S_i is the derived saturation map. A membership μ^* of class c can be back-derived by calculating the distances on the colourwheel:

$$d_c^*(i) = \sqrt{[f_{1c}(i) - f_{1c}]^2 + [f_{2c}(i) - f_{2c}]^2} \qquad i = 1, 2, ...n \tag{6.17}$$

where $d_c^*(i)$ is the distance in the two-dimensional attribute space. This can then be used to derive μ^* as in the Eq. (6.1).

6.2.4 Confusion index based on the CM saturation

The fuzziness of the derived multiple memberships are analysed by calculating the confusion index (CI), which is commonly defined as the difference, or sometimes the ratio, between the first and second highest membership class per pixel:

$$CI_1 = 1 - (\mu_{\max} - \mu_{2nd\,\max}) \tag{6.18}$$
$$CI_2 = \frac{\mu_{2nd\,\max}}{\mu_{\max}} \tag{6.19}$$

where $\mu_{\max}$ is the highest membership and the $\mu_{2nd\,\max}$ is the second highest membership (Burrough *et al.*, 1997). The CI map calculated from one of these equations will show higher values in the areas of transition between different classes in attribute space. The $CI_{1,2}$ are the highest at the transition zones, and thus can be used for their automatic delineatation. However, the $CI_{1,2}$ do not consider how similar are the $\mu_{\max}$ and $\mu_{2nd\,\max}$ classes taxonomically. It can show high confusion although the classes might be taxonomically very similar and vice versa. To asses which classes are more similar, we can calculate the distances between the class centres using formula in equation (6.4) and pooled standard deviation:

$$d_{c,c'} = \sqrt{\sum_{j=1}^{l} \left(\frac{x_{c,j} - x_{c',j}}{\sqrt{s_{x_{c,j}}^2 + s_{x_{c',j}}^2}} \right)^2} \tag{6.20}$$

where $d_{c,c'}$ is the distance between the cth and c'th class centre. From the distances between the class centres, a membership can be derived as:

$$\mu_{c,c'} = \frac{\left(d^2_{c,c'}\right)^{-\frac{1}{(q-1)}}}{\sum_{c=1}^{k}\left(d^2_{c,c'}\right)^{-\frac{1}{(q-1)}}} \tag{6.21}$$

It appears that the between-class memberships should be used to adjust the CI for similarities between the classes. However, the calculation of adjusted CI can be complex as there are many combinations of class centres. On the other hand, the design of fuzzy-metric colourwheel can be regarded as a way of adjusting the confusion. The classes that are closer on the colourwheel, will give a lower distance when used to derive a new colour, which give us similar impression of the confusion. In ILWIS, the colour separation operation can be used to derive the saturation S_i of each pixel from its (R_i, G_i, B_i) values. The saturation can now be expressed as the relative saturation or radial distance on the colourwheel ($CI_{\mathbf{CM}}$) by transforming S to the 0-1 range:

$$CI_{\mathbf{CM}} = 1 - \frac{S}{210} \tag{6.22}$$

In this case, the maximum back-transformed saturation value produced by ILWIS was 210, not 240 as would be expected from the forward transformation. However, due to the truncation of the values used in the HSI to RGB transformation, which is non-linear, we decided to use the lower value for saturation as the more realistic.

6.2.5 Deriving primary boundaries

The geographic transition zones, i.e. pixels were first extracted by calculating a crisp map of most possible classes (defuzzification). This map was then converted to a vector format using raster to vector transformation in GIS, i.e. to a segment map showing the transition zones. The $CI_{\mathbf{CM}}$ can be used to quantify different boundaries. In this case, the boundary pixels with high $CI_{\mathbf{CM}}$ (> 0.7) were classified as first-level or primary boundaries, otherwise as second-level, secondary boundaries or soft lines, which correspond to the boundary types typically used in the photo-interpretation (Buringh, 1960). Areas of high $CI_{\mathbf{CM}}$ outside the transition zones were considered to represent possible intra/extra-grade classes, as they show higher distance from more than two classes at the same time.

6.2.6 Case study

The methodology was tested using the Baranja Hill data set, a 3.5×3.5 km square. This is the central photo from the aerial photo-triplet that had been interpreted for landform classes using a conventional aerial photointerpretation technique (API) — the geo-pedological approach to soil mapping (Zinck & Valenzuela, 1990). Two main landscapes were recognised: a dissected elongated hill with numerous vales and a plain with two terrace levels. To classify the landform classes, we used six attribute maps: relative elevation to groundwater (GWD), slope gradient (SLOPE), wetness index (CTI), profile and tangent curvature (PROFC and TANGC) and viewshed (VSHED) at grid resolution of 25 m. We selected about 30 pixels in each mapping unit to estimate the central values in the attribute space (Table 6.1). This corresponds to the selection of the training set in the supervised classification of remote sensing images.

The memberships of the fuzzy k-means were calculated in ILWIS from the set of six maps. Factor analysis of the attribute space was carried out using the maximum likelihood estimation model of S-PLUS (MathSoft Inc., 1999). The derived memberships are shown in Fig. 6.3.

6.3 Results

6.3.1 Attribute space and selection of colours

The derived H_c based on the class centres from Table 6.1 and factor analysis is shown in Table 6.2. The first factor (F_{1c}) accounted for 37%, and the second (F_{2c}) for 32%, of the total variance. The biplot display in the Fig. 6.4 shows the results of the factor analysis graphically, where arrows represent the proportion of the original variance explained by the two principal components. The direction of the arrows indicates the relative loadings on the first and second principal components. The most correlated landform parameters were SLOPE and CTI, TANGC and PROFC, and GWD and CTI. The closest classes on the circumference of the colourwheel were Pl311 and Pl411, i.e. the low and high terraces of the plain. These classes are also very close in taxonomic space, both being towards the centre of the two-dimensional attribute space formed by the factors. By contrast, the summit and shoulder of the hill (Hi111 and Hi112) are close on the colourwheel (similar angle), but not in the taxonomic space, i.e. Hi111 was much further from the centre. This problem is further discussed in the Conclusions and discussion.

The allocation of classes on the colourwheel can be compared to the similarity distances calculated between the class centres (Table 6.3) and their derived between-class memberships (Table 6.4). The most similar class pairs are: Hi112 and Hi311;

Table 6.1: Landform classes, attribute class centres and sampled variation (σ) around the central values.

Landform classes		SLOPE (σ)	PROFC (σ)	TANGC (σ)	CTI (σ)	VSHED (σ)	GWD (σ)
Code	Description	%	100 m^{-1}	100 m^{-1}	-	-	m
Hi111	Hill (summit)	6.3	-0.54	0.45	5.9	0.89	125.4
		(3.29)	(0.41)	(0.37)	(0.55)	(1.08)	(18.53)
Hi112	Hill (shoulder)	14.2	-0.12	0.28	5.5	1.50	80.1
		(5.13)	(0.44)	(0.40)	(0.47)	(3.39)	(24.26)
Hi211	Escarpment (scarp)	29.4	-0.17	0.24	4.7	-3.71	56.3
		(3.83)	(0.28)	(0.26)	(0.13)	(2.16)	(20.46)
Hi212	Escarpment (colluvium)	25.4	0.38	0.11	4.9	-3.46	31.0
		(2.79)	(0.55)	(0.17)	(0.17)	(2.58)	(11.10)
Hi311	Vale (slope)	23.8	0.24	-0.18	5.2	-0.50	85.3
		(7.94)	(0.58)	(0.75)	(0.76)	(3.46)	(14.50)
Hi312	Vale (bottom)	4.3	0.70	-0.25	7.9	0.00	40.7
		(2.58)	(0.55)	(0.24)	(1.13)	(0.72)	(21.75)
Hi411	Glacis (slope)	7.1	-0.04	0.23	6.1	2.31	60.2
		(1.57)	(0.15)	(0.08)	(0.23)	(0.35)	(2.75)
Pl311	High terrace	0.8	-0.03	-0.04	8.3	0.54	7.8
		(0.29)	(0.04)	(0.05)	(0.30)	(0.21)	(1.52)
Pl411	Low terrace	2.6	0.13	-0.03	8.0	0.04	1.8
		(2.70)	(0.15)	(0.12)	(1.22)	(0.89)	(3.84)

Hi112 and Hi111; and Pl311 and Pl411. The sum of distances gives an idea on the closeness of class centres in the multivariate attribute space. In this case, the most distant class centres are Pl311 and Hi411, while Hi112 and Hi311 proved to be the closest having the lowest total sum of distances. These relations correspond to the one seen in the biplot display.

6.3.2 Comparisons — defuzzification, PM and CM

Fig. 6.5 shows a comparison between the mixed-colour map and alternative techniques: defuzzification and PM. In the case of defuzzified map and PM, the colours

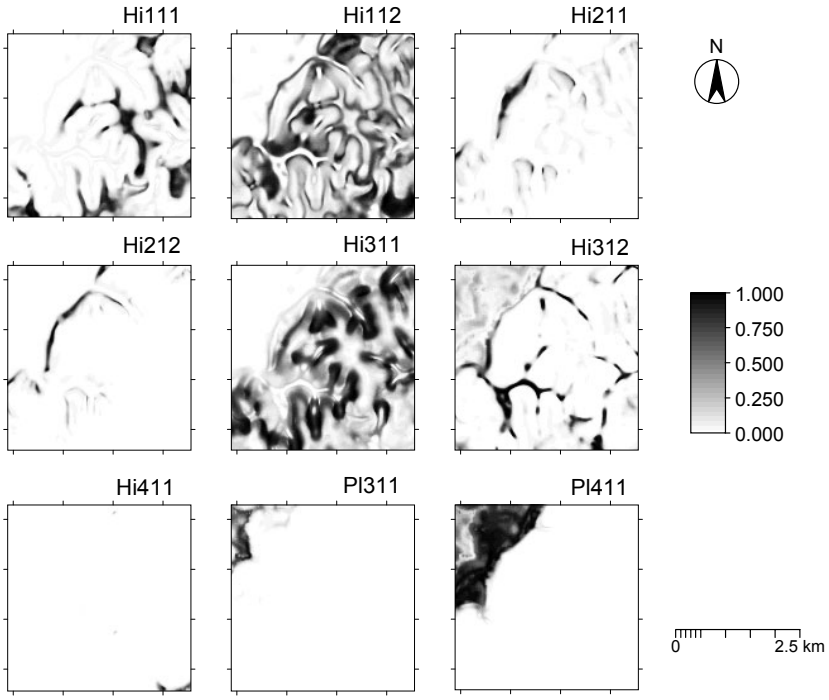

Figure 6.3: Derived membership maps — landform facets.

were selected subjectively to represent a psychological impression of the landform facets: bluer colours were used to represent wetter, lower soils such as the terraces and vales, green to represent steeper slopes, yellow for hilltops, whereas, in the case of the mixed-colour map, colours reflect distances between classes in the attribute space. The defuzzified map in Fig. 6.5a, does not provide any information about confusion or original memberships. The PM allows the viewer to infer about the overall confusion of each class. For example, the vale bottom (Hi311) seems to be the purest delineation (Fig. 6.5b). The mixed-colour map (Fig. 6.5c) shows intermediate colours between similar classes (i.e. along the circumference), which represent intergrades in taxonomic space and therefore are assigned intermediate colour. On the other hand, intergrades between strongly-contrasting classes finished having colours towards the centre of the circle, i.e. being whitish. For example, intermediate pixels between escarpment (Hi211) and the two terrace levels (Pl311 and Pl411) show al-

Table 6.2: Deriving the class Hues (H_c): factor scores (F_{1c} and F_{2c}), Hue angle (φ) and the angular coordinates (f_{1c} and f_{2c}).

Classes	F_{1c}	F_{2c}	φ	H_c	S_c	I_c	f_{1c}	f_{2c}
Hi111	0.88	1.53	60.1	40	240	120	207.8	120.0
Hi112	0.06	0.69	85.0	57	240	120	239.3	18.8
Hi211	-2.03	0.86	157.0	105	240	120	91.8	-221.7
Hi212	-1.23	-0.07	-176.7	122	240	120	-12.6	-239.7
Hi311	-0.62	-1.78	-109.2	167	240	120	-226.2	-80.1
Hi312	0.75	-1.44	-62.5	198	240	120	-213.8	109.0
Hi411	0.92	0.41	24.0	16	240	120	97.6	219.3
Pl311	0.69	-0.08	-6.6	236	240	120	-25.1	238.7
Pl411	0.59	-0.14	-13.3	231	240	120	-56.0	233.4

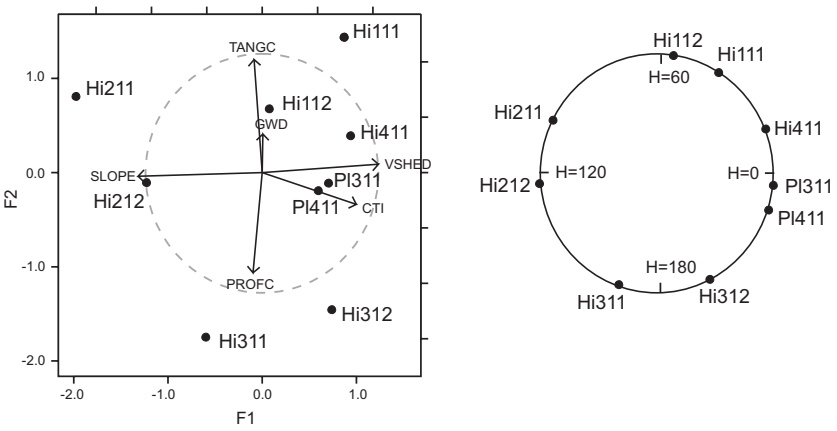

Figure 6.4: The biplot display with factor scorings (left) and the derived Hcs on the fuzzy-metric HSI colourwheel legend (right). The landform attributes are presented with the eigen vectors.

most white colour. This is psychologically appealing: pure white colour represents a mixture of strong contrasts or an undefined class, where a mixture of similar classes results in a transitional colour.

The inverse process of back-deriving the memberships from the mixed-colour map showed that the estimation of the memberships, based on the angular coordi-

Table 6.3: Diagonal distances $d_{c,c'}$ between the class centres in multivariate feature space. Smaller distances indicate more similar classes, while the total sum gives idea how distant is a class from the overall feature space.

	Hi111	Hi112	Hi211	Hi212	Hi311	Hi312	Hi411	Pl311	Pl411
Hi111	0								
Hi112	2.2	0							
Hi211	6.0	3.2	0						
Hi212	6.8	3.2	2.0	0					
Hi311	3.1	1.4	1.8	3.1	0				
Hi312	4.2	3.3	6.7	6.4	3.6	0			
Hi411	3.9	2.0	8.0	7.8	3.1	4.2	0		
Pl311	7.8	6.5	13.7	13.4	7.2	2.7	18.8	0	
Pl411	7.1	4.4	7.2	7.0	6.4	2.2	12.9	2.0	0
Total	41.1	26.2	48.6	49.8	29.7	33.2	60.8	72.0	49.2

Table 6.4: Membership values between-class centres based on the similarity distances.

	Hi111	Hi112	Hi211	Hi212	Hi311	Hi312	Hi411	Pl311	Pl411
Hi111	1.000								
Hi112	0.927	1.000							
Hi211	0.000	0.001	1.000						
Hi212	0.000	0.001	0.317	1.000					
Hi311	0.059	0.928	0.675	0.032	1.000				
Hi312	0.005	0.001	0.000	0.000	0.000	1.000			
Hi411	0.008	0.046	0.000	0.000	0.001	0.005	1.000		
Pl311	0.000	0.000	0.000	0.000	0.000	0.171	0.000	1.000	
Pl411	0.000	0.000	0.000	0.000	0.000	0.770	0.000	0.925	1.000

nates of the reduced feature space[3] (f_1, f_2), can make inaccurate estimates of the

[3]See Eq. 6.17 on page 124.

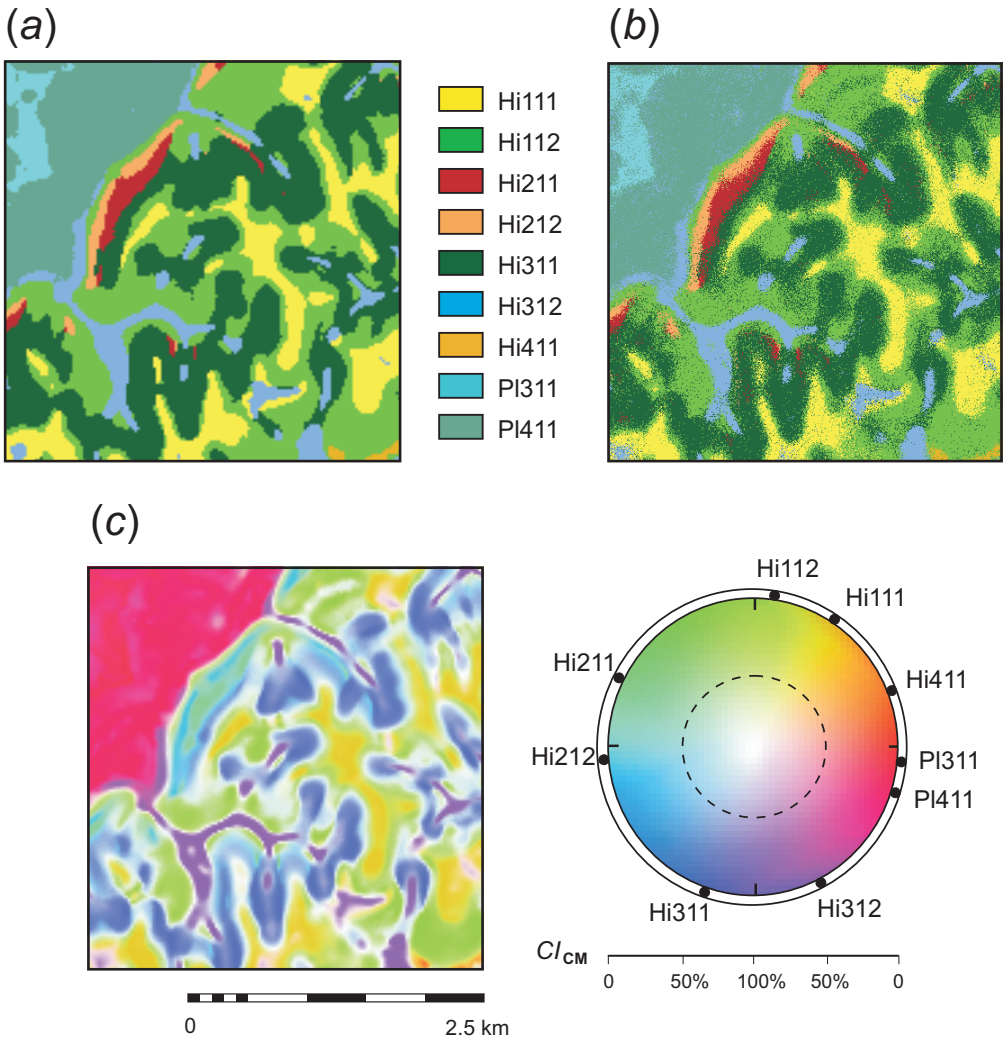

Figure 6.5: Comparison of different cartographic techniques: (a) defuzzification; (b) pixel mixture and (c) colour mixture with the circular fuzzy-metric legend.

original membership in areas of high confusion. This is not the case with the PM technique, where the original memberships can be fairly well estimated by counting the proportion of pixels. In the case of CM, once we projected the class centres onto the circumference of colourwheel, we have irrecoverably lost a part of the information. We compared original and back-derived memberships for all classes and got

strong correspondence with correlation coefficients between the classes ranging from R^2=0.43 (Pl311) to R^2=0.94 (Hi312, Pl411). However, the back-derivation showed systematic discrepancy, as the areas of higher original memberships will usually be overestimated, while the areas of high confusion can give rather different estimates of the original values.

6.3.3 Confusion and boundary index

Differences between $CI_{\mathbf{CM}}$ and the commonly used confusion indices $CI_{1,2}$ can be seen in Fig. 6.6. The CI_1 map shows all transition zones, while the $CI_{\mathbf{CM}}$ map shows only the transition zones between classes that have sufficient angular separation on the colourwheel. In this case some classes were basically merged, and no confusion is shown. For example, the transition zone between the hill summit and shoulder (Hi111 and Hi112) disappeared in the mixed-colour and the $CI_{\mathbf{CM}}$ map. Also the transitions between the high and low terraces and the sloping part and colluvium of the escarpment were ignored. On the other hand, the transitions between the Hi111 and Hi311 and Hi312 were shown as being more abrupt.

The histogram of the $CI_{\mathbf{CM}}$ map showed a small grouping around the higher $CI_{\mathbf{CM}}$ and a clear breakpoint at $CI_{\mathbf{CM}}$=0.7. We selected this as the threshold to separate the primary and secondary boundaries. The derived primary boundaries (transitions between the more contrasting classes) can be seen in Fig. 6.7. The CM technique shows the clear advantage of diminishing the boundaries of fairly similar classes, which can not be achieved by using the common confusion indices (CI_{12}).

The remaining areas of high $CI_{\mathbf{CM}}$ (>0.7), i.e. which were not areas of transition were considered to be inter/extra-grades, again adjusted for the taxonomic similarities. Comparing these patches with the input maps, we found that they are mainly correlated with profile curvature and slope (PROFC and SLOPE). Indeed, both profile and tangent curvature (PROFC and TANGC) did not show correspondence with the delineated geomorphic units in a visual overlay. These are thus areas inside some API units, which differ in from the surrounding unit in their convexity, and should perhaps be considered as different geomorphic classes in a revised API legend.

6.4 Conclusions and discussion

Comparison of the CM and the alternative approaches to visualisation of multiple memberships showed that the there are two major applications that should be emphasized in this work: (1) CM visualises uncertainty of our classification by giving an insight into both taxonomic and geographical confusion and (2) it serves as a method to generalise similar classes. In the case of visualisation of uncertainty we

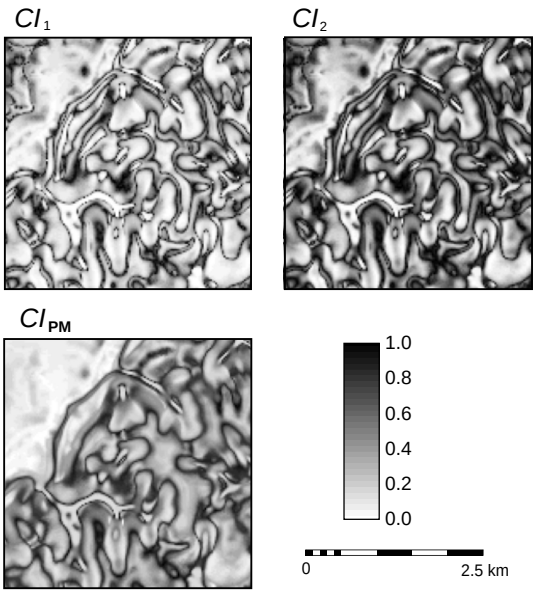

Figure 6.6: Confusion indices calculated using memberships (CI_1 and CI_2) and based on the saturation of mixed-colours map ($CI_{\mathbf{CM}}$). $CI_{\mathbf{CM}}$ in general shows lower values than CI_2 but similar pattern.

used whiteness, which showed better impression of uncertainty than the saturation. This was also confirmed after we printed the mixed-colour map. The confusion index derived from the mixed-colour map allows extraction of primary transition zones between fairly contrasting classes. Moreover, the $CI_{\mathbf{CM}}$ can be used to locate areas of high taxonomic confusion (inter/extra-grades) and therefore be used to redesign the map legend or make additional sampling on the field. The commonly used $CI_{1,2}$ do not account for confusion in taxonomic space. In the case of CM, the selection of colours to represent a class was not based on the mapper's connotation (e.g. yellow for sand, purple for peat), but were assigned by the algorithm and limited to seven colour types. This is a relatively small number considering that there are 627 colour names in the qualitative Colour Notation System (Berk *et al.*, 1982), which will be discouraging for an experienced cartographer. However, taking the rule of thumb that the optimal number of classes that can be perceived and memorised by user is seven (Kraak & Ormeling, 1996), the double continuous approach ensures limited number of colours regardless of the total number of classes.

Although the input number of classes is high, the system will always limit the

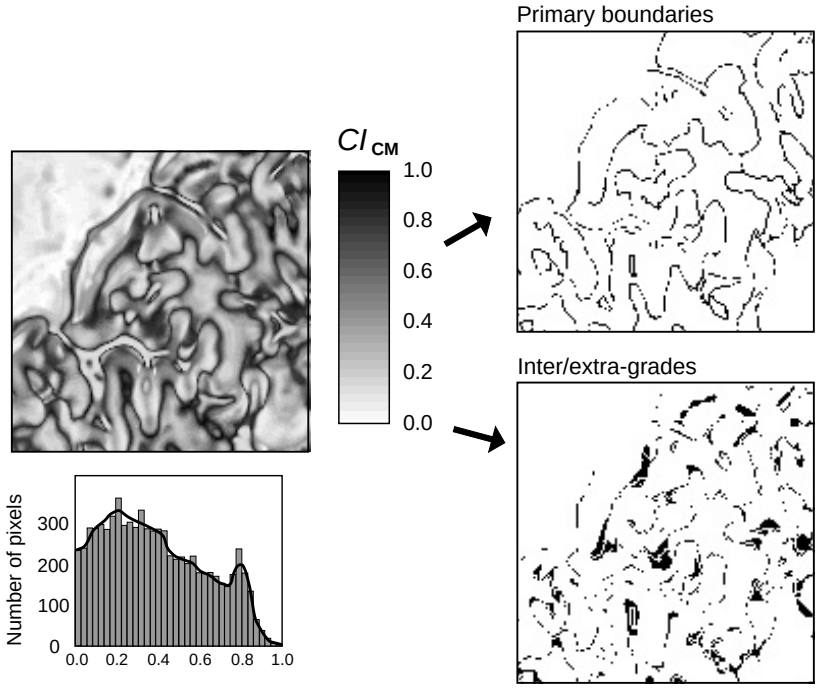

Figure 6.7: Extraction of primary boundaries pixels and possible inter/extra-grade areas derived from relative distance in the colourwheel . Notice the grouping of higher $CI_{\mathbf{CM}}$ values in the upper part of the histogram.

number of colours to seven generic hues: red, yellowish, green, cyan, blue, magenta and white. In that sense, the mixed-colour map can be compared with the pseudo-colour scale used in many GIS packages for visualising continuous variables. The derived saturation of the mixed-colour map calculated using the CM technique visualises both objects and their taxonomic location as fuzzy sets, whereas the defuzzification or PM present classes as crisp definitions. Therefore we address it as a double continuous or double fuzzy approach and the resulting map as a mixed-colour map with a fuzzy-metric legend (Fig. 6.8). In this case the legend indicates not only the category, but also 'amount' of a category, e.g. *"terraceness"* or *"summitness"* in different parts of the map. Although one may argue that the concept of relatively homogeneous delineations, occupying relatively compact geographical areas is not so far from reality, the advantage of the continuous categorical maps is that they are universal. If there is indeed no confusion between the class clusters, a map becomes double-crisp.

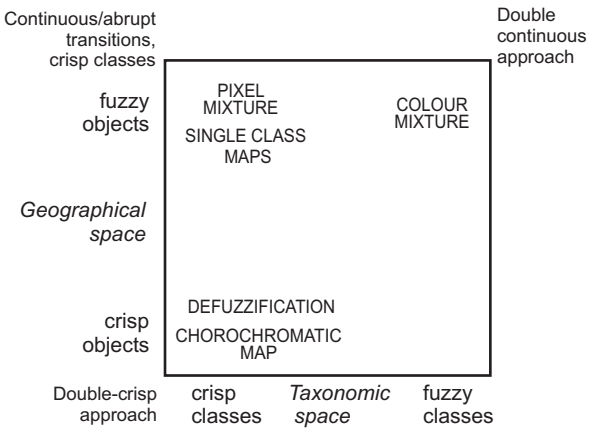

Figure 6.8: Different cartographic techniques for visualisation of multiple memberships: comparison in the relation how are the geographic transition and taxonomic definition modelled.

The limitations of the described CM method are as follows. First, in the case of CM, the colours can still be derived from different combinations of memberships, i.e. the class centres are considered to be as relative as the definition of classes. If the classes from which a new colour is derived are fairly distant on the colourwheel, the CM will produce an ambiguous (whitish) colour. The original memberships can not be accurately back-derived as we do not know from which combination of membership was it derived. The system quantifies the taxonomic space, which means that even if a membership of some class is low, if two closely neighbouring classes show equally high memberships, the derived colour will correspond to an intermediate class colour.

The second limitation of CM is that we were forced to reduce the attribute space to only the first and second components. Thus we have discarded information in this reduced attribute space, as only 67% of total variation was explained with first two components. Also note that for factor analysis, number of classes needs to be larger than the number of attributes. An alternative would be to develop a method to use distances between the class centres (from Table 6.3) to allocate them on the circumference of the colourwheel. In both cases we will have to approximate or average input values, as the number of classes is much lower than the number of possible combinations. In this case a 9×9 matrix gives 36 combinations in total or $\frac{k^2-k}{2}$. Second, in the case that the class centres are maximally spread over the feature space, the biplot display method is fast and gives a good picture about the

relation between them (Fig. 6.9a). A problem will appear if a class is very close to the centre of gravity (Fig. 6.9b). Two classes could have similar angles but one could be much closer to the centre than the other (class C in this case). In fact, this class can easily finish being any colour, which is an unwanted effect. It will appear closer to some class although it's almost equally distant from all classes. We can also imagine a case where a single class, fairly distant from other classes, will distract the centre of gravity (Fig. 6.9c). In this case it is clear that there are more taxonomic clusters, but the produced map will show only few colours.

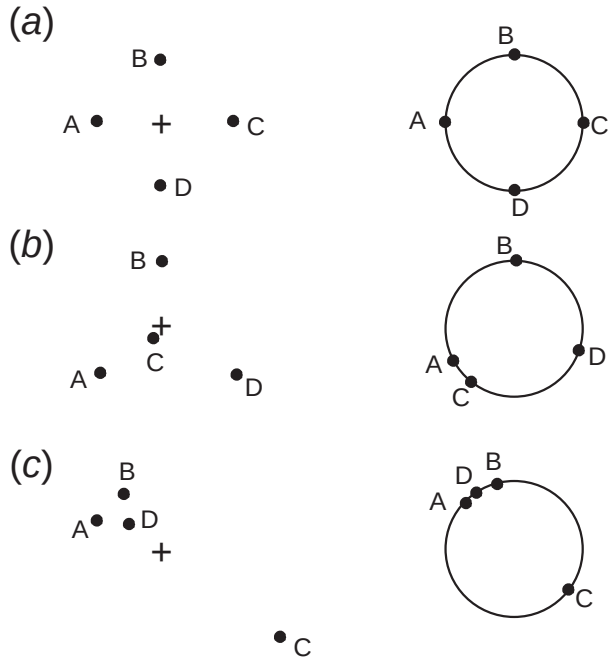

Figure 6.9: Some special cases that may appear in the design of the colourwheel using the biplot display method.

Finally, a limitation of the proposed CM algorithm is that the definition of the HSI colourwheel is not completely perceptually scaled (equal distances measured throughout colour spaces represent equally perceived colour differences). In this case, we were forced to use a geometrical colour space to enable linear calculations with colours. It is possible that some other colour models could provide better solutions for colour mixing and design of the legend.

"There is no longer any need to cling the double-crisp model... we can do much better now" stated Burrough *et al.* (1997, p. 133) half a decade ago. At that time,

the algorithms for fuzzy classifications were applied in experimental studies only, while today they can be can found in many statistical and GIS packages, some free e.g. FuzME (Minasny & McBratney, 1999). In this paper we presented a method to visualise and analyse the multiple memberships by doing GIS calculations on colours. Its development has forced us to think about taxonomic space and how to construct a metric legend for categorical variables. We finally selected the circular form of the legend — categorical classes are not sorted one above each other as they are in the traditional concepts. A circle has no beginning and therefore indicates that the classes are not ordinal. On the other hand, some classes can be located closer and some farther apart.

Operational tools for production and use of continuous categorical maps in large projects are yet to be developed. The example shown here is based on nine classes in a small area. Real county-level soil surveys typically have 20 to 40 map units, which requires a bigger computational effort. The next step will be to test this methodology on data sets developed as a part of large inventory projects. Some parts of the CM principles could be also used in a more interactive way to offer better insight into the classification uncertainty of discrete data. For example, by linking resulting mixed-colour map to the sample set points, so that an mapper could inspect class definition interactively. This could, for example, help optimise definition and number of class centres.

Chapter 7

Grid-based Soil Information System*

"It is now quite possible to combine information derived from DEMs and satellite observation with profile data and numerical models of soil processes to produce a rich, predictive models of the soil to meet both the purposes of research in soil formation and landscape development and practical considerations of land suitability assessment, decision making or the review of development scenarios."

[P.A. Burrough, announcing future research in "Continuous classification in soil survey: spatial correlation, confusion and boundaries", Geoderma, vol. 77(2-4): 115-135]

*based on: Hengl, T., 2004? A hybrid grid-based soil information system based on the mixed model of spatial variation. Geoderma, in review.

7.1 Introduction

A Soil Information System (SIS), also referred to as a Soil Geographical Database (SGDB), is a commonly used term for a thematic GIS specifically designed to provide (geo)information on soils (Burrough, 1991). This is a structured digital version of soil maps and soil survey reports associated with data from laboratory analysis. A Conventional SIS consists of:

1. a polygon map, representing the soil bodies;
2. a point map, representing profile observations, and
3. attribute tables representing sampled descriptive and physical or chemical soil properties.

The polygon map is a class-type map, the classes are soil mapping units (further referred to as SMUs) and the profiles are organized into a relational database and linked to the SMUs via their coordinates or soil types (Zinck & Valenzuela, 1990). This system follows the Discrete Model of Spatial Variation (Heuvelink, 1998). The key function of a SIS is to serve the users for data retrieval, spatial queries, statistical analysis and visualisation of results. The profile data is used to make attribute or thematic maps and statistical representations by averaging the values per SMU type or soil type (Burrough, 1993a). Similarly, the SMU's can be directly linked to interpretation tables e.g. soil suitability classes. The above-described system is also referred to as the "conventional approach" to the soil mapping and has been adopted and used in most of the World today, especially at regional and national scales.

For many GIS professionals, working on data integration, a critical layer in a multi-thematic GIS, particularly when utilized in land management decisions, is soil survey information (Maclean *et al.*, 1993). For other SIS external users, such as agronomists, land use planners or civil engineers, the concepts of soil classes and soil mapping units are often harder to grasp and interpret than the land use types or vegetation types. Instead of the map of soil types, the external users are often more interested into the maps of soil interpretations (e.g. suitability for vine production) or limiting land characteristics (e.g. depth to gleying) or technical properties of the soils (e.g. texture fractions, depth to the cemented layers etc.) (Dent & Young, 1981). Moreover, modern users require soil geoinformation at increasingly finer level of detail and increasingly higher accuracy.

There are several likely reasons that conventional soil maps are unpopular among the external users. First, the concept of soil types is probably the fuzziest from all environmental sciences, as the soil bodies are hidden, often irregular or random in

distribution (Burrough *et al.*, 1997). Second, classification systems have been an object of dispute and it was not until the end of the last century that an official international classification system (FAO, 1998) was accepted. Even today, there is still a high chance that two soil surveyors, working independently in the same pit, will identify two different types of soils. Third, analytical procedures are missing in some phases of soil mapping or are not fully documented. For example, the soil boundaries are drawn by following the mental model in surveyor's head rather than by an objective procedure (Cook *et al.*, 1996). Hence, soil survey is still considered by some to be more of an art than a science (Hudson, 1992). The fourth cause of the general low confidence in the soil maps is that their operational quality, i.e. accuracy, lineage and completeness, has often been proved to be lower than expected (Marsman & de Gruijter, 1987; Burrough, 1993a). Finally, the concept of SMUs and related polygon-based organisation of SIS is not immediately suitable for multi-source data integration and quantitative environmental modelling (Ventura *et al.*, 1996). Some more recent conceptual designs of SGDBs, e.g. by Fernandez & Rusinkiewicz (1993), are often unnecessary too complex and therefore user-unfriendly for external users. This is most probably because: (a) the soil surveyors often produce multiple-component mapping units, which are harder (sometimes impossible) to organize and query and (b) SGDB use several entities at the same time (mapping units, pedons, horizons), which can be connected in several ways, thus confusing the external users.

The above-listed problems with the conventional approach have been a major inspiration for researchers in the last decade or two. In early 90's, McSweeney *et al.* (1994) laid the foundation for a new four-stage framework for modelling the distribution of soils. From then, the following two developments have shown to be especially promising: use of auxiliary or secondary data, such as terrain parameters and remote sensing images (Dobos *et al.*, 2000; McKenzie *et al.*, 2000), and use of new concepts and methods, such as continuous classification to model the soils more successfully (McBratney *et al.*, 1997). The use of auxiliary data to improve mapping of soil variables has been especially prominent in Australia (Carlile *et al.*, 2001). Also in the Netherlands, there has been a significant shift towards the quantitative methods for inventarization and utilization of soil data (Buurman & Sevink, 1995). Even in the USA, where the soil mapping is fully dominated by the U.S. Soil Taxonomy and the Soil Survey Manual, there are more and more alternative systems being developed (Zhu *et al.*, 2001). This, however, does not mean that the photo-interpretation or empirical knowledge on soils should be cast out from operational soil survey. On the contrary, case studies have shown that the purely geostatistical methods do not always give prediction maps better than those obtained by subjective photo-interpretation (van Kuilenburg *et al.*, 1982; Boucneau *et al.*, 1998).

In this chapter a grid-based SIS, which integrates the use of photo-interpretation, auxiliary terrain and remote sensing data, hybrid pedometric techniques, continuous classification and advanced visualisation techniques is described. This connects the methods from the previous chapters into a real soil survey application.

7.2 Methods

Three main aspects determine the design of a SIS: (a) concepts and elements used (entities); (b) organizational structure and operations and (c) format and presentation of products. In the following sections, the key concepts and elements used are listed. First, the relation between the grid size and cartographic scale is explained, then a schematic flow of the methodological steps and explanation of algorithms for interpolation, classification, inference, visualisation and (dis)aggregation of data is given. Note that I refer to the proposed SIS as the *hybrid grid-based SIS* in the further text — the adjective 'hybrid' determines both the use of the mixed model of spatial variation and hybrid interpolation technique.

7.2.1 Key concepts

Two key concepts specifically distinguish the SIS proposed in this paper from other similar grid-based SIS applications: use of quantitative methods in all parts of mapping process and combination of different mapping techniques (including photo-interpretation, kriging and correlation with auxiliary maps). The latter ensures a combination of the abrupt and continous transitions in space, which is referred to as the *Mixed Model of Spatial Variation* (Mowrer & Congalton, 2000). This is a combination of the discrete and continuous models of spatial variation, although one might argue that the continuous model already can adopt both continuous and less-continuous (discrete) transitions. The following concepts define the hybrid grid-based SIS more closely:

- The fundamental spatial entity is a grid cell. All GIS layers are brought to same grid resolution in order to make calculations and data integration possible. The grid size (resolution) determines the effective scale.
- The focus is production of maps of key land characteristics. This means that the soil mappers need to interview their users prior to the actual sampling and select the most important land characteristics, level of detail (grid size) and required accuracy. These wishes are then adjusted to the available funds.
- The SIS includes not only maps of soil variables and tables of soil attributes but also auxiliary (non-soil) variables used to assist soil mapping, as well as

derived classifications and interpretations. This means that a SIS user can get a better insight into the original data and extend it with an additional survey or investigate eventual problems with the data.

- Three types of operations are used to produce soil geoinformation from input layers: interpolation, classification and inference. All these are achieved using the GIS operations on grid maps, rather than table calculations.

- Quantitative methods are used to interpolate soil variables (universal kriging), classify (fuzzy k-means) and retrieve them.

- Soil properties, classes, and interpretations are modelled using the mixed model of spatial variation, so that both discrete and continuous transitions are possible.

- The original soil description and measurements are linked to the spatial predictions and interpretations, so that the latter can be updated if the former is augmented or corrected. This linkage is kept in tables built for this purpose. For example, the `interpolation table` records the number of regression coefficients and kriging parameters derived from the regression and geostatistical analysis. Consequently, each prediction or interpretation map can be updated by updating the input maps or adding the new soil samples.

7.2.2 Selection of a suitable grid size

The grid size, i.e. the length of one side of a grid cell, is linearly related to the cartographic scale. However, there are different ideas about the suitable grid size for a given scale. In conventional soil cartography, the scale is commonly assessed by using either the Maximum Location Accuracy (MLA) or Average Size Area (ASA) of the polygons on the ground. For example, MLA on the ground when divided with MLA on the map (e.g. 0.25 mm for maps produced according to common map accuracy standards) gives the scale denominator (Rossiter, 2001). To assess the scale denominator via the ASA, the square root of the nominator should be used. These cartographic definitions can also be used to estimate the suitable grid size for a given mapping scale. As a rule of thumb, Rossiter (2001) suggest that four grid cells should be considered equivalent the Minimum Legible Delineation (MLD). According to the definition of Vink (1975) the MLD is 0.25 cm^2 on the map. The suitable grid size is then:

$$p = \sqrt{\frac{MLD}{4}} = \frac{\sqrt{SN^2 \cdot 0.000025}}{2} = SN \cdot 0.0025 \tag{7.1}$$

where p is the grid (pixel) size, MLD is the Minimum Legible Delineation area on the ground and SN is the scale denominator. This means that for a 1:50 K scale, MLD is 6.25 ha and suitable grid size is 125 m, which seems fairly coarse. Larger grid sizes (0.5 mm to 3 mm on the map) have also been recommended by Valenzuela & Baumgardner (1990). In remote sensing, the relation of the ground resolution and the cartographic scale is somewhat stricter. For example, the Landsat images of 30 m ground resolution are commonly related to the 1:50 K or 1:100 K scale (Lillesand & Kiefer, 2000). Hence, the ground resolution can be defined as two times the MLA on the ground:

$$p = SN \cdot MLA \cdot 2 = SN \cdot 0.0005 \tag{7.2}$$

so for 1:50 K scale, a suitable grid size is 25 m.

The third criterion for the selection of the suitable grid size is empirical knowledge of spatial variation. Ideally, the grid size should equal the minimum size of a pedon (1 m^2), especially if the soils are varying at short distances (e.g. cockpits in the Karst area). If the soils are homogeneous spatially and show smoother transitions, much larger grid sizes will be adequate for spatial modelling (Thompson *et al.*, 2001). This means that the selection of the suitable grid size should be adjusted to the spatial variability of soils to avoid over-sampling. Florinsky & Kuryakova (2000) suggested that, for soil-terrain modelling, adequate grid size is the one that offers the highest predictive power, i.e. correlation coefficient in their case. The spatial variation of soils can be estimated from the terrain data i.e. contour data. Hengl *et al.* (2003b) suggest that the grid spacing should be at least half the average spacing between the contours to represent the most changes in a terrain.

Although these three criteria give a range of possible values, a rule of thumb *the finer the grid size the better* is suggested in the most cases. The importance of the finer grid size has been proven to play an important role especially if the terrain data is used for spatial modelling of soils (Dietrich *et al.*, 1995; Thompson *et al.*, 2001). With increasingly powerful computers and cheap storage, fine grid sizes are feasible for most study areas.

7.2.3 Interpolation, classification and inference methods

Three operations play key roles in the production of geoinformation in the hybrid grid-based SIS: interpolation, classification and inference. Each is explained in more detail down bellow. A flow diagram of the computational procedures is given in Fig. 7.1. The profile data is first combined with a set of predictors to produce continuous field maps of measured soil variables. These are then classified to membership maps using continuous classification and the predefined class centres. Finally,

the interpolated soil variables, auxiliary predictors and derived memberships can be used to derive soil interpretations, i.e. inferred soil geoinformation.

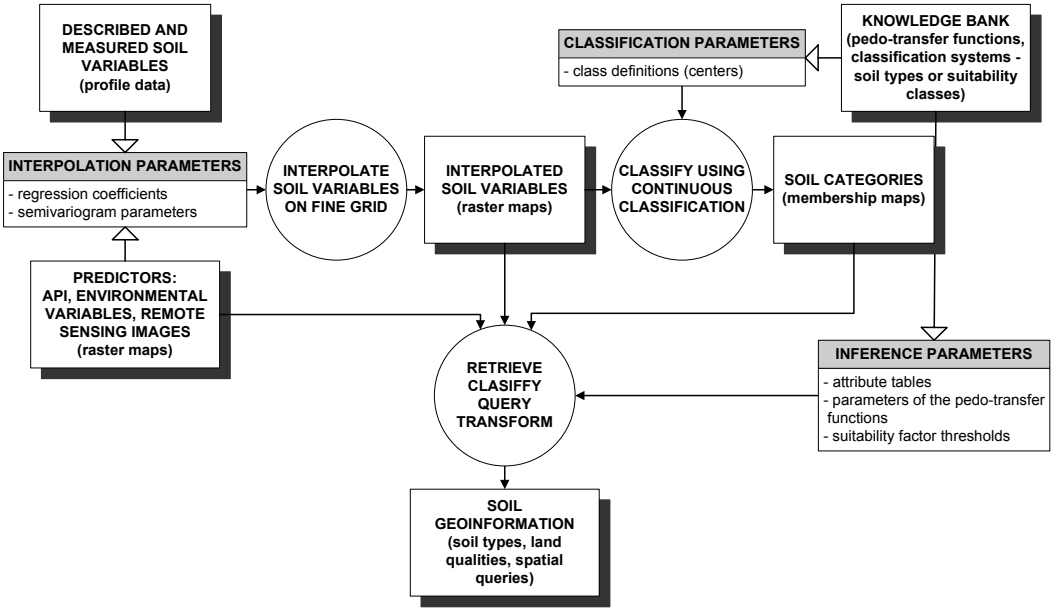

Figure 7.1: Schematic flow of methodological steps.

Interpolation

The generic framework based on the step-wise principal component logistic regression-kriging model, was used to interpolate the soil variables. This algorithm can use information from the photo-interpretation, auxiliary data and spatial auto-correlation at the same time. The algorithm is explained in more detail in chapter 5.

Classification

After all selected soil variables have been interpolated they can be classified using the point observations and the class centres for each category (e.g. soil classes). A flexible classification algorithm is the fuzzy k-means classification, which gives a

membership map for each class. This is the concept of continuous soil mapping, first introduced by **?** and then further on developed by de Gruijter *et al.* (1997). The limitation of their approach, however, is that it employs only geostatistical interpolation while the auxiliary variables are ignored. This approach is somewhat different since first the soil properties are mapped over the whole area and then classified per each grid. This generally means that the produced memberships will follow the pattern of the relief and other predictors, thus giving a more realistic picture. The classification of maps and resulting continuous soil map is explained[2] in more detail in chapter 6.

Inference

The derived memberships, also referred to as similarity values (Zhu *et al.*, 1997), can now be linked to the attribute tables, pedo-transfer functions or suitability ranks (knowledge bank). The key columns can be the soil categories, which is a common way of organizing the SGDB (Zinck & Valenzuela, 1990). The inferred soil attribute is then mapped directly from the membership maps using the linear additive weighting function (Zhu *et al.*, 2001):

$$\hat{S}(i) = \sum_{c=1}^{k} \mu_c(i) \cdot S_c \qquad \sum_{c=1}^{k} \mu_c(i) = 1 \qquad i = 1, .., N \tag{7.3}$$

where $\hat{S}(i)$ is the inferred soil attribute at ith grid position and S_c is the modal value of the inferred soil attribute of the cth category. For example, imagine four membership maps of soil type `A`, `B`, `C` and `D`. The knowledge bank shows that soil type `A` has 10%, `B` 10%, `C` 30% and `D` 40% of clay and the membership values at a grid position are 0.6, 0.2, 0.1 and 0.1, so the Eq. (7.3) will estimate the average clay content of 15%. Note that although the method assumes that a linear weighted average best represents the overall value, the technique can be extended to any aggregation method.

The membership maps can also be used for land suitability assessment. One option is to use the limitation scoring system described by Triantafilis *et al.* (2001). Here, the key issue is to derive limitation scores (or negative points) based on the definition of land qualities and threshold limits. In the case of the hybrid grid-based SIS, the limitation score can be calculated per each pixel by cumulatively using membership maps, interpolated soil variables (e.g. gleying properties) and/or auxiliary variables (e.g. slope):

[2]See also supplementary materials for ILWIS commands.

$$l(i) = \sum_{c=1}^{k} \mu_c(i) \cdot l_c + \sum_{r=1}^{t} S_r(i) \cdot l_r \qquad \sum_{c=1}^{k} \mu_c(i) = 1 \qquad i = 1, .., N \tag{7.4}$$

where l is the accumulated limitation score, l_c is the limitation score of the cth soil type, S_r is the classified auxiliary or soil variable and l_r is the limitation score of the rth class. For example, the same grid position as above (`A`, `B`, `C`, `D`) and the limitations scores 5, 0, 0, 20, give the average limitation score 5. The slope at the same grid position is 10%, which gives 3 more points (9–16%) so that the total accumulative score is 8. The accumulated limitation score, ranging from 0 to ∞ is transformed to continuous land suitability by:

$$L_s = e^{-0.1 \cdot l} \quad L_s \in [0, 1] \tag{7.5}$$

where L_s is the continuous land suitability and l is the accumulated limitation score.

7.2.4 Aggregation and disaggregation

Aggregation or down-scaling is a process of reducing the scale of map and disaggregation is the opposite process. In the grid-based SIS, aggregation means changing towards a coarser resolutions and disagreggation towards finer resolutions, i.e. smaller grid sizes. A schematic example of aggregation and disaggregation in the hybrid grid-based SIS is shown in Fig. 7.2. This models follows the conceptual model of scaling described by McBratney (1998). One advantage of the hybrid grid-based SIS is that the aggregation is easier than with the conventional system where both the soil boundaries and the legend need to be adjusted. In the grid-based SIS, each interpolated continuous soil variable can be resampled to a coarser grid using standard image processing algorithms such as bilinear resampling (Lillesand & Kiefer, 2000). The scaling of the continuous variables is much less problematic than the scaling of categorical variables, such as soil types. The resampling of soil types to a coarser resolution implies that the small local patches will be merged with the dominant types and disappear from the map. Because we deal with maps of soil memberships, we can first resample these to a coarser resolution and then re-standardize them by:

$$\mu_c^{S-}(i) = \frac{\mu_c^{+}(i)}{\sum_{c=1}^{k} \mu_c^{+}(i)} \quad c = 1, 2, ...k \quad i = 1, 2, ...N \tag{7.6}$$

where μ_c^{S-} is the down-scaled membership value and μ_c^{+} is the resampled membership. A longer alternative is to re-calculate soil variables and re-classify soil types from the input maps at finer resolutions.

The hybrid grid-based SIS is also attractive for the purpose of up-scaling, which is in the conventional SIS almost impossible. Because the accuracy of interpolation depends on the quality and detail of auxiliary variables (terrain data, remote sensing images), one can imagine that improving the spatial detail of the predictors will also reflect on the interpolation results. A caution should be made not to 'blow-up' the scale outside the realistic limits defined by the standards. For example, if the inspection density is four observations per km^2 the largest scale that the existing dataset can be disaggregated to is 1:25 K. Additional observations are recommended to achieve larger scales.

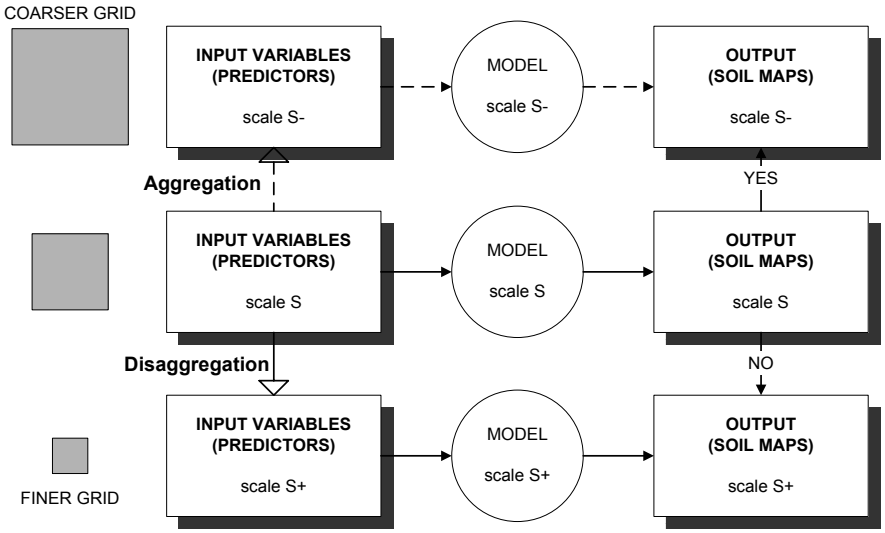

Figure 7.2: Schematic example of aggregation and disaggregation process in the hybrid grid-based SIS. Note that although direct disaggregation of soil maps is possible, it is not recommended. S indicates scale: S^- are smaller scales and S^+ are larger scales.

7.2.5 Case study and data analysis

The methodology was developed and tested using a data set from Baranja hill and a portion of the adjacent Danube terraces in Eastern Croatia. The study area is 3.8×3.8 km square (centred on 45°47'40" N, 18°41'27" E) and corresponds to the

size of a single 1:20 K aerial photo (Fig. 7.3). The main geomorphic facets are hill summits and shoulders, eroded slopes of small vales, vale bottoms and high and low river terraces. The elevations range from 80 to 240 m. I first produced an API map using the geopedological approach of Zinck & Valenzuela (1990). I then made 59 profile observations using a random design (40) and two transect studies (19) (Fig. 7.3c). The boundaries were finally cross-checked on the field to produce a conventional soil map with the legend.

The observed soil types ranged from Calcaric Regosols, Cambisols to Kastanozems (FAO, 1998). The Calcaric Cambisols are the dominant soil type in the hilland, while in the vale bottoms and in the lower floodplain, I observed gleyic properties. At some locations on the hill summits, I observed occurrence of a hypocalcic horizon ($> 15\%$ calcium carbonate equivalent). This layer is neither cemented nor close to the surface so it does not present a limitation for agriculture. I observed the following land use types: vineyards, orchards, natural grasslands, meadows (for animal production), natural forest and woodland (hunting resorts), residential use, fish pond, water control (channels), animal farming and crop fields. The most common crops were maize and wheat, vegetables (manual farming), sugar beet and sunflower.

The most controlling factors for agricultural management in the area are: slope, solum thickness, soil alkalinity and water-saturation conditions. Finally, I selected the following six soil variables as the most important diagnostic land characteristics:

1. Depth to the parent material , i.e. thickness of solum (SOLUM) measured in cm.

2. Occurrence of the gleying properties (GLEY_P) — coded with "0" for not observed, "1" for gleying properties within 50 cm and "0.5" for gleying properties within 50 cm.

3. Occurrence of the Mollic horizon (MOL_H) — coded with "0" for not observed and "1" for observed Mollic horizon.

4. Occurrence of the Calcic horizon (CALC_H) — coded with "0" for not observed and "1" for observed Calcic horizon.

5. Thickness of the topsoil (A_DEPTH) measured in cm.

6. Silt fraction (0.002–0.05 mm) content in topsoil (A_SILT) estimated using the centroids of the textural classes and expressed in percentage. The texture classes ranged from sandy-loam, loam, silt loam to silty clay loam.

Note that the indicator variables GLEY_P, MOL_H and CALC_H have either 0 and 1 value which can not be transformed (see chapter 5, page 94). To avoid

division by zero or ln(0) problems, I introduced a small adjustment of 0.01, so that 0 becomes 0.01 and 1 becomes 0.99. A more optimal approach would be to estimate these threshold iteratively in a statistical package.

The working scale of the project was 1:50 K, hence, a grid size of 25 m, which corresponds to 0.5 mm on the map was selected. For the predictors, I used six terrain parameters (Hengl *et al.*, 2003b): elevation (DEM), slope gradient in % (SLOPE), profile curvature (PROFC), plan curvature (PLANC), wetness index (CTI) and slope insolation (SINS); all derived in ILWIS[3]. As remote sensing-based predictors, I used the intensity (value on the grey scale) of the aerial photo (AP), the standard deviation image filter of the AP map (AP_STD) and NDVI map derived from the Landsat 7 image. The aerial photo was taken in May 1998 and the satellite image in August of 1999. I assumed that these remote sensing-based variables would help explain the occurrence of horizons and depths. The nine maps were first transformed to nine predictive components (SPCs) using factor analysis in ILWIS. This was done to reduce the multicollinearity and optimize the selection of the best subset of predictors[4].

In addition to the SPCs, nine soil mapping units (SMUs) were transformed to nine indicator variables: colluvial footslopes (SMU1), eroded slope (SMU2), floodplain (SMU3), glacis (SMU4), high terrace (SMU5), scarp (SMU6), shoulder (SMU7), summit (SMU8) and vale bottom (SMU9). We also added three land use indicator variables: agricultural land (LU1), natural forest (LU2) and pastures and orchards (LU3). The total number of predictors was 21 (Fig. 7.3a and b). The target soil variables and the predictors were imported to a regression table consisting of 59 observations, 9 target variables and 21 predictors. The 'best' subset of predictors (SPCs, SMUs and LUs) was selected using the stepwise regression in the S-PLUS statistical package (MathSoft Inc., 1999). The regression coefficients and interpolated the residuals were then calculated over the whole study area using the regression-kriging (see chapter 5).

The set of nine interpolated soil maps was further used to classify the whole area. The membership maps were calculated using the supervised fuzzy k-means classification. First the class centres were calculated by averaging the nine soil variables per soil type. For the indicator soil variables the sampled standard deviation was zero, which is unsolvable. The indicator variables follow a binomial distribution, so that the standard deviation can be estimated using:

$$\hat{s}_z = \sqrt{\frac{p \cdot (1-p)}{k}} \tag{7.7}$$

[3]See lecture note "Digital Terrain Analysis in ILWIS", available with supplementary materials.
[4]See chapter 5 for more details.

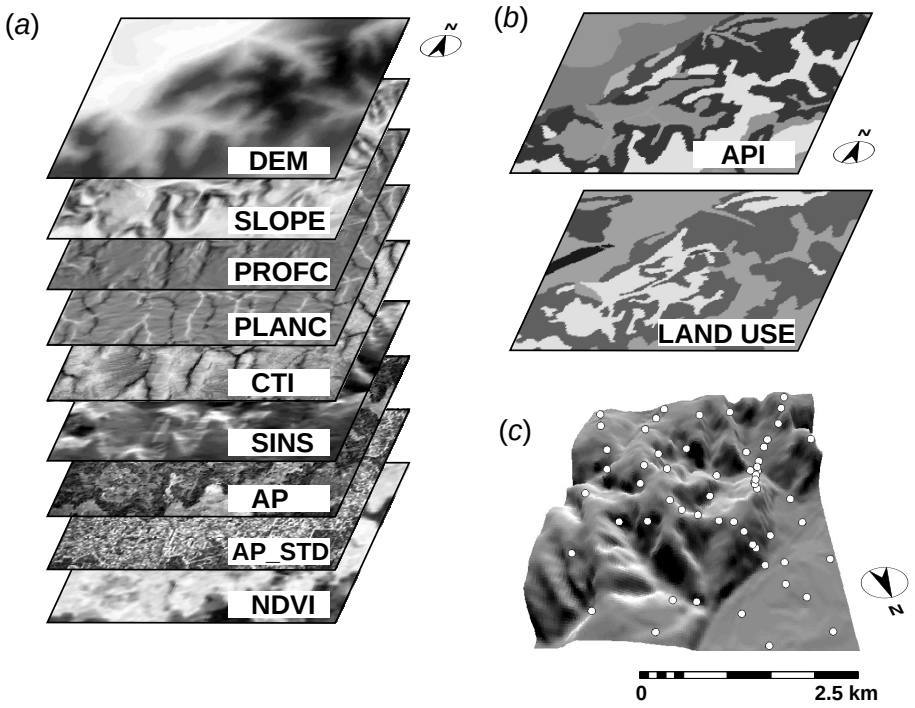

Figure 7.3: Multi-source predictors: (a) auxiliary predictors terrain parameters and remote sensing data; (b) aerial photo-interpretation map (API) and land use map and (c) location of the 59 soil profile observations. DEM – elevation; SLOPE – slope gradient in %; PROFC – profile curvature; PLANC – plan curvature; CTI – wetness index; SINS – slope insolation; AP – intensity of the aerial photo; AP_STD – standard deviation of the AP map and NDVI map derived from the Landsat 7 image.

where p is the threshold probability (e.g. 95%) and k is the number of classes. In the case of MOL_H and CALC_H, the number of classes is two and the standard deviation is 0.15, while in the case of GLEY_P the standard deviation is 0.13.

Membership maps for the six observed soil types were derived: Siltic, Calcisols (CL_s), Calcari-Eutric Cambisols (CM_ce), Gleyi-Calcaric Cambisols (CM_gc), Calcari-Eutric Gleysols (GL_ce), Calci-Siltic Kastanozems (KS_cs) and Calcari-Eutric Regosols (RG_ce). The memberships were then used to derive the limitation score for the land utilisation type (wheat) using the soil types and slope classes as input (Eq. 7.4)). In addition, the membership values were resampled to the 100 m grid using the Eq. (7.6) to demonstrate disaggregation aspects.

7.2.6 Comparison of conventional and hybrid grid-based SIS

The hybrid grid-based SIS was compared with the conventional polygon-based SIS of the same area. I first compared the predictability of SMUs and SPCs, which gives an idea which predictors explain the measured soil variables better. This was done by comparing the correlation coefficient and coefficient of determination between all target variables and predictors. The two systems were also compared for their mapping efficiency using: amount of variation explained and thematic confusion. Amount of variation explained was assessed by calculating the sum of squared residuals, i.e. *RMSE* for each of the six interpolated variables. The lower the *RMSE*, the better is the fitting of the data. The thematic confusion was assessed by calculating the confusion index among each spatial entity:

$$CI = 1 - (\mu_{\max} - \mu_{2nd\,\max}) \tag{7.8}$$

where $\mu_{\max}$ is the highest membership and the $\mu_{2nd\,\max}$ is the second highest membership at the same location (Burrough *et al.*, 1997). The lower the CI, the higher the certainty of the classification system. Note that the CI for SMUs is calculated by first calculating composition of soil types in percentage. The CI value is then attributed to each SMU to derive the overall or average confusion index. In addition to the statistical measures, a summary comparison of the two systems for their cost-effectiveness, flexibility and technical properties was made.

7.3 Results

7.3.1 Mapping soil variables

The factor analysis on the continuous predictors showed that the information overlap is low. The first three SPCs accounted for about 65% of the total variation and it appears that all SPCs need to be taken into account. A first comparison of correlation coefficients between the all combination of SPCs and SMUs with target variables showed that the auxiliary predictors are slightly more correlated with the target soil variables than the SMUs (Fig. 7.4a). However, the amount of variation explained in the multivariate models (adjusted R^2) showed that the SMUs are in general better predictors than the SPCs, except for SOLUM and CALC_H (Fig. 7.4b). In all cases, except for CALC_H, the regression models explained about 40% of variation and were statistically significant (p <0.001). Note that the discrepancy between the univariate correlation coefficients (r) and coefficients of multiple determination (R^2) in Fig. 7.4 is because there is still some thematic overlap in the SPCs. The SMUs (indicator variables) have no overlap by definition so that lower univariate correlations will accumulate more effectively in the multivariate model.

In all cases the step-wise regression selected from 3 to 6 predictors from the 21 possible, or 25% in average (Table 7.1). The best predictors were:

- for SOLUM – SPC1 (CTI, SLOPE), SPC3 (AP_STD) and SMU5 (high terrace);
- for GLEY_P – SMU3 (floodplain area), SMU9 (vale bottom) and SPC9 (CTI)
- for MOL_H – SMU4 (glacis), SMU5 (high terrace) and SPC9 (CTI);
- for CALC_H – LU2 (natural forests) and SMU2 (eroded slope);
- for A_DEPTH – LU1 (agricultural land) and SMU5 (high terrace) and
- for A_SILT – SMU9 (vale bottom), SPC9 (CTI) and SMU3 (floodplain area).

Many predictors, on the other hand, have been ignored by the system, such as SPCs 2,6,7,8, SMUs 1,6,7,8 and LU3. The models in general reflect our empirical idea of the distribution of soils. For example, I observed the gleyic properties in only two mapping units and assumed that these are closely related with the potential of water accumulation, which was also confirmed by the model (SMU3, SMU9 and CTI). In the case of CALC_H, the current predictors are of little help. It seems that this variable is controlled by the parent material and not geomorphology or land use. Note that the adjusted R^2's (Table 7.1) are somewhat higher than the ones in the Fig. 7.4. This is because a lower number of predictors is used for final prediction, which typically means a lower adjusted R^2.

The geostastical analysis of the residuals showed the pure nugget variation for the SOLUM and GLEY_P, fairly long-range spatial dependence for MOL_H and CALC_H and somewhat shorter-range spatial dependence for A_DEPTH and A_SILT (Table 7.1). The pure nugget effect for residuals is reasonable for GLEY_P because most of the variation (70%) has been accounted for by the model. For SOLUM, the pure nugget effect is somewhat more surprising since the residuals are still significant. In this case, only 37% of the total variation has been explained by the regression analysis. This means that SOLUM is much noisier variable and much harder to map, which is probably due to the fuzzy character of the boundary between the solum and parent material. The ordinary kriging of residuals practically 'saved' the prediction of CALC_H, despite the poor regression model. The residuals, however, showed strong spatial dependence, which was sufficient to map it using ordinary kriging.

A visual comparison of the interpolated maps produced using the conventional approach (Fig. 7.5, left) and hybrid interpolation (Fig. 7.5, right) suggests that the hybrid system in general offers more detail and higher contrast. In the case of the

Table 7.1: Soil variables (logit-transforms), selected sub-sample of predictors, adjusted R^2 and estimated variogram parameters.

		Soil variables (logit-transforms)					
		SOLUM^{++}	GLEY_P^{++}	MOL_H^{++}	CALC_H^{++}	A_DEPTH^{++}	A_SILT^{++}
Regression coefficients (predictors)	Intercept	-0.72	-1.953	-2.848	-3.888	-2.014	-0.624
	SPC1	0.0114	-0.002	0.0284	0	-0.002	0.0026
	SPC2	0	0	0	0	0	0
	SPC3	0.0178	0	0	0	-0.008	0
	SPC4	0	-0.004	0	0	0	0.0029
	SPC5	0.013	0	0	0	0	0
	SPC6	0	0	0	0	0	0
	SPC7	0	0	0	0	0	0
	SPC8	0	0	0	0	0	0
	SPC9	-0.013	-0.043	-0.058	-0.018	-0.01	0.0111
	SMU1	0	0	0	0	0	0
	SMU2	-0.18	0	0	1.0332	0	0
	SMU3	0	4.965	0	0	0	-0.622
	SMU4	0	0	8.7559	0	0	0
	SMU5	0.3211	0	8.8444	0	0.7607	0
	SMU6	0	0	0	0	0	0
	SMU7	0	0	0	0	0	0
	SMU8	0	0	0	0	0	0
	SMU9	0	5.9859	0	0	0	-0.777
	LU1	0	0	0	0	0.4484	0
	LU2	0	0	0	1.4065	-0.066	0
	LU3	0	0	0	0	0	0
	R^2_a	0.37	0.70	0.59	0.13	0.41	0.61
Variogram	Variogram model	nugget effect	nugget effect	exponential	exponential	exponential	exponential
	C_0	0.156	3.78	0.27	0	0	0
	C_0+C_1	0.156	3.78	26.2	8.28	0.192	0.122
	R (m)	0	0	10 km	759	194	69

hybrid systems, not only discrete and continuous transitions can be seen, but also the pattern of relief or land use is reflected via the auxiliary maps. This hybrid pattern is especially distinct in the map of A_SILT: the highest values follow the steeper slopes, discrete transitions are visible in the floodplain area but also the kriging pattern with hot spots (Fig. 7.5c, right).

The conventional system is more sensitive to the fairly contrasting inclusions in the mapping unit. For example, the prediction map of the GLEY_P for the conventional system shows a value of 0.1 even at locations where no gleying could have occurred (Fig. 7.5b, left). This is because there was a single profile (inclusion) in this mapping unit, which somehow finished in the neighbouring polygon (probably boundary misplaced during API). This affected then the whole attribute map giving an unrealistic prediction of occurrence of gleying properties.

Comparison of the *RMSE* at observation points for these six variables showed no large difference for SOLUM (17.4 cm vs. 17.8 cm) and GLEY_P (0.18 vs. 0.13), but in all other case was the data better fitted with the hybrid interpolation technique (0.21 vs. 0.02 for MOL_H, 0.26 vs. 0.01 for CALC_H, 8.6 cm vs. 0.7 cm for A_DEPTH and 7.8% vs. 1.1 for A_SILT).

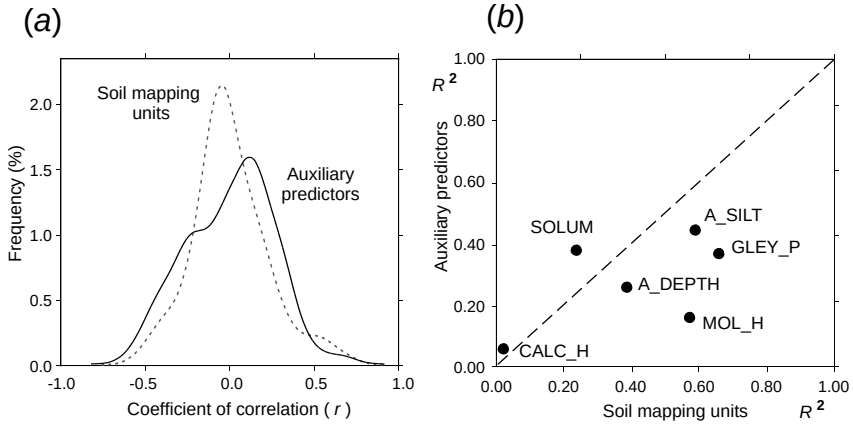

Figure 7.4: Comparison of relationships between the soil variables and soil mapping units and auxiliary predictors: (a) density histograms of the correlation coefficients for univariate models and (b) coefficients of multiple determination for fitted soil variables. SOLUM - depth to the parent material in cm; GLEY_P - occurrence of the gleying properties; MOL_H - occurrence of the Mollic horizon; CALC_H - occurrence of the Calcic horizon; A_DEPTH - thickness of the topsoil in cm; A_SILT - silt fraction content in topsoil.

7.3.2 Classification, down-scaling and inference

The classified map of soil types (Fig. 7.6b) reflects empirical ideas, following the fieldwork experience, about the distribution of the soils. The CM_ce is the dominant soil type covering 61% of the study area, CM_gc and GL_ce occur as expected at lowest convex positions, while the RG_ce occurs more locally (slopes). The CL_s was depicted as the highest membership in only 0.6% of the study area and as the mapping of Calcic horizon was difficult.

From the sampled class centres for the six soil types (Table 7.2), it can be seen that some classes can be distinguished in the attribute space more easily then others. For example, KS_cs is clearly a distinct soil type: deep soil, with occurrence of Mollic horizon and no gleying properties. The factor analysis of class centres also showed that especially CM_ce and RG_ce; and CM_gc and GL_c are similar soil types. This information about the similarity of soils was then used to produce a fuzz-metric legend and then visualise soil taxa and problematic areas as a continuous soil map (Fig. 7.6c). This mixed-colour map indeed shows highest classification uncertainty between the CL_s and KS_cs (note the white patches in Fig. 7.6c). This information can now be used to collect additional samples or cross-check accuracy of our classification system. Also note that the continuous soil map shows three major groups of soil types indicated as bluish (CM_ce, RG_ce and CL_s), greenish (GL_ce, CM_gc) and reddish (KS_cs).

The average confusion index for the conventional SIS, calculated using Eq. (7.8), was 51% (±28%) for the whole map. The confusion index for the hybrid grid-based SIS was 17% (±14%) in average (see the legend in Fig. 7.6a). This means that the spatial confusion between the membership maps is significantly lower ($p < 0.05$) than the confusion within the SMUs for the conventional SIS. After the down-scaling (100 m grid), the less frequent classes did not disappear from the map as we would have expected. For example CL_s occupies about 9 ha in the 100 m scale map, while it occupied 7.9 ha in the 25 m scale map (Fig. 7.6d). This means that the proposed aggregation algorithm retains smaller-size features if their membership is more distinct.

From the membership maps and classified slope map the accumulated limitation score and the resulting continuous land suitability for wheat were derived. The schematic example of the calculation is shown in Fig. 7.7. I used the following limitation scores: 3 (CL_s), 1 (CM_ce), 1 (CM_gc), 9 (GL_ce), 0 (KS_cs) and 9 (RG_ce) for soil types and 0 (0-2%), 1 (2-8%), 3 (9-16%), 9 (17-25%) and 27 (> 25%) for the slope classes.

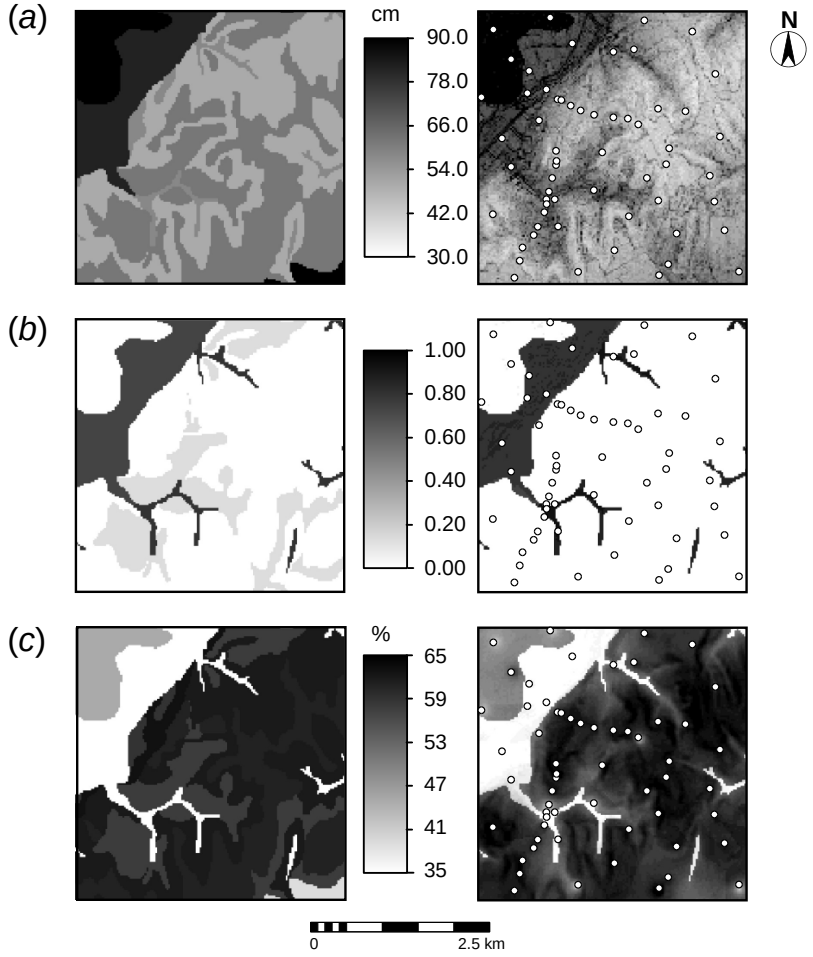

Figure 7.5: Comparison of (a) depth to the parent material (SOLUM); (b) occurrence of the gleying properties (GLEY_P) and (c) silt fraction content in topsoil (A_SILT), interpolated using the mapping units only (left) and the hybrid interpolation algorithm (right).

7.4 Conclusions and discussion

In this chapter I have presented some key concepts, operations and organizational issues of a grid-based SIS as an alternative to the conventional polygon-based SIS and plain geostatistical techniques. The proposed hybrid grid-based SIS was not developed for purpose of replacing conventional techniques and concepts, replac-

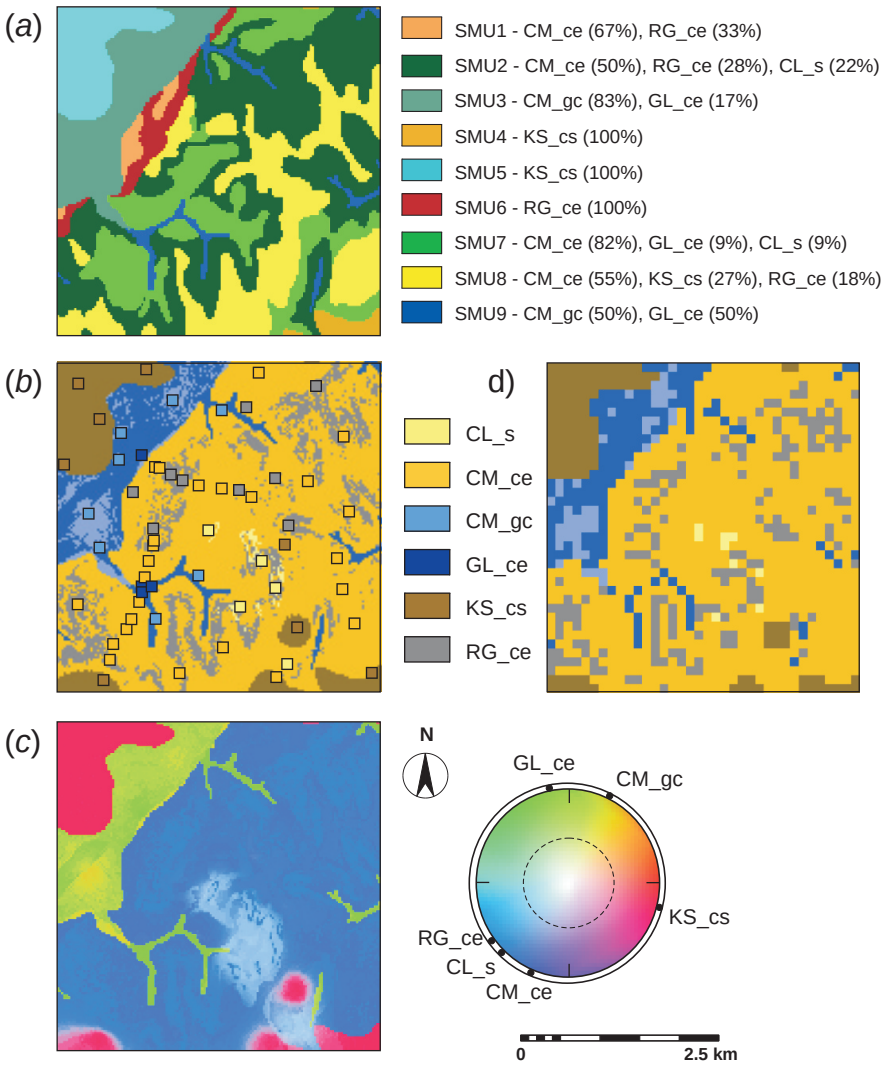

Figure 7.6: Comparison of (a) the conventional soil map with compound composition of mapping units, (b) defuzzified (highest) membership map from the supervised fuzzy k-means classification with freely selected colours; (c) the continuous soil map with a circular legend and (d) down-scaled map to 100 m grid. CL_s - Siltic, Calcisols; CM_ce - Calcari-Eutric Cambisols; CM_gc - Gleyi-Calcaric Cambisols; GL_ce - Calcari-Eutric Gleysols; KS_cs - Calci-Siltic Kastanozems and RG_ce - Calcari-Eutric Regosols.

Table 7.2: Class centres used to classify the six soil types from six attributes.

	Sampled class centres and variation around the central values					
	SOLUM	GLEY_P	MOL_H	CALC_H	A_DEPTH	A_SILT
	(σ)	(σ)	(σ)	(σ)	(σ)	(σ)
	cm	-	-	-	cm	%
CL_s	37.4	0	0	1	17	63
	(11.4)	(0.13)	(0.15)	(0.15)	(12)	(5.1)
CM_ce	60.16	0	0	0	22.48	61
	(16.3)	(0.13)	(0.15)	(0.15)	(6.9)	(8.6)
CM_gc	77.75	0.5	0	0	32.5	37.3
	(14.5)	(0.13)	(0.15)	(0.15)	(14.5)	(3.2)
GL_ce	63.75	1	0	0	23.25	29.5
	(25)	(0.13)	(0.15)	(0.15)	(4.6)	(7.1)
KS_cs	92.88	0	1	0	47.13	51
	(14.2)	(0.13)	(0.15)	(0.15)	(5.5)	(12.8)
RG_ce	36.67	0	0	0	17.22	61.6
	(15.2)	(0.13)	(0.15)	(0.15)	(6.5)	(4.3)

ing existing soil databases or devaluating the importance of photo-interpretation or existing classification systems, but to employ these in a more objective manner. Moreover, the proposed hybrid grid-based SIS is a generalization of the conventional approach. One can imagine that if the within-unit variability is infinitively small and if there is no overlap between class definitions, than the hybrid SIS will show the same, so-called, "double-crisp" form (crips objects and crisp classes) as a conventional map. In fact, in our case study the API units played an important role and the transition of soils was, consequently, more discrete in many parts of the area.

The summary comparison of the two systems can be seen in Table 7.3. The important advantages of the hybrid grid-based SIS that need to be emphasized are:

- It directly offers a map of soil types rather than a map of the soil-mapping units.

- All variables, including the soil types and land suitability are mapped in a continuous manner and on fine grain of detail. In this case study, the average

Table 7.3: Summary comparison between the conventional polygon-based and grid-based SIS. The technical details apply to the study area.

Aspect	Polygon-based	Grid-based
Entity	Polygon	Grid
Detail (average size area)	33.8 ha (581 m)	0.0625 ha (25 m)
Content	Polygon class-type map linked with attribute tables (profile observations)	Set of grid maps linked with attribute tables (regression coefficients, variogram parameters, central values, limitation scores)
Interpolation method	Averaging per SMU or soil type	Regression-kriging
Products	Distribution of soil mapping units with composition; soil profile database; crisp land suitability	Distribution of soil variables (land characteristics), soil types and land suitability with estimated uncertainty
Purity of entities (confusion index)	low (51% in average)	high (17% in average)
Level of detail and reliability of predictions	Only average or modal values; contrasting inclusions may be listed separately	Higher level of detail; the predictions follow the pattern in relief, vegetation or land use, according to factors included in the model
Data input and analysis	API by surveyors conceptual knowledge; lines are digitized; topology is created in GIS ; soil profile observations are organized in a relational database	Auxiliary maps are obtained from secondary sources; computations can be demanding and the end product depends on the quality of the input data and algorithms used for interpolation
Memory use	Single vector map and set of tables (very low); 10 KB per km^2 at 1:50 K	About 21 map of predictors, 9 maps of transformed predictors (SPCs), 6 maps of soil variables, 6 maps of soil types etc. (very high); 400 KB per km^2 at 25 m resolution

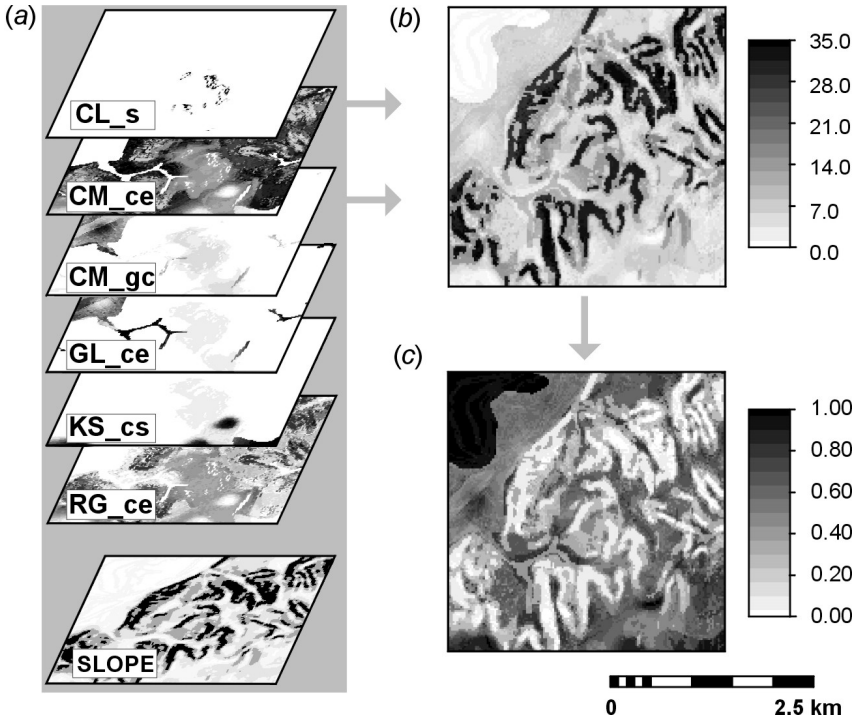

Figure 7.7: Mapping continuous land suitability for wheat: (a) memberships for soil types and slope classes; (b) accumulated limitation score and (c) continuous land suitability. CL_s - Siltic, Calcisols; CM_ce - Calcari-Eutric Cambisols; CM_gc - Gleyi-Calcaric Cambisols; GL_ce - Calcari-Eutric Gleysols; KS_cs - Calci-Siltic Kastanozems; RG_ce - Calcari-Eutric Regosols and SLOPE - slope gradient in %.

size of detail was about 25 times smaller for the grid-based SIS.

- The products of mapping are not only maps of soil variables but also the respective prediction uncertainty (i.e. prediction error or confusion index).
- Maps are more suitable for integration with other geo-data.
- It in general provides more reliable soil geoinformation with lower thematic confusion and higher level of detail than the conventional survey.
- The original soil observations and interpolation/classification parameters are linked to the GIS calculations via the special tables and can be updated.

On the other hand, the disadvantages of the hybrid grid-based SIS are:

- It is computationally demanding as it requires number of GIS, statistical operations with each variable. It also consumes a lot of memory: I estimated that for this case study the memory consumption per km^2 is about 40 times bigger for grid-based SIS.
- It requires number of auxiliary variables, which also means somewhat higher investments.
- Because it is data-driven, it fully depends on the quality of the input data. The prediction maps, however, can always be saved with a good API map and manual correction of problematic features.

The number of observations also plays an important role. In this case study I have dealt with a small case study and relatively small number of profile observations. This caused some problems for the fitting of the data, variogram modelling and factor analysis of the thematic similarity. A much larger number of observations, predictors and soil variables will probably be more satisfactory to the real users. I also experienced problems with interpolation of some variables. In this case study this was occurrence of the calcic horizon, which seems to be difficult with this set of predictors. This feature could have been probably explained better with the use of parent material as auxiliary map.

Also note that some of the applied algorithms, such as the continuous land suitability, are not completely satisfactory. Although this method objectively combines limitations, it depends entirely on the subjective assignment of limitation scores to classes, and also on the concept that a linear combination best expresses suitability.

A more flexible system will be to keep all original data in original cell size (or as sample points) and up or downscale as necessary depending on the algorithm. The input data often comes at different resolutions (multi-source data), for example, terrain data may be available at finer resolution (10 m), satellite data at coarser resolutions (30 m) or very coarse resolutions (1 km). Calculations with raster maps of different resolutions without resampling, however, are still not possible in many GIS packages. Another improvement would be to use the kriging by moving window and not the global estimation of the regression residuals. This would, however, require even more input points and computational power.

Chapter 8

Adequacy of soil resource inventories*

"What is wrong with soil maps? The problem is that the soil scientists make the soil maps for them selves... they spend more time on fighting about the soil classification in the pit, than talking with the farmers!"

[by Boško Miloš, Institute for Adriatic crops in Split, discussion during the fieldwork in Croatia]

*based: on Hengl T., Husnjak S. and Rossiter D.G., 2004? Assessing adequacy and usability of soil resource inventories: The National soil inventory in Croatia. to be submitted.

8.1 Introduction

The USA National Committee for Digital Cartographic Standards defines quality as the *"suitability of the data for the intended use"* (Moellering, 1987). Five main elements determine the overall quality of a map: lineage, positional and attribute accuracy, logical consistency and completeness. These correspond to the quality measures and standards, agreed to by the International Cartography Association and applicable to any GIS (Guptill & Morrison, 1995). Soil surveyors have developed a concept of adequacy of a soil resource inventory, which was first introduced by a group at Cornell University and then further on developed in collaboration with workers from many survey organizations (Soil Resource Inventory Study Group, 1977, 1978; Forbes *et al.*, 1982). This group proposed that adequacy, also called *"fitness for use"*, should be evaluated by using four aspects:

- map scale and texture;
- map legend;
- base map quality, and
- ground truth, also called thematic accuracy.

The last aspect has received attention from Dutch mappers (de Gruijter & Marsman, 1984; Marsman & de Gruijter, 1986).

Recently, the concept of usability has been introduced of which the data quality is considered to be just one element (Wachowicz *et al.*, 2002). The difference between quality, adequacy and usability is that quality is a constant measure, the adequacy changes within the problem-solving context, while usability reflects all these elements in relation to the end-users satisfaction (Fig. 8.1). Adequacy is usually related to the concept of effective scale: a lower quality GIS will show higher adequacy if we use it at smaller scales. Usability goes beyond these concepts and does not necessarily show direct relation to the overall quality, i.e. a dataset of high quality does not have to be usable and vice versa. For example, a high quality soil map can finish being unusable if:

- It requires professional knowledge on how to interpret it, which is not available;
- It is hard to integrate the soil map within an existing GIS;
- It requires specific hardware/software in addition; or
- The price is too high for the given project. A simpler product of lower quality may be preferred by the users and therefore more usable.

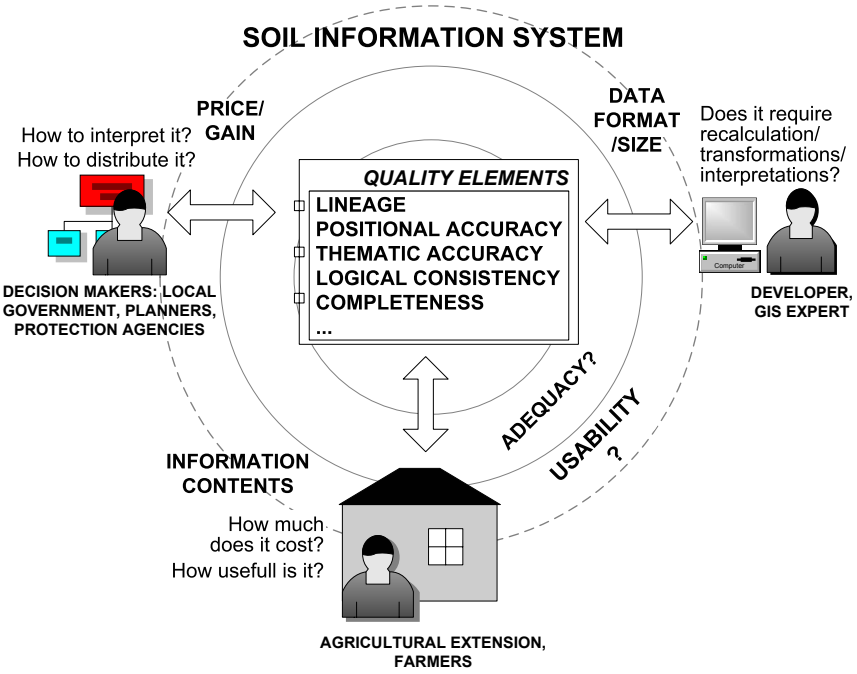

Figure 8.1: Relation between the quality elements, adequacy, usability and users of a SIS.

Several studies in the past thirty years have shown that the technical quality of the soil maps, especially the ones produced through national soil surveys, has often been overestimated. For example in the Netherlands, although the taxonomic purity of the soil delineations was intended to be >70% (Bie & Ulph, 1972), Marsman & de Gruijter (1986) showed that the actual purity is usually considerably lower. Similarly, it has often been emphasized that very few soil profiles inside a well-defined soil mapping unit (further referred to as SMU) actually meet all specifications of the mapping unit (Burrough, 1991). The second serious problem is that the technical aspects of the data quality of soil inventories have often been neglected. Groot (1993) estimated that 80% of the soil information in the world is unusable due to incompleteness, unknown reliability or inconstant spatial referencing.

This chapter gives methodological framework to assess the spatial accuracy of soil boundaries and effectiveness of soil maps. This methodology was used to assess quality elements of the National Soil inventory in Croatia at scale 1:50 K, also called *"Basic Soil Map of Croatia"* (BSMC), which lasted almost 25 years and took about 10 800 profile observations.

8.2 Methods

We selected six map sheets (from the total of 185), performed three control surveys ("Gustirna", "Kalinovac" and "Popovac", each approximate size 4×4 km) and re-described ten profiles of the BSMC in the main landscape regions of Croatia. This was done to estimate the effective map scale, accuracy of map legends, spatial accuracy of soil boundaries, and thematic accuracy of profile observations (Table 8.1 and Fig. 8.2). The control surveys were compared to the original soil map sheets and generalised digital 1:300 K soil map of Croatia (Bogunović *et al.*, 1998). We also used the Croatian soil profile database (Martinović & Vranković, 1997) consisting of 2198 observations to assess the sampling density and thematic contrast of SMU. Considering the spatial accuracy of soil boundaries, we measured about 24 km of soil boundaries or 0.2% from the total length of the boundaries, estimated to be 12 K km. The control surveys are small compared to the original data sets. However, they were well spread over the main geographical regions (Fig. 8.2).

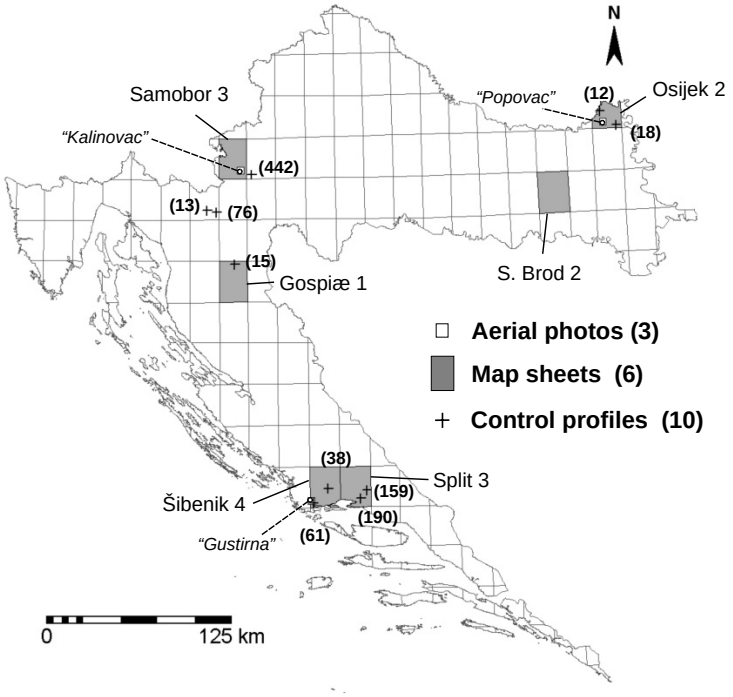

Figure 8.2: Location of map sheets, control survey areas and profile observations (in brackets) used to assess the adequacy elements of the National soil inventory in Croatia.

Table 8.1: Adequacy aspects and data sources and measures used to evaluate them.

Aspect	Data sources (control)	Criteria
Effective map scale	6 map sheets at scale 1:50 K	Average size delineation; Shape complexity index; Inspection density
Accuracy of map legends	3 control survey areas with 20 augerings and 20 minipits per survey area (120 in total)	Chi-square test statistics
Spatial accuracy of soil boundaries	2 master lines (between the most contrasting polygons) in each of the 3 control survey areas (6 in total)	Area of disagreement and mean error
Thematic accuracy of profile observations	10 profile observations	Mean root square error
Thematic contrast of SMUs	2198 soil observations from the soil database together with the 1:300 K digital soil map	Coefficient of variation inside the SMUs and interclass correlation; Average probability of thematic overlap

All data was processed in the ILWIS 3.1 (Unit Geo Software Development, 2001) and ArcView 3.2 GIS packages, while the statistical analysis was done in S-PLUS (MathSoft Inc., 1999). ArcPad 5 and 6 (Environmental Systems Research Institute, 2000) were used for field navigation and initial data processing. This software was running on the iPAQ Compaq pocket PC, to which a GPS receiver (CRUX II GPS PCMCIA card) was attached, which made all together a light, compact and reliable navigation system (further referred to as the mobile GIS).

Profiles were described using the standard national soil survey methodology (Kovačević & Jakšić, 1964) and classification system (Classification of Yugoslav Soils, further referred to as CYS) used in the BSMC and still used in all republics of former Yugoslavia (Škorić *et al.*, 1985). Laboratory analysis was done at the Soil Science Department in Zagreb using the same methods as in the original survey. All this allowed us to expect to describe and measure similar soil properties.

8.2.1 Map scale

The effective map scale was evaluated for six map sheets from different regions in Croatia. We first excluded the non-soil areas, such as urban areas and water bodies. We then used only full polygons, i.e. those completely within the map sheets, in the calculations. From these, the average size delineation (ASD) was calculated as:

$$ASD = \frac{\sum_{j=1}^{m} A_j}{m} \tag{8.1}$$

where A_j is the area of jth polygon and m is the total number of polygons (Forbes *et al.*, 1982). We then calculated the index of maximum reduction (IMR), i.e. factor by which the scale of the map could be reduced before the ASD would be equal to the minimum legible delineation (MLD), which is as a rule of thumb taken to be 0.4 cm^2 on the map:

$$IMR = \sqrt{\frac{ASD}{MLD}} \tag{8.2}$$

From this, the effective scale number (ESN) is computed as:

$$ESN = NSN \cdot \frac{IMR}{2} \tag{8.3}$$

where NSN is the nominal scale number. The factor of two in the denominator ensures that the ASD is four times the size of the MLD, i.e. the arbitrarily-defined optimum legible delineation (OLD) (Forbes *et al.*, 1982). To describe the general geometry of soil polygons, we used a shape complexity index S, which is the perimeter-to-boundary ratio:

$$S = \frac{P}{2r\pi} \quad r = \sqrt{\frac{A}{\pi}} \tag{8.4}$$

where P is the perimeter of polygon, A is the area of polygon and r is the radius of circle with the same surface area (Hole, 1978). A value of S close to 1 means that a polygon is rather compact and simple, while higher values describe narrow and long polygons. A higher complexity index usually relates to a more detailed delineation, which often means higher positional accuracy and larger effective scale (D'Avelo & McLeese, 1998).

8.2.2 Map legends

Accuracy of map legend was assessed using 114 independent sample observations distributed in ten SMU in three control survey areas with two tests: a binomial test for any of the named soils and a multinomial test for the composition of compound map units, which is based on the confusion matrix. These tests were applied at three levels of detail: strict (exact correspondence between named soil and field observation), similar (grouping similar soils based on surveyor opinion), and higher taxonomic (at the level of soil type). The binomial test simply scores a success if any of the named soils is encountered in the field; the estimate of success is the proportion of successes p, with variance for n samples of $\frac{p \cdot (1-p)}{n}$ (Steel & Torrie, 1980, 3.3). The multinomial test is the Pearson's χ^2 statistic, calculated for each SMU over the cells with non-zero expectation:

$$\chi^2 = \sum_{i,j} \frac{\left(y_{ij} - r_{ij} \cdot y_{i+}\right)^2}{r_{ij} \cdot y_{i+}} \quad r_{ij} > 0 \quad df = k \tag{8.5}$$

where y_{ij} is the number of observations in row i, column j, i.e. mapped in legend category i and observed in validation class j, y_{i+} is the total number of observations for legend category i, r_{ij} is the proportion of legend category i that should be in validation class j and df are degrees of freedom based on k comparisons. Note that since observations may not correspond to any named class, the degrees of freedom are equal to the number of classes. The probability of the observed χ^2 shows the statistical significance of the difference between the mapped and observed composition of units, and also can be used to assess the relative accuracy of SMUs. The sum of χ^2 for all tested SMU gives the overall accuracy of compound map units.

In each of the three control areas, we first made 20 mini-pits at randomly selected locations to determine soil types and train our mental model of the soil catenas. We then made two transect studies by taking 20 augerings at approximately equal distances (200 m) in the direction of the most contrasting relief change. The navigation to selected points (mini-pits) and transect studies were operationalized with the help of the mobile GIS system with an ortho-photo in the background. Finally, the set of 120 point observations was used to calculate a confusion matrix and asses the thematic accuracy of map legends. There were in total 18 SMUs in the control survey areas, from which five units (covering 8.5% of the total survey area) had fewer than five observations and were excluded from further calculations. An additional three SMUs were not covered by the random sampling and were also excluded. Finally, 10 SMUs, covering 85% of the total survey area, were evaluated using 114 observations and 39 soil taxa in total (Table 8.2). In addition, the same test was done with the generalized taxa, i.e. at the level of sub-type. In this case, there were 18 soil taxa.

Table 8.2: Mapping units inspected for the accuracy of legend, their composition, size and number of observed points per SMU.

CODE	Map sheet	no.	Original mapping unit composition	Type	Area (ha)	Area (%)	Observed points
SMU1	Osijek 2	7	Cambisols, eutric, typical - Vitisols - Regosols - Chernozems on loess (40:30:20:10)	Association	1036	23.3	29
SMU2	Osijek 2	16	Amphigleys and Hypogleys, mineral, partly hydromeliorated (60:40)	Association	153	3.4	5
SMU3	Samobor 3	9	Dystric Cambisols, illuviated, deep - Luvisols (50:50)	Association	231	5.2	11
SMU4	Samobor 3	20	Alluvial soils, calcareous, gleyic - Humofluvisols, gleyic, non-carbonatic (60:40)	Association	235	5.3	8
SMU5	Samobor 3	23	Pseudogleys, on sloping terrains	Consociation	158	3.5	5
SMU6	Samobor 3	22	Pseudogleys on level terrains - Pseudogley-gley soils (90:10)	Consociation	282	6.3	7
SMU7	Samobor 3	26	Amphigleys, mineral, non-calcareous - Pseudogleys (80:20)	Consociation	344	7.7	8
SMU8	Sibenik 4	8	Terra rossa, typical, shallow, clayey - Calcocambisols, typical, shallow, clayey - Rendzinas, on calcitic dolomite - Calcocambisols, colluvial (20:40:30:10)	Complex	190	4.3	6
SMU9	Sibenik 4	9	Vitosols (on terra rosa and cambisol) - Terra rosa, typical, shallow and colluvial - Calcocambisols, typical, shallow and colluvial (60:20:20)	Complex	850	19.1	20
SMU10	Sibenik 4	16	Vitosols, from terra rossa - Terra rosa, luvic, deep - Terra rosa, colluvial (60:20:20)	Complex	305	6.9	15
			Urban areas		108	2.4	
			Water bodies		86	1.9	
			Total		3977	89.5	114

8.2.3 Soil boundaries

In the case of BSMC, the soil boundaries were delineated manually following the concept of free survey (White, 1997), i.e. by using an irreproducible method. The boundaries were not explicitly drawn using physiographic or geomorphic principles, so that it is not strictly an error if a boundary does not correspond with topography. However, by careful study of the legend and comparison of the definitions of adjacent units, it is possible to infer what topographic features should have been followed. A simple example is where two polygons with contrasting soil types significantly

differ in slope class and there is clear topographic break. These boundaries are often referred to as the primary boundaries or master lines (Buringh, 1960). Although the soil boundaries are traditionally placed subjectively and will always differ, the primary boundaries, drawn independently by several surveyors, should usually match within the soil survey standards (Bie & Beckett, 1973).

To compare the primary soil boundaries we selected six adjacent SMU pairs in three control areas. The procedure can be summarized as follows:

1. make control survey map using stereoscopic photo-interpretation and validate the boundaries on the field;
2. digitise the boundaries from the control and original surveys;
3. identify the most contrasting adjacent units based on topographic or geomorphic features, e.g. slope breaks, change of general landscape; and
4. calculate the deviation between the two lines.

This is done by estimating the area of disagreement (AD), which is the area of polygon produced as the intersection between the original and control survey. The lower is the AD, the better is the positional accuracy of the inspected boundaries. The positional accuracy or the mean absolute error (ME) is then half of the AD width:

$$ME = \frac{AD}{2 \cdot \frac{l+l'}{2}} = \frac{AD}{l+l'} \tag{8.6}$$

where l is the boundary length of the mapped delineations, and l' is the boundary length of the control delineation.

8.2.4 Profile observations

Ten profiles from the original soil survey with positions shown on the soil maps were selected to assess how well the profile data corresponds to the control. We navigated to these points using the mobile GIS with the georeferenced original soil map in the background. It was always possible to be clearly within the square representing the original observation, as the profiles were shown by a 4 mm^2 square on the soil map, representing 100×100 m on the ground, although the surveyors actually could locate the profile within a 2 mm^2 and used the larger square for legibility only. The GPS reported estimated position errors (on the order of ±15 m) and the estimated geo-referencing error when digitising the topo sheets were an order of magnitude smaller. The location of the control observation within the 1 ha square was determined in the field by experienced surveyors, who were looking for the same

type of site as described in the original survey report. We then compared seven physical and chemical soil properties: sand, silt and clay content (%), pH (H2O), pH (MKCl), organic matter (%) and carbonates (%) in all horizons, by calculating the root mean square error (*RMSE*) and relative error:

$$\begin{aligned} RMSE &= \sqrt{\frac{\sum (x - x')^2}{n}} \\ RMSE_r &= \frac{RMSE}{R} \qquad\qquad R \approx 4 \cdot \sigma_x \end{aligned} \tag{8.7}$$

where x is the given value of the soil attribute and x' is the soil attribute measured in the control survey, n is the number of control measurements, R is the range of variation and σ_x is the standard deviation of population. In the case of normally-distributed variables, R can be approximated by four times σ_x (Ott & Longnecker, 2001, p. 92). The range can also be approximated by empirical estimates of the extremes likely to be found in a survey area. For example, in the study area we know from previous experience that clay content can range from 0 to 80%, therefore the range is 80%. The dimensionless $RMSE_r$ allows a comparison of accuracy for variables of different types and with different ranges of variation. Note that the soil samples were not taken at the same depths as in the case of the original survey, so that we needed to first estimate the values at the same depths as the original observations, by estimating values from a depth-vs-property graph.

8.2.5 Soil mapping units (SMUs)

A soil surveyor aims at delineating soil bodies in such a way that contrast between the adjacent SMUs is maximised, which reflects the idea of the maximum amount of information in a system (Finn, 1993). In the case of categorical data, separability of attribute values between the mapping units is the key measure of thematic map quality (Lilburne, 2001). In soil resource inventories, a standard method to assess efficiency of classification is to compare the within-class variances with the between-class and total variances (Webster & Olivier, 1990, pp. 63-70). In this study we used this method to evaluate two aspects of thematic quality of SMUs: (1) *thematic purity* or homogeneity of SMU composition and attribute values within the SMUs and (2) *thematic separability of geographically-adjacent SMUs*. These are different issues as the SMUs can show imprecise distributions of attributes, while at the same time the thematic contrast between the adjacent units can still be fairly high, and vice versa; this is related to the nature of boundaries (Lagacherie *et al.*, 1996).

These were assessed using the 2198 profile observations and the 1:300 K digital soil map of Croatia consisting of 65 SMUs. We first assessed the mapped homo-

geneity of the SMUs by calculating the proportion of the average dominant and associated soil types. This was done for both the 1:300 K Soil map and the three control survey areas. We then calculated mean value and standard deviations for each of the 65 SMUs of the 1:300 K digital Soil map and three soil parameters: clay content (%), pH (measured in H_2O) and organic matter (OM) in topsoil (%). The homogeneity within the SMUs was expressed by the relative standard deviation:

$$s_r\left(x,j\right) = \frac{s_{x,j}}{R} \qquad R \approx 4 \cdot \sigma_x \qquad j = 1,..,k \tag{8.8}$$

where $s_{x,j}$ is the standard deviation of a xth property inside the jth SMU, k is total number of units, and σ_x is the population variance. This is equivalent to the relative variance, commonly used in land resource inventories (Webster & Olivier, 1990, p. 67). A relative standard deviation higher than 25% means that we can measure almost any value of the property inside the tested SMU. In addition, we calculated the interclass correlations as:

$$r_i\left(x\right) = \frac{B - s_W^2}{B + (m-1) \cdot s_W^2} \tag{8.9}$$

where B is the mean square error between the classes, s_W^2 is the mean square error within the classes and m is the number of classes (Webster & Olivier, 1990, p. 67). A value close to 1 means that thematic purity of SMUs is maximum, while a fairly low value of r_i indicates that the variation within the SMUs is close to the total variation.

The thematic separability of adjacent SMUs was assessed using the *thematic overlap*. This was calculated as the average probability of thematic overlap between thematically and geographically adjacent SMUs. The last aspect quantifies the uncertainty in the attribute maps for site-specific decisions. The SMUs were first sorted according to their average value of a property. The probability of thematic overlap was calculated for each neighbouring SMUs as the average probability of overlap. If properties within units are normally distributed, the overlap can be calculated using a t-test to compare difference between two populations assuming equal variances (Ott & Longnecker, 2001), with null hypothesis that the two samples belong to the same population:

$$p\left(x_j \cap x_{j+1}\right) = 2 \cdot \left(1 - p_n\left[\frac{\bar{x}_j - \bar{x}_{j+1}}{\sqrt{s_{x_j}^2 + s_{x_{j+1}}^2}}\right]\right) \tag{8.10}$$

where $p\left(x_j \cap x_{j+1}\right)$ are normal probabilities of thematic overlap, p_n is the one-way normal cumulative probability and $\bar{x}_j$ is the average value of xth property in the jth SMU. Finally, the average probability of overlap $p_{\cap}(x)$ for all SMU for given property is:

$$p_{\cap}(x) = \frac{\sum_{j=1}^{m-1} p\left(x_j \cap x_{j+1}\right)}{m-1} \quad p_{\cap}\left(x\right) \in [0, 100\%] \tag{8.11}$$

where m is the total number of SMUs. The lower is the average probability of overlap, the more SMUs differ among each other, i.e. the more contrasting are the delineations and vice versa. If the $p_{\cap}(x) > 95\%$, the attributes between the adjacent SMUs do not differ significantly, i.e. we can measure similar properties inside adjacent SMUs. This means that the SMUs are over-specified and should be simplified. The neighbouring SMUs, i.e. geographically adjacent polygon pairs were derived in ILWIS using the neighbour polygon operation (Unit Geo Software Development, 2001). We then sorted the pairs of SMU polygons using the longest length of the neighbourhood boundary between the SMUs and calculated overlap using the same statistics as in Eq. (8.11).

8.2.6 Usage and usability

In a GIS, usability can be defined as a property of a given dataset that expresses (1) how well it helps users to arrive at a correct decision within their problem-solving context and (2) how easily can it be accessed and made ready for use. This aspect is subjective and difficult to quantify. We decided to assess usability using two measures: a) number of users compared to the potential number of users; b) degree of their satisfaction; We first made a small inventory of all existing users and then carried out unstructured interviews with several existing users in the land use planning offices of Osijek, Karlovac and Split cities. In addition we discussed the cartographic and GIS issues with cartographic departments in Osijek and Zagreb and a private company in Split. Finally, we discussed the usability problems with the original surveyors, to see the both sides of the story.

8.3 Results

8.3.1 Effective map scale

Table 8.3 shows various measures of map scale for six individual map sheets and the soil suitability map at 1:300 K. The IMR for the 1:50 K sheets ranged from 4 to

7, corresponding to an effective scale between 1:100 K and 1:175 K. Also the shape complexity index shows that the soil polygons in these six sheets are moderately simple, according to the classification of shape complexity suggested by Hole (1953). The map sheets however, can not be simply denominated to a smaller scale, since five of the six map sheets contains polygons smaller than the MLD (10 ha), ranging from 1.4 to 2.4 ha. These polygons should have been integrated with the neighboring SMUs.

Table 8.3: Assessing the effective scale: the average size delineation and inspection density on the test sheets.

		Sheet 1	Sheet 2	Sheet 3	Sheet 4	Sheet5	Sheet 6	Whole
	Statistical measure	Osijek 2	S. Brod 2	Samobor 3	Gospic 1	Split 3	Sibenik 4	Croatia
A_j	Total area of land (ha)	27 923	54 209	53 082	54 820	51 537	53 189	5 566 894
m	Total number of polygons	71	133	264	361	108	112	4312
	Minumum size delineation (ha)	2.3	2.4	1.6	1.4	1.5	10.7	2.0
	Maximum size delineation (ha)	3713.9	8708.9	6334.9	5122.1	6402.8	6761.6	85 443.3
ASD	Average size delineation (ha)	393.3	407.6	201.1	151.9	477.2	474.9	1291.0
	Std. of delineation size (ha)	190.5	215.1	288.7	286.5	178.0	188.6	305.2
NSN	Nominal scale number	50 K	50 K	50 K	50 K	50 K	50 K	300 K
IMR	Index of maximum reduction	6.3	6.4	4.5	3.9	6.9	6.9	1.9
ESN	Effective scale number	156 750	159 606	112 101	97 422	172 698	172 282	284 055
S	Average shape complexity index (±std.)	2.4±1.2	2.9±1.7	2.1±1.3	1.7±0.8	1.9±0.8	1.9±0.7	2.0±1.1
	Number of profiles	37.0	76.0	53.0	41.0	114.0	46.0	10 686
	Inspection density (per 1000 ha)	1.3	1.4	1.0	0.7	2.2	0.9	1.9±2.2

The IMR for the 1:300 K digital map was 1.9, which is almost ideal, showing that the effective scale corresponds to the given scale. However, polygons representing 1% of the total area were smaller than the MLD (360 ha). The shape complexity index shows that the polygons are moderately simple, which is desirable for this fairly small scale. The inspection density for the 1:50 K maps ranged from 0.7 to 2.2 profiles per 1000 ha, which is an order of magnitude smaller than the suggested

minimum of 50 (Avery, 1987, Table 2) for this scale, and two orders of magnitude smaller then the ideal inspection density of 4 observation per map cm^2, i.e. 160 per 1000 ha (Avery, 1987, Table 1). According to the original surveyors, the total inspection density is higher, as there were up to ten times more mini-pits that were mapped but not recorded on the original map sheets. However, we decided not to use this information, as there is no record on it. The inspection density of the 1:300 K digital map was 1.9 per 1000 ha, corresponding to 1.7 per map cm^2 (900 ha at this scale), about half the ideal. However, there was a very large variation among the map sheets making up the national map, so that some areas are well beneath the minimum density even at this scale.

8.3.2 Thematic accuracy of legends

The strict binomial test of the map legend showed that only half (57) of the 114 observations corresponded to one of the named components in their respective SMU; this is 0.5 $\pm$0.05 (see supplementary material for confusion matrix). Accuracy of individual SMU range widely, from 0.09 to 0.97; the upper limit of the 90% confidence interval showed that five of the ten SMU definitely failed the 80% purity test (Marsman & de Gruijter, 1986). Grouping similar soils revealed a much better success rate: 0.79 $\pm$0.04, which is very close to published standards for map unit purity. Here, only one of the ten SMU definitely failed the 80% purity test. At the soil type level, success was even greater: 0.87 $\pm$0.03. Soil types are defined pedogenetically and ignore many soil properties important for interpretations, so this high purity is not useful to most map users.

The binomial test does not measure whether the stated map unit composition is correct; for that the multivariate test was used. The strict test gave a summary χ^2 of 50.14 with 25 d.f. (p <0.001), showing that the legend does not accurately represent the composition of this set of compound map units. Most of the error was concentrated in three map units. In one case many shallow Lithosols on limestone were found in place of mapped shallow Calocambisols (according to CYS, see supplementary material for more details on CYS). In another case deep Terra Rossa were found instead of mapped moderately deep subtypes. In the third case, Luvisols were expected in half the map unit, but not encountered; the process of illuviation was not observed, as there was no significant textural change or visual evidence of clay movement. These errors did not disappear when soils were grouped by taxonomic type; χ^2 of 41.8 (p <0.004). However, grouping similar soils by major interpretations resulted in a more favourable evaluation: χ^2 of 23.26 (p <0.5), so that the legend can not be rejected. Note that the test is not strictly valid, since in one cell the expected frequency is less than one, and there are also a large proportion (15 of 25) of cells with fewer than five which means that the chi-square approximation is

marginally valid, but still indicative (Agresti, 1990).

8.3.3 Spatial accuracy of soil boundaries

The comparison between the soil boundaries in the original and control survey can be seen in Fig. 8.3. The original maps show in general less detail and follow the master lines only approximately. Only in the case of Kalinovac area was the level of detail more or less equal (Fig. 8.3c). The mean absolute error between the six evaluated master lines (Table 8.4), showed that boundaries deviate approximately ±40 m from the reference boundaries delineated in control surveys. From this number we computed the maximum location accuracy of 0.77 mm on the map. The observed accuracy is about three times worse then the typical map accuracy standard for this scale of 0.25 mm (Davies, 1981). Consequently, the effective map scale is approximately three time smaller or 1:150 K. Other, secondary boundaries show even higher discrepancy with the control surveys. Very often we could not conclude on which basis did the surveyor draw the boundaries. Some neighboring SMUs in the Gustirna and Popovac area, for example, had in legend the same soil types, usually only with different composition. This makes the soil boundaries, other than the master lines, even more relative.

Table 8.4: Spatial accuracy of the boundaries for six master lines.

Number	1	2	3	4	5	6	Average
Area of disagreement (ha)	67.4	25.8	29.9	15.4	44.5	12.7	
Summary length ($l + l'$)	10.5	6.6	11.2	5.4	9.5	5.2	
Mean error (± m)	64.2	39.1	26.7	28.5	46.8	24.4	38.3±15.3

8.3.4 Thematic accuracy of profile observations

The comparison of the ten detailed observations showed that the profile data generally corresponds to what was described on the field and measured in the laboratory. The results can be summarized as follows: a) the descriptive data, such as exposition, land cover, rock outcrops, coincided with what we found on the field in most cases; b) the soil types did not match the one found on the field in four cases[2]. For example, profile 442 was classified in the CYS as a Pseudogley, while the high ground water table showed that this is obviously a hypogleyic Eugley (Fig. 8.4a and b).

[2]See profile description data sheets in the supplementary materials.

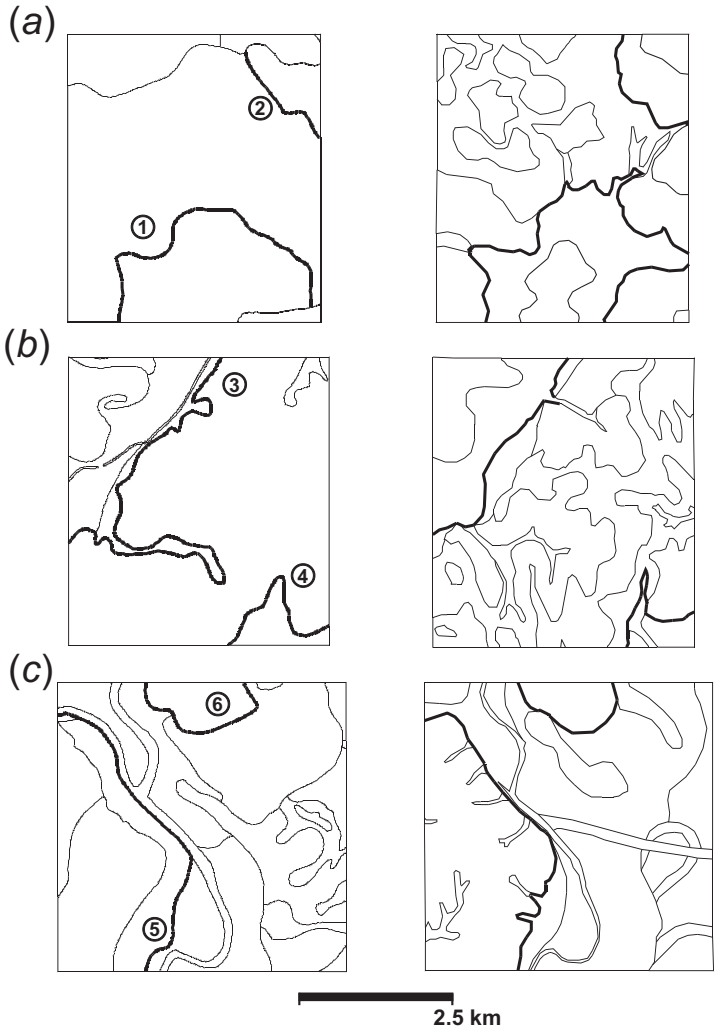

Figure 8.3: Comparison of the original soil boundaries (left) and control survey areas (right): Gustirna (a), Popovac (b) and Kalinovac (c). The master lines are bolded. Compare with the results in table 8.4.

Similarly, profile 76 classified (CYS) as a Calcocambisol on dolomite was described as a Luvisol on dolomite; profile 159 classified as typical (non-luvic) Terra rosa was described as the illuviated sub-type. The biggest discrepancy, however, was profile 13, classified (CYS) as a Calcomelanosol (soils of high-rainfall mountain areas, with a distinct dark epipedon). We were not able to find this type at the given location

nor anywhere in SMU. Instead, the map unit was dominated by Calcocambisols.

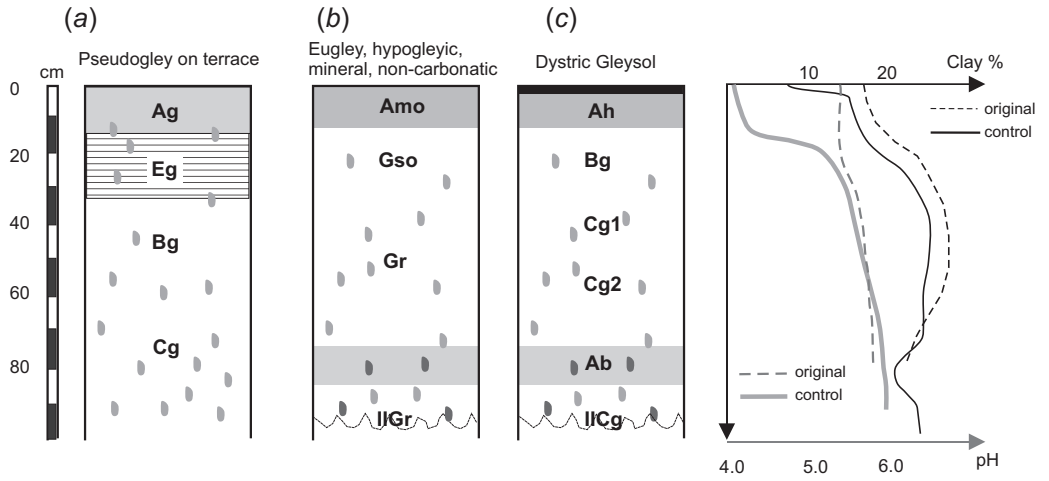

Figure 8.4: Comparison of the original (a) and control profile description using the same methodology (b) and FAO (1998) methodology (c). While the horizon designations do not coincide, the measured properties – clay content (%) and pH in H_2O (depth-vs-property graph) do.

Comparison of the lab data showed a general correspondence with what we described in the field, e.g. texture class, clay content and clay increase (Fig. 8.4c) with an average relative error of ±15.7% (Table 8.5). The highest accuracy of measuring the same property was for the organic matter (±8.7%), while the most inaccurate property was pH.

8.3.5 Homogeneity and thematic contrast between SMUs

The SMUs in the 1:300 K soil map consist of two or more soil types in most cases. We calculated average mapped proportion of the dominant soil type of about 56%, while in the control survey areas we calculated slightly higher composition purity of the SMUs (65%). In the case of the 1:300 K soil map, many of the original SMUs have been merged into bigger units, so that the average number of soil types per SMU is from 2 to 5. When analysed for the categorical type of SMU, based on the USDA Soil survey standards (Soil Survey Division Staff, 1993), we found that some 32% of total area was classified as probable consociations with $> 75\%$ of the dominant soil type, 36% as associations and 32% as complexes. All this indicates that the SMUs have been described as compound and often heterogeneous.

Table 8.5: Summary results on the thematic accuracy of point data with comparison to the expected physical range of variables in Croatia.

Observed parameter	Estimated range of variation (R)	*RMSE* estimated at control points	Number of samples	Precision $RMSE/R$
Sand %	0–80%	12.8%	21	±16.0%
Fine silt %	0–80%	12.3%	21	±15.4%
Clay content (%)	0–80%	12.0%	21	±15.0%
pH (H_2O)	4.5–8.5	0.82	18	±20.5%
pH (MKCl)	3.5–7.5	0.66	18	±16.5%
Organic matter (%)	0–30	2.6%	14	±8.7%
Carbonates (%)	0–30	4.7%	9	±13.8%
Overall				±15.7%

Analysis of the thematic contrast of soil mapping units for clay content (%), pH and OM (%) in the topsoil showed that the SMUs from the 1:300 K Soil map are fairly heterogeneous within the SMUs (Table 8.6). The SMU showed high relative variation of 17% in average for these three variables. These properties also showed fairly low interclass correlations ranging from 0.14 to 0.34. Nevertheless, analysis of variance showed that there are highly significant ($p > 0.01$) differences between the 65 SMUs.

The average probability of thematic overlap showed that the adjacent SMUs after sorting in feature space did not show significant difference ($p_{\cap}(x) > 95\%$). This means that the overall contrast of the constructed map is low. The geographically-adjacent SMUs showed an average 66% overlap of the variation between neighbouring SMUs, which is more satisfactory. These three summary values show that the effective contrast of the 1:300 K soil map is relatively low, which does not have to be due to the poor delineations or legends, but is the effect of the relatively general scale.

The derived attribute map for clay content and relative variation within the units is shown in Fig. 8.5 a and b. In this case, the areas of higher relative variation

Table 8.6: Summary results for the thematic contrast of SMUs for 1:300 K soil map based on the clay content, pH and OM.

	Range of variation calculated from 2198 profiles	Average relative variation	Fisher's test ratio[a]	Interclass correlation	Average probability of thematic overlap	
Property	$\bar{\mu} \pm \sigma(R)$	$\bar{s}_r(x)$	F	$r_i(x)$	$p_\cap(x)$ (sorted polygons)	$p_\cap(x)$ (neighbouring polygons)
Clay content (%)	28.1 ±13.3(53.2)	±19.5%	14.9**	0.19	96.5%	67.4%
pH in H_2O	6.4 ±1.19 (4.8)	±17.1%	31.2**	0.34	95.5%	65.0%
OM %	8.4 ±8.4 (33.6)	±13.8%	10.5**	0.14	96.3%	65.4%
Overall		±17%			96%	66%

[a] ** – Significant at 0.01 level.

(e.g. > 25%), indicated as darker, should not be used to produce attribute maps. Fig. 8.5d shows two examples of high and low thematic overlap for neighboring polygons.

8.3.6 Usability issues

From our interviews we concluded that the most of the government departments involved in physical planning, agricultural extension and environmental protection are not using existing soil data to its fullest potential. The main usability problem seems to be lack of interpreted information at an appropriate scale. Agricultural extensionists, who should use this data routinely, use it rarely and only qualitatively, despite the large amount of analytical data associated with each soil subtype. For use at detailed scales, the problem appears to be two-fold. First, most map units are compound and map users (planners or extensionists) are rarely capable of finding the components in the field by their landscape relations. In some reports the map unit's heterogeneity (degree of internal contrast) is given, so that for more homogeneous units the dominant component can be used for planning. Second, there are no interpretations (land evaluations) of the predicted performance of various Land Utilization Types, or even crops, on each map unit. Such tables can be used directly in decision making. In the current situation, only those users who are trained in the use of primary soils data can make any sense of the maps.

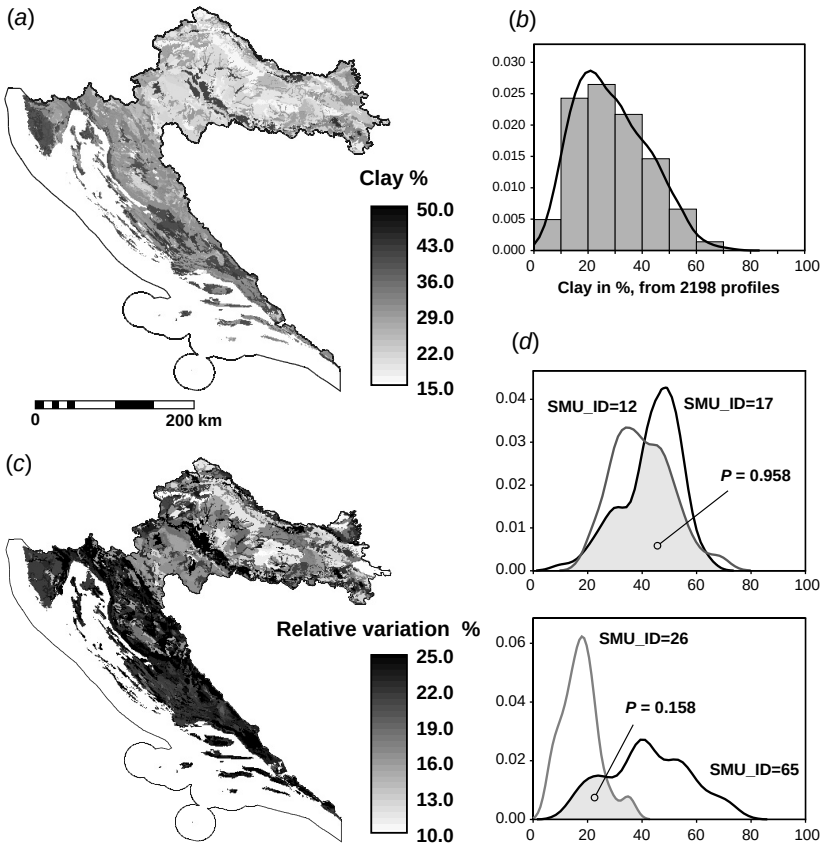

Figure 8.5: Attribute map of clay content (%) (a), density histogram (b), map of relative variation within SMUs (c) and two examples of high and low thematic overlap between the neighbouring SMU pairs (d). P is the probability of thematic overlap ($p_{\cap}(x)$)

At smaller scales, the principal users of soil data in Croatia are:

- Croatian waters, a government agency;
- County planning offices (eight of twenty-one in total); and
- Departments of the Ministry of Nature Protection and Physical Planning.

Together these are only about one quarter of all users that might benefit from the soil geoinformation. Their satisfaction with current products depends mainly on how well their professional background allows them to make their own interpretations from the supplied purely pedologic information and the level of detail needed.

Preconceptions from the user's own field of expertise colour their perceptions of the soils data. For example, land surveyors and GIS professionals were put off by the fact that boundaries in adjacent map sheets do not match. This certainly reduces their confidence in the product.

8.4 Conclusions and discussion

8.4.1 Lineage

The analysis of adequacy and data quality was difficult as the data has not been fully integrated and the detailed metadata is missing. In some cases we were able to understand the exact methodological steps only from the conversation with the surveyors involved. A good example is the problem of changing from the old to the new coordinate system. The soil map of Croatia was produced using the military 1:50 K topo-maps, which were based on the field measurements from the 1930's. This map needed to be transformed to a new system of ground control points. Although the coordinates differ by only a few hundreds of meters on average, transformation from old to new co-ordinate system is rather complex (Radošević, 1979). This means that without the help of land surveyors, the old soil maps can not be accurately integrated into the GIS. We have first overlaid the digitised soil boundaries over the georeferenced original maps and concluded that the discrepancy can be estimated with a systematic shift. We then measured this shift in six points from different parts of the country, which gave us a coordinate correction parameters (dX=135 m, dY=-65 m). This example clearly shows how the lack of metadata can lead to inaccuracies in the GIS that are significant but not easily solvable.

8.4.2 Effective scale

Although the soil maps and survey reports from the national inventory seem to be of high quality and with lot of detail, it is clear that the aimed scale of 1:50 K was not achieved in more than 95% of sampled map sheets. We have estimated the effective scale to be about (Fig. 8.6):

- 1:150 K based on the ASD;
- 1:250 K according to the inspection density;
- 1:150 K according to the spatial accuracy of soil boundaries

Similarly, the shape complexity index showed rather simple geometries of delineations. Thus, the National soil inventory in Croatia can be classified as the small scale or medium intensity survey (Avery, 1987). This means that this is a national

or regional land inventory, which is applicable for evaluation of extensive uses and only general land use planning purposes and not for agricultural extension, civil engineering projects or county level land use planning.

Logical consistency and completeness Considering the consistency and completeness, the National soil inventory shows number of discrepancies. Fig. 8.7a shows how the average density of profiles per 1000 ha varies per map sheet, which indicates rather different sampling densities in two parts of the country. The right part of the same figure shows the printed and not finished reports, i.e. reports that are still kept only as working materials, such as sketches, drawings, handwritings etc. (Fig. 8.7b). These two elements of data quality indicate that the project was not finalised and it requires further comparison of different sub-projects.

Accuracy of map legends The validation of accuracy of legend and its composition showed that there is a significant difference between the original legend and validation. Both for the lowest level (form) and higher level of taxa (subtype), the χ^2 test showed that the composition of the map units was not accurately represented. The BSMC even failed a strict binomial test of accuracy (not considering composition), but was satisfactory when soils were grouped by higher taxa or by similarity. The number of profile observations (114) was not large enough to ensure reliable statistics for ten map units with around 40 taxonomic classes. We estimate that it would take few hundreds random observations to accurately evaluate all 65 SMUs described in Croatia, therefore, these results should be taken with caution. However, we can in general claim that the overall thematic accuracy of legend is lower than mapped, except when soils are grouped by interpretive similarity. There are probably several reasons for this:

- the percentage of soil taxa in the original legends has been subjectively approximated without field sampling such as transects;
- in some cases surveyors placed soil classes that did not appear in the actual soilscape;
- this may be a gross error or perhaps a case where surveyors mapped the expected class without field confirmation;
- soil classes are not defined with exact criteria, which allows several interpretations — these can be grouped by similarity, which is one reason why the relaxed test gives better results.

The last reason makes it especially hard to produce an accurate legend or be certain about the validation test. For example, we noticed in several locations that profiles classified as Dystric Cambisols, illuviated, could have been equally well classified as Luvisols and the legend test would have been less negative. In many

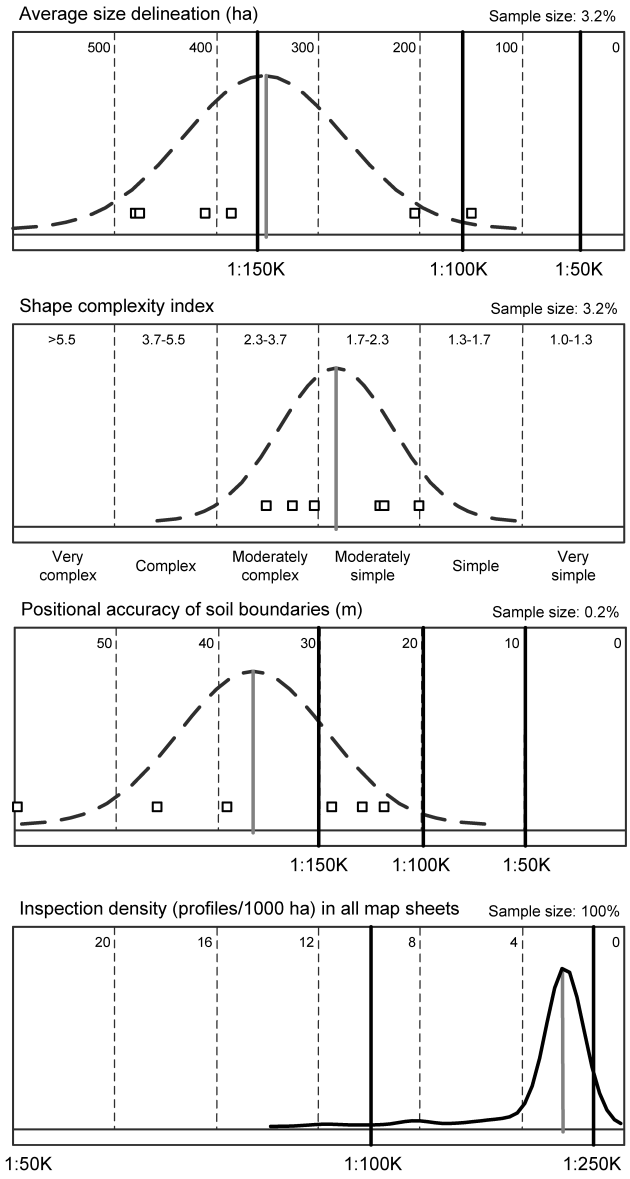

Figure 8.6: Summary adequacy aspects displayed graphically. The normal distribution curves indicate sampled average value and standard deviation of the adequacy measure, while the sample measurements are indicated with boxes.

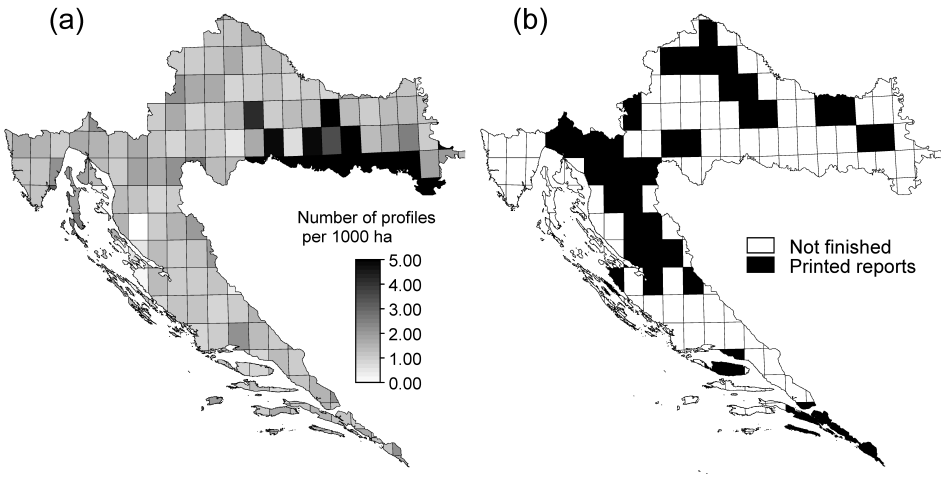

Figure 8.7: Example of methodological inconsistency (a) and incompleteness (b) of the National soil inventory in Croatia.

cases the CYS allows two or more interpretations, which can even be equally good and in fact are grouped in interpretive classifications. For example, less-developed soils on sloping terrains in Popovac area have been classified as Regosols on loess. The same locations could have also been classified as Colluvial soils. All this leads to a conclusion that a) estimated percentages of soil types can not be used for detailed planning and b) the CYS should be either improved or replaced with an analytical system such as the World Reference Base for soil classification (WRB) (FAO, 1998).

8.4.3 Thematic contrast and accuracy of profile observations

A statistical analysis for the 2198 profiles when compared in 65 SMUs using three chemical/physical soil properties showed that the SMUs are rather heterogeneous with small overall contrast of the attribute maps. It would be interesting for a future study to apply these tests to a larger scale survey and compare how heterogeneous will be the delineations between each other and eventually provide some standards.

Similarly, we concluded that precision of the soil attributes attached to soil profiles is not realistic. In this case, we estimated a relative error of $\pm 15.7\%$, which was fairly close to the lowest efficient limit (25%). A user of a soil map, not necessarily introduced in the field of soil science, would expect to find the same soil types and measure similar soil attributes at the given point locations where the original profile observations were made. This is typically not the case in soil survey, where the

local field variation is high. Moreover, a comparative work at ISRIC (van Reeuwijk, 1984), showed that, even the *RMSE* in laboratory results from the same soil samples could easily exceed ±11% for clay content, ±0.2 units for pH and ±20% for CEC. All this leads to conclusion that one can not expect to measure the same soil properties on the field as described in the database and therefore a relative error of ±10% should always be expected. The problem remains of how to communicate this to the survey user.

8.4.4 Usability problems

The reasons why the soil data in Croatia is poorly used and often only quantitatively are more complex than an adequacy analysis can show. Moreover there are no simple measures to quantify the usability, except the number of users and user's satisfaction. Finally, we have identified the three main causes, which inhibit the growth of applications:

1. Missing legislative framework — the accessibility and distribution policy for the soil data in Croatia is not clear neither transparent. Most often, the choice of using or ignoring information on soils is left to the decision makers, i.e. local government or land use planning office leaders. Professional soil scientists or soil surveyors are usually not assigned to land-use planning teams, which are in Croatia mainly guided by architects.

2. Legend and terminological concepts not understandable by end users — especially the soil classification system and the concept of soil types and soil bodies used in Croatia are unknown to most of users and therefore unpopular. The users in principle ask for interpretations, i.e. maps of soil attributes and land characteristics, while the soil maps typically shows only the distribution of soil types with accompanied legend.

3. Inventory products not adjusted to users — a common aphorism used for soil maps *"the soil scientists make maps for them selves"*, applies in Croatia also.

Chapter 9

Conclusions and Discussion

"Scientific progress is like climbing a mountain — you climb and you climb and eventually you have a better overview of the things than the people at lower elevations."

[by A. Stein, during a mid-term meeting]

9.1 Conclusions

The general conclusion of this research is that combined pedometric techniques enhance the practice of soil mapping, making soil maps more objective, detailed and more compatible for integration with other environmental geo-data. Examples from this thesis show that we no longer, *hic et nunc*, need to use the concept of soil mapping units or use double-crisp soil maps. On the other hand, instead of abandoning photo-interpretation, soil classification or empirical knowledge on soils, these methods can be successfully integrated with pedometric techniques. Other specific conclusions relate to the research questions posed in the Chapter 1:

SAMPLING: The allocation of points in both feature and geographical space plays an important role for the efficiency of prediction. Information on feature space and the spatial dependence structure of the predictors can be used prior to the actual soil data collection to design the sampling. A sampling design with equal spreading in both geographical and feature space should be used to optimise the accuracy of prediction for general purpose surveys. If a higher number of auxiliary maps is used for prediction, principal component analysis can be used to reduce multicollinearity and to produce orthogonal variables.

PRE-PROCESSING: Inaccuracies and artefacts in auxiliary maps, especially in terrain parameters, can greatly affect soil mapping. Although these can be hard to detect in the final prediction maps, the quality of DEM-derived maps will be low. Moreover, local errors will commonly propagate to their neighbours and finally occupy larger area than in the original auxiliary maps. Artefacts and outliers in terrain parameters can be systematically reduced by improving the plausibility of a DEM, filtering outliers and by averaging multiple realisations.

PHOTO-INTERPRETATION: The terrain parameters derived from a DEM show strong correspondence with the photo-interpretation units. Hence, supervised classification of terrain parameters can be used to replace the photo-interpretation. However, this correspondence is strong only if the changes in the relief are distinct, which is often not the case in plain areas. This is because the quality of terrain parameters is typically lower in the areas of low relief. The second problem is that some geomorphic features are shaped either by irregular fluvial, geological or climatic processes such as over-flooding, landslides, wind erosion, faulting etc. In this study, the problematic landforms were levees, overflow channels, scarps and point-bar complexes. Such features can still be more accurately mapped by using photo-interpretation (image contrast, texture and pattern), geological data and field checking.

INTERPOLATION: Regression-kriging based on a mixed model of spatial variation is a suitable generic spatial prediction model that can replace plain environmental correlation or ordinary kriging and the use of a continuous or discrete model of spatial variation. The logit transformation of the response variable is a useful for the prediction, especially if the response variable has a skewed distributions. This transformation prevents predictions outside physical limits and, in many cases, ensures the normality of residuals. In addition, transformation of predictors to independent factors typically helps in selecting a smaller (optimal) subset of predictors during the step-wise regression.

VISUALISATION: Multiple membership maps produced by continuous classification can be displayed using the HSI colour model, colour mixing and circular legend. This tool is not only useful to visualise the membership maps (soil types) at once but can also be used as a tool for generalization and visualisation of uncertainty. The circular HSI legend is limited to seven generic colours, hence, the similar soil types will be given a similar colour. Whiteness gives a better impression of classification uncertainty than if only saturation is used. The derived uncertainty can be used to depict transitions and areas of higher confusion. This visualisation tool is applicable to any natural resource inventories where categorical data are used (vegetation types, geoforms etc.).

ORGANIZATION: The key advantages of a hybrid grid-based SIS compared with the conventional double-crisp model are that it is in general more detailed and more accurate. Moreover, it employs auxiliary information in a more systematic manner, it is easier to both manipulate and to (dis)aggregate. It can adopt existing conventional soil databases as it represents a generalization of the conventional model. Selection of the suitable grid size can be related to the cartographic scale, i.e. maximum location accuracy, to minimum legible delineation area, or to the environmental heterogeneity of an area and size of the management units. The described SIS is potentially more attractive to external soil users because it is easier to manipulate and because it aims to map soil types and soil variables at a high level of detail.

QUALITY CONTROL: The true quality of conventional soil resource inventories can differ significantly from the prescribed soil survey standards. This happens especially when subjective methods are used to locate soil samples, delineate soil bodies, classify soils or make legends for mapping units. In the case of the National Soil Inventory in Croatia, the effective scale and usability of the soil data, especially for soil boundaries and legends, has been over-optimistic. Assessment of the adequacy and quality measures has shown that the Croatian data set is of little use for county level and regional level

land use planning. The soil mapping units are of relatively low thematic purity and show high thematic overlap between the adjacent units. Moreover, the concept of soil types and mapping units has remained un-popular among non-specialists for the last 30 years, which also constrains its usage. For future projects, the control of data quality and usability needs to be taken into account as an important step in the production of soil geoinformation.

9.2 Reliable modelling of soil variation

The first step towards reliable modelling of soil variation is to understand the sources and the types of errors inherent in them. The sources of error in (spatial) soil data can be grouped into two main groups: measurement errors and natural spatial variation (Burrough & McDonnell, 1998). Measurement errors typically occur during the positioning in the field, during sampling or the laboratory analysis. These errors should ideally be minimized, because they are not of primary concern for pedometricians. The second step towards reliable modelling is to account for all aspects of natural variation. Although spatial prediction of soil variables is primarily concerned with geographical variability, there are also other aspects of natural soil variation that are often overlooked by many pedometricians. This problem is nicely emphasized by Florinsky *et al.* (2002), for example. Also in this research, I have (unintentionally) focused on the geographical aspect of soil variability only... to discover finally that the mixing of lab data from different seasons, depths and with different support sizes in general means lower predictive power and problems in fully interpreting the results.

In some cases (see, for example, predicting pH in chapter 5), ignoring other sources of soil variability means that no existing interpolation method or auxiliary map can 'save' its mapping. Similarly, even if the data are fitted successfully, e.g. by using large list of auxiliary maps, prediction is not necessarily accurate. This is because a precise prediction does not have to be an accurate one. In fact, an imprecise estimation can sometimes be more accurate than a very precise one (Foote & Huebner, 1995). This thesis certainly does not offer complete instructions for a reliable modelling of soil variation. Some of the problems listed above will still make it difficult to produce accurate, reliable and detailed soil maps. In order to achieve this, one needs to keep in mind the four aspects of natural variation: geographical, depth, temporal and scale. Bellow is an overview of the main concepts and problems associated with each of these, illustrated using some familiar examples (Fig. 9.1).

Geographical variation (2D) Geographical variation is modelled using either a continuous, discrete or mixed model. The results of interpolation are either visualized as 2D maps (Fig. 9.1a) or cross-sections. Some soil variables, such

as thickness of horizons, the occurrence of diagnostic properties or soil types, do not have a third dimension, i.e. they refer to the upper (two) meters of the surface mantle.

Depth — internal or vertical variation (3D) Many soil variables also vary with depth. In many cases, the measured difference between the values is higher at a depth differing by a few centimetres than at geographical distance of few meters (Webster, 2000). Transition between different soil horizons can also be both gradual and abrupt (Fig. 9.1b), which requires a double-mixed model of soil variation for 3D interpolation. Some authors suggest the use of cumulative values on volume (areal) basis to simplify mapping of the 3D variables. For example, McKenzie & Ryan (1999) produced maps of total phosphorus and carbon estimated in the upper 1 m of soil and expressed in tons per hectare, which then simplifies production and retrieval.

Temporal variation Chemical soil variables, such as pH, CEC, nutrients (Fig. 9.1c), water-saturation levels and water content, measured at the same location, can vary over a few years, within a single season or even over a few days (Heuvelink & Webster, 2001). Temporal variability makes the prediction process especially complex and expensive. Maps of soil variables produced for two different time references can differ significantly. This means that a produced map is valid for a certain period (or moment) of time only. However, in many case the seasonal periodicity of soil variables is regular so that prediction does not necessarily require new samples. Another solution is to predict the changes in soil properties by measuring controlling factors. For example, accurate multi-temporal maps of groundwater depth can be used to account for the seasonal variation of chemical soil properties.

Support size Support size is the discretisation level of a geographical surface and is related to the concept of grid size in a grid-based GIS. In the case of regression-kriging, there are two support sizes: the size of the blocks of land sampled, and grid resolution of the auxiliary maps. Soil samples are typically collected as point samples. The support size of the auxiliary maps is commonly much larger than the actual blocks of land sampled, e.g. auxiliary variables are in general averaged (smoothed), while the soil data can present local (micro) features. As a result, the correlation between the auxiliary maps and soil measurements is often low or insignificant. There are two solutions to this problem: a) to up-scale the auxiliary maps or work with super-high resolution/detail data (e.g. IKONOS images of 1 m resolution) or b) to average bulk or composite samples within the regular blocks of land (Patil, 2002). The first approach is more attractive for the efficiency of prediction, but at the cost of more

processing power and storage. The second solution will only result in a better fit, whereas the efficiency of prediction, validated using point observations, may not change significantly.

From the discussion above, the following seven recommendations are taken as essential to the reliable modelling of soil variation[1]:

1. *Spatial soil variability is commonly a result of complex soil processes working at the same time and over long periods of time, rather than an effect of a single realization of a single soil-forming factor.*

2. *The mixed model of spatial variation is better suited than either continuous or discrete models to deal with both the geographical and internal variability of soils.*

3. *Functional relations between soil variables and predictors are in general unknown and the correlation coefficients can differ for different study areas, different seasons and different scales. However, in many cases (see for example models for organic matter in chapter 2 and solum thickness in chapter 3), relations with the environmental predictors often reflect causal linkage: deeper and more developed soils occur at places of higher potential accumulation and lower slope; soils with more organic matter can be found where the climate is cooler and wetter; textural fractions follow the type of parent material etc.*

4. *If the focus of prediction modelling is solely the geographical component (2D), then the samples need to be taken under fixed conditions: same season, same depths, same blocks of land. This also means that each 2D map of a soil variable should always indicate a time reference, referred depth and the sample (support) size.*

5. *The grid size in a grid-based SIS should ideally be a few metres (i.e. the size of a pedon). The smaller support sizes mean more detailed terrain parameters and remote sensing images, which will then also be reflected in the efficiency of prediction.*

6. *For reliable prediction modelling, it is important to analyze the uncertainty 'budget' of the prediction model, i.e. explain the composition of the total variability of the response variable. In the case of regression-kriging, this will show how much of the error due to the measurement error, how much has been*

[1]Note that some of these were not actually proven: these are empirical recommendations, which will need to be refined and tested in the future real case studies.

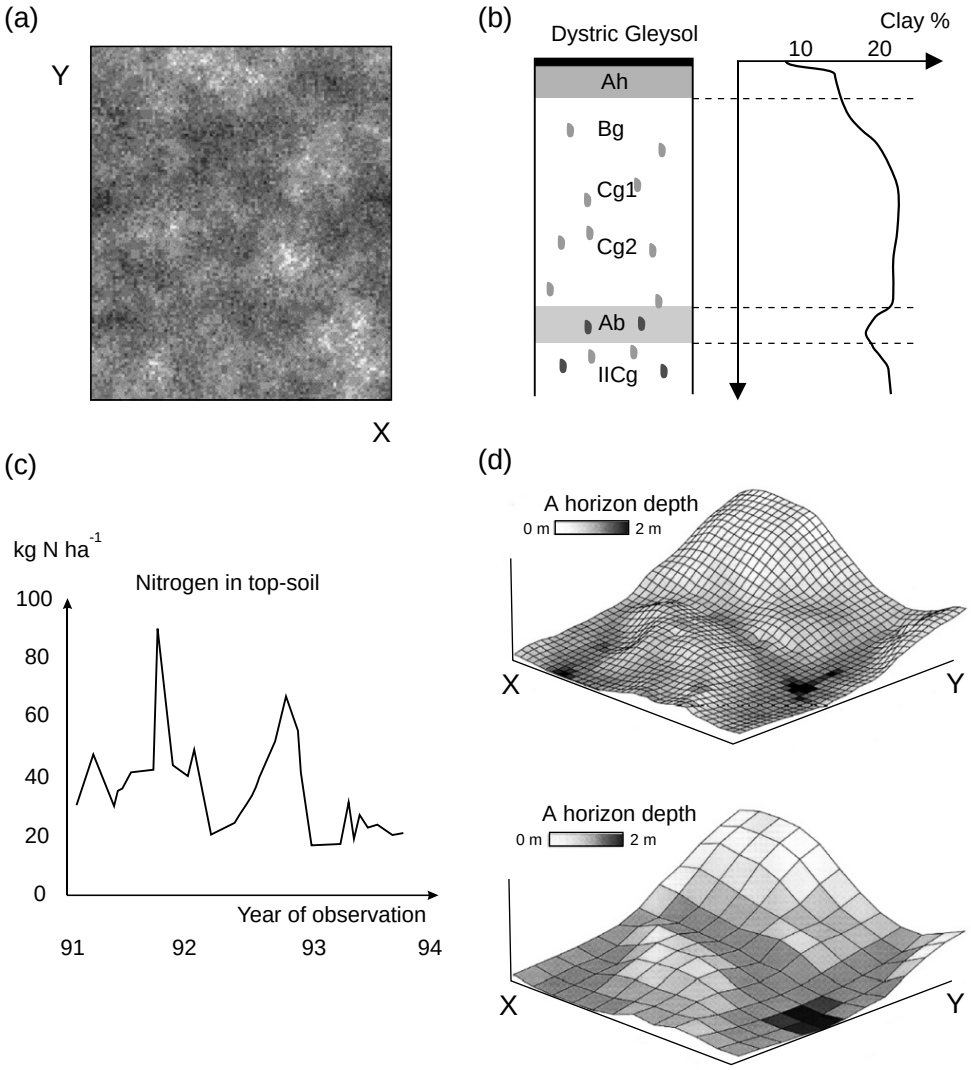

Figure 9.1: Some considerations related to the four aspects of soil variability: (a) Geographical variability — perception of soil variability as a realisation of stationary Gaussian random process visualised as a 2D map — example from Heuvelink & Webster (2001); (b) Stratigraphy or vertical variability — change between the soil horizons can be both continuous and abrupt; (c) Temporal variability — change in Nitrogen content measured at a same location in periods of three months — example from Stenger (1996); (d) Support (grid) size — predictions of the same variable from fine and coarse grid data might give different pictures — example from Thompson *et al.* (2001).

accounted for by the predictors, how much by the kriging and how much is uncorrelated noise. Systematic assessment of the uncertainty budget prevents over-fitting and making biased or over-optimistic predictions.

7. *True uncertainty of prediction can only be assessed using the spatially independent but representative validation set.*

9.3 Pedometric demand-driven soil mapping

It is not only the development of a reliable modelling of spatial distribution of soils that can guarantee a high quality and popular product. Soil mappers will increasingly need to find a balance between the availability of funds, models, tools and users' demands. In fact, the key challenge for future soil mapping projects will be to identify and meet SIS customers' needs (Indorante *et al.*, 1996). This book can be considered as an attempt to adjust pedometric techniques for operational surveys. It can serve as a methodological guide for the production of grid-based SISs at various resolutions. The suggested outline of a pedometric demand-driven grid-based SIS is illustrated in Fig. 9.2. Five steps are common to all scales/intensities: the evaluation of existing data and preparation, design, data organization, soil data collection and the analysis and production of geoinformation. This means that intensive interaction between the users and soil geo-data producers is especially important during the evaluation of quality and usability of existing data, the selection of the key land characteristics and the preparation of the predictors and sampling plan. Note that the circular structure not only suggests re-cycling of the soil data, but also implies a need to periodically update soil geo-information with new interviews, surveys, auxiliary maps etc.

9.4 Further research

Several subjects that have not been studied in depth may be of significance for the future research. In addition, there are number of remaining questions and problems in the area of pedometric mapping that need to be tackled. I have tried to group these to produce a list of the most challenging topics, each of which is probably worth pursuing as a PhD research project.

Sampling optimisation in geo- and feature space Chapter 2 gives an extensive introduction to sampling optimisation, however, many problems are left unsolved. Moreover, it may strike a reader that the there is a significant difference between the theoretical design (Fig. 2.3) and what was really achieved for the data set (Fig. 2.9). Also the number of points used (25) is probably

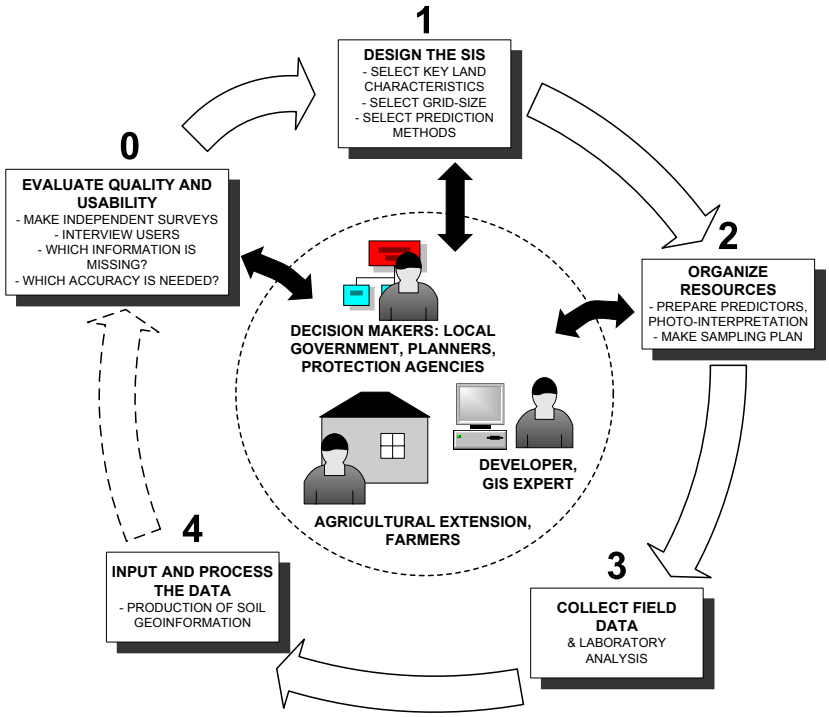

Figure 9.2: Schematic flowchart of a demand-driven production of soil geoinformation.

too low to make some serious conclusions. Hence, more systematic research is needed to clarify these differences and develop a flexible framework for sampling optimisation. Is there any advantage of using the spatial decomposition of principal components? How to integrate geostatistical optimisation techniques and allocation in multivariate feature space? Which combined criteria should be used for this purpose and how will this affect the efficiency of prediction?

Integration of GIS and (geo)statistics As already introduced in chapters 1, 2 and 5, so called CLOPRT techniques and geostatistics are two separate paths to spatial interpolation. Although the theory (universal kriging) for their combination was described by Matheron (1969) some 30 years ago and although there have been many case studies, the number of user-friendly packages for 'universal interpolation' is still low (Goovaerts, 1999b). One package that fully

supports interpolation of point data by employing both the correlation with auxiliary maps and spatial dependence is the GSTAT developed by Pebesma & Wesseling (1998). GSTAT is a General Public License package that can use different GIS extensions from Idrisi to ESRI and GRASS. However, even in GSTAT, a combination of generalized models and kriging cannot be fully employed nor automated. The generic framework for spatial prediction, as described in chapter 5, is a further step towards a full integration of GIS and kriging. The method has been described and tested, and now needs to be integrated within a GIS package. In the case of ILWIS, for example, this means that algorithms normally not available in ILWIS, such as step-wise regression, automated variogram modelling and others, need to be incorporated within the GIS package in order to make the framework operational. The framework can also be extended to the ML, REML or neural network systems. The remaining question is how will this bundle algorithm perform if used with large number of predictors and response variables? What are the limitations and can it be fully automated such that only minimum input is required from the user?

Soil-landscape genesis modelling Instead of merely fitting the soil profile data using some (geo) statistical technique, a more promising approach to the prediction of soil properties and soil behaviours is to make physically-sound quantitative models of soil-landscape genesis (Hoosbeek & Bryant, 1992). With the rapid development of GIS dynamic modelling, these virtual landscapes are often visualised as 3D animations (Mitasova *et al.*, 1997; Burrough & McDonnell, 1998). Minasny & McBratney (2001) have developed a rudimentary model of soil-landscape evolution. Königel (2002) has been working on the predictive model of geo-ecological evolution. Both groups demonstrated the power (and problems) of such systems at the World Congress of Soil Science in Bangkok. The impressions were generally positive, although there were some reluctant voices. The two soil genesis models mainly describe processes controlled by relief. These are then used to make inference about the chemical weathering or development of soil taxa. A future step will certainly be to include the other 'letters' from the `CLORPT` equation. One can imagine that the quantitative simulation of soil genesis might turn out to be as non-linear as with the long-term weather forecast. In fact, quantitative predictive modelling of climatic features was one of the first real-life proofs of chaos theory. It could also be the case that the integrative modelling of soil genesis will finish in a similar dead-end: although the modelling of deposition/accumulation processes may seem easy, the influence of organisms and climate is often random or non-linear (Phillips, 1994). As emphasized by Webster (1994): *"Solving the full system of multi-*

variate equations needed to describe the products of soil genesis in individual regions, let alone globally, remains one of the biggest challenges for pedometricians". How can specific soil features/processes, such as buried horizons, fossil soils or karstic features be modelled? Is it still the case that mapping these by means of API and radar-based remote sensing is much easier than predictive soil modelling? Will it ever be possible to reconstruct the observed distribution of soils and their properties? How should the knowledge-based algorithms be integrated with the spatial prediction models?

Continous soil-type maps The methodology for producing continuous categorical maps explained in chapter 6 and its application to soil mapping described in chapter 7, needs to be tested in operational mapping projects with tens of soil types and larger areas. Moreover, it would be interesting to link this classification and visualisation tool directly with the selection of training sets and the definition of classes. This would help in selecting the optimal number and definition of classes. Furthermore, it would be interesting to compare the way the double continuous approach with the fuzzy-metric legend behaves in those special cases when the number of categories is very high or when categories are fairly distinct and vice versa? What are the constraints of the continuous soil maps and are they more user-friendly than the conventional double-crisp maps? Do we still prefer the hard maps? How suitable is the visualisation of uncertainty with whiteness and how well do users interpret the circular legend? Does the taxonomic uncertainty reflect true extra-grade classes or is this an error in the sampling procedure or in the definition of classes?

Advancing the grid-based SIS for land use planning The suggested hybrid grid-based SIS (chapter 7) is the core of this book as it brings together theory from different chapters and shows its application to soil mapping and evaluation. However, many questions still remain. How universal is this SIS and can it be used in operational survey? How much processing power and time will it need for a large study area or if it is used at large scales? How does it is affect applications? Is there an alternative GIS format that could incorporate mixed model of variation while saving digital storage? How to produce a true 3D grid-based GIS of soilscapes? Ventura *et al.* (1996) suggested that voxels might be used to achieve this goal. Grunwald & Barak (2003) recently developed a number of Virtual Reality Modeling Language tools for the visualisation of 3D soilscapes via the world wide web. How practical are voxels and should we limit ourselves to 3D drapes of 2D surfaces and cross-sections? Should these examples only be considered by scientists and teachers, or are they also of interest to engineers and farmers?

Assessing the usability of soil geographical databases In recent years, there has been considerable attention aimed at the development of a methodology for the assessment of the usability of geographical databases. The Centre for Geoinformation at the Wageningen University has hosted several workshops and seminars on this topic, in one of which I also participated. A report of this workshop (Wachowicz *et al.*, 2002) emphasized four aspects of usability: (1) data quality (accuracy, completeness, logical consistency); (2) data form(at); (3) data accessibility and price and (4) quality of the metadata. The importance of each of these aspects may differ from user to user. For example, for environmental modellers the incompatibility, low thematic contrast and detail of multi-source environmental geo-data will especially militate against their full usage (Lilburne, 2001). Although more and more technical measures are introduced to quantify the adequacy of soil geoinformation, a methodology that could be used to quantify 'usability', i.e. objectively measure usability, is still missing. Is this merely a question of the number of users and their satisfaction? The assessment of usability is much more difficult since it needs to take into account a number of technical, organisational and sociological aspects at the same time. How does the price and accessibility of the products reflect on usability of SIS products and how could the usability of an existing SIS be improved?

Bibliography

Addink, E.A., & Stein, A. 1999. A comparison of conventional and geostatistical methods to replace clouded pixels in NOAA-AVHRR images. *International Journal of Remote Sensing*, **20**(5), 961–977.

Agresti, A. 1990. *Categorical Data Analysis*. New York: Wiley.

Amstrong, M., & Dowd, P.A. (eds). 1993. *Geostatistical simulations*. Vol. vol. 7. Dodrecht: Proceedings of the Geostatistical Simulation Workshop, Fontainebleau, France, 27-28 May, Kluwer Academic Publishers.

Atkinson, G.L., & Donev, A.N. 1992. *Optimum Experimental Designs*. Oxford Statistical Science Series, vol. 8. Oxford: Clarendon Press.

Atkinson, P.M. 1997. Mapping Sub-pixel Boundaries from Remotely Sensed Images. *Pages 166–180 of:* Bristol, P.A. (ed), *Innovations in GIS 4*. Taylor & Francis Inc.

Avery, B.W. 1987. *Soil survey methods: a review*. Technical Monograph No. 18. Silsoe: Soil Survey & Land Resource Centre.

Banks, J. (ed). 1998. *Handbook of Simulation - Principles, Methodology, Advances, Applications, and Practice*. New York: John Wiley & Sons.

Beckett, P.H.T. 1976. Soil survey. *Agriculture Progress*, **51**, 33–49.

Beers, T.W., Dress, P.E., & Wensel, L.C. 1966. Aspect transformation in site productivity research. *Journal of Forestry*, **64**, 691–692.

Bell, J.C., Butler, C.A., Thompson, J.A., Robert, P.C., Rust, R.H., & Larson, W.E. 1994. Soil-terrain modeling for site-specific agricultural management. *Pages 209–227 of: Site specific management for agricultural systems: Proceedings of Second International Conference*.

Bell, J.C., Grigal, D.F., & Bates, P.C. 2000. A soil-terrain model for estimating spatial patterns of soil organic carbon. *Pages 295–310 of:* Wilson, J.P., & Gallant, J. (eds), *Terrain analysis : principles and applications*. New York: Wiley & Sons.

Berk, T., Brownston, L., & Kaufman, A. 1982. A new colour-naming system for graphics languages. *IEEE Computer Graphics & Applications*, **2**(3), 37–44.

Bie, S.W., & Beckett, P.H.T. 1973. Comparison of four independent soil surveys by air-photo interpretation, Paphos area (Cyprus). *Photogrammetria*, **29**, 189–202.

Bie, S.W., & Ulph, A. 1972. The economic value of soil survey information. *Journal of Agricultural Economics*, **23**(3), 285–297.

Birkeland, W.P. 1999. *Soils and Geomorphology*. Third edn. New York: Oxford University Press.

Bishop, T.F.A., & McBratney, A.B. 2001. A comparison of prediction methods for the creation of field-extent soil property maps. *Geoderma*, **103**(1-2), 149–160.

Bognar, A. 1984. An outline of geomorphological characteristics on Baranja. *Pages 67–76 of: Annales Universitas Scientarum de Rolando Eotvos Nominate*, vol. 18-19.

Bogunović, M., Vidaček, Ž., Husnjak, S., & Sraka, M. 1998. Inventory of Soils in Croatia. *Agriculturae Conspectus Scientificus*, **63**(3), 105–112.

Bolstad, P.V., & Lillesand, T.M. 1992. Improved classification of forest vegetation in northern Wisconsin through a rule-based combination of soils, terrain, and Landsat TM data. *Forest Science*, **38**(1), 5–20.

Boucneau, G., van Meirvenne, M., Thas, O., & Hofman, G. 1998. Integrating properties of soil map delineations into ordinary kriging. *European Journal of Soil Science*, **49**(2), 213–229.

Bourennane, H., & King, D. 2003. Using multiple external drifts to estimate a soil variable. *Geoderma*, **In Press, Corrected Proof**.

Bourennane, H., King, D., Chery, P., & Bruand, A. 1996. Improving the kriging of a soil variable using slope gradient as external drift. *European Journal of Soil Science*, **47**(4), 473–483.

Bourennane, H., King, D., & Couturier, A. 2000. Comparison of kriging with external drift and simple linear regression for predicting soil horizon thickness with different sample densities. *Geoderma*, **97**(3-4), 255–271.

Brown, A., & Feringa, W. 1999. *A colour handbook for GIS users and cartographers*. International Institute for Geo-information science and Earth Observation.

Brown, D.G. 1998. Mapping historical forest types in Baraga County Michigan, USA as fuzzy sets. *Plant Ecology*, **134**(1), 97–111.

Brown, D.G., & Bara, T.J. 1994. Recognition and reduction of systematic error in elevation and derivative surfaces from 7 1/2 -minute DEMs. *Photogrammetric Engineering and Remote Sensing*, **60**(2), 189–194.

Brus, D.J., & 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.

Buol, S.W., & Hole, F.D. 1980. *Soil genesis and classification*. Ames: Iowa State Univ. Press.

Buringh, P. 1960. The application of aerial photographs in soil surveys. *Page 633 of: Manual of Photographic Interpretation*. Washington DC: American Society of Photogrammetry.

Burrough, P.A. 1991. Soil information systems. *Pages 153–169 of:* Maguire, D.J., Goodchild, M.F., & Rhind, D.W (eds), *Geographical information systems*, vol. 2: Applications. Harlow: Longman Scientific and Technical.

Burrough, P.A. 1993a. Soil variability: a late 20th century view. *Soils and fertilizers*, **56 (May)**, 529–562.

Burrough, P.A. 1993b. The technologic paradox in soil survey: new methods and techniques of data capture and handling. *ITC Journal*, 15–22.

Burrough, P.A., & McDonnell, R.A. 1998. *Principles of geographical information systems*. Oxford: Oxford University Press.

Burrough, P.A., Bouma, J., & Yates, S.R. 1994. The state of the art in pedometrics. *Geoderma*, **62**(3), 311–326.

Burrough, P.A., Gaans, P.F.M., & van Hootsmans, R. 1997. Continuous classification in soil survey: spatial correlation, confusion and boundaries. *Geoderma*, **77**(2-4), 115–135.

Burrough, P.A., van Gaans, P.F.M., & MacMillan, R.A. 2000. High-resolution landform classification using fuzzy k-means. *Fuzzy Sets and Systems*, **113**, 37–52.

Buurman, P., & Sevink, J. (eds). 1995. *From soil map to information system: collection and use of information on Dutch soils (in Dutch)*. Wageningen: Wageningen Press.

Carlile, P., Bui, E., Moran, C., Minasny, B., & McBratney, A.B. 2001. *Estimating particle size distributions and percent sand, silt and clay for six texture classes using the Australian Soil Resources Information System point database*. Canberra: CSIRO Land and Water.

Chiles, J.P., & Delfiner, P. 1999. *Geostatistics: modeling spatial uncertainty*. New York: John Wiley & Sons.

Christensen, R. 1990. *Linear Models for Multivariate, Time Series, and Spatial Data*. Springer Verlag.

Cochran, W.G., & Cox, G.M. 1992. *Experimental designs, second edition*. New York: John Wiley & Sons.

Congalton, R.G., & Green, K. 1999. *Assessing the accuracy of remotely sensed data: principles and practices*. Boca Raton, FL: Lewis.

Cook, S.E., Corner, R.J., Grealish, G., Gessler, P.E., & Chartres, C.J. 1996. A rule-based system to map soil properties. *Soil Science Society of America Journal*, **60**(6), 1893–1900.

Cressie, N.A.C. 1993. *Statistics for Spatial Data, revised edition*. New York: John Wiley & Sons.

D'Avelo, T.P., & McLeese, R.L. 1998. Why are those lines placed where they are?: An investigation of soil map recompilation methods. *Soil Survey Horizons*, **39**(4), 119–126.

Davies, R.E. 1981. *Surveying: Theory and Practice*. 6th edn. Vol. XV. New York: McGraw-Hill.

de Bruin, S., & Stein, A. 1998. Soil-landscape modelling using fuzzy c-means clustering of attribute data derived from a digital elevation model (DEM). *Geoderma*, **83**(1-2), 17–33.

de Gruijter, J.J., & Boogaard, H.L. 2001. *Fuzzy sets voor zachte klassegrenzen. Toepassing op het landevaluatiesysteem BODEGA*. Tech. rept. 346. ALTERRA.

de Gruijter, J.J., & Marsman, B.A. 1984. Transect sampling for reliable information on mapping units. *Pages 150–163 of:* Nielson, D.R., & Bouma, J. (eds), *Soil spatial variability: proceedings of a workshop of the ISSS and SSSA*. Las Vegas: PUDOC.

de Gruijter, J.J., & McBratney, A.B. 1988. A modified fuzzy k-means method for predictive classification. *Pages 97–105 of:* Bock, H.H. (ed), *Classification and Related Methods of Data Analysis*. Elsevier Science Publishers B.V.

de Gruijter, J.J., Walvoort, D.J.J., & van Gaans, P.F.M. 1997. Continuous soil maps - a fuzzy set approach to bridge the gap between aggregation levels of process and distribution models. *Geoderma*, **77**(2-4), 169–195.

Dent, D., & Young, A. 1981. *Soil survey and land evaluation*. Vol. xiii. London, England: George Allen & Unwin.

Deutsch, C.V., & Journel, A.G. 1992. *Geostatistical Software Library and User's Guide*. New York: Oxford University Press.

Dietrich, W.E., Reiss, R., Hsu, M.L., & Montgomery, D.R. 1995. A Process-Based Model For Colluvial Soil Depth And Shallow Landsliding Using Digital Elevation Data. *Hydrological Processes*, **9**(3-4), 383–400.

Dobos, E. 2002. The use of DEM and satellite data for regional scale soil databases. Paper no. 649. *In: Proceedings of the 17th World Congress of Soil Science*. Bangkok: IUSS.

Dobos, E., Micheli, E., Baumgardner, M.F., Biehl, L., & Helt, T. 2000. Use of combined digital elevation model and satellite radiometric data for regional soil mapping. *Geoderma*, **97**(3-4), 367–391.

Draper, N., & Smith, H. 1998. *Applied Regression Analysis, 3rd Edition*. New York: John Wiley & Sons, Inc.

Eastman, J.R., & Fulk, M. 1993. Long sequence time series evaluation using standardized principal components. *Photogrammetric Engineering and Remote Sensing*, **59**(8), 1307–1312.

Environmental Systems Research Institute. 2000. *Using ArcPad*.

Evens, I.S., & Cox, N.J. 1999. Relations between land surface properties: altitude, slope and cruvature. *Pages 13–45 of:* Hergarten, S., & Neugebauer, H.J. (eds), *Process Modelling and Landform Evolution*. Berlin: Springer Verlag.

FAO. 1998. *World reference base for soil resources*. World soil resources reports. Rome: FAO, ISRIC, ISSS.

Felicisimo, A.M. 1994. Parametric statistical method for error detection in digital elevation models. *ISPRS Journal of Photogrammetry and Remote Sensing*, **49**(4), 29–33.

Fernandez, R.N., & Rusinkiewicz, M. 1993. A conceptual design of a soil database for a geographical information system. *International Journal of Geographical Information Systems*, **7**(6), 525–539.

Finn, T. 1993. Use of the Average Mutual Information Index in Evaluating Classification Error and Consistency. *International Journal of Geographical Information Systems*, **7**, 349–366.

Fisher, P.F. 1993. Visualizing uncertainty in soil maps by animation. *Cartographica*, **30**(2-3), 20–27.

Fisher, P.F., & Wood, J. 1998. What is a mountain? or the Englishman who went up a Boolean geographical concept and realised it was fuzzy. *Geography*, **83**, 247–256.

Florinsky, I.V. 1998. Accuracy of local topographic variables derived from digital elevation models. *International Journal of Geographical Information Science*, **12**(1), 47–62.

Florinsky, I.V., & Kuryakova, G.A. 2000. Determination of grid size for digital terrain modelling in landscape investigations exemplified by soil moisture distribution at a micro-scale. *International Journal of Geographical Information Science*, **14**(8), 815832.

Florinsky, I.V., Eilers, R.G., Manning, G., & Fuller, L.G. 2002. Prediction of soil properties by digital terrain modelling. *Environmental Modelling and Software*, **17**, 295–311.

Foote, K.E., & Huebner, D.J. 1995. *Error, Accuracy, and Precision*. University of Colorado.

Forbes, T.R., Rossiter, D., & van Wambeke, A. 1982. *Guidelines for evaluating the adequacy of soil resource inventories*. 1987 printing edn. SMSS Technical Monograph 4. Ithaca, NY: Cornell University Department of Agronomy.

Gabriel, K.R. 1971. The biplot graphical display of matrices with applications to principal component analysis. *Biometrika*, **58**, 453–467.

Gauch, Jr., & Hugh, G. 1993. Prediction, parsimony & noise. *American Scientist*, **81**, 468–478.

Gaylor, D.W., & Sweeny, S. (eds). 1978. *International encyclopedia of statistics*. Confidence intervals and bands. New York: Free Press.

Gessler, P.E., Moore, I.D., McKenzie, N.J., & Ryan, P.J. 1995. Soil-landscape modelling and spatial prediction of soil attributes. *International Journal of Geographical Information Systems*, **9**(4), 421–432.

Gobin, A. 2000. *Participatory and spatial-modelling methods for land resources analysis*. PhD thesis, Katholik Universiteit Leuven.

Gobin, A., Campling, P., & Feyen, J. 2001. Soil-landscape modelling to quantify spatial variability of soil texture. *Physics and Chemistry of the Earth Part B: Hydrology Oceans and Atmosphere*, **26**(1), 41–45.

Goodchild, M., Chih-Chang, L., & Leung, Y. 1994. Visualizing Fuzzy Maps. *Pages 158–167 of:* Hearnshaw, H.M., & Unwin, D.J. (eds), *Visualisation in Geographical Information Systems*. London: John Wiley and Sons Ltd.

Goodchild, M.F., Parks, B., & Staeyert, L. (eds). 1993. *Environmental modelling with GIS*. New York: Oxford University Press.

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

Goovaerts, P. 1999a. Geostatistics in soil science: State-of-the-art and perspectives. *Geoderma*, **89**(1-2), 1–45.

Goovaerts, P. 1999b. Using elevation to aid the geostatistical mapping of rainfall erosivity. *Catena*, **34**(3-4), 227–242.

Gordon, A.D. 1981. *Classification: Methods for Exploratory Analysis of Multivariate Data*. London: Chapman and Hall.

Gorte, B.G.H., & Koolhoven, W. 1990. Interpolation between isolines based on the Borgefors distance transform. *ITC Journal*, **1**(3), 245–247.

Gotway, C.A., & Stroup, W.W. 1997. A Generalized Linear Model Approach to Spatial Data Analysis and Prediction. *Journal of Agricultural, Biological, and Environmental Statistics*, **2**(2), 157–198.

Groot, R. 1993. Making information technology work. *ITC Journal*, **3**, 228–235.

Grunwald, S., & Barak, P. 2003. 3D Geographic Reconstruction and Visualization Techniques Applied to Land Resource Management. *Transactions in GIS*, **7**(2), 231–241.

Guptill, S.C., & Morrison, J.L. (eds). 1995. *Elements of Spatial Data Quality*. ICA Comission on Spatial Data Quality. Elsevier Science Ltd.

Hall, G.B., Wang, F., & Subaryono. 1992. Comparison of Boolean and fuzzy classification methods in land suitability analysis by using geographical information systems. *Environmental Planning*, **A24**, 497–516.

Hansen, M.C., & Reed, B. 2000. A comparison of the IGBP DISCover and University of Maryland 1 km global land cover products. *International Journal of Remote Sensing*, **21**(6-7), 1365–1373.

Hartemink, A.E., McBratney, A.B., & Cattle, J.A. 2002. Developments and trends in soil science: 100 volumes of Geoderma (1967-2001). *Geoderma*, **100**, 217-268.

Hartigan, J.A. 1975. *Clustering algorithms*. New York: Wiley.

Hengl, T., Walvoort, D.J.J., & Brown, A. 2002. Pixel (PM) and Colour mixture (CM): GIS techniques for visualisation of fuzziness and uncertainty of natural resource inventories. *Pages 300–309 of:* Hunter, Gary J., & Lowell, Kim (eds), *Proceedings of the 5th International Symposium on Spatial Accuracy Assesment in Natural Resources and Environmental Sciences (Accuracy 2002)*.

Hengl, T., Heuvelink, G.B.M, & Stein, A. 2003a. *Comparison of kriging with external drift and regression-kriging. Technical report*. Enschede: International Institute for Geo-information Science and Earth Observation (ITC).

Hengl, T., Gruber, S., & Shrestha, D.P. 2003b. *Digital Terrain Analysis in ILWIS*. Enschede: Lecture notes, International Institute for Geo-Information Science & Earth Observation (ITC).

Heuvelink, G.B.M. 1998. *Error propagation in environmental modelling with GIS*. London, UK: Taylor & Francis.

Heuvelink, G.B.M., & Webster, R. 2001. Modelling soil variation: past, present, and future. *Geoderma*, **100**(1), 269–301.

Hole, F.D. 1953. Suggested terminology for describing soil as three-dimensional bodies. *Soil Science Society of America Procceedings*, **17**, 131–135.

Hole, F.D. 1978. An approach to landscape analysis with emphasis on soils. *Geoderma*, **21**, 1–23.

Holmes, K.W., Chadwick, O.A., & Kyriakidis, Ph.C. 2000. Error in a USGS 30m digital elevation model and its impact on digital terrain modeling. *Journal of Hydrology*, **233**, 154–173.

Hoosbeek, M.R., & Bryant, R.B. 1992. Towards the quantitative modeling of pedogenesis - a review. *Geoderma*, **55**, 183–210.

Hootsmans, R.M. 1996. *Fuzzy sets and series analysis for visual decision support in spatial data exploration*. PhD thesis, University of Utrecht.

Horn, B.K.P. 1981. Hill shading and the reflectance map. *Proceedings IEEE*, **69**(1), 14–47.

Hu, Z., Shafiqul, I., Cheng, Y., Hu, Z.L., Cheng, Y.Z., & Islam, C. 1997. Statistical characterization of remotely sensed soil moisture images. *Remote Sensing of Environment*, **61**(2), 310–318.

Hudson, B.D. 1992. The soil survey as paradigm-based science. *Soil Science Society of America Journal*, **56**(3), 836–841.

Hudson, G., & Wackernagel, H. 1994. Mapping temperature using kriging with external drift: theory and an example from Scotland. *International Journal of Climatology*, **14**(1), 77–91.

Hutchinson, M.F. 1989. A new procedure for gridding elevation and stream line data with automatic removal of spurious pits. *Journal of Hydrology*, **106**, 211–232.

Indorante, S.J., McLeese, R.L., Hammer, R.D., Thompson, B.W., & Alexander, D.L. 1996. Positioning soil survey for the 21st century. *Journal of Soil and Water Conservation*, **51**(1), 21–28.

Irvin, B.J., Ventura, S.J., & Slater, B.K. 1997. Fuzzy and isodata classification of landform elements from digital terrain data in Pleasant Valley, Wisconsin. *Geoderma*, **77**, 137–154.

Isaaks, E.H., & Srivastava, R.M. 1989. *Applied Geostatistics*. New York: Oxford University Press.

Janssen, L.L.F., & Huurneman, G.C. 2001. *Principles of Remote Sensing*. ITC Educational Textbook Series. Enschede: ITC.

Jenny, H. 1941. *Factors of soil formation - a system of quantitative pedology*. New York: McGraw-Hill.

Jenny, H. 1980. *The soil resource: origin and behavior*. K. clayton edn. Ecological studies, vol. 37. New York: Springer-Verlag.

Jiang, B. 1996. *Fuzzy Overlay Analysis and Visualization in Geographic Information Systems*. PhD thesis, University of Utrecht.

Jiang, B., Kainz, W., & Ormeling, F.J. 1995. A modified HLS system used in the visualisation of uncertainty. *Pages 701–712 of: International Symposium on RS, GIS and GPS in Sustainable Development of Environmental Monitoring*.

Kitanidis, P.K. 1994. Generalized covariance functions in estimation. *Mathematical Geology*, **25**, 525–540.

Knotters, M., Brus, D.J., & Voshaar, J.H.O. 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.

Königel, C.J. 2002. The role of pedogenesis in modelling geo-ecological evolution. *In: Proceedings of the 17th World Congress of Soil Science*. Bangkok: IUSS.

Kovačević, P., & Jakšić, V. 1964. *Handbook for Soil Survey methodology (in Croatian)*. Zagreb: Inter-institutional comission for the production of the Soil map of Yugoslavia.

Kraak, M.J., & Ormeling, F. 1996. *Cartography: Visualisation of spatial data*. Essex: Addison Wesley Longman.

Lagacherie, P., Andrieux, P., & et al. 1996. Fuzziness and uncertainty of soil boundaries: from reality to coding in GIS. Geographic objects with indeterminate boundaries. *Pages 275–286 of:* A., Burrough P., & U., Frank A. (eds), *Geographic objects with indeterminate boundaries*. London: Taylor & Francis.

Lane, P.W. 2002. Generalized linear models in soil science. *European Journal of Soil Science*, **53**, 241–251.

Lark, R.M. 2000. Regression analysis with spatially autocorrelated error: Simulation studied and application to mapping of soil organic matter. *International Journal of Geographical Information Science*, **14**(3), 247–264.

Lesch, S.M., Strauss, D.J., & Rhoades, J.D. 1995. Spatial prediction of soil salinity using electromagnetic induction techniques 2. An efficient spatial sampling algorithm suitable for multiple linear regression model identification and estimation. *Water Resource Research*, **31**, 387–398.

Li, Z. 1994. A comparative study of the accuracy of digital terrain models (DTMs) based on various data models. *ISPRS Journal of Photogrammetry and Remote Sensing*, **49**, 2–11.

Lilburne, L. 2001. *The Scale Matcher: a framework for assessing scale compatibility of environmental data and models*. PhD thesis, University of Otago.

Lillesand, T.M., & Kiefer, R.W. 2000. *Remote Sensing and Image Interpretation*. Vol. 4th Ed. John Wiley and Sons.

Looney, S.W., & Gulledge, Jr., T.R. 1985. Use of the Correlation Coefficient with Normal Probability Plots. *The American Statistician*, **39**, 75–79.

Lopez, C. 2000. Improving the elevation accuracy of digital elevation models: A comparison of some error detection procedures. *Transactions in GIS*, **4**(1), 43–64.

MacEachren, A.M., & Kraak, M.J. 1997. Exploratory cartographic visualization: advancing the agenda. *Computers & Geosciences*, **23**(4), 335–344.

Maclean, A.L., D'Avello, T.P., & Shetron, S.G. 1993. The use of variability diagrams to improve the interpretation of digital soil maps in a GIS. *PE and RS, Photogrammetric Engineering and Remote Sensing*, **59**(2), 223–237.

MacMillan, R.A. 2000. *A protocol for preparing digital elevation (DEM) data for input and analysis using the landform segmentation model (LSM) programs, prepared for: the Soil Variability Analysis to Enhance Crop Production (SVAECP) Project*. LandMapper Environmental Solutions.

MacMillan, R.A., Pettapiece, W.W., Nolan, S.C., & Goddard, T.W. 2000. A generic procedure for automatically segmenting landforms into landform elements using DEMs, heuristic rules and fuzzy logic. *Fuzzy Sets and Systems*, **113**, 81–109.

Marsman, B.A., & de Gruijter, J.J. 1986. *Quality of soil maps : a comparison of soil survey methods in a sandy area*. Soil survey papers 15. Wageningen: Soil Survey Institute.

Marsman, B.A., & de Gruijter, J.J. 1987. Quality of soil maps: a comparison of soil survey methods in a sandy area. *Soil Survey Papers, Netherlands Soil Survey Institute, Wageningen*, **15**(103).

Martinoni, D. 2002. *Models and Experiments for Quality Handling in Digital Terrain Modelling*. PhD thesis, University of Zurich.

Martinović, J., & Vranković, A. (eds). 1997. *Croatian Soil database (in Croatian)*. Vol. I. Zagreb: Ministry of Environmental protection and Physical planning.

Matheron, G. 1969. *Le krigeage universel*. Vol. 1. Fontainebleau: Cahiers du Centre de Morphologie Mathematique, Ecole des Mines de Paris.

MathSoft Inc. 1999. *S-PLUS 4 Guide to Statistics*. Vol. 1 and 2. Seattle: MathSoft Inc.

McBratney, A.B. 1998. Some considerations on methods for spatially aggregating and disaggregating soil information. *Nutrient Cycling in Agroecosystems*, **50**, 51–62.

McBratney, A.B., & Walvoort, D.J.J. 2001. Generalised Linear Model Kriging: A generic framework for kriging with secondary data. *In:* Meirvenne, Marc Van (ed), *4th Conference of the Working Group on Pedometrics*. Ghent, Belgium: University of Ghent.

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

McBratney, A.B., Odeh, I.O.A., de Gruijter, J.J., & McSweeney, K. 1997. Application of fuzzy sets in soil science: fuzzy logic, fuzzy measurements and fuzzy decisions. *Geoderma*, **77**(2-4), 85–113.

McBratney, A.B., Odeh, I.O.A., Bishop, T.F.A., Dunbar, M.S., & Shatar, T.M. 2000. An overview of pedometric techniques for use in soil survey. *Geoderma*, **97**(3-4), 293–327.

McBratney, A.B., Mendonça Santos, M.L., & Minasny, B. 2003. On digital soil mapping. *Geoderma*, **(in press)**.

McGwire, K., Friedl, M., & Estes, J.E. 1993. Spatial structure, sampling design and scale in remotely-sensed imagery of California savanna woodland. *International Journal of Remote Sensing*, **14**(11), 2137–2164.

McKenzie, N.J., & 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., & Ryan, P.J. 1999. Spatial prediction of soil properties using environmental correlation. *Geoderma*, **89**(1-2), 67–94.

McKenzie, N.J., Gessler, P.E., Ryan, P.J., & O'Connell, D.A. 2000. The Role of Terrain Analysis in Soil Mapping. *Pages 245–265 of:* Wilson, John P., & Gallant, John C. (eds), *Terrain Analysis: Principles and Applications*. John Wiley & Sons, Inc.

McSweeney, K., Gessler, P.E., Slater, B.K., Hammer, R.D., Bell, J.C., & Petersen, G.W. 1994. Towards a new framework for modeling the soil-landscape continuum. *Pages 127–145 of:* Bartels, J.M. (ed), *Factors of soil formation*. Denver: Madison: Soil Science Society of America.

Minasny, B., & McBratney, A.B. 1999. *FuzME, Program for Fuzzy k-means with Extragrades clustering*. McMillan Building A05, The University of Sydney, NSW 2006: Australian Centre for Precision Agriculture.

Minasny, B., & McBratney, A.B. 2001. A rudimentary mechanistic model for soil formation and landscape development II. A two-dimensional model incorporating chemical weathering. *Geoderma*, **103**, 161–179.

Minasny, B., McBratney, A.B., & Whelan, B.M. 2002. *VESPER version 1.5.* McMillan Building A05, The University of Sydney, NSW 2006: Australian Centre for Precision Agriculture.

Mitas, L., & Mitasova, H. 1999. Spatial interpolation. *Pages 481–492 of:* Longley, P., Goodchild, M.F., Maguire, D.J., & Rhind, D.W. (eds), *Geographical Information Systems: Principles, Techniques, Management and Applications*, vol. 1. Wiley.

Mitasova, H., Hofierka, J., Zlocha, M., & Iverson, L. R. 1996. Modelling topographic potential for erosion and deposition using GIS. *International Journal of Geographical Information Systems*, **10**(5), 629–641.

Mitasova, H., Brown, W. M., Mitas, L., & Warren, S. 1997. Multi-dimensional GIS environment for simulation and analysis of landscape processes. *Page 19 pp. of: ASAE Annual International Meeting, Minneapolis, 10-14 August, 1997.*

Moellering, H. 1987. *A Draft Proposed Standard for Digital Cartographic Data. Report no. 8.* USA National Committee for Digital Cartographic Standards, American Congress on Surveying and Mapping.

Moore, I.D., Gessler, P.E., Nielsen, G.A., & Peterson, G.A. 1993. Soil attribute prediction using terrain analysis. *Soil Science Society of America Journal*, **57**(2), 443–452.

Moran, M.S., Hymer, D.C., Qi, J., & Kerr, Y. 2002. Comparison of ERS-2 SAR and Landsat TM imagery for monitoring agricultural crop and soil conditions. *Remote Sensing of Environment*, **79**(2-3), 243–252.

Mowrer, H.T., & Congalton, R.G. (eds). 2000. *Quantifying Spatial Uncertainty in Natural Resources: Theory and Application for GIS and Remote Sensing*. Chelsea, MI: Ann Arbor Press.

Neter, J., Kutner, M.H., Nachtsheim, C.J., & Wasserman, W. (eds). 1996. *Applied Linear Statistical Models*. 4th edn. The McGraw-Hill Companies.

Niblack, W. 1986. *An Introduction to Digital Image Processing*. Prentice/Hall International.

Odeh, I.O.A., & McBratney, A.B. 2000. Using AVHRR images for spatial prediction of clay content in the lower Namoi Valley of eastern Australia. *Geoderma*, **97**(3-4), 237–254.

Odeh, I.O.A., McBratney, A.B., & Chittleborough, D.J. 1990. Design of optimal sample spacings for mapping soil using fuzzy k-means and regionalized variable theory. *Geoderma*, **47**(1-2), 93–122.

Odeh, I.O.A., McBratney, A.B., & Chittleborough, D.J. 1994. Spatial prediction of soil properties from landform attributes derived from a digital elevation model. *Geoderma*, **63**(3-4), 197–214.

Odeh, I.O.A., McBratney, A.B., & Chittleborough, D.J. 1995. Further results on prediction of soil properties from terrain attributes: heterotopic cokriging and regression-kriging. *Geoderma*, **67**(3-4), 215–226.

Opsomer, J.D., Ruppert, D., Wand, M.P., Holst, U., & Hssjer, O. 1999. Kriging with Nonparametric Variance Function Estimation. *Biometrics*, **55**(3), 704710.

Ott, R.L., & Longnecker, M. (eds). 2001. *An Introduction to Statistical Methods and Data Analysis*. 5th edn. Duxbury press.

Papritz, A., & Stein, A. 1999. Spatial prediction by linear kriging. *Pages 83–113 of:* Stein, A., van der Meer, F., & Gorte, B. (eds), *Spatial statistics for remote sensing*. Dodrecht: Kluwer Academic Publishers.

Park, S.J., & Vlek, P.L.G. 2002. Environmental correlation of three-dimensional soil spatial variability: a comparison of three adaptive techniques. *Geoderma*, **109**(1-2), 117–140.

Patil, G.P. 2002. Composite sampling. *Pages 387–391 of:* El-Shaarawi, Abdel H., & Piegorsch, Walter W. (eds), *Encyclopedia of Environmetrics*, vol. 1. Chichester, UK: John Wiley & Sons.

Pebesma, E.J. 2003. *Gstat: multivariable geostatistics for S*. Vienna, Austria: Distributed Statistical Computing, working papers, available via [http://www.ci.tuwien.ac.at/Conferences/DSC-2003/].

Pebesma, E.J., & Wesseling, C.G. 1998. Gstat, a program for geostatistical modelling, prediction and simulation. *Computers and Geosciences*, **24**(1), 17–31.

Phillips, J.D. 1994. Deterministic uncertainty in landscapes. *Earth Surface Processes and Landforms*, **19**, 389–401.

Pilouk, M. 1992. *Fidelity improvement of DTM from contours*. MSc thesis, ITC.

Quinn, P., Beven, K., Chevallier, P., & Planchon, O. 1991. The prediction of hillslope flow paths for distributed hydrological modelling using digital terrain models. *Hydrological processes*, **5**, 59–79.

Raaflaub, L., & Collins, M.J. 2002. The Effect Errors in Gridded Digital Elevation have on Derived Topographic Parameters Using Monte Carlo Simulation: A Comparison of Algorithms. *Page 279 of:* Hunter, Gary J., & Lowell, Kim (eds), *Proceedings of the 5th International Symposium on Spatial Accuracy Assesment in Natural Resources and Environmental Sciences (Accuracy 2002)*.

Radošević, N. 1979. Pre-war military map 1:100 000 (1:50 000) and todays triangulation (in Serbocroatian). *VGL*, 129–148.

Rossiter, D.G. 2001. *Methodology for Soil Resource Inventories*. 2nd edn. ITC Lecture Notes SOL.27. Enschede, the Netherlands: ITC.

Rossiter, D.G., & Hengl, T. 2002. *Technical note: Creating geometrically-correct photo-interpretations, photomosaics, and base maps for a project GIS*. Tech. rept. ITC, Soil Science Division.

Schneider, B. 1998. *Geomorphologically Plausibel Reconstruction of the Digital Representation of Terrain Surfaces from Contour Data (in german)*. PhD thesis, Universtiy of Zurich.

Scull, P., Franklin, J., Chadwick, O.A., & McArthur, D. 2003. Predictive soil mapping: a review. *Progress in Physical Geography*, **27**(2), 171–197.

Shary, P.A., Sharaya, L.S., & Mitusov, A.V. 2002. Fundamental quantitative methods of land surface analysis. *Geoderma*, **107**(1-2), 1–32.

Skidmore, A.K., Varekamp, C., Wilson, L., Knowles, E., & Delaney, J. 1997. Remote sensing of soils in a eucalypt forest environment. *International Journal of Remote Sensing*, **18**(1), 39–56.

Soil Resource Inventory Study Group. 1977. Soil resource inventories: Proceedings of a workshop held at Cornell University April 4-7, 1977. *Pages 1–353 of: Soil resource inventories*. Agronomy Mimeo No. 77-23. Ithaca, NY: Cornell University Department of Agronomy.

Soil Resource Inventory Study Group. 1978. Soil resource inventories and development planning: Proceedings of a workshop organized by the Soil Resource Inventory Study Group at Cornell University December 11-15, 1978. *Pages 1–332 of: Soil resource inventories and development planning*. Agronomy Mimeo No. 79-23. Ithaca, NY: Cornell University Department of Agronomy.

Soil survey Division staff. 1993. *Soil survey manual*. Washington: United States Department of Agriculture.

Sokal, R.R., & Sneath, P.H.A. 1976. *Principles of Numerical Taxonomy*. San Francisco: W.H. Freeman.

Steel, R.G.D., & Torrie, J.H. 1980. *Principles and procedures of statistics: a biometrical approach*. 2nd edn. New York: McGraw-Hill.

Stein, A. 1991. *Spatial interpolation*. PhD thesis, Wageningen Agricultural University, Wageningen.

Stenger, R. 1996. *Dynamics of Soil Mineral Nitrogen within a Landscape section: - Monitoring - Process Studies - Simulations (in German)*. PhD thesis, Technical University Munich.

Tang, G., Shi, W., & Zhao, M. 2002. Evaluation on the Accuracy of Hydrologic Data Derived From DEMs of Different Spatial Resolution. *Pages 204–213 of:* Hunter, Gary J., & Lowell, Kim (eds), *Proceedings of the 5th International Symposium on Spatial Accuracy Assesment in Natural Resources and Environmental Sciences (Accuracy 2002)*.

Tempfli, K. 1999. DTM accuracy assesment. *Pages 1–11 of: ASPRS Annual Conference*.

Thomas, A.L., King, D., Dambrine, E., Couturier, A., & Roque, J. 1999. Predicting soil classes with parameters derived from relief and geologic materials in a sandstone region of the Vosges mountains (Northeastern France). *Geoderma*, **90**(3-4), 291–305.

Thompson, J.A., Bell, J.C., & Butler, C.A. 1997. Quantitative soil-landscape modeling for estimating the areal extent of hydromorphic soils. *Soil Science Society of America Journal*, **61**(3), 971–980.

Thompson, J.A., Bell, J.C., & Butler, C.A. 2001. Digital elevation model resolution: effects on terrain attribute calculation and quantitative soil-landscape modeling. *Geoderma*, **100**, 67–89.

Triantafilis, J., Ward, W.T., & McBratney, A.B. 2001. Land suitability assesment in the Namoi Valley of Australia, using a continuous model. *Australian Journal of Soil Research*, **39**, 273–290.

Tucker, L., & MacCallum, R. 1997. *Exploratory Factor Analysis*. Columbus: Ohio State University.

Unit Geo Software Development. 2001. *ILWIS 3.0 Academic user's guide*. Enschede, available at [http://www.itc.nl/ilwis/]: ITC.

USGS - NASA Distributed Active Archive Centre. 2001. *FTP access to Global AVHRR 10-day composite data*. US Geological Survey.

Valenzuela, C., & Baumgardner, M.F. 1990. Selection of appropriate cell sizes for thematic maps. *ITC Journal*, **3**, 219–224.

van Groenigen, J.W., Siderius, W., & Stein, A. 1999. Constrained optimisation of soil sampling for minimisation of the kriging variance. *Geoderma*, **87**(3-4), 239–259.

van Kuilenburg, J., de Gruijter, J.J., Marsman, B.A., & 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 Reeuwijk, P. 1984. *Exchangeable bases, base sturation and pH*. Idem: Part II. International Soil and Reference Centre.

Ventura, S.J., Irvin, B.J., Slater, B.K., & McSweeney, K. 1996. Data structures for representation of soil stratigraphy. *Pages 63–68 of:* et al., M. Goodchild (ed), *GIS and Environmental Modeling: Progress and Research Issues*. Boulder: GIS World Books.

Vink, A.P.A. 1975. *Land use in advancing agriculture*. Vol. x. New York: Springer-Verlag.

Škorić, A., Filipovski, G., & Čirić, M. 1985. *Classification of yugoslav soils*. Sarajevo: Academy of Sciences and Arts of Bosnia and Hercegovina.

Wachowicz, M., Riedemann, C., Vullings, W., Surez, J., & Cromvoets, J. 2002. Workshop report on spatial data usability. *Page 8 of: Agile 2002 Conference*.

Wackernagel, H. 1998. *Multivariate geostatistics: an introduction with applications*. 2nd edition edn. Springer-Verlag.

Warrick, A.W., & Myers, D.E. 1987. Optimisation of sampling locations for variogram calculations. *Water Resource Research*, **23**, 496–500.

Webster, R. 1994. The development of pedometrics. *Geoderma*, **62**(1/3), 1–15.

Webster, R., & Beckett, P.H.T. 1968. Quality an usefulness of soil maps. *Nature*, **219**, 680–682.

Webster, R., & Olivier, M.A. 1990. *Statistical Methods in Soil and Land Resource Survey*. Spatial Information Systems. Oxford: Oxford Uni. Press.

Western, S. 1978. *Soil survey contracts & quality control*. Mongraphs on Soil Survey. Oxford: Clarendon Press.

White, R.E. 1997. *Principles and Practice of Soil Science*. 3rd edn. The Soil as a Natural Resource. Oxford: Blackwell Science Ltd.

Wilson, J.P., Repetto, P.L., & Snyder, R.D. 2000. Effect of Data Source, Grid Resolution, and Flow-Routing Method on Computed Topographic Attributes. *Pages 133–161 of:* Wilson, John P., & Gallant, John C. (eds), *Terrain Analysis: Principles and Applications*. John Wiley & Sons, Inc.

Wise, S.M. 2000. Assessing the quality for hyrdological applications of digital elevation models derived from contours. *Hydrological Processes*, **14**(11-12), 1909–1929.

Woodcock, C.E., & Gopal, S. 2000. Fuzzy set theory and thematic maps: Accuracy assessment and area estimation. *International Journal of Geographical Information Science*, **14**(2), 153–172.

Yost, R.S., Uehara, G., & Fos, R.L. 1982. Geostatistical analysis of soil chemical properties of large land areas. II. Kriging. *Soil Science Society of America Journal*, **46**(5), 1033–1037.

Zadeh, L. 1965. Fuzzy sets. *Information and Control*, **8**, 338–353.

Zevenbergen, L.W., & Thorne, C.R. 1987. Quantitative analysis of land surface topography. *Earth Surface Processes Landforms*, **12**, 47–56.

Zhu, A.X., Band, L.E., Dutton, B., & Nimlos, T. 1996. Automated soil inference under fuzzy logic. *Ecological Modeling*, **90**, 123–145.

Zhu, A.X., Band, L., Vertessy, R., & Dutton, B. 1997. Derivation of soil properties using a soil land inference model (SoLIM). *Soil Science Society of America Journal*, **61**(2), 523–533.

Zhu, A.X., B., Hudson, J., Burt, K., Lubich, & D., Simonson. 2001. Soil mapping using GIS, expert knowledge, and fuzzy logic. *Soil Science Society of America Journal*, **65**, 1463–1472.

Zinck, J.A. 1988. *Physiography & Soils*. ITC Lecture Notes, SOL.41. Enschede, the Netherlands: ITC.

Zinck, J.A., & Valenzuela, C.R. 1990. Soil geographic database: structure and application examples. *ITC Journal*, **1990**(3), 270–294.

Summary

Hengl, T. 2003. Pedometric mapping: bridging the gaps between the conventional and pedometric approaches. PhD thesis, Wageningen University.

In recent years, digital soil mapping has faced rapid development of new and economic methods, mainly due to the increasing sources of auxiliary maps. The main objective of this research was to develop a methodology for pedometric mapping that can be used to bridge gaps between the mechanistic pedometric and conventional techniques. The thesis covers seven methodological aspects of soil mapping: sampling, pre-processing, photo-interpretation, interpolation, visualisation, organisation and quality control.

SAMPLING: This chapter evaluates spreading of observations in feature and geographical spaces as a key to sampling optimisation for spatial prediction by correlation with auxiliary maps. Although auxiliary data are commonly used for mapping soil variables, problems associated with the design of sampling strategies are rarely examined. When generalized least squares estimation is used, the overall prediction error depends upon spreading of points in both feature and geographical space. Allocation of points uniformly over the feature space range proportionally to the distribution of predictor (equal range stratification or ER design) is suggested as a prudent sampling strategy when the regression model between the soil and auxiliary variables is unknown. An existing 100-observation sample from a 50×50 km soil survey in central Croatia was used to illustrate these concepts. It was re-sampled to 25-point datasets using different experimental designs: ER and two response surface designs (minmax and D2). The designs were compared for their performance in predicting soil organic matter from elevation (univariate example) using the overall prediction error as an evaluation criterion. The ER design gave similar overall prediction error as the minmax design, suggesting that it is a good compromise between accurate model estimation and minimisation of spatial autocorrelation of residuals. In addition, the ER design was extended to the multivariate case. Four predictors (elevation, temperature, wetness index and NDVI) were transformed to standardised principal components. The sampling points were then assigned to the components in proportion to the variance explained by a principal component analysis and following the ER design.

PRE-PROCESSING: Quality of DEMs and DEM-derived products directly affects the quality of terrain analysis applications. Three approaches to the reduction of errors in

DEM and DEM-derived products have been described: (a) by using empirical knowledge, e.g. to adjust elevations using medial axes or stream networks; (b) by applying filtering operations and (c) by error propagation. Filtering operations are used to replace erratic values or reduce outliers using the spatial dependence structure and probability of exceeding a value estimated from the neighbours. In the case of error propagation, the errors are reduced by calculating the average value of multiple realisations. The methods were tested using a 3.8×3.8 km sample area covering two distinct landscapes: hilland and plain with terraces. The contour data was interpolated using the linear interpolation. The proportion of artefacts (padi terraces) in the unfiltered DEM was 17.3%. After the addition of medial axes, filtering of outliers and adjustment of elevation for streams, the proportion of padi terraces was reduced to 2.2%. Remaining errors in terrain parameters such as undefined pixels and local outliers were reduced using filtering with iterations and by error propagation. The proportion of outliers in all terrain parameters did not exceed 2% of the total area. Both the filtering approach and error propagation give somewhat smoother maps of terrain parameters. The advantage of filtering of outliers is that it employs the structure of the spatial dependence. The advantage of error propagation is that it can be easier automated. The reduction of errors improved the mapping of landform facets (classification) and solum thickness (regression). The classification accuracy increased from 51.3% to 72% and the R^2 of the regression model for the prediction of the solum thickness increased from 0.27 to 0.40.

PHOTO-INTERPRETATION: A method to enhance manual landform delineation using photo-interpretation to map a larger area is described. Conventional aerial photo-interpretation (API) maps using a geo-pedological legend of 21 classes were prepared for six sample areas totaling 111 km^2 in Baranja region, eastern Croatia. Nine terrain parameters extracted from a digital elevation model (ground water depth, slope, plan curvature, profile curvature, viewshed, accumulation flow, wetness index, sediment transport index and the distance to nearest watercourse) were used to extrapolate photo-interpretation over the entire survey area (1062 km^2). The classification accuracy was assessed using the error matrix, calculated by comparing both the whole API maps and point samples, with the results of classification. The first results, using a maximum-likelihood classifier, were 58.2% (hill land), 39.1% (plain), and 45.3% (entire area) reproducibility of the training set. Six classes in the plain were responsible for a large proportion of the misclassifications, due to an insufficiently detailed digital elevation model and the complex nature of landforms (point bar complexes, levees, active channel banks), which can not be explained with the terrain parameters only. Reproducibility for a simplified legend of 15 classes over the study area was improved to 65.8% (plain), 58.2% (hill land) and 63.4% (entire area) using the whole-API training set. After the simplification of legend (15) and with the iterative (3) selection of point-sample training set, classification was able to reproduce 97.6% (hill land), 86.7% (plain), and 90.2% (entire area) of the training set. The supervised classification showed fine details not achieved by photo-interpretation. The number of manual photo-interpretations that had to be prepared was reduced from 84 to 6.

INTERPOLATION: A methodological framework for spatial prediction based on regression-

kriging is described and compared versus ordinary kriging and plain regression. The data are first transformed using logit transformation for target variables and factor analysis for continuous predictors (auxiliary maps). The target variables are then fitted using step-wise regression and residuals interpolated using kriging. A generic visualisation method is used to simultaneously display predictions and associated uncertainty. The framework was tested using 135 profile observations from the national survey in Croatia, divided into interpolation (100) and validation sets (35). Three target variables: organic matter, pH in topsoil and topsoil thickness were predicted from six relief parameters and nine soil mapping units. Prediction efficiency was evaluated using the mean error and root mean square error (RMSE) of prediction at validation points. The results show that the proposed framework improves efficiency of predictions. Moreover, it ensured normality of residuals and enforced prediction values to be within the physical range of a variable. For organic matter, it achieved lower relative RMSE than ordinary kriging (53.3% versus 66.5%). For topsoil thickness, it achieved a lower relative RMSE (66.5% versus 83.3%) and a lower bias than ordinary kriging (0.15 versus 0.69 cm). The prediction of pH in topsoil was difficult with all three methods. This framework opens a possibility to develop a bundle algorithm that can be implemented in a GIS to interpolate soil profile data from existing datasets.

VISUALISATION: A method to visualise multiple membership maps, called *"Colour mixture"* (CM) is described and compared to alternative techniques: defuzzification and Pixel mixture. Six landform parameters were used to derive the landform classes using supervised fuzzy k-means classification. The continuous categorical map is derived by GIS calculations with colours, where colour values are considered to represent the taxonomic space spanned by the attribute variables. Coordinates of the 9 class centres (landform facets) were first transformed from multivariate to two-dimensional attribute space, and then projected on the Hue Saturation Intensity (HSI) colour-wheel. The taxonomic value was coded with the Hue and confusion with Saturation. To improve visual impression, saturation was replaced with whiteness. Classes that were closer in attribute space were merged into similar generic colours. The CM technique limits the derived mixed-colour map to seven generic hues independently of the total number of classes, which provides basis for automated generalisation. Saturation derived from the mixed-colour map was used to derive primary boundaries and to locate areas of higher taxonomic confusion.

ORGANIZATION: The key concepts, operations and organizational structure of a grid-based Soil Information System (SIS) are compared to a conventional polygon-based SIS and illustrated with a case study of a 3.8×3.8 km area in eastern Croatia. The key spatial entity in this system is a grid cell and all GIS layers were brought to the same grid resolution (25 m in this case). The soil variables were modelled using the mixed model of spatial variation, so that both discrete and continuous transitions were possible. The SIS, in this case study, included 21 predictor maps (photo-interpretation map, terrain parameters and remote sensing images), six maps of soil variables (solum thickness, occurence of the mollic, calcic and gleyic horizon, toposil thickness and topsoil silt content) and six derived maps of soil types. Each soil variable was interpolated using a hybrid interpolation technique (regression-kriging). The interpolated maps

were then classified using a continuous classifier (fuzzy k-means) to produce membership maps. These were then used to derive land suitability for wheat production on a continuous scale (0–1), as an example of interpretation that can be derived from the SIS. The photo-interpretation map was shown to be a somewhat better predictor of the listed soil variables than the terrain and remote sensing maps. Comparison of goodness of fit and thematic confusion showed that the grid-based SIS gives in general better fit to the original data, higher level of detail and more reliable predictions than the conventional (polygon-based) SIS. The advantages of the proposed SIS, compared to a conventional survey, are: (1) it offers a map of soil types rather than of the soil-mapping units; (2) all variables are mapped as continuous spatial fields at fine grain of detail; (3) it offers a measure of uncertainty for both input and derived maps; (4) both discrete and continuous transitions are possible and (5) the original soil observations and interpolation/classification parameters are stored in tables as a part of the SIS, so that derived maps can be updated. The disadvantages are: (1) it is computationally demanding and requires a large amount of storage; (2) it is more costly (collection and pre-processing of auxiliary variables) and (3) SIS is sensitive on the quality of the input data.

QUALITY CONTROL: Methodology to assess the quality and adequacy of a national soil resource inventory and to evaluate its usability is described. Six 1:50 K map sheets (of 185 total), three control surveys (each of size 4×4 km) and ten full profile descriptions in the main landscape regions of Croatia were used to estimate the effective map scale, accuracy of map legends and thematic accuracy of profile observations. In addition, the existing digital data sets (soil map of Croatia at scale 1:300 K and database with 2198 profiles) were evaluated for thematic purity and contrast for clay content, pH and organic matter. New methods were developed and tested to assess the spatial accuracy of soil boundaries and the thematic overlap among map units. In the case study, the average polygon size and the positional accuracy of primary soil boundaries (about ±40 m) correspond to the 1:150 K scale, while the inspection density corresponds to the 1:250 K scale. Mapping units are heterogeneous with an average relative variation of 17% within units and a mean thematic overlap of 66% among geographically-adjacent units. There is a large difference between the original legend and the validation sample when considered as taxonomic classes, but much less so when classes are grouped by similarity. The inventory is adequate for small-scale applications but not in general at detailed scales. The major usability problems are compound map units, lack of specific interpretations corresponding to user needs, and lack of legal clarity on ownership and use.

The general conclusion is that the proposed pedometric mapping methodology enhances the practice of soil mapping making the soil maps more objective, detailed and more compatible for integration with other environmental geo-data. There is no need to use the concept of soil mapping units or use double-crisp soil maps anymore. On the other hand, instead of abandoning photo-interpretation, soil classification or empirical knowledge on soils, these methods can be successfully integrated with pedometric techniques.

Samenvatting

Hengl, T. 2003. Pedometrische kartering: overbrugging de kloof tussen de traditionele- en pedometrische benaderingen. Doctoraal proefschrift, Universiteit Wageningen.

Digitale bodemkartering heeft de afgelopen jaren een snelle ontwikkeling doorgemaakt door nieuwe- en economische methoden, als gevolg van het beschikbaar komen van vele soorten secondaire kaarten (terrein parameters, satellitbeelden etc.). Het hoofddoel van dit onderzoek was om een methodogie te ontwikkelen voor pedometrische kartering die gebruikt kan worden om de kloof te dichten tussen de mechanistische pedometrische- en de traditionele benaderingen. De thesis beslaat zeven methodologische benaderingen van bodemkartering: bemonstering, voorbewerking, foto-interpretatie, interpolatie, visualisatie, organisatie en kwaliteitscontrole. Deze hoofdstukken zijn ingediend als wetenschappelijke artikelen in internationale tijdschriften.

BEMONSTERING: Dit hoofdstuk behandelt de verspreiding van waarnemingen in de data- en geografische ruimte als sleutel tot de optimalisatie van bemonstering voor de ruimtelijke voorspelling tijdens de correlatie met secondaire kaarten. Ofschoon deze kaarten veel gebruikt worden voor de kartering van bodemvariabelen worden de problemen die te maken hebben met het ontwerp van een bemonsterings schema zelden onderzocht. Wanneer de gegeneraliseerde kleinste kwadraten wordt toegepast hangt de totale voorspellingsfout af van de verspreiding van de observatie punten over de data- en de geografische ruimte. Het verdelen van bemonsteringspunten over de dataruimte in proportie tot de verdeling van de voorspeller (*equal range* stratificatie of ER ontwerp) wordt voorgesteld als een verstandige bemonsterings strategie wanneer het regressie model tussen de bodem en de ondersteunende variabelen niet bekend is. Om dit concept te illustreren werd een een gebied van 50×50 km in centraal Kroatië met 100 waarnemingen gebruikt. Het gebied werd herbemonsterd tot 25 databestanden met verschillende experimentele ontwerpen: ER en twee responsie-oppervlak ontwerpen (minmax en D2). Het ER gaf dezelfde voorspellingsfout als het minmax ontwerp, wat aangeeft dat het een goed compromis is tussen nauwkeurige schatting van het model en de minimalisatie van de ruimtelijke autocorrelatie van de restwaarden. Ook werd het ER ontwerp uitgebreid naar een multi-variatie studie. Vier voorspellers (hoogte, temperatuur, vochtigheidsindex en NDVI) werden getransformeerd tot hoofdcomponenten. De bemonsteringspunten werden vervolgens toegewezen aan

de componenten in verhouding tot de variantie die bepaald werd door de Hoofdcomponenten Analyse en volgens het ER ontwerp.

VOORBEWERKING: De kwaliteit van de toepassingen van de terrein analyse hangt direct af van de kwaliteit van de digitale hoogte modellen (DEMs) en de hiervan afgeleide data. Drie benaderingen voor reductie van fouten in DEMs en afgeleide data werden beschreven: (a) het gebruik van empirische kennis, b.v. aanpassing van de hoogte d.m.v. mediale assen en het drainage netwerk; (b) data filtreren en (c) foutenvoortplanting. Data filtreren werd gebruikt om onregelmatige waarden te vervangen of om uitschieters te reduceren door middel van de ruimtelijke correlatie struktuur en de waarschijnlijkheid van het overschrijden van een waarde zoals geschat door naburige waarden. Bij de foutenvoortplanting benadering worden de fouten teruggebracht door berekening van de waarden van verschillende realisaties. Deze methode werd getest in een proefgebied van 3.8×3.8 km waarin twee landschappen voorkomen: heuvelland en een vlakte met terrassen. Het gedeelte aan artefacten (padi-terrassen) in de onbehandelde DEM was 17.3%. Na toevoeging van mediale assen, het filtreren van uitschieters, en de aanpassing van de hoogten van het drainagepatroon, werd het aandeel van de padi-terrassen teruggebracht tot 2.2%. Uitschieters bestreken over het gehele gebied niet meer dan 2%. Zowel filtreren als de foutenvoortplanting benadering resulteerden in een wat gelijkmatiger kaartbeeld van de terrein parameters. Het voordeel van filtering van uitschieters is dat het gebruik maakt van de structuur van de ruimtelijke afhankelijkheid. Het voordeel van de foutenvoortplanting benadering is dat het gemakkelijker geautomatiseerd kan worden.Vermindering van de fouten verbeterde de kartering van de landschap facetten (classificatie) en van de dikte van de solum (regressie). De nauwkeurigheid van de classificatie nam toe van 51.3% tot 72% en de R^2 van het regressie model voor de voorspelling van de solumdikte verbeterde van 0.27 tot 0.40.

FOTO-INTERPRETATIE: Een methode wordt beschreven die de handmatige omlijning van landvormen door middel van foto-interpretatie verhoogt voor het karteren van een groter gebied. Traditionele API- kaarten met een geo-pedologische legenda van 21 klassen werden vervaardigd voor zes testgebieden met een totale oppervlakte van 11 sq.km in de Baranja regio van oost Kroatië. Negen terrein parameters werden ontleend aan een digitaal hoogte model, (diepte tot het grondwater, terrein kromming, karakter van de helling, *viewshed*, stroomgebied, vochtigheids index, sediment transport index, en de afstand tot de dichtstbijzijnde waterloop). Deze werden gebruikt om de luchtfoto-interpretatie to extrapoleren over het gehele gebied (1062 sq.km). De nauwkeurigheid van de classificatie werd vastgesteld door de fouten matrix die berekend werd door de vergelijking van de hele API kaarten en bemonsteringspunten met de resultaten van de classificatie. De eerste resultaten berekend door een hoogste aannemelijkheid (*maximum-likelyhood*) classificatie waren een reproduceerbaarheid van 58.2% voor heuvelland, 39.1% voor vlakte en 45.3% voor het gehele gebied. Zes klassen in de vlakte waren verantwoordelijk voor een groot aandeel foute classificaties. Dit was het gevolg van een onvoldoende gedetailleerd digitaal hoogte-model en het complexe karakter van de landvormen (kronkelwaarden en aktieve rivierlopen), die niet verklaard kunnen worden door alleen gebruik te maken van de terrein parameters.

Door gebruik te maken van een vereenvoudigde legenda van 15 klassen voor het onderzoeksgebied verbeterde de reproduceerbaarheid tot 65.8% voor vlakte, 58.2% voor heuvelland en tot 63.4% voor het gehele gebied. Na vereenvoudiging van de legenda tot 15 klassen en door gebruik te maken van de iteratieve selectie (3 stappen) van het oefenbestand van de bemonsteringspunten werd de classificatie verhoogd tot 97.6% heuvelland, 86.7% vlakte en 90.2% voor het hele gebied. De afgeleide classificatie was in staat om kleine details te laten zien die niet konden worden waargenomen m.b.v. luchtfoto-interpretatie. Het aantal handmatige luchtfoto-interpretaties werd teruggebracht van 84 naar 6. De methodologie kan worden toegepast door bodemkarteerders om bestaande kaarten te verfijnen en voor verbetering of vervanging van de luchtfoto-interpretatie voor nieuwe kartering.

INTERPOLATIE: In dit hoofdstuk wordt een methodologische benadering beschreven voor ruimtelijke voorspellling gebaseerd op regressie-kriging, en wordt deze vergeleken met gewone kriging en normale regressie. De gegevens werden eerst omgezet door logit transformatie voor doelvariabelen en factor analyse voor de continue voorspellers (secondaire kaarten). De doel variabelen werden vervolgens ingepast door middel van stapsgewijze regressie en restwaarden werden geïnterpoleerd door middel van kriging. Een generieke visualisatie methode wordt gebruikt om gelijktijdig de voorspellingen en de hiermee geassocieerde onzekerheid te visualiseren. Deze benadering werd getest voor 135 waarnemingspunten van de nationale gegevensbank van Kroatië en verdeeld in 100 interpolatie- en 35 validatie punten. Drie doel variabelen (organische stof, pH van de bovengrond en dikte van de bovengrond) werden voorspeld op basis van zes relief parameters en negen bodemkarteringseenheden. De voorspellings efficientie werd geevalueerd door gebruik te maken van de gemiddelde fout en de gemiddelde gekwadrateerde fout op de validatie punten. De resultaten tonen aan dat de voorgestelde structuur de efficiëntie van de voorspellingen verbeterd. Bovendien garandeert het de normaliteit van de restwaarden en dwong het de voorspelbare waarden binnen de fysieke grenzen van de variabele te blijven. Voor organische stof werd een lagere wortel van de gestandaardiseerde gemiddelde gekwadrateerde voorspelfout bereikt dan met gewone kriging (53.3% versus 66.5%). Dit was ook het geval voor de dikte van de bovengrond (66.5% versus 83.3%), waar ook een kleinere afwijking werd gevonden dan met met gewone kriging (0.15 cm versus 0.69 cm). De voorspelbaarheid van de pH was moeilijk met alle drie methoden. Deze benadering biedt de mogelijkheid om een samengesteld algoritme te ontwikkelen dat kan worden toegepast in een GIS om gegevens van bodemobservatie punten te interpoleren van bestaande databestanden.

VISUALISATIE: Een methode om meervoudige lidmaatschappen (multiple memberships) kaarten te visualiseren, nl. kleuren meng (*Colur Mixture* of CM) methode, wordt beschreven, en vergeleken met twee alternatieve technieken: defuzzifikatie en pixel menging (*Pixel Mixture*). Zes landvorm parameters werden gebruikt om de de landvorm klassen te definieren met gebruik making van de *fuzzy k-means* classificatie. De continue categorische kaart is gebaseerd op GIS berekeningen met kleuren, waarbij de kleurwaarden (Hue) verondersteld worden om de taxonomische ruimte van de variabelen te representeren. Eerst werden de centrale punten van de negen landvormen getransformeerd van de multivariate naar de twee-dimensionale data-ruimte door factor

analyse en vervolgens geprojecteerd op het nl. Kleurwaarde-Verzadiging-Intensiviteit (*Hue Saturation Intensity* of HSI) kleurenschema. De taxonomische waarde werd uitgedrukt met de kleurwaarde en de verwarring met verzadiging. De laatste waarde werd vervangen door 'witheid' om de visualisatie te verbeteren. Klassen die dichter bij elkaar lagen in de data ruimte werden samengevoegd in gelijkwaardige generieke kleuren. De CM techniek beperkt de afgeleide gemengde kleurenkaart tot zeven generieke kleurwaarden, onafhankelijk van het totaal aantal klassen, en vormt de basis voor de automatische generalisatie. De index voor verwarring (*confusion index*) die werd afgeleid van de gemengde kleurenkaart werd gebruikt om primaire grenzen te bepalen en om gebieden met hogere taxonomische onnauwkeurigheid te lokaliseren.

ORGANISATIE: In dit hoofdstuk worden basis concepten, handelingen en de struktuur van een raster gebaseerd bodem-informatie systeem vergeleken met een traditioneel polygon gebaseerd bodeminformatie systeem (*Soil Information System* of SIS). Dit wordt geïllustreerd door een studie van een gebied van 3.8×3.8 km in oostelijk Kroatië. De ruimtelijke eenheid in dit systeem is een raster cel en alle GIS kaarten werden teruggebracht tot dezelfde resolutie, namelijk 25 m. Bodem-variabelen werden gemodelleerd met behulp van een gemengd model van ruimtelijke variatie zodat zowel discrete als continue variatie mogelijk was. Het bodeminformatie systeem (SIS) bevatte 21 bestaande kaarten (luchtfoto interpretatie, terrain parameters en satellite beelden), alsmede zes kaarten van bodem variabelen (bodemdikte, aanwezigheid van een mollic, calcic of gleyic horizon, dikte van de bovengrond en percentage silt van deze laag), alsmede zes afgeleide kaarten van bodem typen. Iedere bodem-variabele werd geïnterpoleerd door middel van regressie-kriging. De geïnterpoleerde kaarten werden vervolgens geclassificeerd door gebruik te maken van een continue classificator (*fuzzy k-means*) om lidmaatschapskaarten te vervaardigen. Deze kaarten werden vervolgens gebruikt om de geschiktheid voor tarwe produktie te evalueren op een schaal van 0–1. Het bleek dat de luchtfoto-interpretatie kaart beter in staat was om de aangegeven bodemvariabelen te voorspellen dan de terrein parameters en satellietbeelden. Een vergelijking gebaseerd op de aanpassingsmaat en de thematische verwarring toonde aan dat het raster gebaseerde SIS over het algemeen beter overeenkomt, met meer detail geeft en een hogere mate van betrouwbaarheid heeft dan het traditionele polygon-gebaseerde SIS.

KWALITEITSCONTROLE: Een methodologie om de kwaliteit en de geschiktheid van een nationale bodeminventarisatie te bepalen en om haar bruikbaarheid te evalueren, worden beschreven in dit hoofdstuk. Daarvoor werden gebruikt: zes kaartbladen schaal 1:50 K (uit een totaal van 180), drie proefopnamen (elk met een grootte van 4×4 km) en tien complete profielbeschrijvingen van de hoofdlandschappelijke regio's van Kroatië. Deze gegevens werden gebruikt om de effectieve kaartschaal te bepalen, de nauwkeurigheid van de kaartlegenda's te controleren en de thematische nauwkeurigheid van de bodemobservatie punten te evalueren. Ook werden bestaande digitale bodeminformatie bestanden geevalueerd op hun thematische zuiverheid en op kleigehalte, pH en organische stof getoetst. Nieuwe methoden werden ontwikkeld en getest om de ruimtelijke nauwkeurigheid van de bodemgrenzen en de thematische overlap tussen kaarteenheden te onderzoeken. In het proefgebied kwamen de gemid-

delde grootte van de polygon en de positie nauwkeurigheid van de bodemgrenzen (ongeveer ±40 m) overeen met de 1:150 K kaartschaal, terwijl de correlatie dichtheid correspondeerde met de 1:250 K kaartschaal. Kaarteenheden waren heterogeen met een gemiddelde relatieve variatie van 17% binnen de eenheden en een gemiddelde thematische overlap van 66% tussen geografisch aangrenzende eenheden. Er is een groot verschil tussen de originele legenda en de proefbemonstering ter validatie als het gaat om de taxonomische classificatie, maar dit is minder het geval als de klassen worden gegroepeerd op basis van overeenkomstigheid. De voornaamste gebruiksproblemen betroffen samengestelde kaarteenheden, gebrek aan specifieke interpretaties die overeenkwamen met de behoeften van de gebruiker en het gebrek aan duidelijk legale wetgeving op het gebied van landeigendom en het -gebruik.

De algemene conclusie is dat de voorgestelde pedometrische kaarterings methodologie bodemkartering verbetert omdat het de bodemkaarten meer objectief, gedetailleerder en meer vergelijkbaar maakt voor integratie met andere ruimtelijke geodata. Er is verder geen noodzaak om nog gebruik te maken van het concept van kaarteenheden of de polygon-gebaseerde bodemkaarten. Daar tegenover staat dat inplaats van het afschaffen van de luchtfoto-interpretatie, de bodem classificatie en de empirische kennis aangaande de bodem, deze methoden succesvol geïntegreerd kunnen worden met pedometrische technieken.

Sažetak

Hengl, T. 2003. Pedometrijsko kartiranje: prevladavanje razlika izmedju tradicionalnog i pedometrijskog pristupa. Doktorska teza, Sveučilište u Wageningenu.

Posljednjih godina, digitalno kartiranje tala doživjelo je ubrzani razvitak novih i ekonomičnih metoda, uglavnom zbog veće dostupnosti tzv. pomoćnih karata (parametri reljefa, sateliski snimci itd.). Glavni cilj ovog istraživanja bio je razviti metodologiju za pedometrijsko kartiranje, koje se može koristiti za prevladavanje razlika izmedju tradicionalnih i pedometrijskih tehnika. Teza pokriva sedam metodoloških područja kartiranja tala: uzorkovanje, predobradu, fotointerpretaciju, interpolaciju, vizualizaciju, organizaciju i kontrolu kvalitete.

UZORKOVANJE: U ovom poglavlju razmatrana su načela razmještaja observacija u tematskom i geografskom prostoru kao ključ za optimizaciju odabira uzorkovanja za potrebe prostorne predikcije temeljene na korelaciji sa pomoćnim kartama. Iako je uporaba pomoćnih karata u kartiranju tala uobičajena praksa, problemi vezani uz dizajn uzorkovanja često se zanemaruju. Ukupna pogreška predikcije, u slučaju generalizirane metode najmanjih kvadrata, ovisi o položaju točaka u oba prostora — tematskom i geografskom. Ujednačeni razmještaj točaka, proporcionalno distribuciji prediktora (tzv. *Equal range* stratifikacija ili ER dizajn), predložen je kao 'najsigurnija' strategija uzorkovanja kada je regresijski model izmedju pedoloških i pomoćnih varijabli nepoznat. Postojećih uzorak od 100 pedoloških obzervacija na 50×50 km velikom području u središnjoj Hrvatskoj korišten je za ilustraciju ovih načela. Uzorak je bio podijeljen u podskupove od 25 obzervacija koristeći različite eksperimentalne dizajne: ER, te dva regresijska dizajna (minmax i D2). Dizajni su usporedjeni po uspjehu u predikciji sadržaja organske tvari (univarijatni primjer) uz pomoć nadmorske visine, koristeći ukupnu pogrešku predikcije kao kriterij procjene. ER dizajn postigao je sličnu ukupnu pogrešku kao i minmax dizajn, pa se može zaključiti da ER predstavlja dobar kompromis izmedju točne procjene modela i minimizacije prostorne autokorelacije reziduala. ER dizajn je takodjer proširen na multivarijatni slučaj. Četiri prediktora (nadmorska visina, temperatura, indeks vlažnosti i NDVI) prvo su transformirani u standardizirane komponente (*principal components*). Zatim su točke uzorkovanja rasporedjene po komponentama i to proporcionalno varijanci objašnjenoj u komponentnoj analizi, te propocionalno distribuciji prediktora.

PREDOBRADA: Kvaliteta digitalnih modela reljefa (DMR) i reljefnih parametara neposredno utječe na kvalitetu pripadajućih aplikacija. Opisana su tri pristupa redukciji pogrešaka u DMR-ima i reljefnim parametrima: (a) rabeći empirijsko znanje, npr. korekcija DMR-a korištenjem linija reljefnih lomova i karte tokova; (b) rabeći GIS operacije filtriranja, te (c) rabeći metodologiju propagacije pogrešaka. Operacije filtriranja korištene su za uklanjanje grubih grešaka tj. redukciju outliers-a. Temelje se na strukturi prostorne autokorelacije, te usporedbi promatranih vrijednosti i vrijednosti procijenjenih uz pomoć susjednih piksela. Metodom propagacije pogrešaka, pogreške se smanjuju računanjem prosječnih vrijednosti iz nekoliko simulacija. Metode su testirane na 3.8×3.8 km području istraživanja koje obuhvaća dva različita tipa reljefa: brdo i ravan sa terasama. Digitalizirane konture su interpolirane koristeći linearnu interpolaciju. Udio artefakata (tzv. *padi* terase) u nefiltriranom DMR-u bio je 17.3%. Nakon dodatka linija loma, filtriranja outliersa, te korekcije DMR-a prema karti tokova, udio *padi* terasa se smanjio na 2.2%. Preostale greške u parametrima reljefa, kao što su nedefinirani pikseli te lokalni outliers-i, reducirani su uporabom filtriranja sa iteracijama i propagacije pogrešaka. Udio outliers-a u parameterima reljefa nije prelazio 2% ukupne površine. Oba pristupa redukciji pogrešaka — filtriranje i propagacija pogrešaka, daju nešto generaliziranu sliku parametara reljefa. Prednost filtriranja outliers-a je da metoda direktno rabi strukturu prostorne autokorelacije. Prednost metode propagacije pogrešaka je da može biti lako automatizirana. Redukcija pogrešaka poboljšala je kartiranje geomorfoloških jedinica (klasifikacija), te dubine tla (regresija). Točnost klasifikacije povećala se sa 51.3% na 72%, dok je R^2 regresijskog modela za predikciju dubine tla narastao sa 0.27 na 0.40.

FOTO-INTERPRETACIJA: Ovo poglavlje nudi opis metode koja poboljšava manualnu foto-interpretaciju geomorfoloških jedinica pri kartiranju većih područja. Tradicionalne aero foto-interpretacijske (AFI) karte pripremljene su korištenjem geopedološke legende (21 klasa) za šest test područja u Baranji, istočna Hrvatska, ukupne površine 111 km^2. Devet parametara reljefa izlučenih iz DMR-a (dubina podzemne vode, nagib terena, horizontalna kurvatura, vertikalna kurvatura, sjenčanje, drenažna površina, indeks vlažnosti, indeks transporta sedimenta i udaljenost do najbliže vodene površine), korišteni su za ekstrapolaciju foto-interpretacije na cijelom području istraživanja (1062 km^2). Točnost klasifikacije procijenjena ili odredjena je uporabom matrice grešaka (*error matrix*), izračunate usporedbom svih AFI karata i točkastih uzoraka sa rezultatima klasifikacije na svim test područjima. Prvi rezultati klasifikacije, metodom maksimalne uvjetne vjerojatnosti (*maximum-likelihood*), dali su 58.2% (brdo), 39.1% (ravan), te 45.3% (cijelo područje) podudaranja sa test područjima. Šest klasa u ravnici prouzročile su velike pogreške u klasifikaciji, vjerojatno zbog nedovoljno detaljnog DMR-a te kompleksne naravi geomorfoloških klasa (riječni obalni kompleksi, rubovi rijeka, aktivni kanali), koji ne mogu biti objašnjeni samo uz pomoć parametara reljefa. Pomoću simplificirane legende sa 15 klasa, točnost klasifikacije povećala se na 65.8% (ravan), 58.2% (brdo) i 63.4% (cijelo područje) za sve AFI karte. Nakon simplifikacije legende (15), te sa iterativnom (3) selekcijom točkastih uzoraka, klasifikacija je pokazala 97.6% (brdo), 86.7% (ravan), i 90.2% (cijelo područje) podudaranja sa test područjima. Vodjena (*supervised*) klasifikacija

pokazala je detalje koje nije bilo moguće izlučiti uz pomoć aero foto-interpretacije. Broj manualnih foto-interpretacija, koje je trebalo pripremiti, smanjen je sa 84 na 6. Ova metodologija može koristiti pedo-kartografskim timovima za korekciju i nadopunu postojećih karata te za poboljšanje ili zamjenu API-ja u novim pedo-kartografskim projektima.

INTERPOLACIJA: Metodološka shema za prostornu predikciju temeljena na regresijskom-krigingu opisana je i usporedjena sa ordinarnim krigingom i čistom regresijom. Podaci su prvo transformirani korištenjem *logit* transformacije za ciljne varijable te faktor analize za kontinuirane prediktore (pomoćne karte). Ciljne varijable su zatim modelirane korištenjem *step-wise* regresije, a reziduali su interpolirani ordinarnim krigingom. Generična metoda vizualizacije razvijena je za simultani prikaz vrijednosti predikcije i pripadajuće nepouzdanosti. Interpolacijski algoritam testiran je koristeći 135 pedoloških profila iz Baze tala Republike Hrvatske, podijeljenih u interpolacijske (100) i kontrolne točke (35). Tri ciljne varijable: organska tvar, pH u oraničnom horizontu i debljina oraničnog horizonta bile su interpolirane koristeći šest parametera reljefa te devet pedo-kartografskih jedinica. Točnost predikcije procijenjena je uz pomoć srednje pogreške (ME) i korijena srednjeg kvadratnog odstupanja (RMSE) izračunatih na kontrolnim točkama. Rezultati su pokazali da predloženi interpolacijski algoritam povećava uspješnost predikcije. Štoviše, algoritam je osigurao normalnost reziduala, te spriječio pojavu vrijednosti izvan fizičkog raspona varijabli. U slučaju predikcije organske tvari, interpolacijski algoritam postigao je manji standardizirani RMSE nego ordinarni kriging (53.3% *versus* 66.5%). U slučaju predikcije debljine oraničnog horizonta, postigao je manji standardizirani RMSE (66.5% *versus* 83.3%), te manji sistematski otklon (ME) nego ordinarni kriging (0.15 *versus* 0.69 cm). Sve tri metode nisu bile uspješne u predikciji pH u tlu . Ovaj interpolacijski algoritam otvara mogućnost razvitka integriranog algoritma, koji bi se mogao korisiti za GIS-temeljenu poluautomatsku interpolaciju podataka iz postojećih baza tala.

VIZUALIZACIJA: Opisana je metoda za vizualizaciju višestrukih pripadnosti (*memberships*), zvana "mješač boja" (*Colour Mixture* ili CM) i usporedjena sa alternativnim tehnikama: defuzifikacijom i "mješačem piksela" (*Pixel Mixture*). Šest parametara reljefa korišteno je za klasifikaciju geomorfoloških jedinica metodom vodjene *fuzzy k-means* klasifikacije. Kontinuirana kategorička karta proizvedena je uporabom GIS kalkulacija s bojama, gdje boja predstavlja taksonomski prostor odredjen prediktorima. Koordinate 9 taksonomskih centara (geomorfološke jedinice) su prvo tranformirane iz multivarijantnog u dvodimenzionalni atributni prostor, i zatim projicirane na tzv. HSI (*Hue Saturation Intensity*) kružnu paletu boja (*colourwheel*). Taksonomska dimenzija kodirana je tipom boje (*Hue*), a konfuzija klasifikacije zasićenošću (*Saturation*). Kako bi pospješili vizualnu impresiju nepouzdanosti (*uncertainty*), zasićenost je zamijenjena bijelom bojom, tj. 'izbjeljivanjem'. Klase koje su bile bliže u atributnom prostoru sjedinjene su u sličnu generičnu boju. CM tehnika ograničava izračunatu miješanu boju, neovisno o ukupnom broju klasa, na sedam generičkih boja, što konačno omogućuje automatsku generalizaciju klasa. Zasićenost izlučena iz karte miješanih boja, rabljena je za detekciju primarnih granica te prostornu lokaciju područja velike taksonomske konfuzije.

ORGANIZACIJA: Glavna načela, operacije i organizacijska struktura grid-temeljenog Sustava Informacija o Tlu (*Soil Information System* -SIS) usporedjeni su sa tradicionalnim poligonskim SIS-om. Ta usporedba ilustrirana je koristeći 3.8×3.8 km test područje u istočnoj Hrvatskoj. Ključni prostorni element u grid-temeljenom SIS-u je kvadrat (*grid cell*) i svi GIS slojevi bili su podešeni na istu terensku rezoluciju (25 m u ovom slučaju). Pedološke varijable modelirane su koristeći tzv. "miješani model prostorne varijabilnosti" (*mixed model of spatial variation*), koji omogućuje i diskretne i kontinuirane prijelaze u prostoru. U ovom test području, SIS se sastojao od 21 prediktora (foto-interpretacijska karta, parametri reljefa, te satelitski snimci), šest karata pedoloških varijabli (debljina tla, učestalost moličnog, kalcičnog i glejičnog horizonta, debljina oraničnog horizonta, te sadržaj praha), te šest karata pripadnosti (*memberships*) za svaki tip tla. Svaka pedološka varijabla interpolirana je metodom hibridne interpolacije (regresijski-kriging). Interpolirane karte su zatim klasificirane pomoću kontinuirane klasifikacije (*fuzzy k-means*), kako bi se dobile karte pripadnosti. Karte pripadnosti su zatim poslužile za izračun kontinuirane pogodnosti (0–1) za proizvodnju pšenice, kao primjer interpretacije koju ju moguće postići uz pomoć opisanog SIS-a. Foto-interpretacijska karta se pokazala kao nešto bolji prediktor ciljnih pedoloških varijabli nego parametri reljefa i satelitski snimci. Usporedba uspješnosti modeliranja (*goodness of fit*), te tematske konfuzije, pokazali su da grid-temeljeni SIS, u pravilu, postiže veću detaljnost i pouzdaniju predikciju nego tradicionalni (poligonski) SIS. Prednosti dizajniranog SIS-a, u usporedbi sa tradicionalnim kartiranjem tala, su: (1) završava kartom tipova tala, a ne kartom pedo-kartografskih jedinica; (2) sve pedološke varijable su kartirane kao kontinuirana polja visokog detalja; (3) pruža mjeru nepouzdanosti, kako za ulazne, tako za izlazne varijable; (4) i diskretni i kontinuirani prijelazi u prostoru su mogući; (5) originalne pedološke observacije i interpolacijski/klasifikacijski parametri pohranjeni su u posebnim tablicama, koje su dio SIS-a, tako da je moguće obnoviti izračunate karte. Nedostaci su: (1) sustav je računarski zahtjevan i traži puno memorije; (2) troškovi su veći (nabavka i obrada podataka) i (3) sustav je osjetljiv na kvalitetu ulaznih podataka.

KONTROLA KVALITETE: Opisana je metodologija za procjenu kvalitete i pogodnosti, te uporabljivosti nacionalne inventarizacije tala. Šest 1:50 K listova karata (od 185 ukupno), tri kontrolna kartiranja (svako veličine 4×4 km), te deset detaljnih pedoloških profila u glavnim reljefnim regijama u Hrvatskoj, korišteni su za procjenu efektivnog mjerila, točnosti legendi, te tematsku točnost točkastih opažanja. Takodjer je procijenjena tematska čistoća i kontrast postojeće Baze tala (2198 profila) i osnovne pedološke karte Republike Hrvatske (u mjerilu 1:300 K), koristeći podatke o sadržaju gline, pH i organskoj tvari. Razvijene su i testirane nove metode za procjenu prostorne točnosti pedoloških granica i tematskog preklapanja izmedju susjednih kartografskih jedinica. Razultati ove studije pokazali su da srednja veličina poligona, te prostorna točnost pedoloških granica (oko ±40 m) odgovaraju efektivnom mjerilu od 1:150 K, dok gustoća profila odgovara mjerilu od 1:250 K. Pedo-kartografske jedinice su heterogene sa srednjom relativnom varijacijom od 17% unutar jedinica, te sa srednjim preklapanjem od 66% izmedju susjednih poligona. Uočena je velika razlika izmedju originalnih legendi i kontrolnih karata na nivou taksonomskih klasa,

te znatno manja razlika nakon grupiranja klasa prema sličnosti. Ova Nacionalna inventarizacija pogodna je za nacionalna planiranja sitnog mjerila, ali ne i za detaljna mjerila. Glavni problemi uporabljivosti su postojanje heterogenih kartografskih jedinica, nedostatak specifičnih interpretacija potrebnih korisnicima, te neriješena pitanja vlasništva i uporabe.

Osnovni zaključak je da predložena metodologija pedometrijskog kartiranja pospješuje pedo-kartografsku praksu, čineći pedološke karte objektivnijima, detaljnijima, te kompatibilnijima za integraciju sa drugim okolišnim geo podacima. Štoviše, više nema potrebe za korištenjem koncepta pedo-kartografskih jedinica ili tradicionalnih poligonskih karata. Sa druge strane, umjesto napuštanja foto-interpretacije, klasifikacije tala ili empirijskog znanja o tlima, ove se metode mogu uspješno integrirati sa pedometrijskih tehnikama.

Curriculum vitae

Tomislav Hengl was born in Osijek, Croatia on January 15th, 1974. He completed gymnasium for Natural sciences and mathematics in 1992 and started his study of Forestry at Faculty of Forestry, University of Zagreb. In 1996, he was awarded with The Rectors price for the best student scientific work in the area of bio-technical sciences. After graduation he applied for scholarship for post-graduated study abroad given by Croatian Ministry of Science and technology. The scholarship was established towards the rebuilding of the University of J.J. Strossmayer located in Osijek, which was partially destroyed in the war. In 1997 he was chosen for scholarship, signed a contract and started working as a young scientist at the Faculty of Agriculture, University of J.J. Strossmayer in Osijek. He started his MSc study on ITC in September 1998. within course *"Geoinformation for sustainable soil resource management"* and graduated with distinction in February 2000. He immediately continued with the PhD study, within the same institute.
During the three and half years of his PhD research, he has been involved with production and publication of scientific materials, but also as an assistant lecturer for *"Digital Terrain Analysis"*, within the Soil Information Systems specialisation. He also finished the advanced courses: *"Presentation skills"*, *"Scientific writing"* and *"Advanced statistics"*, orgnized by ITC and WAU's Production Ecology and Environmental Conservation postgraduate school. He participated in two Pedometric and two Accuracy conferences, one World Congress of Soil Science and numerous workshops and meetings.
Upon completion of the PhD study, you can contact him at the address down-bellow:

Tomislav Hengl
AGIS centre
Faculty of Agriculture
Trg Sv. Trojstva 3, 31000 Osijek, Croatia
Tel: +385-31-224288
Fax: +385-31-207017

Email: hengl@pfos.hr
Email 2nd: hengl@itc.nl
Home Page: http://www.pfos.hr/∼hengl

List of ITC PhD students

1. **Akinyede,** 1990, Highway cost modelling and route selection using a geotechnical information system
2. **Pan He Ping,** 1990, 90-9003757-8, Spatial structure theory in machine vision and applications to structural and textural analysis of remotely sensed images
3. **Bocco Verdinelli, G.,** 1990, Gully erosion analysis using remote sensing and geographic information systems: a case study in Central Mexico
4. **Sharifi, M.,** 1991, Composite sampling optimization for DTM in the context of GIS
5. **Drummond, J.,** 1991, Determining and processing quality parameters in geographic information systems
6. **Groten, S.,** 1991, Satellite monitoring of agro-ecosystems in the Sahel
7. **Sharifi, A.,** 1991, 90-6164-074-1, Development of an appropriate resource information system to support agricultural management at farm enterprise level
8. **Zee, D. van der,** 1991, 90-6164-075-X, Recreation studied from above: Air photo interpretation as input into land evaluation for recreation
9. **Mannaerts, C.,** 1991, 90-6164-085-7, Assessment of the transferability of laboratory rainfall-runoff and rainfall - soil loss relationships to field and catchment scales: a study in the Cape Verde Islands
10. **Ze Shen Wang,** 1991, 90-393-0333-9, An expert system for cartographic symbol design
11. **Zhou Yunxian,** 1991, 90-6164-081-4, Application of Radon transforms to the processing of airborne geophysical data
12. **Zuviria, M. de,** 1992, 90-6164-077-6, Mapping agro-topoclimates by integrating topographic, meteorological and land ecological data in a geographic information system: a case study of the Lom Sak area, North Central Thailand
13. **Westen, C. van,** 1993, 90-6164-078-4, Application of Geographic Information Systems to landslide hazard zonation
14. **Shi Wenzhong,** 1994, 90-6164-099-7, Modelling positional and thematic uncertainties in integration of remote sensing and geographic information systems
15. **Javelosa, R.,** 1994, 90-6164-086-5, Active Quaternary environments in the Philippine mobile belt
16. **Lo King-Chang,** 1994, 90-9006526-1, High Quality Automatic DEM, Digital Elevation Model Generation from Multiple Imagery
17. **Wokabi, S.,** 1994, 90-6164-102-0, Quantified land evaluation for maize yield gap analysis at three sites on the eastern slope of Mt. Kenya
18. **Rodriguez, O.,** 1995, Land Use conflicts and planning strategies in urban fringes: a case study of Western Caracas, Venezuela
19. **Meer, F. van der,** 1995, 90-5485-385-9, Imaging spectrometry & the Ronda peridotites
20. **Kufoniyi, O.,** 1995, 90-6164-105-5, Spatial coincidence: automated database updating and data consistency in vector GIS
21. **Zambezi, P.,** 1995, Geochemistry of the Nkombwa Hill carbonatite complex of Isoka District, north-east Zambia, with special emphasis on economic minerals
22. **Woldai, T.,** 1995, The application of remote sensing to the study of the geology and structure of the Carboniferous in the Calañas area, pyrite belt, SW Spain
23. **Verweij, P.,** 1995, 90-6164-109-8, Spatial and temporal modelling of vegetation patterns: burning and grazing in the Paramo of Los Nevados National Park, Colombia

24. **Pohl, C.,** 1996, 90-6164-121-7, Geometric Aspects of Multisensor Image Fusion for Topographic Map Updating in the Humid Tropics
25. **Jiang Bin,** 1996, 90-6266-128-9, Fuzzy overlay analysis and visualization in GIS
26. **Metternicht, G.,** 1996, 90-6164-118-7, Detecting and monitoring land degradation features and processes in the Cochabamba Valleys, Bolivia. A synergistic approach
27. **Hoanh Chu Thai,** 1996, 90-6164-120-9, Development of a Computerized Aid to Integrated Land Use Planning (CAILUP) at regional level in irrigated areas: a case study for the Quan Lo Phung Hiep region in the Mekong Delta, Vietnam
28. **Roshannejad, A.,** 1996, 90-9009284-6, The management of spatio-temporal data in a national geographic information system
29. **Terlien, M.,** 1996, 90-6164-115-2, Modelling Spatial and Temporal Variations in Rainfall-Triggered Landslides: the integration of hydrologic models, slope stability models and GIS for the hazard zonation of rainfall-triggered landslides with examples from Manizales, Colombia
30. **Mahavir, J.,** 1996, 90-6164-117-9, Modelling settlement patterns for metropolitan regions: inputs from remote sensing
31. **Al-Amir, S.,** 1996, 90-6164-116-0, Modern spatial planning practice as supported by the multi-applicable tools of remote sensing and GIS: the Syrian case
32. **Pilouk, M.,** 1996, 90-6164-122-5, Integrated modelling for 3D GIS
33. **Duan Zengshan,** 1996, 90-6164-123-3, Optimization modelling of a river-aquifer system with technical interventions: a case study for the Huangshui river and the coastal aquifer, Shandong, China
34. **Man, W.H. de,** 1996, 90-9009-775-9, Surveys: informatie als norm: een verkenning van de institutionalisering van dorp - surveys in Thailand en op de Filippijnen
35. **Vekerdy, Z.,** 1996, 90-6164-119-5, GIS-based hydrological modelling of alluvial regions: using the example of the Kisaföld, Hungary
36. **Pereira, Luisa,** 1996, 90-407-1385-5, A Robust and Adaptive Matching Procedure for Automatic Modelling of Terrain Relief
37. **Fandino Lozano, M.,** 1996, 90-6164-129-2, A Framework of Ecological Evaluation oriented at the Establishment and Management of Protected Areas: a case study of the Santuario de Iguaque, Colombia
38. **Toxopeus, B.,** 1996, 90-6164-126-8, ISM: an Interactive Spatial and temporal Modelling system as a tool in ecosystem management: with two case studies : Cibodas biosphere reserve, West Java Indonesia: Amboseli biosphere reserve, Kajiado district, Central Southern Kenya
39. **Wang Yiman,** 1997, 90-6164-131-4, Satellite SAR imagery for topographic mapping of tidal flat areas in the Dutch Wadden Sea
40. **Asun Saldana-Lopez,** 1997, 90-6164-133-0, Complexity of soils and Soilscape patterns on the southern slopes of the Ayllon Range, central Spain: a GIS assisted modelling approach
41. **Ceccarelli, T.,** 1997, 90-6164-135-7, Towards a planning support system for communal areas in the Zambezi valley, Zimbabwe; a multi-criteria evaluation linking farm household analysis, land evaluation and geographic information systems
42. **Peng Wanning,** 1997, 90-6164-134-9, Automated generalization in GIS
43. **Lawas, C.,** 1997, 90-6164-137-3, The Resource Users' Knowledge, the neglected input in Land resource management: the case of the Kankanaey farmers in Benguet, Philippines
44. **Bijker, W.,** 1997, 90-6164-139-X, Radar for rain forest: A monitoring system for land cover Change in the Colombian Amazon
45. **Farshad, A.,** 1997, 90-6164-142-X, Analysis of integrated land and water management practices within different agricultural systems under semi-arid conditions of Iran and evaluation of their sustainability
46. **Orlic, B.,** 1997, 90-6164-140-3, Predicting subsurface conditions for geotechnical modelling
47. **Bishr, Y.,** 1997, 90-6164-141-1, Semantic Aspects of Interoperable GIS
48. **Zhang Xiangmin,** 1998, 90-6164-144-6, Coal fires in Northwest China: detection, monitoring and prediction using remote sensing data
49. **Gens, R.,** 1998, 90-6164-155-1, Quality assessment of SAR interferometric data
50. **Turkstra, J.,** 1998, 90-6164-147-0, Urban development and geographical information: spatial and temporal patterns of urban development and land values using integrated geo-data, Villaviciencia, Colombia
51. **Cassells, C.,** 1998, Thermal modelling of underground coal fires in northern China

52. **Naseri, M.,** 1998, 90-6164-195-0, Characterization of Salt-affected Soils for Modelling Sustainable Land Management in Semi-arid Environment: a case study in the Gorgan Region, Northeast, Iran
53. **Gorte, B.G.H.,** 1998, 90-6164-157-8, Probabilistic Segmentation of Remotely Sensed Images
54. **Tenalem Ayenew,** 1998, 90-6164-158-6, The hydrological system of the lake district basin, central main Ethiopian rift
55. **Wang Donggen,** 1998, 90-6864-551-7, Conjoint approaches to developing activity-based models
56. **Bastidas de Calderon, M.,** 1998, 90-6164-193-4, Environmental fragility and vulnerability of Amazonian landscapes and ecosystems in the middle Orinoco river basin, Venezuela
57. **Moameni, A.,** 1999, Soil quality changes under long-term wheat cultivation in the Marvdasht plain, South-Central Iran
58. **Groenigen, J.W. van,** 1999, 90-6164-156-X, Constrained optimisation of spatial sampling: a geostatistical approach
59. **Cheng Tao,** 1999, 90-6164-164-0, A process-oriented data model for fuzzy spatial objects
60. **Wolski, Piotr,** 1999, 90-6164-165-9, Application of reservoir modelling to hydrotopes identified by remote sensing
61. **Acharya, B.,** 1999, 90-6164-168-3, Forest biodiversity assessment: A spatial analysis of tree species diversity in Nepal
62. **Akbar Abkar, Ali,** 1999, 90-6164-169-1, Likelihood-based segmentation and classification of remotely sensed images
63. **Yanuariadi, T.,** 1999, 90-5808-082-X, Sustainable Land Allocation: GIS-based decision support for industrial forest plantation development in Indonesia
64. **Abu Bakr, Mohamed,** 1999, 90-6164-170-5, An Integrated Agro-Economic and Agro-Ecological Framework for Land Use Planning and Policy Analysis
65. **Eleveld, M.,** 1999, 90-6461-166-7, Exploring coastal morphodynamics of Ameland (The Netherlands) with remote sensing monitoring techniques and dynamic modelling in GIS
66. **Yang Hong,** 1999, 90-6164-172-1, Imaging Spectrometry for Hydrocarbon Microseepage
67. **Mainam, Félix,** 1999, 90-6164-179-9, Modelling soil erodibility in the semiarid zone of Cameroon
68. **Bakr, Mahmould,** 2000, 90-6164-176-4, A Stochastic Inverse-Management Approach to Groundwater Quality
69. **Zlatanova, Z.,** 2000, 90-6164-178-0, 3D GIS for Urban Development
70. **Ottichilo, Wilber K.,** 2000, 90-5808-197-4, Wildlife Dynamics: An Analysis of Change in the Masai Mara Ecosystem
71. **Kaymakci, Nuri,** 2000, 90-6164-181-0, Tectono-stratigraphical Evolution of the Cankori Basin (Central Anatolia, Turkey)
72. **Gonzalez, Rhodora,** 2000, 90-5808-246-6, Platforms and Terraces: Bridging participation and GIS in joint-learning for watershed management with the Ifugaos of the Philippines
73. **Schetselaar, Ernst,** 2000, 90-6164-180-2, Integrated analyses of granite-gneiss terrain from field and multisource remotely sensed data. A case study from the Canadian Shield
74. **Mesgari, Saadi,** 2000, 90-3651-511-4, Topological Cell-Tuple Structure for Three-Dimensional Spatial Data
75. **Bie, Cees A.J.M. de,** 2000, 90-5808-253-9, Comparative Performance Analysis of Agro-Ecosystems
76. **Khaemba, Wilson M.,** 2000, 90-5808-280-6, Spatial Statistics for Natural Resource Management
77. **Shrestha, Dhruba,** 2000, 90-6164-189-6, Aspects of erosion and sedimentation in the Nepalese Himalaya: highland-lowland relations
78. **Asadi Haroni, Hooshang,** 2000, 90-6164-185-3, The Zarshuran Gold Deposit Model Applied in a Mineral Exploration GIS in Iran
79. **Raza, Ale,** 2001, 90-3651-540-8, Object-oriented Temporal GIS for Urban Applications
80. **Farah, Hussein,** 2001, 90-5808-331-4, Estimation of regional evaporation under different weather conditions from satellite and meteorological data. A case study in the Naivasha Basin, Kenya
81. **Zheng, Ding,** 2001, 90-6164-190-X, A Neuro-Fuzzy Approach to Linguistic Knowledge Acquisition and Assessment in Spatial Decision Making
82. **Sahu, B. K.,** 2001, Aeromagnetics of continental areas flanking the Indian Ocean; with implications for geological correlation and Gondwana reassembly
83. **Alfestawi, Y.,** 2001, 90-6164-198-5, The structural, paleogeographical and hydrocarbon systems analysis of the Ghadamis and Murzuq Basins, West Libya, with emphasis on their relation to the intervening Al Qarqaf Arch

84. **Liu, Xuehua,** 2001, 90-5808-496-5, Mapping and Modelling the Habitat of Giant Pandas in Foping Nature Reserve, China
85. **Oindo, Boniface Oluoch,** 2001, 90-5808-495-7, Spatial Patterns of Species Diversity in Kenya
86. **Carranza, Emmanuel John,** 2002, 90-6164-203-5, Geologically-Constrained Mineral Potential Mapping
87 **Rugege, Dennis,** 2002, 90-5808-584-8, Regional Analysis of Maize-Based Land Use Systems for Early Warning Applications
88 **Liu, Yaolin,** 2002, 90-5808-648-8, Categorical Database Generalization in GIS
89 **Ogao, Patrick Job,** 2002, 90-6164-206-X, Exploratory Visualization Of Temporal Geospatial Data Using Animation
90. **Abadi, Abdulbaset M.,** 2002, 906164-205-1, Tectonics of the Sirt Basin - Interferences from tectonic subsidence analysis, stress inversion and gravity modeling
91. **Geneletti, Davide,** 2002, ISSN 0169-4839, Ecological evaluation for environmental impact assessment
92. **Sedogo, Laurent G.,** 2002, ISBN 90-5808-751-4, Integration of Participatory Local and Regional Planning for Resources Management Using Remote Sensing and GIS
93. **Montoya, Ana Lorena,** 2002, ISBN 90-6164-2086, Urban Disaster Management: A Case Study of Earthquake Risk Assessment in Cartago, Costa Rica
94. **Ahmad, Mobin-ud-Din,** 2002, ISBN 90-5808-761-1, Estimation of net groundwater use in irrigated river basins using geo-information techniques: A case study in Rechna Doab, Pakistan
95. **Said, Mohammed Yahya,** 2003, ISBN 90-5808-794-8, Multiscale perspective of species richness in East Africa
96. **Schmidt, Karin S.,** 2003, ISBN 90-5808-830-8, Hyperspectral Remote Sensing of Vegetation Species Distribution in a Saltmarsh
97. **Binnqüist, Rosaura C. L.,** 2003, ISBN 9036519004, The Endurance of Mexican Amate Paper: Exploring Additional Dimensions to the Sustainable Development Concept
98. **Zhengdong, Huang,** 2003, ISBN 90-6164-211-6, Data Integration for Urban Transport Planning
99. **Jianquan, Cheng,** 2003, ISBN 90-6164-212-4, Modelling Spatial and Temporal Urban Growth
100. **Campos dos Santos, José Laurindo,** 2003, ISBN 90-6164-214-0, A Biodiversity Information System in an Open Data/Metadatabase Architecture