Geographical Information Modelling for Land Resource Survey

Promotoren: dr. ir. M. Molenaar
hoogleraar in de geo-informatica en de
ruimtelijke gegevensinwinning

dr. ir. A.K. Bregt
hoogleraar in de geo-informatiekunde met
bijzondere aandacht voor geografische
informatiesystemen

NN03201 , 2793

Geographical Information Modelling for Land Resource Survey

Sytze de Bruin

PROEFSCHRIFT

ter verkrijging van de graad van doctor
op gezag van de rector magnificus
van Wageningen Universiteit,
dr. C.M. Karssen,
in het openbaar te verdedigen
op dinsdag 30 mei 2000
des namiddags te vier uur in de Aula.

BIBLIOTHEEK

Cover design: Henk de Bruin

Sytze de Bruin

Geographical Information Modelling for Land Resource Survey
Thesis Wageningen University
with summary in Dutch
ISBN 90-5808-211-3

Printed by: Ponsen & Looijen bv, Wageningen

BIBLIOTHEEK
LANDBOUWUNIVERSITEIT
WAGENINGEN

ONDERZOEK, 2799

STELLINGEN

1. Het toenemende gebruik van geografische informatiesystemen maakt aanpassing van traditionele bodemkundige en agronomische karteringsactiviteiten zowel mogelijk als noodzakelijk (*dit proefschrift*).

2. Aangezien fuzziness een eigenschap is van de werkelijkheidsperceptie moet het tot uiting worden gebracht in het conceptuele model waarmee geografische verschijnselen beschreven worden. Onzekerheden die betrekking hebben op onnauwkeurigheden en/of fouten komen daarentegen voor in elke terreinbeschrijving, ongeacht het conceptuele gegevensmodel (*dit proefschrift*).

3. Het kwantificeren van onzekerheden in geografische databestanden dient te geschieden aan de hand van ruimtelijke modellen in plaats van met de nog veel gebruikte globale indices (*dit proefschrift*).

4. In discussies over het al dan niet fuzzy zijn van vegetatietypen in ruimte en tijd dient te worden bedacht dat onzekerheid geen eigenschap is van een landschap, maar een kenmerk van onze kennis en perceptie van dat landschap.

   Droesen, W.J. (1999) Spatial modelling and monitoring of natural landscapes (Thesis Wageningen University).

   Sanders, M.E.(1999) Remotely sensed hydrological isolation: a key factor predicting plant species distribution in fens (Thesis Wageningen University).

5. De opmars van desktop GIS-producten, die de koppeling van geografische databestanden en bijvoorbeeld tekstbestanden realiseren via paginageoriënteerde ingebedde objecten, onderstreept de kracht van de papieren kaart als metafoor voor de geografische werkelijkheid.

6. Een lokaal positioneringssysteem (LPS) is in de precisielandbouw meer op zijn plaats dan een globaal positioneringssysteem (GPS).

7. Een wetenschapsgebied waarin men medeonderzoekers ziet als rivalen in plaats van collega's is nodig toe aan een nieuw paradigma.

8. Het verwoorden van de stelligheid van een uitspraak met de term *met aan zekerheid grenzende waarschijnlijkheid* laat veel ruimte voor interpretatie, omdat deze term uitsluitend zekerheid uitsluit.

9. De eco-toeristenindustrie zou vliegreizigers moeten weigeren.

10. Universitaire internetsites geven vaak blijk van een grotere zorg voor vorm dan voor inhoud.

11. De salariëring, aard van werkzaamheden en het grotendeels ontbreken van verdere loopbaanmogelijkheden op de universiteit rechtvaardigen vervanging van de acroniemen *AIO* en *OIO* door *OTO* (Onderbetaald Tijdelijk Onderzoeker).

12. Het promotieonderzoek van een echtgenoot en vader vergt grote inzet van vrouw en kind.

Stellingen behorende bij het proefschrift
Geographical Information Modelling for Land Resource Survey
Sytze de Bruin, Wageningen, 30 mei 2000.

Acknowledgements

It is a great pleasure to thank the many persons who have contributed in one way or another to the completion of this dissertation. First of all, I gratefully acknowledge my family for their support, encouragement and patience. Particularly my wife Shirley, but also my children Stefan and Natalia and my father and mother have been deeply involved in this effort.

I am very grateful to my supervisors:- Martien Molenaar, for initiating the research and for his confidence and support along the way. The support and enthusiasm of Arnold Bregt, my second supervisor who became involved at a later stage of the project, have been very stimulating.

Willem Wielemaker and John Stuiver were largely responsible for the beginning of my involvement in Geographical Information Science. Willem arranged a study leave at the then Department of Land Surveying and Remote Sensing when I returned from abroad in May 1993. On that occasion, John Stuiver provided a tailor-made course which satisfied my desire to learn about GIS and photogrammetry. Thank you both! A few years later, some of the initial ideas were published in a paper (co-authored by Willem Wielemaker and Martien Molenaar), which forms the basis of Chapter 3 of this dissertation.

Special thanks go out to Alfred Stein for his contributions to Chapter 4, his prompt and constructive comments on several other parts of the manuscript and for co-ordinating the Methodology discussion group of the C.T. de Wit Graduate School Production Ecology. Alfred and my fellow participants in the discussion group are acknowledged for the many interesting meetings and for discussing my work.

I am grateful to Ben Gorte for his contribution to Chapter 5.

I would like to thank Elisabeth Addink (my room-mate), Johan Bouma, Jaap de Gruijter, Inakwu Odeh, Eric van Ranst, and several anonymous reviewers for their useful comments on parts of the manuscript.

The Consejería de Medio Ambiente of the Junta de Andalucía, Sevilla is acknowledged for providing the digital land cover data that was used for the case study reported in Chapter 5.

My colleagues at the Centre for Geo-Information are acknowledged for their collegiality, the technical and administrative assistance provided, and for their patience while queuing after one of my time-consuming print jobs. Thank you very much!

Finally, I am very grateful to my brother Henk who designed the cover of this book.

Sytze de Bruin
Wageningen, March 2000

Contents

1 Introduction

1.1 Background

For many years, land resource survey was regarded as the recognition and subsequent mapping of different types of soil, vegetation, rocks, landforms or other land resources (Webster and Oliver, 1990, p. 1). The introduction of computer techniques initially did not change this, as it merely resulted in manual cartographic tasks being replaced by automation. The capabilities of early spatial analysis systems that emerged along with the map-making tools went little further than raster overlaying and subsequent visualisation using crude line printer graphics (Burrough and McDonnell, 1998). The poor graphical quality of these prints prevented them from being accepted as cartographic products.

Pushed by technological developments and increased awareness of the importance of being able to manipulate large quantities of spatial information, geographical information systems (GIS) have become widely accepted in today's world (Burrough and McDonnell, 1998; Longley *et al.*, 1999). This has had, and will continue to have, major implications for land resources survey. No longer is the paper map, which previously dictated the form of spatial representation, the default data store and end-product of a survey. Geographic information theory provides surveying disciplines with a conceptual framework to formulate alternative and richer spatial representations that can be mapped onto data models provided by computer technologists (Molenaar, 1989, 1996; Raper, 1999). Furthermore, digital technology has improved the accessibility of ancillary data (e.g. digital elevation models, remotely sensed imagery, postcode areas) and enables their utilisation in target database production. (e.g. Molenaar and Janssen, 1994; Gorte and Stein, 1998; Goovaerts, 1999). Unfortunately, there are disciplinary gaps between the different fields of study involved, so that new opportunities are not yet fully exploited in land resource survey. This stresses the need for more comprehensive studies exploring the utility of new concepts and methods.

Another consequence of the common acceptance of GIS is that land resource databases are increasingly being used beyond disciplinary boundaries, for example, to support decision making (Goovaerts, 1997, 1999; Gorte, 1998; Eastman 1999). Likewise, they are used in combination with other data sets by environmental scientists engaged in modelling and monitoring physical processes on or near the earth's surface.

The greater distance between data producers and data consumers (Veregin, 1999) and integrated use of multiple data sets and physical response models (e.g. Heuvelink, 1993, 1998a) raise the issue of uncertainty.

Land resource databases are certainly not error free. Surveyors have to resort to sampling to obtain data on phenomena of interest. Exhaustively sampled data are usually only available in the form of non-exact, (weakly) correlated secondary data. Vague class definitions may contribute to further uncertainty. Although uncertainty modelling for spatial data has been the subject of much recent research (e.g. Foody *et al.*, 1992; Goodchild *et al.*, 1992; Altman, 1994; Hunter and Goodchild, 1995; Fisher, 1998; Worboys, 1998; Kyriakidis *et al.*, 1999), proposed methods and measures are only sparsely used in applied environmental research (Goovaerts, 1999). Additionally, two types of uncertainty (i.e. fuzziness and inaccuracy) are commonly confused in literature, although they differ in several key respects (Manton *et al.*, 1994; Fisher, 1996; Lark and Bolam, 1997).

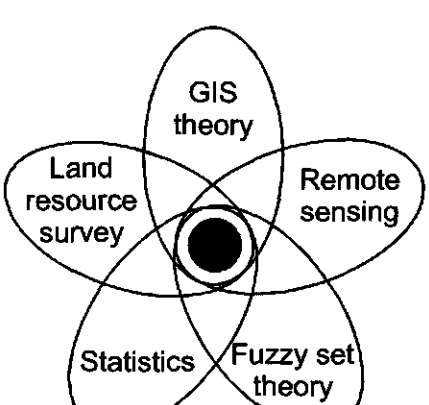


Figure 1.1 Research on the interface between five fields of study.

1.2 Aim and scope

As observed above, the increasing use of GIS has at least three major implications for land resources survey:

- Alternative models for spatial representation have become available;
- Increasingly, ancillary data can be used to support target database generation;
- There is greater need for uncertainty analysis.

However, owing to disciplinary gaps, the resulting opportunities and requirements are far from being fully adopted in practice. Against this background, the overall objective of this research is to explore and demonstrate the utility of new concepts and tools for improved land resource survey. This requires investigations on the interface between several fields of study, five of which are included in the current research (see Figure 1.1): land resource survey, geographic information theory, remote sensing, statistics, and fuzzy set theory. Capitalising on my own background in soil science and my colleagues'

experience in land cover mapping, the research concentrates on the survey of soil and land cover.

Even with these restrictions, the subject remains too broad for comprehensive coverage in a study of this size. Therefore, the shaded inner circle in Figure 1.1, indicating the scope of this study, is smaller than the complete area of overlap of the five ellipses. Its actual size is not intended to reflect the relative contribution of this research, though. Several choices had to be made to keep the subject within manageable proportions, the most important of which are listed below.

- The study focuses on representation of the terrain in a GIS database and on querying that database. It does not include, for example, dynamic process modelling in GIS;
- The research deals with data uncertainty rather than data quality. The latter also concerns fitness for purpose (Unwin, 1995; Veregin, 1999) and would require analysis of the use of data, for example in risk-based policy;
- The research does not deal with all aspects of uncertainty but focuses on fuzziness of class intensions and assessment of thematic accuracy. Their effect on the spatial extent of geographical features is also considered.
- Terrain descriptions are essentially two dimensional (2D), or 2.5D at most. The only way the third *spatial* dimension is considered is by treating it (elevation) as an attribute. Temporal aspects are captured using a snapshot approach (Peuquet and Duan, 1995), i.e. by time stamping a sequence of spatial state descriptions;
- Most concepts and tools are explored and demonstrated in either a soil survey or a land cover mapping context, but not both.

The overall objective was broken down into various partial goals that are being addressed in Chapters 2 to 7 as indicated below and detailed in the introductions to the respective chapters.

1.3 Outline of the thesis

The core of this thesis (Chapters 3-7) is based on a series of five papers, by myself as the principal author, that have been or will be published in international peer-reviewed journals. These chapters cover different concepts and tools for improved land resource survey from the perspective of GIS use. Each chapter is introduced separately by stating its partial research goals and the relation to other research in the field. They are preceded by a general introduction to spatial modelling concepts and tools that are relevant to land resources survey (Chapter 2). These are only briefly discussed in Chapter 2, as they are further explored and exemplified by case studies in Chapters 3 to 7.

Chapter 3 formulates and demonstrates a methodological framework that takes advantage of GIS capabilities to interactively formalise soil-landscape knowledge using stepwise image interpretation and inductive learning of soil-landscape relationships. It involves terrain description at successive levels of detail, information transfer between these levels, and explicit representation of expert decisions.

Chapter 4 describes a method to improve conventional soil-landscape modelling by representing fuzzy transition zones between soil-landscape units. The method uses fuzzy c-means clustering of attribute data derived from a digital elevation model and employs a new procedure for cluster validity evaluation.

Chapter 5 presents a probabilistic method to improve the accuracy of remotely sensed image classifications. First, an image is stratified using GIS-stored ancillary data. Next, *a priori* class probability estimates for each stratum are iteratively improved using intermediate classification results. The chapter also shows how posterior probability vectors can be used to represent local uncertainty in image classifications and in the results of subsequent analysis.

Chapter 6 introduces the concept of spatial uncertainty, i.e. joint uncertainty about a spatial phenomenon at several locations taken together. It explores the use of two geostatistical tools, i.e. collocated indicator co-kriging and stochastic simulation, to evaluate uncertainty in area estimates derived from classified remotely sensed imagery and sampled reference data.

Chapter 7 first explains the difference between membership grade and probability of membership and then exemplifies how these uncertainty measures can be combined to handle GIS queries expressed in verbal language. Such queries typically involve a mixture of uncertainties in the outcome of events that are governed by chance and in the meaning of linguistic terms.

Finally, Chapter 8 concludes the thesis with a summary of the main findings and suggestions for further research.

1.4 Location of the case studies

The spatial modelling tools and concepts are demonstrated by five case studies from a common study area located around the village of Alora in the province of Malaga, southern Spain (see Figure 1.2). The Alora region is within the Betic Cordillera, the most western of the European Alpine mountain ranges, and includes part of the drainage basin of the river Guadalhorce. The climate is dry Mediterranean with an average annual precipitation of 531 mm and a dry period of 4.5 months (De Leon *et al.*, 1989). There is great variation in geology, landscapes and soils within short distances, and a variety of crops are grown.

For the past nine years, Alora has provided the setting for a field training project of Wageningen University in which students and lecturers from several disciplines come together around the central theme of sustainable land use. Thanks to this project I could count on local expertise as well as free access to several relevant data sets, such as a digital elevation model, remotely sensed imagery, aerial photography and orthophotography. Hence the choice of area. Details of the study area and descriptions of the used data sets are provided in Chapters 3 to 7.

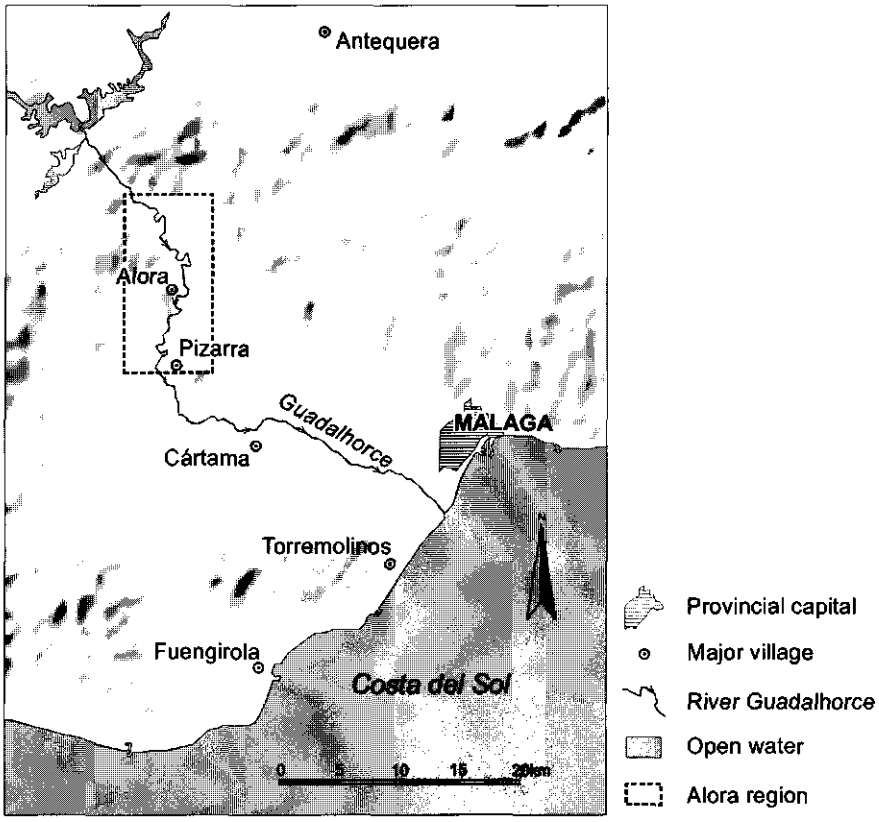
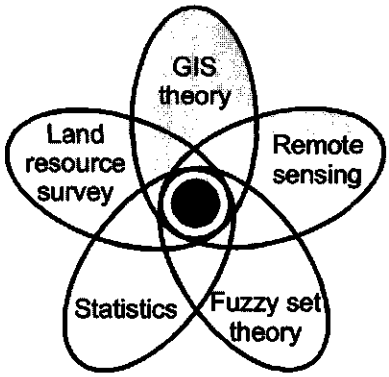


Figure 1.2 Location of the Alora region (indicated by the broken line) in the province of Malaga, southern Spain.


GIS
theory
Land
resource
survey
Remote
sensing
Statistics
Fuzzy set
theory

2 Spatial modelling concepts

2.1 Introduction

The purpose of this chapter is to introduce several spatial modelling concepts that are relevant to land resources survey. The concepts are only briefly discussed, as they are further explored and exemplified in the case studies presented in Chapters 3 to 7 of this thesis.

Acquisition of geo-information is always done with a particular view or model of real-world phenomena in mind. This view affects how geographic data is modelled in the computer and the way in which it can be used for further analysis. Therefore, I will start with a brief section on data models. Treatment of this subject is limited to the level of conceptual data modelling (Molenaar, 1996, 1998) and does not involve either logical data schemas or physical implementation of these on the computer. Next, there is a section on data acquisition and predictive mapping of land resources. The chapter ends with a section on uncertainty modelling. Frequently, reference is made to later chapters where more explanation is given and example applications are described.

2.2 Data modelling

2.2.1 Conceptual models of geographic phenomena

A terrain description is inevitably an abstraction, or, in other words, a model of the real terrain it represents. Until recently, two fundamentally different conceptual models were used for representing geographic phenomena: the *discrete object model*[1] and the *continuous field model*[2] (Burrough, 1996). The discrete object model views the world as being composed of well-defined spatial entities. A key feature of this view is that each entity is assigned to only one of a set of clearly distinct categories or classes. Each object has an identity, occupies space and has properties. Objects are homogeneous within their boundaries, at least with respect to some properties (Frank, 1996). Examples are buildings, runways, farm lots, railways, etc. The continuous field model, on the other

[1] Also known as *crisp object* or *exact object model.*

[2] Also known as *surface model.*

hand, views geographic space as a – not necessarily smooth – continuum. It assumes that every point in space can be characterised in terms of a set of attribute values measured at geometric coordinates in a Euclidean space (Burrough and Frank, 1995). Examples are elevation and slope in an undulating landscape, concentration of algal chlorophyll in surface water, green leaf area index in an agricultural field, etc.

These two data models are too restrictive when it comes to modelling phenomena that are conceived as nameable objects but without the object classes having clear-cut boundaries. Zadeh (1965) first introduced the concept of fuzzy sets to deal with classes that do not have sharply defined boundaries. Fuzzy sets are characterised by membership functions that assign grades of membership in the real interval [0, 1] to elements. The membership grade expresses the degree to which an element is similar to the concept represented by a fuzzy set. Membership in a fuzzy set is thus not a matter of yes or no but of a varying degree. Consequently, an element can partially belong to multiple fuzzy sets. Fuzzy set theory allows geographic phenomena to be modelled as objects whose boundaries are not exactly definable. Geographic space is then seen to be composed of elementary units that belong to classes having diffuse boundaries in attribute space. Presence of spatial correlation among these units – in fact a necessity for any kind of mapping (Journel, 1996) – ensures that they form spatially contiguous regions (Burrough *et al.*, 1997). Each of these fuzzily connected regions represents an *object with indeterminate boundaries* or *fuzzy object*. The spatial extent of fuzzy objects can be determined by evaluating class membership functions in combination with adjacency relationships between geographic elements (Molenaar, 1998).

Examples of phenomena that have been modelled using fuzzy set theory are: climatic regions (McBratney and Moore, 1985); polluted areas (Hendricks-Franssen *et al.*, 1997), soils (Burrough *et al.*, 1997), soil-landscapes (De Bruin and Stein, 1998; see Chapter 4), vegetation (Foody, 1992; Droesen, 1999), and coastal geomorphology (Cheng, 1999).

2.2.2 GIS data structures

The nature of digital computers imposes that computerised geographic data are always stored in a discretised form. There are two basic data structures to store geographic data in the computer: the vector structure and the raster structure. A third structure, based on object-orientated programming languages (see Burrough and McDonnell, 1998, pp. 72-74) is not treated here separately, because in essence it recurs to the basic structures. Besides, to date the implementation of object-oriented databases in GIS has been limited (Burrough and McDonnell, 1998).

Figure 2.1 Point, line and polygon of the vector structure.

The vector structure uses points, lines and polygons to describe geographic phenomena (see Figure 2.1). The geometry of these elementary units is explicitly and precisely defined in the database. Points are geometrically represented by an (x, y) coordinate pair, lines consist of a series of points connected by edges, and polygons consist of one or more lines that together form a closed loop. The thematic attribute data of a vector unit reside in one or more related records.

The vector structure is very suited to represent discrete geographic objects. It also lends itself to represent continuous fields and fuzzy objects (see Figure 2.2). For example, a triangular irregular network (TIN) based on a Delauney triangulation of irregularly spaced points provides a vector data model of a continuous field (Burrough and McDonnell 1998).

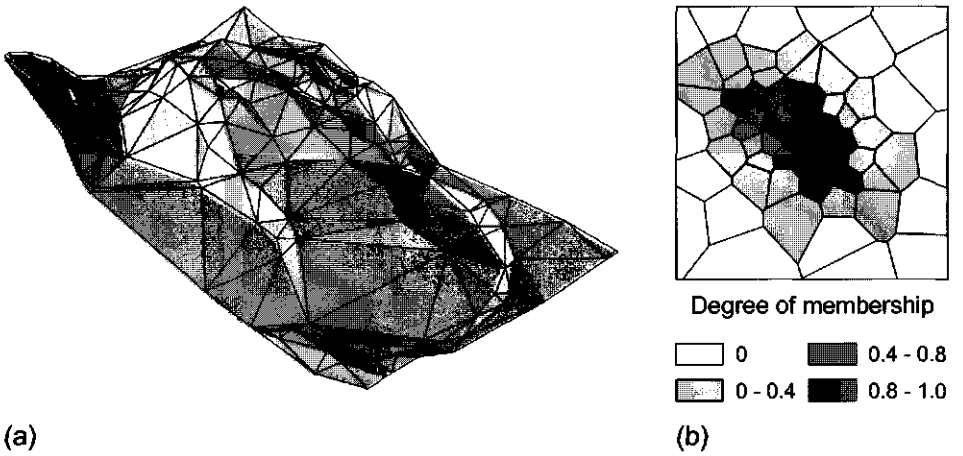


Figure 2.2 Vector representations of a continuous field (a) and a fuzzy object (b). Figure 2.2(a) is a perspective view of a TIN-based digital elevation model. Figure 2.2(b) shows Thiessen polygons that are shaded according to the degree to which they are part of the fuzzy object.

The raster data structure comprises a grid of n rows × m columns. Each element of the grid holds an attribute value or a pointer to a record storing multiple attribute data of a geographic position. The raster structure has two possible interpretations (Figure 2.3): the point or lattice interpretation and the cell interpretation (ESRI, 1994a; Fisher, 1997; Molenaar, 1998). The former represents a surface using an array of mesh points at the intersections of regularly spaced grid lines. Each point contains an attribute value (e.g. elevation). Attribute values for locations between mesh points can be approximated by interpolation based on neighbouring points (Figure 2.3a). The cell interpretation corresponds to a regular tessellation of the surface. Each cell represents a rectangular area using a constant attribute value (Figure 2.3b).

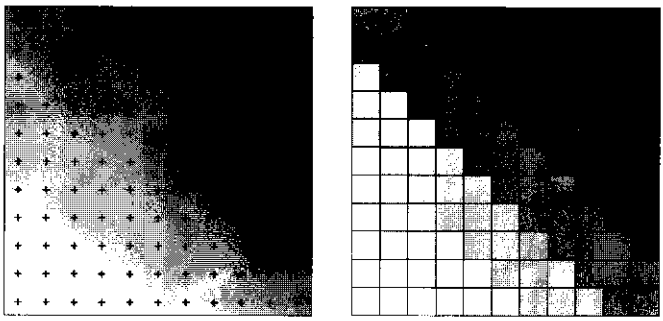


Figure 2.3 Point interpretation (a) and cell interpretation (b) of the raster structure.

The spatial resolution of a raster refers to the step sizes in x (column) and y (row) directions. In the case of a point raster these define the distances between mesh points in the terrain. In a cell raster they define the size of the sides of the cell. Given the coordinates of the raster origin, its spatial resolution and information on projection, the geographic position of a raster element is referred to implicitly by means of the row and column indices.

Like the vector structure, the raster structure is capable of representing all three conceptual models described in Section 2.2.1. Figure 2.3 shows raster representations of a continuous field. Figure 2.4 shows examples of cell rasters representing a discrete object and a fuzzy object.

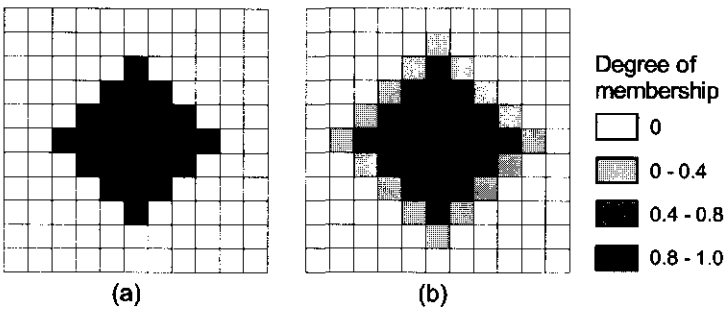


Figure 2.4 Cell raster representations of a discrete object (a) and a fuzzy object (b).

The choice of using either the raster structure or the vector structure to model geographic information used to be an important conceptual and technical issue. At present, the data structures are no longer seen as mutually exclusive alternatives (Unwin, 1995; Burrough and McDonnell, 1998). Molenaar (1998) showed that the vector and raster structures have similar expressive powers. Table 2.1 summarises how both structures enable representation of all three conceptual models of geographic phenomena. In addition, earlier problems regarding the quality of graphical output and data storage requirements of raster systems have largely been overcome with today's computer

hardware and software. Many GIS now support both structures and allow for conversion between them. Yet, if a GIS analysis involves multiple data sets these are usually required to be in the same structural form (Burrough and McDonnell, 1998).

Table 2.1 Possible implementations of the three conceptual models of geographic phenomena in vector structure and raster structure.

Conceptual model	Vector structure	Raster structure
Continuous field	Create a TIN by means of a Delauney triangulation of irregularly spaced sample points	Discretise field into point raster or cell raster; assign attribute values to raster elements
Exact object	Assign object identifier to geometrical element(s) belonging to the object; this is equivalent to relating geometrical element(s) to the object via *part of* links that are valued either zero or one	Assign object identifier to raster cells belonging to the object; this is equivalent to relating raster cells to the object via *part of* links that are valued either zero or one
Fuzzy object	Relate geometrical elements to fuzzy object via *part of* links in [0, 1] interval	Relate raster cells to fuzzy object via *part of* links in [0, 1] interval

2.2.3 Classification and geometric partitioning

Irrespective of the data structure, spatial modelling always requires geographic space to be partitioned into a finite number of geometrical elements. If these elements, denoted x_j, are disjoint, they together constitute the geometric universe of the spatial model M, or more briefly, the map geometry, $\mathbf{G}_M = \{x_1, x_2, ..., x_n\}$. Each elementary unit is linked to a single thematic description consisting of a one or more valued attributes. If the attribute data is denoted $\mathbf{x}_j$, with index j referring to the jth element in $\mathbf{G}_M$, then $\mathbf{X}_M = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ denotes the attribute space or feature space of M. Objects in M can be distinguished because they have dissimilar descriptions[1]. For many GIS applications the differences will be primarily thematic. Contiguous geometrical elements sharing the same thematic description then belong to one object, at least for the purpose of the survey. Classification is a helpful tool to check for this condition. In this context, elements are considered to belong to one and the same (data) class if they are described using the same set of attributes and if they have similar attribute values.

[1] Molenaar (1994, 1998) introduces the concept map universe, $\mathbf{U}_M$, as the set of all objects occurring in a map M. Reference to a $\mathbf{U}_M$ at this stage assumes a set of known objects. This is an unrealistic assumption in a surveying context where objects are yet to be established. Moreover, the geometry of objects having an uncertain extent is modelled in $\mathbf{G}_M$ rather than $\mathbf{U}_M$ (see Molenaar, 1998, p. 198).

The term *class intension* refers to the definition of a class as given by the properties that determine class membership. A class is *crisp* if its intension is clearly defined. In that case there are well-defined criteria to determine whether an element should be considered a member of the class (Molenaar, 1998). This results in a crisp membership function:

$$\mu_{A_i}(\mathbf{x}_j) = \begin{cases} 1 & \text{if } \mathbf{x}_j \text{ meets the criteria for membership in } A_i \\ 0 & \text{otherwise.} \end{cases}$$

A system of c classes for which $\sum_{i=1}^{c} \mu_{A_i}(\mathbf{x}_j) = 1 \quad \forall \quad j \in \{1,\ldots,n\}$, i.e. the classes A_i are disjoint and exhaustive with respect to the elements $\mathbf{x}_j$, leads to a thematic partition of $\mathbf{X}_M$. The one-to-one link between elements in $\mathbf{X}_M$ and those in $\mathbf{G}_M$ (see above) implies that a thematic partition of $\mathbf{X}_M$ generates a geometric partition of $\mathbf{G}_M$ (cf. Molenaar, 1998, pp. 141-142).

In Section 2.2.1 fuzzy sets were introduced as a means to deal with spatial objects with indeterminate boundaries. A fuzzy set has a weakly defined intension, i.e. the criteria that define whether an element is a member of the set, or class, are vague. Consequently, membership in a fuzzy set A_i is allowed to be partial: $0 \le \mu_{A_i}(\mathbf{x}_j) \le 1$. If $\mu_{A_i}(\mathbf{x}_j) = 1$, element $\mathbf{x}_j$ has properties that completely match the central notion represented by A_i. If $\mu_{A_i}(\mathbf{x}_j) = 0$, the properties of $\mathbf{x}_j$ definitely exclude it from membership in A_i. Otherwise the membership function takes an intermediate value. A system of c fuzzy classes for which $\sum_{i=1}^{c} \mu_{A_i}(\mathbf{x}_j) = 1 \quad \forall \quad j \in \{1,\ldots,n\}$ generates a fuzzy thematic pseudopartition of $\mathbf{X}_M$ (Klir and Yuan, 1995), and hence a fuzzy geometric pseudopartition of $\mathbf{G}_M$. Presence of spatial correlation of data from nearby elements leads to their grouping into spatially contiguous regions. The latter can be interpreted as objects with a fuzzy extent after evaluating adjacency relationships between geographic elements (Molenaar, 1998).

Methods for constructing membership functions can be divided into expert judgement-based and data-driven approaches. The Keys to Soil Taxonomy (Soil Survey Staff, 1996) are a well-known system of crisp membership functions that have been constructed on the basis of expert knowledge. Partitional or hierarchical cluster analysis of a multivariate data set (e.g. Van Ryzin, 1977; Gordon, 1981) can be used to obtain data-dependent crisp membership functions. The former divide the entire data set of n elements into a specified number (c) of disjoint groups. The latter produce hierarchically nested sets of thematic partitions (see Figure 2.5). The partitional fuzzy c-means clustering algorithm (Bezdek, 1981) is frequently used to construct fuzzy membership functions. On the other hand, Klir and Yuan (1995) describe several direct and indirect methods to construct fuzzy membership functions on the basis of expert knowledge. Membership functions derived from expert knowledge are also known as semantic import models (Burrough and McDonnell, 1998).

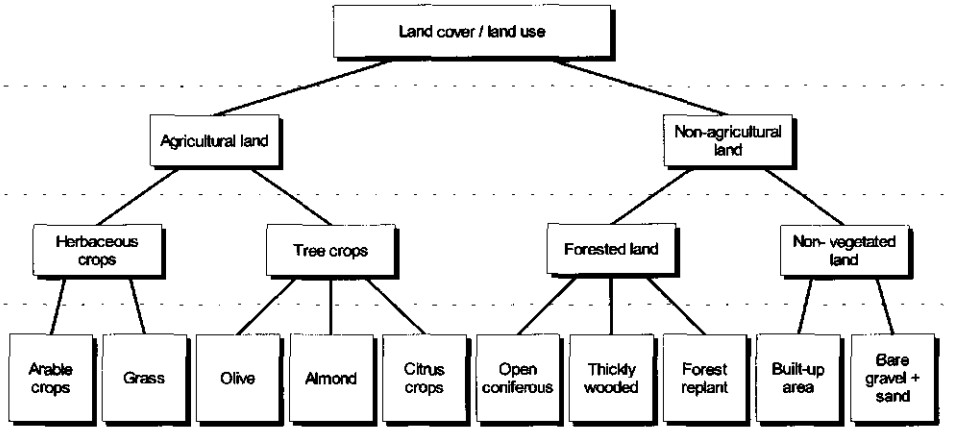


Figure 2.5 Example of a classification hierarchy. The dashed lines separate hierarchical levels of the classification system.

2.2.4 Classification hierarchy

A classification hierarchy can be represented as an inverted tree showing relations between nested thematic partitions (see Figure 2.5). Sectioning a crisp classification hierarchy at any level, as illustrated by the dashed lines in Figure 2.5, will produce a partition of the elements into disjoint groups. Each class of a lower level partition is wholly contained within a single class of a higher level partition (Gordon, 1981). In a downward direction along the tree class intensions become more specific, so that the elements' descriptions are specialised. In the opposite case the descriptions of the elements become more generalised.

Hierarchical cluster analysis creates a classification hierarchy by analysing the data using some measure of thematic proximity (Gordon, 1981). Classification hierarchies can also be obtained by dissection or agglomeration of classes on the basis of expert judgement. Similarly, fuzzy classes belonging to a pseudopartition of $\mathbf{X}_M$ can be combined to generate a fuzzy pseudopartition at a higher hierarchical level (e.g. De Bruin and Stein, 1998; see Chapter 4). Whereas membership in the union of crisp classes is uniquely determined by the membership grades in the individual classes, there exist many fuzzy union operators that have validity in different contexts (Klir and Yuan, 1995). It can be checked that agglomeration of fuzzy classes by standard fuzzy union (i.e. $\mu_{A_1 \cup A_2}(\mathbf{x}_j) = \max[\mu_{A_1}(\mathbf{x}_j), \mu_{A_2}(\mathbf{x}_j)]$) does not necessarily produce a higher level fuzzy pseudopartition of $\mathbf{X}_M$ and hence $\mathbf{G}_M$. In that context the bounded sum operator (i.e. $\mu_{A_1 \cup A_2}(\mathbf{x}_j) = \min[1, \mu_{A_1}(\mathbf{x}_j) + \mu_{A_2}(\mathbf{x}_j)]$) is more appropriate. Note that for this particular purpose the upper bound (unity) is non-restrictive so that $\mu_{A_1 \cup A_2}(\mathbf{x}_j) = \mu_{A_1}(\mathbf{x}_j) + \mu_{A_2}(\mathbf{x}_j)$.

2.2.5 Hierarchical object relationships

Just as there exist spatial objects that are composed of several related geometrical elements, there exist composite objects made up of multiple elementary objects. The upward relationships between elementary units and higher level objects are expressed in *part of* links. In general, these links are established on the basis of rules that evaluate two types of criteria (Molenaar, 1993, 1998):

- Criteria specifying the classes of the elementary units that are considered for aggregation;
- Criteria specifying the geometric and topological relationships among these elements.

Connectivity (of line segments) and adjacency (of area elements) are important topological relationships in this respect. For example (see Figure 2.5), adjacent areas classified as *open coniferous forest*, *thickly wooded land* and *forest replant* may be aggregated to represent a contiguous forest object. In this particular example aggregation conformed to a classification hierarchy[1]. Often this is not the case as classification hierarchies and aggregation hierarchies are quite different.

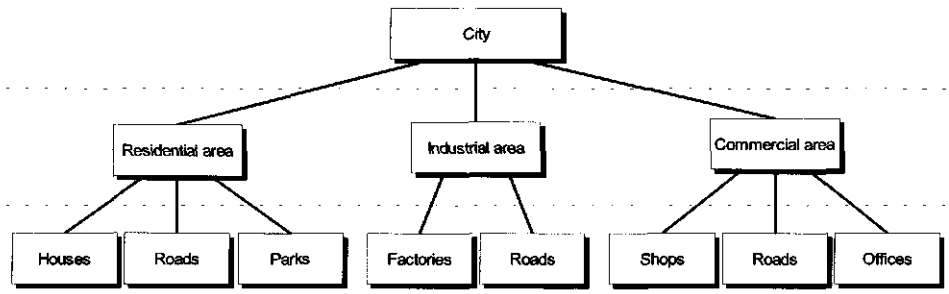


Figure 2.6 Hierarchical relationships between elementary and aggregated objects. The dashed lines separate aggregation levels (after Molenaar, 1993).

Figure 2.6 shows an example of a functional aggregation hierarchy[2]. The figure illustrates the semantic difference between upward links in a classification hierarchy and those in an aggregation hierarchy. In a classification hierarchy, classes are linked to higher level classes by *is a* links. For example, a citrus crop *is a* tree crop, and land covered by a tree crop *is* agricultural land (see Figure 2.5). The links are valid wherever the citrus crop is located, irrespective of the neighbouring crops. On the other hand, upward links in an aggregation hierarchy are *part of* links. For example, a road segment *R* can be *part of* a residential, an industrial, or a commercial area, each of which is *part of* the city (see Figure 2.6). To determine the type of area of which *R* actually forms a

[1] This type of aggregation is referred to as class driven aggregation (Molenaar, 1998).

[2] Functional aggregation, on the other hand, requires completely different thematic description of aggregate objects, so that other classes should be defined (see Figure 2.6).

part it is necessary to evaluate its adjacency relationships with objects of the type *house*, *park*, *factory*, *office* and *shop*.

Spatial objects that are considered elementary at one scale may be regarded as composite objects at larger scales, whereas they may hold too much detail for representation and analysis at smaller scales. When elementary objects are aggregated, so will part of their attribute values. At the same time some data may be discarded as they hold no significance for the composite objects. Usually, the geometric description of lower level objects is lost as a result of merging. Consequently, a terrain description at a higher aggregation level contains less detail than a description of elementary objects. In the opposite direction, disaggregation of composite objects requires that additional information be included in the terrain description (De Bruin *et al.*, 1999; see Chapter 3).

2.3 Data acquisition and mapping

2.3.1 Primary and secondary data

After choosing a conceptual data model (i.e. continuous field, discrete object or fuzzy object), a desired level of spatial detail (i.e. resolution or aggregation level), and the thematic attributes for which data are to be recorded, systematic data collection can commence. In a land resources survey this typically involves collecting a small sample of precisely measured primary data (ground truth) as well as a larger or even exhaustive sample of related secondary data.

Because soil is hidden below the surface, it can only be examined at a limited number of locations. Predictive mapping of soil properties at unvisited locations may well benefit from complementary data on external indicators such as landscape morphology, vegetation and surface colours (e.g. Hall and Olson, 1991; Soil Survey Division Staff, 1993; Slater *et al.*, 1994). Land cover, on the other hand, is readily visible on the surface. Yet, if large areas of land are to be mapped it is not feasible to obtain complete area coverage by field survey methods alone (e.g. Gillespie *et al.*, 1996). As satellite remote sensing provides a synoptic view of the Earth's surface it allows for timely and consistent acquisition of regional and global land cover data (Barnsley *et al.*, 1997).

2.3.2 Soil survey

Using the soil-landscape model, soil surveyors classify and delineate bodies of soil on the landscape by directly examining $<<0.1\%$ of the soil below the surface (Hudson, 1990, 1992). The conventional soil-landscape model adopts the discrete object view. It is built on the concept of soil-landscape objects. These are terrain units resulting from the interactions of the five factors affecting soil formation, i.e. parent material, climate, organisms, relief and time (Jenny, 1941; Hall, 1983; Hudson, 1990, 1992; Hall and Olson, 1991; Hewitt, 1993). They are conceived as being spatially organised in larger landscape units according to an aggregation hierarchy (Soil Survey Division Staff, 1993, pp. 9-11). Boundaries between soil-landscape objects can be recognised and mapped as discontinuities on the earth's surface, and usually coincide with abrupt changes in the soil

cover. Visual interpretation of aerial photography may play a substantial role here (De Bruin *et al.*, 1999; see Chapter 3). The relevance of the boundaries for soil mapping is checked using field observations such as widely spaced augerings and soil pits. Soil-landscape objects are grouped into a limited number of classes, often referred to as map units (Soil Survey Division Staff, 1993), each with a characteristic soil cover. The soil cover is usually described with reference to some system of soil classification (e.g. Soil Survey Staff, 1996). Burrough *et al.* (1997) called this conceptual model the 'double crisp' model because identified soil groups are assumed to be crisply delineated in both taxonomic space and in geographic space.

The discrete object view adopted in the original soil-landscape model is an approximation and a simplification of a more complex pattern of variation. Boundaries between soil-landscape units are often transition zones rather than sharp boundaries. It is inappropriate to assign sites within a transition zone to any single soil-landscape unit. Rather, these sites should be assigned partial membership in two or more units (Lagacherie *et al.*, 1996). This can be achieved by adopting a fuzzy object view. De Bruin and Stein (1998), see Chapter 4, explored the use of fuzzy *c*-means clustering of attribute data derived from a digital elevation model (DEM) to represent transition zones in the soil-landscape.

Another modification of the soil-landscape model is based on viewing landscape and target soil properties as correlated continuous fields. The modification relies on Jenny's (1941) factors of soil formation, but rather than viewing the soil-landscape as being composed of discrete objects it adopts a continuous fields view. One approach has been to generate multilinear regression models relating a sparse sample of soil data to an exhaustive set of attribute data derived from a DEM. The regression models are then used to predict the target variables to the grid nodes of the DEM (Moore *et al.*, 1993; Odeh *et al.*, 1994; Gessler *et al.*, 1995). A serious drawback of using simple regression for spatial prediction is that it takes no account of the spatial dependence among locations. Response variables are estimated from local explanatory variables using global regression equations. These equations are not exact inasmuch as they do not honour measured data values at their locations. Additionally, any information from nearby sites is ignored. Therefore, regression does not make full use of the data (Atkinson *et al.*, 1994).

On the contrary, geostatistical methods exploit rather than ignore spatial dependence of sample data. In geostatistics, spatial variability of a property is considered as a realisation of a random function that can be represented by a stochastic model (e.g. Isaaks and Srivastava, 1989, pp. 198-236). The geostatistical method of spatial prediction is called kriging. At its simplest, kriging is no more than a method of weighted averaging of the sampled values of a property Z within a neighbourhood n (Webster and Oliver, 1990). However, there are several kriging methods that allow the incorporation of secondary data in the interpolation process (e.g. Goovaerts, 1997, 1999). Some kriging variants are specially adapted to predict categorical variables (e.g. soil classes). In this thesis the use of these methods is explored in the context of land cover mapping rather than soil surveying (Chapter 6).

2.3.3 Land cover classification

Satellite remote sensing has become an important tool in land cover mapping, providing an attractive supplement to relatively inefficient ground surveys. The elementary unit of a remotely sensed image is the pixel (picture element). A recorded pixel value is primarily a function of the electromagnetic energy emitted or reflected by the section of the earth's surface that corresponds to the sensor's instantaneous field of view (IFOV). Sensor systems typically collect data in several spectral bands (e.g. the Thematic Mapper sensor on Landsat 5 has seven spectral bands). It is usually assumed that the energy flux from the IFOV is equally integrated over adjacent, non-overlapping rectangular cells; the pixels' ground resolution cells. In practice, most sensors are centre biased such that the energy from the centre of the IFOV has most influence on the value recorded for a pixel (Fisher, 1997). The IFOV of a sensor can also be smaller or larger than the ground resolution cell. However, in a well-designed sensor system the ground resolution cell will approximate the instantaneous field of view of the instrument (Strahler *et al.*, 1986).

A common approach to extract land cover data from remotely sensed imagery is by multispectral classification. The usual assumptions are that the image scene is composed of discrete, crisply bounded, homogeneous land cover regions that are larger than the sensor's ground resolution cells (*H*-resolution: Strahler *et al.*, 1986). However, several classifiers allowing alternative assumptions have been proposed (Robinove, 1981; Wang, 1990a,b; Foody, 1992, 1997; Eastman, 1997), but these will not be discussed in this thesis. In conventional supervised image classification, a pixel is regarded as a sample from one of a known number (c) of land cover populations (classes), each having a characteristic spectral response pattern. The aim is to assign the pixel to the correct class, in which it has full membership. Spectral response patterns are obtained from training data for which the true classes are known. Usually sample means and sample variance matrices are used as the parameters of normal class probability densities.

Bayes' classification rule assigns a pixel, x, characterised by its spectral feature vector $\mathbf{x}$, to the category C_i for which it attains maximum posterior probability $P(x \in C_i|\mathbf{x})$, or more briefly, $P(C_i|\mathbf{x})$:

$$P(C_i|\mathbf{x}) = \frac{P(\mathbf{x}|C_i)P(C_i)}{\sum_{i=1}^{c} P(\mathbf{x}|C_i)P(C_i)} \tag{2.1}$$

where $P(\mathbf{x}|C_i)$ is the probability of $\mathbf{x}$, conditional to C_i and $P(C_i)$ is the prior probability of C_i irrespective of $\mathbf{x}$ (Duda and Hart, 1973). The prior probability $P(C_i)$ is an initial estimate of the proportion of pixels that belongs to a particular category C_i. Classification can benefit from stratification of the image, particularly if prior probabilities estimates are available for each stratum (Strahler, 1980; Hutchinson, 1982). Gorte and Stein (1998) developed an algorithm that uses intermediate classification results to iteratively adjust prior probabilities related to spatial strata. De Bruin and Gorte (2000), see Chapter 5, used this algorithm to improve land cover classification after stratifying Landsat TM imagery on the basis of geological map units.

Image classifiers typically ignore the spatial component of data or even assume that data vectors in neighbouring pixels are independent, but clearly this is not so. Failure to account for spatial dependencies can result in increased classification error rates and representations that are patchier than the true scene (Cressie, 1991, pp. 501-504). Chapter 6 presents a geostatistical method to update image derived class probabilities of type (2.1) by conditioning on a sample of high accuracy land cover data.

2.4 Uncertainty modelling

2.4.1 Types of uncertainty

The fact that any landscape description is a model based on a limited sample of measured target attribute data implies that it is never completely certain. One kind of uncertainty already referred to concerns *fuzziness* of the class intensions used in a landscape description. Fuzziness is directly related to the fuzzy object world view (see Section 2.2.1).

Uncertainty may also denote a recognition of possible error in the reported value (Couclelis, 1996). In this respect it is closely related to *accuracy*, which is usually defined as closeness of estimates to values accepted to be true (Unwin, 1995). Regardless of the conceptual model, any terrain description is affected by the latter kind of uncertainty. Consider, for example, a statement of the type $x \in A_1$, or $\mu_{A_1}(x)=1$, i.e. element x belongs to set A_1 (Molenaar, 1993, 1996, 1998). An example of such a statement is: location x belongs to a high region. Fuzziness then concerns the definition of A_1 (high). Is the class intension crisply defined, e.g. by an elevation exceeding 500 m, or is it defined by a fuzzy membership function? Regardless of the definition of A_1, a statement $\mu_{A_1}(x)=1$ may be inaccurate because the attribute value of x contains measurement error and/or there is insufficient evidence to assign x to A_1. For example, the elevation of x may be derived from a digital elevation model so that it is likely to be in error. Or, instead of elevation, air pressure is measured using a precision instrument. In that case the evidential support for definite assignment of x to A_1 may be lacking.

A third kind of uncertainty is due to lack of *precision.* Precision refers to the granularity or resolution at which an observation is made, or information is presented (Worboys, 1998). It can be expressed in terms of number of bits, or significant digits or level of generalisation of a classification system. High precision certainly does not imply a high level of accuracy (Unwin, 1995). In this thesis, the fuzziness (Chapters 4 and 7) and error or accuracy related (Chapters 5-7) aspects of uncertainty are explored. In the remainder of this section they are referred to as *fuzziness* and *inaccuracy* respectively.

2.4.2 Error modelling for inaccuracy assessment

Map inaccuracies cannot be calculated for complete landscape descriptions, since this would require knowledge of accurate values for every mapped location. If this were the case, inaccuracy could simply be eliminated by substitution. Error modelling, on the other hand, allows an indication of the possible magnitude or distribution of inaccuracies

for spatial attributes to be given (Isaaks and Srivastava, 1989, pp. 489-497; Goodchild *et al.*, 1992; Heuvelink, 1993, 1998a).

Measures commonly used in error modelling are *error variances*, *confidence intervals*, and *probability distributions*. In a terrain description, an error variance represents the expected squared deviation from a reported local value; i.e. the variability component not accounted for by the model. A confidence interval reports an interval, rather than a single estimate, as well as a probability that the true value falls within this interval. Probability distributions specify ranges of possible values, each with an associated probability of occurrence. They also allow error modelling for random categorical variables. These are random variables on a nominal scale, taking only one from an unordered set of discrete values[1]. Probability distributions provide considerably more information than error variances or confidence intervals as they model the extent and distribution of possible departure from reported values. Combined with a loss (or utility) function, probability distributions allow the risk involved in alternative decisions, made on the basis of landscape descriptions that are likely to contain error, to be evaluated (Isaaks and Srivastava, 1989; Goovaerts, 1997, 1999; Gorte, 1998; Kyriakidis, 1999).

As the term implies, error modelling always requires a model specifying prior concepts (decisions) about the spatial phenomenon under study (Goovaerts, 1997, p. 442). Therefore, error modelling is to some extent a subjective enterprise, with different models giving different results. In this thesis, an example from remotely sensed image classification is used to illustrate implications of some modelling choices on error estimation (Chapter 6).

2.4.3 Inaccuracy of classified imagery

Remotely sensed image classifiers typically report only the most likely class for each pixel. Classification output thus does not differentiate between pixels being spectrally similar to a single class and those presenting spectral similarity with two or more classes (Foody *et al.*, 1992; see Figure 2.7). Usually, an accuracy statement is provided in the form of an overall classification accuracy measure (producer's accuracy or user's accuracy) or a confusion matrix, also known as a misclassification or error matrix. The producer's accuracy indicates the probability that a reference pixel is correctly classified, and so is a measure of omission error. The user's accuracy, on the other hand, is an experimental estimate of the probability that a classified pixel actually represents the reported category on the ground, and is thus related to commission error. The confusion matrix allows these and other inaccuracy measures for individual categories to be calculated (Aronoff, 1982; Congalton *et al.*, 1983 Rosenfield and Fitzpatrick-Lins, 1986; Story and Congalton, 1986; Congalton, 1991).

An obvious shortcoming of confusion matrix-derived measures is their implicit assumption of homogeneity over the mapped area (Goodchild *et al.*, 1992). Conversely, a model of local inaccuracies is obtained by viewing the unknown class of a pixel as a

[1] Categorical variables and crisp sets are related in the sense that a category is a crisp set. Thus, if x is an element and the random categorical variable $S(x)$ takes the value s_i for x, then x is a member of a crisp set C_i, i.e. $C_i = \{x \mid S(x) = s_i\}$.

random variable. The vector of posterior probabilities from Bayes' classification rule (Eq. 2.1) may then be used to provide an estimate of its conditional distribution, given the remotely sensed spectral response (Goodchild *et al.*, 1992; Foody *et al.*, 1992; Van der Wel *et al.*, 1998). This approach, which implicitly assumes that the random variables in neighbouring pixels are independent, has also been demonstrated by De Bruin and Gorte (2000; see Chapter 5).

Besides neglecting spatial dependence between pixels, the approach based on Equation 2.1 does not make full use of available reference data as it ignores their spatial component. It does not consider data locations nor does it use spatial dependence models that may be derived from the reference data. Kyriakidis (1999) and De Bruin (2000), see Chapter 6, proposed geostatistical methods to update image derived class probabilities by conditioning on a sample of high accuracy data. These methods not only enable improved modelling of local classification inaccuracies, but also allow assessment of *spatial inaccuracy*, i.e. the joint uncertainty about the class label at several pixels taken together (e.g. objects). Spatial inaccuracy is modelled by stochastic simulation, i.e. generating multiple equiprobable realisations of the joint distribution of attribute values in space (Zhu and Journel, 1993; Journel, 1996; Goovaerts, 1997, 1999).

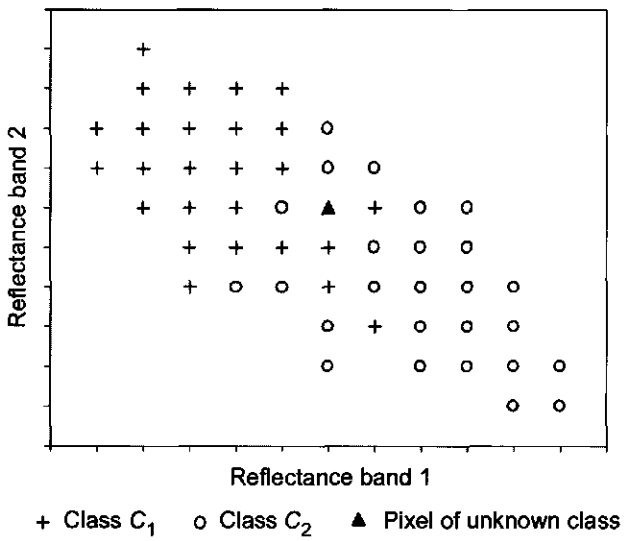


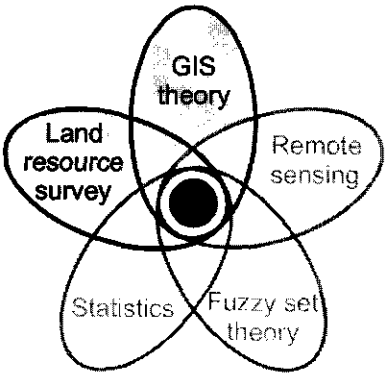
Figure 2.7 Overlap between two classes C_1 (+) and C_2 (o) in a two band spectral space. Based on spectral data alone, the pixel of unknown class (▲) cannot unambiguously be assigned to either class.

2.4.4 Combining fuzziness and inaccuracy

The above error models assume that each geometric element belongs to a single class that can be positively identified once sufficient data has been collected. Presence of mixed pixels invalidates this assumption. In this thesis, multiple class membership at the

pixel level is further explored insofar as it is due to fuzzy class intensions[1]. Fuzzy class intensions impose extra modelling efforts as the inaccuracy and fuzziness aspects of uncertainty will co-occur. Fuzzy set theory and probability can be used together to model both aspects of uncertainty in combination. In Chapter 7 this is demonstrated by calculating the expectation of a fuzzy membership function defined on a random variable. Chapter 7 also introduces the concept of a fuzzy probability qualifier (or fuzzy probability) to deal with vague selection criteria in answering queries on probabilistic data.

[1] Mixed pixels may also be due to discrete object boundaries crossing a pixel's ground resolution cell or presence of sub-pixel objects (Fisher, 1997; De Bruin and Molenaar, 1999).


GIS
theory
Land
resource
survey
Remote
sensing
Statistics
Fuzzy set
theory

3 Formalisation of soil-landscape knowledge through interactive hierarchical disaggregation[1]

Abstract

The soil-landscape model strongly depends on scarcely documented expert knowledge. In this paper a methodological framework is formulated that takes advantage of a GIS to interactively formalise soil-landscape knowledge using stepwise image interpretation and inductive learning of soil-landscape relationships. It examines topology to keep record of potential *part of* relationships between terrain objects denoting discontinuities in soil formation regimes. The relationships are used to visualise the pathway along which terrain objects have been derived. They can be applied in similar areas to facilitate image interpretation by restricting possible lower level terrain objects. The framework may adopt different methods to describe soil variation in relation to a terrain description. It is illustrated using stratification of soil texture data according to terrain object classes in a case study within the Guadalhorce basin in southern Spain. The degree of association between terrain object classes and particle size classes increased from 6% to 38% in three steps of image interpretation.

3.1 Introduction

The soil-landscape model (Hudson, 1990, 1992) regards the landscape as a mosaic of soil-landscape objects that can be grouped into a limited number of classes, each with a characteristic soil cover. Boundaries between soil-landscape objects can be recognised and mapped as discontinuities on the earth's surface, usually coinciding with abrupt changes in the soil cover. Visual image interpretation plays a substantial role in soil-landscape modelling. Remotely sensed images provide a synoptic view of the survey area, in which an interpreter can detect zones of rapid change in one or more soil forming factors.

---

[1] Based on: De Bruin, S., Wielemaker, W.G., and Molenaar, M., 1999. Formalisation of soil-landscape knowledge through interactive hierarchical disaggregation. Geoderma 91, 151-172.
© 1999 Elsevier Science B.V.

Recent work by Bell *et al.* (1994), Deka *et al.* (1995), McLeod *et al.* (1995), Wright (1996) and others demonstrates the continuing success of the soil-landscape model. Criticism, however, mainly focuses on the following problems:

1. Its geographical model, the exact object model, cannot properly represent spatial variation of soil properties (Lagacherie *et al.*, 1996; Burrough *et al.*, 1997).
2. Soil survey reports are difficult to update with new information and incapable of responding to specific customers' demands (Bouma and Hoosbeek, 1996; Indorante *et al.*, 1996).
3. The soil-landscape model largely relies on tacit knowledge (Hudson, 1992). Tacit knowledge is difficult to communicate, which explains soil survey's general failure to communicate about the methods and models employed in deriving map units and statements about their content (Hewitt, 1993).

The first problem has received much attention in recent years. Burrough *et al.* (1997) identified two major phases along which a new paradigm of soil classification and mapping is evolving from the exact object model. These are the introduction of geostatistics in the 1980's, and the introduction of fuzzy set methods in the 1990's. Burrough *et al.* (1997) concluded that when applying these tools "... primary boundaries and zonations based on important differences in lithology, landform or drainage must be taken into account ..." Finding these boundaries and zonations heavily relies on the surveyor's tacit knowledge (problem 3).

Both phases emerged parallel to the advent of geographic information systems (GISs). These enable user access of computer stored soil data, providing new opportunities for data actualisation, analysis, and interaction with customers (Indorante *et al.*, 1996). Yet, as long as the GIS merely contains a copy of the traditional soil map, problem 3 remains. A GIS employed during data acquisition, however, may capture expert rules. The soil database would accumulate tacit knowledge, making it available for others and for application in similar areas.

This paper formulates a methodological framework that takes advantage of modern GIS capabilities to interactively formalise soil-landscape knowledge. Several recent studies proposed methods to infer soil characteristics from environmental data (e.g. Cook *et al.*, 1996; Thompson *et al.*, 1997; Zhu *et al.*, 1996; 1997), or explored the use of soil pattern knowledge in automated survey (Lagacherie *et al.*, 1995; Domburg *et al.*, 1997). In contrast, our framework focuses on image interpretation for soil-landscape modelling. It is compatible with the common practice of remotely sensed image interpretation, and may adopt different methods to describe soil variation in relation to a terrain description. It involves terrain description at successive levels of detail, information transfer between these levels, and explicit representation of expert decisions. The framework is illustrated with a case study within the Guadalhorce basin in southern Spain.

3.2 Methodological background

3.2.1 Terrain objects

In conformity with the soil-landscape paradigm we regard the landscape as a mosaic of spatial objects. The elementary object is the facet, which corresponds to the smallest

landscape segments that can be discerned on large scale (e.g. 1: 10 000) aerial photographs (cf. Dent and Young, 1981). We assume that facets are homogeneous as for lithology, morphogenetic origin, curvature and relative position in the slope sequence, and have a narrow range in slope gradients. An area of 400 m^2 is adopted as the lower size limit of a facet. Facets are similar to geomorphological sites (Wright, 1996), except that the lower size limit is higher so that they can be distinguished on aerial photographs.

Higher level terrain objects are composites of multiple facets that satisfy certain aggregation rules. For example, an alluvial terrace may consist of two adjacent facets, one being an abandoned flood plain and the other a descending slope. An *aggregation hierarchy* defines how to compose objects from elementary objects and how to combine these to build more compound objects, and so on (Molenaar, 1996). The above alluvial terrace could be part of a river valley that comprises the river channel, the present floodplain and several differently aged terraces. Within an aggregation hierarchy each lower level object belongs to exactly one higher level object, while the objects of each single aggregation level compose the entire survey area. The aggregation hierarchy thus corresponds to a series of nested spatial partitions.

In this paper the term terrain object refers to an object at any level of the aggregation hierarchy. Its boundaries correspond with zones of rapid change in one or more soil forming factors over short distance. We assume that all terrain objects belong to some class, while each terrain object belongs to exactly one class. Each level of the aggregation hierarchy has therefore a thematic partition that comprises the complete set of necessary terrain object classes.

A *soil-landscape object* is a terrain object accompanied by a description of the soil cover. In a full soil-landscape model the soil cover is described in terms of many properties and with reference to the entire soil, for example using soil types characterised by modal profiles. A partial soil-landscape model refers to one or a few individual properties and/or only part of the soil profile.

3.2.2 Image interpretation

We assume that the terrain objects are interpreted from aerial photography following the stepwise interpretation method described by Olson (1973) and Estes *et al.* (1983). The interpretation results in various division levels that form a hierarchy of nested spatial partitions or, in other words, a disaggregation hierarchy.

During disaggregation, attribute values of higher level objects are to be decomposed into lower level data (cf. Molenaar, 1996). If lower level objects would again be combined (aggregation), the original attribute values of the composite should be recovered. Attribute values of higher level objects therefore constrain the domain of attribute values of lower level objects. These domain constraints take the form of rules specifying possible types of lower level terrain objects given the higher level object class. The hillslope model with summit, shoulder, backslope, footslope and toeslope (Ruhe, 1960) is an example of such a rule.

The rules restrict the type of evidence needed to establish lower level terrain objects, enabling a better directed exploration of available information sources. Stepwise image interpretation thus provides a mechanism to streamline the identification of terrain objects at the lowest level of interest. However, working exclusively from general to

specific considerations may lead to biased results (Olson, 1973; Estes *et al.*, 1983). An error introduced in the first disaggregation step, if not corrected, will propagate through the hierarchy. At any time, the lowest level of a disaggregation hierarchy contains mere hypotheses about the terrain objects identified at that level. These hypotheses must be confirmed using feedback of evidence obtained from subsequent levels. Therefore, image interpretation is iterative, both inductive and deductive, rather than a one-way deductive process.

3.2.3 Topological relationships

Disaggregation of terrain objects concerns thematic, geometric and topological elements. Domain constraints on attribute data are rules with respect to the thematic description of terrain objects. For example, a piedmont may be composed of differently aged alluvial fans and colluvial slopes. A mountain crest cannot occur within that piedmont. Geometry related rules refer to the size, shape and position of terrain objects. They may, for example, specify the allowable size of impurities when decomposing a terrain object into its components. Important topological relationships are containment, adjacency and overlap (Figure 3.1).

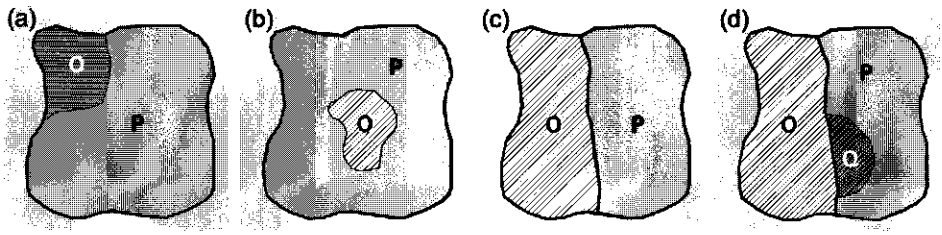


Figure 3.1 Topological relationships between terrain objects. Thick lines denote objects delineated in the first disaggregation step; a thin line delimits an object recognised in the next step. (a) Containment; O is within P and O is part of P. (b) Containment; O is an island within P because O is not part of P. (c) Adjacency. (d) Overlap; O overlaps P because Q is part of O.

Containment

The containment relationship (Figure 3.1a, b) asserts whether a terrain object, O, is within another object, P. If O is contained in P and if the attribute domain constraints for decomposing P are satisfied, then O is part of P (Figure 3.1a). An object O is considered to be an island in P if these domain constraints are not satisfied (e.g. Figure 3.1b). In that case O is adjacent to P, but not part of it.

Adjacency

The adjacency relationship (Figure 3.1c) indicates whether two terrain objects border. Adjacent objects have to be thematically different. An alluvial terrace can only be distinguished from an adjacent terrace if it has differentiating properties. Adjacent terrain objects may also be associated. For example, an alluvial fan is associated with the

uplands from which it receives sediment. Adjacency can help uncover repeating associations that are related to functional dependencies between terrain objects.

Overlap

Overlap occurs if a terrain object, Q, is contained in P while it is part of an adjacent object, O (Figure 3.1d). Image interpretation produces a hierarchy of nested spatial partitions, each having a thematic partition. Overlap is therefore not possible. It may occur, though, as a consequence of the use of incomplete evidence during image interpretation. For example, an alluvial terrace, P, may be found to contain a colluvial footslope, Q, which according to an earlier determined relationship is part of an adjacent hillslope, O (Figure 3.1d). The overlap indicates a misinterpretation of the boundary at the previous disaggregation level.

This does not imply that a terrain object cannot be part of different aggregates. Distinct user contexts result in different aggregation hierarchies, which may assign an object to different aggregates (Molenaar, 1996). Image interpretation, however, should result in a single hierarchy. Other hierarchies may be formulated once the terrain objects have been interpreted.

3.2.4 The soil cover of terrain objects

A terrain description for soil-landscape modelling allows to predict the soil cover based upon relationships between soils and terrain object classes, possibly in combination with other spatial data. Except for some generalities (for example, steep, sparsely vegetated slopes have shallow soils), the relationships are determined from field observations like widely spaced augerings and soil pits. Several methods exist to derive relations among spatial data, such as classification according to spatial features (Webster and Oliver, 1990, pp. 67-70), linear regression (Moore *et al.*, 1993), Bayesian inductive modelling (Cook *et al.*, 1996), multivariate discriminant analysis (Bell *et al.*, 1994), generalised linear modelling (McKenzie and Austin, 1993), stratified kriging (Stein *et al.*, 1988; McBratney *et al.*, 1991), fuzzy soil-land inference (Zhu *et al.*, 1996; 1997) and classification trees (Lagacherie and Holmes, 1997). Any of these methods can be used to derive a soil-landscape model.

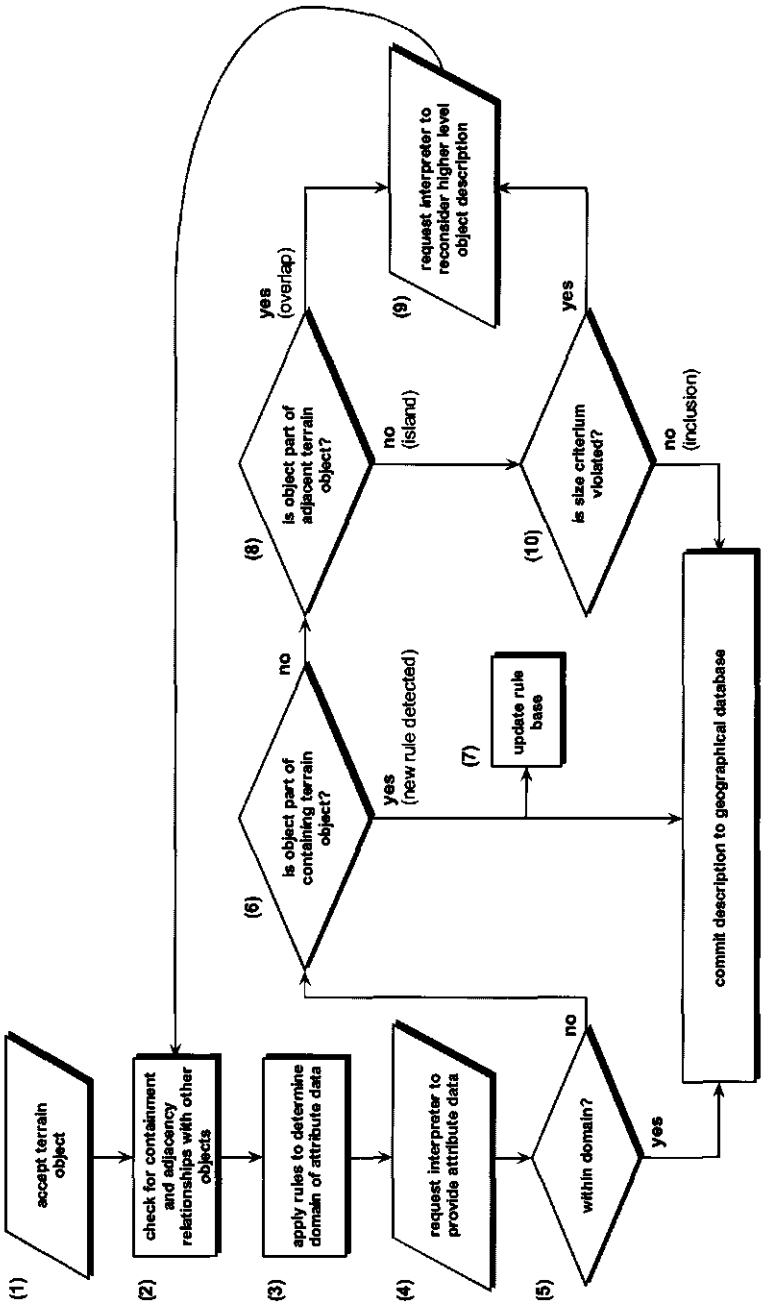


Figure 3.2 Process flow of GIS-assisted image interpretation. The numbers between brackets are explained in Section 3.3.1.

Soil sample data are also used to assess the predictive power of the terrain description. The data used for this purpose may either be the same that were used to estimate the model parameters, or belong to an independent subset to obtain an unbiased estimate of model accuracy. Depending upon its accuracy, the original soil-landscape model may need to be revised by application of a different inductive method or further image interpretation. Image interpretation thus involves another iteration loop that includes assessment of the degree of association between the terrain description and soil properties.

3.3 Framework

We now present a framework for deriving a soil-landscape model. It involves integration of GIS-assisted data entry and analysis tools into soil survey supported by image interpretation. It has been divided in two parts. The first part concerns the image interpretation. The second part derives a soil-landscape model using the terrain description from the first part.

3.3.1 GIS-assisted stepwise image interpretation

Image interpretation starts with demarcation of the survey area. It proceeds stepwise (section 3.2.2), running the process represented in Figure 3.2 for each newly identified terrain object. It presumes the existence of an initial set of rules describing possible *part of* relationships between lower level and higher level terrain objects. Some rules may already be in computer interpretable notation whereas others are formalised interactively during interpretation.

First the geometry of a terrain object is digitised using photographic imagery or a digitised thematic map layer (1). Its containment and adjacency relationships with other objects are determined (2). These topological relationships are used to derive a preliminary set of valid attribute data (3). These may concern landform class, lithology, relative age, surface topography, vegetation, etc. Containment activates a rule specifying possible values given the attribute data of the higher level terrain object. Adjacency excludes the attribute values assigned to neighbouring terrain objects.

Next, the interpreter must specify the attribute data pertinent to the terrain object (4). These can either be within the preliminarily determined domain, or in conflict with it (5). In the first case the object description is committed to the geographical database; in the latter case progress depends on whether the interpreter considers the terrain object to be part of the containing object (6). This admittedly subjective decision reflects the interpreter's opinion regarding the thematic significance of a *part of relationship* and its potential relevance in similar areas. For example, if a large piedmont contains frequent hills (e.g. terrace remnants) these might be considered to be part of that piedmont. If so, a new part of relationship is added to the rule base (7) and the terrain object description is committed to the geographical database. Else, a further check concerns the presence of overlap with adjacent objects (8). In case of overlap, the descriptions of the corresponding higher level objects are revised (9). This may have consequences for lower level terrain objects. Otherwise, the new object is an inclusion when considering

the previous level of description. If the total area of inclusions within the higher level object does not exceed some pre-established limit (10), the new object description is committed to the database. Otherwise, the thematic description of the higher level object, its geometry, or both require revision (9).

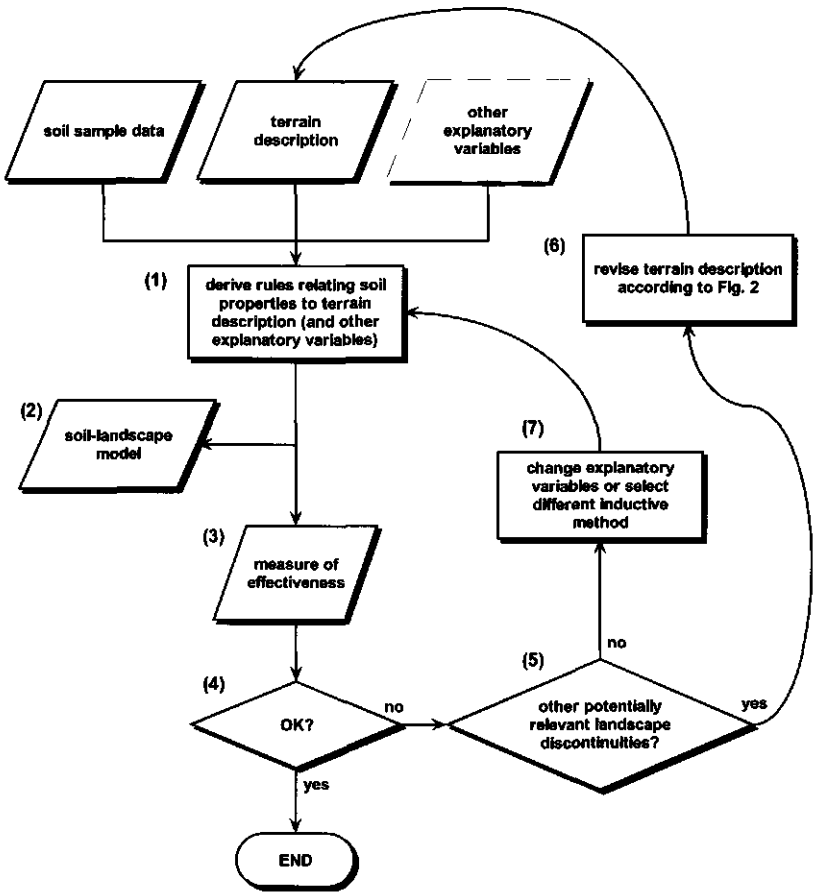


Figure 3.3 Process flow for constructing a soil-landscape model from a set of soil sample data and a terrain description. The numbers between brackets are explained in Section 3.3.2.

3.3.2 Derivation of the soil-landscape model

Figure 3.3 shows how to derive a soil-landscape model from a set of location specific soil sample data, a terrain description resulting from image interpretation, and other explanatory variables, if applicable. The latter may include, for example, attribute data derived from a digital elevation model (e.g. Moore *et al.*, 1993; Thompson *et al.*, 1997; De Bruin and Stein, 1998, see Chapter 4). The type of soil sample data to be used in this exercise depends upon the objective of modelling, i.e. development of a full soil-landscape model or a partial (e.g. single property) soil-landscape model.

Through overlaying, sample sites are labelled according to the terrain object within which they are situated. An inductive method is chosen and the soil sample data and explanatory variables are analysed to estimate the corresponding model parameters (1). The initial assumptions are that the used data are reliable, and that both the inductive method and the terrain description are suitable for soil-landscape modelling. The results are a soil-landscape model (2), and a measure of the effectiveness of that model (3). If these are satisfactory the soil-landscape model is accepted (4). Otherwise, if other relevant landscape discontinuities are expected, the terrain description is revised (6) by means of further image interpretation (Figure 3.2). Else, the method of analysis is adjusted in an attempt to improve the soil-landscape model (7).

3.4 Case study

3.4.1 Materials and methods

The case study concerns part of the drainage basin of the river Guadalhorce, north of the village of Alora in Malaga province, southern Spain. The total survey area is 38.5 km^2, centred on 36°51′29″N, 4°42′43″W. Elevation varies between 100 and 630 m above sea level. A sample area of approximately 200 ha was selected for more detailed study (see Figure 3.4).

We used a standard mirror stereoscope to interpret stereo aerial photography at approximate scales of 1: 20 000 (October 1992) and 1: 25 000 (June 1990). The photographs covering the 200 ha sample area were enlarged to approximately 1: 5 000 to allow a more detailed interpretation and a precise plotting of terrain object boundaries. Image interpretation was according to the stepwise method outlined in section 3.2.2. A geological map digitised from the Alora (IGME, 1978) and Ardales (ITGE, 1991) map sheets served as ancillary information to the image interpretation. The terrain objects were digitised on-screen with a digital orthophoto produced from the October 1992 photography in the background. For visualisation purposes we used a 10 m resolution digital elevation model (DEM) derived from elevation data produced by automated image matching of the October 1992 photography.

Soil data were collected in the sample area from 55 soil pits and auger holes. Sampling involved qualitative description and field estimation of several soil properties (parent material, drainage class, root restricting depth, identification and depth of horizons, matrix and mottle colours, texture, structure (pits only), amount, size and type of rock fragments, and reaction to HCl). In the following we will use particle size classes of the control section according to the family differentiae of Soil Taxonomy (Soil Survey Staff, 1996) to characterise soil variability. We decided to use particle size classification because it provides a practical grouping of the soils in the study area, which could be conveniently derived from the field observations. Other soil properties were less discriminating (drainage class, matrix and mottle colours, reaction to HCl) or could not be consistently determined in the field (root restricting depth). As such, the case study serves to demonstrate the derivation of a partial soil-landscape model according to the framework formulated in this paper.

We used contingency tables to analyse soil-landscape relationships since the explanatory variable (terrain object class) and the soil classes built from textural differentiation are both nominal variables. The degree of association between the variables was assessed using the λ of Goodman and Kruskal (1954). The employed measure, λ_{PSC}, denotes the relative decrease in probability of error in predicting the soil particle size class as between the terrain object class unknown (1) and known (2):

$$\lambda_{PSC} = \frac{(\text{Prob. of error in case } 1) - (\text{Prob. of error in case } 2)}{(\text{Prob. of error in case } 1)}$$

The GIS software used for digitising, building and maintaining topology, overlaying and visualisation was ARC/INFO version 7.1.1 running on a DEC Alpha under Digital UNIX. As yet, operations related to the use and maintenance of the rule base (Figure 3.2, steps 3, 5-8, 10) were carried out manually. Cross tabulation was performed using SPSS 7.5 on a Windows 95 platform.

3.4.2 Stepwise image interpretation

Figure 3.4 shows our initial segmentation of the study area. Delineations reflect major differences in relief and result from interpretation of the aerial photographs without using additional information sources. In the subsequent steps of image interpretation we always operated in downhill direction, starting from the highest locations.

Figure 3.5 shows the terrain description after the second step of image interpretation, now supported by lithological data obtained by overlaying the digitised geological map on the orthophoto. The terrace remnant-capped hills and different transportational slopes of the mountain with table-shaped crest (Figure 3.5) were identified during a brief field visit. Boundaries correspond to breaks of slope and often differ from the boundaries on the geological map. Naming of terrain objects was also based on our own observational data, because the geological map is likely to contain local inaccuracies as a consequence of generalisation and correlation with areas outside our survey area.

Starting from an initial set of potential *part of* relationships derived from the nine unit landsurface model (Dalrymple *et al.*, 1968; Conacher and Dalrymple, 1977) (Table 3.1), we needed to update the rule base during the second step of image interpretation. The class labels assigned to terrain objects considered to be part of previous level objects did not fit the (empty) domains determined with these relationships (Figure 3.2, steps 2-6). The new domain specifications added to our rule base are listed in Table 3.2.

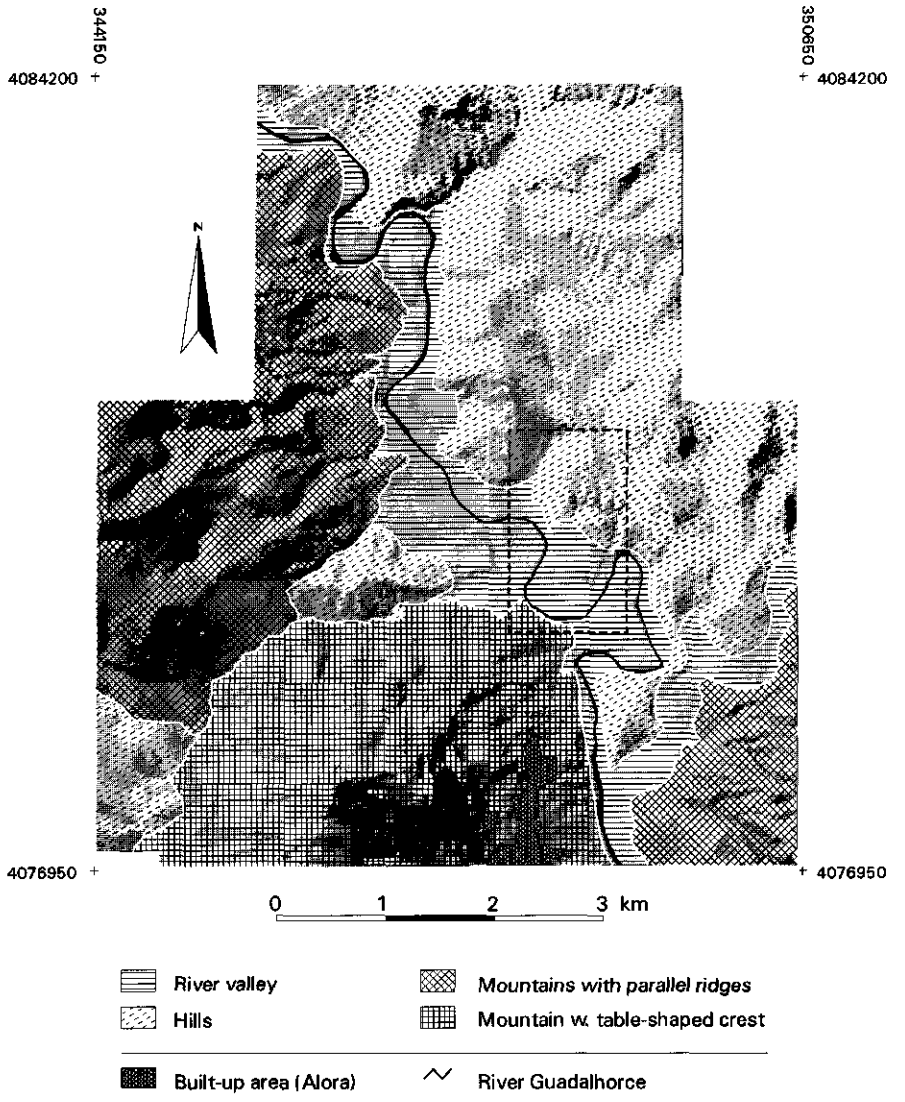


Figure 3.4 First step of the image interpretation against a shaded relief backdrop artificially illuminated from the north-west. The dashed black line to the right of the centre of the image delimits the 200 ha sample area that was studied in more detail. The co-ordinates (m) correspond to UTM zone 30.

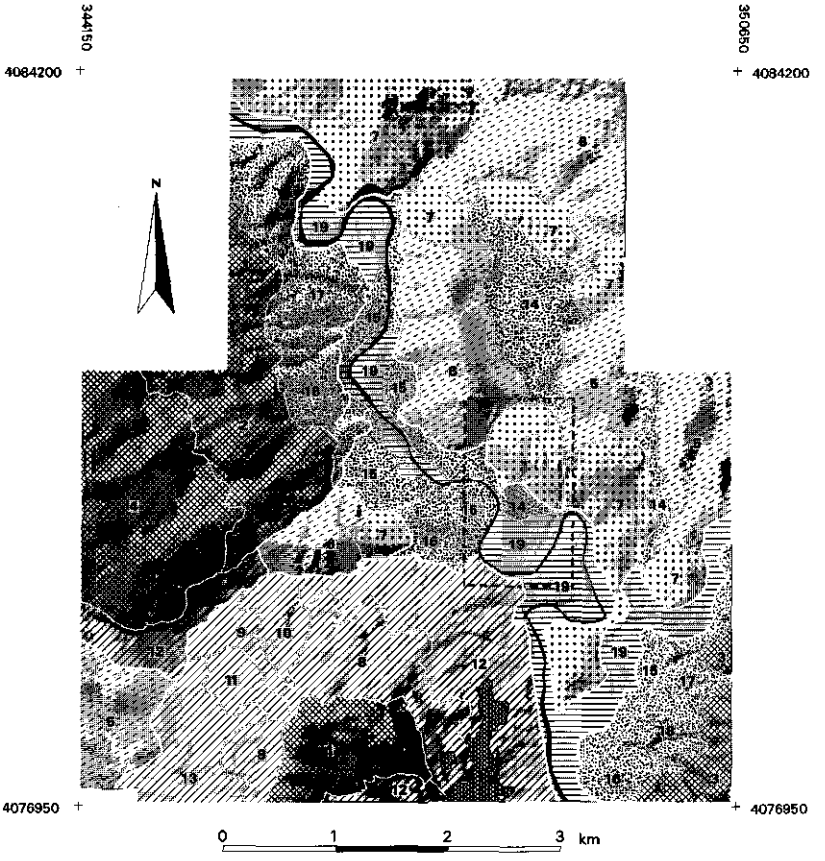
Table 3.1 Potential part of relationships derived from the nine unit landsurface model (Dalrymple *et al.*, 1968; Conacher and Dalrymple, 1977).

Higher level terrain object	Possible lower level objects
Hill / Mountain	Interfluve
	Seepage slope
	Convex creep slope
	Fall face
	Transportational midslope
	Colluvial footslope
River valley	Alluvial toeslope
	Channel wall
	Channel bed

In the first instance, we encountered overlapping terrain objects during further segmentation of the previously delineated river valley. Within the valley we found differently aged alluvial fans adjacent to the mountains (Figure 3.5, labels 15, 16) and a footslope with colluvium derived from marl and limestone-rich materials neighbouring the hills (Figure 3.5, label 14). According to rules established during segmentation of the mountains and hills, the fans are part of the mountains, whereas the colluvial footslope is part of the hills (see Table 3.2). Following Figure 3.2 (steps 8, 9), we adjusted the higher level objects' geometry by joining the fan area to the mountain land and the footslope to the hills. The boundary between the footslope and the terrace remnant-capped hill was slightly modified to represent geometry as determined during the second step of image interpretation. The fan area on the east side of the river does not overlap with the adjacent hills, because Table 3.2 contains no entry for alluvial fans in the domain specification of hill components.

The terrain description after the third step of image interpretation is given in Figure 3.6. The interpretation concerns the sample area, here shown in perspective view from the south-west to improve perceptibility of surface topography. It is based on visual analysis of the enlarged stereo air photos and on parent material data from the 55 sample locations. At this level, the terrain description differentiates the regimes of soil formation according to contrasts in landform, lithology and relative age of the landforms. However, the terrain objects still encompass important topographic variation.

Figure 3.5 Second step of the image interpretation. Thick white lines denote terrain objects delineated during the first disaggregation step. The thinner white lines correspond to the second disaggregation step.



HILLS IN HARD ROCK

1 Conglomerate (summit)
2 Gneisses
3 Phyllites
4 *Serpentinite*

HILLS IN MARL

5 Undifferentiated
6 With sandstone
7 Terrace remnant-capped

TRANSPORTATIONAL SLOPES

Conglomerate blocks over marl
8 Large blocks
9 Small blocks
10 Rockslides

11 Debris avalanche and rockcreep

Debris over marl
12 Undifferentiated
13 With landslides

FOOTSLOPES

Mainly colluvial deposits
14 Marl + limestone-rich materials

Alluvial fans
15 Young
16 Old
17 *Very old, with terrace remnants*
18 Very old, strongly incised

FLUVIAL TERRACES

19 Limestone-rich deposits

Built-up area (Alora)

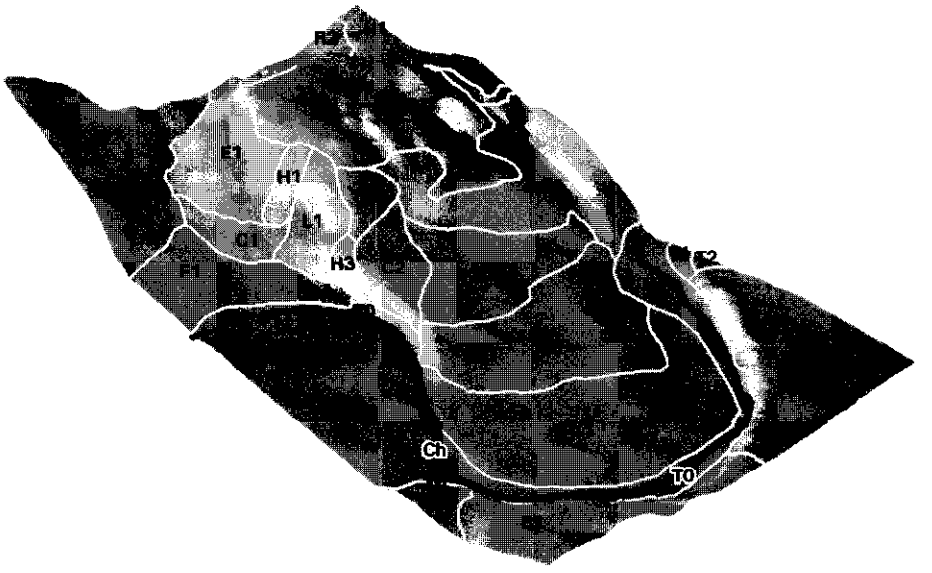
River Guadalhorce

Boundary of sample area

Table 3.2 Domains of class labels for terrain objects after the second step of image interpretation.

Higher level terrain object	Possible lower level objects (parts)
Mountain with table-shaped crest	Summit in conglomerate Transportational slope, large conglomerate blocks over marl Transportational slope, small conglomerate blocks over marl Transportational slope, rockslide of conglomerate over marl Transportational slope, debris avalanche and rockcreep, conglomerate over marl Transportational slope, conglomerate debris over marl Transportational slope with landslides, conglomerate debris over marl
Mountains with parallel ridges	Hills in gneiss Hills in phyllites Hills in serpentinites Young alluvial fans Old alluvial fans Very old alluvial fans with terrace remnants Very old, strongly incised alluvial fans
Hills	Hills in marl Hills in marl with sandstone Terrace remnant-capped hill in marl Footslope with colluvium derived from marl and limestone-rich materials
River valley	Limestone-rich fluvial terraces Young alluvial fans River channel

We subdivided the terrace remnant-capped hill, the colluvial footslope and the alluvial plain of the previous step, while the geometry of the other previous level objects remained unchanged. On the south-west and west slopes of the terrace remnant-capped hill we found residual hills (Figure 3.6, labels H1 and H3), which we did not regard as common constituents of the higher level object class. However, because they occupied only 4.3% of the area they were permitted as inclusions in the terrace remnant-capped hill (Figure 3.2, steps 8, 10). Figure 3.7 provides an overview of all successive splits of the hierarchical breakdown of the sample area into its lowest level terrain objects, and the links between spatially nested terrain objects.



C1	Footslope, mixed calcareous colluvium	H3	Residual hill, calcareous sandstone
C2	Footslope, colluvium derived from marl	L1	Landslide, marl
Ch	River channel, limestone-rich	R1	Terrace remnant, mixed calcareous
E1	Transportational slope, calcareous materials over marl	R2	Terrace remnant, calcareous mat. over marl
E2	Transportational slope, conglomerate debris over marl	T0	Floodplain, limestone-rich
F1	Recent alluvial fan, mixed calcareous	T1	Recent terrace, mixed calcareous
F3	Old alluvial fan, mixed calcareous	T2	Subrecent terrace, mixed calcareous
H1	Residual hill, marl	T3	Moderately old terrace, mixed calcareous
H2	Residual hill, marl and sandstone		

Figure 3.6 Third step of the image interpretation. The interpretation concerns the sample area, here shown in perspective view from the south-west.

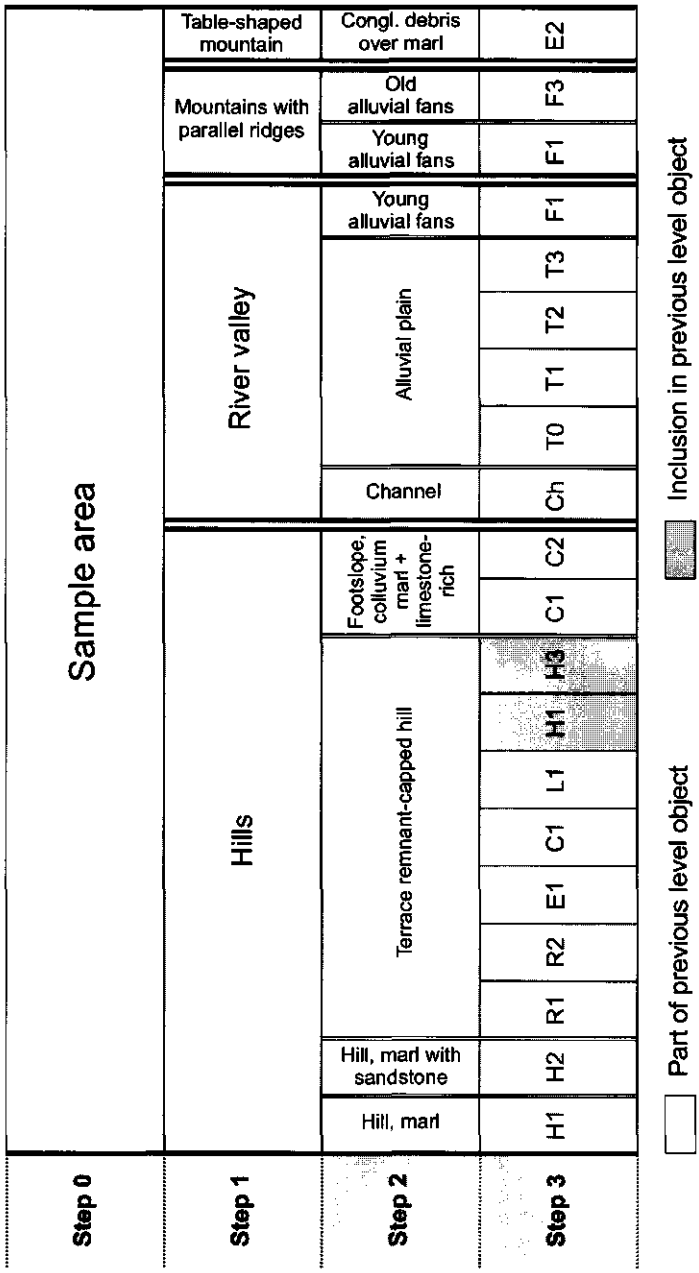


Figure 3.7 Diagram representing the hierarchical breakdown of the sample area, after removal of overlap. The labels in each box denote object classes, not individual objects. Repeating lower level – higher level object relationships are indicated only once. See Figure 3.6 for an explanation of the abbreviations (step3).

3.4.3 Soil-landscape relationships

Table 3.3 is a contingency table of particle size class by terrain object class after the first step of image interpretation (Figure 3.4). The table represents a preliminary partial soil-landscape model, where percentages between brackets are conditional probabilities of particle size classes given the terrain object class.

Within the river valley five out of the seven particle size classes were observed, while as much as six classes were found within the hilland. There were no observations within the mountain with table-shaped crest. At this stage, λ_{PSC} was equal to 0.06, indicating that knowledge of the terrain object class would reduce the probability of error in predicting a single particle size class by 6%. As the first segmentation did not account for all landscape discontinuities, we expected that a more elaborate terrain description would improve the soil-landscape model (Figure 3.3, step 6).

Table 3.3 Contingency table of particle size class by terrain object class after the first segmentation of the sample area.

Particle size class[a]	Terrain object class[b]				Total[b]	
	River valley		Hills			
	n	%	*n*	%	*n*	%
CL	5	(22.7)	3	(9.1)	8	(14.5)
CL/F			2	(6.1)	2	(3.6)
COAR	4	(18.2)	17	(51.5)	21	(38.2)
F	6	(27.3)	9	(27.3)	15	(27.3)
F/S			1	(3)	1	(1.8)
FL	6	(27.3)	1	(3)	7	(12.7)
FL/S	1	(4.5)			1	(1.8)
Total	22	(100)	33	(100)	55	(100)

[a] CL = coarse loamy; CL/F = coarse loamy over clayey; COAR includes the classes loamy-skeletal, sandy and sandy-skeletal; F = fine; F/S = clayey over sandy; FL = fine loamy; FL/S = fine loamy over sandy.

[b] The numbers on the left are frequency counts; the numbers between brackets denote percentages of the total number of observations within the terrain object class.

Using the terrain descriptions resulting from the second and third steps of image interpretation, λ_{PSC} improved to 0.24 and 0.38, respectively. Table 3.4 is the contingency table of particle size class by terrain object class after the third step of image interpretation. The soil-landscape model could be simplified by combining the terrain object classes C2, H1 and L1 (landforms developed in marls and clays), without affecting λ_{PSC}. The recent depositional terrain object classes C1 and F1 still exhibit much internal variation in particle size distribution, probably as a result of sediment sorting. Variability within the classes E1, H2 and R2 can probably be attributed to varying proportions of the

different parent materials within terrain objects of these classes. To further improve the partial soil-landscape model, it should account for textural variation related to gradual topographic differences or employ spatial interpolation of data from additional observation points. There are no clear topographic discontinuities that allow unambiguous delineation of more detailed terrain objects.

Table 3.4 Contingency table of particle size class by terrain object class after the third step of image interpretation.

Particle-size class[a]		Terrain object class[b]											Total
		C1	C2	E1	F1	H1	H2	L1	R1	R2	T1	T3	
CL	*n*	1		1	2				2		2		8
	%	14.3		12.5	25				13.3		66.7		14.5
CL/F	*n*			1						1			2
	%			12.5						50			3.6
COAR	*n*	3		3	2		1		12				21
	%	42.9		37.5	25		50		80				38.2
F	*n*	1	1	3		2	1	1		1		5	15
	%	14.3	100	37.5		100	50	100		50		83.3	27.3
F/S	*n*								1				1
	%								6.7				1.8
FL	*n*	2			3						1	1	7
	%	28.6			37.5						33.3	16.7	12.7
FL/S	*n*				1								1
	%				12.5								1.8
Total	*n*	7	1	8	8	2	2	1	15	2	3	6	55
	%	100	100	100	100	100	100	100	100	100	100	100	100

[a] Notation as in Table 3.3.

[b] C1 = footslope, mixed calcareous colluvium; C2 = footslope, colluvium derived from marl; E1 = transportational slope, calcareous materials over marl; F1 = recent alluvial fan, mixed calcareous; H1 = residual hill, marl; H2 = residual hill, marl and sandstone; L1 = landslide, marl; R1 = terrace remnant, mixed calcareous; R2 = terrace remnant, calcareous materials over marl; T1 = recent terrace, mixed calcareous; T3 = moderately old terrace; mixed calcareous.

3.5 Discussion and conclusions

Analysis of topological and thematic relationships among terrain objects resulting from stepwise image interpretation provides a basis for interactive formalisation of soil-landscape knowledge. We formulated and demonstrated a framework in which this concept has been worked out. Although a full integration of image interpretation with database updating was not implemented in our case study, it is in principle possible due to recent advances in software engineering (Woodsford, 1996). Especially the move to open system architectures would enable effective implementation of the framework.

The essential qualities of the framework are listed below.

(1) It creates a rule base with possible *part of* relationships between nested terrain objects. This rule base, being of great help for image interpretation, can be transferred for application in similar areas, for example after studying reference areas (cf. Lagacherie *et al.*, 1995). Although the rule base will probably require updating, it may provide a useful starting-point for interpretation of similar landscapes.

(2) The framework allows visualisation of the pathway along which terrain objects have been derived.

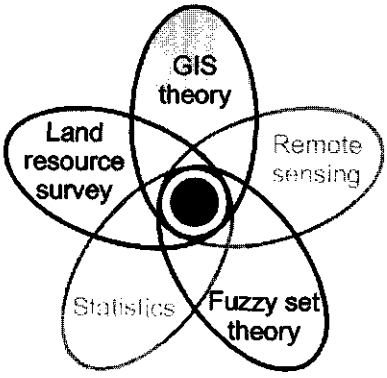
(3) Though not demonstrated in our case study, it is expected that the framework can support different methods for analysing and describing soil variation in relation to a terrain description. Our case study used stratification according to terrain object classes. To confirm the framework's flexibility, future research should include testing it with other methods.

Major limitations of the framework are: (1) It does not explicitly deal with object identification criteria (e.g. which evidence supports identification of an alluvial fan), nor does it handle rules for deciding upon the exact location of object boundaries; (2) the image interpretation method cannot handle ambiguity. Decisive moments indicated in Figure 3.2 require straight yes or no answers, each terrain object must be definitely assigned to one class, and object boundaries are sharp lines.

Limitation 1 is partly inherent to the hardly understood process of human image interpretation in which complex spatial analyses play an essential role (Campbell, 1983). In part it is compensated for by the GIS visualisation tools that provide the means to create images (e.g. perspective views) illustrating the landscape configuration of interpreted terrain objects. These images may serve as examples for terrain object delineation elsewhere. Though not included in this paper, limitation 1 may be overcome by complementing the description of terrain objects with information on the evidence used for their identification.

As a consequence of limitation 2, the interpretation method is of most utility when there is no doubt about the identity and boundary locations of terrain objects. In stepwise image interpretation these circumstances prevail at higher aggregation levels, where object identity is evident and boundaries usually agree with obvious breaks of slope. At lower aggregation levels there is more confusion involved in terrain description, since topographic variability appears to be gradual or the information needed for proper object identification is lacking.

An obvious solution to limitation 2 is therefore to apply the interpretation method outlined in Figure 3.2 only to higher aggregation levels. Lower aggregation levels require different techniques that can deal with gradual variation and/or take into account the ambiguity involved in terrain description. Figure 3.3 indicates that such techniques, if formulated as an inductive method, can be integrated into the general framework presented in this paper. Examples of such an integration are stratified spatial interpolation (e.g. Stein *et al.*, 1988; McBratney *et al.*, 1991; Boucneau *et al.*, 1998), or fuzzy *c*-means clustering to represent transition zones within crisp terrain objects (e.g. De Bruin and Stein, 1998, see Chapter 4). This suggests that our framework is compatible with the evolving soil survey paradigm recognised by Burrough *et al.* (1997). However, further research is required to support decision making as to the aggregation level at which crisp terrain description should be abandoned in favour of a continuous model.


GIS
theory
Land
resource
survey
Remote
sensing
Statistics
Fuzzy set
theory

4 Soil-landscape modelling using fuzzy *c*-means clustering of attribute data derived from a DEM[1]

Abstract

This study explores the use of fuzzy *c*-means clustering of attribute data derived from a digital elevation model to represent transition zones in the soil-landscape. The conventional geographic model used for soil-landscape description is not able to properly deal with these. Fuzzy *c*-means clustering was applied to a hillslope within a small drainage basin in southern Spain. Cluster validity evaluation was based on the coefficient of determination of regressing topsoil clay data on membership grades. The resulting clusters occupied spatially contiguous areas. A high degree of association with measured topsoil clay data ($r_a^2 = 0.68$) was found for three clusters and a weighting exponent of 2.1. Location of the clusters coincided with observable terrain characteristics. Therefore it was concluded that the coefficient of determination of regressing soil sample data on membership grades efficiently supports deciding upon the optimum fuzzy *c*-partition. The study confirms that fuzzy *c*-means clustering of terrain attribute data enhances conventional soil-landscape modelling, as it allows representation of the fuzziness inherent to soil-landscape units.

4.1 Introduction

The soil-landscape model enables soil scientists to accurately predict soil types and their associated properties using the relation between soils and landscape features and a limited set of soil observations. The model is built on the concept of soil-landscape units. These are natural terrain units, with observable form and shape, resulting from the interactions of the five factors affecting soil formation, namely parent material, climate, organisms, relief and time (Jenny, 1941; Hudson, 1990, 1992). In conventional soil

[1] Based on: De Bruin, S., and Stein, A., 1998. Soil-landscape modelling using fuzzy c-means clustering of attribute data derived from a Digital Elevation Model (DEM). Geoderma 83, 17-33.
© 1998 Elsevier Science B.V.

survey, the surveyors traverse the landscape looking for areas where landscape features change rapidly within a relatively short distance. Such areas are marked as boundaries of soil-landscape units, since a concomitant change in soil properties typically occurs at the same zone and within the same lateral distance (Hudson, 1990, 1992). Readers are referred to Hudson (1990) for a more detailed description of the soil-landscape model.

The conventional geographic model used in conjunction with the soil-landscape model is the *discrete object model* (cf. Burrough *et al.*, 1997). The landscape is represented as a series of discrete, interlocking soil volumes of various sizes and shapes (Hole and Campbell, 1985). The model suggests that the soil-landscape consists of internally homogeneous units separated by sharp boundaries. However, such a representation is an approximation and a simplification of a more complex pattern of variation (Lagacherie *et al.*, 1996). Boundaries between soil-landscape units are transition zones rather than sharp boundaries. It is therefore inappropriate to assign a site within a transition zone to any single soil-landscape unit.

In this paper we explore the use of fuzzy pattern recognition to deal with this problem. Alternatives like canonical correlation or regression of soil data on terrain attributes have not been used, as our primary aim is to generalise and modify the crisp soil-landscape model. Zadeh (1965) first introduced the concept of fuzzy sets to deal with classes that do not have sharply defined boundaries. Fuzzy sets are characterised by a membership function that assigns to each element a grade of membership ranging from zero to one. Therefore membership in a fuzzy set is not a matter of yes or no but of a varying degree. Consequently, an element can partially belong to multiple fuzzy sets. Zadeh's paper initiated an important paradigm shift, which affects virtually all sciences since it challenges the firmly established two-valued logic (Klir and Yuan, 1995, pp. 30-32). Readers are referred to McBratney and Odeh (1997) for a broad overview of the application of fuzzy sets in soil science. Recently, fuzzy pattern recognition has been applied to classify, among other things, soil survey data (McBratney and De Gruijter, 1992; Odeh *et al.*, 1992a, 1992b; De Gruijter *et al.*, 1997), environmental pollution data (Hendricks-Franssen *et al.*, 1997) and landforms (Irvin *et al.*, 1997). To obtain complete area coverage, the approach has been either to classify interpolated data (Hootsmans, 1996), or to interpolate class membership grades of measured points (Odeh *et al.*, 1992b; De Gruijter *et al.*, 1997).

The objective of this study is to improve soil-landscape modelling by using fuzzy sets. More specifically, we use the fuzzy *c*-means method to recognise fuzzy patterns in a set of attribute data derived from a digital elevation model (cf. Irvin *et al.*, 1997). We argue that the fuzzy *c*-means method, in combination with a special procedure for cluster validity evaluation, identifies structures that provide important information about the soil-landscape. As opposed to discrete object based modelling, the resulting soil-landscape model preserves the fuzzy nature of soil-landscape units. The approach, as applied to a hillslope within a small drainage basin in southern Spain, is described and discussed.

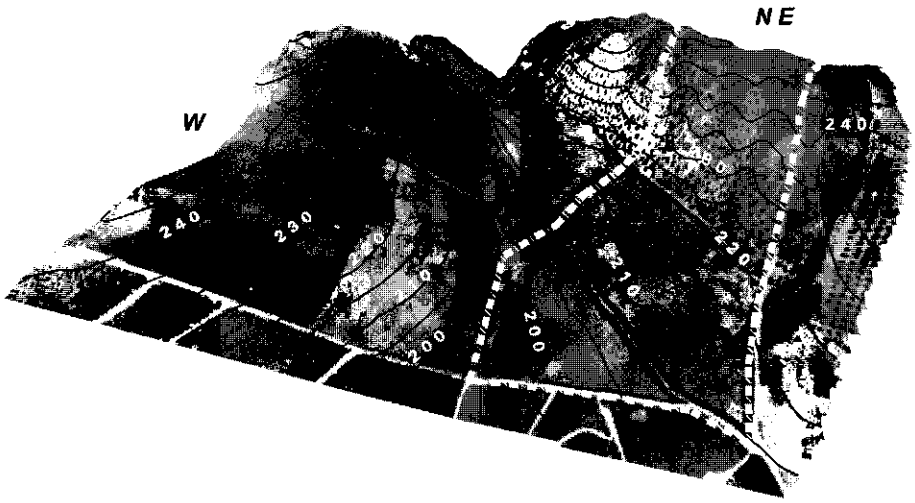


Figure 4.1 Perspective view of the study area, indicated by the thick broken line, and its immediate surroundings. The continuous black lines are the 5 m interval elevation contour lines; the numbers in white indicate the elevation above mean sea level.

4.2 Materials and methods

4.2.1 Study area

The study area is approximately 20 ha, centred on 4°41'55"W 36°52'33"N. The area is about 5.5 km north of Alora. Figure 4.1 provides a perspective view of the study area (indicated by the thick broken line) and its immediate surroundings. It is in the upper part of the drainage basin of a small tributary of the river Guadalhorce. On the west side (marked 'W' in Figure 4.1), the crest of an elongated hill that mainly formed in folded fine textured flysch deposits of early Tertiary age constitutes the watershed. The sharper parts of this crest predominantly consist of sandstone (ITGE, 1991). On the north-east side (marked 'NE' in Figure 4.1) the drainage divide is formed by a hill capped by a severely eroded Pleistocene alluvial terrace remnant, composed of carbonate rich sands and gravels. The crest is underlain by the already mentioned flysch deposits, which also constitute the bedrock of the towards the north-west narrowing valley between the hills (ITGE, 1991). In the north-west an inversion of the general direction of slope delimits the drainage basin. Land use consists of an olive and almond orchard on the terrace remnant, whereas the valley is used for rainfed arable cropping.

The study area has interesting features from the point of view of soil-landscape modelling. First, there are two remarkably different landforms with contrasting lithology. These landforms are easily recognised and mapped, since the transition from the terrace remnant to the valley is rather abrupt. Second, there are areas within the two landforms that are not easily mapped. Alluvium from the terrace remnant is unequally distributed over the adjacent valley. However, it is not at all obvious to draw lines between areas

with different sedimentation patterns. Moreover, within the terrace remnant it is impossible to definitely separate gullies from interfluves. Apparently, their boundaries are not clear, which makes the area particularly suitable to illustrate the fuzzy set approach.

4.2.2 Terrain attribute data

There are several efficient methods to obtain elevation data for development of accurate, high resolution Digital Elevation Models (DEMs). We created a 5 m resolution DEM from elevation data obtained by automated image matching of scanned panchromatic aerial photographs. This was done by interpolation using the ARC/INFO procedure TOPOGRID. This procedure employs an algorithm developed by Hutchinson (1989) to create hydrologically sound DEMs. We made use of TOPOGRID's optional drainage enforcement routine to automatically remove spurious sinks. In addition we digitised the location and flow direction of the major stream in the valley and used it as an extra input for the interpolation procedure. The resulting DEM comprised 8412 elevation points on a regular grid.

Terrain analysis algorithms enable the calculation of topographic attribute data from a DEM (e.g. Moore *et al.*, 1991; Quinn *et al.*, 1995; Blaszczynski, 1997). The relative magnitudes of many hydrological, geomorphologic and biological processes active in the landscape are sensitive to topographic position (Moore *et al.*, 1991). The spatial distribution of terrain attributes can therefore be used as an indirect measure of the spatial variability of these processes. Odeh *et al.* (1991, 1994) found that slope, plan curvature (contour curvature), profile curvature (slope profile curvature) and upslope distance and area accounted for much of the soil variation in their study area. Moore *et al.* (1993) and Gessler *et al.* (1995) also found that particularly attributes that characterise the distribution of hydrological processes were significantly correlated with soil properties. These findings support that the soil catena develops in response to the way water flows through the landscape (Moore *et al.*, 1993). In our analysis we used elevation, slope, plan curvature, profile curvature, stream power index and wetness index to characterise surface topography. Within the study area, elevation is indicative of lithological differences (see Figure 4.1). The other attributes were adopted because their spatial distributions are often related to spatial variability of erosion, sediment transport and sedimentation (Moore *et al.*, 1991).

The attributes slope, plan curvature and profile curvature were calculated with ARC/INFO's GRID function CURVATURE, which implements an algorithm developed by Zevenbergen and Thorne (1987). The stream power and wetness indices were respectively calculated using the equations:

$$\text{stream power index} = a \tan \beta \tag{4.1}$$

and

$$\text{wetness index} = \ln\left(\frac{a}{\tan \beta}\right) \tag{4.2}$$

where β is the slope angle and a is the upslope area per unit width of contour (Moore *et al.*, 1993). Calculation of a was based the multiple-direction contributing area algorithm of Freeman (1991).

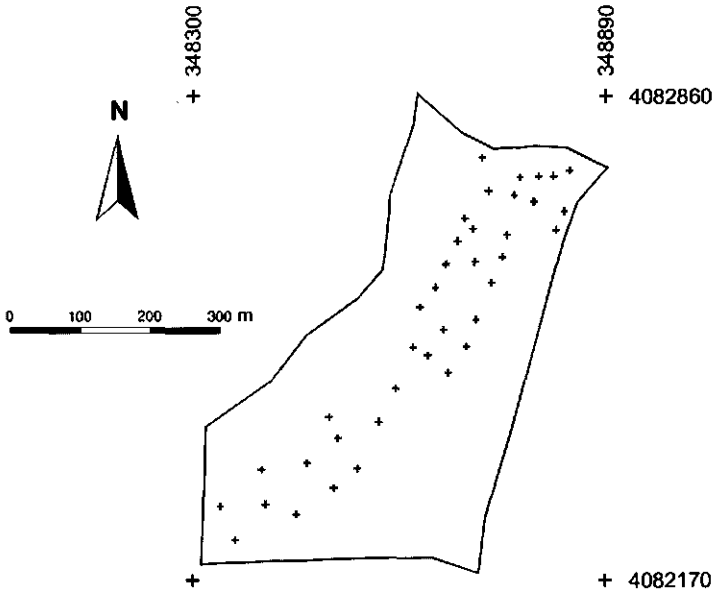


Figure 4.2 Location of the sample points in UTM projection. The co-ordinates (m) correspond to UTM zone 30.

4.2.3 Topsoil clay data

A total of 38 samples from 0-15 cm depth was collected within a 125 m wide and 710 m long transect strip, oriented with its longest axis parallel to the general direction of slope (Figure 4.2). The sample locations were chosen at representative sites (i.e., different sites within valley, gullies and interfluves), to reflect both general and local topsoil variations (purposive sampling). Such a sampling strategy is often used in landscape-based soil surveys, especially when the number of samples is small in comparison with terrain complexity (Webster and Oliver, 1990, pp. 29-30).

The clay content of the fine earth fraction (< 2 mm) was measured in the laboratory using the pipette method after wet combustion of organic matter and removal of carbonates (Van Reeuwijk, 1992). We selected topsoil clay content to characterise soil variability because it may reflect differences in lithology as well as erosion and sedimentation. As such it serves as an indicator for a full soil-landscape model. In a previous study topsoil clay content was found to be closely related with surface topography in the study area (De Bruin and Wielemaker, 1996).

4.2.4 Fuzzy c-means clustering

Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ denote a data set with n elements, each represented by a vector of measurements on p continuous variables. Our aim is to subdivide $\mathbf{X}$ into c partially overlapping, non-empty subsets. A collection of these subsets is a fuzzy c-partition of $\mathbf{X}$, denoted by $P = \{A_1, A_2, ..., A_c\}$, if it satisfies the following conditions:

$$0 \le \mu_{A_i}(\mathbf{x}_j) \le 1 \qquad \forall \quad i \in \{1, ..., c\}, \quad j \in \{1, ..., n\} \tag{4.3}$$

$$\sum_{i=1}^{c} \mu_{A_i}(\mathbf{x}_j) = 1 \qquad \forall \quad j \in \{1, ..., n\} \tag{4.4}$$

$$\sum_{j=1}^{n} \mu_{A_i}(\mathbf{x}_j) > 0 \qquad \forall \quad i \in \{1, ..., c\} \tag{4.5}$$

where $\mathbf{x}_j$ is the data vector of the kth element of $\mathbf{X}$ and $\mu_{A_i}(\mathbf{x}_j)$ represents the membership grade of individual $\mathbf{x}_j$ in subset A_i (Klir and Yuan, 1995). The combination of the first two conditions implies that elements of $\mathbf{X}$ can partially belong to different subsets simultaneously. This distinguishes fuzzy c-partitions from hard partitions, where the membership grades are either zero or one so that the subsets of $\mathbf{X}$ are disjoint. The conditions (Eqs. 4.3 and 4.4) also suggest that $\mu_{A_i}(\mathbf{x}_j)$ can be interpreted as the conditional probability of subset A_i given element $\mathbf{x}_j$. However, membership grades are often combined in ways that are not restricted to established statistical methodology (Laviolette *et al.*, 1995; Rousseeuw, 1995). Moreover, the probability that element $\mathbf{x}_j$ is a member of subset A_i assumes a crisp definition of A_i. Conversely, the membership grade of $\mathbf{x}_j$ in A_i expresses the degree with which $\mathbf{x}_j$ meets the central concept of A_i (Burrough *et al.*, 1997; see also Chapter 7). The third condition (Eq. 4.5) excludes the null set from fuzzy c-partitions.

Fuzzy clustering is aimed at finding a fuzzy c-partition and the associated subset prototypes, being cluster centres in a p-dimensional space, that best represent the structure present in the data set. There are several fuzzy clustering methods, which are different in the criteria that are used to identify the optimum fuzzy partition of the data set (Bezdek, 1981). A commonly applied method is the fuzzy c-means (FCM) procedure, defined by the generalised least-squares objective function:

$$J(P) = \sum_{j=1}^{n} \sum_{i=1}^{c} [\mu_{A_i}(\mathbf{x}_j)]^{\phi} ||\mathbf{x}_j - \mathbf{v}_i||^2 \tag{4.6}$$

where ϕ is a weighting exponent, $\phi \ge 1$, and $||\mathbf{x}_j - \mathbf{v}_i||^2$ is the distance (dissimilarity) between the element $\mathbf{x}_j$ and cluster centre $\mathbf{v}_i$ according to some inner product-induced norm. The data vector of the cluster centres $\mathbf{v}_i$, $i \in \{1, ..., c\}$, are found by iteration with:

$$\mathbf{v}_i = \frac{\sum_{j=1}^{n} [\mu_{A_i}(\mathbf{x}_j)]^{\phi} \mathbf{x}_k}{\sum_{j=1}^{n} [\mu_{A_i}(\mathbf{x}_j)]^{\phi}} \tag{4.7}$$

Distance norms commonly used in combination with Equation 4.6 are the Euclidean norm, the diagonal norm and the Mahalonobis norm. The Euclidean norm assigns equal weight to all measured variables and is therefore appropriate when the clusters in **X** have the general shape of hyper-spheres. The diagonal norm compensates for distortions in this spherical shape caused by dissimilarities in variances among the measured variables. The Mahalonobis norm compensates for both dissimilarities in variances and linear correlations among the measured variables (Bezdek, 1981).

The degree of fuzziness in the resulting partition is controlled by the weighting coefficient ϕ. When ϕ approaches 1, the partition becomes hard. With large values for ϕ the cluster centres tend towards the centroid of the data set and the partition becomes fuzzier. Currently there is no theoretical basis for an optimum choice for the value of ϕ (Klir and Yuan, 1995). In this paper we will present an empirical way to decide upon the value of ϕ.

Bezdek (1981) developed an iterative FCM algorithm that converges to local minima of $J(P)$, for $\phi > 1$. We implemented this algorithm in DEC Pascal for Alpha AXP and ran the program on an Alpha AXP under DEC OSF/1. The terrain attribute data described in Section 4.2.2 constituted the data set for fuzzy c-means clustering. To account for the dissimilarities of variances and statistical dependencies among the six terrain attributes, the distance calculations were based on the Mahalonobis norm.

4.2.5 Cluster validity

For different specified numbers of clusters, $c \in \{2, 3, ..., n\}$ and virtually any value of the weighting exponent, $\phi > 1$, the fuzzy c-means algorithm produces cluster centres and membership grades. Several cluster validity functionals have been proposed to help decide whether the algorithmically suggested clusters provide a meaningful partition of the data set. Bezdek (1981) and Roubens (1982) describe a number of scalar measures of partition fuzziness to estimate the optimum value of c. A commonly used measure is the normalised partition entropy (Roubens, 1982; Odeh *et al.*, 1992a), a Shannon entropy-like measure defined by:

$$H' = \frac{-\sum_{i=1}^{c}\sum_{j=1}^{n} \mu_{A_i}(\mathbf{x}_j) \log_b \mu_{A_i}(\mathbf{x}_j) / n}{\log_b c} \qquad (b > 1) \tag{4.8}$$

For a discussion on classification entropy we refer to Bezdek (1981, pp. 109-118). McBratney and Moore (1985) suggest a measure based on a derivative of the objective function, $-\frac{d}{d\phi} J(P)\sqrt{c}$, to evaluate the optimum combination of c and ϕ simultaneously.

Since all these measures are based on information obtained from within the clustering process itself, they are so-called *internal criterion* measures (Milligan, 1996). Conversely, *external criterion* analysis is based on variables not used in the cluster analysis (Milligan, 1996). Internal criterion analysis is not always sufficient to identify the most suitable partition of a data set. A partition is useful only if the results can be understood within the context of the research question at issue (Molenaar, 1993; Milligan, 1996). In this way, the fuzzy *c*-partition of our set of terrain attribute data is most useful if spatial distribution of the membership grades shows a strong relationship with spatial variability of soil properties. Internal criterion analysis cannot test for this relationship. We therefore propose to evaluate validity of the fuzzy *c*-partitions based on their relation with external soil sample data. The approach is a generalisation of external criterion analysis used to evaluate hard partitions of a data set.

Standard hypothesis testing techniques such as analysis of variance provide means to evaluate and compare hard classifications (Webster and Oliver, 1990; Milligan, 1996). However, the analyses are only valid when based on variables that are not used in the clustering analysis (Milligan, 1996). In such cases, partitioning the sum of squared deviations from the general mean allows calculation of the coefficient of determination (r^2), which expresses the proportion of variation in the external variable accounted for by classification. The r^2 coefficient is the complement of the ratio of the within-class sum of squares to the total sum of squares:

$$r^2 = 1 - \frac{\text{within-class sum of squares}}{\text{total sum of squares}} \tag{4.9}$$

The within-class sum of squares is the sum of squared deviations from the mean per class to which the observations belong. Hence, r^2 is a measure of the goodness of fit of the least-squares response of the variable of interest to the classification. The response model can be written as a linear regression equation, where the variable of interest is a function of the binary class memberships, and the means per class are the sample regression coefficients.

We used a similar procedure to evaluate and compare the fuzzy *c*-partitions of the set of terrain attribute data. Considering our objective to identify soil-landscape units, a suitable cluster validity measure is the fuzzy *c*-partition's degree of association with clay content measured at the sample locations. The r^2 coefficient expresses this degree of association for a particular response model (Eq. 4.9). Since fuzzy c-partitions have membership grades ranging from zero to one (Eq. 4.3), the coefficients of the response model cannot simply be estimated from the sample means per cluster. Instead, we used least-squares regression analysis to estimate model parameters. We assumed a linear response of topsoil clay content to the cluster membership grades, employing the regression model:

$$topsoil\ clay = \beta_0 + \beta_1 \mu_{A_1}(\mathbf{x}) + \ \ldots\ + \beta_c \mu_{A_c}(\mathbf{x}) + \varepsilon \tag{4.10}$$

where β_0, β_1, ..., β_c are the regression coefficients and ε is a random error. Substituting *residual sum of squares* for *within-class sum of squares* in Equation 4.9, we calculated r^2

for fuzzy c-partitions corresponding to different values for c and ϕ. To compensate for different numbers of explanatory variables in the regression equation (Eq. 4.10), r^2 was adjusted according to:

$$r_a^2 = r^2 - \frac{c(1-r^2)}{n_{smp} - c - 1} \tag{4.11}$$

where n_{smp} denotes the number of sample points (n_{smp} = 38). The optimum fuzzy c-partition would be the one with the largest r_a^2, i.e., the largest proportion of variation in clay content accounted for by regression.

4.3 Results and discussion

Figure 4.3 shows the effect of different values of ϕ on fuzzy 4-means clustering of the terrain attribute data set. For reason of graphical representation, each element $\mathbf{x}_j$ was labelled according to the cluster to which it has the highest membership grade. That is, for each $\mathbf{x}_j$:

$$\mu_{A_i}(\mathbf{x}_j) = \begin{cases} 1 & \text{if } \mu_{A_i}(\mathbf{x}_j) = \max(\mu_{A_1}(\mathbf{x}_j), \ldots, \mu_{A_c}(\mathbf{x}_j)) \\ 0 & \text{else} \end{cases}, \quad i \in \{1, \ldots, c\} \tag{4.12}$$

The degree of fuzziness associated with this labelling can be read from the grey shade, which is proportional to a confusion index value. The latter was calculated as one minus the difference between the largest and the second largest membership grade of element $\mathbf{x}_j$, as suggested by Burrough (1996). Accordingly, the confusion index is largest when the difference in membership grades of the two most important clusters for the element is smallest, i.e., when there is much confusion about the cluster to which the element should be assigned (Burrough *et al.*, 1997).

Figure 4.3 clearly shows the effect of the value for ϕ on partition fuzziness. In Figure 4.3a (ϕ = 1.4), areas with maximum confusion are confined to the gully bottoms and interfluvial crests within the terrace remnant, and rather narrow boundary zones between spatially contiguous regions with relatively homogeneous membership grades. In Figure 4.3d (ϕ = 2.2) the two clusters corresponding to the terrace remnant (A_1 and A_2) are completely confused, while only cluster A_3 in the valley differs substantially from the other clusters. The best choice of ϕ is not clear from Figure 4.3, though.

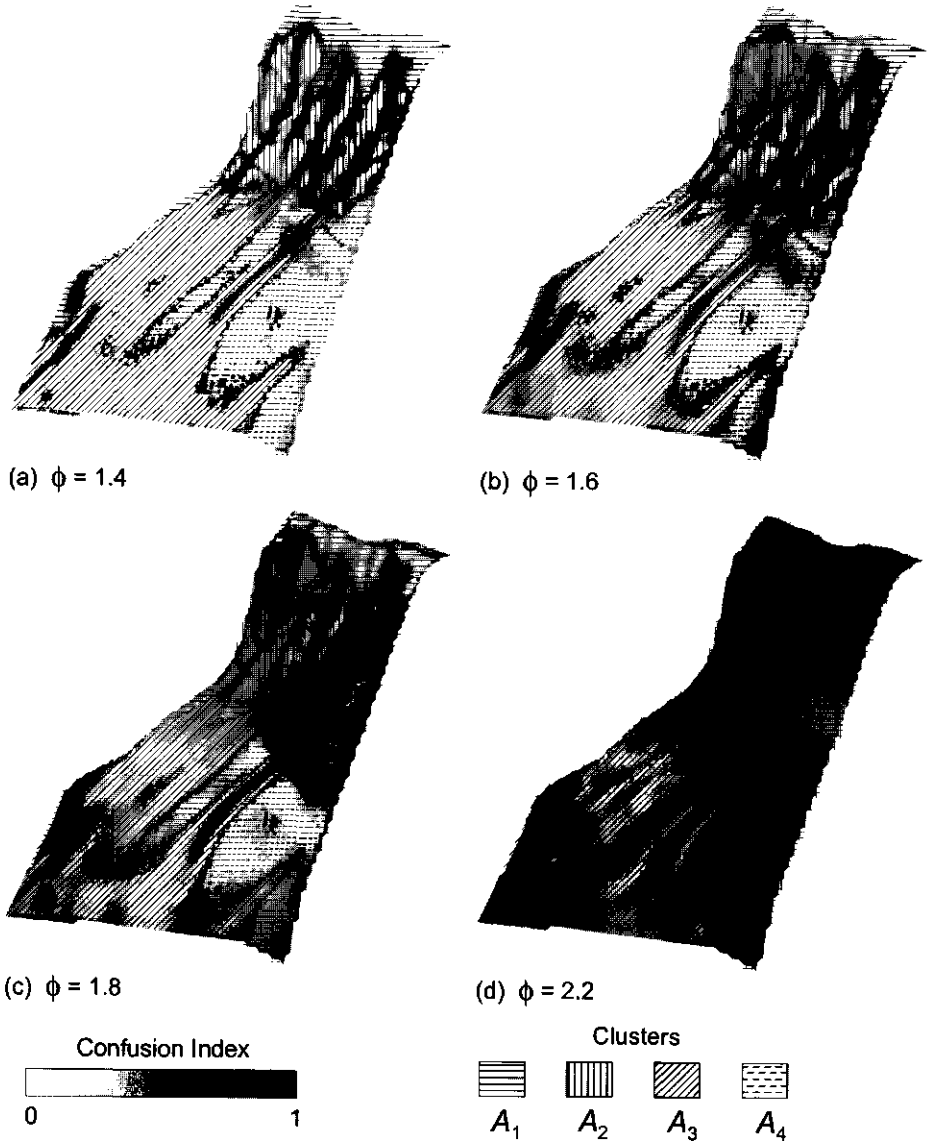


Figure 4.3 Effect of different values of the weighting coefficient ϕ on fuzzy 4-partitions of the terrain attribute data set, in perspective view.

Figure 4.4 is a plot of the adjusted coefficient of determination of regressing topsoil clay content on cluster membership grades (see Section 4.2.5). The plotted data points correspond to fuzzy c-partitions that resulted from clustering with different combinations of ϕ and c. The optimum fuzzy c-partition is obtained with $c = 4$ and $\phi = 2.1$, at which the

coefficient of determination attains its maximum value ($r_a^2 = 0.70$), i.e., 70 per cent of variation in clay content is accounted for by regression.

Table 1 lists the cluster centres of this fuzzy 4-partition. Data precision of the table does not allow to distinguish between the centres of the clusters A_1 and A_2; at least four decimal places would be required to show the differences between the centres. Similarly, Figure 4.3 shows that the optimum fuzzy 4-partition resembles a situation where the clusters A_1 and A_2 are completely confused. It may therefore be questioned whether the difference between these clusters should be maintained in the soil-landscape model. Note, however, that fuzzy 3-means clustering may not identify identical cluster centres, since the FCM algorithm converges to local minima of J(*P*), which need not be the same (Bezdek, 1981).

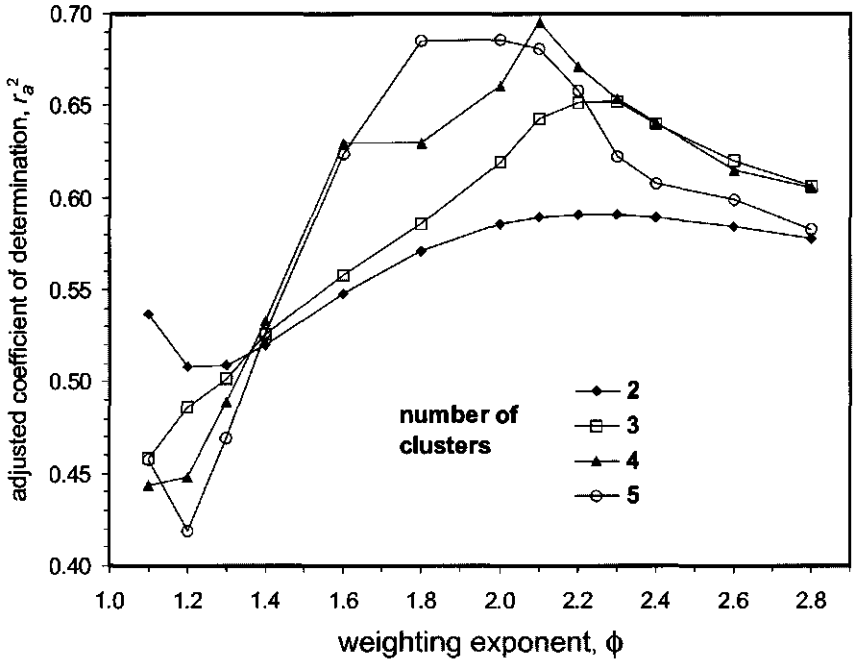


Figure 4.4 Adjusted coefficient of determination of the regression of topsoil clay content as a function of cluster membership grades of fuzzy *c*-partitions corresponding to different values for ϕ and *c*.

Table 4.1 Cluster centres after fuzzy 4-means clustering of the terrain attribute data set with $\phi = 2.1$.

Attribute	General mean	Cluster			
		A_1 [a]	A_2 [a]	A_3	A_4
Elevation (m)	221	231	231	211	219
Slope (deg)	4.65	5.39	5.39	3.67	4.63
Prof. curvature (m^{-1})[b]	$-2.06\ 10^{-5}$	$-2.24\ 10^{-4}$	$-2.24\ 10^{-4}$	$7.86\ 10^{-5}$	$1.48\ 10^{-4}$
Plan curvature (m^{-1})[c]	$-2.06\ 10^{-5}$	$4.59\ 10^{-5}$	$4.59\ 10^{-5}$	$-1.32\ 10^{-4}$	$1.53\ 10^{-4}$
Wetness index (-)	7.19	6.41	6.41	8.42	6.89
Sream power index (-)	106	90.3	90.3	141	72.9

[a] Data precision of the table does not allow to distinguish between the centres of the clusters A_1 and A_2. At least four decimal places are required to indicate differences between the cluster centres.
[b] Positive values indicate convex profile curvatures, implying acceleration of surface flow; negative values indicate concave profile curvatures, implying flow deceleration.
[c] Positive values indicate convex plan curvatures, implying concentration of surface flow; negative values indicate concave plan curvatures, implying flow divergence.

We repeated the regression analysis of Section 4.2.5, substituting the membership grades in the clusters A_1 and A_2 by the membership in the union of these clusters, $(\mu_{A_1} \cup \mu_{A_2})(\mathbf{x}_j)$. To comply with the partition constraint (Eq. 4.4), the union was calculated employing the bounded sum t-conorm (Klir and Yuan, 1995), which owing to Equation 4.4 reduces to:

$$(\mu_{A_1} \cup \mu_{A_2})(\mathbf{x}_j) = \mu_{A_1}(\mathbf{x}_j) + \mu_{A_2}(\mathbf{x}_j) \tag{4.13}$$

Figure 4.5 shows the adjusted coefficient of determination of the regression of topsoil clay content as a function of the membership grades in $A_1 \cup A_2$, A_3 and A_4. Again, the optimum fuzzy c-partition is obtained with $\phi = 2.1$. The regression now accounts for 68 per cent of the variation in topsoil clay content, which is similar to the 70 per cent obtained with the fuzzy 4-partition. From an information-theoretic point of view, preference goes to the partition that reveals most information about the presence of substructures in the terrain attribute data set. This corresponds to the partition with least uncertainty regarding the cluster memberships, such as indicated by entropy-like measures (Bezdek, 1981). Evaluation of the normalised partition entropy (Eq. 4.8) yields $H'' = 0.889$ and $H'' = 0.912$ for the fuzzy 3-partition and the fuzzy 4-partition respectively, therefore the fuzzy 3-partition is preferred.

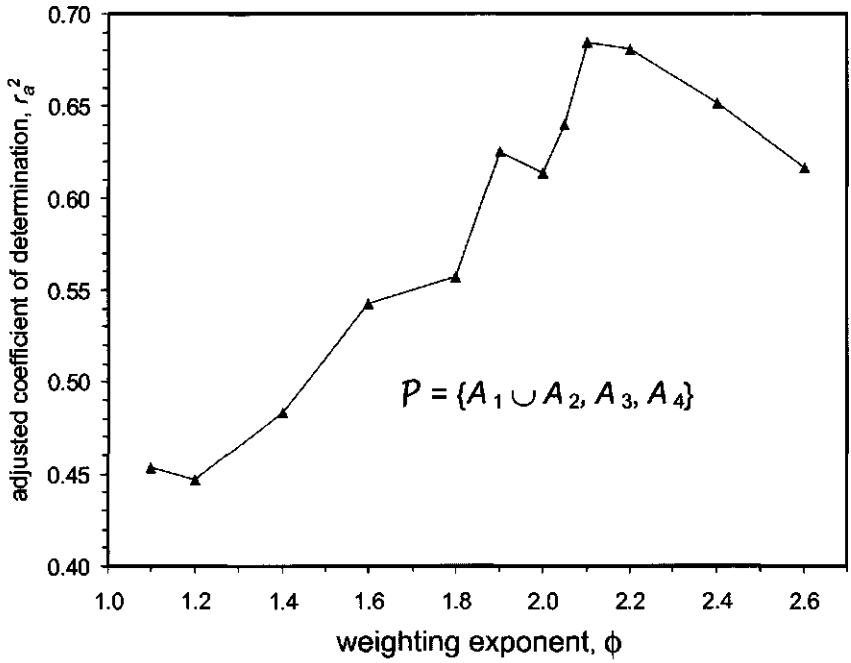


Figure 4.5 Coefficient of determination of the regression of topsoil clay content as a function of $(\mu_{A_1} \cup \mu_{A_2})(\mathbf{x}_j)$, $\mu_{A_3}(\mathbf{x}_j)$ and $\mu_{A_4}(\mathbf{x}_j)$ for different values of ϕ.

Figure 4.6 is a spatial representation of the final fuzzy 3-partition of the terrain attribute data set. For reason of graphical representation, the partition was defuzzified using Equation 4.12. Figure 4.6a shows spatially contiguous areas of elements with relatively high membership grades delimited by zones with large confusion index values. This indicates a strong spatial correlation of the attribute data, being a necessary condition for fuzzy set approaches to work well (Burrough *et al.*, 1997). The combination of Figure 4.6b and Table 1 allows a physical interpretation of the fuzzy clusters. The union of A_1 and A_2 represents the higher part of the study area, corresponding to the terrace remnant. Cluster A_3 corresponds to the lower, relatively wet part of the valley, where surface flow tends to diverge. Cluster A_4 coincides with the higher, relatively dry parts of the valley. If evaluation of Equation 4.12 causes an element to be assigned to a not expected cluster, then this is accompanied by a high level of confusion, as indicated by the confusion index (Figure 4.6a). Figure 4.7 shows the topsoil clay content predicted from regression of the sample data on the membership grades in $A_1 \cup A_2$, A_3 and A_4. The spatial distribution of predicted topsoil clay reflects the landscape pattern as indicated above. The lowest values are on the terrace remnant, intermediate values in the drier parts of the valley and high values in the lower, relatively wet parts of the valley.

Considering the high degree of association with measured topsoil clay data and the coincidence with observable terrain characteristics, the final fuzzy 3-partition provides important information about the soil-landscape. We expect that incorporation of full soil

profile data would further improve the utility of fuzzy *c*-means clustering for soil-landscape modelling.

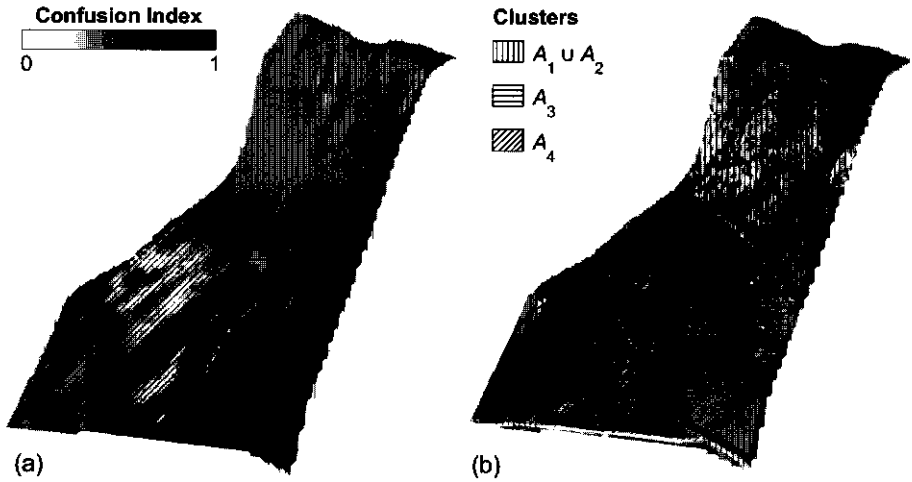


Figure 4.6 Defuzzified optimum fuzzy *c*-partition of the terrain attribute data set over a background of: (a) a map of the confusion index, and (b) a panchromatic photographic image of the study area, in perspective view.

4.4 Conclusions

The results of this study confirm that the fuzzy set approach improves conventional soil-landscape modelling. It allows representation of the fuzziness inherent to soil-landscape units. Fuzzy *c*-means clustering of attribute data derived from a digital elevation model reveals spatial patterns that provide important information about the soil-landscape. The coefficient of determination of regressing soil sample data on membership grades efficiently supports deciding upon the optimum fuzzy *c*-partition. It helps to determine the optimum number of clusters and the best value for the fuzziness coefficient in the context of the actual study. In addition, it provides a measure of the accuracy of the derived soil-landscape model. Future challenges include incorporation of soil profile data other than topsoil data, to further enhance the quality of the soil-landscape model. Given the relatively low price and increasing availability of accurate digital elevation models, the procedure supports generation of reliable soil-landscape models with complete area coverage at low costs. The procedure does not require large sets of soil data since it makes use of the relation between soils and landscape features.

In the case study presented here, the optimum fuzzy *c*-partition showed a high degree of association with measured topsoil clay data, and coincided with observable terrain characteristics. The final pseudopartition comprised three fuzzy units, which were interpreted as:

- The higher part of the study area, which corresponds to an alluvial terrace remnant.

- The lower, relatively wet part of the valley, where surface flow tends to diverge.
- The higher, relatively dry parts of the valley.

Corresponding with the configuration of the physical landscape, the boundaries between these units were represented as transition zones.

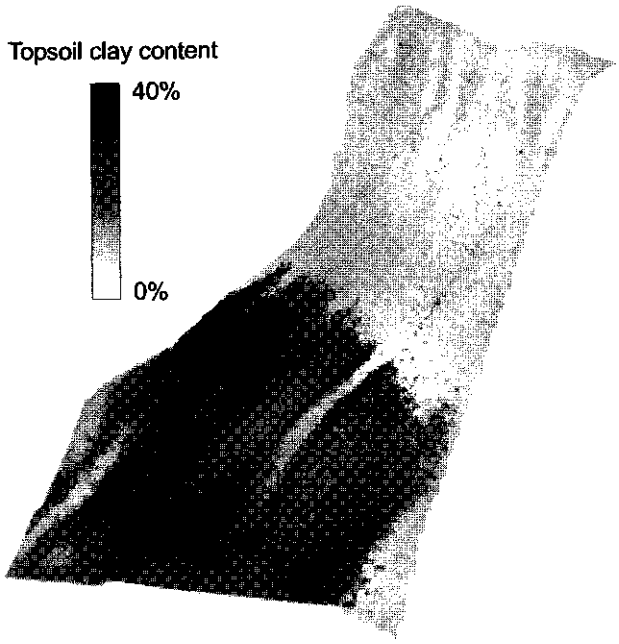
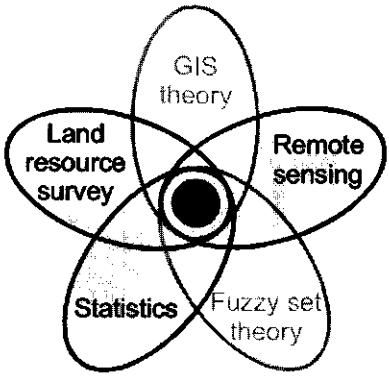


Figure 4.7 Topsoil clay content predicted from regression of the sample data on the membership grades in $A_1 \cup A_2$, A_3 and A_4.


GIS
theory
Land
resource
survey
Remote
sensing
Statistics
Fuzzy set
theory

5 Probabilistic image classification using geological map units applied to land cover change detection[1]

Abstract

This paper describes how probabilistic methods provide a means to integrate analysis of remotely sensed imagery and geo-information processing. In a case study from southern Spain, geological map units were used to improve land cover classification from Landsat TM imagery. Overall classification accuracy improved from 76% to 90% (1984) and from 64% to 69% (1995) when using stratification according to geology combined with iterative estimation of prior probabilities. Differences between the two years were mainly due to extremely dry conditions during the 1995 growing season. Per-pixel probabilities of class successions and entropy values calculated from the classification's posterior probability vectors served to quantify uncertainty in a post-classification comparison. It is concluded that iterative estimation of prior probabilities provides a practical approach to improve classification accuracy. Posterior probabilities of class membership provide useful information about the magnitude and spatial distribution of classification uncertainty.

5.1 Introduction

Land cover classification from multispectral imagery is essentially a matter of deciding about a pixel's cover category on the basis of a limited amount of spectral information. The popular Maximum Likelihood (ML) decision rule requires that evidential support for each class be expressed in probabilistic terms. It assigns a pixel to the class having the largest probability of membership, thus minimising the risk of misclassification (Duda and Hart, 1973).

---

[1] Based on: De Bruin, S., and Gorte, B.G.H. Probabilistic image classification using geological map units applied to land cover change detection. Accepted for publication in the International Journal of Remote Sensing. © 2000 Taylor & Francis Ltd.

Strahler (1980) showed how additional information can be effectively incorporated into ML classification through the use of modified prior probabilities. This is particularly attractive when there is spectral overlap amongst classes, i.e. when classification based on spectral data alone is ambiguous. In an experiment involving classification of natural vegetation, the use of prior probabilities related to elevation and aspect classes resulted in a considerably improved classification accuracy over that obtained with spectral data alone (Strahler, 1980). Janssen and Middelkoop (1992) used a similar approach to incorporate historical land cover data stored in a Geographical Information system (GIS) and knowledge on crop rotation schemes in the classification procedure. They also reported improved classification accuracy when prior probabilities were used. Gorte and Stein (1998) developed an extension to ML classification that uses classification results to iteratively adjust prior probabilities related to spatial strata. Their procedure does not require the priors to be known at the outset, but estimates them from the image to be classified. In this study we used the procedure to incorporate geological map units into land cover classification from Landsat TM imagery.

In spite of its dependence upon incomplete knowledge, standard output of ML classification disregards the uncertainty involved in class assignment, as it usually comprises only the most likely class of membership (Foody *et al.*, 1992). Measures of accuracy of a classified image are conventionally derived from an error matrix that compares classified cover types with sampled ground truth (e.g. Rosenfield and Fitzpatrick-Lins, 1986; Congalton, 1991). These measures give insight into the overall classification accuracy or the performance on a per-class basis but fail to represent spatial distribution of classification uncertainty (Klinkenberg and Joy, 1994). This also applies if the accuracy measures are used to simulate errors for dynamic display in a classified image (see Fisher, 1994b).

Conversely, the posterior probabilities used in ML classification provide information on the uncertainty in assigning individual pixels to a class (Foody *et al.*, 1992, Van der Wel *et al.*, 1998). Storage of these probabilities is particularly valuable if classified images are to be used for further analysis in a GIS. Uncertainties present in the inputs to an analysis will accumulate and affect the end-result (Heuvelink *et al.*, 1989; Heuvelink, 1998a; Burrough and McDonnell, 1998). For example, a common method to detect land cover change is to compare independently classified images by so-called post-classification comparison (Singh, 1989; Jensen *et al.*, 1997; Miller *et al.*, 1998). Obviously, every error in the individual classifications will also be present in the change map. The posterior probability vectors of compared classifications can be used to answer GIS queries about the degree of uncertainty attached to mapped land cover changes, as we will show in this paper.

The objective of this study is to demonstrate:

1. The use of stratification combined with iterative estimation of prior probabilities to improve classification accuracy.
2. The use of posterior probability vectors to represent uncertainty in image classifications and in the results of subsequent analysis.

We present a case study from Alora, southern Spain in which a GIS-stored geological map was used to improve classification of Landsat TM imagery acquired in 1984 and 1995. Posterior probabilities of class membership were fed back to the GIS to visualise uncertainty in a post-classification comparison.

5.2 Methods

5.2.1 Probabilistic image classification using map units

Consider an image classification problem in which each pixel, x, is to be assigned to one of a set of c mutually exclusive and exhaustive categories. The ML decision rule assigns each pixel to the class C_i, $i \in \{1,\ldots, c\}$, having the largest posterior probability of membership given the spectral values contained in its feature vector $\mathbf{x}$. Calculation of the posterior probabilities, $P(x \in C_i|\mathbf{x})$, or more briefly, $P(C_i|\mathbf{x})$, is based on Bayes' Rule:

$$P(C_i|\mathbf{x}) = \frac{P(\mathbf{x}|C_i)P(C_i)}{P(\mathbf{x})} \tag{5.1}$$

where $P(\mathbf{x}|C_i)$ is the conditional probability that $\mathbf{x}$ occurs, given category C_i, $P(C_i)$ is the prior probability of C_i, irrespective of $\mathbf{x}$, and $P(\mathbf{x}) = \sum_{i=1}^{c} P(\mathbf{x}|C_i)P(C_i)$ (Duda and Hart, 1973).

In a supervised classification the conditional probabilities, $P(\mathbf{x}|C_i)$, are estimated from training data for the c classes. Usually sample means and (co-)variances are used as the parameters of normal class probability densities. In this study we used the non-parametric k-nearest neighbour method to derive $P(\mathbf{x}|C_i)$ (Fukunaga and Hummels, 1987; Therrien, 1989; Gorte and Stein, 1998).

In the absence of prior knowledge about the terrain, all classes are initially assumed to be equally probable, i.e. $P(C_i) = 1/c \;\forall\; i \in \{1,\ldots, c\}$. However, classification accuracy is improved if estimates of the area covered by each class C_i are incorporated into the classification process (Strahler, 1980). Such strategy can be combined effectively with stratification of the image on the basis of suitable map units. A prerequisite is that correct class area estimates for each stratum are available. Therefore, we employed a procedure developed by Gorte and Stein (1998) that allows the assessment of relative class areas directly from the image being classified. Starting from equal prior probabilities for all classes, Equation 5.1 is repeatedly evaluated, each time substituting the set of prior probabilities by the normalised sums of posterior probabilities from the previous iteration step. Iteration stops when none of the priors changes more than some pre-established tolerance threshold. Gorte and Stein (1998) proved that the procedure converges to statistically correct area estimates.

5.2.2 Post-classification comparison and uncertainty

Post-classification comparison is the most commonly used method of land cover change detection (Jensen *et al.*, 1997). It deals with independently classified images (Singh, 1989; Bruzzone and Serpico, 1997). In other words, the posterior probability that a pixel x belongs to category C_j at time $t2$ is calculated irrespective of the class and feature vector at the previous time, $t1$:

$$P(C_{j,t2}|C_{i,t1},\mathbf{x}_{t1},\mathbf{x}_{t2}) = P(C_{j,t2}|\mathbf{x}_{t2}) \tag{5.2}$$

Hence, given the feature vectors $\mathbf{x}_{t1}$ and $\mathbf{x}_{t2}$, the probability of a succession of classes at the level of the individual pixel is given by:

$$P(C_{i,t1},C_{j,t2}|\mathbf{x}_{t1},\mathbf{x}_{t2}) = P(C_{i,t1}|\mathbf{x}_{t1})\,P(C_{j,t2}|\mathbf{x}_{t2}) \tag{5.3}$$

Shannon's information-theoretic entropy (Shannon and Weaver, 1949; Applebaum, 1996) provides a suitable measure of the per-pixel uncertainty in a post-classification comparison (cf. Maselli *et al.*, 1994; Foody, 1996; Van der Wel *et al.*, 1998). With c and d representing the number of classes at $t1$ and $t2$, Shannon's entropy (expressed in bits) is calculated as:

$$\begin{aligned} H &= -\sum_{i=1}^{c}\sum_{j=1}^{d} P_{ij}\log_2(P_{ij}) = \\ &-\sum_{i=1}^{c} P(C_{i,t1}|\mathbf{x}_{t1})\log_2(P(C_{i,t1}|\mathbf{x}_{t1})) - \sum_{j=1}^{d} P(C_{j,t2}|\mathbf{x}_{t2})\log_2(P(C_{j,t2}|\mathbf{x}_{t2})) \end{aligned} \tag{5.4}$$

where $P_{ij} = P(C_{i,t1},C_{j,t2}|\mathbf{x}_{t1},\mathbf{x}_{t2})$, and $P_{ij}\log_2(P_{ij}) = 0$ for $P_{ij} = 0$. Equation 5.4 shows that the entropy of a post-classification comparison equals the sum of the entropies of the individual classifications. The value of H ranges from 0, in case of total certainty about a class succession, to $\log_2(cd)$ in case all class successions have equal probability, $1/cd$, i.e. complete uncertainty. It renders the amount of additional information (in bits) that is required to change ambiguity regarding the class succession into definite assignment. Note that entropy expresses uncertainty *according to* the vectors of posterior probabilities. It does not involve uncertainty *concerning* the probabilities; these are assumed to be correct (Gorte, 1998).

Rather than being confronted with the overall classification uncertainty on a pixel basis, users of land cover information may be interested in the per-pixel probability of a particular land cover transition. If such transition concerns a succession of single classes from the classification scheme, its probability is readily calculated using Equation 5.3. If, on the other hand, the transition involves composite classes, the pixel's probabilities of membership to these composites have to be calculated prior to multiplication. For example, a monitoring agency may require per-pixel estimates of the probability of deforestation over the past ten years. Yet, the classification scheme it employs has separate classes for coniferous forest (C_1) and deciduous forest (C_2). A class for both types of forest would correspond to the union of C_1 and C_2, i.e. $C_1 \cup C_2$. If C_1 and C_2 are mutually exclusive classes, the probability of membership to their union is calculated as:

$$P(C_1 \cup C_2|\mathbf{x}) = P(C_1|\mathbf{x}) + P(C_2|\mathbf{x}) \tag{5.5}$$

The probability of absence of a class C_i equals the probability of membership to the complement of that class, $\overline{C_i}$:

$$P(\overline{C_i}|\mathbf{x}) = 1 - P(C_i|\mathbf{x}) \tag{5.6}$$

5.3 Alora case study

5.3.1 Study area

The study area covers approximately 110 km^2 and is centred on the village of Alora in Malaga province, southern Spain (Figure 5.1). Elevation varies between 80 and 735 m above sea level. Agricultural land use in the study area is closely related to geology and topography, as can be observed in Figure 5.1. Below is a general description of this relationship; the colours in parentheses refer to the Landsat TM 4-5-3 colour composite shown in Figure 5.1.

Cultivation of surface-irrigated citrus (red) is concentrated on the level floodplain and terraces of the river Guadalhorce. The increased use of drip irrigation in recent years, however, has promoted citrus cultivation on sloping landforms (Siderius and Elbersen, 1986). Rainfed arable cropping (light shades) is the principal land use on gentle slopes and rolling hills in the flysch deposits of Cretaceous and Tertiary age. High parts and steeper slopes in this landscape are often covered by olive (greyish green). The mountains in Miocene conglomerates and Paleozoic metamorphic rocks are partly used to grow olive, almond and other tree crops (greyish green and brownish purple). Other parts have herbaceous vegetation with shrubs and scattered trees (green), and are used for extensive grazing. On the peridotites and serpentinites a reforestation project is being carried out.

The apparent relation between land cover and geology in the study area suggests that the use of geological map data may improve the accuracy of land cover classification from remotely sensed imagery.

5.3.2 Imagery

We used Landsat TM images acquired on 3 September 1984 and 13 May 1995. The 1984 image was registered to the 1995 image using a first order polynomial transformation with nearest neighbour resampling. The root mean square error (RMSE) of the 10 used control points was below 0.35 pixels in x and y direction. Additionally we digitised 20 ground control points (GCP) to model the relation between 1995 image coordinates and UTM zone 30 coordinates with a first order polynomial. The GCPs had an east-west RMSE of 13 m and a north-south RMSE of 10 m.

Image processing was restricted to TM bands 3, 4 and 5, as the software implementing the k-nearest neighbour algorithm (Gorte, 1998; Gorte and Stein, 1998) handled a maximum of 3 spectral bands. This band subset is widely accepted as standard in vegetation studies (Conese and Maselli, 1993). However, the use of more spectral bands would probably have resulted in better classification accuracies.

5.3.3 Geological map units

We used a generalised geological map derived from the digitised Alora (Instituto Geológico y Minero de España [IGME], 1978) and Ardales (Instituto Tecnológico GeoMinero de España, [ITGE] 1991) map sheets as an additional information source for land cover classification. Generalisation of the original map information served to reduce the number of units, to increase their minimum size, and to harmonise the level of detail of the two adjoining map sheets.

The generalisation process comprised five subsequent steps:

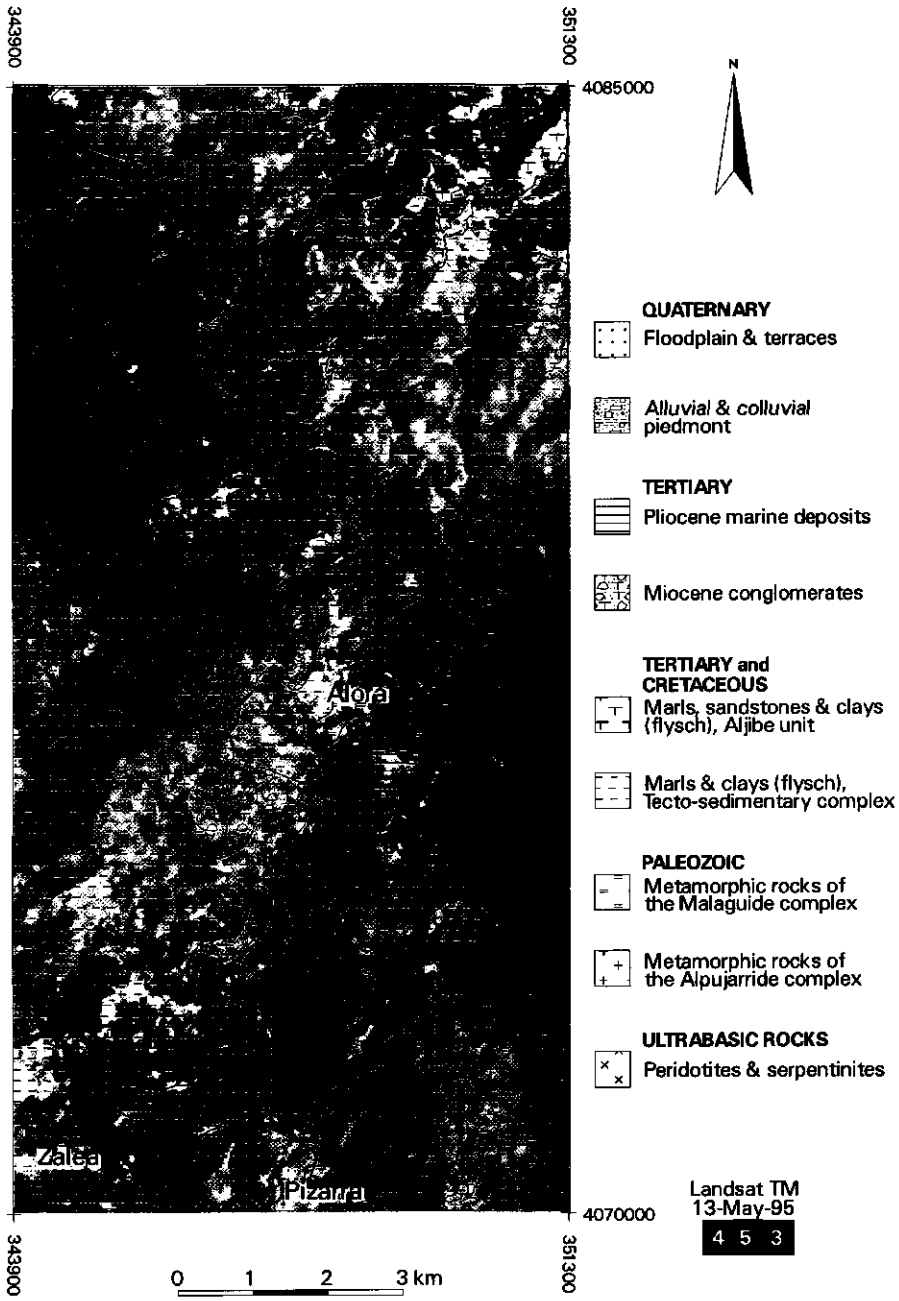
1. Sandstone ridges smaller than 25 ha were merged with adjacent units belonging to the tecto-sedimentary complex.
2. Thematic description was generalised to the nine classes listed in the legend of Figure 5.1; boundaries between similar units were removed.
3. The narrow floodplains of two minor tributaries and the upstream part of the river Guadalhorce were eliminated.
4. Piedmont units smaller than 25 ha were merged with adjacent upslope units.
5. Remaining units smaller than 25 ha were merged with the adjacent unit having the largest shared border.

The resulting coverage, consisting of 31 units (see Figure 5.1), was rasterised and registered to the 1995 Landsat TM image.

5.3.4 Land cover classification and change analysis

The 1984 and 1995 images were classified independently using iteratively estimated prior probabilities per geological unit as described in Section 5.2. We used 1995 digital land cover data provided by the Consejería de Medio Ambiente of the Junta de Andalucía (Sevilla) and fieldwork executed in March 1998 to collect 87 reference sites with known land cover. These constituted the reference data for classification of the 1995 image. Reference data for the 1984 classification were derived from this set by retaining only those sites being consistent with land cover maps of 1977 (Ministerio de Agricultura, 1978) and 1987 (Junta de Andalucía, 1995; supplied in digital format by the Centro Nacional de Información Geográfica, Madrid). Some boundaries were modified to fit field geometry as apparent from the 1984 image. Both sets of reference data were split up into a training set and an evaluation set. Table 1 details the numbers of pixels used for training and evaluation. Note that the 1984 classification scheme did not include *forest clearings and replants*. That is because according to the 1977 and 1987 surveys this class was previously not present within the study area.

Figure 5.1 Geological map (generalised from Instituto Geológico y Minero de España (IGME) (1978) and Instituto Tecnológico GeoMinero de España (ITGE) (1991)) superimposed on a rectified Landsat TM 4-5-3 colour composite (13-May-1995) of the study area. Coordinates (m) correspond to UTM zone 30.


343900
351300
4085000
4070000
N
Alora
Zalea
Pizarra
0 1 2 3 km
QUATERNARY
Floodplain & terraces
Alluvial & colluvial piedmont
TERTIARY
Pliocene marine deposits
Miocene conglomerates
TERTIARY and CRETACEOUS
Marls, sandstones & clays (flysch), Aljibe unit
Marls & clays (flysch), Tecto-sedimentary complex
PALEOZOIC
Metamorphic rocks of the Malaguide complex
Metamorphic rocks of the Alpujarride complex
ULTRABASIC ROCKS
Peridotites & serpentinites
Landsat TM
13-May-95
4 5 3

The classified images were submitted to change analysis by means of post-classification comparison. We used Equations 5.3-5.6 to calculate per-pixel probabilities of class successions and to assess the uncertainty in the post-classification comparison as expressed by Shannon's entropy measure (Shannon and Weaver, 1949; Applebaum, 1996).

Table 5.1 Numbers of pixels used for training and evaluation.

Land cover class	1984		1995	
	Training	Evaluation	Training	Evaluation
Citrus	715	346	709	461
Arable land	1190	476	1190	476
Built-up area	251	70	179	173
Olive	695	400	761	460
Almond mixed with other woody species	621	128	433	320
Herbs with shrubs and scattered trees	865	220	1007	739
Open coniferous woodland	259	119	216	220
Dense woodland	95	-	128	-
Clearings and replants	-	-	174	102
Gravel and sand without vegetation	72	-	86	21

5.4 Results

5.4.1 Land cover classifications

Figure 5.2 shows the land cover classifications from the 1984 and 1995 imagery obtained by the iterative procedure. The error matrices corresponding to these classifications are given in Table 2 and Table 3. Overall accuracy of the 1984 classification improved from 76%, with equal priors, to 90% when using stratification according to generalised geology combined with iterative estimation of prior probabilities (Table 4). Overall accuracy of the 1995 classification improved from 64% to 69% (Table 4).

The rather disappointing results of the latter classification (Table 3, 4) can be largely attributed to extremely dry conditions during the 1995 growing season. Most rainfed arable crops failed early in the season, leaving the soil for the greater part without vegetative cover. Also the sparse undergrowth in olive plantations gave rise to much reflection from bare soil. This resulted in spectral overlap and thus confusion among the classes of arable crops, olive and built-up areas (see Table 3).

Table 5.2 Error matrix, user's accuracy (UA) and producer's accuracy (PA) of the 1984 classification.

Classif. Data	Reference data (evaluation set)									Total	UA (%)
	Cit	Ara	Bld	Oli	Alm	Her	Opn	Den	Gra		
Cit	346									346	100
Ara		412	17	9		2				440	94
Bld		17	37							54	69
Oli		26	4	366		12	1			409	89
Alm			1		128	21	7			157	82
Her		19	2	9		185				215	86
Opn				16			111			127	87
Den										0	
Gra		2	9							11	
Total	346	476	70	400	128	220	119	0	0	1759	
PA (%)	100	87	53	92	100	84	93				

Cit = citrus, Ara = arable crops, Bld = built-up area, Oli = olive, Alm = almond mixed with other woody species, Her = herbs with shrubs and scattered trees, Opn = open coniferous woodland, Den = dense woodland, Gra = gravel and sand without vegetation. Overall accuracy = 90%.

Table 5.3 Error matrix, user's accuracy (UA) and producer's accuracy (PA) of the 1995 classification.

Classif. Data	Reference data (evaluation set)										Total	UA (%)
	Cit	Ara	Bld	Oli	Alm	Her	Opn	Den	Clr	Gra		
Cit	458			5							463	99
Ara		465	56	209		110					840	55
Bld		2	49							2	53	92
Oli		9	59	163		12	6		12	5	266	61
Alm	3		3	16	292	242				4	560	52
Her			5	54	12	348	5		2		426	82
Opn					15	24	183		14		236	78
Den							6				6	
Clr				13	1	3	20		74		111	67
Gra			1							10	11	91
Total	461	476	173	460	320	739	220	0	102	21	2972	
PA (%)	99	98	28	35	91	47	83		73	48		

Clr = clearings and replants; other codes as in Table 5.2. Overall accuracy = 69%.

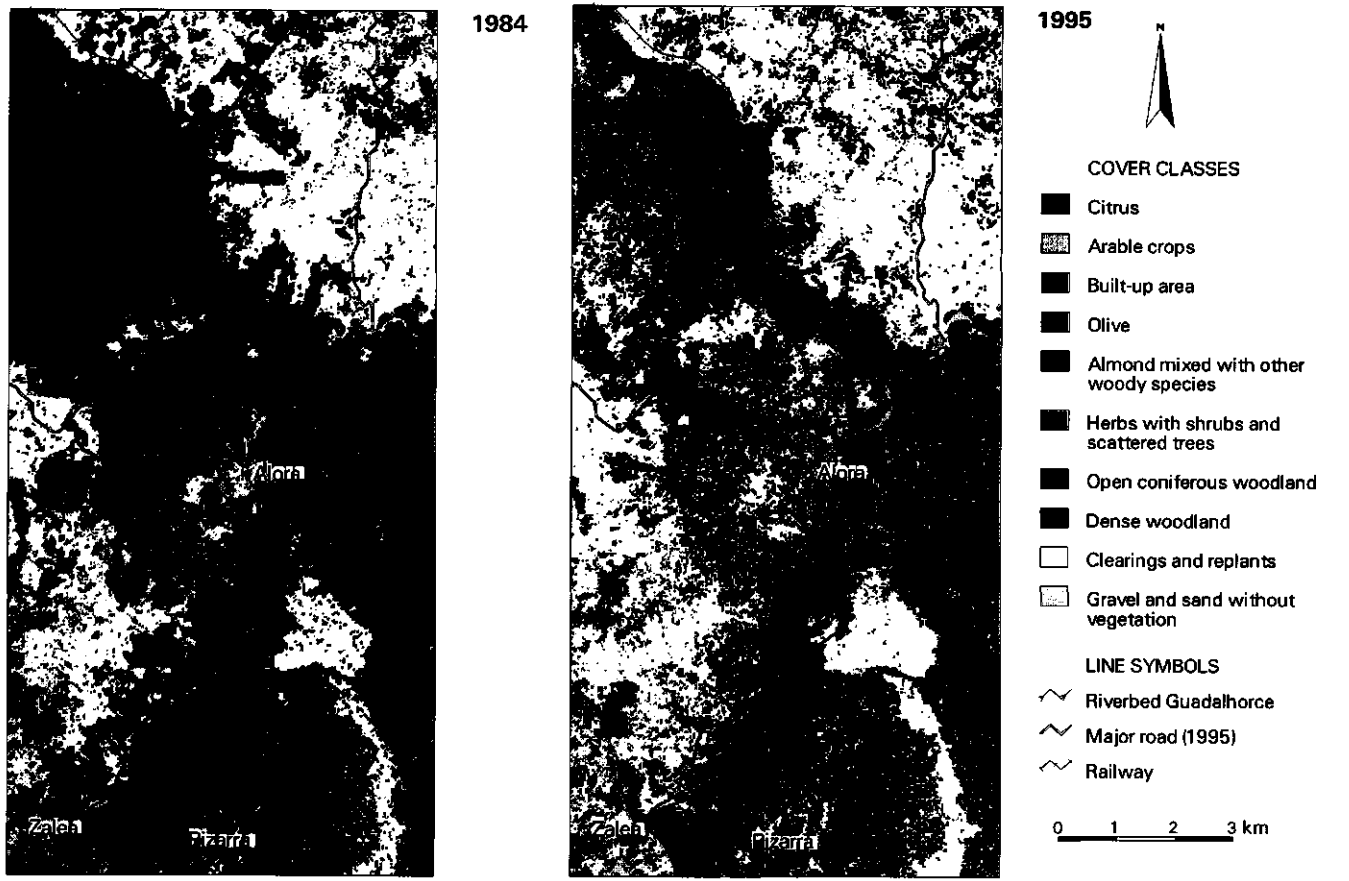

1984
1995
N
COVER CLASSES
Citrus
Arable crops
Built-up area
Olive
Almond mixed with other woody species
Herbs with shrubs and scattered trees
Open coniferous woodland
Dense woodland
Clearings and replants
Gravel and sand without vegetation
LINE SYMBOLS
Riverbed Guadalhorce
Major road (1995)
Railway
0 1 2 3 km
Alora
Zalea
Pizarra

Table 5.4 Overall accuracy of the 1984 and 1995 classifications with and without using iteratively estimated prior probabilities per geological delineation.

Land cover classification	Overall accuracy (%)	
	Equal priors	Iteratively estimated priors
1984	76	90
1995	64	69

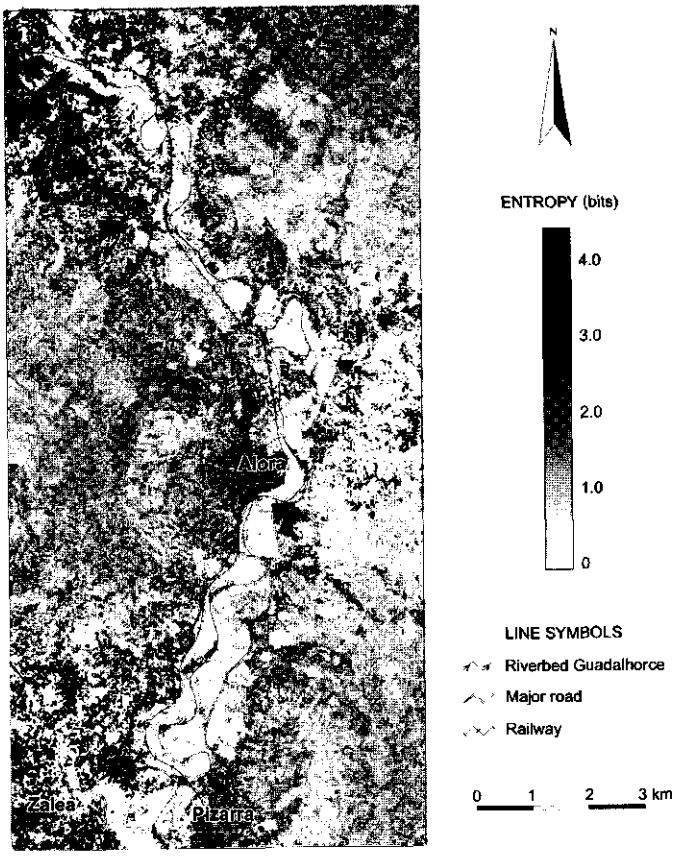


Figure 5.3 Uncertainty in the post-classification comparison as expressed by Shannon's entropy measure.

← Figure 5.2 Land cover classifications from the 1984 and 1995 Landsat TM imagery.

5.4.2 Post-classification comparison

Comparison of the classifications shown in Figure 5.2 suggests that in the period 1984 - 1995 large areas with olive and open coniferous woodland changed into other land cover types. Such inference must however be interpreted with great care, since it is based on uncertain classifications. Figure 5.3 shows the uncertainty in the post-classification comparison as expressed by Shannon's entropy measure (Eq. 5.4). It appears that particularly the areas that seem to have lost olive cover are characterised by high entropy values, i.e. much uncertainty regarding the class succession. On the contrary, sites that according to both classifications were covered by citrus mostly have entropy values close to zero.

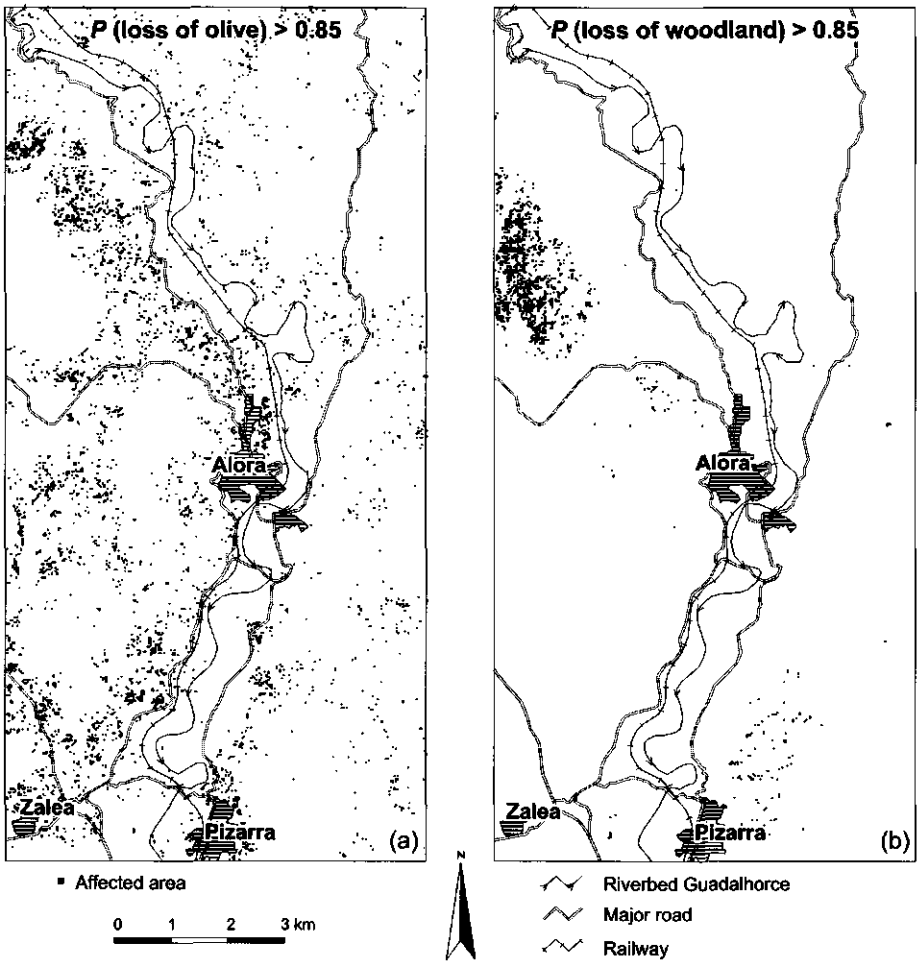


Figure 5.4 Sites having a probability exceeding 0.85 to have lost Olive (a) and Woodland (b) in the period 1984-1995.

Users of land cover information may be interested in the level of confidence they can put on particular mapped land cover changes. As an example, Figure 5.4 shows maps of sites that have lost olive cover (a) and dense woodland or open coniferous woodland (b), with probabilities exceeding 0.85. The per-pixel probability of woodland in the 1984 and 1995 classifications was calculated as $P(\text{Woodland} \mid \mathbf{x}) = P(\text{Dense woodland} \mid \mathbf{x}) + P(\text{Open coniferous woodland} \mid \mathbf{x})$, according to Equation 5.5. Note, that apart from a somewhat larger cluster near the north-west corner of the study area, highly probable loss of olive cover is mapped as a dispersed pattern of small patches and isolated pixels. This challenges the hypothesis that large areas with olive changed into other land cover types in the period 1984 - 1995. Conversely, sites with highly probable loss of woodland are more clustered in an area approximately 4 km north-west of the village of Alora.

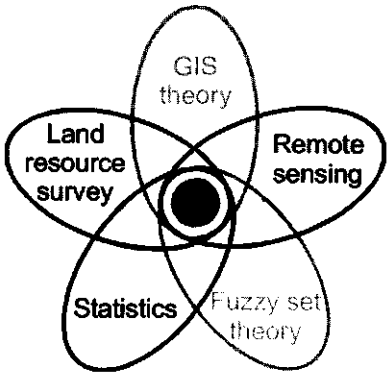
5.5 Concluding remarks

The case study reported here exemplifies how probabilistic methods provide a means to integrate geo-information processing and analysis of remotely sensed imagery. GIS-stored map units were used to improve probabilistic classification of remotely sensed images. Feedback of posterior probabilities to the GIS served to analyse uncertainty in the classification results. The classification accuracies we obtained using iterative estimation of prior probabilities confirm the usefulness of the method as reported in other studies (Gorte and Stein, 1998; Gorte 1998). The latter studies were carried out in the Netherlands and involved stratification on the basis of postcode areas and automatically extracted image segments.

Our use of post-classification comparison was intended solely as a straightforward example of further use of image classification results. It is certainly not promoted as the most appropriate method of change detection (cf. Singh, 1989; Bruzzone and Serpico, 1997; Jensen *et al.*, 1997). Other types of analysis can also benefit from uncertainty information regarding classified remotely sensed images. For example, Monte Carlo simulation (e.g. Fisher, 1991; Heuvelink, 1998a) allows to model propagation of uncertainty through virtually any kind of analysis (also see Chapter 6).

In summary we conclude that:

- Iterative estimation of prior probabilities after stratification according to map units carrying information relevant to the classification theme provides a practical approach to improve classification accuracy.
- Posterior probabilities of class membership provide useful estimates of the magnitude and spatial distribution of local uncertainty in classification results. Such estimates are particularly valuable if classified images are to be used for further analysis in a GIS (cf. Foody *et al.*, 1992).


GIS
theory
Land
resource
survey
Remote
sensing
Statistics
Fuzzy set
theory

6 Predicting the areal extent of land cover types using classified imagery and geostatistics[1]

Abstract

Remote sensing is an efficient means of obtaining large-area land cover data. Yet, remotely sensed data are not error-free. This paper presents a geostatistical method to model spatial uncertainty in estimates of the areal extent of land cover types. The area estimates are based on exhaustive but uncertain (soft) remotely sensed data and a sample of reference (hard) data. The method requires a set of mutually exclusive and exhaustive land cover classes. Land cover regions should be larger than the pixels' ground resolution cells. Using sequential indicator simulation a set of equally probable maps are generated from which uncertainties regarding land cover patterns are inferred. Collocated indicator co-kriging, the geostatistical estimation method employed, explicitly accounts for the spatial cross-correlation between hard and soft data using a simplified model of coregionalisation. The method is illustrated using a case study from southern Spain. Demonstrated uncertainties concern the areal extent of a contiguous olive region and the proportion of olive vegetation within large pixel blocks. As the image-derived olive data were not very informative, conditioning on hard data had a considerable effect on the area estimates and their uncertainties. For example, the expected areal extent of the contiguous olive region increased from 65 ha to 217 ha when conditioning on the reference sample.

6.1 Introduction

Current concerns about environmental changes have lead to an increased demand for land cover data at regional to global scales (e.g. DeFries and Townshend, 1994; Vogelmann *et al.*, 1998). Satellite remote sensing is an efficient means of obtaining these data in a timely and consistent manner. Yet, remotely sensed land cover data are not error-free, as they rely largely on the spectral responses of land cover types that may not

[1] Based on: De Bruin, S. Predicting the areal extent of land cover types using classified imagery and geostatistics. Accepted for publication in Remote Sensing of Environment.
© 2000 Elsevier Science Inc.

all be spectrally distinguishable. Data accuracy may further degrade as a result of errors in the source data and imperfect image processing. If remotely sensed land cover data are used to evaluate environmental changes one should, therefore, account for the uncertainties in these data.

Foody *et al.* (1992), Maselli *et al.* (1994), Van der Wel *et al.* (1998), De Bruin and Gorte (2000; see Chapter 5) and others explored how posterior probability vectors, being a by-product of probabilistic image classification, can be used to represent *local uncertainty* about class labels of individual pixels. This paper goes one step further and presents a geostatistical approach to assess *spatial uncertainty* (Goovaerts, 1997; 1999; Deutsch and Journel, 1998), i.e. the joint uncertainty about land cover at several pixels taken together. This is particularly useful in regional analyses that require spatially aggregated land cover data. Examples of these are assessments of the areal extent of land cover types over spatial units with fixed geometry (e.g. political units or square cells) or the size of contiguous regions having one vegetation cover. Sequential indicator simulation (SIS) enables the generation of multiple maps that honour the available data and allow spatial patterns and uncertainties in the mapped land cover to be inferred. Because in SIS, spatial structures are described in terms of variograms, the approach is notably different from the one proposed by Canters (1997) who used image segmentation to derive spatial structures.

Recently, Kyriakidis (1999) used SIS to map thematic classification accuracy through integration of image-reported (soft) and higher accuracy (hard) class labels. Data integration was accomplished by using simple indicator kriging with varying local means (SKlm) (Goovaerts and Journel, 1995; Goovaerts, 1997) obtained from spatially degraded classified imagery. In this study, the soft indicator data are derived from an image classifier's posterior probability vectors. Data integration is based on a collocated co-kriging approach (Almeida and Journel, 1994; Goovaerts and Journel, 1995) that, unlike SKlm, explicitly accounts for the spatial cross-correlation between hard and soft data. As a consequence, collocated co-kriging estimates are potentially less influenced by sharp local contrasts in the soft data, which are very common in classified imagery (speckling).

This paper explores the use of SIS with collocated indicator co-kriging to evaluate uncertainty in area estimates derived from classified remotely sensed imagery. First, the consequences of spatial uncertainty on area predictions are explained. Next, two sections briefly outline the methods of collocated co-kriging of indicator data and SIS. Finally, the approach is illustrated by predicting the areal extent of a contiguous olive region around a given point, and within pixel blocks covering a study area in southern Spain.

6.2 Area prediction under uncertainty

An obvious way to derive area estimates over a region from remotely sensed imagery is by counting the number of pixels that have been assigned to a given land cover. Bayes' decision rule, which is sometimes referred to as *maximum likelihood rule*, assigns each pixel to the class having the largest conditional probability of membership (Duda and Hart, 1973). It typically leads to an over-representation of the most frequent class and under-representation of less frequent categories (Goovaerts, 1997). Soares

(1992) developed a classification algorithm that does not have this drawback. However, if the only aim is to estimate class areas over regions that contain a large number of pixels there is no need for class allocation altogether, provided that the conditional probability vectors are available.

The regional proportion $q(R;s_k)$ of category s_k over a region R equals the number of pixels where s_k occurs divided by the total number (N) of pixels in R: $q(R;s_k) = \frac{1}{N}\sum_{i=1}^{N} q(\mathbf{u}_i;s_k)$, where $q(\mathbf{u}_i;s_k)$ is defined by:

$$q(\mathbf{u}_i;s_k) = \begin{cases} 1 & \text{if } s(\mathbf{u}_i) = s_k \\ 0 & \text{otherwise} \end{cases} \tag{6.1}$$

with $\mathbf{u}_i$ denoting the ith pixel location, $i = 1,\ldots,N$, and $s(\mathbf{u}_i)$ being the land cover class at $\mathbf{u}_i$. As the true category $s(\mathbf{u}_i)$ is unknown, it is modelled by the random variable (RV) $S(\mathbf{u}_i)$. Consequently, $q(\mathbf{u}_i;s_k)$ is modelled by the RV $Q(\mathbf{u}_i;s_k)$. The (conditional) expectation ($E[.]$) and variance ($Var[.]$) of each $Q(\mathbf{u}_i;s_k)$ are given by[1]:

$$\begin{aligned} E[Q(\mathbf{u}_i;s_k)] &= 1 \cdot P(S(\mathbf{u}_i) = s_k) + 0 \cdot P(S(\mathbf{u}_i) \neq s_k) \\ &= P(\mathbf{u}_i;s_k|\mathbf{x}_i) \text{ and} \end{aligned} \tag{6.2}$$

$$\begin{aligned} Var[Q(\mathbf{u}_i;s_k)] &= P(S(\mathbf{u}_i) = s_k) \cdot P(S(\mathbf{u}_i) \neq s_k) \\ &= P(\mathbf{u}_i;s_k|\mathbf{x}_i) \cdot (1 - P(\mathbf{u}_i;s_k|\mathbf{x}_i)) \end{aligned} \tag{6.3}$$

where $P(\mathbf{u}_i;s_k|\mathbf{x}_i)$ is an estimate of the conditional probability for class s_k to occur at location $\mathbf{u}_i$ given the corresponding spectral feature vector $\mathbf{x}_i$. The expected regional proportion $E[Q(R;s_k)]$ equals the sum of N expectations from Equation 6.2 divided by N:

$$E[Q(R;s_k)] = \frac{1}{N} \cdot \sum_{i=1}^{N} P(\mathbf{u}_i;s_k|\mathbf{x}_i) \tag{6.4}$$

Calculation of the variance of $Q(R;s_k)$ is more involved though, as will be illustrated below.

Figures 6.1a-c represent pixel blocks of 100 pixels each. The pixels are shaded according to their values for $P(\mathbf{u}_i;s_k|\mathbf{x}_i)$. In all three cases the expected regional proportion $E[Q(R;s_k)]$ equals 0.5. If the pixels were to be independent from each other $Var[Q(R;s_k)]$ would equal the sum of the variances (Eq. 6.3) of the 100 individual pixels divided by 100. The results are shown in the first row of Table 6.1. Spatial

[1] $Q(\mathbf{u}_i;s_k)$ has a Bernoulli distribution; see e.g. Snedecor and Cochran (1989, Ch. 7).

independence, however, rarely occurs in image scenes and would seriously restrict the usefulness of remotely sensed imagery in a land resource survey.

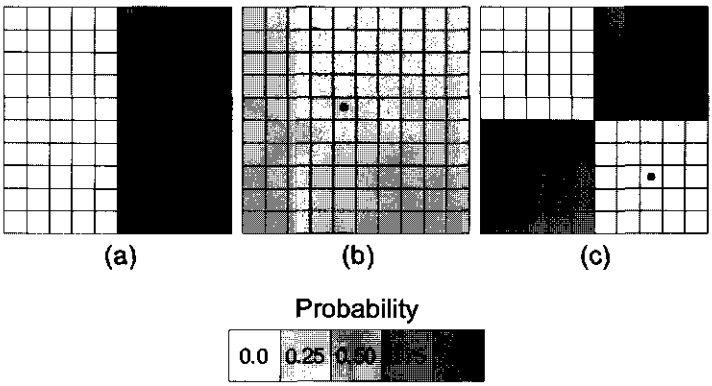


Figure 6.1 Pixel blocks (regions) showing conditional probabilities for a land cover type s_k given the pixels' spectral feature vectors. In all three cases the expected proportion of s_k covered pixels equals 0.5. The black dots in (b) and (c) indicate sample locations.

Assume that each grey shade in Figure 6.1 represents an independent object with a homogeneous land cover (e.g. an agricultural field). Figure 6.1b thus represents one object, corresponding to an extreme case of spatial dependence among pixels. The expectation $E[Q(R;s_k)]$ still equals 0.5, but now $Var[Q(R;s_k)]$ amounts to 0.25, being 100 times larger than for independent pixels (cf. Goodchild *et al.*, 1992; Canters, 1997). This is not surprising as $Q(R; s_k)$ can only take the value zero or one. On the other hand, $Var[Q(R;s_k)]$ would reduce to zero if the true land cover were to be sampled at the pixel locations indicated by black dots in Figures 6.1b-c. The variance reduction in the case of independent pixels would amount to only 1% and 4% respectively (see Table 6.1), as knowledge of the land cover at the sample locations would not affect the uncertainty at other locations.

Table 6.1 Variance of the areal proportion of class s_k over region R, $Var[Q(R; s_k)]$, for different situations indicated in Figure 6.1.

	Situation	Figure 6.1a	Figure 6.1b	Figure 6.1c
(1)	Independent pixels	0	2.5×10^{-3}	9.375×10^{-4}
(2)	Multi-pixel object(s)	0	0.25	2.344×10^{-4}
(3)	As (1) but with sampled ground truth	0	2.475×10^{-3}	9.0×10^{-4}
(4)	As (2) but with sampled ground truth	0	0	0

The geostatistical methods presented hereafter use prior models of spatial correlation to describe spatial continuity of land cover types. They assume the existence of an exhaustive sample of soft data derived from the probability vectors from an image classification and a relatively small sample of hard reference data. Area predictions are conditioned on both data types and on the spatial correlation models that tie the data together. The methods not only deal with area proportions within spatially confined units but also enable uncertainty in the geometry of contiguous regions of a given land cover to be modelled.

6.3 Indicator co-kriging

6.3.1 Indicator approach

The above example illustrates that uncertainty in area estimates from remotely sensed imagery can be considerably reduced if the estimates are conditioned on sampled ground truth (hard data). The example does not show that such conditioning involves updating the image-derived conditional probabilities. Indicator kriging provides a framework to generate posterior conditional probabilities by integrating hard and soft indicator data (Journel, 1986; Zhu and Journel, 1993; Goovaerts, 1997).

Indicator kriging of a categorical variable (e.g. land cover class) requires that all data be coded as local prior probability values. Precise measurements of category s_k at hard data locations $\mathbf{u}_\alpha$ are coded into a set of K binary (hard) indicator data defined as:

$$i(\mathbf{u}_\alpha; s_k) = \begin{cases} 1 & \text{if } s(\mathbf{u}_\alpha) = s_k \\ 0 & \text{otherwise} \end{cases} \qquad k = 1, \ldots, K \tag{6.5}$$

These measurements are often supplemented by a large amount of indirect data such as class probabilities conditioned on remotely sensed spectral responses. These are expressed as soft indicator data with values between 0 and 1, thereby indicating uncertainty about the actual category at the soft data location $\mathbf{u}_i$. For example:

$$y(\mathbf{u}_i; s_k) = P(\mathbf{u}_i; s_k | \mathbf{x}) \tag{6.6}$$

cf. Section 6.2.

Next, local prior probabilities are updated into posterior distributions using nearby hard and soft data. Collocated indicator co-kriging is an updating procedure which incorporates exhaustively sampled soft data by using only the soft indicator datum that is collocated with the location being estimated. It has important advantages over full co-kriging in that it avoids instability problems caused by highly redundant soft information and significantly simplifies modelling of spatial correlation (Almeida and Journel, 1994; Goovaerts and Journel, 1995).

6.3.2 Collocated indicator co-kriging

The ordinary collocated indicator co-kriging (ocICK) estimate of the posterior probability vector of a categorical variable is:

$$[P(\mathbf{u};s_k|(n))]_{ocICK} = \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha^{OCK}(\mathbf{u};s_k)\cdot i(\mathbf{u}_\alpha;s_k) + \lambda_{n(\mathbf{u})+1}^{OCK}(\mathbf{u};s_k)\cdot y(\mathbf{u};s_k) \quad k=1,...,K \qquad (6.7)$$

where (n) denotes the nearby hard and the collocated soft data. Using models of spatial dependence, the weights $\lambda_\alpha^{OCK}(\mathbf{u};s_k)$ and $\lambda_{n(\mathbf{u})+1}^{OCK}(\mathbf{u};s_k)$ are determined by solution of an ordinary co-kriging (OCK) system under the unbiasedness condition:

$$\sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha^{OCK}(\mathbf{u};s_k) + \lambda_{n(\mathbf{u})+1}^{OCK}(\mathbf{u};s_k) = 1 \qquad (6.8)$$

(e.g. Isaaks and Srivastava, 1989, pp.400-416, Goovaerts, 1997, p.313). Any posterior probability outside the interval [0, 1] is reset to the closest bound, zero or one. Subsequently, the estimates $P(\mathbf{u};s_k|(n))$, $k = 1, ..., K$ are standardised by their sum to meet the condition:

$$\sum_{k=1}^{K} P(\mathbf{u};s_k|(n)) = 1 \qquad (6.9)$$

(Goovaerts, 1997; Deutsch and Journel, 1998). Note that condition 6.8 guarantees unbiasedness only if the hard and soft indicator variables have the same mean within each search neighbourhood.

Unlike full co-kriging, solution of the OCK system by ocICK does not require a spatial dependence model for the soft indicator data, but only for the hard indicator data and the cross-correlation between hard and soft data. Spatial dependence modelling is usually done by fitting functions through sample semivariance values. These are half the average squared difference of paired observations in a number of direction and distance classes, the latter a vector $\mathbf{h}$ apart. The linear model of (co)regionalisation is then used to ensure positive definiteness of the covariance matrix in the kriging system (e.g. Isaaks and Srivastava, 1989; Goovaerts, 1997; but Yao and Journel, 1998). If $\gamma_I(\mathbf{h};s_k)$ denotes the variogram of hard indicator data, the value $2\gamma_I(\mathbf{h};s_k)$ indicates how often two locations a vector $\mathbf{h}$ apart belong to different categories $s_{k'} \neq s_k$ (Goovaerts, 1997; 1999).

Since ocICK requires the covariance of the soft indicator data only at $\mathbf{h} = 0$, the only constraint the linear model of coregionalisation must satisfy is that

$$\left| sill[\gamma_{IY}(\mathbf{h};s_k)] \right| \leq \sqrt{sill[\gamma_Y(\mathbf{h};s_k)]\cdot sill[\gamma_I(\mathbf{h};s_k)]} \qquad (6.10)$$

(Goovaerts, pers. comm.), where *sill*[.] denotes the semivariance for distances larger than the range, i.e. the distance where the variogram levels off, I denotes the hard indicator and Y is the soft indicator. Modelling can be further simplified using a Markov-type assumption, which states that dependence of the soft indicator on the hard indicator is limited to the collocated hard indicator datum (Zhu and Journel, 1993; Almeida and Journel, 1994; Goovaerts, 1997). The cross-variogram between hard and soft indicator data, $\gamma_{IY}(\mathbf{h};s_k)$, is then inferred directly from $\gamma_I(\mathbf{h};s_k)$, using a coefficient obtained from calibrating the soft data to the hard data. The validity of this approximation must be checked (see e.g. Goovaerts and Journel, 1995). Note that ocICK with a Markov coregionalisation model is equivalent to ordinary kriging of the residuals when the drift given by the cross-correlation coefficient between hard and soft indicator data has been subtracted (Coléou, 1999).

6.4 Sequential indicator simulation (SIS)

The posterior probabilities $P(\mathbf{u};s_k|(n))$, $k = 1,..., K$, computed by indicator kriging, model the *local* uncertainty about the category that occurs at each interpolated location. As opposed to the kriging variance, which is independent of data values (e.g. Goovaerts, 1997; 1999), measures derived from these distributions reflect the uncertainty that is due to both data geometry and data values. Regional analyses, however, often require spatially aggregated data. This implies that local uncertainties must be combined to reflect joint uncertainty at several locations taken together. Such spatial uncertainty can be modelled by stochastic simulation, i.e. generating multiple equiprobable realisations of the joint distribution of attribute values in space (Zhu and Journel, 1993; Journel, 1996; Goovaerts, 1997; 1999).

Simulation of multiple realisations of a categorical variable can be performed using SIS. Such simulation proceeds as follows (Gómez-Hernández and Srivastava, 1990; Goovaerts, 1997; Deutsch and Journel, 1998; Kyriakidis, 1999):

1. Define a random path through all nodes (pixels) to be simulated, visiting each node only once;
2. At each node $\mathbf{u}$ along this random path:
   (a) Determine the posterior probability $P(\mathbf{u};s_k|(n))$ for each category s_k, $k = 1,..., K$, conditional to the neighbouring hard and soft indicator data, for example using ocICK (Eq. 6.7).
   (b) Generate a value $s^{(l)}(\mathbf{u}) = s_k^{(l)}$ via Monte Carlo sampling of the above distribution. The simulated value is added to the conditioning data set to be used as a hard datum in all subsequent determinations;
3. Move to another node along the random path and repeat step 2.
   The realisation is completed when all nodes have been given a simulated value.

The set of realisations generated by SIS provides an uncertainty model of the spatial distribution of (categorical) attribute values. Spatial features, such as contiguous nodes (pixels) assigned to the same category, are considered certain if seen in all realisations. Conversely, features are deemed uncertain if seen only on a few simulated maps.

Returning to the problem of area prediction referred to in Section 6.2, this model of spatial uncertainty can be used to assess uncertainty in area estimates derived from remotely sensed imagery. This will be demonstrated below.

6.5 Case study

6.5.1 Study area, data and methods

The case study concerns part of the drainage basin of the river Guadalhorce in the province of Malaga, southern Spain. The area is approximately 110 km^2 in extent and centred around the village of Alora. The major part of the study area is covered by digital colour orthophotography derived from aerial photographs taken in 1996. The latter were supplied by the Instituto de Cartografía de Andalucía. De Bruin and Gorte (2000), see Chapter 5, did a land cover classification of the study area using 1995 Landsat TM imagery. The classification scheme distinguished ten mutually exclusive and exhaustive land cover classes. The per-pixel class membership probabilities conditional to the remotely sensed spectral responses were stored to enable further analyses of local classification uncertainty. Here, in the first instance, we will consider the three main crop types in the area: citrus fruits, arable crops and olive. Later on, attention is focused on the olive crop.

Hard land cover indicator data (Eq. 6.5) were collected by visual interpretation of the digital colour orthophotography. First, an equilateral triangular grid with a spacing of 420 m was superimposed over the area having orthophoto coverage. At each grid node the land cover category was determined within a square cell of 900 m^2. The cells precisely matched ground resolution cells of the georeferenced 1995 Landsat TM image. Only cells in which a unique land cover category could be clearly identified were retained (514 cells). Subsequently the grid was densified for improved variogram estimation at short distances. The locations of 200 additional sample points were optimised using spatial simulated annealing (Van Groenigen and Stein, 1998). The objective was to have at least 100 point pairs in distance class 90-180 m and 400 point pairs in distance class 180-270 m, in each of two direction classes ($0 \pm 45°$ and $90 \pm 45°$). Five points were lost because they were positioned within a cell that was also sampled by another point. In another 21 cells the land cover could not be properly determined. The total reference set thus amounted to 688 cells with high accuracy (hard) land cover data (Figure 6.2).

The image-derived land cover class probabilities (De Bruin and Gorte, 2000), see Chapter 5, were calibrated against the hard indicator data by means of logistic regression so as to approximate validity of unbiasedness condition 6.8. The thus transformed class probabilities served as soft indicator data in all subsequent analyses. Indicator variogram modelling for the three main crop types was done using GSTAT 2.0 (Pebesma, 1998; Pebesma and Wesseling, 1998). The Markov coregionalisation model was used to infer the cross-variograms between hard and soft indicator data. The resulting models were visually checked against sample cross-variogram values. If the Markov approximation was inappropriate, constraint 6.10 was used to fit a linear model of coregionalisation.

The error matrix of the 1995 classification as reported by De Bruin and Gorte (2000), see Chapter 5, illustrates the difficulty of correctly classifying olive from remotely sensed imagery. The class had 65% omission errors and included 39% false commissions as a result of spectral confusion with other land cover classes. Therefore, the olive class was selected to demonstrate the effect of using hard data in geostatistical estimation and simulation.

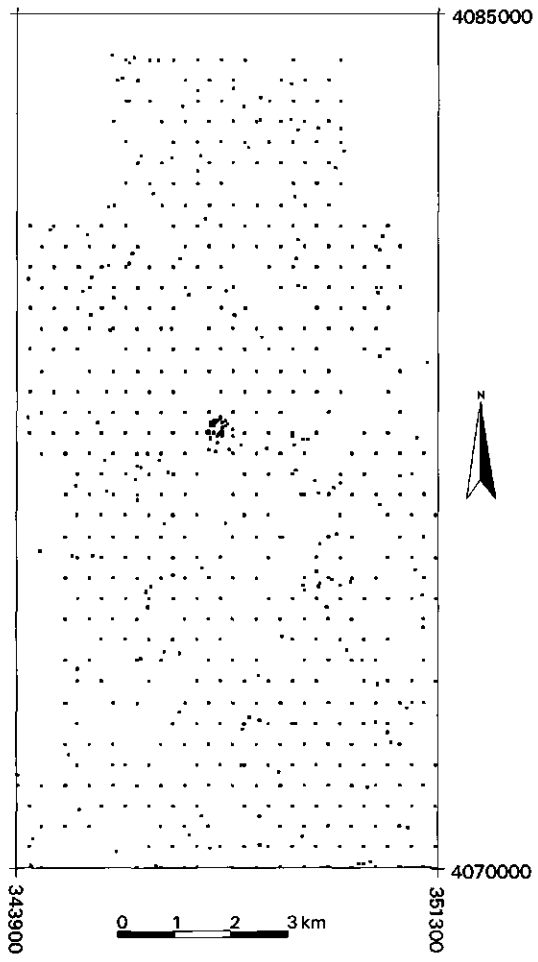


Figure 6.2 Locations of the 688 land cover samples. Co-ordinates (m) correspond to UTM zone 30.

The expanded GSLIB co-kriging program *newcokb3d* (Ma and Journel, 1999) was used to implement ocICK for estimating local probabilities of the occurrence of olive. Sequential indicator simulation with ocICK was performed using the GSLIB program *sisim*. The latter program was modified to enable the use of fitted linear models of coregionalisation. Estimates of the spatial uncertainty about the presence or absence of

olive vegetation were obtained from 500 SIS realisations both with and without conditioning on the hard indicator data.

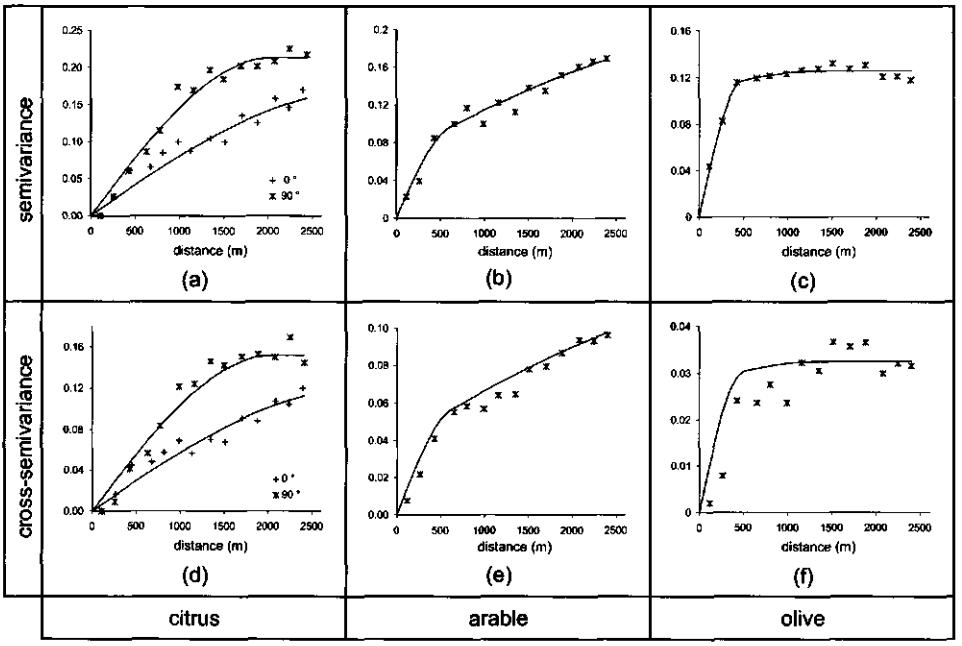


Figure 6.3 Experimental indicator (cross-)variograms (symbols), fitted variogram models (a-c) and Markov models of the indicator cross-variograms (d-f) for the three main crop types in the study area.

6.5.2 Results

Figure 6.3 shows the indicator (cross-)variograms for the three main crop types in the study area. The continuous curves in the upper three plots (Figures 6.3a-c) were obtained by fitting positive linear combinations of spherical functions through the sample semivariances. The indicator variogram for citrus (Figure 6.3a) is anisotropic, i.e. the pattern of spatial connectivity changes with direction; the axis of greatest spatial continuity being in a north-south direction. The solid lines in Figures 6.3d-f are Markov models of the indicator cross-variograms $\gamma_{IY}(\mathbf{h};s_k)$, s_k = *citrus*, *arable*, *olive*. The models show good correspondence with the experimental data for citrus and arable crops, but the approximation does not fit the olive data. The cross-variogram $\gamma_{IY}(\mathbf{h};olive)$ cannot be considered as being proportional to $\gamma_I(\mathbf{h};olive)$. A better fit, obtained with a linear model of coregionalisation, is shown in Figure 6.4. This model puts most weight (73%) on the relatively unimportant long range component (1350 m) of $\gamma_I(\mathbf{h};olive)$. Besides having low overall predictive ability (relatively low cross-variogram values), the image-derived soft indicator data particularly fail to detect short range variations in the presence or absence of olive vegetation.

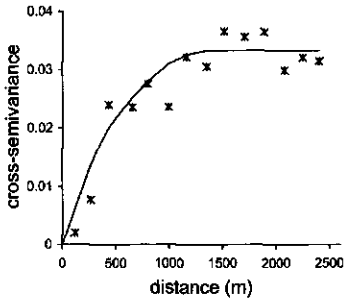


Figure 6.4 Experimental indicator cross-variogram for olive (symbols) and linear model of coregionalisation fitted under constraint (6.10).

As can be observed in Figure 6.5, conditioning on hard olive indicator data has a considerable effect on the estimated local probabilities of occurrence. In the neighbourhood of hard indicator data, the short range variability of the kriging estimates (Figure 6.5b) is lower than that of the image-derived probabilities (Figure 6.5a). At the same time the local uncertainty is lower. Beyond the range of influence of the hard indicator data (see Figure 6.2) Figures 6.5a and 6.5b are identical.

The effect of the hard indicator data on estimating spatial uncertainty is even more pronounced. Figure 6.6 illustrates some results of 500 maximally conditioned SIS computations, i.e. simulations honouring both the hard and the soft indicator data. The attribute of interest concerns the area of a contiguous olive covered region around one of the sample locations (point #213). This location was known to be covered by olive vegetation. The area estimate is subject to spatial uncertainty because it depends on the land cover at multiple locations taken together. Therefore, it cannot be directly calculated from a probability field (e.g. Figure 6.5b), but an approximate answer can be obtained from the statistics of a set of equiprobable realisations. The results of the multiple SIS computations are summarised in a histogram (Figure 6.6a) and a cumulative distribution graph (Figure 6.6b) of the simulated area. Olive labelled pixels were considered connected if they were within the immediate 8-pixel neighbourhood (eight nearest neighbours) of each other. The mean area amounted to 217 ha and the (sample) variance was 7638 ha^2. However, the latter figure is of little practical value as the area distribution exhibits bimodality with distinct peaks around 150 ha and 330 ha. This bimodality is caused by two regions being connected or not in the individual simulations.

The SIS computations were repeated without conditioning on the 688 hard indicator data. The results are summarised in Figures 6.7a-b. The mean area and variance now amounted to 65 ha and 3513 ha^2 respectively. The area distribution has a high peak at 0 ha and a second, lower peak around 70 ha. The first peak is due to uncertainty about the land cover at location #213 itself. In 31% of the simulations it was classified as not having olive vegetation. The difference with the distribution of Figure 6.6 is a consequence of the absence of hard indicator data that via the model of coregionalisation relate the uncertain image-derived data to locations having known land cover.

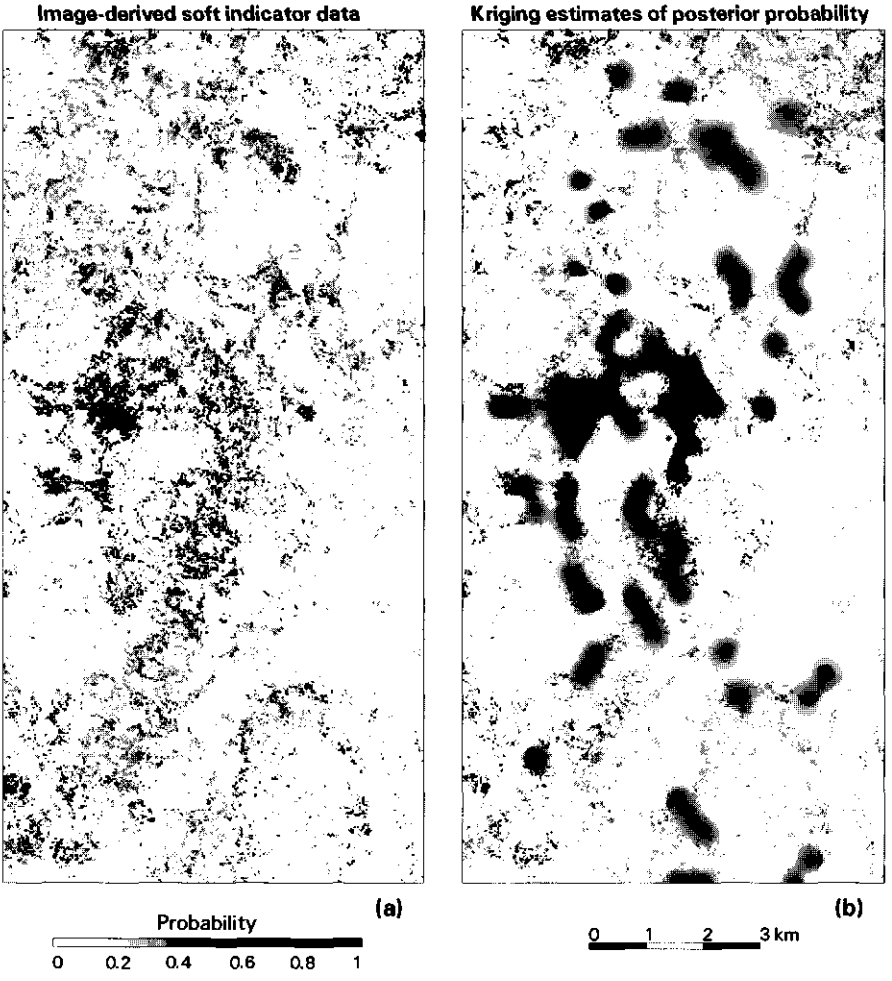


Figure 6.5 Image-derived soft indicator data for olive (a) and ocICK estimates of the local probabilities of occurrence conditional to the nearby hard and the collocated soft indicator data (b).

As indicated earlier, area predictions over spatial units with fixed geometry also involve spatial uncertainty. Figure 6.8 shows estimates of the proportions of olive vegetation and their variances in square pixel blocks of 100 pixels (9 ha) each. The former were calculated as block averages of the ocICK posterior probability estimates $P(\mathbf{u};olive|(n))$ shown in Figure 6.5b. Alternatively, they could have been obtained from some form of block kriging (Isaaks and Srivastava, 1989; Goovaerts, 1997; Deutsch and Journel, 1998). The variances were calculated from the 500 maximally conditioned SIS realisations. Note that, unlike block kriging variances, these conditional variances reflect the uncertainty that is due to both data geometry and data values.

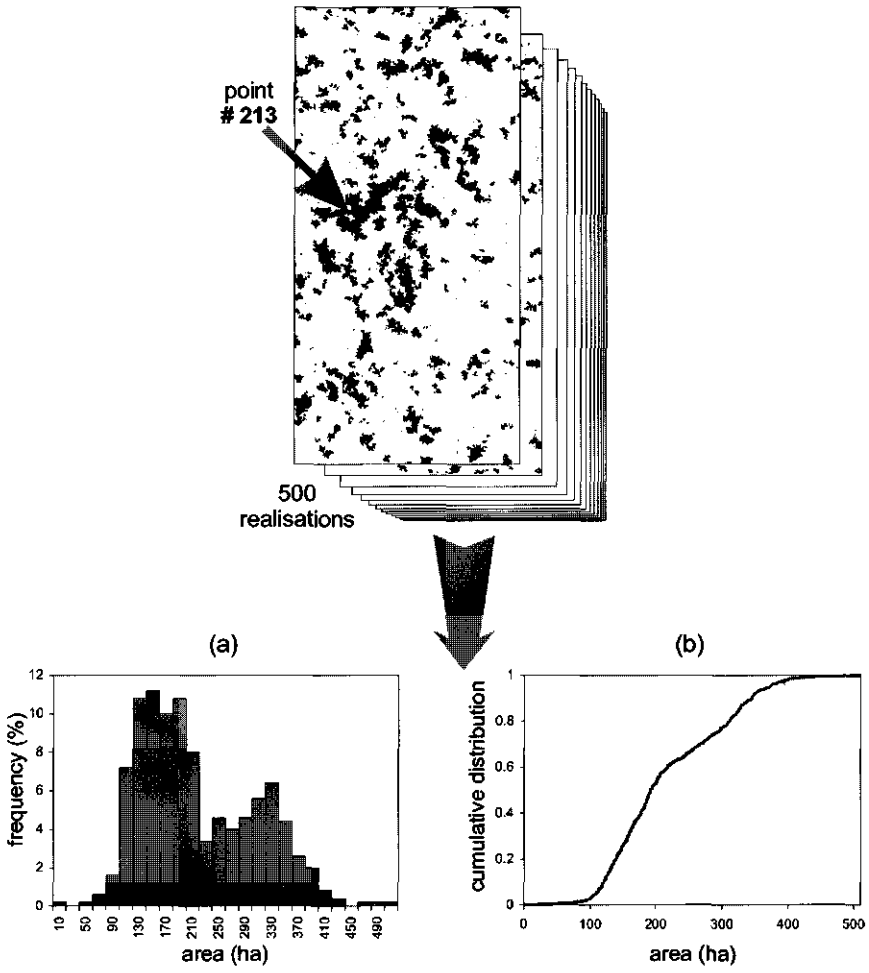


Figure 6.6 Histogram (a) and cumulative distribution (b) of the area of a contiguous region with olive vegetation (around sample #213). The distribution was calculated from 500 SIS realisations; all conditioned on nearby hard and collocated soft indicator data.

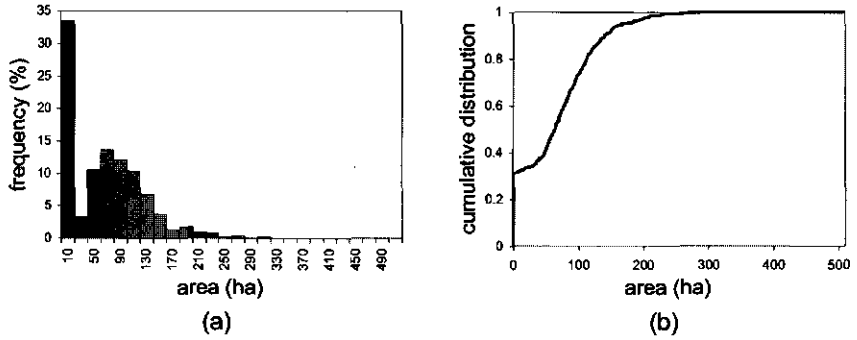


Figure 6.7 Histogram (a) and cumulative distribution (b) of the area of a contiguous region with olive vegetation (around sample #213). The distribution was calculated from 500 semi-conditional SIS realisations, i.e. without considering the hard indicator data.

6.6 Concluding remarks

This paper presents a geostatistical method to model uncertainties in image-derived estimates of the areal extent of land cover types. These uncertainties have a spatial character and may concern, for example, the size of land cover regions (e.g. habitats) or the proportion of land cover types within spatial units (e.g. pixel blocks). The area estimates are based on exhaustive but uncertain (soft) remotely sensed data and a sample of exact (hard) data. The latter data are particularly important if the image-derived data are not very informative. Collocated indicator co-kriging allows the updating of soft probabilistic data using a simplified model of coregionalisation between hard and soft data. A Markov-type assumption may further alleviate the modelling efforts. The case study, however, demonstrated that the Markov approximation does not always fit the experimental cross-variogram. Sequential indicator simulation enables the generation of a set of alternative equiprobable maps, from which uncertainties regarding land cover patterns can be inferred. The method can be implemented using public domain software (Deutsch and Journel, 1998; Pebesma, 1998; Ma and Journel, 1999).

Assessment of uncertainty about land cover is rarely a goal in itself. More often, the variable of interest is an ecological response variable that may ultimately be used in developing land use policies. Estimates of the uncertainty in such a variable can be obtained by using multiple SIS-generated land cover maps as input to ecological response models. The uncertainty estimates thus obtained can then be used in risk-based policy (Goovaerts, 1999; Kyriakidis, 1999).

The indicator approach presented in this paper requires the land cover regions to be considerably larger than the pixels' ground resolution cells (*H*-resolution, Strahler *et al.*, 1986). In the opposite case, it may be relevant to model vegetation quantities as continuous variables so that an approach similar to that proposed by Dungan (1998) could be adopted. Alternatively, a mixed pixel view may be more appropriate, in which case geostatistical estimation could be performed using some form of compositional kriging (De Gruijter *et al.*, 1997).

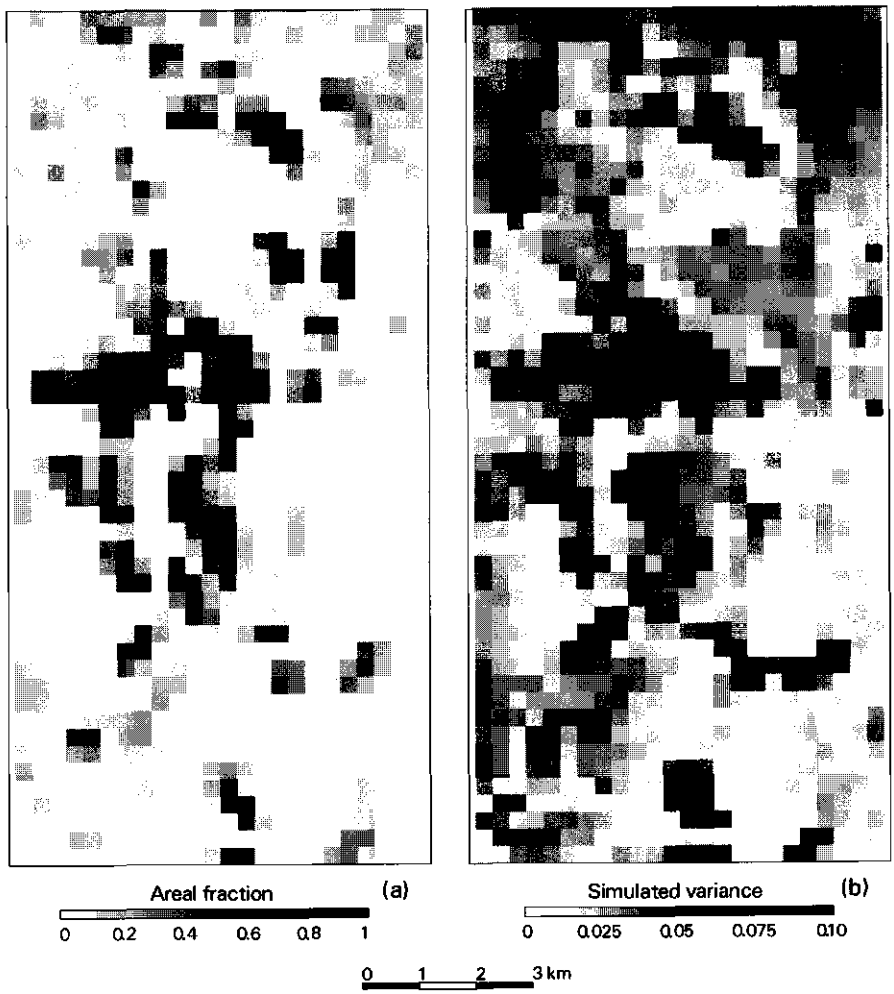
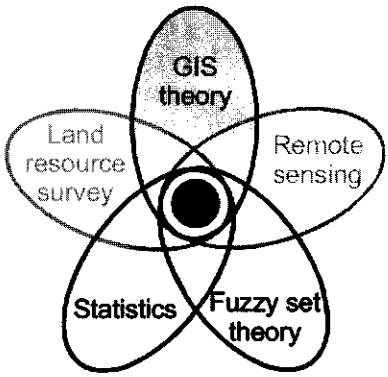


Figure 6.8 Estimates of the proportions of olive vegetation (a) and their variances (b) in pixel blocks of 10×10 pixels (9 ha).

The method requires an exhaustive set of mutually exclusive land cover classes (e.g. olive vs. non-olive). Uncertainty is due to incomplete data about the true land cover type but does not concern the class definitions; these should be clear-cut. If the latter is not the case, the concept of expected membership in a fuzzy set can be used to combine uncertainty about values of random variables with uncertainty about the class intentions. Examples of such a combination of fuzziness and probabilistic uncertainty in areas other than land cover mapping have been reported by Lark and Bolam (1997), see also Chapter 7.

Finally, uncertainty about a spatial phenomenon always depends on the decisions made to model that phenomenon. Important decisions in the present study are the

stationarity decision, which is very common in geostatistics, and the models of spatial dependence. Model decisions may be inappropriate and sometimes difficult to validate. Although cross-validation or validation against a separate evaluation set may be helpful, they also suffer from severe restrictions. Hence the importance of clearly documenting all aspects of the model. (Goovaerts, 1997, pp. 105-106, 442). The frequent assumption of independent pixels obviously impedes proper assessment of spatial uncertainty in image-derived area estimates. Using well-documented geostatistical methods, the modelling alternative presented in this paper exploits spatial dependence rather than ignoring it.


GIS
theory
Land
resource
survey
Remote
sensing
Statistics
Fuzzy set
theory

7 Querying probabilistic land cover data using fuzzy set theory[1]

Abstract

Queries expressed in verbal language often involve a mixture of uncertainties in the outcomes of events that are governed by chance and in the meaning of linguistic terms. This study exemplifies how probability and fuzzy sets can work together to deal with such queries in the spatial domain. It involves site selection on the basis of accessibility (travel time) estimates and per-pixel probabilities of land cover change derived from remotely sensed imagery. Relationships between probabilities and fuzzy sets were established on the basis of a linguistic probability qualifier (high probability) and the expectation of a membership function defined on stochastic travel time. Fuzzy query processing was compared with crisp processing to emphasise the difference between grade and probability of membership. Fuzzy set theory allowed to deal with the vague meanings of linguistic terms. The fuzzy query response contained more information than the crisp response, namely the degree to which individual locations matched the selection criteria. This illustrates the gain in expressive power provided by combining probability and fuzzy sets.

7.1 Introduction

Both probabilistic methods and methods based on fuzzy set theory are currently being used to deal with uncertainty in the classification of remotely sensed imagery (e.g. Canters, 1997; Eastman, 1997; Foody, 1997; Gorte and Stein, 1998). Yet, the distinction between probability of class membership and degree of membership in a fuzzy class appears to be an object of confusion in the geo-information and remote sensing community. For example, Foody and Trodd (1993, p. 343; emphasis added) stated that probabilities derived from maximum likelihood classification and fuzzy membership functions from fuzzy *c*-means classification are '... measures of the strength of

[1] Based on: De Bruin, S. Querying probabilistic land cover data using fuzzy set theory. Accepted for publication in the International Journal of Geographical Information Science.
© 2000 Taylor & Francis Ltd.

membership to "*discrete classes*" ...'. They continued by saying that '... both approaches imply *partial class membership* ... '. These statements are not only erroneous, but they are also contradictory. Fuzzy *c*-means classification allows partial class membership, but discrete classes do not have partial members. Similar examples of confusion can be found in works on the unmixing of coarse pixel signatures (e.g. Schowengerdt, 1996; Bastin, 1997).

In general, confusion arises because there is a tendency to interpret posterior probabilities of a classification as analogous to class assignment function parameters (Manton *et al.*, 1994). They are not analogous though. The probabilistic classifiers implemented in most commercially available image processing systems assume that an element is a member of only one crisp class, i.e. a class that sharply distinguishes between members and non-members. Classification uncertainty lies in the inability to identify the class to which the element belongs. The posterior probabilities of a classification are estimates of the likelihood of full membership in each class and *not* the grade of membership in these classes (Manton *et al.*, 1994). Usually, membership is assigned to the class with the highest likelihood. Thus, the class assignment function takes the value one for this class and zero for all other classes. Fuzzy classification, on the other hand, is based on the concept of fuzzy sets (Zadeh, 1965). In the fuzzy set model, the class assignment function attributes to each element a grade of membership in the real interval [0, 1] for every defined set. This grade of membership corresponds to the degree to which the element is similar to the concept or prototype represented by that set. Accordingly, fuzzy sets enable representation of imprecisely defined classes such as vague concepts expressed in verbal language.

The relationship between the theories of probability and fuzzy sets is a matter of much controversy among scientists (e.g. Laviolette *et al.*, 1995; Nguyen, 1997). In clarifying the difference between probability and degree of class membership to geographers, Fisher (1994a, 1996) referred to two forms of viewshed regions that can be derived from a digital elevation model. The probable viewshed represented uncertainty about the existence of a direct line of sight between observer and location; the fuzzy viewshed indicated the degree to which any potentially visible target would be discernible. Although the example elucidated differences in concepts, the work did not demonstrate the combined use of both approaches. It has been claimed that such a combination may significantly improve the modelling of human knowledge (e.g. Zadeh, 1995; Nguyen, 1997).

The objective of this paper is to exemplify how probability and fuzzy sets can work together to handle imprecisely formulated queries on uncertain spatial data. The example involves site selection on the basis of accessibility estimates and per-pixel probabilities of land cover change. The latter were derived from classified remotely sensed imagery. Relationships between probabilities and fuzzy sets are established using the concepts of fuzzy probability qualifiers (Wallsten *et al.*, 1986) and expectation of membership functions defined on stochastic variables (Zadeh, 1968; Kandel, 1986).

7.2 The example query

Geographical information systems (GIS) are frequently used for site selection purposes. For example, Cadwell *et al.* (1995) used a GIS to select optimal sites for

planting seedlings of big sagebrush in order to restore shrub cover in burned areas. Other applications are described in Hendrix and Buckley (1992), Fleischer *et al.* (1998) and Wright *et al.* (1998). Although in the latter two articles reference is made to data currency and accuracy as required for successful data integration, none of the listed papers deals explicitly with the issue of uncertainty in site selection.

This study is concerned with such uncertainty, both in spatial input data and in selection criteria. A query was submitted to a spatial database of a study area, approximately 110 km^2 in extent, centred on the village of Alora in Malaga province, southern Spain (see Figure 7.1). I considered the following query, the output of which could be used to support site selection for an experiment requiring a surface area of less than 900 m^2.

Query:

Show locations that with high probability have lost forest cover over the period 1984 to 1995 and that are easily accessible from a major road.

This query is illustrative of problems that involve a mixture of uncertainties in the outcomes of events that are governed by chance and in the meaning of subjective concepts. These problems frequently occur when dealing with concepts expressed in verbal language.

De Bruin and Gorte (2000), see Chapter 5, analysed Landsat Thematic Mapper imagery acquired in 1984 and 1995 to identify land cover changes within the study area by post-classification comparison. Post-classification comparison means that the images are classified separately and then overlaid to determine changes at the pixel level (Singh, 1989). For the purpose of the present study the original 10 land cover classes were rearranged into forest cover (comprising open coniferous forest and thickly wooded land) and non-forest cover (the other eight classes). Figure 7.1 shows loss of forest cover according to a comparison of the 1984 and 1995 classifications in which each pixel was assigned to the class having maximum posterior probability. Obviously, such a comparison disregards any uncertainty in the individual classifications. Therefore, De Bruin and Gorte (2000), see Chapter 5, proposed to make use of the posterior probabilities of class membership in a post-classification comparison. Under the assumption of independence, a pixel's conditional probability of forest loss, given its spectral vectors $\mathbf{x}_{t1}$ (1984) and $\mathbf{x}_{t2}$ (1995), can be calculated as:

$$P(\text{loss of forest cover} \mid \mathbf{x}_{t1}, \mathbf{x}_{t2}) = P(\text{forest} \mid \mathbf{x}_{t1})\, P(\text{non-forest} \mid \mathbf{x}_{t2})$$

where $P(a|b)$ denotes the probability of a conditional on b. A grid in which these probabilities are represented at 30 m spatial resolution served as the main constituent of the queried database. Each cell of this grid meets the surface area requirement of an experimental site. It is implicitly assumed that forest loss either occurred or did not occur at the cell level, i.e. partial loss of forest cover is not considered.

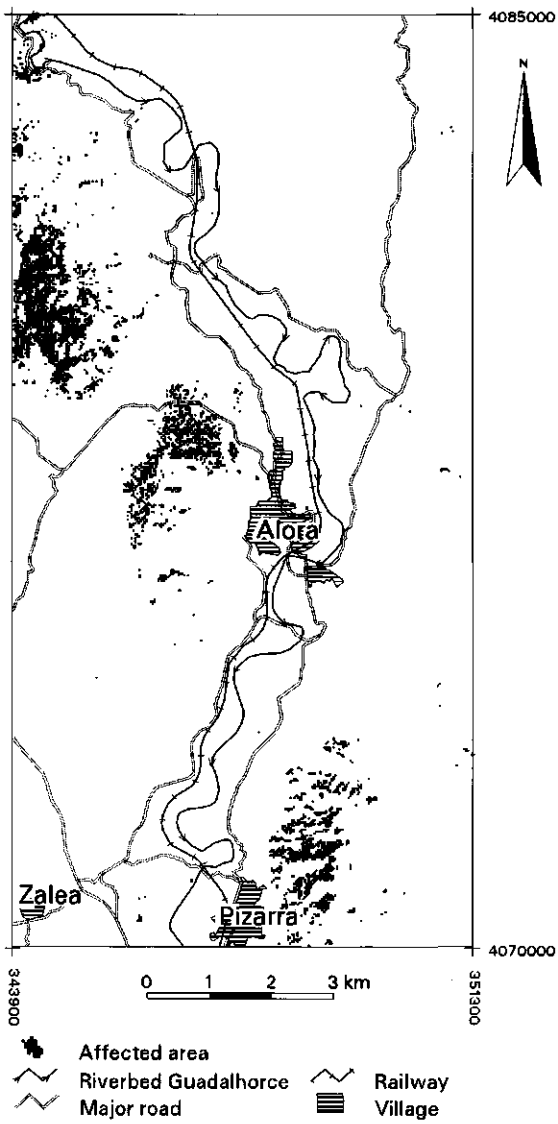


Figure 7.1 Loss of forest cover in the study area according to a post classification comparison following maximum posterior probability class assignment. Coordinates (m) correspond to UTM zone 30.

Apart from being concerned with stochastic uncertainty about the occurrence of forest loss, the above query contains two imprecise selection criteria. These are *high probability* and *easily accessible*. In order to handle these terms in a structured way the query was divided in two sub-queries that are being addressed in Sections 7.3.1 and

7.3.2. Section 7.3.3 explains how the results of the sub-queries were combined to produce the complete response.

7.3 Query processing

7.3.1 Fuzzily qualified probability of forest loss

The phrase 'locations that with high probability have lost forest cover' contains a linguistic expression. What should be understood by the term *high probability*? In an effort to introduce numerical conversions of probability terms, Mosteller and Youtz (1990) provided measures of central tendency for 52 probability terms (among which *high probability*). Using a direct translation, as suggested in that work, one could set a threshold, for example $P = 0.85$, which if exceeded implies membership in the set of desired locations x. The corresponding membership function, $\mu_A(x)$, would then be defined as:

$$\mu_A(x) = \begin{cases} 1 & \text{if } P(\text{loss of forest cover} \mid \mathbf{x}_{t1}, \mathbf{x}_{t2}) > 0.85 \\ 0 & \text{otherwise} \end{cases} \tag{7.1}$$

(see Figure 7.2b), where A denotes the set of grid cells x having high probability of forest loss.

Cliff (1990) and Wallsten and Budescu (1990) questioned the desirability of such a quantification of probabilistic expressions because, among other things, it disregards the inherent fuzziness of these concepts. Wallsten *et al.* (1986) showed that the meanings of probability terms can be represented by means of fuzzy membership functions over the [0, 1] probability interval. These functions take the value zero for probabilities not at all in the vague concept represented by the term, one for probabilities that are perfect exemplars of the concept, and intermediate values otherwise. Membership functions for the same term may differ substantially over people (Wallsten *et al.*, 1986) and depend upon context and communication direction (Wallsten and Budescu, 1995). In this study, a function similar to the one used by Klir and Yuan (1995, p. 223) for the term *very likely* was adopted:

$$\mu_A(x) = \begin{cases} e^{-(P-1)^2/0.035} & \text{if } P(\text{loss of forest cover} \mid \mathbf{x}_{t1}, \mathbf{x}_{t2}) > 0.5 \\ 0 & \text{otherwise} \end{cases} \tag{7.2}$$

(see Figure 7.2a).

For each grid cell the degree of membership in the set of desired locations was determined by applying Function (7.2) to the conditional probabilities of forest loss. To compare this procedure with a crisp set approach, the calculations were also performed using Function (7.1). The probability threshold for crisp membership ($P = 0.85$) corresponds approximately to the crossover point of the fuzzy membership function. The

crossover point is the point at which the membership grade in a fuzzy set is 0.5 (Kandel, 1986).

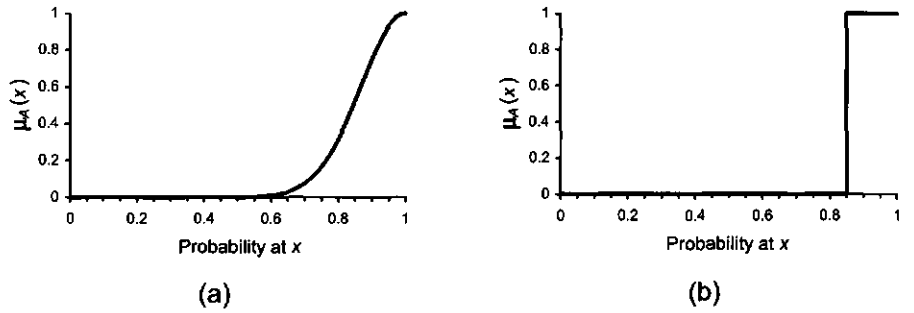


Figure 7.2 Fuzzy (a) and crisp (b) representation of the term high probability.

7.3.2 Easy accessibility

The accessibility of a location can be expressed as the time cost of reaching that location from a major road. To calculate this cost the COSTDISTANCE function of ARC/INFO (Environmental Systems Research Institute [ESRI], 1994b) was used. It assigns to each cell in a grid the cost to reach a source cell via the least-accumulative-cost path across an impedance surface. The latter defines the impedance to move planimetrically through each grid cell. As will be explained below, the calculation required multiple impedance grids that were derived from a digitised topographic map (Servicio Geográfico del Ejército, 1995) and slope data. A set of grid cells covered by major roads constituted the source cells.

The time required to reach a location may fluctuate stochastically as a result of weather-induced road conditions. On the basis of some experience in the area, it was assumed that during the experiment the average travel speed on minor roads is normally distributed with mean = 30 km/h and standard deviation = 6.5 km/h. The probability distribution over the speed range [10.5, 49.5] km/h was discretised into 13 intervals of equal width. The probability of each interval was assigned to its midpoint (see Table 7.1). It was assumed that the same conditions apply to all minor roads. Off the road the average speed was taken to vary according to slope steepness: $\bar{v} = \max(0.1, 3(1-\tan\beta))$ (km/h), where β is the slope angle. Slope angle data were derived from a digital elevation model provided by the Servicio Geográfico del Ejército, Madrid.

By taking the inverse of average speed, 13 impedance grids were prepared, each with regard to a different travel speed on minor roads. Where there is no bridge, the river Guadalhorce imposed a barrier in the accessibility calculations. For simplicity, fences and other obstructions were ignored. Subsequent COSTDISTANCE calculations resulted in 13 grids that were assigned the probabilities from Table 7.1. This set of grids rendered a spatial representation of the probability distribution of local least time cost from major roads, $p(t_x)$.

Table 7.1 Discretised probability distribution of average travel speed on minor roads.

Speed (km/h)	12, 48	15, 45	18, 42	21, 39	24, 36	27, 33	30
$P(.)$	0.006	0.013	0.034	0.071	0.120	0.165	0.182

Easy accessibility is a fuzzy concept that for the purpose of this study was defined subjectively by the membership function:

$$\mu_B(t_x) = \begin{cases} 1 & \text{if } t_x \le 5 \text{ min} \\ (25 - t_x)/20 & \text{if } 5 < t_x \le 25 \text{ min} \\ 0 & \text{otherwise} \end{cases} \tag{7.3}$$

(see Figure 7.3a). Note that $\mu_B(t_x)$ is defined on a stochastic variable. If Z is a continuous stochastic variable with a probability density function $f(z)$ then the expectation of a fuzzy membership function $\mu_B(Z)$ is:

$$E[\mu_B(Z)] = \int_{-\infty}^{\infty} \mu_B(z) f(z) dz. \tag{7.4}$$

Zadeh (1968) equalled $E[\mu_B(Z)]$ to the probability of a fuzzy event, but the correctness of such an interpretation of (7.4) is questionable (see Toth, 1992). Given the discrete probability distribution $p(t_x)$ of local least time cost from major roads, the expectation of membership Function (7.3) at location x was calculated as:

$$E[\mu_B(x)] = \sum_{t_x \in T_x} \mu_B(t_x) p(t_x) \tag{7.5}$$

where T_x denotes the universal set of travel times at that location and $\mu_B(t_x)$ is defined in (7.3). Figure 7.4 illustrates the above procedure. Rounded rectangles represent processing steps, normal rectangles represent data sets and arrows represent data flow.

To compare this procedure with a crisp set (D) approach, the calculations were repeated by setting a membership threshold at the crossover point (t_x = 15 min) of Function (7.3) (see Figure 7.3b). Subsequently, all locations having $E[\mu_B(x)] > 0.5$ were assigned full membership in D.

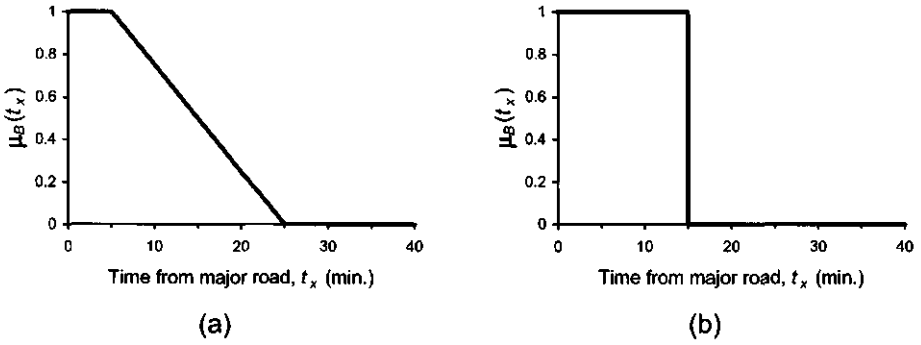


Figure 7.3 Fuzzy (a) and crisp (b) representation of the term easy accessibility.

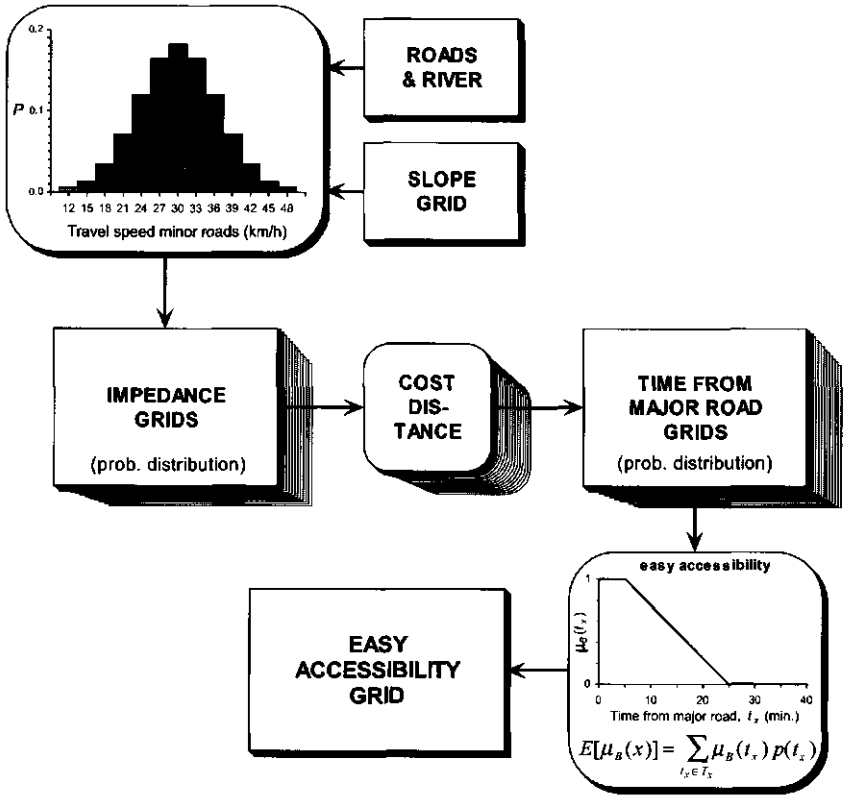


Figure 7.4 Outline of the procedure to compute local expectation of easy accessibility.

7.3.3 *Combination of the sub-query results*

The procedures described in Sections 7.3.1 and 7.3.2 result in two fuzzy sets. The first, denoted A, represents locations that with high probability have lost forest cover over the period 1984 to 1995. The second, B, represents locations that are expected to be easily accessible from a major road. The intersection of these fuzzy sets, $F = A \cap B$, provides the answer to the query of Section 7.2. It was calculated using the standard operator of fuzzy intersection (Klir and Yuan, 1995):

$$\mu_F(x) = \mu_{(A \cap B)}(x) = \min\{\mu_A(x),\ E[\mu_B(x)]\} \tag{7.6}$$

7.4 Query results

7.4.1 *Highly probable loss of forest cover*

Figure 7.5 shows locations that with high probability have lost forest cover over the period 1984 to 1995 according to crisp selection (Figure 7.5a) and fuzzy selection (Figure 7.5b) using membership Functions (7.1) and (7.2) respectively. The grids are displayed at 50 m spatial resolution to allow a better representation of grey shades. When Figure 7.5a is compared with figures 1 and 5b it can be seen that a shift of the threshold value in membership Function (7.1) may yield quite different results. Yet, the threshold (P = 0.85) was set more or less arbitrarily. This crisp selection criterion discriminated sharply between members and non-members. In either case it did not matter whether the probability of forest loss was near or far off the threshold value. Fuzzy selection, on the other hand, resulted in a grid representing the degree to which the probability of forest loss is compatible with the fuzzy concept *high probability*.

7.4.2 *Local expectation of easy accessibility*

Figure 7.6a shows expected memberships in the crisp set of locations that can be reached within 15 minutes from a major road. Figure 7.6b shows the expectations of membership in the fuzzy set of easily accessible locations. Where there are no proximate minor roads (e.g. 5 km north-northwest of Alora), travel time from major roads is assumed to be completely determinate, i.e. not subject to random variation. Consequently, uncertainty about the ease of accessibility is entirely attributable to the fuzzy definition of the concept (3). Otherwise, the uncertainty expressed in the expected memberships is due both to fluctuation of average speed on minor roads (Table 7.1) and to fuzziness of the event (see Figure 7.6).

The effect of the barrier imposed by the river Guadalhorce is most prominent in the upper left corner of Figure 7.6b, where expected membership grades decrease sharply from one to zero. This illustrates that crisp transitions are preserved when using a fuzzy set representation. To save space, the results obtained by thresholding the expectations of the crisp membership grades are not shown here.

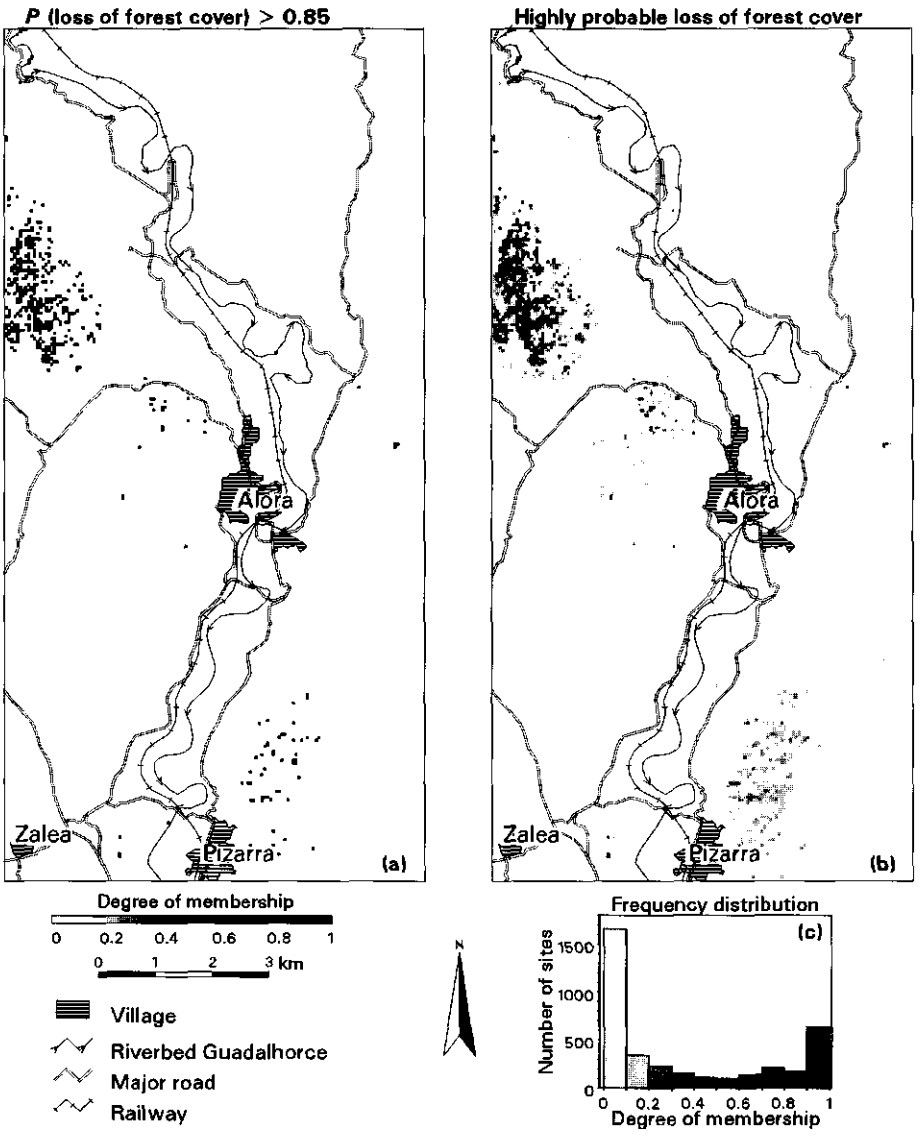


Figure 7.5 Crisp (a) and fuzzy (b) representation of locations that with high probability have lost forest cover over the period 1984 to 1995. Histogram (c) represents the frequency distribution of fuzzy membership grades (b).

7.4.3 Complete query result

Figure 7.7 visualises the intersection according to equation (7.6) of the fuzzy sets represented in figures 5b and 6b. It is displayed at 50 m spatial resolution to allow a better representation of grey shades. Although at the original resolution the number of

grid cells having non-zero membership exceeds 3000, there are only a few with membership grades close to one. These represent the locations that most closely match the vague concepts included in the query.

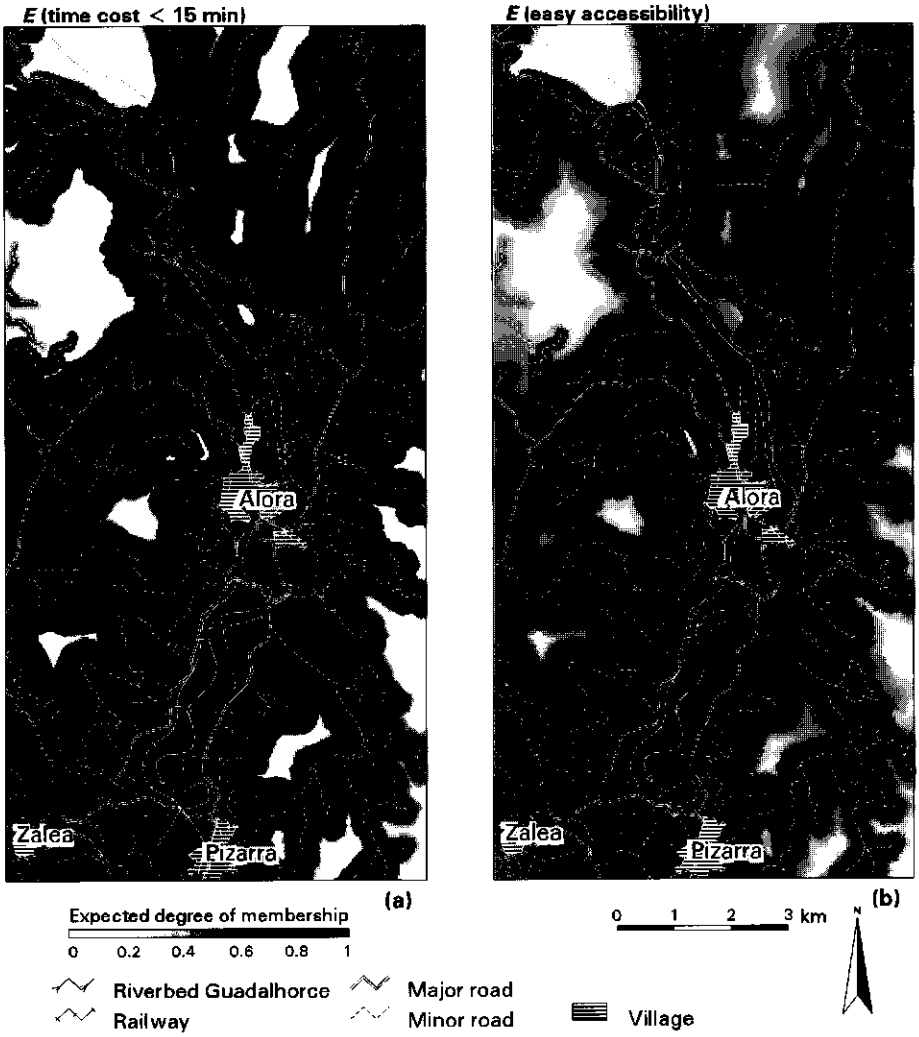


Figure 7.6 Easy accessibility; (a) expected membership in the crisp set of locations that can be reached within 15 minutes from a major road, (b) expected membership in the fuzzy set of easily accessible locations.

7.4.4 Fuzzy selection versus crisp selection

In the end a location will either be selected or rejected for conducting an experiment. A definite decision in this respect will probably be preceded by field inspection.

Selection of candidate sites that are to be checked in the field could be based on the degree to which they meet formulated preferences.

A strong α-cut of a fuzzy set, denoted ${}^{\alpha+}F$, is a crisp set that contains all elements whose membership grades in F exceed the given value of α. For example, ${}^{0+}F$, is the set that contains all elements having nonzero membership grade in F, i.e. the support of F. The cardinality of a strong α-cut, denoted $|{}^{\alpha+}F|$, specifies the number of elements in ${}^{\alpha+}F$. Table 7.2 lists cardinalities for different strong α-cuts of the sets constructed by fuzzy (F) and crisp (C) processing of the query. In the crisp query response the cardinality of ${}^{\alpha+}C$ is invariant for $\alpha \in [0, 1)$, because membership grades can only take the value zero or one. Within class variability is not expressed in the membership grades. In order to find preferred locations it would be necessary to evaluate P(loss of forest cover) and $P(t_x < 15$ min.) for the selected cells, assuming that the criterion $t_x < 15$ min. is appropriate. This could be achieved in a decision-analytical way after assigning a utility value, if a site were selected, to each of the four possible compound events involving forest loss and travel time. Preferred sites would be those having highest expected utility (see e.g. Von Winterfeldt and Edwards, 1986). However, this approach would involve additional processing steps that have not been addressed in the query.

The fuzzy query response, on the other hand, provides membership grades that express the degree to which the selection criteria are met. Sub-selection of preferred sites would just be a matter of choosing an appropriate α-cut. This approach takes advantage of the fact that subjective preferences had already been formulated in the membership functions for the vague concepts *high probability* and *easily accessible*. Rather than immediately forcing these concepts into crisp approximations, fuzzy set theory allowed to preserve their imprecise meanings until the final step preceding crisp action.

Table 7.2 Cardinalities of strong a-cuts of fuzzy (F) and crisp (C) query responses.

α	$\lvert{}^{\alpha+}F\rvert$	$\lvert{}^{\alpha+}C\rvert$
0	3452	412
0.20	1226	412
0.40	661	412
0.60	210	412
0.80	64	412
0.90	41	412
0.95	33	412

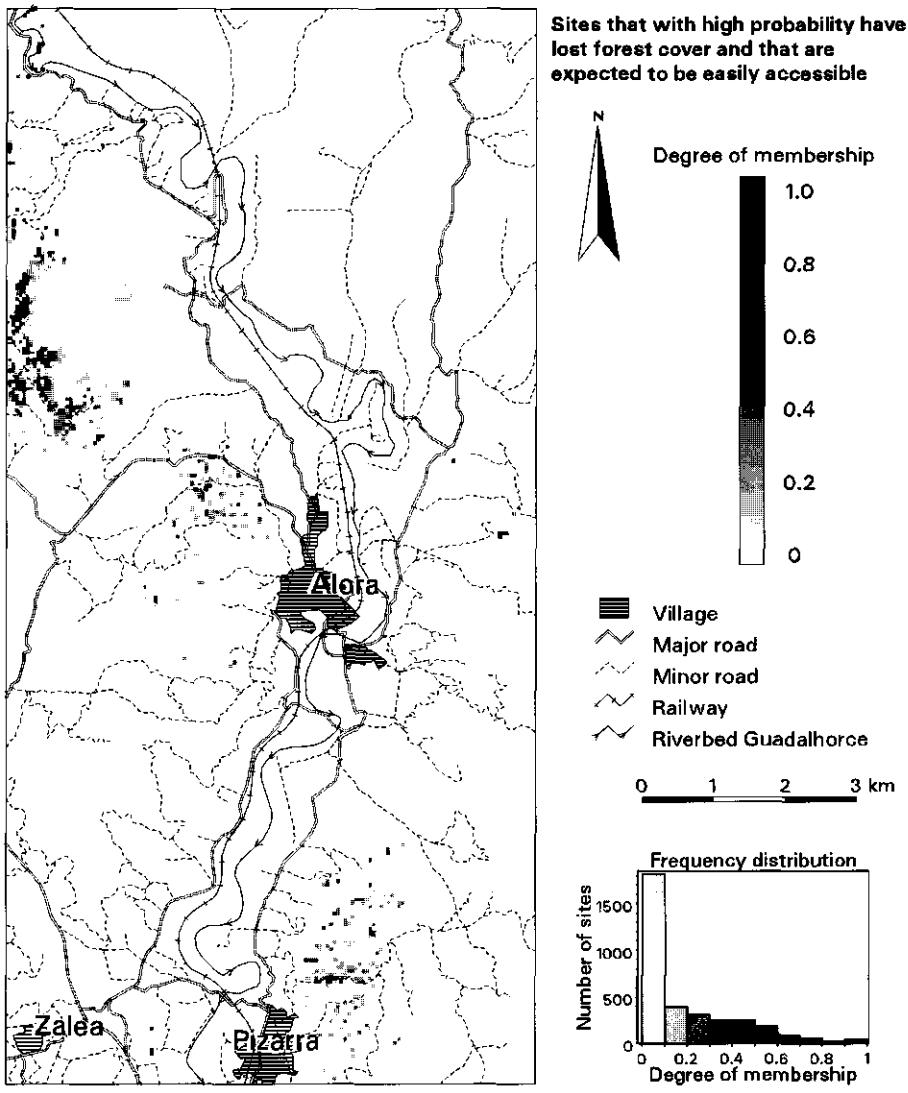


Figure 7.7 Map showing the complete query result.

7.5 Concluding remarks

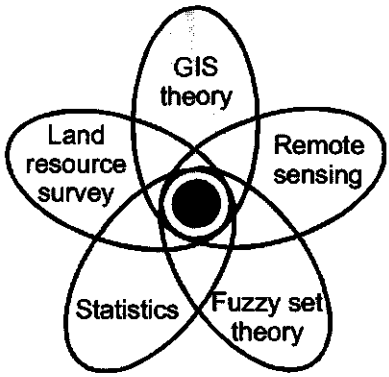
The example query demonstrated that fuzzy membership grades and probabilities of class membership can be combined to handle imprecisely formulated queries on uncertain spatial data. Query processing using fuzzy membership functions was compared with crisp processing to emphasise the difference between the two measures of uncertainty. In both cases, loss of forest cover was considered a crisp event at the cell

level, with a given probability of occurrence. The difference was in the definitions of the concepts *high probability* and *easy accessibility*.

Using crisp membership functions, the imprecise qualifier *high probability* was replaced by a probability threshold and *easy accessibility* was considered a crisp event. Consequently, the intrinsic uncertainty contained in these terms was ignored in the query response. On the contrary, fuzzy set theory allowed to deal with the vague meanings of these terms throughout query processing, but at the expense of having to construct fuzzy membership functions. Numerous methods exist for constructing membership functions on the basis of subjective judgement (e.g. Wallsten *et al.*, 1986; Klir and Yuan, 1995, Ch. 10 and the references therein). Although these methods require more information than is needed for constructing crisp membership functions, this is amply compensated by the extra sensitivity in data analysis (Burrough and McDonnell, 1998).

In the example, the combination of fuzzy membership grades and probabilities enabled processing of (1) the fuzzily qualified probability of a crisp event, and (2) membership in a fuzzy set defined on a stochastic variable. As a result, the fuzzy query response contained more information than the crisp response, namely the degree to which individual locations match the selection criteria. These degrees allowed to sub-select the most preferred sites from the complete set of selected locations. This illustrates the gain in expressive power provided by fuzzy set theory (e.g. Zadeh, 1995).

An additional advantage of fuzzy membership functions is that they are less sensitive to small data errors near critical class boundaries (Heuvelink and Burrough, 1993). In like manner, fuzzy membership functions may do more justice to computed variables that depend on many assumptions. An example of such a variable was the travel time from major roads. Although not demonstrated in this paper, a change in the assumptions used to calculate travel time will result in a change in the support of the set of easily accessible locations. If easy accessibility is defined by a crisp threshold on travel time, a number of locations will completely reverse set membership, even after only a slight shift in the assumptions. The change in degree of membership will be less pronounced, however, if easy accessibility is defined by a fuzzy membership function.


GIS
theory
Land
resource
survey
Remote
sensing
Statistics
Fuzzy set
theory

8 Concluding remarks

The general purpose of this research was to explore and demonstrate the utility of new concepts and tools for improved land resource survey. The study was motivated by the increasing use of GIS, having at least three major implications for land resource survey:

- GIS allows alternative and richer representation than is possible with the traditional paper map;
- Digital technology has improved the accessibility of ancillary data (e.g. digital elevation models, remotely sensed imagery, postcode areas) and the possibilities of incorporating these into database production;
- Owing to the greater distance between data producers and consumers and the increased use of spatial data sets in physical response models there is greater need for uncertainty analysis.

The research's main contributions with respect to each of these issues are summarised below. The chapter concludes with some suggestions for future extensions to this work.

8.1 Alternative conceptual model

In Chapter 2 the theory of fuzzy sets (Zadeh, 1965) was introduced to deal with thematic classes that are partially indistinct as a result of vague class intensions. Presence of one-to-one links between geometrical elements and their thematic description allows direct mapping of thematic fuzzy pseudopartitions to geometric space. Spatial correlation of data from nearby elements leads to their grouping into spatially contiguous regions that can be interpreted as objects with a fuzzy spatial extent.

Chapter 4 explored the use of fuzzy sets to represent transition zones in a soil-landscape. The discrete object model, being the conventional conceptual model used for soil-landscape description, is not able to deal with these. Fuzzy c-means clustering of attribute data derived from a DEM resulted in spatially contiguous fuzzy regions that were closely related to topsoil clay variability. The coefficient of determination of regressing soil sample data on fuzzy membership grades was used to decide upon the optimum fuzzy pseudopartition in terms of the number of clusters and the amount of

fuzziness. Indeed, fuzzy landscape descriptions were more closely related to sampled topsoil clay data than their (nearly) crisp counterparts.

Chapter 7 explained the difference between class indistinctness due to vague class intensions and confusion of classes as a result of inability to identify the class to which an element belongs. The former type of uncertainty was handled using fuzzy set membership grades, whereas probabilities of membership to discrete classes were used to deal with the latter. Both uncertainty measures were combined to handle a site selection query involving a mixture of uncertainties in the outcomes of random events and in the meaning of linguistic terms. A comparison of crisp and fuzzy query responses demonstrated that the latter contained more information, as it preserved the fuzziness contained in linguistic terms rather than ignoring it. An additional advantage of fuzzy class intensions is that class membership becomes less sensitive to small data errors near critical class boundaries (Heuvelink and Burrough, 1993).

8.2 Use of secondary data

In a land resource survey, data acquisition typically involves collecting a small sample of precisely measured primary data as well as a larger or even exhaustive sample of related secondary data.

Soil survey often relies on soil-landscape relationships to allow efficient mapping of soil properties. Yet, soil surveyors generally fail to communicate about the methods and models employed in deriving map units and statements about their content (Hudson, 1992; Hewitt, 1993). Chapter 3 formulated and demonstrated a methodological framework that takes advantage of a GIS to interactively formalise soil-landscape knowledge using stepwise image interpretation and inductive learning of soil-landscape relationships. It examines topology to record potential *part of* relationships between hierarchically nested terrain objects corresponding with distinct soil formation regimes. These relationships can be applied in similar areas to facilitate image interpretation by restricting possible lower level objects. GIS visualisation tools are used to create images (e.g. perspective views) illustrating the landscape configuration of interpreted terrain objects. The framework is expected to support different methods for analysing and describing soil variation in relation to a terrain description, including those requiring alternative conceptual data models. Chapter 3 though only demonstrated its use with the discrete object model.

Satellite remote sensing has become an important tool in land cover mapping, as it provides an attractive supplement to relatively inefficient ground surveys. A common approach to extract land cover data from remotely sensed imagery is by multispectral classification. Additional information can be incorporated into such classification through the use of modified prior class probabilities (Strahler, 1980; Hutchinson, 1982). This is particularly advantageous in the case of spectral overlap among target classes, i.e. when unequivocal class assignment based on spectral data alone is impossible. Chapter 5 demonstrated a procedure described by Gorte and Stein (1998) that uses intermediate classification results to iteratively adjust prior probabilities related to spatial strata. A case study, concerning land cover classification from Landsat TM imagery and image stratification based on geological map units, confirmed the method's ability to improve

classification accuracy. The study also demonstrated the use of conditional probabilities to represent uncertainty in class assignments (see below).

8.3 Uncertainty

The fact that any landscape description is a model based on a limited sample of measured target attribute data implies that it is never completely certain. One kind of uncertainty concerns fuzziness of the class intensions used in a landscape description. Fuzziness is directly related to the fuzzy object conceptual model of geographic phenomena (see section 8.1). Uncertainty may also denote a recognition of possible error or inaccuracy in the reported value. In Chapter 2 it was argued that, regardless of the conceptual model, any terrain description is affected by the latter kind of uncertainty. Error modelling gives an indication of the possible magnitude or distribution of inaccuracies for spatial attributes. In this thesis, error modelling was applied to land cover classification from remotely sensed imagery.

Chapter 5 explored the use of class probabilities conditional to spectral data, which are intermediate results of image classification, to estimate the magnitude and distribution of local uncertainty in classified imagery. A case study demonstrated the implication of such uncertainty on change analysis involving multiple classifications. A major shortcoming of the approach is that it implicitly assumes data in neighbouring pixels to be independent. Moreover, it does not make full use of available reference data as it ignores their spatial component. It does not consider data locations nor does it use spatial dependence models that may be derived from the reference data.

The assumption of independent pixels obviously impedes proper assessment of spatial uncertainty, such as joint uncertainty about the land cover class at several pixels taken together. Chapter 6 presented a geostatistical method to model spatial uncertainty in estimates of the areal extent of crisp land cover types. It employs collocated indicator co-kriging to update soft, image-derived conditional probabilities by conditioning these on sampled (hard) reference data. Unlike full co-kriging, solution of a collocated indicator co-kriging system only requires spatial dependence models for the hard data and for the cross-correlation between hard and soft data. A Markov-type assumption, stating that spatial dependence of soft data on hard data is limited to the collocated hard indicator datum, may further alleviate modelling efforts. Sequential indicator simulation was demonstrated to enable the generation of a set of data conditioned realisations, from which uncertainties regarding the spatial extent of land cover features (e.g. objects) may be inferred.

As indicated above, error or inaccuracy and fuzziness may co-occur. Chapter 7 introduced two concepts that may be used to combine the two types of uncertainty: the linguistic probability qualifier and the expected membership grade. Combining inaccuracy and fuzziness extends the expressiveness of statistical uncertainty analysis as it preserves the fuzziness contained in linguistic terms rather than ignoring it.

8.4 Further research

As anticipated in Chapter 1, the current research theme has many aspects that could not be covered in the present study. Several aspects were intentionally omitted so as to demarcate the scope of this research. These may serve as a point of departure for future extensions to this work. Three limitations mentioned in the introduction to this thesis are of particular interest in this context:

1. The research dealt with data uncertainty rather than data quality;
2. The research did not deal with all aspects of uncertainty but focused on fuzziness of class intensions and assessment of thematic accuracy;
3. Terrain descriptions were essentially two dimensional (2D), or 2.5D at the most.

8.4.1 Data quality

Unlike uncertainty, data quality explicitly refers to fitness-for-use. Yet, geographical data sets may be used many times and for widely different purposes. Today, it is quite possible that creators and users of data share little in the way of common disciplinary background, leaving the data set open to misunderstanding and misinterpretation. Easy access to data provided by the Internet and various geographical data archives further increases the possibility of inappropriate use (Goodchild and Longley, 1999).

Data quality includes such components as lineage, accuracy, precision, consistency and completeness, where the latter may be subdivided into data completeness and model completeness (Brassel *et al.*, 1995; Vegerin, 1999). Model completeness refers to the agreement between database specification and the abstract universe that is required for a particular application, i.e. data fitness (Brassel *et al.*, 1995). It is particularly this latter component which is difficult to include in data quality statements. Future research could aim at finding ways to communicate data fitness. One possibility would be to encapsulate geographic data with (references to) appropriate methods of analysis (Goodchild and Longley, 1999) and append high accuracy sample data (Fisher, 1998; Kyriakidis, *et al.*, 1999) to enable error propagation modelling up to a decision stage (e.g. Goovaerts, 1999).

8.4.2 Imprecision and resolution

An important component of uncertainty and data quality is the imprecision[1] resulting from the resolution (both spatial and thematic) at which data are represented. This thesis did not deal with imprecision. Using ideas from rough set theory, Worboys (1998) developed a formal framework for handling spatial imprecision at multiple resolutions. In future work this formalism could be extended to deal with thematic imprecision[2]. Also its usefulness in practical studies such as land resource surveys remains to be demonstrated. In this context, the original focus on discrete geographic objects probably needs to be broadened. It is also worth noting that a change in spatial resolution not only affects

[1] Also known as non-specificity (Klir and Yuan, 1995).

[2] Observed spatial variability largely depends on the level of thematic precision.

imprecision, but, due to data aggregation, may also alter the results of statistical analyses[1] (e.g. Cressie, 1996, 1998; Heuvelink, 1998b).

8.4.3 Multidimensional GIS

To further advance environmental modelling, future research should bring the methods presented in this thesis beyond the 2.5D limit (Raper, 1995). Recent literature points to several examples of implemented 3D representation tools (see Raper, 1995, 1999; Verbree *et al.*, 1999) but to date the integration of temporal processes with spatial databases is only in its infancy (Cheng, 1999; Egenhofer *et al.*, 1999; Peuquet, 1999). On the other hand, progress is being made in developing methods for statistical analysis of space-time data (Stein *et al.*, 1998; Wikle *et al.*, 1998).

[1] This effect is known as the modifiable areal unit problem (MAUP).

REFERENCES

Almeida, A.S., and Journel, A.G., 1994. Joint simulation of multiple variables with a Markov-type coregionalization model. Mathematical Geology 26, 565-588.

Altman, D., 1994. Fuzzy set theoretic approaches for handling imprecision in spatial analysis. International Journal of Geographical Information Science 8, 271-289.

Applebaum, D., 1996. Probability and Information: An Integrated Approach. Cambridge University Press, Cambridge.

Aronoff, S., 1982. Classification accuracy: a user approach. Photogrammetric Engineering and Remote Sensing 48, 1299-1307.

Atkinson, P.M., Webster,R.,and Curran, P.J. 1994. Cokriging with airborne MSS imagery. Remote Sensing of Environment 50, 335-345.

Barnsley, M.J., Barr, S.L., and Tsang, T., 1997. Scaling and generalisation in land cover mapping from satellite sensors. In: P.R. Van Gardingen, G.M. Foody and P.J. Curran. (Editors), Scaling-up: From Cell to Landscape. Cambridge University Press, Cambridge, pp. 173-199.

Bastin, L., 1997, Comparison of fuzzy c-means classification, linear mixture modelling and MLC probabilities as tools for unmixing coarse pixels. International Journal of Remote Sensing 18, 3629-3648.

Bell, J.C., Cunninghamk, R.L., and Havens, M.W., 1994. Soil drainage class probability mapping using a soil-landscape model. Soil Science Society of America Journal 58, 464-470.

Bezdek, J.C., 1981. Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York.

Blaszczynski, J.S., 1997. Landform characterization with geographic information systems. Photogrammetric Engineering and Remote Sensing 63, 183-191.

Boucneau, G., Van Meirvenne, M., Thas, O., and Hofman, G., 1998. Integrating properties of soil map delineations into ordinary kriging. European Journal of Soil Science 49, 213-229.

Bouma, J., and Hoosbeek, M.R., 1996. The contribution and importance of soil scientists in interdisciplinary studies dealing with land. In: R.J. Wagenet and J. Bouma (Editors), The Role of Soil Science in Interdisciplinary Research. SSSA Special Publication 45, American Society of Agronomy/Soil Science Society of America, Madison, WI, pp. 1-15.

Brassel, K., Bucher, F., Stephan, E.M., and Vckovski, A., 1995. Completeness. In: S.C. Guptill and J.L. Morrison (Editors), Elements of Spatial Data Quality. Elsevier Science, Amsterdam, pp. 81-108.

Bruzzone, L., and Serpico, S.B., 1997. An iterative technique for the detection of land-cover transitions in multitemporal remote sensing images. IEEE Transactions on Geoscience and Remote Sensing 35, 858-867.

Burrough, P.A., 1996. Natural objects with indeterminate boundaries. In: P.A. Burrough and A.U. Frank (Editors), Geographic Objects with Indeterminate Boundaries. Taylor & Francis, London, pp. 3-28.

Burrough, P.A., and Frank, A.U., 1995. Concepts and paradigms in spatial information: are current geographical information systems truly generic? International Journal of Geographical Information Science 9, 101-116.

Burrough, P.A., and McDonnell, R.A., 1998. Principles of Geographical Information Systems. Oxford University Press, Oxford.

Burrough, P.A., Van Gaans, P.F.M., and Hootsmans, R., 1997. Continuous classification in soil survey: spatial correlation, confusion and boundaries. Geoderma 77, 115-135.

Cadwell, L.L., Downs, J.L., Phelps, C.M., Nugent, J. J., Marsh, L., and Fitzner, L., 1996, Sagebrush restoration in the shrub-steppe of south-central Washington. Proceedings Shrubland Ecosystem Dynamics in a Changing Environment, Las Cruces, 23-25 May 1995. General Technical Report No. INT-GTR-338, USDA Forest Service, Ogden, UT, pp. 143-145.

Campbell, J.B., 1983. Mapping the Land: Aerial Imagery for Land Use Information. Resource Publications in Geography. Association of American Geographers, Washington, D.C.

Canters, F., 1997. Evaluating the uncertainty of area estimates derived from fuzzy land-cover classification. Photogrammetric Engineering and Remote Sensing 63, 403-414.

Cheng, T., 1999. A Process-Oriented Data Model for Fuzzy Spatial Objects. ITC Publication Series No. 68, International Institute for Aerospace Survey and Earth Sciences (ITC), Enschede.

Cliff, N., 1990, Comment on: F Mosteller, and C. Youtz, Quantifying probabilistic expressions. Statistical Science 5, 16-18.

Coléou, T., 1999. Links between external drift, Bayesian kriging, collocated cokriging. CGG-Petrosystems Hints, http://www.cgg.com/software/pts/hints/KR1.html.

Conacher, A.J., and Dalrymple, J.B., 1977. The nine-unit landsurface model: an approach to pedogeomorphic research. Geoderma 18, 1-154.

Conese, C., and Maselli, F., 1993. Selection of optimum bands from TM scenes through mutual information analysis. ISPRS Journal of Photogrammetry and Remote Sensing 48, 2-11.

Congalton, R.G., 1991. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sensing of Environment 37, 35-46.

Congalton, R.G., Oderwald, R.G., and Mead, R.A., 1983. Assessing Landsat classification accuracy using discrete multivariate analysis statistical techniques. Photogrammetric Engineering and Remote Sensing 49, 1671-1678.

Cook, S.E., Corner, R.J., Grealish, G., Gessler, P.E., and Chartres, C.J., 1996. A rule-based system to map soil properties. Soil Science Society of America Journal 60, 1893-1900.

Couclelis, H., 1996. Towards an operational typology of geographic entities with ill-defined boundaries. In: P.A. Burrough and A.U. Frank (Editors), Geographic Objects with Indeterminate Boundaries. Taylor & Francis, London, pp. 45-55.

Cressie, N., 1991. Statistics for Spatial Data. John Wiley & Sons, New York.

Cressie, N., 1996. Change of support and the modifiable areal unit problem. Geographical Systems, 3, 159-180.

Cressie, N., 1998. Aggregation and interaction issues in statistical modeling of spatiotemporal processes. Geoderma, 85, 133-140.

Dalrymple, J.B., Blong, R.J., and Conacher, A.J., 1968. An hypothetical nine unit landsurface model. Zeitschrift für Geomorphologie 12, 60-76.

De Bruin, S., 2000. Querying probabilistic land cover data using fuzzy set theory. International Journal of Geographical Information Science (*in press*).

De Bruin, S., and Wielemaker, W.G., 1996. GIS applications in land resource studies. Lecture notes, Department of Soil Science and Geology, Wageningen Agricultural University, The Netherlands.

De Bruin, S., and Stein, A., 1998. Soil-landscape modelling using fuzzy c-means clustering of attribute data derived from a digital elevation model (DEM). Geoderma 83, 17-33.

De Bruin, S., and Molenaar, M., 1999. Remote sensing and geographical information systems. In: A. Stein, F.D. Van der Meer, and B.G.H. Gorte (Editors), Spatial Statistics for Remote Sensing. Kluwer Academic Publishers, Dordrecht, pp. 41-54.

De Bruin, S., and Gorte, B.G.H., 2000. Probabilistic image classification using geological map delineations applied to land cover change detection. International Journal of Remote Sensing (*in press*).

De Bruin, S., Wielemaker, W.G., and Molenaar, M., 1999. Formalisation of soil landscape knowledge through interactive hierarchical disaggregation. Geoderma 91, 151-172.

De Gruijter, J.J., Walvoort, D.J.J., and 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, 169-195.

De Leon, A., Arriba, A., and De La Plaza, M., 1989. Caracterización Agroclimática de la Provincia de Málaga. Ministerio de Agricultura, Pesca y Alimentación, Madrid.

DeFries, R.S., and Townshend, J.R.G., 1994. NDVI-derived land cover classifications at a global scale. International Journal of Remote Sensing 15, 3567-3586.

Deka, B., Sawhney, J.S., Sharma, B.D., and Sidhu, P.S., 1995. Soil-landscape relationships in Siwalik hills of the semiarid tract of Punjab, India. Arid Soil Research and Rehabilitation 10, 149-159.

Dent, D., and Young, A., 1981. Soil Survey and Land Evaluation. George Allen & Unwin, London.

Deutsch, C.V., and Journel, A.G., 1998. GSLIB Geostatistical Software Library and User's Guide, 2nd Edition. Oxford University Press, New York.

Domburg, P., De Gruijter, J.J., and Van Beek, P., 1997. Designing efficient soil survey schemes with a knowledge-based system using dynamic programming. Geoderma 75, 183-201.

Droesen, W.J., 1999. Spatial modelling and monitoring of natural landscapes with cases in the Amsterdam Waterworks Dunes. Thesis Wageningen University.

Duda, R.O., and Hart, P.E., 1973. Pattern Classification and Scene Analysis. John Wiley & Sons, New York, pp. 10-38.

Dungan, J.L., 1998. Spatial prediction of vegetation quantities using ground and image data. International Journal of Remote Sensing 19, 267-285.

Eastman, J.R., 1997. Idrisi for Windows User's Guide Version 2.0. Clark University, Worcester, MA.

Eastman, J.R., 1999. Multi-criteria evaluation and GIS. In: P.A. Longley, M.F. Goodchild, D.J. Maguire and D.W. Rhind (Editors), Geographical Information Systems - Volume 1, Principles and Technical Issues, 2nd Edition. John Wiley & Sons, New York, pp. 493-502.

Egenhofer, M.J., Glasgow, J., Günter, O., Herring, J.R., and Peuquet, D.J., 1999. Progress in computational methods for representing geographical concepts. International Journal of Geographical Information Science 13, 775-796.

ESRI, 1994a. Cell-based Modeling with GRID. Environmental Systems Research Institute, Redlands, CA.

ESRI, 1994b. GRID Commands. Environmental Systems Research Institute, Redlands, CA.

Estes, J.E., Hajic, E.J., and Tinney, L.R., 1983. Fundamentals of image analysis: analysis of visible and thermal infrared data. In: D.S. Simonett and F.T. Ulaby (Editors), Manual of Remote Sensing - Volume I, 2nd Edition. American Society of Photogrammetry, Falls Curch, VA, pp. 987-1124.

Fisher, P.F., 1991. Modelling soil map-unit inclusions by Monte Carlo simulation. International Journal of Geographical Information Systems, 5, 193-208.

Fisher, P.F., 1994a. Probable and fuzzy models of the viewshed operation. In: M.F. Worboys (Editor), Innovations in GIS 1: Selected Papers from the First Conference on GIS Research UK. Taylor & Francis, London, pp.161-175.

Fisher, P.F., 1994b. Visualization of the reliability in classified remote sensing images. Photogrammetric Engineering & Remote Sensing 60, 905-910.

Fisher, P.F., 1996. Boolean and fuzzy regions. In: P.A. Burrough and A.U. Frank (Editors), Geographic Objects with Indeterminate Boundaries. Taylor & Francis, London, pp. 87-94.

Fisher, P.F., 1997. The pixel: a snare and a delusion. International Journal of Remote Sensing, 18, 679-685.

Fisher, P.F., 1998. Improved modeling of elevation error with geostatistics. GeoInformatica 2, 215-233.

Fleischer, P., Bowles, F.A., and Richardson, M.D., 1998. Identification of potential sites for deep-ocean waste isolation with a geographic site-selection model. Journal of Marine Systems 14, 241-271.

Foody, G.M., 1992. A fuzzy sets approach to the representation of vegetation continua from remotely sensed data: an example from lowland heath. Photogrammetric Engineering & Remote Sensing 58, 221-225.

Foody, G.M., 1996. Approaches for the production and evaluation of fuzzy land cover classifications from remotely-sensed data. International Journal of Remote Sensing 17, 1317-1340.

Foody, G.M., 1997. Fully fuzzy supervised classification of land cover from remotely sensed imagery with an artificial neural network. Neural Computing & Applications 5, 238-247.

Foody, G.M., and Trodd, N.M., 1993. Non-classificatory analysis and representation of heathland vegetation from remotely sensed imagery. GeoJournal 29, 343-350.

Foody, G.M., Campbell, N.A., Trodd N.M., and Wood, T.F., 1992. Derivation and application of probabilistic measures of class membership from the maximum-likelihood classification. Photogrammetric Engineering and Remote Sensing 58, 1335-1341.

Frank, A.U., 1996. The prevalence of objects with sharp boundaries in GIS. In: P.A. Burrough, and A.U. Frank (Editors), Geographic Objects with Indeterminate Boundaries. Taylor & Francis, London, pp. 29-40.

Freeman, T.G., 1991. Calculating catchment area with divergent flow based on a regular grid. Computers and Geosciences 17, 413-422.

Fukunaga, K., and Hummels, D.M., 1987. Bayes error estimation using Parzen and k-NN procedures. IEEE Transactions on Pattern Analysis and Machine Intelligence 9, 634-643.

Gessler, P.E., Moore, A.W., Mckenzie, N.J., and Ryan, P.J., 1995. Soil-landscape modelling and spatial prediction of soil attributes. International Journal of Geographical Information Systems 9, 421-432.

Gillespie, M.K., Howard, D.C. Ness M.J., and Fuller, R.M., 1996. Linking satellite and field survey data, through the use of GIS, as implemented in Great Britain in the Countryside Survey 1990 project. Environmental Monitoring and Assessment 39, 385-398.

Gómez-Hernández, J.J., and Srivastava, R.M., 1990. ISIM3D: an ANSI-C three-dimensional multiple indicator conditional simulation program. Computers and Geosciences 16, 395-440.

Goodchild, M.F., and Longley, P.A., 1999. The future of GIS and spatial analysis. In: P.A. Longley, M.F. Goodchild, D.J. Maguire and D.W. Rhind (Editors), Geographical Information Systems - Volume 1, Principles and Technical Issues, 2nd Edition. John Wiley & Sons, New York, pp. 567-580.

Goodchild, M.F., Sun, G., and Yang, S., 1992. Development and test of an error model for categorical data. International Journal of Geographical Information Systems 6, 87-104.

Goodman, L.A., and Kruskal, W.H., 1954. Measures of association for cross classifications. Journal of the American Statistical Association 49, 732-764.

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

Goovaerts, P., 1999. Geostatistics in soil science: state-of-the-art and perspectives. Geoderma 89, 1-45.

Goovaerts, P., and Journel, A.G., 1995. Integrating soil map information in modelling the spatial variation of continuous soil properties. European Journal of Soil Science 46, 397-414.

Gordon, A.D., 1981. Classification. Chapman and Hall, London.

Gorte, B.G.H., 1998. Probabilistic Segmentation of Remotely Sensed Images. ITC Publication Series No. 63, International Institute for Aerospace Survey and Earth Sciences (ITC), Enschede.

Gorte, B.G.H., and Stein, A., 1998. Bayesian classification and class area estimation of satellite images using stratification. IEEE Transactions on Geoscience and Remote Sensing 36, 803-812.

Hall, G.F., 1983. Pedology and geomorphology. In: L.P. Wilding, N.E. Smeck, and G.F. Hall (Editors), Pedogenesis and Soil Taxonomy. Elsevier Science Publishers, Amsterdam, pp. 117-140.

Hall, G.F., and Olson, C.G., 1991. Predicting variability of soils from landscape models. In: M.J. Mausbach, and L.P. Wilding (editors), Spatial Variabilities of Soils and Landforms. SSSA Special Publication No. 28, Soil Science Society of America, Madison, WI, pp. 9-24.

Hendricks-Franssen, H.J.W.M., Van Eijnsbergen, A.C., and Stein, A., 1997. Use of spatial prediction techniques and fuzzy classification for mapping of soil pollutants. Geoderma 77, 243-262

Hendrix, W.G., and Buckley, D.J.A., 1992. Use of a geographic information system for selection of sites for land application of sewage waste. Journal of Soil and Water Conservation 47, 271-274.

Heuvelink, G.B.M., 1993. Error Propagation in Quantitative Spatial Modelling: Applications in Geographical Information Systems. Netherlands Geographical Studies No. 163, Koninklijk Nederlands Aardrijkskundig Genootschap, Utrecht.

Heuvelink, G.B.M., 1998a. Error propagation in environmental modelling with GIS. Taylor & Francis, London.

Heuvelink, G.B.M., 1998b. Uncertainty analysis in environmental modelling under a change of spatial scale. Nutrient Cycling in Agroecosystems 50, 255-264.

Heuvelink, G.B.M., and Burrough, P.A., 1993. Error propagation in cartographic modelling using Boolean logic and continuous classification. International Journal of Geographical Information Systems 7, 231-246.

Heuvelink, G.B.M., Burrough, P.A., and Stein, A., 1989. Propagation of errors in spatial modelling with GIS. International Journal of Geographical Information Science 3, 303-322.

Hewitt, A.E., 1993. Predictive modelling in soil survey. Soils and Fertilizers 56, 305-314.

Hole, F.D., and Campbell, J.B., 1985. Soil Landscape Analysis. Routledge & Kegan Paul, London.

Hootsmans, R.M., 1996. Fuzzy Sets and Series Analysis for Visual Decision Support in Spatial Data Exploration. Koninklijk Nederlands Aardrijkskundig Genootschap / Faculteit Ruimtelijke Wetenschappen Universiteit Utrecht, Utrecht.

Hudson, B.D., 1990. Concepts of soil mapping and interpretation. Soil Survey Horizons 31, 63-72.

Hudson, B.D., 1992. The soil survey as paradigm-based science. Soil Science Society of America Journal 56, 836-841.

Hunter, G.T., and Goodchild, M.F., 1995. Dealing with error in Spatial databases: a simple case study. Photogrammetric Engineering & Remote Sensing 61, 529-537.

Hutchinson, C.F., 1982. Techniques for combining Landsat and ancillary data for digital classification improvement. Photogrammetric Engineering and Remote Sensing 48, 123-130.

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.

IGME, 1978. Mapa Geológico de España 1: 50.000, No. 1052 (Alora) (Map sheet and explanatory text) Instituto Tecnológico Geominero de España, Madrid.

Indorante, S.J., McLeese, R.L., Hammer, R.D., Thompson, B.W., and Alexander, D.L., 1996. Positioning soil survey for the 21st century. Journal of Soil and Water Conservation 51, 21-28.

Irvin, B.J., Ventura, S.J., and 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., and Srivastava, R.M., 1989. Applied Geostatistics. Oxford University Press, Oxford.

ITGE, 1991. Mapa Geológico de España 1: 50.000, No. 1038 (Ardales) (Map sheet and explanatory text) Instituto Tecnológico Geominero de España, Madrid.

Janssen, L.L.F., and Middelkoop, H., 1992. Knowledge-based crop classification of a Landsat Thematic Mapper image. International Journal of Remote Sensing 13, 2827-2837.

Jenny, H., 1941. Factors of Soil Formation - A System of Quantitative Pedology. McGraw-Hill, New York.

Jensen, J.R., Cowen, D., Narumalani, S., and Halls, J., 1997. Principles of change detection using digital remote sensor data. In: J.L. Star, J.E. Estes, and K.C. McGwire (Editors), Integration of Geographic Information Systems and Remote Sensing. Cambridge University Press, Cambridge, pp. 37-54.

Journel, A.G., 1986. Constrained interpolation and qualitative information - the soft kriging approach. Mathematical Geology 18, 269-286.

Journel, A.G., 1996. Modelling uncertainty and spatial dependence: stochastic imaging. International Journal of Geographical Information Systems 10, 517-522.

Junta de Andalucía, 1995. Usos y Coberturas Vegetales del Suelo en Andalucía: Seguimiento a través de Imágenes de Satélite. Consejería de Medio Ambiente - Junta de Andalucía, Sevilla.

Kandel, A., 1986. Fuzzy Mathematical Techniques with Applications. Addison-Wesley Publishing Company, Reading.

Klinkenberg, B., and Joy, M., 1994. Visualizing uncertainty: succession or misclassification? GIS/LIS Proceedings 25-27 October 1994, Phoenix, Arizona. American Society for Photogrammetry and Remote Sensing, Bethesda, pp. 494-503.

Klir, G.J., and Yuan, B., 1995. Fuzzy Sets and Fuzzy Logic: Theory and Applications. Prentice-Hall, Upper Saddle River, NJ.

Kyriakidis, P.C., 1999. Stochastic imaging for assessing the impact of imprecise spatial information on ecological models. Conference on Spatial Statistics for Production Ecology, April 19-21, 1999, Wageningen.

Kyriakidis, P.C., Shortridge, A.M., and Goodchild, M.F., 1999. Geostatistics for conflation and accuracy assessment of digital elevation models. International Journal of Geographical Information Science 13, 677-707.

Lagacherie, P., and Holmes, S., 1997. Addressing geographical data errors in a classification tree for soil unit prediction. International Journal of Geographical Information Science 11, 183-198.

Lagacherie, P., Legros, J.P., and Burrough, P.A., 1995. A soil survey procedure using the knowledge of soil pattern established on a previously mapped reference area. Geoderma 65, 283-301.

Lagacherie, P., Andrieux, P., and Bouzigues, R., 1996. Fuzziness and uncertainty of soil boundaries: from reality to coding in GIS. In: P.A. Burrough and A.U. Frank (Editors), Geographic Objects with Indeterminate Boundaries. Taylor & Francis, London, pp. 275-286.

Lark, R.M., and Bolam, H.C., 1997. Uncertainty in prediction and interpretation of spatially variable data on soils. Geoderma 77, 263-282.

Laviolette, M., Seaman, J.W., Barrett, J.D., and Woodall, W.H., 1995. A probabilistic and statistical view of fuzzy methods (with discussion). Technometrics 37, 249-292.

Longley, P.A., Goodchild, M.F., Maguire, D.J., and Rhind, D.W., 1999. Introduction. In: P.A. Longley, M.F. Goodchild, D.J. Maguire and D.W. Rhind (Editors), Geographical Information Systems - Volume 1, Principles and Technical Issues, 2nd Edition. John Wiley & Sons, New York, pp. 1-20.

Ma, X., and Journel, A.G., 1999. An expanded GSLIB cokriging program allowing for two Markov models. Computers and Geosciences 25, 627-639.

Manton, K.G., Woodbury, M.A., and Tolley, H.D., 1994. Statistical Applications Using Fuzzy Sets. John Wiley & Sons, New York.

Maselli, F., Conese, C., and Petkov, L., 1994. Use of probability entropy for the estimation and graphical representation of the accuracy of maximum likelihood classifications. ISPRS Journal of Photogrammetry and Remote Sensing 49, 13-20.

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

McBratney, A.B., and De Gruijter, J.J., 1992. A continuum approach to soil classification by modified fuzzy k-means with extragrades. Journal of Soil Science 43, 159-175.

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

McBratney, A.B., Hart, G.A., and McGarry, D.,1991. The use of region partitioning to improve the representation of geostatistically mapped soil attributes. Journal of Soil Science 42, 513-532.

McKenzie, N.J., and Austin, M.P., 1993. A quantitative Australian approach to medium and small scale surveys based on soil stratigraphy and environmental correlation. Geoderma 57, 329-355.

McLeod, M., Rijkse, W.C., and Dymond, J.R., 1995. A soil-landscape model for close-jointed mudstone, Gisborne-East Cape, North Island, New Zealand. Australian Journal of Soil Research 33, 381-396.

Miller, A.B., Bryant, E.S., and Birnie, R.W., 1998. An analysis of land cover changes in the Northern Forest of New England using multitemporal Landsat MSS data. International Journal of Remote Sensing 19, 245-265.

Milligan, G.W., 1996. Clustering validation: results and implications for applied analyses. In: P. Arabie, L.J. Hubert and G. de Soete (Editors), Clustering and Classification. World Scientific Publishers, River Edge, NJ, pp. 341-375.

Ministerio de Agricultura, 1978. Evaluación de Recursos Agrários : Mapa de Cultivos y Aprovechamiento, Map sheets 1038 (Ardales) and 1052 (Alora). Ministerio de Agricultura, Madrid.

Molenaar, M., 1989. Towards a geographic information theory. ITC Journal 1989, 5-11.

Molenaar, M., 1993. Object hierarchies and uncertainty in GIS, or why is standardisation so difficult? Geo-Informations-Systeme 6(4), 22-28.

Molenaar, M., 1994. A syntax for the representation of fuzzy spatial objects. In: M. Molenaar and S. De Hoop (Editors), Advanced Geographic Data Modelling: Spatial Modelling and Query Languages for 2D and 3D Applications. Publications on Geodesy No. 40, Nederlandse Commisie voor Geodesie, Delft, pp. 155-169.

Molenaar, M., 1996. The role of topologic and hierarchical spatial object models in database generalization. In: M. Molenaar (Editor), Methods for the Generalization of Geo-Databases. Publications on Geodesy 43. Netherlands Geodetic Commision, Delft, pp. 13-35.

Molenaar, M., 1998. An Introduction to the Theory of Spatial Object Modelling for GIS. Taylor & Francis, London.

Molenaar, M., and Janssen, L.L.F., 1994. Terrain objects, their dynamics and their monitoring by the integration of GIS and remote sensing. In: H. Ebner, C. Heipke and K. Eder (Editors), Spatial Information from Digital Photogrammetry and Computer Vision, Symposium ISPRS Commision III, Munich, September 5-9, 1994. SPIE-The International Society for Optical Engineering, Bellingham, WA, pp. 585-591.

Moore, I.D., Gessler, P.E., Nielsen G.A., and Petersen, G.A., 1993. Soil attribute prediction using terrain analysis. Soil Science Society of America Journal 57, 443-452.

Moore, I.D.,. Grayson R.B and Ladson, A.R., 1991. Digital terrain modelling: a review of hydrological, geomorphological, and biological applications. Hydrological Processes 5, 3-30.

Mosteller, F., and Youtz, C., 1990. Quantifying probabilistic expressions (with discussion). Statistical Science 5, 2-34.

Nguyen, H.T., 1997. Fuzzy sets and probability. Fuzzy Sets and Systems 90, 129-132.

Odeh, I.O.A., Chittleborough, D.J., and McBratney, A.B., 1991. Elucidation of soil-landform interrelationships by canonical ordination analysis. Geoderma, 49, 1-32.

Odeh, I.O.H., McBratney, A.B., and Chittleborough, D.J., 1992a. Soil pattern recognition with fuzzy-*c*-means: application to classification and soil-landform interrelationships. Soil Science Society of America Journal 56, 505-516.

Odeh, I.O.H., McBratney, A.B., and Chittleborough, D.J., 1992b. Fuzzy-*c*-means and Kriging for mapping soil as a continuous system. Soil Science Society of America Journal 56, 1848-1854.

Odeh, I.O.H., McBratney, A.B., and Chittleborough, D.J., 1994. Spatial prediction of soil properties from landform attributes derived from a digital elevation model. Geoderma 63, 197-214.

Olson, C.E., 1973. What is photographic interpretation? In: R.K. Holz (Editor), The Surveillant Science - Remote Sensing of the Environment. Houghton Mifflin Company, Boston, MA, pp. 95-102.

Pebesma, E.J., 1998. Gstat user's manual. Utrecht University, Department of Physical Geography, http://www.geog.uu.nl/gstat/.

Pebesma, E.J., and Wesseling, C.G., 1998. GSTAT: a program for geostatistical modelling, prediction and simulation. Computers and Geosciences 24, 17-31.

Peuquet, D.J., 1999. Time in GIS and geographical databases. In: P.A. Longley, M.F. Goodchild, D.J. Maguire and D.W. Rhind (Editors), Geographical Information Systems - Volume 1, Principles and Technical Issues, 2nd Edition. John Wiley & Sons, New York, pp. 91-103.

Peuquet, D.J., and Duan, N., 1995. An event-based spatiotemporal data model (ESTDM) for temporal analysis of geographical data. International Journal of Geographical Information Systems 9, 7-24.

Quinn, P.F., Beven, K.J., and Lamb, R., 1995. The ln(a/tanb) index: how to calculate it and how to use it within the TOPMODEL framework. Hydrological Processes 9, 161-182.

Raper, J.F., 1995. Making GIS multidimensional. JEC-GI '95 Proceedings (Volume 1) From Research to Application through Cooperation. Joint European Conference and Exhibition on Geographical Information, The Hague, March 26-31, pp. 232-240.

Raper, J.F., 1999. Spatial representation: the scientist's perspective. In: P.A. Longley, M.F. Goodchild, D.J. Maguire and D.W. Rhind (Editors), Geographical Information Systems - Volume 1, Principles and Technical Issues, 2nd Edition. John Wiley & Sons, New York, pp. 61-70.

Robinove, C.J., 1981. The logic of multispectral classification and mapping of land. Remote Sensing of Environment 11, 231-244.

Rosenfield, G.H., and Fitzpatrick-Lins, K., 1986. A coefficient of agreement as a measure of thematic classification accuracy. Photogrammetric Engineering and Remote Sensing 52, 223-227.

Roubens, M., 1982. Fuzzy clustering algorithms and their cluster validity. European Journal of Operational Research 10, 294-301.

Rousseeuw, P.J., 1995. Fuzzy clustering at the intersection. In: Discussion of: M. Laviolette, J.W. Seaman, J.D. Barrett and W.H. Woodhall, A probabilistic and statistical view of fuzzy methods. Technometrics 37, 283-286.

Ruhe, R.V., 1960. Elements of the soil landscape. Transactions of 7th International Congress of Soil Science. International Society of Soil Science, Madison, WI, pp. 165-170.

Schowengerdt, R.A., 1996. On the estimation of spatial-spectral mixing with classifier likelihood functions. Pattern Recognition Letters 17, 1379-1387.

Servicio Geográfico del Ejército, 1995. Cartografía Militar de España, Serie L, Escala 1:50.000, Map sheets 16-43 (Ardales) and 16-44 (Alora). Servicio Geográfico del Ejército, Madrid.

Shannon, C. E., and Weaver, W., 1949. The Mathematical Theory of Communications. University of Illinois Press, Urbana.

Siderius, W., and Elbersen, G.W.W., 1986. Drip irrigation as a method for soil and water conservation in sloping areas: a case study from Malaga province, Spain. In: W. Siderius (Editor), Land Evaluation for Land Use Planning and Conservation in Sloping Areas - International Workshop, Enschede, The Netherlands, 17-21 December 1984, International Institute for Land Reclamation and Improvement, Wageningen, pp. 263-289.

Singh, A., 1989. Digital change detection techniques using remotely-sensed data. International Journal of Remote Sensing 10, 989-1003.

Slater, B.K., McSweeney, K., Ventura, S.J., Irvin, B.J., and McBratney, A.B., 1994. A spatial framework for integrating soil-landscape and pedogenetic models. In: R.B. Bryant and R.W. Arnold (Editors), Quantitative Modeling of Soil Forming Processes. Soil Science Society of America, Madison, WI, pp. 169-185.

Snedecor, G.W., and Cochran, W.G., 1989. Statistical Methods, 8th Edition. Iowa State University Press, Ames.

Soares, A., 1992. Geostatistical estimation of multiphase structures. Mathematical Geology 24,149-160.

Soil Survey Division Staff, 1993. Soil Survey Manual. United States Department of Agriculture, Washington.

Soil Survey Staff, 1996. Keys to Soil Taxonomy, 7th Edition. Technical Monograph 19. Pocahontas Press, Blacksburg, VA.

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

Stein, A., Van Groenigen, J.W., Jeger M.J., and Hoosbeek, M.R., 1998. Space-time statistics for environmental and agricultural related phenomena. Environmental and Ecological Statistics 5, 155-172.

Story, M., and Congalton, R.G., 1986. Accuracy assessment: a user's perspective. Photogrammetric Engineering and Remote Sensing 52, 397-399.

Strahler, A.H., 1980. The use of prior probabilities in maximum likelihood classification of remotely sensed data. Remote Sensing of Environment 10, 1135-163.

Strahler, A.H., Woodcock C.E., and Smith, J.A., 1986. On the nature of models in remote sensing. Remote Sensing of Environment 20, 121-139.

Therrien, C.W., 1989. Decision, Estimation and Classification. John Wiley & Sons, New York.

Thompson, J.A., Bell, J.C., and Butler, C.A., 1997. Quantitative soil-landscape modeling for estimating the areal extent of hydromorphic soils. Soil Science Society of America Journal 61, 971-980.

Toth, H., 1992. Probabilities and fuzzy events: an operational approach. Fuzzy Sets and Systems 48, 113-127.

Unwin, D.J., 1995. Geographical information systems and the problem of 'error and uncertainty'. Progress in Human Geography 19, 549-558.

Van der Wel, F.J.M., Van der Gaag, L.C., and Gorte, B.G.H., 1998. Visual exploration of uncertainty in remote sensing classification. Computers and Geosciences 24, 335-343.

Van Groenigen, J.W., and Stein, A., 1998. Constrained optimization of spatial sampling using continuous simulated annealing. Journal of Environmental Quality 27, 1078-1086.

Van Reeuwijk, L.P., 1992. Procedures for Soil Analysis, 3rd Edition.. Technical Paper No. 9, International Soil Reference and Information Centre, Wageningen.

Van Ryzin, J., 1977. Classification and Clustering : Proceedings of an Advanced Seminar, the University of Wisconsin at Madison, May 3 - 5, 1976. Academic Press, New York.

Verbree, E., Van Maren, G., Germs, R., Jansen, F., and Kraak, M-J., 1999. Interaction in virtual world views - linking 3D GIS with VR. International Journal of Geographical Information Science 13, 385-396.

Veregin, H., 1999. Data quality parameters. In: P.A. Longley, M.F. Goodchild, D.J. Maguire and D.W. Rhind (Editors), Geographical Information Systems - Volume 1, Principles and Technical Issues, 2nd Edition. John Wiley & Sons, New York, pp. 177-189.

Vogelmann, J.E., Sohl, T.L., Campbell, P.V., and Shaw, D.M., 1998. Regional land cover characterization using Landsat Thematic Mapper data and ancillary data sources. Environmental Monitoring and Assessment 51, 415-428.

Von Winterfeldt, D., and Edwards, W., 1986. Decision Analysis and Behavioral Research. Cambridge University Press, Cambridge.

Wallsten, T.S., and Budescu, D.V., 1990. Comment on: F. Mosteller, and C. Youtz, Quantifying probabilistic expressions. Statistical Science 5, 23-26.

Wallsten, T.S., and Budescu, D.V., 1995. A review of human linguistic probability processing: general principles and empirical evidence. The Knowledge Engineering Review 10, 43-62.

Wallsten, T.S., Budescu, D.V., Rapoport, A., Zwick, R., and Forsyth, B., 1986. Measuring the vague meanings of probability terms. Journal of Experimental Psychology: General 115, 348-365.

Wang, F., 1990a. Fuzzy supervised classification of remote sensing images. IEEE Transactions on Geoscience and Remote Sensing 28, 194-201.

Wang, F., 1990b. Improving remote sensing image analysis through fuzzy information representation. Photogrammetric Engineering and Remote Sensing 56, 1163-1169.

Webster, R., and Oliver, M.A., 1990. Statistical Methods in Soil and Land Resource Survey. Oxford University Press, Oxford.

Wikle, C.K., Berliner, L.M., and Cressie, N., 1998. Hierarchical Bayesian space-time models. Environmental and Ecological Statistics 5, 117-154.

Woodsford, P.A., 1996. Spatial database update - A key to effective automation. International Society for Photogrammetry and Remote Sensing 18th Congress. International Archives of Photogrammetry and Remote Sensing, 31(B4). RICS Books, Coventry, UK, pp. 955-961.

Worboys, M.F., 1998. Imprecision in finite resolution spatial data. GeoInformatica 2, 257-279.

Wright, R.L., 1996. An evaluation of soil variability over a single bedrock type in part of southeast Spain. Catena 27, 1-24.

Wright, R., Ray, S., Green, D.R., and Wood, M., 1998. Development of a GIS of the Moray Firth (Scotland, UK) and its application in environmental management (site selection for an 'artificial reef'). The Science of the Total Environment 223, 65-76.

Yao, T., and Journel, A.G., 1998. Automatic modeling of (cross) covariance tables using fast Fourier transform. Mathematical Geology 30, 589-615.

Zadeh, L.A., 1965. Fuzzy sets. Information and Control 8, 38-353.

Zadeh, L.A., 1968. Probability measures of fuzzy events. Journal of Mathematical Analysis and Applications 23, 421-427.

Zadeh, L.A., 1995. Probability theory and fuzzy logic are complementary rather than competitive, In: Discussion of: M. Laviolette, J.W. Seaman, J.D. Barrett, and W.H. Woodall, A probabilistic and statistical view of fuzzy methods. Technometrics 37, 271-276.

Zevenbergen, L.W., and Thorne, C.R., 1987. Quantitative analysis of land surface topography. Earth Surface Processes and Landforms 12, 47-56.

Zhu, H., and Journel, A.G., 1993. Formatting and integrating soft data: stochastic imaging via the Markov-Bayes algorithm. In: A. Soares (Editor), Geostatistics Tróia '92, Volume 1. Kluwer Academic Publishers, Dordrecht. pp. 1-12.

Zhu, A.X., Band, L.E., Dutton, B., and Nimlos, T.J., 1996. Automated soil inference under fuzzy logic. Ecological Modelling 90, 123-145.

Zhu, A.X., Band, L.E., Vertessy, R., and Dutton, B., 1997. Derivation of soil properties using a soil land inference model. Soil Science Society of America Journal 61, 523-533.

Abstract

The increasing popularity of geographical information systems (GIS) has at least three major implications for land resources survey. Firstly, GIS allows alternative and richer representation of spatial phenomena than is possible with the traditional paper map. Secondly, digital technology has improved the accessibility of ancillary data, such as digital elevation models and remotely sensed imagery, and the possibilities of incorporating these into target database production. Thirdly, owing to the greater distance between data producers and consumers there is a greater need for uncertainty analysis. However, partly due to disciplinary gaps, the introduction of GIS has not resulted in a thorough adjustment of traditional survey methods. Against this background, the overall objective of this study was to explore and demonstrate the utility of new concepts and tools within the context of pedological and agronomical land surveys. To this end, research was conducted on the interface between five fields of study: geographic information theory, land resource survey, remote sensing, statistics and fuzzy set theory. A demonstration site was chosen around the village of Alora in southern Spain.

Fuzzy set theory provides a formalism to deal with classes that are partly indistinct as a result of vague class intensions. Fuzzy sets are characterised by membership functions that assign real numbers from the interval [0, 1] to elements, thereby indicating the grade of membership in that set. When fuzzy membership functions are used to classify attribute data linked to geometrical elements, presence of spatial dependence among these elements ensures that they form spatially contiguous regions. These can be interpreted as objects with indeterminate boundaries or fuzzy objects. Fuzzy set theory thus adds to the conventional conceptual data models that assume either discrete spatial objects or continuous fields.

This thesis includes two case studies that demonstrate the use of the fuzzy set theory in the acquisition and querying of geographical information. The first study explored the use of fuzzy *c*-means clustering of attribute data derived from a digital elevation model to represent transition zones in a soil-landscape model. Validity evaluation of the resulting terrain descriptions was based on the coefficient of determination of regressing topsoil clay data on membership grades. Vaguely bounded regions were more closely related to the observed variation of clay content ($r^2_{corr} \approx 0.68$) than crisply bounded units as used in a conventional soil survey ($r^2_{corr} \approx 0.5$).

The second case study involved the use of the fuzzy set theory in querying uncertain geographical data. It explains differences between fuzziness and stochastic uncertainty on the basis of an example query concerning loss of forest and ease of access. Relationships between probabilities and fuzzy set memberships were established using a linguistic probability qualifier (high probability) and the expectation of a membership function defined on a stochastic travel time. Fuzzy query processing was compared with crisp processing. The fuzzy query response contained more information because, unlike the

crisp response, it indicated the degree to which individual locations matched the vague selection criteria.

In a land resource survey, data acquisition typically involves collecting a small sample of precisely measured primary data as well as a larger or even exhaustive sample of related secondary data. Soil surveyors often rely on soil-landscape relationships and image interpretation to enable efficient mapping of soil properties. Yet, they generally fail to communicate about the knowledge and methods employed in deriving map units and statements about their content.

In this thesis, a methodological framework is formulated and demonstrated that takes advantage of GIS to interactively formalise soil-landscape knowledge using stepwise image interpretation and inductive learning of soil-landscape relationships. It examines topology to record potential *part of* links between hierarchically nested terrain objects corresponding to distinct soil formation regimes. These relationships can be applied in similar areas to facilitate image interpretation by restricting possible lower level objects. GIS visualisation tools can be used to create images (e.g. perspective views) illustrating the landscape configuration of interpreted terrain objects. The framework is expected to support different methods for analysing and describing soil variation in relation to a terrain description, including those requiring alternative conceptual data models. In this thesis, though, it is only demonstrated with the discrete object model.

Satellite remote sensing has become an important tool in land cover mapping, providing an attractive supplement to relatively inefficient ground surveys. A common approach to extract land cover data from remotely sensed imagery is by probabilistic classification of multispectral data. Additional information can be incorporated into such classification, for example by translating it into Bayesian prior probabilities for each land cover type. This is particularly advantageous in the case of spectral overlap among target classes, i.e. when unequivocal class assignment based on spectral data alone is impossible.

This thesis demonstrates a procedure to iteratively estimate regional prior class probabilities pertaining to areas resulting from image stratification. This method thus allows the incorporation of additional information into the classification process without the requirement of known prior class probabilities. The demonstration project involved Landsat TM imagery from 1984 and 1995. Image stratification was based on a geological map of the study area. Overall classification accuracy improved from 76% to 90% (1984) and from 64% to 69% (1995) when employing iteratively estimated prior probabilities.

The fact that any landscape description is a model based on a limited sample of measured target attribute data implies that it is never completely certain. The presence of error or inaccuracy in the data contributes significantly to such uncertainty. Usually, the accuracy of land survey datasets is indicated using global indices (e.g. see above). Error modelling, on the other hand, allows an indication of the spatial distribution of possible map inaccuracies to be given. This study explored two approaches to error modelling, which are demonstrated within the context of land cover analysis using remotely sensed imagery.

The first approach involves the use of local class probabilities conditional to the pixels' spectral data. These probabilities are intermediate results of probabilistic image classification and indicate the magnitude and distribution of classification uncertainty. A case study demonstrated the implication of such uncertainty on change detection by

comparing independently classified images. A major shortcoming of this approach is that it implicitly assumes data in neighbouring pixels to be independent. Moreover, it does not make full use of available reference data as it ignores their spatial component. It does not consider data locations nor does it use spatial dependence models that can be derived from the reference data.

The assumption of independent pixels obviously impedes proper assessment of spatial uncertainty, such as joint uncertainty about the land cover class at several pixels taken together. Therefore, the second approach was based on geostatistical methods, which exploit spatial dependence rather than ignoring it. It is demonstrated how the above conditional probabilities can be updated by conditioning on sampled reference data at their locations. Stochastic simulation was used to generate a set of 500 equally probable maps, from which uncertainties regarding the spatial extent of contiguous olive orchards could be inferred.

Future challenges include studies on other quality aspects of land survey datasets. The present research was limited to uncertainty analysis, so that, for example, data precision and fitness for use were not addressed. Other potential extensions to this work concern full inclusion of the third spatial dimension and modelling of temporal aspects.

Samenvatting

De groeiende populariteit van geografische informatiesystemen (GIS) heeft tenminste drie belangrijke consequenties voor de landschapsmodellering. Ten eerste zijn er nieuwe gegevensmodellen en presentatievormen beschikbaar gekomen, waarmee de mogelijkheden van de traditionele papieren kaart, die vroeger zowel de database als het eindproduct van een kartering vormde, aanzienlijk worden uitbreid. Ten tweede kan bij de vervaardiging van geografische gegevensbestanden in verhoogde mate gebruik gemaakt worden van secundaire data, zoals remote sensing-beelden en digitale hoogtebestanden. Ten derde maakt de grotere afstand tussen de producenten en consumenten van geografische databases een grondige analyse van onzekerheden in de gegevens noodzakelijk. Mede door onbekendheid met de nieuwe begrippen en methodes heeft de introductie van GIS echter nog niet geleid tot een ingrijpende bijstelling van traditionele karteringsactiviteiten. Het algehele doel van deze studie was dan ook nieuwe concepten en gereedschappen op hun nut te onderzoeken en ze te demonstreren in de context van bodemkundige en agronomische landschapsanalyses. Hiertoe is onderzoek verricht op het raakvlak van een vijftal wetenschapsgebieden, te weten: geografische informatietheorie, bodemkundige en agronomische landinventarisatie, remote sensing, statistiek en de theorie van de vage verzamelingen (fuzzy set theory). Als proefgebied is een gebied rond het dorp Alora in zuid Spanje gebruikt.

De theorie van vage de verzamelingen biedt een formalisme om te kunnen werken met elkaar gedeeltelijk overlappende klassen die het gevolg zijn van onscherpe klassedefinities. De theorie kan van nut zijn in landschapsanalyses, omdat men daar vaak te maken heeft met niet scherp gedefinieerde begrippen. Vage verzamelingen (fuzzy sets) worden gekenmerkt door lidmaatschapsfuncties die elementen een reëel getal uit het $[0,1]$ interval toekennen en daarmee de mate van lidmaatschap in de betreffende verzameling aanduiden. Wanneer aan geometrische elementen gekoppelde attribuutwaarden worden geclassificeerd met behulp van vage lidmaatschapfuncties, vormen zich, ten gevolge van hun onderlinge ruimtelijke afhankelijkheid, samenhangende structuren, die geïnterpreteerd kunnen worden als vaag omgrensde ruimtelijke objecten. De theorie levert daarmee een aanvulling op de bestaande conceptuele gegevensmodellen die uitgaan van discrete ruimtelijke objecten of continue velden.

In dit proefschrift wordt het gebruik van fuzzy sets in de acquisitie en het bevragen van geografische gegevens aan de hand van twee voorbeeldstudies gedemonstreerd. De eerste studie behelst het gebruik van fuzzy c-means clustering van terrein-attribuutwaarden afgeleid uit een digitaal hoogtemodel, met als doel het beschrijven van overgangsgebieden in een bodem-landschapsmodel. De bruikbaarheid van de resulterende terreinbeschrijvingen werd bepaald met behulp van lineaire regressie van bemonsterde kleigehaltes van de bovengrond. Vaag omgrensde zones bleken de geobserveerde variatie in kleigehalte beter te kunnen verklaren ($r^2_{corr} \approx 0.68$) dan scherp

omlijnde eenheden, zoals gebruikt in de traditionele landschappelijke bodemkartering ($r^2_{corr} \approx 0.5$).

De tweede voorbeeldstudie betreft het gebruik van fuzzy sets in de bevraging van onzekere geografische gegevens. Aan de hand van een vraag omtrent het verdwijnen van bosachtige vegetatie en de bereikbaarheid van locaties werd het verschil tussen vaagheid (fuzziness) en stochastische onzekerheid aangetoond. Bij de beantwoording van de vraag werden beide typen onzekerheid gecombineerd, waarbij gebruik gemaakt werd van een vage waarschijnlijkheidsmaat en de verwachtingswaarde van een vage lidmaatschapsfunctie. Het resulterende antwoord werd vergeleken met een scherp (niet vaag) antwoord. Het vage antwoord omvatte meer informatie, aangezien het tevens de mate waarin voldaan wordt aan de vage selectiecriteria weergaf.

Bodemkundige en agronomische analyses baseren zich doorgaans op een kleine steekproef van gemeten doelgegevens, die wordt aangevuld met een grote of zelfs gebiedsdekkende verzameling secundaire gegevens. Zo wordt in de bodemkunde veelvuldig gebruik gemaakt van bodem-landschapsrelaties en beeldinterpretatie om een efficiënte kartering van bodemeigenschappen mogelijk te maken. De daarbij gebruikte expertkennis en methoden blijven echter grotendeels verborgen.

In dit proefschrift wordt een methodologisch raamwerk geformuleerd en gedemonstreerd waarmee, op interactieve wijze, bodem-landschappelijke kennis kan worden gemodelleerd in een GIS. Daarbij wordt gebruik gemaakt van stapsgewijze beeldinterpretatie door de expert en het inductief leren van bodem-landschapsrelaties. De methode is gedeeltelijk gebaseerd op de analyse van topologische relaties die kunnen duiden op *deel-van* (PART_OF) relaties tussen geneste terreinobjecten met verschillende bodemvormende regimes. De gevonden relaties kunnen toegepast worden in vergelijkbare gebieden. Daar kunnen ze de beeldinterpretatie vereenvoudigen door het aantal mogelijke objectklassen op lagere hiërarchische niveaus steeds in te perken. Tevens kan de landschappelijke configuratie van geinterpreteerde terreinobjecten gevisualiseerd worden, bijvoorbeeld met behulp van perspectivische afbeeldingen. Het ligt in de verwachting dat het raamwerk kan functioneren met diverse methodes voor de analyse en beschrijving van bodem-landschapsrelaties, in combinatie met verschillende conceptuele gegevensmodellen. In dit proefschrift wordt het echter alleen gedemonstreerd in een discrete objecten context.

Satelliet-remote sensing wordt veelvuldig ingezet in landbedekkingskarteringen, waar het een aantrekkelijke aanvulling vormt op de relatief inefficiënte veldopnamen. De gewenste informatie wordt meestal uit de satellietbeelden verkregen door middel van probabilistische classificatie van spectrale waarden uit de beeldelementen. Daarbij kan gebruik gemaakt worden van additionele gegevens, bijvoorbeeld als deze vertaald worden naar Bayesiaanse *a priori* kansen van de verschillende landbedekkingsklassen. Dit is vooral zinvol als de klassen spectrale overlap vertonen, waardoor eenduidige klassetoekenning gebaseerd op alleen spectrale informatie onmogelijk is.

In dit proefschrift wordt een procedure gedemonstreerd voor het iteratief schatten van regionale *a priori* kansen voor deelgebieden ontstaan door stratificatie van het beeld. De methode maakt het daarmee mogelijk additionele informatie in het classificatieproces op te nemen, zonder dat de te gebruiken *a priori* kansen vooraf bekend moeten zijn. In de casestudie werd gebruik gemaakt van Landsat TM beelden uit 1984 en 1995.

Beeldstratificatie was gebaseerd op een geologische kaart van het proefgebied. De iteratief geschatte *a priori* kansen verbeterden de globale nauwkeurigheid van de landbedekkingsclassificaties van 76% naar 90% (1984) en van 64% naar 69% (1995).

Het feit dat landschapsbeschrijvingen altijd abstracties van de werkelijkheid zijn, die bovendien gebaseerd zijn op een beperkte steekproef van gemeten doelgegevens, impliceert dat ze nooit volledig zeker zijn. De aanwezigheid van fouten of onnauwkeurigheden in de database draagt in belangrijke mate bij aan deze onzekerheid. Meestal wordt de nauwkeurigheid van een landschapsbeschrijving alleen aangeduid met een globale index (zie bijvoorbeeld hierboven). Om een beeld te krijgen van de ruimtelijke spreiding van onnauwkeurigheden moet men gebruik maken van foutenmodellering. In dit proefschrift is dit op twee manieren uitgewerkt binnen de context van remote sensing-ondersteunde landbedekkingsanalyse.

De eerste benadering betreft het gebruik van lokale voorwaardelijke kansen van landbedekkingsklassen, gegeven de spectrale waarden van de beeldelementen. Deze kansverdelingen zijn een tussenresultaat van het classificatieproces en verschaffen een ruimtelijk beeld van de lokale classificatieonzekerheden. Een casestudie liet zien dat deze onzekerheden grote consequenties hebben wanneer men temporele veranderingen in landbedekking wil opsporen door meerdere classificaties met elkaar te vergelijken. Een duidelijke tekortkoming van de methode is dat er impliciet vanuit gegaan wordt dat de gegevens uit naburige beeldelementen onafhankelijk van elkaar zijn. Bovendien wordt de ruimtelijke component van de referentiedata verwaarloosd.

De veronderstelling van onafhankelijke beeldelementen is duidelijk onhoudbaar wanneer men geïnteresseerd is in ruimtelijke onzekerheden. Hiermee wordt gedoeld op de gezamenlijke onzekerheid in een groep beeldelementen. De tweede benadering is daarom gebaseerd op geostatistische methoden, die juist gebruik maken van ruimtelijke afhankelijkheid, in plaats van deze te verwaarlozen. In een casestudie werd aangetoond hoe de eerdergenoemde voorwaardelijke kansen kunnen worden herzien, door deze afhankelijk te maken van nabijgelegen referentiedata. Met behulp van stochastische simulatie werd een reeks van 500 even waarschijnlijke landbedekkingskaarten gemaakt, waaruit onder andere onzekerheden betreffende de oppervlakte van aaneengesloten olijfboomgaarden konden worden afgeleid.

Uitdagingen voor de toekomst zijn studies naar andere kwaliteitsaspecten van landschapsgerelateerde datasets. Dit proefschrift beperkte zich tot onzekerheidsanalyses, waardoor bijvoorbeeld resolutie en geschiktheid voor gebruik niet aan bod kwamen. Andere potentiële uitbreidingen van het onderzoek betreffen het beschrijven van de derde ruimtelijke dimensie en vooral het modelleren van temporele aspecten.

Curriculum vitae

Sytze de Bruin was born on March 15, 1964 in Arnhem, The Netherlands. In 1983 he started his study at Wageningen Agricultural University (WAU). He graduated in 1989, majoring in soil science and land evaluation. The practical training (soil survey) and part of the thesis work (land evaluation and remote sensing) were done at the Atlantic Zone Programme in Costa Rica.

From 1989 till 1992 he worked in Costa Rica as a research assistant with the Atlantic Zone Programme and as a researcher at Palma Tica, which at that time was a subsidiary of Chiquita Brands. From 1992 till mid 1993 he was stationed at the National Agricultural University (UNA) in Managua, Nicaragua, where he worked as a soil expert within the framework of a UNA/WAU university co-operation project. Back in Wageningen, a subsequent study leave was used to learn the principles of geographical information systems. Thereafter, he was employed at the WAU Department of Soil Science and Geology to develop a course on GIS applications in land resource studies.

In April 1995 he moved to the then Department of Land Surveying and Remote Sensing where he started the research that resulted in this dissertation. In 1996 this research was halted for six months while he worked on a project surveying the present and potential significance of remote sensing for the Dutch Ministry of Agriculture, Nature Management and Fisheries.

His current research at Centre for Geo-Information of Wageningen University and Research Centre aims at operationalising the fitness-for-use component of data quality statements.