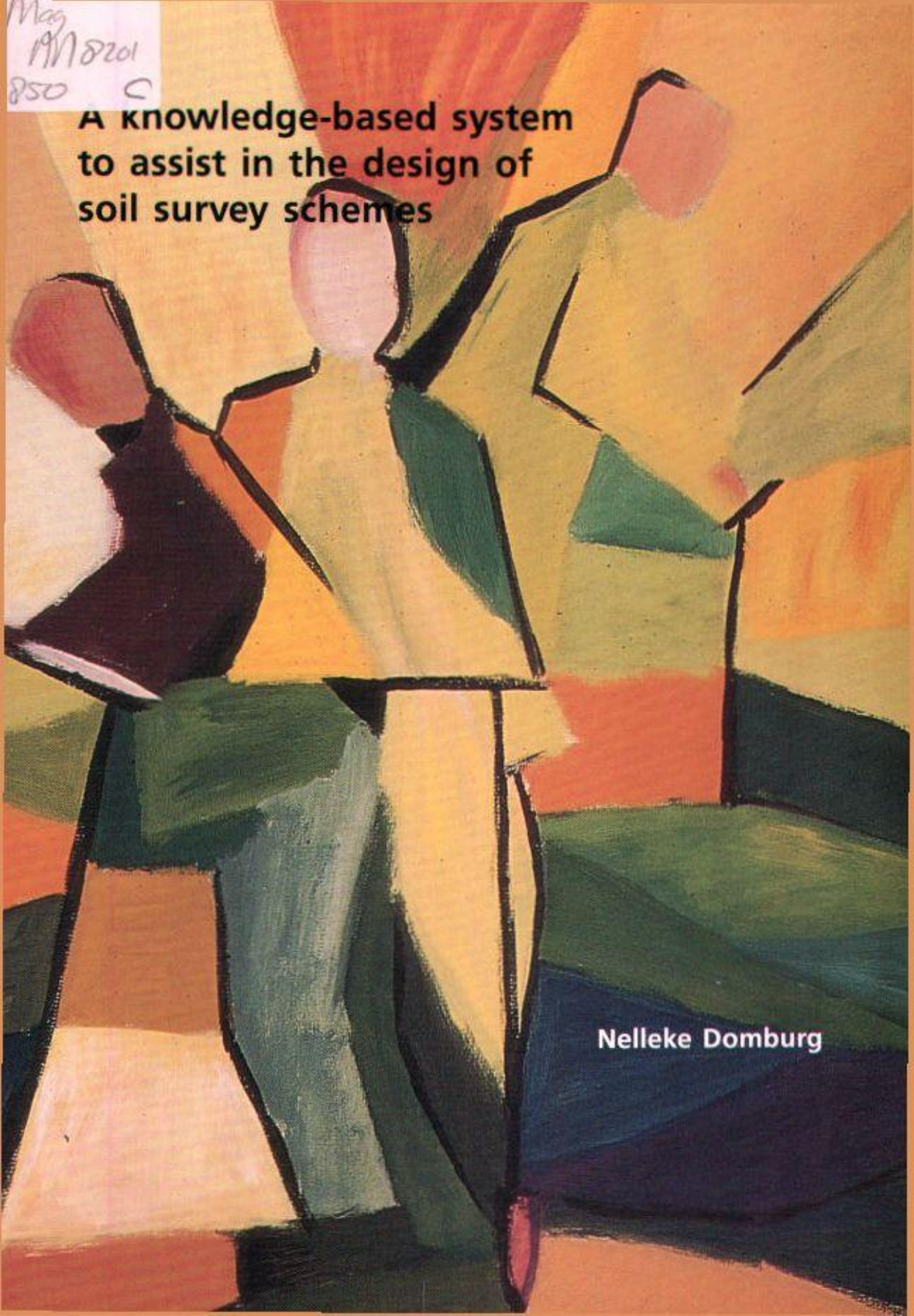
A knowledge-based system to assist in the design of soil survey schemes



Nelleke Domburg

1. Een efficiënt plan voor bodeminventarisatie is geen garantie voor de kwaliteit van bodeminventarisatie.

   Dit proefschrift.

2. De waarde van gekwantificeerde nauwkeurigheid van bodeminventarisaties wordt vaak onderschat.

3. De efficientie van het oriënterend en het nader onderzoek naar homogeen verdeelde verontreinigingen kan worden vergroot door bij het definiëren van de strata en bij de bepaling van het aantal bemonsteringspunten per stratum gebruik te maken van de informatie uit het vooronderzoek.

   Lamé, F.P.J. & Bosman, R. (1993)
   Protocol voor het oriënterend onderzoek: naar de aard en concentratie van verontreinigende stoffen en de plaats van voorkomen van bodemverontreiniging. Den Haag, SDU.
   Lamé, F.P.J. & Bosman, R. (1993)
   Protocol voor het nader onderzoek deel 1: naar de aard en concentratie van verontreinigende stoffen en de omvang van bodemverontreiniging. Den Haag, SDU.

4. Van de ontwikkeling van een kennissysteem wordt iemand wijzer.

5. AI steunt op gezond verstand.

   Globaal (1994), tijdschrift van Amnesty International.
   Hayes-Roth, F. , Waterman, D. & Lenat, D.B. (1983)
   Building Expert Systems, London, Addison-Wesley Publishing Company.

6. Door het drinken van Max Havelaar koffie wordt in elk geval tijdens de koffiepauze een bijdrage geleverd aan het oplossen van maatschappelijke problemen.

7. Door het toelaten van nieuwe kansspelen geeft de Nederlandse overheid te kennen de gokverslaving blijvend aan te willen pakken.

8. Als de veranderingsmanagers bij het ministerie van Landbouw, Natuurbeheer en Visserij geïsoleerd komen te staan is het tijd voor veranderingen.

   Waal, D. de (1994)
   Het is te eenzaam aan de top van LNV. Met Name Weekblad, Weekblad van het ministerie van Landbouw, Natuurbeheer en Visserij, 26.

9. De toename van het eco-toerisme maakt natuur minder natuurlijk.

   Boo, E. (1992)
   The Ecotourism Boom, Planning for Development and Management. In: Wildlands and Human Needs, A Program of World Wildlife Fund.

10. Als de ene tijd is verstreken breken er andere tijden aan.

11. Ontwikkeling van een beslissing-ondersteunend systeem vraagt om ondersteuning van veel beslissingen.

12. Multi-disciplinair onderzoek vereist extra discipline van de onderzoeker.

Stellingen bij het proefschrift 'A knowledge-based system to assist in the design of soil survey schemes' van P. Domburg. Wageningen, 28 oktober 1994.

A knowledge-based system to assist in the design of soil survey schemes

Promotor ir. M.S. Elzas
hoogleraar in de Informatica

Co-promotor dr. ir. J.J. de Gruijter
hoofd van de afdeling Landinventarisatiemethoden,
DLO-Staring Centrum, Instituut voor Onderzoek van het Landelijk Gebied
(SC-DLO)

P. Domburg

A knowledge-based system to assist in the design of soil survey schemes

Proefschrift
ter verkrijging van de graad van doctor
in de landbouw- en milieuwetenschappen
op gezag van de rector magnificus,
dr. C.M. Karssen,
in het openbaar te verdedigen
op vrijdag 28 oktober 1994
des namiddags te vier uur in de Aula
van de Landbouwuniversiteit te Wageningen

The research presented in this thesis was funded by The Netherlands Integrated Soil Research Programme, and performed at The DLO Winand Staring Centre for Integrated Land, Soil and Water Research (SC-DLO), P.O. Box 125, 6700 AC Wageningen, Netherlands.

CIP-DATA KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Domburg, Petronella
A knowledge-based system to assist in the design of soil survey schemes / Petronella Domburg. - [S.l. : s.n.]
Thesis Wageningen. - With index, ref. - With summary in Dutch.
ISBN 90-5485-312-3
Subject headings: knowledge-based systems / soil survey / sampling strategies.

Aan Hans,
Heit en Mem

Abstract

Domburg, P., 1994. **A knowledge-based system to assist in the design of soil survey schemes.** Doctoral thesis, Wageningen Agricultural University, Wageningen, Netherlands, (xiv) + 192 pp.

Soil survey information with quantified accuracy is relevant to decisions on land use and environmental problems. To obtain such information statistical strategies should be used for collecting and analysing data. A survey project based on a statistical sampling strategy requires a soil survey scheme specifying which sites are to be sampled, which data are to be recorded and how they are to be analysed statistically. The efficiency of such a scheme is determined by the accuracy of the survey results and the cost of operation. This accuracy and cost depend mainly on the method of determination and the sampling design in the scheme.

This study aimed at formulating the basic design considerations of a knowledge-based system (KBS) to assist in the design of soil survey schemes. This system should incorporate pedological and statistical knowledge. The domain of the system has provisionally been limited to surveys for which a design-based approach, i.e. the use of classical sampling theory, is appropriate.

Initially, the domain of the system has been structured in three layers: (i) an entity structure clarifying the position of the system in a soil survey project; (ii) a model describing the design process as a number of interrelated steps, and (iii) a conceptual framework defining the main concepts and their relations.

Further analysis made it possible to specify the tasks in which the KBS should assist: definition of the survey request, selection of prior information, design of outlinear schemes, evaluation and optimization of outlinear schemes, generation of a report, and evaluation *a posteriori*.

The system will primarily assist in the statistical decisions in the design process. Since there was no suitable classification of sampling designs available, a hierarchical framework of sampling designs has been constructed, in which sampling designs are grouped into types of designs, and types are grouped into classes of designs. Furthermore the main classes of sampling designs treated in the literature have been ordered in a taxonomy. Decision trees have been developed to guide the selection of an appropriate sampling approach (design-based versus model-based), and, in the case of a design-based approach, to guide the search for an appropriate class of sampling designs.

To ensure that the available means for a project, such as budget, personnel, and equipment, are used adequately schemes should be evaluated and optimized beforehand. Models related to the features of sampling designs have been developed for predicting the accuracy and cost of survey schemes, the so-called prior evaluation. Furthermore the use of dynamic programming is proposed to search for the optimal sampling design within an outlinear scheme. The procedure enables objective comparison of schemes taking into account differences in spatial variability or sampling cost among sub-regions.

Finally, basic design considerations are presented consisting of an initial requirements

definition, a description of the intended use of the KBS, and a specification of the components for an actual KBS. Five components are distinguished: a database, a knowledge base, a model base, a problem-solving model, and a user interface. The system will assist in its own maintenance through continuous storage of knowledge from executed projects. This will facilitate the re-use of information. A KBS which is based on these basic design considerations will assist in controlling the quality of soil survey projects.

Additional index words: artificial intelligence, dynamic programming, design-based approach, domain structuring, prediction of accuracy, prediction of cost

Woord vooraf

Er zijn veel mensen die op de een of andere manier hebben bijgedragen aan de totstandkoming van dit proefschrift en ik wil hen graag op deze plaats daarvoor bedanken.

In de eerste plaats natuurlijk Jaap de Gruijter, die het idee voor dit onderzoek heeft ontwikkeld en tijdens het project zowel de rol van statistisch adviseur als die van begeleider op zich heeft genomen. Dankzij hem heb ik inzicht gekregen in de mogelijkheden van het gebruik van statistiek bij bodemonderzoek en heb ik de kans gekregen mijn eigen weg te zoeken als onderzoeker.

Prof. Elzas wil ik bedanken voor het geregelde overleg en de positief kritische adviezen. Door zijn visie op het belang van de verschillende deelaspecten in dit onderzoek heb ik de kans gekregen me in verschillende disciplines te verdiepen zonder het geheel uit het oog te verliezen.

De leden van de begeleidingscommissie bedank ik voor de tijd die zij hebben vrijgemaakt om de voortgang van het onderzoek te bespreken: Prof. P. van Beek (Vakgroep Wiskunde, LUW), Prof. J. Bouma (Vakgroep Bodemkunde en Geologie, LUW), A. Breeuwsma (SC-DLO), A.K. Bregt (SC-DLO), Prof. H.J. van den Herik (Vakgroep Informatica, Rijksuniversiteit Limburg), en A.S. Stein (Vakgroep Bodemkunde en Geologie, LUW). Prof. Van Beek wil ik speciaal bedanken voor zijn begeleiding bij de deelstudie naar de mogelijkheid om plannen voor bodeminventarisatie te optimaliseren (hoofdstuk 7).

In de afgelopen vier jaar hebben ook enkele studenten van de LUW vanuit verschillende vakgebieden aan het project meegewerkt: Heleen Kamermans (afstudeervak Voorlichtingskunde), Ester Jaarsma (afstudeervak Bodemhygiëne en Bodemverontreiniging), Jan Willem van Groenigen (afstudeervak Informatica) en Sicco Steijaert (afstudeervak Operationele Analyse). Bedankt voor jullie enthousiasme en inzet.

Mijn collega's van het DLO-Staring Centrum wil ik bedanken voor de bereidwilligheid mij in te wijden in de praktijk van het bodemonderzoek, in het bijzonder: John Mulder, Joop ten Cate, Reind Visschers, Ben Marsman, Mirjam Hack-ten Broeke, Dick Brus, Martin Knotters en Wim te Riele.

Hans Vierbergen bedank ik voor het bijhouden van de administratieve rompslomp van een 'derde-geldstroom AIO op afstand'.

Tot slot wil ik Hans bedanken, die me heeft geholpen op z'n tijd het werk te relativeren. Bedankt voor het opfleuren, je (computer-)steun en het zelf-gebakken brood op zaterdag.

Nelleke Domburg
Renkum, juli 1994.

Contents

Chapter 1

Introduction

Parts of this chapter have been published in:

Domburg, P. & Gruijter, J.J. de (1992)
A framework of concepts for soil survey using probability sampling. Report 55. Wageningen, DLO Winand Staring Centre for Integrated Land, Soil and Water Research.

Domburg, P., Gruijter, J.J. de & Brus, D.J. (1994)
A structured approach to designing soil survey schemes with prediction of sampling error from variograms. Geoderma, 62(1-3), pp. 151-164.

1 Introduction

1.1 Soil survey

Soil may be characterized by many properties showing various degrees of spatial variation and correlations. Since for most soil properties it is impossible to continuously observe the whole land surface, soil survey usually aims at describing or mapping soil properties from sample data. Using the method of *free survey*, which is the conventional method to produce soil maps, surveyors divide the land into distinct types from observations of various landscape features (e.g. vegetation, land use, or elevation) using prior information (e.g. on geology, geomorphology, or hydrology) and then describe each soil type by sampling at some sites (Steur, 1961). The surveyor chooses the locations where augerings are made, and determines the delineations of the mapping units. This procedure is rather subjective and provides mainly qualitative information. The results usually contain only limited information on the variability of soil properties and on the accuracy of the results. Such information is not sufficient for all conceivable objectives.

There is a growing need for quantitative soil survey information of which the accuracy can also be quantified; researchers and those commissioning survey projects are not only interested in information on soil, but also in the accuracy of this information. Such information is relevant to decisions on the suitability of soils for different types of land use (e.g. for agricultural use), or on environmental issues (e.g. the production of drinking-water). The risks of inappropriate decisions being taken is influenced by the quality of the available information, which is frequently inadequate when dealing with new requests. As a result, a fresh field survey is often needed.

Information with quantified accuracy can be obtained by using statistical sampling strategies for collecting and analysing data. Before field work starts a *soil survey scheme* should be designed, specifying which sites are to be sampled, which data are to be recorded, and how they are to be analysed statistically. Apart from quantifying the accuracy of the final results, statistics enable the efficiency of possible sampling strategies to be quantified. The *efficiency* of a given sampling design p can be defined as the ratio of the sampling variance of a reference design (often simple random sampling) to the sampling variance of p, at same sample size or at same cost.

In this thesis three kinds of requests for soil survey using sampling strategies are distinguished. These are related to the types of results required (Fig. 1.1). This thesis does not consider variation in time, which would involve other requests and sampling strategies, but focuses on spatial variation.

Requests for 'how much'
First, there is a demand for studies concerning *how much* of a soil property is present, for example requests for estimating values of statistical parameters, such as mean, median, or areal proportion, for a given soil property. An example of this kind of soil survey is a study to determine the areal proportion of a region where the soil is saturated with phosphate. In

the case where a single property is of interest, the result of the study is a single value for the region as a whole indicating the areal proportion, accompanied by its quantified accuracy.

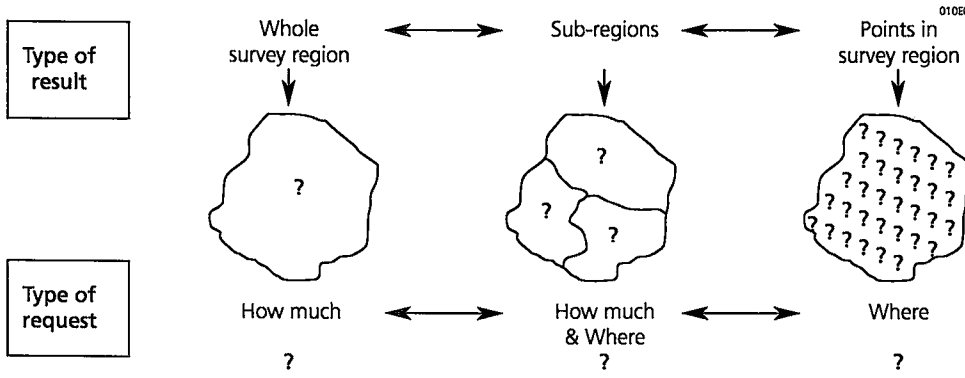


Figure 1.1 Relations between types of results and types of requests

Requests for 'where'
Second, there is a demand for soil surveys with the emphasis on *where* specific soil properties are present. Such studies usually result in maps, for example a map representing the spatial pattern of a soil property such as the organic matter content of the topsoil, or the moisture supply capacity. These maps give values of soil properties at individual points in the survey region. Of course, the answer to a *where* request implies the answer to a corresponding *how much* request, but the reverse is not true. Generally, answering a *where* request requires greater effort in data collection than answering a *how much* request.

Requests for 'how much & where'
Between these two extreme categories exists a third group of requests which is a combination of *how much* and *where*. One example is a study of the mean phosphate content of the topsoil in a region which incorporates three large land use units. If, besides a result for the whole region, accurate estimations of the phosphate content are also required for each of the land use units, both *how much* and *where* have implications for the choice of the sampling strategy.

1.2 Statistical approaches to soil surveying

In this thesis attention is focused on requests requiring a statistical approach to soil surveying. It should however be noted that there are also survey requests for which a statistical approach is not meaningful, e.g. surveys to demonstrate the presence of contaminated spots, or surveys for which the allowable number of sample points is very small. Under these conditions the approach to survey will be *purposive sampling*, i.e. sample points are deliberately selected, and the results are not analysed statistically.

Examining the use of sampling strategies for soil surveying, two approaches can be distinguished: the use of classical sampling theory (*design-based* approach) and the use of

geostatistical techniques (*model-based* approach) (Särndal, 1978; De Gruijter & Ter Braak, 1990). Classical sampling theory has been used in soil sampling for many years. During recent decades the use of geostatistical techniques has increased and knowledge concerning the usefulness of these techniques is expanding (e.g. Journel & Huijbregts, 1978; Webster & Oliver, 1990).

In the design-based approach the emphasis is on answering requests for *how much* is present, whereas the major strength of the model-based approach lies in determining *where* given soil properties are present. Figure 1.2 shows the emphasis of the statistical approaches on the types of results and types of requests.

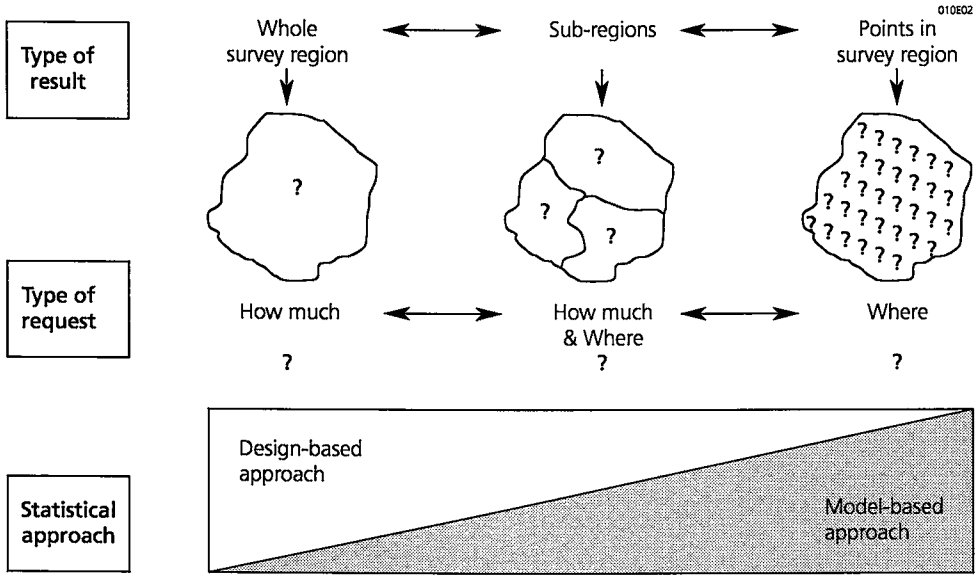


Figure 1.2 Emphasis of design-based and model-based approach on different types of survey requests

The relative importance of *how much* and of *where* influences the way in which data should best be collected and analysed, i.e. it suggests which sampling strategy seems most appropriate. However, it is impossible to divide all soil survey requests strictly into *how much* and *where* requests, or into requests requiring a design-based or a model-based approach. The distinction is more like a continuum with two extremes. In many cases the emphasis is focused on one side, and then one aspect mainly influences the choice of the sampling strategy.

1.2.1 Design-based approach

Although the literature on classical sampling theory focuses on the *how much* type of surveys, this approach is often also applicable to requests for both *how much & where*. In the design-based approach the concept of population is essential. The term *population* means the complete set of elements under study in a particular instance. In a soil survey project the population may consist of the complete set of possible sample points in the survey region.

Values of the soil properties at all locations are considered to be unknown but fixed, i.e. not random. A list of elements is selected from the population for sampling and the sampling design assigns a probability of selection to every subset. The sampling design also determines whether the sample points are mutually independent. Estimations of parameters are based on the design and possibly on auxiliary variables.

The design-based approach is also referred to as *probability sampling.* Cochran (1977: p. 9) characterizes probability sampling with four mathematical properties:

- it is possible to define a set of distinct samples, S_1, S_2... S_v, which the procedure is capable of selecting if applied to a specific population; this means that it is possible to indicate precisely which sampling elements belong to a particular sample;
- each possible sample S_l has assigned to it a known probability of selection π_l;
- one of the S_l s is selected by a random process in which each S_l receives its appropriate probability π_l of being selected;
- the method for computing the estimate from the sample is stated and leads to a unique estimate for any specific sample; it may be declared, for example, that the estimate is to be the average of the measurements on the individual elements in the sample.

In the literature on statistics, a survey using probability sampling is often referred to as a *sample survey* (e.g. Cochran, 1977: p. 2-4; Krishnaiah & Rao, 1988: p. 17, 47). Here, the term *sample* does not mean a single observation element taken in the field, which is also often referred to as a sample, but indicates the whole list or collection of (locations of) the elements to be observed. The term *sample survey* will also be used throughout this thesis. The use of statistical sampling to collect data for a survey is called *survey sampling* (e.g. Krishnaiah & Rao, 1988; Cassel et al., 1977). Prior information on the survey region and on the spatial variation in the region can be used for the design of sample surveys in soils.

1.2.2 Model-based approach

An important distinctive property of the model-based approach compared with the design-based approach is that the sampling elements to be observed need not be selected at random. In contrast, the elements are selected with a special purpose in mind, based on assumptions of the spatial dependence of the soil property in the survey region. The existence and modelling of spatial dependence in soils is the central theme of this approach: sites close to each other are more similar than sites further apart. In the model-based approach, data are therefore often collected at a fixed regular grid, while randomness and independence of sample points are the main characteristics of samples from the design-based approach.

The data collected with the model-based approach are normally used to predict the value of the soil property of interest at unvisited points. For this purpose many methods of spatial interpolation are available. For example, Van Kuilenburg et al. (1982), Burrough (1986), and Webster and Oliver (1990) have all reported on interpolation techniques as applied to soil survey. These techniques can, for example, be used to produce soil maps in cases where there is no obvious relation between soil types and landscape features.

Prior information on the spatial variation in the soil is required to design an optimal scheme for data collection using a model-based approach. The term optimal may concern the relation between the spacing and orientation of a grid, and the maximum prescribed sampling error

(McBratney et al., 1981), or a satisfactory compromise between a complete record of soil boundaries and sampling effort (Burgess & Webster, 1984). In the following the term 'optimal survey scheme' is related to efficiency with respect to accuracy of results and operational cost.

1.3 Computerized support

The conventional use of computer systems is for performing large calculations based on algorithms which lead to a correct solution. During recent decades other applications of computer systems stimulated by rapid developments in computer technology, are also being developed.

An interesting development is the growing use of computers for storing and providing information. For these objectives the development of database systems which enable systematic storage of large amounts of data and easy retrieval of specified selections started. At the same time, during the 1960s and 1970s, other researchers worked on the development of geographical information systems (GISs) (Burrough, 1986). GISs provide a powerful set of tools for storing, retrieving, transforming, and displaying spatial data from the real world for a growing number of purposes. In the field of soil survey these systems were initially used mainly to support mapping with a great deal of attention to cartographic accuracy and visual quality, and to support the spatial analysis of maps (e.g. Burrough, 1982; Rogoff, 1982; Burrough, 1986). In the last decade the interest in the use of GISs in combination with models of spatial processes has increased (e.g. Burrough, 1993). Steube and Johnston (1990) and Vieux (1991), for example, present studies on the linkage of GISs with hydrologic models. GISs can be used in different stages of modelling: development, testing, and application.

Besides developments in databases and GISs, researchers worked on systems providing advice and supporting decisions. From the 1950s onwards there was a growing interest in capturing human knowledge in computer systems, i.e. develop systems that possess *artificial intelligence* (AI). Waterman (1986: p. 388) defined AI as:

"... the subfield of computer science concerned with developing intelligent computer programs. This includes programs that can solve problems, learn from experience, understand language, interpret visual scenes, and, in general, behave in a way that would be considered intelligent if observed in a human."

Computer programs in which human expertise is captured are called *expert systems* (ESs); a more general class is formed by all those programs that operate on previously stored knowledge: *knowledge-based systems* (KBSs). The domain knowledge in an ES is separated from the procedural knowledge which determines how this factual knowledge is used. Advantages of these systems are that they make expert knowledge available to a large group of users and that they are able to explain to the user how a specific answer or a piece of advice has been derived. It may be a disadvantage that the domain of these systems is generally limited and that they can handle only problems that fall within their limited scope, whereas human experts are often capable of solving a number of different problems.

Another discipline dealing with supporting decisions is *operations research* (OR). According

to Hillier and Lieberman (1990: pp. 5-6):

"... operations research is concerned with optimal decision making in, and modeling of deterministic and probabilistic systems that originate from real life. These applications, which occur in government, business, engineering, economics, and the national and social sciences, are characterized largely by the need to allocate limited resources."

In the early years OR was mainly related to managers' decision-making activities, but nowadays it is applied to a wide variety of applications. Techniques from the field of OR are also used in automated support systems, which are generally referred to as *decision support systems* (DSSs) (e.g. Finlay, 1990). In these systems the knowledge used to solve a problem is represented as a mathematical model.

Although AI and OR were developed independently for many years there are similarities between these disciplines, and they may complement each other. There is a growing awareness that it may be profitable to consider the potential of both fields to solve problems, which may result in a choice for one of the two or for a combination of techniques (e.g. Simon, 1987; Finlay, 1990; Ignizio, 1990).

Various types of software packages have been developed to assist in statistical domains (e.g. Hand, 1984). Most of these packages are conventional programs for data analysis. However, there is a growing interest in the use of more advanced techniques (like AI techniques) to assist in the selection of statistical analysis techniques, or in the design of experiments or survey samples (Hand, 1984; Schach, 1986; Van den Berg, 1992). As far as known, no attempts have been made to develop a system to assist in the design of schemes for soil survey using statistical sampling strategies.

1.4 The practice of soil surveying using statistics

In the Netherlands large amounts of soil survey information have been collected and stored in soil maps, reports, databases, and GISs. This information is not always sufficient to answer a new survey request, but it should be utilized during the design of survey schemes. Knowledge about soil properties and soil survey can be referred to as *pedological* knowledge. Besides this pedological knowledge, statistical knowledge should be used to design soil survey schemes.

The development of a scheme for soil survey using probability sampling can be considered as a design process. At present this process takes place during one or more consultations between a researcher, or a research group, and a statistician who conducts the process by helping the researchers to make their aim more explicit and by trying to recover all relevant information. The statistician can be regarded as an expert in using sampling strategies in spatial sampling. The researcher may negotiate aspects of the survey scheme with people who commission a survey.

At present, this design process is often hampered by the following obstacles:

- no structured approach to designing survey schemes has been prescribed, which hampers the re-use of knowledge from historical projects, and quality control of surveys;

- existing information is not easily available, and must be gathered from different sources, e.g. maps, reports, databases and GISs. If available, soil databases and GISs may provide some computerized support for retrieving prior information;
- there is only limited information available on the variability of soil properties and on the accuracy of survey results;
- the number of possible schemes with respect to combinations of type of sampling design and method of determination, is virtually unlimited, whereas the time available to design a scheme is not;
- general procedures for evaluating the accuracy and cost of survey schemes are lacking.

Due to the last two points schemes are hardly ever compared. At present, schemes are often constructed *ad hoc* using the available knowledge and experience of those involved. This interferes with the aim of designing schemes in a reproducible, i.e. verifiable, way. If the way in which schemes are produced is known, this may provide insight into the quality of collected data, and the re-usability of information may increase.

1.5 Project aim and research questions

The aforementioned obstacles (Section 1.4) and the growing need for information with quantified accuracy have raised the question how the design of survey schemes can be supported by a computer system. Therefore a project has been launched which aims at:

developing a KBS in which pedological and statistical knowledge are integrated, to assist in the process of designing soil survey schemes.

The **aim** of this thesis is to identify basic design considerations, on which such a system needs to be based. To achieve this aim the following **research questions** should be answered.

1. *How can the design of soil survey schemes be structured?*
   A structured approach is required to develop a KBS. Before a system is built it will enable the verification of schemes and improve the comparability of surveys.
2. *What are the main decision problems during the design of soil survey schemes?*
   When the decision problems have been analysed the tasks to be supported can be specified.
3. *How can relevant knowledge and prior information be stored, selected and used to design schemes?*
   Pedological and statistical knowledge should be easily available in a computer system, enabling the selection of relevant knowledge in a limited period.
4. *How can schemes be evaluated in advance with respect to accuracy and cost?*
   Models are required to predict the accuracy and cost of survey schemes, i.e. models for prior evaluation, enabling objective comparison of possible schemes.
5. *Can an optimal soil survey scheme be found?*
   When appropriate methods of determination and appropriate sampling strategies are

selected the most efficient scheme should be searched.

6. *How should a system to assist the design of soil survey schemes be constructed?* Finally, the answers to the above questions should be integrated in basic design considerations of the whole KBS.

If these questions can be answered, the obstacles enumerated in the previous section will have been solved. Some parts of the system can be based on human expertise and on previously stored knowledge, e.g. from literature. For some other parts knowledge has to be generated.

The research questions require the contributions of different disciplines to the project. The minimum set of these disciplines is: soil science, statistics, computer science, and operations research. The former two provide knowledge about the design of soil survey schemes, the latter two about useful techniques to develop a system for computerized support of the design process, including the search for a (semi-)optimal scheme.

The system may be meaningful to different parties involved in soil surveys: both researchers and those commissioning projects may derive benefit from the system during the design of a scheme and during negotiations concerning the survey project. Furthermore, the system may be useful to statisticians involved in soil sampling.

1.6 Outline of the thesis

The next chapter deals with the scope of the project, including the domain of the system, knowledge acquisition and generating knowledge. Chapter 3 describes the theoretical background of the use of AI and OR in systems providing computerized support, and some developments in statistical support systems. This results in a rough structure of a system to assist in the design of soil survey schemes. Thereafter, Chapter 4 goes into the approach used for domain structuring and its results. Chapter 5 discusses the decision problems during the design of soil survey schemes, leading to a specification of the tasks to be supported. Chapter 6 focuses on the required knowledge about methods of determination and statistical knowledge. Then, the methods for prior evaluation and optimization of survey schemes are introduced and elaborated in Chapter 7. In Chapter 8 the results of the preceding chapters are integrated in the basic design considerations of the KBS. Finally, Chapter 9 presents the concluding remarks.

In this thesis quotations are in italics, between blank lines. Information on historical cases that is used to illustrate the text is also represented in italics between blank lines. Important concepts in the text are mostly in italics and can be retrieved using the subject index. The symbols in the equations are defined when they are used for the first time. Greek letters are used according to (geo-)statistical conventions. Besides, vectors are printed in bold italics, scalars in normal italics, and functions and symbols on the nature of quantities are normal upright characters. Sub- and superscripts are typographically treated as separate symbols.

The references, a glossary, a list of abbreviations, a list of symbols and a subject index have been added.

Chapter 2

Scope of the project

Parts of this chapter have been published in:

Domburg, P. & Elzas, M.S. (1994)
Structuring the Domain of a Complex System: a basis for a knowledge-based system supporting soil survey design. In: Beulens, A.J.M., Doležal, J. & Sebastian, H-J. (Eds.), Optimization-Based Computer-Aided Modelling and Design, Proceedings of the second Working Conference of the IFIP TC 7.6 Working Group, Dagstuhl, Germany, 1992. Leidschendam, Lansa Publishing, pp. 181-195.

Domburg, P. & Gruijter, J.J. de (1992)
A framework of concepts for soil survey using probability sampling. Report 55. Wageningen, DLO Winand Staring Centre for Integrated Land, Soil and Water Research.

2 Scope of the project

2.1 Background

Given the aim and constraints of a particular soil inventory study, different types of data, information and knowledge are needed to design a soil survey scheme. Distinctions can be made between *data*, *information*, and *knowledge* depending on the importance of the context. For data the context is unimportant, for information the context is of some importance, and for knowledge it is very important. For example, a table representing the results of a chemical analysis contains data. A soil map contains pedological information: the data collected in the field are described and classified. If a soil map is available during the design of a soil survey scheme, pedological knowledge about the way to interpret and to use that map is required. In general, data and information are explicit; knowledge must also be made explicit to be used in a computer program. This study concentrates on the use of knowledge and information, as a restricted type of knowledge, in a knowledge-based system; from now on the term knowledge will be used as a concept comprising both.

This chapter specifies the scope of the system. Therefore, first the domain of the support system is delimited (Section 2.2), i.e. the set of survey requests for which assistance will be provided during the design of survey schemes is specified. Then the sources of knowledge and the methods of knowledge acquisition are described (Section 2.3). Finally, attention is paid to the objective of supporting some tasks which cannot be performed at present, and for which knowledge has to be generated (Section 2.4).

2.2 Domain

To successfully develop a knowledge-based system the domain or scope of the system needs to be delimited (e.g. Waterman, 1986). This implies that there should be a limited area of problems for which the system can provide support, but also that a large number of problems cannot be solved using the system. A limited domain is vital during system development; however, it may be extended later. Two other requirements placed on the domain are: the knowledge should be reliable, and the knowledge should be static (Stefik et al., 1983b). Thus, consensus about the domain knowledge is very important.

At an early stage of this study three decisions on the limitation of the domain have been made with respect to: the statistical approach, the type of sampling, and the dimensions of the requests to be assisted. The following sub-sections deal with these decisions.

2.2.1 Classical sampling theory

The system will at first support the use of classical sampling theory (design-based approach), although the desirability to include geostatistics (model-based approach) at a later stage will be taken into account. This implies that the system focuses on supporting *how much* and *how much & where* requests (see Chapter 1). There are several reasons for this decision.

Initially, an exploratory soil survey often starts with a *how much* request and the *where*

request appears at a later stage. This is, for example, the prescribed approach to soil pollution surveys in the Netherlands. At an early stage it should be investigated whether a region is to be considered as being polluted (*how much* request). If so, at the following stages the 'exact' location of polluted soil should be determined (*where* request), whereafter soil sanitation may be able to start.

Secondly, the classical sampling approach has been used in many fields for many decades, and there is broad consensus on the applicability and characteristics of different types of sampling designs. In the practice of soil surveying, this approach has been used for many years. During recent decades the development of geostatistical techniques has increased and knowledge on the applicability of these techniques to soil surveying is expanding (e.g. Journel & Huijbregts, 1978; Webster & Oliver, 1990; Stein, 1991). Different geostatistical methods are being developed but so far there is little consensus on which method suits best in a given situation. Englund (1990) has investigated the differences between geostatisticians in their approaches to analysis and interpretation of data. The variation between geostatisticians turned out to be considerable. A model-based approach requires assumptions on the spatial dependence in a survey region which leads to a subjective choice of a sampling strategy. Since there is much more consensus in classical sampling theory on the applicability of strategies, this statistical knowledge is more suitable for use in a knowledge-based system.

2.2.2 Point sampling in the plane

The most common type of soil sampling can be referred to as *point sampling in the plane.* It is frequently the case that the sampling elements are augerings or profile pits, which can be regarded as points in a plane. The fact that this study is confined to point sampling in the plane does not imply that variation with depth is ignored. The latter is in fact often accounted for in the definition of the soil property of interest, i.e. the *target variable*, e.g. the average phosphate content to a given depth, or the depth to a given soil layer.

2.2.3 Single criterion requests

Requests for spatial inventories of soils may relate to one or more (single or multiple) target quantities, like a mean value, or an areal proportion, and to one or more target variables, like the mean highest groundwater level, or the cadmium content. The way in which these quantities and soil properties vary in space differs. However, a scheme which is efficient for one target quantity and soil property is not necessarily efficient for other quantities and properties. Nevertheless, if a survey aims at more than one soil property, these properties may be spatially related.

At present questions with multiple criteria (i.e. target quantities and/or target variables) are reduced to single criterion problems by determining which quantity and which variable is most important, and developing a scheme for this restricted request. There is knowledge on and experience of developing survey schemes for this type of problem. As far as known, there is no knowledge about multiple criteria problems.

In this study the emphasis is on single criterion surveys and there will be no investigation of multiple criteria problems. This restriction also implies that no attention will be paid to inventories for monitoring temporal changes, nor to inventories aiming at the determination of regression parameters.

There is no special restriction on the nature of the soil properties examined; they may be chemical or physical, estimated or measured, cheap or expensive, susceptible to inaccurate or accurate determination. These differences have no consequences for the way survey schemes should be developed. The same applies to the scale at which the properties are measured: nominal, ordinal, interval, or ratio. Only, in the first two cases, the target quantities are confined to proportions. In the last two cases target quantities may also be, for example, means, quantiles, or tolerance intervals. The system to be developed does not impose restrictions on the size of the survey area, which may vary, nor on its shape, which may be contiguous or non-contiguous.

2.3 Knowledge acquisition

The term *knowledge acquisition* as used in the specific field of computer science known as knowledge engineering refers to the process of extracting, structuring, and organizing knowledge from different sources, usually including human experts, so that it can be used in a computer program (Waterman, 1986). This section discusses the sources and methods used for knowledge acquisition.

2.3.1 Sources of knowledge

At present, the accessibility of knowledge needed to design survey schemes is restricted because it is scattered, and an overview of existing knowledge is lacking. The relevant sources are discussed in this section.

Human experts

It is often profitable to make use of knowledge based on practical experience, e.g. of surveyors, or statisticians. Therefore it must be clear who are the experts, and it must be possible to consult them.

Maps

In the Netherlands a multi-purpose soil map is available of the whole country, scale 1:50 000. This map displays information on a large number of soil properties, e.g. the distinction between different layers, the clay content and organic matter content of the layers, and the depth to the groundwater. Soil maps on larger scales are only available for parts of the country, since they are produced for separate surveys.

Besides multi-purpose maps there are single purpose maps, e.g. a map displaying only information on groundwater classes, and interpretive maps. The latter are deduced from readily available data on soil properties using so-called pedo-transfer functions. These functions relate basic soil data (e.g. on texture, organic matter content, or oxalate-extractable iron) to derived soil properties (e.g. the fertility, or the phosphate sorption capacity) (e.g. Breeuwsma et al., 1986). Interpretive maps are often the results of physical land evaluation.

Maps are generally accompanied by legends and reports which provide additional information. Sometimes there may be indistinctness concerning the interpretation of maps, as the information provided is mainly qualitative. Part of the data collected for producing maps is stored in databases nowadays, and many maps are stored in a GIS database.

Literature

Theoretical knowledge documented in books or articles needs to be studied before it can be used in practice. A problem is that this knowledge is often not easily accessible for novice users. If time is lacking, only part of the knowledge can be used. Once a strategy has proven to be usable it is re-used repeatedly without considering other strategies. It is often difficult to compare new developments little tested in the field with those that have already proved their value in practice. This hampers the application of more sophisticated methods in practice.

Domains of knowledge

Two domains of knowledge that are relevant to the design of soil survey schemes are: pedological knowledge, i.e. knowledge about soil properties and soil surveying, and statistical knowledge. Knowledge within these domains can be subdivided. Table 2.1 displays the nature and sources of knowledge needed to design soil survey schemes.

Table 2.1 Nature and sources of knowledge to design soil survey schemes

010E11

Nature of knowledge	Source of knowledge
Pedological knowledge	
- general pedological knowledge	- literature, (soil) maps, reports, databases, GISs
- knowledge on *spatial variability* (1)	- profile descriptions (reports, databases), researchers, soil surveyors
- knowledge on methods of determination	- laboratory handbooks, researchers soil chemists/physicists
- knowledge on *logistical aspects of field work* (2)	- researchers, soil surveyors
Statistical knowledge	
- general statistical knowledge	- literature
- knowledge on application of statistics in soil survey	- literature, researchers, statistical consultants active in soil survey

(1) knowledge on variation of soil properties in space

(2) e.g. some soil properties cannot be measured in winter time, or the suitability of a method of determination is sometimes related to certain soil types.

Besides the types of knowledge shown in Table 2.1 knowledge on how to construct schemes (i.e. knowledge of the design process) utilizing pedological and statistical knowledge, is important. As far as known, the knowledge about designing soil survey schemes has not been formalized until now, i.e. this design process has not been described. To gain insight into the whole problem domain, a number of knowledge-acquisition methods have been used.

2.3.2 Acquisition methods

In this project, knowledge is mainly collected by studying the literature, by interviewing a statistical consultant experienced in soil surveying, and by examining historical cases. *Historical cases* are sample surveys in soil that were executed before the development of the KBS started.

Studying literature

General knowledge of classical sampling theory and definitions of statistical terms is derived from the books of Cassel et al. (1977), Cochran (1977), and Krishnaiah & Rao (1988). There are slight differences in terminology between these handbooks; sometimes no crisp definition is given or a particular specification is indicated as 'recommended use'. Sometimes a choice has been made between alternative definitions or a definition has been adjusted to the domain considered here. This literature also provides knowledge on the characteristics of sampling designs, from which rules may be derived to support the choice of applicable designs. Literature on the application of statistical methods in soil survey contains descriptions of statistical terminology and examples of their meaning in soil survey (e.g. Webster & Oliver, 1990), but does not provide a detailed approach to constructing soil survey schemes on which the design of the proposed system can be based.

Interviewing

In order to describe the process of designing schemes and to discover relevant concepts, a statistician with experience in designing soil survey schemes (an *expert*) was interviewed in about ten sessions. During these sessions specific questions were asked about the process of designing a soil survey scheme in order to identify key concepts needed to find a solution for any given request, and to identify the steps in the design process. This process, which starts with a request for a soil survey and ends with a soil survey scheme, was fully considered.

The full texts of the interviews were discussed with the expert and adapted where necessary. The knowledge obtained from the interviews was structured and organized and the results were discussed with the expert. Through this interaction the knowledge could be formalized.

The first sessions started with exploratory interviews and the questions gradually became more selective. At later stages there were still regular discussions with the statistician. There have also been discussions on historical cases with other people involved in these surveys.

A technique for structuring knowledge frequently used in knowledge engineering is *card-sorting*. Related concepts are written on cards and the expert is asked to sort the cards and explain the reasons for a specific order; different sequences may be possible. Card-sorting has been used to structure statistical knowledge.

Examining historical cases

After the exploratory interviews, an overview was made of 23 historical cases of sample surveys in soil carried out at the DLO Winand Staring Centre. The descriptions were used to check and adjust the structures derived from the literature and the interviews, and to evaluate their applicability in practice. Furthermore, these cases were used to analyse decision problems arising during the design process and to describe the relevant domain knowledge

acquired from pedology and statistics.

2.4 Generating knowledge

As stated earlier (Section 1.5) the system proposed should be able to perform some tasks which at the moment are not performed in the practice of soil surveying: prior evaluation of the accuracy and of the cost, and optimizing schemes. If models are available to predict accuracy and cost, schemes can be compared objectively, and it will probably become possible to search for the most efficient scheme. To achieve these objectives knowledge needs to be generated. In this study, models have been developed in which the consequences of different sampling strategies for accuracy and cost are made explicit. Furthermore the possibilities for using these models combined with techniques from the field of OR to support the search for an efficient scheme have been investigated. So, in this study generating knowledge does not refer to the development of new techniques, but rather to find out how existing techniques can be used for new tasks.

Chapter 3

Computerized support: approaches and applicability

3 Computerized support: approaches and applicability

3.1 Outline

In the first chapter some attention was paid to computerized support (Section 1.3). This chapter discusses in more detail the use of AI and OR techniques in computer systems and the development of statistical support systems. AI and OR are two distinct fields in which techniques are developed to support decision making. Both have their own specific potentials. Sections 3.2 and 3.3 give brief reviews of the developments in AI and OR in the past years. Then, Section 3.4 discusses the combination of techniques from both fields to solve problems. Section 3.5 focuses on the development of statistical support systems: the domains and techniques used in these systems are considered. Thereafter, Section 3.6 deals with the advantages and disadvantages of the techniques discussed for use in a system to assist in designing soil survey schemes. Finally, Section 3.7 describes the rough structure of the proposed system.

3.2 Artificial Intelligence

AI is a very wide research area within the field of computer science. It is difficult to provide a definition of such a complex area which can be accepted by everyone. Barr and Feigenbaum (1981: p.3) give the following description, which corresponds with Waterman's definition quoted in Section 1.3:

"Artificial Intelligence (AI) is the part of computer science concerned with designing intelligent computer systems, that is, systems that exhibit the characteristics we associate with intelligence in human behavior - understanding language, learning, reasoning, solving problems, and so on."

Rich (1983: p. 1) gives a more terse definition outlining what AI constitutes:

"Artificial intelligence (AI) is the study of how to make computers do things at which, at the moment, people are better."

These definitions pay no attention to how problems should be solved. Van den Herik (1988) states that in the field of AI computer programs should be developed that do things which humans can do, and that humans must be able to understand and perform the tasks in the way the computer does. This is a more extensive objective than that of Rich. According to Van den Herik (1988):

AI is the study of how to improve the understanding of human thinking.

3.2.1 History

In the early years of AI research researchers aimed at developing universal mechanisms for problem solving. Dreyfus & Dreyfus (1988) describe two opposite visions of what computers could be. These visions developed from the early 1950s.

One approach considered computers as symbol manipulators that require a symbolic representation of the world. Newell and Simon were two leading adherents of this vision. They concluded that everything, even numbers, could be represented as symbols (Newell, 1983). These symbols could be manipulated by means of formal rules thus generating computer behaviour that resembled in part the behaviour of an intelligent problem solver. This approach started from a perspective of automated problem solving; the first attempts concentrated on solving mathematical games and executing a task of medical diagnosis.

The second approach sought to create artificial intelligence by modelling the brain. Computers should simulate how a network of neurons (*neural network*) learns to distinguish patterns and respond correctly. This approach started from the idea of machine learning: a neural network program, i.e. a program modelled on the human brain, can be trained with examples and so learn, for example, to distinguish between certain types of patterns like the patterns of speech or images. Rosenblatt (1958) developed such a network, which he called a perceptron, and trained it to classify sets of patterns as similar or distinct.

Both approaches aimed at developing a system that could handle all kinds of problems: universality of mechanisms. However, AI demonstrations of both approaches solve specific tasks in clear-cut parts of the world: micro-worlds (Papert, 1988). When it turned out that systems could be developed that yielded reasonable solutions in a micro-world, it was concluded that a universal problem solving mechanism might be too ambitious in the short term. From the 1960s and 1970s part of the AI researchers concentrated on building tools and practical expert aids, while others continued working on more abstract, universal mechanisms (Elzas, 1986). The growing attention to practical systems, which may seem less ambitious, resulted in the development of valuable KBSs and ESs in a growing number of domains, e.g. finance, management, failure analysis, and scheduling (Hayes-Roth & Jacobstein, 1994).

3.2.2 Expert systems

Waterman (1986: p. 11) has defined an *expert system* (ES) as:

"A computer program using expert knowledge to attain high levels of performance in a narrow problem area."

Earlier Waterman and Hayes-Roth (1983: p. 169) stated that:

"An expert system is a computer program that embodies the expertise of one or more experts in some domain and applies this knowledge to make useful inferences for the user of the system."

This second definition stresses an important characteristic of ES: namely, that human expertise should be involved. The system should be skilful at applying this knowledge and it should possess robustness, i.e. having both specific and general knowledge on the domain.

Another characteristic is that an ES uses symbolic reasoning, i.e. problem solving based on manipulating symbols that stand for domain concepts. So, knowledge should be represented symbolically. A third characteristic of an ES is that it has depth. This means that the system operates in a narrow, but complex, domain which implies that the rules in the system are also complicated. Finally, an ES should possess self-knowledge, which refers in the first instance to knowledge on its own reasoning process (*meta-knowledge*). Secondly, an ES should be able to explain how it arrives at its solutions, and therefore possess an *explanation facility*.

The four characteristics mentioned above (expertise, symbolic reasoning, depth, and self-knowledge) distinguish an ES from a conventional program (Waterman, 1986).

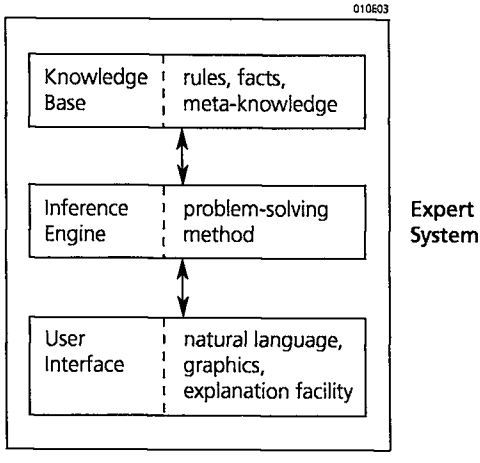


Figure 3.1 The structure of an expert system

Figure 3.1 shows the structure of an ES. An important feature of the structure of an ES is that knowledge about the problem domain is separated from the other knowledge in the system. This domain knowledge is stored in the *knowledge base* as facts and rules or as a hierarchy of frames, which reflect factual relations and are thus applicable to the domain. Rules are often in the form: IF <condition> THEN <conclusion>. A frame is a data structure in which a number of facts or attributes about an object can be stored, in so-called *slots*. Besides slots, frames can possess *demons*, procedural mechanisms that activate some process if a certain slot-filling condition becomes true, e.g. to compute a new value using the values of other slots. Frames are organized in a hierarchical structure, therefore each frame refers to its parent (preceding frame). A frame may for example be defined as follows:

FRAME: T
PARENT: Thing
SLOT 1: Value 1
SLOT 2: Value 2
SLOT 3: Method 1, Value 3
IF ADDED: perform computation New Value, Method 1 (Value 1, Value 2)

Besides knowledge represented as rules or frames, meta-knowledge that guides and bounds the possible activation of rules is also stored in the knowledge base. The problems ESs deal with are often difficult and poorly understood and can generally not be solved with practical, exact-solution algorithms. Instead *heuristic rules*, which are based on human experience, are often used to support the effective search for solutions. Heuristics are as rules of thumb or short cuts which produce an acceptable solution in most cases within an acceptable period of time (see e.g. Barr & Feigenbaum, 1981: pp. 28-30; Korf, 1987). Heuristics can, for example, be used if the available factual knowledge is insufficient or if an exact approach would be too time-consuming.

The *inference engine* contains generic problem solving knowledge. Barr and Feigenbaum (1982: p. 189) define the inference engine, or the systems reasoning process, as:

"The mechanism used to draw conclusions based on the rules in the knowledge base and the data for the current case ..."

The inference engine finds the rules that are satisfied by the information available, then selects rules that will actually be executed, and executes these rules (Brownston et al., 1985). This part of the system, which uses meta-knowledge or heuristics, controls the search for solutions.

The *user interface* guides communication with users; it determines how the system is perceived. The language of the dialogue is generally close to natural language. The system should be easy to handle, for which graphical representations are often used. A distinguishing feature of the user interface of an ES is the explanation facility, which is crucial to acceptance of the system (Swartout, 1987). Explanations may be based on the executed rules, which should therefore be translated into natural language (e.g. SINCE <condition> THEREFORE <conclusion>).

3.2.3 Knowledge engineering

Knowledge Engineering (KE) is the process of building ESs. According to Hayes-Roth (1987: p. 293), knowledge engineers are concerned with extracting knowledge from human experts and integrating it in an overall knowledge system architecture. He uses the term 'knowledge system' as shorthand for 'knowledge-based expert system'. Knowledge engineers convert knowledge into applicable forms. Hayes-Roth distinguishes four KE activities: knowledge acquisition, knowledge system design, knowledge programming, and knowledge refinement.

Knowledge acquisition

Knowledge acquisition (KA) refers to the process of eliciting knowledge from experts and other sources (like literature) to recover the basic concepts of the domain. The relevant KA methods and knowledge sources in this project have already been introduced in the preceding chapter (Section 2.3).

Knowledge system design

The second activity, knowledge system design, concentrates on producing a framework for the system and on selecting an appropriate scheme for knowledge representation. So, in fact knowledge system design consists of two sub-activities: knowledge structuring and knowledge

representation.

Firstly, knowledge structuring aims at distinguishing the relations in the knowledge domain and constructing a model of the problem solving strategy. The KADS methodology, which aims at providing a framework for KE activities (Breuker & Wielinga, 1989), focuses on knowledge acquisition and modelling. Initially, the acronym KADS stood for Knowledge Analysis and Documentation System, later on other interpretations have been given such as Knowledge Analysis and Design Support. An essential element of the KADS methodology is to develop at an early stage a model that represents the inference structure (a so-called *interpretation model*). Knowledge acquisition and modelling are driven by this model. However, construction of an adequate model at an early stage may be difficult. Chandrasekaran et al. (1992) describe an other approach to knowledge modelling: task-structure analysis. A task stands for a type of problem-solving goal, e.g. diagnosis or design, and is related to the types of knowledge needed to accomplish it: the methods. The task structure can be represented as a tree of tasks, methods and sub-tasks that should be applied recursively. For various types of tasks structures can be evolved based on historical analysis. Then they can facilitate knowledge modelling of the same type of task in an other domain (Chandrasekaran et al., 1992). The development of general task structures that are flexible enough to be useful in various domains may take quite some time.

Secondly, knowledge representation includes different options e.g. formal logic, semantic networks, frames, or rules (see e.g. Waterman, 1986; Shapiro, 1987). A representation should be suitable for storing the available knowledge and facilitating the search and inference required for problem solving.

Knowledge programming

Knowledge programming deals with transforming human knowledge into a knowledge base with a corresponding inference engine. It is hard to characterize an inference engine in a general way. The structure of the inference engine depends on the nature of the domain of interest and on the way in which knowledge is represented and organized in the system. Two inference methods which may affect the structure of the inference are forward reasoning and backward reasoning. *Forward reasoning* starts from the facts in a given situation to establish new facts, and finally reaching a conclusion. *Backward reasoning* starts with what has to be proved, and the system tries to find and evaluate the facts needed for the proof. Systems do not always deal with exact knowledge; there may be uncertainty in the facts or rules a system has to deal with. A variety of approaches to reasoning with uncertainty are being developed in AI. One approach is to add *certainty factors* to facts, rules, or conclusions. These are numbers that measure the certainty or confidence in the validity of a fact, rule, or conclusion. In the 'Encyclopedia of Artificial Intelligence' (Shapiro, 1987) a distinction is made between numerical and symbolic approaches, and a number of selected approaches are illustrated and discussed, e.g. Bayes' Rule, Confirmation Theory (using certainty factors), Necessity and Probability Theory (dealing with fuzzy values), and the Theory of Endorsement (Bonissone, 1987). It is beyond the scope of this chapter to deal with these approaches in detail.

Knowledge refinement

The last engineering activity according to Hayes-Roth (1987) is knowledge refinement, which

continues until the system has achieved an adequate level of performance.

Knowledge maintenance
Another engineering activity may be added to the four described above: knowledge maintenance or knowledge management. When an ES operates it needs maintenance, just like conventional programs. In the course of time it may be necessary to adjust the knowledge base or to add new knowledge. Maintenance of the knowledge of a human expert depends on an ability to learn. Machine learning is therefore an important topic in AI, which tries to imitate human intelligence. Carbonell and Langley (1987) give an overview of the main lines of research on this topic. In spite of all the efforts made, the number of operational machine learning systems is small. However, the knowledge base must be kept up to date. The knowledge engineer may be responsible for this maintenance if the system created does not have a learning ability itself.

3.2.4 Knowledge-based systems and expert database systems

An ES cannot be developed for every problem domain. It is sometimes advantageous to combine expert-system techniques with other approaches or techniques. Two results of such developments are considered here: knowledge-based systems and expert database systems. The techniques used for developing these systems do correspond roughly with those for ESs.

Knowledge-based systems
An ES should provide the user with clear conclusions and the system should take decisions independently of the user at any one time. There are many problem domains for which it is impossible to develop an ES that can derive conclusions autonomously. For practical reasons it may be desirable that some decisions in the problem solving process are taken by the user, or, if there are a number of possibilities at a certain moment, the user may be asked to indicate an order of importance for that particular case. This ranking may be so situation specific that general criteria cannot be derived easily. Development of an ES may also be hampered by the lack of a single expert who has a thorough command of the whole domain. Then knowledge has to be derived from different sources, possibly including experts on parts of the domain.

Attempts to develop computer systems to assist in these domains, using expert-system-like techniques have lead to the development of KBSs. KBSs are programs that operate on previously stored knowledge. According to Waterman (1986) the main characteristic of a KBS is that the domain knowledge is separated from the systems' other knowledge. These systems constitute a more general class of systems than ESs; the main components of a KBS correspond with those of an ES. A KBS requires the user to provide more input than just facts; the user also has to take decisions. The answers a KBS provides should be considered as suggestions.

Expert database systems
Another class of systems combines the advantages of ESs with these of database management systems (DBMSs): *expert database systems* (EDSs). Smith (1986: p. 5) defines EDS as:

"... a system for developing applications requiring knowledge-directed processing of shared information."

An ES has the advantage that the methods of knowledge representation facilitate operating in complex and ill-structured domains. An additional advantage is the explanation facility.

DBMSs aim to develop applications when multiple users require access to the same (often large) collection of information. The DBMSs protect such information by providing consistency control, recovery control, concurrency control, and security control. Other advantages of DBMSs are that they can control data redundancy and distribution, and that they facilitate the development and maintenance of application programs.

To characterize systems in which the advantages of ESs and DBMSs are combined the term 'expert database system' has been introduced (Kerschberg, 1986).

3.3 Operations research

3.3.1 History

OR was established during the Second World War, and at that time the main field of application was military. Just after the war many developments in OR were related to decisions of a management nature. In the 1950s OR rapidly spread to a variety of applications, e.g. in administration and production, in business and industry, and nowadays it is applied to a large variety of applications. However, the term OR is still often substituted for or associated with management science (MS) (e.g. Wagner, 1975; Keen & Scott Morton, 1978). Wagner (1975: p. 2) describes OR as:

"... a scientific approach to problem-solving for executive management."

A general description of OR (Hillier and Lieberman, 1990: pp. 5-6), which was also quoted in Section 1.3, is:

"... operations research is concerned with optimal decision making in, and modeling of, deterministic and probabilistic systems that originate from real life. These applications, which occur in government, business, engineering, economics, and the natural and social sciences, are characterized largely by the need to allocate limited resources."

As pointed out in this description the problems OR deals with are mainly concerned with the allocation of scarce resources. Van Beek and Hendriks (1985: p. 3) refer in their description of OR also to the fact that OR is part of the field of mathematics:

OR is concerned with developing, analysing, and implementing mathematical models which are used for assisting a decision-making process.

Wagner (1975: p. 3) mentions the relation with mathematics as one of four qualities which characterize an OR approach:

- there should be a primary focus on decision making;
- there should be an appraisal of criteria for assessing economic effectiveness: feasible solutions should be compared using measurable values, e.g cost, or profit;
- the procedure should rely on a formal mathematical model: repetition of a process using the same data should yield the same results;
- the approach should require the use of an electronic computer.

It is obvious that without the rapid growth of computer technology OR would not be applied so frequently to real-life problems. Wagner added this fourth characteristic because of the need to use a computer for solving OR problems due to either the complexity of the mathematical model, the amount of data, or the magnitude of the computations. For simple problems no OR approach is needed to find a solution.

Hillier and Lieberman (1990: pp. 16-25) distinguish six phases in an OR study.

- Formulate the problem: the objectives and relevant conditions and constraints need to be specified at an early stage.
- Construct a mathematical model: the problem has to be reformulated in a form that is convenient for analysis. Decisions should be represented as quantifiable decision variables (e.g. production level: number of products to be produced) and the measure of performance (e.g. profit) as a mathematical function of these variables (*objective function*). Constraints can also be expressed mathematically by inequalities or equations that represent restrictions on the values of the decision variables (e.g. limited amount of raw material to produce product X).
- Derive a solution: many OR procedures aim at finding the best, or optimal solution. These solutions, however, are optimal only with respect to the model being used. It should be recognized that real-life problems are often extremely complex and that models that can be mathematically manipulated always simplify reality. Sometimes the time and cost required to search for an optimal solution are unrealistic; then the user may have to live with a solution which satisfies the constraints reasonably well. The process of searching for such a solution is called *satisficing*. This term is a combination of satisfactory and optimizing.
- Test both model and solution: the model and its solution should be evaluated. It should be checked whether the model functions appropriately and whether it should be improved.
- Check the solution: when the model is used in practice it should continuously be checked to see whether conditions that are changing in the real world require the model to be modified. A plan should be developed to detect such changes and to ensure that adaptations are made.
- Implement the results: the solutions of the OR procedure should be implemented in practice; only then can it be proved that the OR study produces benefits. The decision makers should be aware of the applicability of the solution to their decision problems.

With such an approach OR and MS originally concentrated on solving well-structured problems. According to Cyert (1981) OR should put more effort into poorly-structured problems, e.g. in the area of strategic planning or in organizational design.

3.3.2 Techniques

A number of OR procedures have been developed to solve different types of problems. Some examples of standard tools are linear programming (LP), dynamic programming (DP), and network models (e.g. Hillier & Lieberman, 1990; Winston, 1991).

LP is applied to problems in which there is a need to allocate resources to activities. This tool is applicable when both the objective function and the constraints can be expressed as linear functions. DP uses backward induction and is particularly suitable for problems that can be divided in a number of stages, often time stages. It assists in determining an optimal combination of decisions. In this case it is not essential that the mathematical functions in the model are linear. Network models can be applied to, for example, transportation problems, problems in facilities location, planning, and in a large variety of other fields. These models are very useful to depict relationships and connections between the components of systems. Generally decisions need to be made on the best way to conduct a flow through the network.

There are OR techniques that deal with deterministic problems, but also techniques that deal with probabilistic problems. The second category offers the possibility of dealing with uncertainty, if assumptions can be made about probability distributions.

There are a number of practical problems for which the search for an optimal solution is impractical. For these problems *heuristic algorithms* may be used which will generally fairly quickly find reasonably adequate, feasible solutions (satisficing) that are not necessarily optimal. Dannenbring and Starr (1981: p. 456) describe a heuristic procedure as:

"... any method of solving a problem that does not guarantee that the solution is optimal."

They describe different categories of heuristic approaches to problem solving, e.g. approximation heuristics, solution generating heuristics, and solution improvement heuristics. In general, heuristics apply logically developed rules. It cannot be proved (mathematically) that they are certain to result in a feasible solution, but they control the progress of the search for a solution.

3.3.3 OR programs and decision support systems

The relation between the developments in computer technology and the application of OR has already been mentioned (Sub-section 3.3.1). On the one hand computer programs are being developed for specific OR techniques, on the other systems are being developed in which OR techniques are combined with other approaches and techniques. A prominent example of the last group is a decision support system.

OR programs

For some OR techniques standard software packages have been developed, e.g. for LP. The mathematical formulations for other techniques, like DP, are so specific to a particular problem that it is impossible to develop standard packages. Then repeated implementation of *ad hoc* programs for each particular problem is required.

Decision support systems

Computer programs have also been developed in which mathematical models are combined with other techniques to support decision processes. These systems are called *decision*

support systems (DSSs). According to Keen & Scott Morton (1978: p. 97) DSSs provide computer-based support for management decision-makers dealing with semi-structured problems. Bennett (1983: p. 1) defines DSS as:

"... a coherent system of computer-based technology (hardware, software, and supporting documentation) used by managers as an aid to their decision making in semistructured decision tasks."

Figure 3.2 depicts the structure of a DSS (see e.g. Grünwald & Fortuin, 1989; Hendriks, 1990). Facts that are required for the decision making process are stored in the database. The model base includes mathematical models that help to find solutions. Grünwald and Fortuin (1989) state that it is not obligatory to use OR models in DSSs; there should, however, be some decision models. Presentation, communication, and manipulation of data by the user is facilitated by the user interface.

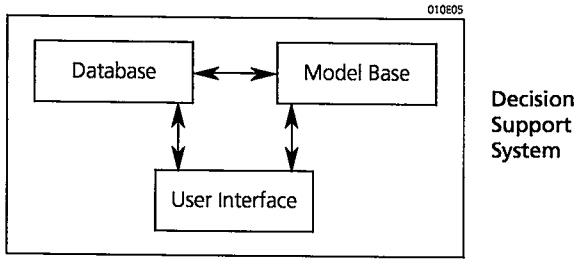


Figure 3.2 The structure of a decision support system

Keen and Scott Morton (1978: p. 2) have pointed out three main differences between OR and DSS.

- The impact of OR has mostly been on structured problems (rather than tasks), where the tasks can be pre-specified, whereas the impact of DSS is on decisions that are sufficiently structured to use computer and analytic support, but where managers' judgement is essential.
- The payoff of OR to the organization has been in generating better solutions, whereas DSSs aim at improving the effectiveness of users by extending the range and capability of their decision processes.
- OR has provided managers with detailed recommendations and new approaches for dealing with complex problems. The relevance of DSSs is that they create supportive tools, under the control of the momentaneous user.

It is obvious that DSS is not synonymous with OR or with an OR program. However, in practice, models developed in the OR field are often incorporated in DSSs.

3.4 Combining AI and OR

For many years AI and OR have developed independently and their techniques have been implemented separately. It is obvious from the foregoing sections that there are a number of similarities and differences between AI and OR. Both profit from the developments in computer technology, and both aim at providing some kind of decision support. In both fields a structured approach is generally used to develop a system for solving a particular problem. Such an approach starts with collecting knowledge and information and defining the problem. Thereafter, a model has to be constructed that needs to be elaborated before it can be used in practice. When the model is operational, it needs to be maintained. Both fields provide techniques for dealing with uncertainties. Differences between these fields are that OR focuses mostly on solving well-structured (often linear) problems in a mathematical way, whereas AI deals with poorly-structured and rather complex problems by imitating reasoning processes. Therefore AI systems try to capture human experience to solve problems, whereas OR systems use mathematical models. An additional difference is that in general AI systems are required to explain their reasoning process, whereas OR systems cannot be so required.

It is not always obvious which approach best suits a given problem; sometimes either one of the approaches can be used. O'Keefe et al. (1986) give an example of a knowledge-based approach to a problem to which quantitative (OR) methods have been applied: the problem that bankers have when analysing company accounts, with a view to extending a loan. However, they also address the possibility of integrating techniques from both fields in one system. There is a growing awareness that combining AI and OR techniques may be profitable. AI sometimes deals with optimization problems (e.g. Amarel, 1987) and with search methods that make it possible to find optimal solutions (Barr & Feigenbaum, 1981). In Bennett's book on DSSs (1983) attention is paid to both the integration of optimization models with DSSs and the combination of AI with DSSs. O'Keefe (1985) clearly points out how co-operation between the two fields can produce mutual benefits: e.g. OR can profit from the addition of knowledge-based methods to quantitative approaches, for instance in terms of heuristics, and expert systems may increasingly employ optimization techniques. Simon also recognizes these benefits, as can be concluded from the title of a paper on this subject 'Two heads are better than one: the collaboration between AI and OR' (Simon, 1987). He states that researchers from both fields should adopt a problem-oriented point of view: the main objective in practice should be to solve problems using the most appropriate techniques, instead of letting the techniques determine which problems are dealt with. This view is adhered to in this study.

3.5 Statistical support systems

The objective of the computer system to be developed in this thesis is to assist in the design of schemes for soil surveys using statistical sampling strategies. The system can therefore be considered as a statistical support system. This section deals with the use of computer systems for statistical work.

Hand (1984) distinguishes four types of software packages for statistical work:

- simple packages (for calculators or computers) that can easily be used for data analysis;
- complex packages (for computers) that require professional statistical knowledge for data analysis;
- simple or complex packages with an interface that facilitates access for statistically naive users and helps to prevent misuse of analysis techniques by guiding the users;
- expert systems that should be able to give advice on both experimental design and data analysis.

A domain for statistical ESs, which is not mentioned by Hand but which is especially relevant in this thesis, is sample design. The first two types of statistical software distinguished by Hand are based on arithmetic expertise and the user has to determine which statistical techniques can be meaningfully applied. The second two contain statistical knowledge supporting intelligent use. These systems try to apply AI to statistics in two directions: to make statistical software more applicable to statistically naive users, or to make a statistical assistant. This is a computer program which contains statistical experience, questions the user about his aims, and assists in the design of experiments or samples and/or in data analysis. Developments of systems in the last direction are especially relevant to this study, and in this section some developments are discussed. Most references deal with envisaged systems, or with prototypes. According to Adèr (1992) there are no operational statistical ESs; most systems are in an experimental phase or have been abandoned before completion.

Systems that serve as statistical assistants have two main objectives: to support the design of experiments, or to support of the choice of a technique for statistical analysis of data. Jones (1980) discusses some aspects related to the design of statistical experiments but his findings are not based on implemented systems. He concludes that a computer can be used most efficiently for designing familiar experiments, for which the necessary information can be specified in advance. A computer program is based on formalized knowledge, which can be acquired more easily for familiar experiments than for exceptional experiments. There is also more experience of and consensus on familiar experiments. An advantage of a computerized support system will be that it ensures that all important aspects will be considered during the design, but it will never be able to replace an experienced consultant, because it is programmed in advance and lacks the versatility of a human consultant, who is able to illuminate unnoticed aspects of problems and to deal with closely related problems. Jones stresses the importance of an explanation facility, especially for someone who is not a statistician. The system he aims at may be an ES or a KBS in a limited domain.

De Greef (1991) deals with the development of a statistical consultation system to assist in the planning of data collection and statistical analysis for psychological surveys. Two prototype systems have been developed using the KADS methodology, an emerging methodology for the development of KBSs (see Sub-section 3.2.3). In his paper De Greef focuses on the importance of co-operation with the intended users, and the statistical domain is used only as a case study to develop a methodology.

Gale (1986) and Adèr (1992) refer to statistical ESs for data analysis. These systems are based on human expertise, and it is recognized that they are in an experimental phase. It is concluded that consultation systems in data analysis are feasible, but that further effort should be put into extending explanation facilities and formalizing the domain knowledge. With respect to statistical domain knowledge, Adèr states that it has been hardly investigated

systematically and he refers in particular to the differences between experts. These differences make it difficult to use a single strategy in the design of systems. Knowledge formalization may help to unify knowledge from different experts. According to Gale, support for knowledge acquisition is required to enhance the value of knowledge-based consultation systems in data analysis.

Schach (1986) deals with ideas for a system which is most closely related to the system aimed at here: computer support for the design and analysis of survey samples. However, as far as known such a system has not yet been developed. Schach aims at sampling from a human population, e.g. for market research or opinion research. The computer is expected to become a partner in a statistical dialogue, which issues recommendations on data collection and analysis. It will ensure that all relevant aspects are considered before reaching a final decision. Schach doubts suggestions of letting a computer system that uses strict mathematical criteria make decisions. In his experience, there are very few real statistical problems for which one single course of action is possible. Then, strict mathematical criteria might yield no solution. He states that a life statistician should always be involved to weigh the consequences of possible options for which not all conditions are satisfied.

Jöckel (1986) and Van den Berg and Visser (1990) have noted that besides statistical knowledge, knowledge from other disciplines in the domain also influences the problem solving process and should be incorporated in the statistical support system. As far as known at the moment, there are no (experimental) systems available containing both pedological and statistical knowledge to assist in the design and analysis of soil surveys using statistical sampling strategies.

3.6 Applicability of techniques in this study

This section deals with the applicability of each of the techniques discussed above to develop a system to assist in the design of soil survey schemes. In the first chapter it has already been noted that at least the following disciplines are involved in this project: soil science, statistics, computer science, and operations research. This indicates that each discipline seems individually incapable of fully assisting in the design of soil survey schemes. In the following sub-sections the possibilities of using only AI, OR, or statistics are considered.

3.6.1 Why not just AI?

It seems attractive to use AI for the problem at hand, since human expertise on designing soil survey schemes needs to be captured in the system. However, the system cannot be based on the knowledge of human experts only, knowledge from different sources has to be used. Therefore, a KBS, which operates on previously stored knowledge, seems to be appropriate. When not all knowledge which is necessary to provide clear conclusions on problems in the domain can be specified in a system in advance, the user is often obliged to make certain decisions in the problem solving process. For the system aimed at, it seems attractive that the user can make or adjust some of the decisions during the design process. There is a wide variety of surveys for which the system could provide assistance, and it seems impossible to provide a system - at an early stage of development - with relevant decision criteria for each situation. Each soil survey project, with its own specific conditions, will require

input from the user. If the system is provided with a learning capability, the user may be asked to explain his or her decisions so that additional knowledge can be continuously stored.

As mentioned in Section 2.4, the system should be able to perform some tasks which at the moment are not performed during the design of soil survey schemes, namely prior evaluation of accuracy and cost, and searching for an optimal scheme. At the moment there is no human experience on performing these tasks. Algorithmic procedures seem to be required to provide objective, quantitative results for these tasks, instead of using symbolic reasoning. So, besides AI techniques other techniques should be used.

An option might be to develop an EDS; knowledge of historical cases could then be stored, and retrieved and re-used when similar cases arise. Such an approach is referred to as case-based reasoning (e.g. Carbonell & Langley, 1987). This approach seems inappropriate to start with for three main reasons. In the first place, the number of usable, historical cases is very limited, since statistical methods have not been used frequently in soil survey practice. In the second place, if statistical methods have been used, the design of the survey scheme, and especially the outcome and a justification, is very hard to reconstruct. Thirdly, the system should be able to deal with a large variety of cases, e.g. different survey regions, different soil properties of interest, different constraints. It seems hardly possible and impractical to collect and store all knowledge which is only specific to a particular case.

It seems attractive to use various techniques to develop the system; AI techniques will certainly be useful, for example, to support problem specification, to support knowledge-based selection of applicable statistical methods, and to develop a user interface with an explanation facility. Considering the nature of the domain knowledge, it seems desirable that the system should be able to provide the user, if needed, with background information, and to explain how it reached its solution.

3.6.2 Why not just OR?

The system aimed at here cannot be developed using a pure OR approach, since soil survey schemes cannot be designed using only mathematical models. This is due to the fact that not all decision variables are quantifiable, e.g. the choice of a type of sampling design is not only determined by quantifiable criteria, but is related to the type of request, among other things.

It is, however, attractive to use OR for one task in the design process, namely the search for a scheme that satisfies the constraints as well as possible, and results, if possible, in an optimal solution. The main constraints are the budget available and the accuracy required. The aim may be to find the most accurate scheme for a given financial constraint, or to find the scheme with the lowest cost for a given accuracy constraint. Before the search for an optimal scheme can start, a number of decisions have to be made, e.g. on a suitable type of design, and on the method of determination to be used. Moreover, to enable this search, mathematical models are required for predicting accuracy and cost of schemes. These models must be available before an appropriate OR technique can be chosen. The search for an optimal scheme is a problem of allocating scarce resources, e.g. the available budget or the available time for the survey project, for which an OR approach is attractive.

3.6.3 Why not just a statistical package?

It goes without saying that the system should provide statistical support, but in addition knowledge on at least one other part of the domain, namely soil survey, should be incorporated. A simple statistical package is insufficient to support fully the design of soil survey schemes. Selection of applicable statistical methods to collect and analyse data is a task that requires statistical expertise. Therefore, a statistical ES or a statistical KBS seems more appropriate, albeit that the topics that have to be addressed exceed the features of these systems.

3.7 A system to assist in the design of soil survey schemes

How can the system aimed at here be characterized? The main objective of the system is to assist in the process of designing soil survey schemes. Therefore, the system should have both pedological and statistical knowledge at its disposal. So, the system should possess characteristics of a statistical KBS. Furthermore, the system should be able to perform some additional tasks, which at the moment, are not handled by a human expert, and for which no procedures are available. To fulfil these tasks knowledge has to be generated. At the first place, models are required to predict the accuracy and cost of survey schemes, i.e. models for prior evaluation. At present, accuracy and cost are only roughly assessed, since there are no well-founded evaluation models. Secondly, when evaluation models are available in which the influence of the type of sampling design is explicit, these can be used to support the search for a scheme that satisfies the constraints as well as possible.Techniques from the field of OR may be useful to support this second task.

So, at least the following disciplines need to be involved in development of the system: soil science, statistics, computer science (in particular AI), and OR. The system will be indicated as a *knowledge-based system to assist in the design of soil survey schemes.* Figure 3.3 depicts the rough structure of this system.

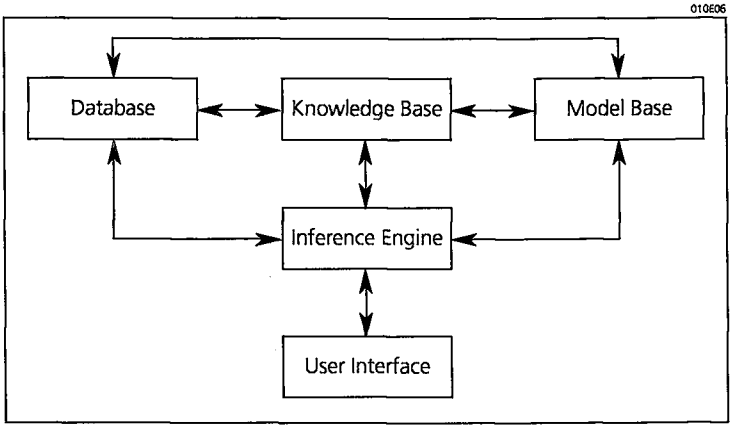


Figure 3.3 Rough structure of a knowledge-based system to assist in the design of soil survey schemes.

Soil survey data should be stored in the database, pedological and statistical knowledge in the knowledge base, and evaluation and optimization models in the model base. Precisely which knowledge and information need to be stored and how it should be stored, will be dealt with in the following chapters. The inference engine should control the design process of soil survey schemes. The final inference structure is related to the structure of the design process and the way in which the knowledge can be structured and stored. In this study evaluation and optimization procedures are developed. The user interface should guide the communication with the user, and provide explanation when required. This study focuses on the knowledge that should be incorporated in the system, and only limited attention will be paid to exactly how the system should appear to the user.

Chapter 4

Structuring the domain of soil survey projects

Parts of this chapter have been published in:

Domburg, P. & Elzas, M.S. (1994)
Structuring the Domain of a Complex System: a basis for a knowledge-based system supporting soil survey design. In: Beulens, A.J.M., Doležal, J. & Sebastian, H-J. (Eds.), Optimization-Based Computer-Aided Modelling and Design, Proceedings of the second Working Conference of the IFIP TC 7.6 Working Group, Dagstuhl, Germany, 1992. Leidschendam, Lansa Publishing, pp. 181-195.

Domburg, P. & Gruijter, J.J. de (1992)
A framework of concepts for soil survey using probability sampling. Report 55. Wageningen, DLO Winand Staring Centre for Integrated Land, Soil and Water Research.

4 Structuring the domain of soil survey projects

4.1 Background

In the field of AI ESs and KBSs are considered to be computer programs that respectively solve or assist in solving problems. In the preceding chapter five KE activities were mentioned: knowledge acquisition, knowledge system design (i.e. knowledge structuring and knowledge representation), knowledge programming, knowledge refinement, and knowledge maintenance. These activities can start after identification of the characteristics of the problem to be solved. These include the definition of the problem, and the specification of the knowledge-sources, the required resources, and the goals or objectives for building the system. Chapter 1 and Chapter 2 dealt with this identification.

This thesis primarily deals with the first two KE activities: knowledge acquisition, and knowledge system design. During knowledge system design no strict decision is made on the tools to be used during the implementation. It is difficult to determine beforehand which tool suits best; the early choice of a tool for implementation may influence the form of the final system (e.g. Waterman, 1986). If an inappropriate tool is chosen, one may be inclined to rework the problem so that it fits the capabilities of the tool, in which case the final system will support a different problem. A tool chosen at an early stage may also turn out to lack some relevant features, e.g. to create interfaces with existing information systems, or to provide efficient explanation facilities.

This chapter presents the results of knowledge structuring. Knowledge acquisition has been used to gain insight into the domain. Section 4.2 deals with the approach used for domain structuring. Thereafter three subsequent domain levels are described (Sections 4.3 to 4.5).

4.2 Approach to domain structuring

The characteristics of the domain (Sub-section 4.2.1) and knowledge on methods of domain structuring gave rise to distinguish layers as a basis for structure (Sub-section 4.2.2). Besides two cases are introduced which will be used in this and the following chapters as illustrations.

4.2.1 Domain characteristics

The development of a KBS which has to function smoothly in an existing working environment, e.g. an organization or a research field, requires understanding of the position of the system and its potential users in the environment. Besides, the tasks of the system need to be specified, and the concepts used in the domain need to be defined. The domain of interest in this thesis is complex because:

- the process to be supported can be viewed as a design process; the difficulties of supporting design processes are generally recognized; Stefik et al. (1983a) enumerate six key problems related to design problems in general:

1. consequences of design decisions cannot be assessed immediately, therefore design possibilities must be explored tentatively;
2. constraints on a design come from many sources, and usually are hard to integrate with design choices;
3. design problems often have to be divided into sub-problems, which are, however, seldom independent;
4. since it is often hard to assess the impact of a change in part of the design, a design system should be able to explain all decisions taken in different sub-systems;
5. if re-design is required, a picture should be provided of the total problem, to avoid attention being focused on local optima;
6. design problems often require reasoning about spatial relations; since it is difficult to reason approximately or qualitatively about such relationships, considerable computational resources are needed.

All these problems also apply to the design of soil survey schemes. It will be shown in this thesis that: (1) the design process is iterative and different options should be explored and compared objectively (e.g. this chapter, Chapter 8); (2) constraints on accuracy and cost and logistical constraints should be considered during the whole design process (e.g. Chapter 6, Chapter 7); (3) the steps in the design process are interrelated (e.g. this chapter, Chapter 6, Chapter 8); (4) explanation facilities are required to gain a clear understanding of the whole process (Chapter 8); (5) if there is a change in one of the steps distinguished, the consequences for the whole problem should be considered and not only the effects on a particular step, since the process consists of interrelated steps (e.g. this chapter, Chapter 8); (6) the spatial component in soil survey is obvious; selection of samples and prediction of the accuracy of schemes requires considerable computational capacity (see Chapter 7).

- there are several disciplines involved; of course, pedological and statistical knowledge are required to design a survey scheme; if this process is to be supported by a computer system, contributions from the fields of computer science are also required; furthermore, to enable the search for an optimal design, techniques from the field of OR may be useful;
- at present, no structured approach to designing soil survey schemes is in use; however, KBSs are only manageable with a structured approach;
- some tasks are to be supported which cannot be performed at present and for which procedures need to be developed; the knowledge to be generated is vital to the whole system, but performance of the proposed procedures is hard to assess in advance;
- there is not a single expert for the whole problem domain, but there are people experienced in various parts of this domain. In practice, these experts do not always agree;
- the number of possible combinations of type of sampling design and method of determination is virtually unlimited, whereas one efficient scheme should be constructed in a short period from this initially unbounded solution space.

Such a complex domain can be structured by looking at it from different angles. The KADS methodology, which mainly supports knowledge acquisition and modelling for the development

of KBSs, is in particular usable for domains in which the structure of the problem-solving process (i.e. the so-called inference layer) can be easily discovered. However in this thesis, the structure of the domain of interest was not clear at all when the project started. Therefore, a situation specific approach to domain structuring was developed. This approach is partly based on existing methods.

4.2.2 Layers as a basis for structure

Three interrelated layers are distinguished to describe the complex soil survey domain from a broad to a more detailed level. The relation between these three layers is shown in Figure 4.1. In KADS different layers are also distinguished (Breuker & Wielinga, 1989), but as indicated above, a somewhat different approach is applied in this thesis.

At the highest level, a soil survey project is described as a system consisting of a hierarchy of entities and aspects, i.e. an *entity structure*. The *aspects* refer to a process or stage of a project, and the *entities* are characteristics of a particular aspect. The usefulness of descriptions of systems using entity structures for AI application development has been shown by Elzas (1989). The systematic design of complex systems requires rational choices. Entity structuring is used here to describe the scope of a survey project as a whole: the position of the system under the present working environment is depicted. Thereafter the scope of the KBS can be specified.

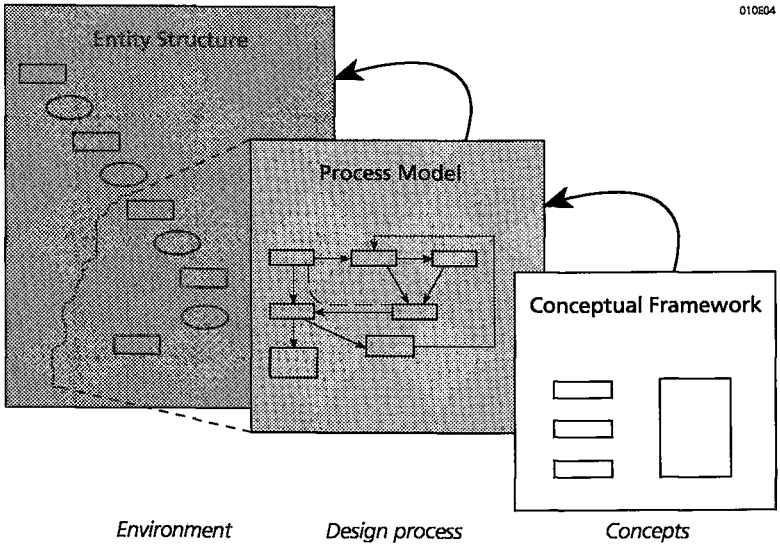


Figure 4.1 Three interrelated domain layers

The design process of soil survey schemes is described as a process model (intermediate level). In this model, different phases, and connections between the phases in the design process, are specified. This model serves as a basic structure for the design of the system proposed.

At the lowest level, a conceptual framework is defined for effective communication between

the future system and a user, and also for communication within the system. All concepts which are important during the design process are included in this framework.

The three layers are described in the following sections. This approach to domain structuring serves as a firm basis for the development of a computer system to assist in this complex domain. It is applicable to, and has advantages for, the structuring of complex systems in general. The approach ensures attention is paid to:

- the position of the system which has to provide computerized support in an existing environment;
- the dynamic process to be supported;
- explicit and unambiguous definitions of domain concepts.

4.2.3 Two cases

This section outlines the backgrounds and aims of two soil survey cases using probability sampling which will fulfil a guiding role in the following chapters. These cases are somewhat schematic versions of original surveys carried out at the DLO Winand Staring Centre. The historical cases are simplified to avoid irrelevant complications and to fit into the limited scope of this thesis.

Case A: *Phosphate saturation in the Ootmarsum region*

This case is part of a survey project commissioned by the Province of Overijssel (the Netherlands), which aimed at quantifying the areal proportion of cultivated soil saturated with phosphate in two regions in the province (Hack-ten Broeke et al., 1990). The purpose of the project was to quantify the areal proportion of soil saturated with phosphate in two regions representative of the eastern sandy regions of the Netherlands, given a particular definition of saturation. At a higher level, results of this study were used to support decisions on the control of groundwater and surface water pollution.

Background. *In many rural regions in the Netherlands a great deal of animal manure is produced. If the soil is fertilized with too much manure it becomes saturated with phosphate from the manure. Phosphate then leaches and pollutes the groundwater and surface water. The phosphate sorption capacity, or the maximum areic mass of phosphate sorbed by soil (P_{max}) varies between soils. It is related to the oxalate-extractable iron (Fe) and aluminium (Al) in the soil, the density of the soil, and the depth of the mean highest water table (MHW) (e.g. Schoumans et al., 1989: p. 201). The relative mass of phosphate sorbed by soil (P_{rel}) is calculated by dividing areic mass of phosphate sorbed by soil (P), i.e. the actual phosphate content, by P_{max}, both summed over depth to the MHW. If P_{max} and P_{rel} are known, this information can be used to support decisions on the control of groundwater and surface water pollution.*

The Ootmarsum region in Overijssel was one of the regions selected to be surveyed. Part of this region has a high elevation and deep groundwater tables. Roughly thirty percent of this region is used for forest with a nature conservation role. Many valleys cross the region. The region consist largely of sandy soils, but there are clay soils in some places. The spatial inventory was confined to the areas used for agriculture.

Case A includes one region from the original survey: the Ootmarsum region. The purpose of this case is to estimate the proportion of the region in which the soil is saturated with

phosphate, according to a given definition of saturation. The survey region consists of 2252 ha of agricultural land near the village of Ootmarsum. Certain specific features must be taken into account while designing a scheme for a soil survey in this region for the objectives mentioned above. One feature is the presence of dry and wet sub-regions, the latter being relatively small in relation to the whole survey region. Wet areas are more sensitive to phosphate leaching than drier ones and therefore particularly accurate information is required about them. Another feature is that correlations are assumed to exist between map units of the available soil map (scale 1:50 000) and land use categories on the one hand and P_{max} and P on the other. Both these features have an important impact on the design of the survey scheme, apart from the usual constraints concerning the available budget and required accuracy.

Case B: *Mean highest water table in a map unit of the 1:50 000 soil map*
This case is derived from the project 'National Sampling Map Units' carried out at the DLO Winand Staring Centre. The purpose of the project is to upgrade the national soil map of the Netherlands, scale 1:50 000, by collecting detailed quantitative information on the spatial variability of soil properties within the map units. The first sample of this project relates to map unit Hn21-VI (Veldpodzol-gronden in groundwater class VI) (Visschers, 1993).

Background. *The national soil map of the Netherlands is a multi-purpose map. The 62 map sheets, mainly subdivided into West and East, were produced by the free survey method, so without the use of statistical methods. This project was started about 30 years ago and is now nearly finished. The map sheets have extensive legends and notes which contain mainly qualitative information and only limited quantitative information on the spatial variability of soil properties; see for example Damoiseaux et al. (1990), and Vleeshouwer & Damoiseaux (1990). The project 'National Sampling Map Units' aims at satisfying the growing need for soil information with quantified accuracy by upgrading the existing national soil map.*

In the original study, data on all soil properties generally relevant to sandy regions have been collected, whereas case B focuses on collecting information related to the MHW. The domain of the system proposed is primarily limited to inventory studies in which one soil property is of primary interest (Sub-section 2.2.3) so, only this single property should be considered when designing the survey scheme. The MHW is selected as a property of interest because it is highly relevant to many other research projects, particularly for environmental and land evaluation studies. The purpose of case B is to estimate the spatial mean of the MHW in map unit Hn21-VI. The survey region contains all delineations on the 1:50 000 national soil map classified as map unit Hn21-VI.

The geometry of the survey region, with map delineations of Hn21-VI being distributed all over the Netherlands, makes it impossible to visit locations in all delineations. This would be too time-consuming and would result in high costs of travel. Besides this constraint, which is related to both logistics and financial aspects, the budget available for the project is limited and some minimum accuracy of the results is required. These constraints all affect decisions related to the design of the survey scheme.

4.3 Soil survey project as entity structure

A soil survey is often only a part of a larger research project. In such a case the specific aim of the survey is embedded into the broader purposes of the project. The project purpose of case A is for example to quantify the areal proportion of soil saturated with phosphate in a region representative of eastern sandy regions of the Netherlands, in order to support the control of groundwater and surface water pollution in all eastern sandy regions. The aim of the survey is more specifically to determine the areal proportion of the Ootmarsum region (2252 ha of agricultural land) in which the soil should be considered saturated with phosphate, given a particular definition of saturation.

The entity structure depicted in Figure 4.2 takes into account the whole scope of a soil survey project.

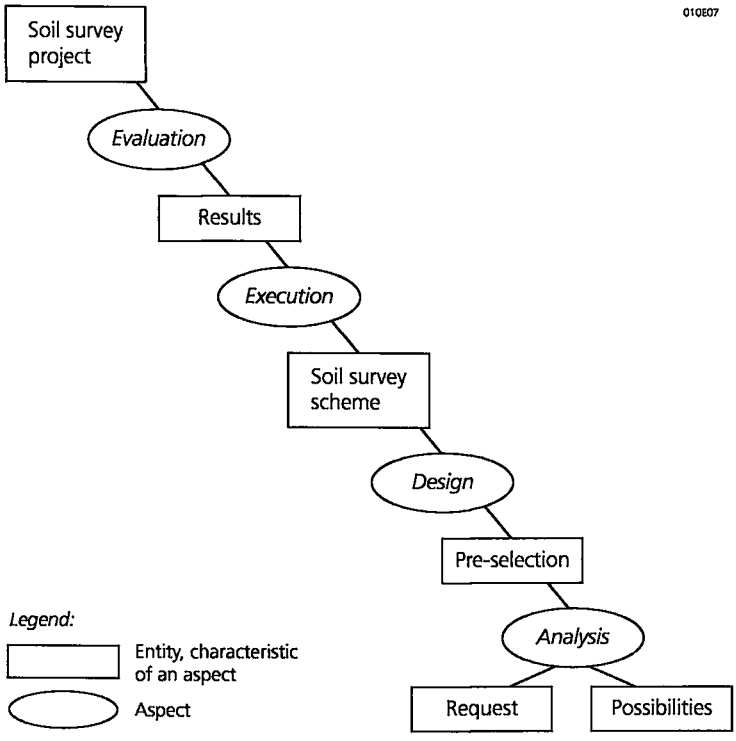


Figure 4.2 Entity structure of a soil survey project

A survey starts with a request (specifying aim and constraints) on the one hand, and possibilities (concerning sampling techniques, available prior information, and methods of determination) on the other hand. Both the request and the possibilities have to be analysed to provide a pre-selection of applicable methods of determination and design types. This is the basis for the design, which results after a number of iterations in a soil survey scheme. The execution of this scheme (i.e. collecting and analysing data) provides the survey results.

After evaluation of the results, including a comparison with the original scheme, the soil survey project is completed. The evaluation aims at collecting and storing information resulting from the project, so that it can be utilized in future projects. At present, the use of information and experience from historical surveys is limited. During the construction of this entity structure the need for a system that learns from its use was discovered. To improve the design process continuously, completed projects should be evaluated. Adding this step to soil survey projects ensures continuous improvement of the system and increase of the knowledge. It is clear that the success of *evaluation a posteriori* is highly dependent on the willingness of the user to give feedback. The KBS proposed in Chapter 1 will focus on the process which starts with definition of a request and ends with the generation of a report on a soil survey scheme. The design process, as described below, is a dynamic instantiation of a part of this static entity structure.

4.4 Model of the design process

The model of the design process, i.e. the process model, (Fig. 4.3) is primarily based on the actual procedure of constructing a survey scheme. In addition, some components which will also be supported by the future system, are included in the model. The main additional aspects are the prior evaluation of schemes, and the comparison of schemes.

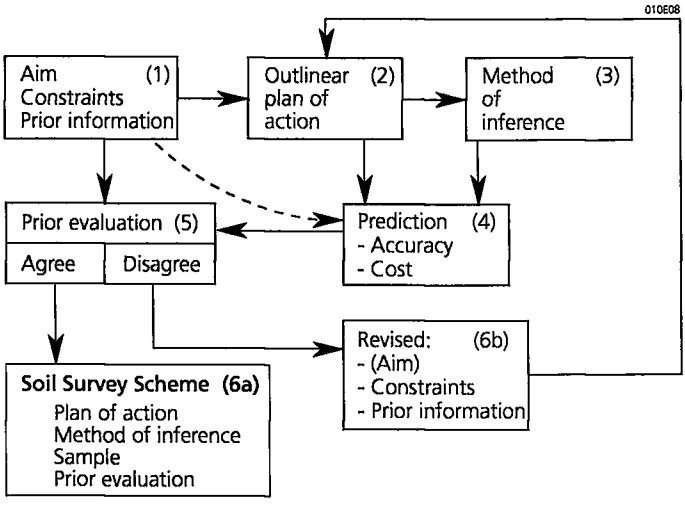


Figure 4.3 Model of the design process

It is worthwhile evaluating a scheme before field work starts, so that it is possible to check whether the researchers and those commissioning the project agree with the consequences of the scheme. If they disagree with the predicted accuracy or with the cost, the scheme can be adapted before actually starting data collection.

In the present situation only limited attention is paid to evaluating soil survey schemes beforehand. The scheme that is finally proposed is assumed to fulfil the aim and constraints

reasonably well, but at present the time and means to compare possible schemes are generally lacking. If models to predict cost and accuracy are available, a better comparison of potential schemes is possible. The ultimate goal of the design process is to construct an efficient scheme.

Different steps can be distinguished during the design process, which is an iterative process. (1) It starts with defining aim and constraints, and searching for adequate prior information. (2) The next step entails the construction of an outlinear plan of action, including preliminary choices of method of determination and type of sampling design. The term *outlinear* is used since at this stage only a few of the elements which need to be specified in the final plan of action are considered. (3) The third step focuses on the method which will be used to analyse the data statistically. This is determined by the statistical method of inference, which is related to the type of sampling design. (4) Once these parts (i.e (2) and (3)) of a scheme are known, tentative predictions of accuracy and cost can be made using prior information. (5) These predictions have to be compared with the original aim and constraints (*prior evaluation*). By repeating this evaluation for different numbers of observation points, it is possible to search for a scheme that satisfies the constraints as well as possible, i.e. search for an optimal scheme. (6a) If there is sufficient matching, the final scheme can be elaborated. (6b) Otherwise, the process has to start again, e.g. by looking for another sampling design or by revising the constraints. Of course, there may be more than one possible outlinear plan at step (2). Then these plans should be evaluated independently. If there is disagreement at step (5), it may be advisable to evaluate the other outlinear plans first, before revising the original input (6b). In general, only minor changes in the aim will be acceptable, but it may be possible to revise the constraints, e.g. by providing additional budget or more accurate prior information.

4.5 Conceptual framework

4.5.1 Important concepts

Explicit and unambiguous definitions of the concepts used are required for effective communication on the problem domain. Furthermore they are the basis for a structured approach to soil survey, which is one of the principal requirements for the development of a KBS. The conceptual framework depicted in Figures 4.4 and 4.5 is based on the literature on statistics concerning sampling (Cassel et al., 1977; Cochran, 1977; Krishnaiah & Rao, 1988) and on general experience in soil surveying. The concepts in the framework bear a slight resemblance to the principal steps in a sample survey as described by Cochran (1977: p. 4-8), who writes about sampling techniques in general. As this thesis focuses on the application of sampling strategies in soil survey, some additional concepts are needed. The order of the concepts is based on mutual relationships, and on the order in which they appear during the design of a sample survey.

The development of the conceptual framework has revealed that there is sometimes confusion about the applicability and the meaning of certain concepts. Also differences in terminology are found in the literature, and often no crisp definitions are given. Sometimes, a choice has been made from different definitions, or a definition has been adjusted to the framework.

Besides the use of the framework in the system aimed at, it may facilitate negotiations concerning aims and constraints of soil surveys between contracting partners in general. The absence of clear concepts can cause ambiguity and confusion among researchers, or between researchers and decision-makers (e.g. policy-makers). The use of unambiguously defined concepts supports effective communication between all parties involved in a soil survey project.

The figures below together show the conceptual framework. Figure 4.4 can be interpreted as the input of the support system. The user will play an important role during the specification of these concepts. Figure 4.5 shows the constituent parts of the output of the design process. At present, the documentation on survey schemes is often incomplete, but all the elements depicted in Figure 4.5 usually appear during the development of a survey scheme. In future, attention will need to be paid to all concepts, if a structured approach to soil survey is to be used.

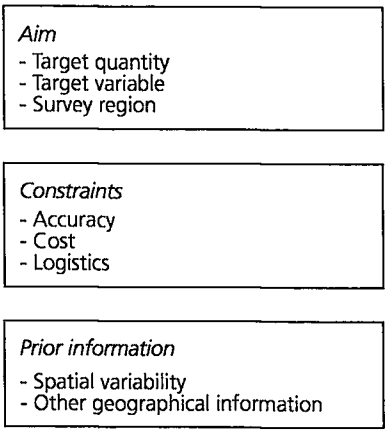


Figure 4.4 Factors governing the construction of a soil survey scheme

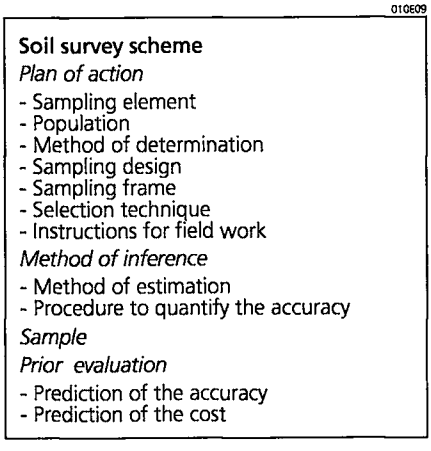


Figure 4.5 Structure of a soil survey scheme

In designing a sample survey the possibilities are always bounded by various constraints. The design of a survey scheme starts with specifying the *aim* and *constraints* of the survey. These two factors will then guide the search for relevant *prior information* from previous (soil) survey projects, for example, information on how the property of interest varies in space, or areal information represented on soil maps or stored in a GIS database. Stein (1994) deals with the use of prior information in spatial statistics, including both interpolation and sampling. Besides information on variability and areal information, he introduces models to determine values of soil properties as a source of prior information. He refers to models based on chemical and physical laws, which describe relationships or processes. In this thesis information on these models is not distinctly considered as a source of prior information, but it has to be incorporated in the methods of determination.

The design process starts from the available pedological and statistical knowledge, and from the explicit specifications of the aim, constraints and prior information. The final result

of this process is a scheme specifying:

- the principal steps in organizing survey sampling: a *plan of action*;
- the way the data are to be analysed statistically: the *method of inference*;
- the selected set of sampling elements to be observed: the *sample*;
- predictions of the accuracy and cost expected to result from implementing the scheme: the *prior evaluation*.

The main decisions to be made during the design of soil survey schemes are the choice of a method of determination and the choice of a sampling strategy. A *sampling strategy* is the combination of a sampling design and an estimator - these concepts are defined in Subsection 4.5.2 -. In the present situation decisions on the estimator, which is part of the method of inference, are often not made at the same time as decisions on the sampling design, but at a later stage. Initially this caused the separation of the parts of a sampling strategy in the structure of a soil survey scheme. Furthermore, the present structure of a scheme distinguishes between organization of data collection and the statistical analysis (method of inference), which is a practical distinction. Besides, the method of inference includes not only the estimator but also the procedure to quantify the accuracy of the estimator from the sample data.

4.5.2 Definition of concepts

The concepts that are used are defined below and illustrated with examples from the cases introduced in Section 4.2. Since no structured approach to designing and describing soil survey schemes has been available previously, the available information on historical studies is often incomplete. Such information could not be recovered retrospectively due to limited documentation. Descriptions of constraints and the prior information used are almost always lacking. Information on elements of the plan of action, such as selection technique and instructions for field work are also rarely reported. Furthermore, the prior evaluation of a scheme for the survey is never described. Therefore, some concepts cannot be illustrated with the cases.

The **aim** of a survey consists basically of the following three elements: the target quantity, the target variable and the survey region.

- *Target quantity.* The target quantity is the quantity to be estimated or predicted from the sample survey data. Examples are: means, proportions (of the region having a given condition), quantiles, tolerance intervals, and measures of dispersion. Such parameters can be estimated from observed values of elements of the population. Note that the whole frequency distribution can also be estimated by calculating the areal proportions for a sequence of increasing threshold values. In the event of a geostatistical approach to a soil survey, the target quantity may be stochastic and may have different possible values in a given situation. With sample surveys (the cases considered in this thesis) the target quantity is a parameter.

Case A: proportion (of the region where, according to a given definition, the soil is saturated with phosphate).

Case B: mean (spatial mean of the MHW in map unit Hn21-VI).

- *Target variable.* Target variables are soil properties (e.g. the highest groundwater, the clay content, or the moisture supply capacity) of which a target quantity is to be determined by the survey. Although values of quantitative variables may be measurable, it sometimes suffices to record them as only present or absent (Webster & Oliver, 1990: p. 6); for example a certain location in the field may be recorded as being saturated or non-saturated with phosphate.

  Case A: a variable, indicating for any given point in the area whether or not P_{rel}, defined as P divided by P_{max}, both averaged over depth to the MHW, exceeds 0.25.

  Case B: depth of MHW in centimetres.

- *Survey region.* The survey region is the geographical region to be surveyed. The boundaries and location of the region are important here. As stated earlier, this thesis considers survey regions as planes, i.e. two dimensional. These regions may be spatially contiguous or non-contiguous.

  Case A: 2252 ha of agricultural land near the village of Ootmarsum as indicated by the authority commissioning the project.

  Case B: all delineations on the 1:50 000 national soil map of the Netherlands classified as map unit Hn21-VI.

Requests for a soil survey are always accompanied by **constraints** concerning the following three aspects.

- *Accuracy.* There are two issues to consider with respect to the accuracy of the survey results. First, it may have to meet a minimum requirement. If, for instance, accuracy is defined as the mean squared error of estimate, that quantity might be required not to exceed a given value. Such a constraint controls the quality of the result. This quality can be improved by taking larger samples, by using more efficient sampling designs, or by using more accurate methods of determination, but any of these will usually also increase time and cost.

  Second, it may be required that the accuracy of the survey results can be quantified from the sample data alone, i.e. without recourse to assumptions about the nature of the spatial variation. Such a requirement will diminish the class of admissible designs, since there are sampling designs for which no variance formulae can be constructed, like systematic sampling.

  Decisions on accuracy requirements should be made by those who will be using the survey results.

Case A: the survey region consists of dry and wet sub-regions, of which the latter represent a relatively small proportion of the whole survey region. Accurate information is especially required on the wet regions, because these are more sensitive to phosphate leaching than drier regions. One should strive for maximum accuracy given the available budget. The accuracy of the result must be quantifiable from the sample data.

Case B: one should strive for maximum accuracy given the available budget. The accuracy of the result must be quantifiable from the sample data.

- *Cost.* The cost of a spatial inventory is mainly determined by the sampling design and method of determination. The available budget, which is almost always limited, influences the choice of these two elements of the plan of action.
- *Logistics.* A third category of constraints are those of a logistical nature. Restricted capacity in a laboratory, or a limited period in which the field work may be done, are examples of this category. Such constraints limit the maximum allowable sample size, if no additional capacity can be made available.

All three categories of constraints generally affect the design of a soil survey scheme by restricting the number of possible solutions.

Once aim and constraints have been established, attention should be paid to what prior information is available from previous studies. The following main categories of **prior information** are distinguished.

- *Spatial variability.* Information on the spatial variability of soil properties, i.e. the way in which the properties vary in space, can be used to support the design of an efficient soil survey scheme. It may, for example, assist in determining whether it is attractive to divide the survey region into sub-regions (see Chapter 6). Prior information on spatial variability is required to predict the accuracy for a given soil survey scheme (see Chapter 7). If available, information on the target variable in the survey region should be used. Otherwise information on a *co-variable*, i.e. a variable known to be related to the target variable, or information about similar regions elsewhere may be useful. Some information on spatial variability can be derived from soil maps.

  Case A: information on the spatial variability of the same target variable (P_{rel} at points) in a comparable survey region may be useful. For case A information and experience from a comparable soil survey project in the Province of Gelderland (Breeuwsma et al., 1989) could be utilized. If available, information on spatial variability of co-variables in the region related to the P_{max} could also be used; for example information on the oxalate-extractable Fe and Al, and on the depth of MHW. In this study information on the MHW is derived from the soil map, scale 1:50 000.

  Case B: information on the spatial variability of the MHW in sandy regions could be useful.

- *Other geographical information.* Apart from information on spatial variability other geographical information could also be useful to set up a soil survey scheme. Examples are: soil maps, land use maps, soil survey data from reports, databases and GISs, data

on vegetation, geomorphology, etc. This information may be used in setting up a sampling design. Besides, it might be helpful in dealing with logistic constraints. For example, the units on a soil map could be used for stratification and soil survey information from databases or reports could support the choice of the strata, for example by combining map units, or by combining units of a soil map and units of a land use map.

Case A: national soil map, scale 1:50 000, map sheet 28 East; land use map; topographical map.

Case B: national soil map, scale 1:50 000, all map sheets with delineations classified as map unit Hn21-VI.

The specific answers to various questions arising during the design process can be regarded as elements of a soil survey scheme. At present such schemes are not documented in full detail, but the elements mentioned here are all relevant to soil survey, and will need to be made explicit in the future if the design process is supported by a computer system.

A **soil survey scheme** consists of a plan of action (including the sampling design and the method of determination), the method of inference, a specification of the selected sample, and the prior evaluation of the scheme. Before field work starts all these elements need to be specified.

The **plan of action** includes the following items.

- *Sampling element.* Sampling elements of interest are defined as all (possible) objects that are identifiable and that are elements for the method of determination. Only a subset of sampling elements can be observed in a sample survey. Examples of sampling elements in a soil survey are: a particular soil pit, an augering, or a soil sample.

  Case A: standard augering to the MHW, with a maximum of 1 metre.

  Case B: standard augering to the mean lowest water table (MLW), with a minimum of 1.5 metre.

- *Population.* The population is the aggregate of sampling elements of interest, existing in a specified region (the survey region) at a specified point in time (during a specified period of time) (Krishnaiah & Rao, 1988: p. 19). In soil survey practice it is important to distinguish non-soil from soil, because usually only locations identifiable as soil are of interest for the study. Farmyards, ditches, and roads, are examples of items considered as non-soil. What in a particular survey should be considered as non-soil has to be specified explicitly, in short, all those areas which do not belong to the population of interest. The specifications of soil and non-soil may differ between surveys, since the population of interest may differ.

  The definition of the population of interest must be usable in practice. The surveyor must be able to decide in the field, without much hesitation, whether or not an element belongs to the population (Cochran, 1977: p. 5).

  Case A: aggregate of all possible augering locations identifiable as soil in the survey

region of this case.

Case B: similar to case A.

- *Method of determination.* The method of determination specifies how the values of the target variable are determined for given sampling elements, i.e. the method of measurement, observation or estimation in the field, laboratory analysis, and sometimes model calculations using co-variables. It often occurs that one or more co-variables, correlated with the target variable, are measured instead of the target variable, because they are cheaper and easier to determine. Values of the target variable are then to be estimated from the data collected on the co-variables.

  Case A: P_{rel} at sample points is related to the content of oxalate-extractable Fe and Aℓ, the density of the soil, P, and the depth to the MHW. The content of oxalate-extractable Fe and Aℓ, and P are determined by laboratory analysis; information on the density of the soil is based on literature (previous survey). The groundwater level in auger holes should be measured the day after augering. These values should be compared with those in reference tubes (i.e. with known values of MHW) measured on the same day. Values of the MHW in the survey region can be derived from this comparison. A regression model should be used to calculate P_{rel}.

  Case B: values of the MHW at sample points are based on field estimations related to profile and field characteristics. These estimations should be corrected by comparing measurements of the groundwater depth at 18 auger points with measurements of the groundwater depth in reference tubes (with known values of MHW) nearby; both are measured at the moment the water table in the reference tubes is near MHW level. The MHW value of a sampling point can be estimated by linear regression of the measurements at that point on those of a reference point with known values of MHW (e.g. Van der Sluijs & De Gruijter, 1985).

- *Sampling design.* The sampling design is a mathematical function assigning a probability of selection to every possible sample. A *sample* is a list of sampling elements to be observed. Depending on the design, sampling elements may occur in this list several times. The *sample size* is the number of components in the list. If this is fixed and pre-determined then it is implied by the sampling design, as any list of a different size will be assigned probability zero making up the sample. Different classes of designs can be distinguished, for example simple random sampling, stratified sampling and cluster sampling (e.g. Cochran, 1977). All these classes can be subdivided into more specific designs. Each design class has characteristics of its own, for example concerning its usability under specific conditions, or its applicability to answer a particular request (see Chapter 6).

  Case A: Stratified sampling with simple random sampling in each stratum. Strata: seven combinations of map units on the national soil map 1:50 000, combined with land use categories (arable, grass) and drainage areas. Total number of strata: 26. Design within strata: simple random sampling with replacement and with equal probabilities. Allocation to strata: proportional to size, however, twice as many in strata defined as 'wet', and at least two per stratum. Total sample size: 116.

Case B: Stratified two-stage sampling with simple random sampling in both stages. Strata: map sheets. Design within strata: two-stage with simple random sampling in both stages. First stage: random selection of two map delineations per stratum with replacement and probabilities proportional to size (i.e. area of delineations). Second stage: random selection of four points per selected map delineation by simple random sampling with replacement and with equal probability. Total sample size: 264.

- *Sampling frame.* A list of all sampling elements in the population used to select elements to be sampled (the *sample points*) is referred to as a sampling frame (Krishnaiah & Rao, 1988: p. 21). This list should enable selection of elements from sub-populations (e.g. strata), or selection of sub-populations (e.g. primary units) if this is prescribed by the sampling design. Efforts should be made to find or construct a sampling frame which fits the design. Sometimes a design requires more than one frame, e.g. a two-stage design may require different frames for selections in the first and second stages.

  The term *list* is to be taken in a broad sense: it may be an enumeration of sampling elements, i.e. a list in the literal sense, or it may be a map of the survey region containing all elements of the population. Nowadays the sampling frame is often available in machine readable form, for example stored in a database, or in a GIS database.

  The sampling frame should correspond as well as possible to the population of interest. In soil survey practice, however, the frame often contains elements defined as non-soil, and therefore not belonging to the population of interest. If the frame contains elements of which the non-soil status can only be established in the field, there should be instructions on how to act when such elements are encountered.

Case A: an overlay of a soil map, scale 1:50 000 on a land use map was used to select the sample points. A topographical map was used in the office to check whether the selected elements were located on agricultural land (and not on roads, farmyards etc.).

Case B: first stage: for each stratum a list of all map delineations belonging to map unit Hn21-VI with their areas; second stage: cartographic representations of the selected map delineations.

- *Selection technique.* The selection technique is the operational method by which sampling elements are selected to be included in the sample, with predetermined probabilities according to the sampling design. Computerized selection techniques utilize random number generators to select for example the primary units, and the co-ordinates identifying the elements to be included in the sample. Generally, selection according to a given design can be realized by different techniques, which may vary in operational usefulness.
- *Instructions for field work.* As stated before, the sampling frame is often imperfect. Therefore, instructions should be given on how to act if sampling elements appear to be located in non-soil. Furthermore, instructions are desired concerning situations in which a sample point is inaccessible (e.g. because of crops). The ways to register these points (coding) and those of which the values are outside the range of measurement, need to be established before the field work starts. If other difficulties are anticipated, the schemes should include instructions on how to cope with these as well.

The **method of inference** consists of the method of estimation of the target quantity (*estimator*) and the procedure to quantify the accuracy of the estimator from the sample data. In the case of classical sampling theory the procedure to quantify the accuracy coincides with the variance formula (which is related to the method of estimation). The statistical method of inference is related to the sampling design. If the survey aims at common target quantities and if standard types of designs are used, the inference methods are readily available from statistical handbooks.

Sometimes an estimate can be improved by means of an *auxiliary variable* (e.g. Krishnaiah & Rao, 1988: p. 26), correlated with the target variable. In such cases the target quantity is for instance estimated by a ratio or a regression estimator.

Case A: standard formulae for stratified sampling with simple random sampling in each stratum. The proportion (i.e. the target quantity) of the region where the soil should be considered as being phosphate saturated can be estimated from information on the phosphate saturation at sample points (see the method of determination). Proportions are estimated in the same way as spatial means of quantitative variables.

Case B: standard formulae for stratified two-stage sampling with simple random sampling in both stages.

The **sample** is the random result from applying the selection technique to the sampling frame. It consists of a list of the sample points. In a soil survey these sample points may be represented as numerical co-ordinates or as points on a map.

The **prior evaluation** shows the predictions of both accuracy and cost of the scheme proposed (see Chapter 7). The prior evaluation of the accuracy is based on the sampling design, the method of determination, the procedure to quantify the accuracy, and the prior information on spatial variability. The cost of a scheme can be roughly predicted from the sampling design, the method of determination, and information on the survey region.

Although the framework concentrates on soil survey using probability sampling, i.e. using classical sampling theory, part of it may be applicable to soil survey in general or to other spatial sample surveys. Other types of soil survey will probably need additional concepts for their formal description.

Chapter 5

Problems in designing soil survey schemes

Parts of this chapter have been published in:

Domburg, P. & Elzas, M.S. (1994)
Structuring the Domain of a Complex system: a basis for a knowledge-based system supporting soil survey design. In: Beulens, A.J.M., Doležal, J. & Sebastian, H-J. (Eds.), Optimization-Based Computer-Aided Modelling and Design, Proceedings of the second Working Conference of the IFIP TC 7.6 Working Group, Dagstuhl, Germany, 1992. Leidschendam, Lansa Publishing, pp. 181-195.

5 Problems in designing soil survey schemes

5.1 Scope

Some obvious problems during the design of soil survey schemes were mentioned in the first chapter: a structured approach is lacking, the accessibility of existing information is often insufficient, information on spatial variation and on accuracy of soil information is limited, the available time for design is limited, and general procedures to evaluate schemes are lacking. A further analysis of these and other problems arising during this particular design process is based on interviews with an experienced statistical consultant and on a description of historical cases.

The model of the design process introduced in Section 4.4 is the starting point for the problem analysis (Section 5.2). This process model is primarily based on current practice as became apparent from the interviews. Further analysis of the problems in the steps distinguished in the process model, also considering the historical cases (Section 5.3), resulted in a specification of the tasks to be supported (Section 5.4). These tasks are not related one-to-one to the steps in the process model. Some shifts were necessary to distinguish clear, non-overlapping tasks. This tasks structure made it possible to specify the role of the different disciplines involved in the development of the system (Section 5.5).

5.2 Problems in the design process

In the model of the design process the following steps were distinguished (Section 4.4):

- define the request, i.e. aim and constraints, and select adequate prior information;
- construct an outlinear plan of action, including preliminary choices of type of sampling design and method of determination;
- specify the statistical method of inference;
- make tentative predictions of accuracy and cost, using prior information;
- compare the predictions of accuracy and cost with the original request;
- elaborate the final scheme;
- return to another possible outlinear plan of action in step two, or revise the request and/or provide additional prior information and return to step two.

The decision problems encountered during these steps are discussed in the following subsections. The decision problems are evidently related to the concepts defined in the conceptual framework (Section 4.5). Before starting the analysis an overview is given of the use of concepts in the design steps (Table 5.1). The first time a concept arises it needs to be specified, thereafter its specified instantiation is usually used, or adapted, in one or more of the subsequent steps, and finally reported. The table shows that some of the steps distinguished need only a few concepts, and the presence of concepts in some other steps is nearly completely overlapping.

Table 5.1 Use of concepts in the design steps

010612

Concept		Step in the design process						
		1	2	3	4	5	6a	6b
Aim	Target quantity	s	u	u			r	
	Target variable	s	u				r	
	Survey region	s	u				r	a
Constraints	Accuracy	s	u			u	r	a
	Cost	s	u			u	r	a
	Logistics	s	u			u	r	a
Prior information	Spatial variability	s	u		u		r	a
	Other geographical information	s	u		u		r	a
Plan of action	Sampling element	s	u				r	a
	Population	s	u				r	a
	Method of determination		s		u		r	a
	Sampling design		s	u	u		r	a
	Sampling frame	s	u		u		r	a
	Selection technique				s / a + u		r	
	Instructions for field work						s + r	
Method of inference	Method of estimation			s / a			r	
	Procedure to quantify accuracy			s / a	u		r	
Sample							s + r	
Prior evaluation	Prediction of the accuracy				s / a	u	r	
	Prediction of the cost				s / a	u	r	

Steps in the design process:

1 Define aim and constraints, and select adequate prior information;
2 Construct an outlinear plan of action (method of determination + type of sampling design);
3 Specify the statistical method of inference;
4 Make tentative predictions of accuracy and cost, using prior information;
5 Compare the predictions of accuracy and cost, with the original request;
6a Elaborate the final scheme;
6b Return to another outlinear plan of action in step two, or revise the request and/or provide additional prior information and return to step two.

Legend:

s = specify or define a concept for the first time;
u = use an earlier specified or defined concept;
a = adapt earlier specified or defined concepts;
r = report a concept;
+ = and;
/ = or.

5.2.1 Aim, constraints, prior information

The three aspects that need to be dealt with in the first step, which together constitute the input of the system, are not put in arbitrary order. First, the request, i.e. aim and constraints, needs to be defined, and second, adequate prior information should be searched. This selection of prior information is guided by the definition of the request. The decision problems often already start at the definition phase.

Define request

Researchers or people who commission surveys often find it difficult to distinguish between different types of requests: *how much*, *where*, or a hybrid type comprising both *how much & where*? Since traditionally the common aim of a soil survey was to produce a map, it is often assumed that the survey outcome should be a map, e.g. a map that depicts the phosphate saturated areas in the survey region. However, sometimes a numerical type of result will suffice, e.g. the areal proportion of the survey region which is to be considered as phosphate saturated. The objective for which the survey results have to be used determines the type of result required. If the aim is defined incorrectly the survey results may be inappropriate, and effort and money wasted. Researchers could profit from an overview of the applicability and consequences of different sampling approaches.

As stated in Sub-section 2.2.3, the system will support single criterion requests. In practice several soil properties are usually observed or measured in the same survey. The user of the proposed system should specify the most important property as the target variable.

The definition of the survey region is determined by the objective of the soil survey project as a whole. The size of the survey region may need to be limited because of constraints, e.g. a large region requires more time and budget to be surveyed than a smaller region.

The constraints can often more easily be specified than the aim. The user should be forced to specify the constraints at the beginning of the design process since this will limit the solution space, and so influence the progress of the design process. There may be constraints concerning accuracy, cost, and logistics. The user may specify a demand for a minimum accuracy, or a maximum sampling variance. This will influence the number of applicable types of sampling designs and the sample size. In the case of a *how much* request it should also be specified at an early stage whether the accuracy of the survey results should be quantifiable from the survey data alone. If this is required, any systematic sampling is to be excluded. The available budget for the project as a whole is often fixed; consequently only a limited budget for spatial inventory will be available: a financial constraint. Logistic constraints are, for example, the available time for field work or the laboratory capacity for chemical analysis.

Select prior information

In the previous chapter two main groups of prior information were distinguished: spatial variability, and other geographical information. Prior information on the spatial variability of the target variable, or a co-variable, in the survey region is required to decide on the distribution of observation points over the survey region, e.g. in a region or sub-region with little variation an additional observation point provides less information than an additional observation point in a region or sub-region with a lot of variation. Furthermore, this prior information is needed to predict the accuracy of survey schemes; this will be shown in Chapter

7. Geographical information, e.g. a soil map, can be used to decide on a type of sampling design, e.g. using map units as strata. When, later on in the design process, observation points are to be selected, prior information that can be used as a sampling frame should be provided. This may, for example, be derived from a GIS database.

At present, selection of appropriate prior information is often difficult. This is on the one hand due to the fact that the existing information is scattered over several data sources and not easily retrievable. Another problem is that there is little quantitative information on the variability of soil properties and on the accuracy of survey results. So, besides the availability, the quality of the information is a problem.

5.2.2 Outlinear plan of action

An *outlinear plan of action* is outlinear in the sense that, at an early stage of the design process, only tentative decisions are made on some of the elements which need to be specified in the final plan of action. Two elements on which at an early stage preliminary decisions are made are the method of determination and the sampling design.

The number of possible methods to determine a particular soil property is generally limited. One method is often the favourite beforehand, e.g. when it conforms to standards. In other cases, the price and accuracy of the method of determination are important to the choice of an appropriate method. Information on the accuracy of methods of determination is not always quantitative, but when there are a number of possible methods, the relative accuracy can often be assessed. Both information on cost and on accuracy should be taken into account when selecting an appropriate method.

To be able to design outlinear plans of action it should first be considered whether a classical sampling approach or a geostatistical approach is most appropriate. As explained in the first chapter this depends on the type of request and the type of result required. If the first approach is adopted, applicable types of sampling designs and methods of determination need to be selected. These selections should be guided by the defined aim and constraints, and by the selected prior information. If, for example, the survey aims at determining the mean organic matter content of the top-soil in sub-regions with different types of agricultural use, land use type can be used as a criterion for stratification and a stratified sampling design can be applied. If the geostatistical approach is adopted, decisions are also needed on the method of determination and on the method of data collection. Provisionally the system will not assist in the use of geostatistics (Sub-section 2.2.1).

The search for sampling designs is based on the definition of the request and the prior information available, and requires decisions on the sampling element and on the population. At present, selection of applicable types of designs is often determined by the experience of those involved in survey design. Time to call in experts or to consult statistical handbooks is generally very limited. The availability of procedures for sample selection and/or for analysis of results is generally also a decisive factor. Thus, the choice of types of designs is partly influenced by practical conditions. It may therefore occur that more appropriate types of designs are not considered.

Combinations of selected methods of determination and selected types of designs which seem to cause no conflicts with the definition of the request result in outlinear plans of action. Initially there are several possible plans, from which one has to be selected during the further design process.

5.2.3 Method of inference

At present it sometimes happens that specification of the method of inference is postponed to the analysis phase, after data collection. It may then appear that the way in which the data were collected was not the most appropriate. As noted earlier (Section 4.5), the selection of the statistical method of inference should be guided by the sampling design. If complicated sampling designs are constructed, or if unusual target quantities are to be determined, the corresponding method of inference may not be directly available, but may need to be developed *ad hoc*. In any case, the specification of the method of inference should be linked with the design of an outlinear plan to assure that the consequences of a particular sampling design on the possible analysis methods are considered in time. The combination of these steps can be called the design of *outlinear survey schemes* instead of outlinear plans of action. A survey scheme is called outlinear as long as not all elements which need to be specified in the final scheme are made explicit. At an early stage tentative decisions are only made on the method of determination, the sampling element, the type of sampling design and the corresponding method of estimation. During the design process the scheme is further elaborated and final decisions are made.

5.2.4 Prediction: accuracy, cost

In the present situation requirements specified in advance for the accuracy of survey results are rare. In general, no quantitative prediction of the accuracy is given either. Researchers generally proceed pragmatically in survey design by taking into account the results obtained in comparable surveys in the past, by demanding a minimum number of observation points per sub-region to assure reasonable accuracy, or by taking the maximum allowable budget as a constraint and accepting the resulting accuracy, whatever that may be.

The costs of soil survey projects are currently roughly assessed, and the cost of spatial inventory is generally not specified separately. The relations between types of sampling designs and inventory cost are never made explicit, which hampers comparison of sampling designs.

There is a general lack of detailed information on the economic side of spatial inventories. To be able to assess a proposed survey scheme there need to be procedures to predict accuracy and cost .

5.2.5 Prior evaluation

The predicted accuracy and cost need to be compared with the defined aim and constraints (prior evaluation). At present, there are no objective evaluation procedures in use to predict the accuracy and cost of schemes. If there were such procedures, it might even be possible to search for an optimal scheme. In this study such procedures have been developed (Chapter 7).

5.2.6 Revising: aim, constraints, prior information

If the evaluation and possible optimization does not result in a scheme which satisfies the original aim and constraints, another outlinear scheme should be evaluated. If none of the outlinear schemes results in a satisfactory scheme the options for revising the original input, i.e. the aim, the constraints, and the prior information, should be considered. It is illogical to change the target quantity or the target variable, because this severely changes the aim

of the survey so that revision would result in another survey. It may, however, be acceptable to reduce the size of the survey region. It may also be possible to relax the constraints, e.g. by increasing the budget, or to provide additional prior information enabling a better survey scheme to be constructed. The time and effort it takes to provide this information may be economically worthwhile. The adapted request and additional information are the starting point for a new run through the design process.

5.2.7 Soil survey scheme

If an outlinear scheme is designed which is efficient with respect to accuracy and cost, the final scheme can be elaborated. A report has to be generated describing the decisions of the previous steps, including, if necessary, specific instructions for field work. Such instructions should prescribe, e.g. how to cope with observation points that appear to be located in non-soil, or values of the target variable which are outside the range of measurement. It should be considered which possible problems require specific instructions for field work. Experience from historical surveys can be very useful when specifying such instructions.

The reports of historical surveys have often turned out to be incomplete, and a justification for the final scheme is hardly ever reported. Therefore, it is difficult to utilize knowledge from these surveys for future projects. This problem can be overcome by generating automatically structured reports of soil survey schemes in the course of using a structured, computer supported, procedure to design soil survey schemes.

5.3 Problems in historical cases

This section considers the problems in the two historical cases introduced in the previous chapter. Problems of choice encountered in these cases are not reported anywhere. People involved in these surveys have contributed to reviews of the surveys which made it possible to recover the main decision problems and conditions that hampered the design of the survey schemes.

5.3.1 Case A

The aim of spatial inventory was specified as determining the areal proportion of the survey region which is saturated with phosphate. Besides an overall value of the saturated areal proportion for the whole survey region, specific information was required on the situation in a number of sub-regions, which were based on combinations of land use categories (arable, grass) and drainage areas. So, the request was of the *how much & where* type. Specification of relevant and useful sub-regions required some discussion. The number of sub-regions had to be limited to assure sufficiently accurate results per sub-region.

However, before the discussion on the sub-regions could start, the survey region itself had to be defined. The choice of the Ootmarsum region as survey region in the province of Overijssel was not predetermined. From the beginning it was assumed that both the aim and the observation density had to be the same as in another historical survey on phosphate saturation, i.e. one observation per 20 ha (Breeuwsma et al., 1989). Given the budget and this observation density, the allowable size of the survey region was determined. It might

have been a point of discussion whether the specified aim was fully in accordance with the objective of the commissioner and whether this observation density was actually required in this survey. The aim had been specified before a statistical advisor was consulted. This may have resulted in too little effort being spent on interpreting the objective of the commissioner, who was probably more interested in the relative mass of phosphate sorbed by soil in the whole province. If so, a lower observation density would have made it possible to take a sample of the whole province instead of sampling sub-regions with a higher observation density. Furthermore, the survey has also been used to test a procedure that was being developed to predict the relative mass of phosphate sorbed by soil generally in regions. This goal of researchers at the institute may have influenced the way in which the aim of the survey was specified, and has probably also resulted in the conclusion that an observation density of one point per 20 ha was necessary. In retrospect, it is not easy to discover precisely what the justification for the specification of the aim was. It is, however, important to realize the general point here namely that an incorrectly defined aim tends to produce inappropriate survey results, and that sufficient effort should be spent to clarify the aim.

The main constraint was budgetary which strongly influenced both the time and number of observations permitted. The accuracy had to be maximal for the available budget and quantifiable from the sample data. Since wet areas are more sensitive to phosphate leaching than drier regions, accurate information was especially required on the wet parts of the survey region. This had to be taken into account when allocating the observation points. Due to the lack of procedures for evaluating survey schemes in advance, the efficiency was only roughly assessed. It was unclear whether the efficiency could have been improved by (minor) changes in the final scheme, e.g. an other type of sampling design, sample size, or method of determination.

Useful prior information was collected from different sources, e.g. information on spatial variability from a comparable project, from the national soil map and from a topographical map, scale 1:50 000. A land use map had to be produced in advance to be able to distinguish between grassland and land for maize. For these two land use types special estimates were required.

The type of sampling design, i.e. stratified sampling with simple random sampling in each stratum with replacement and equal probabilities, was chosen because of the experience with this type of design and because of the relatively simple statistical method of inference. No other types were considered. Experience from a comparable survey in the past was used to determine combinations of map units for stratification, however, the discussion on the stratification took quite some time and, retrospectively, the final outcome could not be fully justified.

The main decision problem with respect to the choice of the method of determination was how to decide on the procedure for estimating the depth of the MHW on sample points. The final choice was mainly influenced by the available budget and the available time. A more accurate method would have cost too much time (and money) or the number of observation points would have had to be drastically reduced. However, even a smaller sample size combined with a more accurate method of determination might be more efficient, i.e. provide more accurate results by the same type of sampling design.

A computer program was used to select the sample, according to the sampling design,

from a GIS database (in ARC/INFO). The topographical map was used to check manually at the office whether selected points belonged to the population, or were unfortunately located in non-soil, e.g. on roads, in ditches. Spare points were selected for points that turned out to be located in non-soil.

In summary, the main problems of survey design in case A were:

- a lack of clarity in specifying the aim;
- limited availability of knowledge and prior information;
- a lack of procedures for prior evaluation and optimization of schemes.

5.3.2 Case B

This survey project aimed at answering a *how much* request, namely it aimed at determining the spatial mean of MHW in map unit Hn21-VI of the national soil map of the Netherlands, scale 1:50 000. This map unit is found scattered all over the country: from north to south. The main constraints were the available capacity of personnel and the budget. A preliminary estimate of the maximal allowable sample size was deduced from these conditions, and influenced further survey design.

Map sheets of the national soil map of the Netherlands with their corresponding notes were used as prior information. In addition statistical soil survey knowledge based on experience of soil surveyors was used to design a survey scheme.

It took a great deal of discussion to decide on the type of sampling design. The selection of the possibilities discussed was based on the knowledge and experience of the people involved. At an early stage it was decided to stratify. The choice of stratification, namely to use map sheets as strata, was based on operational advantages: information on areas of map units was easily available for map sheets, and the use of map sheets facilitated the procedure for sample selection. The final design was stratified two-stage sampling with simple random sampling in both stages. The use of map sheets as strata resulted in observation points located all over the country; the two-stage approach within strata effected groups of observation points within a stratum. In the first stage two map delineations were selected with probabilities proportional to size and with replacement, i.e. one map delineation could be selected several times. In the second stage four points were selected within selected delineations by simple random sampling with equal probability. The geographical concentration of groups of observation points caused by this design seemed attractive from an operational point of view. The final sampling design could not be objectively compared with other possible designs, due to the lack of procedures for evaluating accuracy and cost.

With respect to the choice of a method to determine the value of the MHW, again, as in case A, considerations of accuracy and cost had to be weighed against each other.

A computer program was developed to assist in:

- processing prior information on (combinations of) map units that were to be used as strata (e.g. number of units, total area);
- storage and retrieval of sample data in a database;
- selecting data from sub-populations of observations in a sample and analysing these data statistically.

The main problems in case B were:

- the limited availability of knowledge and information, in particular on the application of various types of sampling designs to soil surveying;
- a lack of procedures for prior evaluation (and optimization) of schemes.

5.4 Tasks to be supported

The analysis of problems in designing survey schemes resulted in the identification of six main tasks, the support of which will facilitate the design process. These tasks should not be considered as a fixed succession: most tasks are related to several preceding tasks and the whole design process is iterative. Figure 5.1 shows the main tasks during the design of soil survey schemes with their mutual relations represented by arrows. When these tasks are supported, the most significant general problems in the design of survey schemes will be overcome.

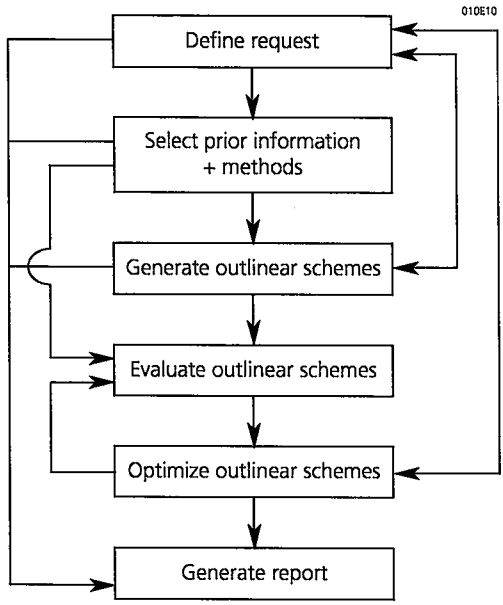


Figure 5.1 Main tasks during the design of soil survey schemes with mutual relations

Definition of the request for soil survey
There is a relation between the type of request and the applicability of sampling strategies. At the start of a consultation the user of the system should become aware of the possible types of results of surveys, ranging from a single value of the target quantity for the whole region to a map of the survey region, which displays information on points in the region. The last type of result usually requires much more effort than the first. A proper specification of

the request at the start of the design process helps to determine whether classical sampling theory or geostatistics is most appropriate. It can limit the number of applicable types of sampling strategies, i.e. restrict the solution space. At the start of the design the user should also be forced to specify the constraints as clearly as possible, since the constraints are also decisive for the further course of the design process.

Although the request should be defined with due care, one should be allowed to adapt the original input at a later stage if no satisficing scheme can be found, e.g. the survey regions may be redefined, or it may be possible to enlarge the budget.

Selection of prior information, methods of determination and types of sampling designs
In Fig. 5.1 this task is mentioned for short: Selection of prior information and methods. After the definition of the request several selections need to be made. Two steps can be distinguished in this task: (i) selecting relevant prior information, (ii) selecting applicable methods of determination and types of sampling designs.

Initially, prior information on spatial variability or other geographical information, e.g. soil maps, should be selected. Therefore, the storage and retrieval of this information needs to be improved so that a computer can assist in this selection. Attention should also be paid to updating this information continuously by storing results from surveys to which the proposed system has contributed. The importance of this information to the design process has already been stressed in Sub-section 5.2.1.

Thereafter, applicable methods of determination and types of sampling designs need to be selected, taking into account the prior information selected. This selection restricts the space of possible solutions. For a given soil property of interest (the target variable), the number of possible methods of determination is generally limited. When selecting appropriate methods the cost and accuracy of these methods need to be taken into account. Whether this information should be stored in a computer system or provided by the user will be discussed in the next chapter.

Statistical knowledge and knowledge on how to make a proper selection of types of sampling designs for a given request has to be well-organized in the system. If in the future a computer system assists in the selection of applicable types of sampling designs, selection of samples and analysis of the results may also need to be supported by a computer system (not necessarily the same system). Otherwise, the system might give rise to a bottle-neck at the execution stage of a soil survey project, which would reduce the advantage of assistance in the design stage. A selected sample should be part of the final scheme, and analysis of data collected according to an efficient sampling design should not be too complicated for the user. In a first prototype of the system only the most common types of sampling designs need to be available.

Generation of outlinear schemes
Possible outlinear schemes have to be constructed from the selected types of designs and methods of determination using the available prior information and taking into account the defined request. At an early stage, a number of possible outlinear schemes should be considered; the choice for one particular scheme can be postponed to a later step in the design process.

Evaluation of outlinear schemes
To enable objective comparison of schemes models have been developed to predict the accuracy and the cost of survey schemes. Therefore, the number of sampling elements in the sample needs to be specified. The models for prior evaluation take into account the characteristics of various sampling designs. Information on spatial variability and on various cost components is required to enable prior evaluation (see Chapter 7).

Optimization of outlinear schemes
Procedures are required to assist in the search for an efficient scheme, if possible an optimal scheme, which maximizes the accuracy for a given budget or minimizes the cost of operation for a given accuracy constraint. These optimization procedures can be directed by the constraints and the models for prior evaluation (see Chapter 7). When the results of the optimization of an outlinear scheme are acceptable for the user, the main elements of the final survey scheme are known.

Report generation
Finally, the instructions for field work need to be specified and the sample needs to be selected. Then, the final scheme, complete with justifications for the decisions made, has to be presented in a report using the results of previous tasks. This report should not only serve as a clear guide for the further course of the survey, but may also be useful when providing future support. It should therefore be stored properly.

The use of concepts in the tasks distinguished is depicted in Table 5.2. This table shows that every task needs a number of concepts, and the same concept may have different roles in a task during the design process. This table gives a more compact image than Table 5.1.

In addition to these tasks the system may assist in one more task in a soil survey project: evaluation *a posteriori*. Survey projects should be evaluated afterwards with the help of the system itself enabling continuous collection and storage of new knowledge. So, re-use of experience from the history of surveys can be improved and the system can contribute to its own maintenance: a simple form of a *self-learning* system. Furthermore, as stated above, selection of samples and processing of collected data should also be supported by a computer program (not necessarily the same system).

After the tasks to be supported are specified, the knowledge required to perform these tasks needs to be structured. Some tasks can be performed using existing knowledge, e.g. definition of the request, selection of prior information and methods, and generation of possible schemes. For other tasks knowledge has to be generated, e.g. for evaluation and optimization of outlinear schemes. The following chapters elaborate on the ways in which these tasks should be supported.

Table 5.2 Use of concepts for the tasks

010E13

Concept		Task					
		D	S	G	E	O	R
Aim	Target quantity	s	u	u			r
	Target variable	s	u	u			r
	Survey region	s / a	u	u			r
Constraints	Accuracy	s / a	u	u		u	r
	Cost	s / a	u	u		u	r
	Logistics	s / a	u	u		u	r
Prior information	Spatial variability		s / a	u	u	u	r
	Other geographical information		s / a	u			r
Plan of action	Sampling element	s / a	u	u			r
	Population	s / a	u	u			r
	Method of determination		s / a	u	u	u	r
	Sampling design		s / a	u	u	u	r
	Sampling frame		s / a	u	u		r
	Selection technique				s / a + u		r
	Instructions for field work						s + r
Method of inference	Method of estimation		s / a	u			r
	Procedure to quantify accuracy		s / a	u	u	u	r
Sample							s + r
Prior evaluation	Prediction of the accuracy				s / a	u	r
	Prediction of the cost				s / a	u	r

Tasks:

D Define request;
S Select prior information, methods of determination, and types of sampling designs;
G Generate outlinear schemes;
E Evaluate outlinear schemes;
O Optimize outlinear schemes;
R Generate a Report of the final soil survey scheme;

Legend:

s = specify or define a concept for the first time;
u = use an earlier specified or defined concept;
a = adapt earlier specified or defined concepts;
r = report a concept;
+ = and;
/ = or.

5.5 Input from various disciplines

The development of the support system requires a multi-disciplinary approach with contributions from the field of soil science, statistics, computer science, and operations research. The formal description of the domain and the specification of tasks made it possible to specify the role of the various disciplines. Table 5.3 shows the relations between tasks and the contributions of the various disciplines.

Pedological and statistical knowledge are the basic types of knowledge in the system. Operations research may provide techniques to support the search for an optimal scheme. It is obvious that for the development of a KBS computer science will be used to support each distinct task. The system aimed at here requires that a number of techniques from the field of computer science are used, the most important being:

- AI techniques for: the inference mechanism, the structure of the knowledge base, the user interface, and the self-learning mechanism;
- database techniques for: storage and retrieval of information on soil survey projects;
- GIS for: storage and retrieval of spatial data on survey regions.

Table 5.3 Relations between system tasks and contributions of disciplines

010E14

Discipline	Task					
	Define request	Select prior information + methods	Generate outlinear schemes	Evaluate outlinear schemes	Optimize outlinear schemes	Generate report
Soil Science	X	X	X			X
Statistics	X	X	X	X		
Computer Science	X	X	X	X	X	X
Operations Research					X	

Chapter 6

Knowledge about methods of determination and statistics

Parts of this chapter have been submitted for publication:

Domburg, P., Gruijter, J.J. de & Beek, P. van (submitted)
Designing Efficient Soil Survey Schemes with a Knowledge-Based System using Dynamic Programming. Environmetrics.

6 Knowledge about methods of determination and statistics

6.1 Outline

The third research question in Section 1.5 was related to the availability of knowledge in the system. This chapter deals with the knowledge required to design an outlinear scheme, including knowledge about methods of determination and statistical knowledge. The main elements to be specified in an outlinear survey scheme are: the method of determination, the sampling element, the type of sampling design and the corresponding estimator of the target quantity.

Section 6.2 discusses current knowledge about methods of determination and how appropriate methods should be selected. The sampling element is generally related to the method of determination. Thereafter, Section 6.3 deals with the statistical knowledge required to select an appropriate statistical approach and, when a design-based approach is applicable, to select possible types of sampling designs. Therefore, the required statistical knowledge needed to be structured. This chapter ends with some remarks on the design of outlinear survey schemes (Section 6.4).

6.2 Methods of determination

Data on the soil property of interest (the target variable) can be obtained by observation, by measurements in the field or in a laboratory, or can be derived from data on other soil properties. There are often several possible methods of determining the value of a soil property, which differ in accuracy and cost.

Sometimes, values of the soil property of interest can be estimated in the field. In general, such field estimations are relatively cheap because no expensive equipment is needed and little time is needed. Disadvantages of field estimations are that they may depend on the surveyor, and that the accuracy is rarely specifiable. Nevertheless, in traditional soil-mapping surveys field estimations are frequently used as the method of determination because no quantification of accuracy of the survey results is required and because it saves time and money. For example, in general the clay content and organic-matter content in a soil profile are estimated. The error in the estimations of a particular surveyor can be quantified by comparing his estimates with the results of laboratory analysis; however, there may be differences in estimated values and in errors in these estimations between various surveyors. For environmental surveys, e.g. for soil protection and soil sanitation, field estimations are rarely used to estimate the concentration of pollutants, e.g. the phosphate content, or the cadmium content. Then, instead of estimating in the field, samples are taken which are analysed in the laboratory. This is due to the fact that no reliable methods are available to accurately estimate such target variables in the field, and because the accuracy of specific results is often required. These surveys, on which risk assessments are often based, require accurate and reliable (or reproducible) results. Laboratory analysis is more expensive than field estimation, but advantages of laboratory analysis are that the accuracy can be more

accurately quantified and the procedures are objective and reproducible.

There are a wide range of soil properties which can be determined in surveys, but for a particular property the number of feasible methods of determination is often limited. Researchers are often well-acquainted with available methods of determination. It is, therefore, impractical to store all these methods in a system. The proposed system will mainly assist in statistical aspects in the design of soil survey schemes, and will provide limited assistance in selecting the most appropriate method(s) of determination. If information from previous surveys is stored in the system, the system may provide the user with information on methods used in comparable surveys in the past. Then, the user should decide whether this method is appropriate for the current survey and, if so, check whether the information about this method is still up-to-date. Although the selection of methods of determination requires very specialist knowledge, and certainly should be given thorough attention, the user will be responsible for providing information on applicable methods of determination. This decision does not imply that the selection of methods of determination is less important. Both accurate sampling as well as analysis are important in soil surveying. Assuming that the selected methods are unbiased, i.e. they result in correct answers on average, and that the measurement error does not depend on the level of the measured values, the system will be able to evaluate possible methods where needed. Therefore, information on accuracy (preferably in the form of the standard error) and cost of appropriate methods is also required (see Chapter 7).

Selection of appropriate methods should be guided by the aim of the survey, the purpose for which the results will be used, the accuracy required, and the budget and time available. It is often the case that a method that conforms to standards is the favourite beforehand, leaving other possibilities out of consideration. However, it should be noted, as Barcelona (1988) remarks, that an analytical method developed and validated for a particular purpose is not automatically applicable to surveys of a different nature, and research projects may also require the use of more specific and particularly sensitive types of analytical procedures than those in 'standard' references.

The choice of a method of determination generally influences the definition of the sampling element, since this is the object on which the method of determination operates. For example, if an auger hole (in the field) or a soil sample (to be analysed in the laboratory) must meet the requirements of the method of determination, the volume of the sample is important.

The selection of methods of determination in the two cases is discussed below.

Case A. *The aim was to estimate the areal proportion of soil saturated with phosphate. A regression model was used to calculated the relative mass of phosphate sorbed by soil. This model required values for the content of oxalate-extractable Fe and Al and for the areic mass of phosphate sorbed by soil (P) at sample points, and values of the density of the soil and the MHW at these points. The content of oxalate-extractable Fe and Al, and P were determined by laboratory analysis and therefore a composite sample was taken over depth to MHW at each sample point. The sampling element was a standard augering to MHW, with a maximum depth of 1 metre.*

It is also possible to measure the maximum areic mass of phosphate sorbed by soil (P_{max}) of samples in a laboratory, and to use these values to determine the relative mass of phosphate sorbed by soil (P_{rel}). This laboratory analysis, however, is much more expensive and therefore a regression model has been used for which the necessary laboratory analysis

was cheaper.

The information required on the density of the soil was based on the results of historical surveys. The day after augering, the groundwater level in auger holes was measured and values of the MHW were estimated by comparing these values with those in reference tubes (i.e. with known values of MHW) measured on the same day. Other methods of determining values of MHW which provide more accurate results would have been too time-consuming.

***Case B**. The aim was to estimate the spatial mean of the MHW in map unit Hn21-VI of the national soil map of the Netherlands. The values of the MHW at sample points were based on field estimations related to profile and field characteristics. These estimations were corrected by comparison with measurements of the groundwater level in 18 tubes; the values in these tubes were compared with the groundwater level in reference tubes (with known values of MHW) in the neighbourhood; both measured at the moment the water table in the reference tubes is near MHW level.*

The time required and the cost of possible methods to determine the MHW at sample points have influenced the choice of this method of determination. The sampling element was a standard augering to the MLW, with a minimum depth of 1.5 metre. (This choice was also related to an interest in some other soil properties.)

6.3 Statistical knowledge

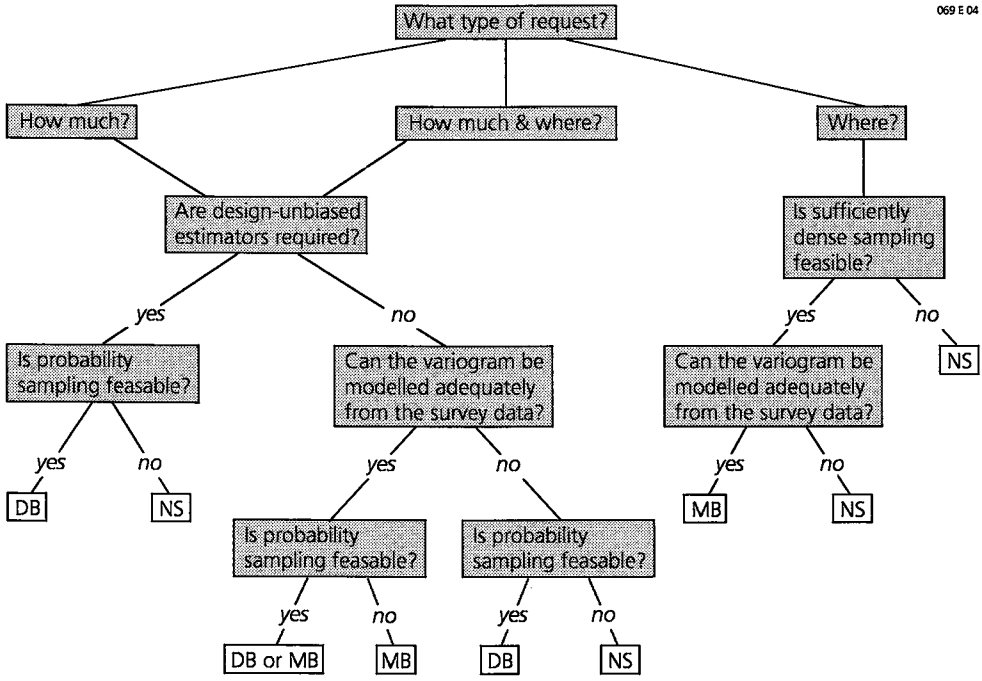
The type of sampling design is an important element in the outlinear survey scheme. Before this can be selected, it should be clear whether a design-based approach is appropriate (Sub-section 6.3.1). To facilitate the selection of types of designs existing knowledge of sampling designs needed to be structured. Sub-section 6.3.2 presents a hierarchical framework to classify sampling designs and a taxonomy of the main classes of sampling designs. Sub-section 6.3.3 deals with the selection of types of designs.

6.3.1 Selecting statistical approaches

In Sub-section 2.2.1 it was explained why the system initially only assists in the use of classical sampling theory, i.e. a design-based approach, but that it would be desirable to apply it to the use of geostatistics, i.e. a model-based approach, in future. Therefore, the selection of an appropriate sampling approach should take place at a high level in the system. Figure 6.1 summarizes the selection of appropriate statistical approaches in a decision tree. As explained in Chapter 1, the appropriateness of sampling approaches is related to the type of request. The distinction between types of request is made at the root of the tree. Brus (1993) also discusses the appropriateness of sampling approaches. Besides the design-based and the model-based approach, he introduces the model-assisted approach. Provisionally, the system proposed will focus on choosing from the first two approaches.

For *how much* and *how much & where* requests it should be further decided whether a model-independent quantification of the accuracy is required. In that case a design-based approach should be followed. However, before designing a survey scheme using probability sampling, it should be checked whether that type of sampling is feasible. Sometimes, random selected sample points are not accessible. Such situations cannot always be foreseen at the office, and are not immediately disastrous if the number of inaccessible points is limited. However, if access to large parts of the survey region is problematic, probability sampling

is not feasible. Then, the only alternative is purposive sampling, i.e. to use the available prior information to select sample points without using a random process. Unless the requirement of model-independent quantification of accuracy is dropped, no statistical approach is appropriate. In this thesis no attention is paid to assistance in the use of non-statistical approaches.



Legend:
DB : Design-Based approach
MB: Model-Based approach
NS : No Solution

Figure 6.1 Decision tree for selecting statistical approaches

If the accuracy does not need to be quantified independently of a model a model-based approach may be appropriate. The model-based approach requires that a variogram is modelled, based on the survey data. Therefore, it should be possible to measure sufficient sample points. Pronouncements upon adequate sampling for estimating variograms can be found in the literature. Webster and Oliver (1992), for example, state that at least 100 data are needed to compute an acceptable variogram. An estimate of the sample size is needed to determine whether a variogram can be modelled adequately. When an appropriate method of determination has been selected before choosing a statistical approach, the sample size allowed can be estimated roughly based on the corresponding analysis cost and practical experience of the surveyors.

In the case of a *where* request, the survey results should allow meaningful predictions of intermediate points. Prior information on spatial variability and a rough estimate of the

sample size should be used to check whether sufficiently dense sampling is possible. This depends on the variogram.

The following sub-sections focus on assistance in the use of classical sampling theory.

6.3.2 Structuring knowledge of sampling designs

In this study a hierarchical framework of sampling designs has been constructed in which sampling designs are grouped into types, and types are grouped into classes (Fig. 6.2).

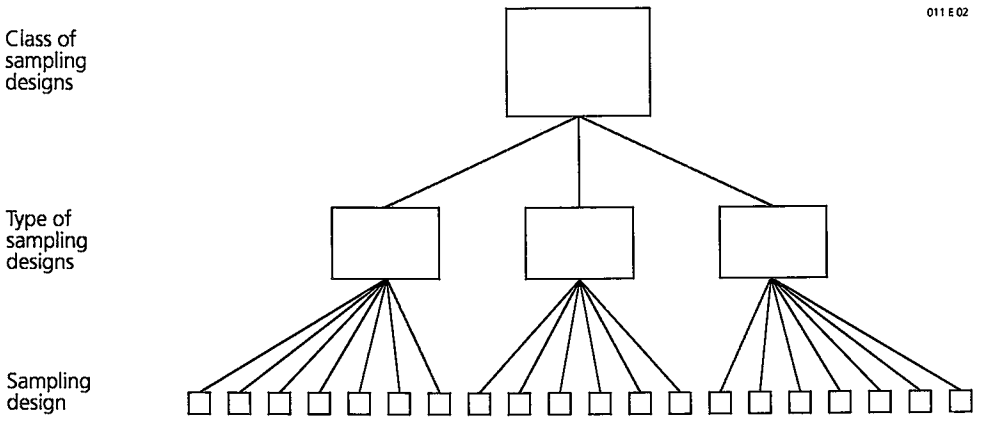


Figure 6.2 Hierarchical framework of sampling designs

At the top level *classes of sampling designs* are distinguished. The distinction between classes of designs is related to partitioning of the population. A partition divides the population into a number of subsets, such that each element of the population belongs to exactly one subset. The subsets can be used in sampling as strata, primary units or clusters, depending on the class of designs. Also, there may be more that one partition, together forming a hierarchical subdivision of the population. For example, the population may be first subdivided into strata, and then elements within these strata may be grouped into clusters. The number and sequences of partitioning and sampling or subsampling differ between classes (e.g. stratified two-stage sampling versus two-stage stratified sampling). In this thesis, the main classes of sampling designs dealt with in classical sampling theory (e.g. Cochran, 1977; Särndal et al., 1992) have been ordered in a taxonomy (Fig. 6.3). The abbreviations of classes of designs used here are in accordance with those used in Särndal et al. (1992). In each class a number of *types of sampling designs* can be specified. Types of sampling designs within the same class differ only with respect to whether selection is made with or without replacement, and with equal or unequal probabilities of inclusion. The type of random selection affects the estimator and its variance. Types of random selection are discussed at the end of this sub-section. Data from designs of the same type can be analysed statistically using the same method of inference. The lowest level of the hierarchical framework comprises the sampling designs. A *sampling design* is a mathematical function that assigns a probability of selection to every possible subset of sampling elements. Designs within a type differ only with respect to the number of elements or sets of elements that are to be selected; the type of design

prescribes the way in which these elements or sets of elements are selected.

Simple random sampling	*SI*
Simple random cluster sampling	*SIC*
Systematic sampling	*SY*
Stratified sampling	
Stratified sampling with *SI* sampling in each stratum	*STSI*
Stratified sampling with *SIC* sampling in each stratum	*STSIC*
Stratified sampling with *SY* sampling in each stratum	*STSY*
Stratified two-stage sampling	
Stratified two-stage sampling with *SI* sampling in both stages	*STSI, SI*
Stratified two-stage sampling with *SI* sampling in the first stage, and *SIC* sampling in the second stage	*STSI, SIC*
Stratified two-stage sampling with *SI* sampling in the first stage, and *SY* sampling in the second stage	*STSI, SY*
Two-stage sampling	
Two-stage sampling with *SI* sampling in both stages	*SI, SI*
Two-stage sampling with *SI* sampling in the first stage, and *SIC* sampling in the second stage	*SI, SIC*
Two-stage sampling with *SI* sampling in the first stage, and *SY* sampling in the second stage	*SI, SY*
Two stage stratified sampling	
Two-stage sampling with *SI* sampling in the first stage, and *STSI* sampling in each stratum	*SI, STSI*
Two-stage sampling with *SI* sampling in the first stage, and *STSIC* sampling within each stratum	*SI, STSIC*
Two-stage sampling with *SI* sampling in the first stage, and *STSY* sampling within each stratum	*SI, STSY*

Figure 6.3. Taxonomy of main classes of sampling designs, with abbreviations

The classes of sampling designs are composed of one or more of five basic procedures: simple random, cluster, systematic, stratified, and two-stage. Two-stage sampling is a frequently used instance of multi-stage sampling; the preparation and execution of three-or-more-stage sampling is more complex, and is therefore hardly used in soil surveying. Here, two-stage sampling is the only form of multi-stage sampling considered. Consequently, the main classes of sampling designs are combinations of at most three of these procedures. Initially, the design may be simple random, cluster, systematic, stratified, or two-stage. When, for example, the class of designs starts with stratified or two-stage sampling, the design within strata or primary units may be simple random, cluster, systematic, two-stage (if 'stratified' at the first level) or stratified (if 'two-stage' at the first level). When both 'stratified' and 'two-stage' are used in a design, the design may finally be simple random, cluster or systematic. The meaning and characteristics of the five basic procedures with respect to soil sampling are described below.

Simple random (SI) sampling means that each possible sample which may result from the sampling design has an equal probability of being selected. *SI* sampling within a region or sub-region leads to an irregular scattering of sample points over this region or sub-region. Visiting independently selected points will be rather time-consuming in comparison with a design in which the sample points are grouped, e.g. in clusters, but the estimated means for the region in the case of *SI* sampling may be more accurate (dependent on the spatial

variability in the survey region). When randomly selecting a sample point each sampling element in the population should have an equal chance of being drawn. Therefore, it should be possible to select random numbers, i.e.

" ... stochastic variables which are uniformly distributed on the interval [0,1] and show (stochastic) independence." (Kleijnen, 1974: p. 6)

Nowadays computers are generally used to select such numbers using a deterministic procedure. These procedures however do not result in random numbers, but in *pseudo-random numbers.*

"Pseudorandom numbers are generated by applying a deterministic algebraic formula which results in numbers that for practical purposes are considered to behave as random numbers, i.e., to be also uniformly distributed and mutually independent." (Kleijnen, 1974: p. 8)

In practice, most people rely on the results of the random-number generators which are available as if they were true randomly selected numbers. However, these results are always pseudo-random and one should be aware of the following problems:

- the procedures that calculate the numbers require a starting value or seed. This seed often influences the result; after a number of calculations the same 'random' number or numbers may be generated;
- the values of parameters in the deterministic procedure, the type of procedure, and the type of computer (word length) influence the maximum number of pseudo-random numbers that can be generated before the same sequence of numbers is repeated within a certain resolution margin.

Uniformity and independence of pseudo-random numbers should be tested. Many tests have been developed to check the results of generators. Kleijnen and Van Groenendaal (1988) discuss some of these methods, but they add a warning: these tests are often unable to assure that a generator performs correctly. Advantages of the use of computers to produce random numbers are:

- if the same seed and the same parameter values are used repeatedly the same sequence of numbers can be reproduced;
- there is no need to store random numbers;
- therefore, the computer needs no time to read in random numbers from an external source, which is usually a slow process (Kleijnen, 1974).

Here, the fact is accepted that pseudo-random numbers will always be used for sample selection instead of true random numbers. Therefore, from now on the terms random and random number stand for pseudo-random and pseudo-random number in this thesis.

Cluster sampling means that groups or *clusters* of sampling elements are selected instead of individual elements. These clusters have to be defined in advance. In spatial sampling, the configuration, direction and often also the size (i.e. number of elements) of the clusters

are specified before a sample can be selected. An example of a cluster is an east-west oriented transect, with five equidistant sample points. A cluster may also consist of a grid of a fixed number of points, or of two perpendicular transects. In this study only simple random cluster (*SIC*) sampling is considered. Then, clusters are selected by drawing a random starting point for each cluster from which the other points in the cluster can be found by pacing. Clusters may or may not have equal sizes. The distances between sample points can be manipulated by the definition of clusters. A common reason for cluster sampling is to reduce survey costs, assuming that locating and visiting points clustered in the field is less time-consuming than locating and visiting the same number of independently selected points. The use of *SIC* sampling for operational advantages may, however, result in less accurate survey results than random sampling. Clustering can also be applied to influence the minimum distance between sample points, which may be required when the spatial variation of the target variable is known beforehand. If the soil property of interest hardly changes over short distances, nearby sample points provide little additional information and should therefore be avoided. In such a case, the definition of the cluster may be used to ensure that sample points are not located at short distances. Then, clusters may no longer have an operational advantage, but the accuracy of the results would increase.

A *systematic (SY) sample* covers a whole region or sub-region with a systematic pattern of sample points, e.g. a square grid or a triangular grid. The co-ordinates of the sample points are determined by randomly selecting one starting point from which the other sample points can be found following the systematic pattern. *SY* sampling can therefore be seen as a special case of cluster sampling, in the sense that only one cluster is selected. If the distance between adjacent sample points is not too large, a *SY* sample may have operational advantages: the next sample point can be located easily. *SY* sampling produces the most even spreading of sample points over the region, which may be advantageous in some cases, e.g. with irregular spatial variation, but disadvantageous in others, e.g. in populations with periodic variation in space. Soil properties can vary in an area with a period of several metres. For example, there may be periodicity in the groundwater level of a field that has been underdrained. If the period of a grid coincides with this periodicity in the field, which cannot easily be observed, the results will be much less accurate than would be expected from the sample data. However, if the period of variation and its direction are known, a grid can be defined with unrelated spacing and orientation. *SY* sampling generally facilitates field work, but a disadvantage is the lack of a method for estimating the sampling variance objectively from the sample data, i.e. without introducing assumptions.

Stratified sampling indicates a division of the population, e.g. a division of the survey region into sub-regions or *strata* in each of which a number of sample points or sets of sample points is selected. In soil surveying, stratification is often used. One reason for this may be that data with known accuracy are required for the sub-regions (*how much & where* request). Then, it may be possible to treat each stratum as a population in its own right. Stratification can also be used to obtain a more accurate estimate for the survey region as a whole (*how much* request). When the whole survey region is rather heterogeneous, it may be divided into relatively homogeneous strata. Then, accurate estimates of the strata can be obtained from relatively small samples, which can be combined into an accurate estimate for the whole region.

In the case of stratification, there is no obligation to use the same type of design within

strata. Eventually, the most efficient design should be determined for the survey region as a whole or for the separate strata. In both situations, the use of different types of design within different strata may be most efficient, e.g. *SI* sampling in one stratum and *SIC* sampling in another. For simplicity, the overview of the main classes of designs is restricted to those with the same approach within strata.

A *two-stage sampling* procedure implies that sampling elements must be selected in two stages. Initially, the population is subdivided into primary units from which in the first stage a number are selected at random. From the primary units chosen in this way second-stage units are selected: the sample points or sets of sample points. When, for example, *SI* sampling is used in both stages, the design is referred to as *SI, SI* sampling. The motivation for using multiple stages is the cost reduction which can be achieved by grouping the sample points. Comparable to the use of cluster sampling for cost reduction, this may reduce accuracy, because sample points are forced near to each other in some parts of the survey region whereas in other parts of the survey region sample points are lacking.

069 E 03

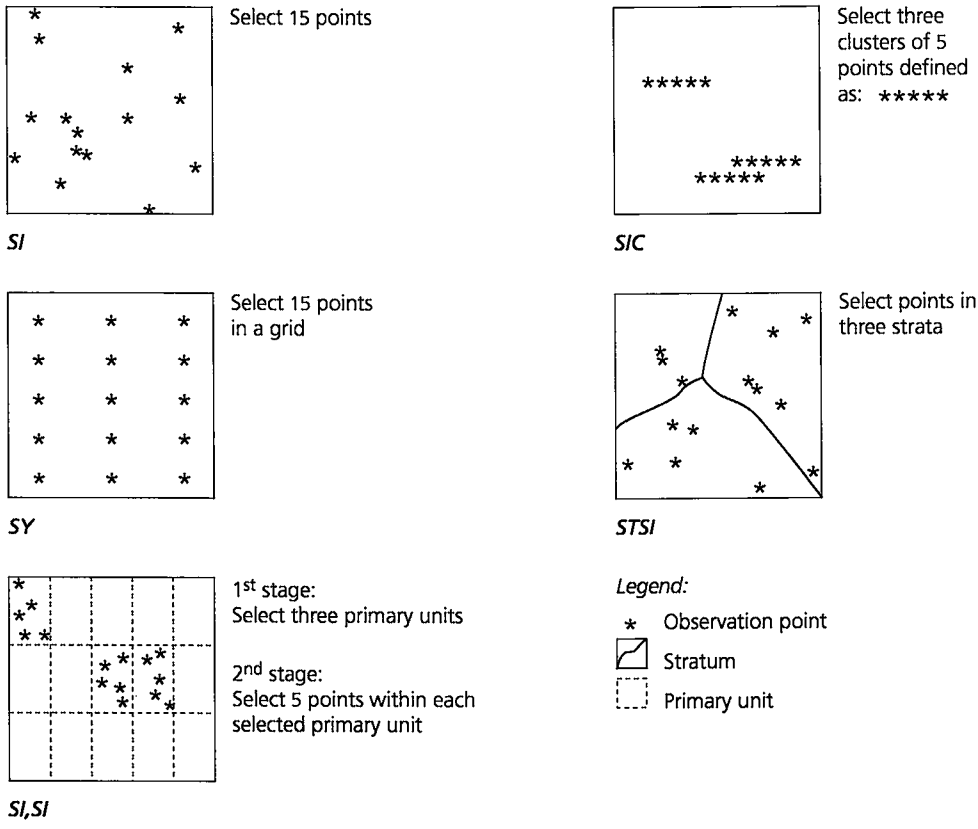


Figure 6.4 Results of five procedures to select 15 sample points

Figure 6.4 shows the possible results of selecting 15 sample points in a survey region

using these five procedures. Before samples are selected, the parts of the survey region that are used for selection need to be defined, e.g. the strata, the primary units, the clusters, the sample points. The prior information may embody partitions of the population suitable for sampling. Thereafter, first the type or types of random selection must be specified, resulting in the type of sampling design. Then the numbers of elements or subsets of elements to be selected need to be determined.

For each case in a class of designs for which elements or subsets of elements need to be randomly selected, it is essential to specify whether the primary units, clusters, or sample points will be selected with or without replacement and with equal or unequal probabilities of inclusion. Types of sampling designs within the same class differ only with respect to the type or types of random selection. Sampling with replacement means that an element may be selected more than once. Cochran (1977, p. 18) states that the formulae for the variances and estimated variances of estimates made from the sample are often simpler when sampling is with replacement. Therefore sampling with replacement is often used, although it seems disadvantageous to have the same sample point two or more times in the sample. In spatial surveys the chance of selecting the same sample point twice is negligible because the collection of possible locations, e.g. defined by pairs of an *x* and a *y* co-ordinate, is generally very large in relation to the number of sample points to be selected. Therefore, selection with replacement is often preferable. In the case of *SI, SI* sampling first sub-regions are selected (*primary units*) within which sample points are selected at the next stage. The number of primary units may be limited, which often makes selection without replacement preferable in order to achieve a spread of groups of sample points over the survey region. Sample points within these units may then be selected with replacement.

With respect to the probability of inclusion it should be decided whether some elements should have a greater chance of being included in the sample than others, and whether this probability should be proportional to their size, or to an auxiliary variable. Selection of units or clusters with a probability proportional to size means selection with a probability proportional to the number of elements in them. If all elements or subsets of elements from which a number has to be selected have equal sizes, elements or subsets are usually selected with equal probabilities. However, if elements or subsets of elements do not have equal sizes, it may be more efficient to select elements or subsets with unequal probabilities. Probability proportional to size is often used to obtain so-called 'self-weighting' samples, hence simple formulae for estimation. For example, if clusters of unequal sizes have to be selected, and these clusters are selected by drawing a random starting point in the survey region, larger clusters, which consist of more elements, have a larger probability of inclusion than smaller clusters. Selection with probability related to an auxiliary variable is not related to the sizes of the (subsets of) elements. It is used to reduce the sampling variance. Table 6.1 shows the six possible types of random selection.

Decisions on both replacement and probability of inclusion have to be made for each selection except for *SY* sampling, where only one 'grid' has to be selected per region or sub-region and therefore only the probability of inclusion is relevant. One example of a type of design is: *STSI* sampling with per stratum selection of sample points with replacement and equal probability. Another example is: *SI, SIC* sampling with random selection of primary units without replacement and equal probability, and selection of clusters of variable sizes with replacement and probability proportional to their sizes.

069 E 06

Probability of inclusion		Replacement	
		with	without
Equal		X	X
Unequal	- Proportional to size	X	X
	- Proportional to auxiliary variable	X	X

Table 6.1 Types of random selection

Each type of design has its own statistical method of inference. The formulae for the main types of design are available from the literature on classical sampling theory. For more advanced types of designs, methods of inference need be derived special, which are beyond the scope of this thesis. For instance, the population mean from *STSI* sampling with per stratum selection of sample points with replacement and equal probability, is estimated according to:

$$\hat{z} = \sum_{h=1}^{L} \{w_h \cdot \frac{1}{n_h} \cdot \sum_{i=1}^{n_h} z_i\} \qquad (1)$$

where:
$\hat{z}$ = estimator of the spatial mean of a property z in the survey region;
L = number of strata;
h = stratum number;
w_h = weight of stratum h, equal to its areal proportion in the survey region;
n_h = number of sample points in stratum h;
z_i = the value at the ith sample point.
The variance of the estimator is:

$$V(\hat{z}) = \sum_{h=1}^{L} \frac{w_h^2 \cdot S_h^2}{n_h} \qquad (2)$$

where:
S_h^2 = variance among elements in stratum h.

When applied to spatial sampling, the types of designs considered here start from selection of a sample point by random selection of a combination of an x and a y co-ordinate. There are however a large number of more sophisticated types of designs for sampling a two-dimensional area which result from a combination of sampling procedures in two perpendicular directions. Koop (1990) gives, for example, a detailed description of the sampling theory for twenty-one possible methods of sampling a plane area with random points, resulting from a combination of random, stratified, and systematic sampling in two perpendicular directions, with or without alignment of the sampled points. For each method described the variance functions are derived. Webster and Oliver (1990) give a clear example of the possibility of

combining the advantages of two types of sampling designs: they discuss the principle of stratified systematic unaligned sampling, a design in which the advantages of a regular grid and randomization are combined. This thesis deals only with the construction of the most significant types of designs.

Below, the main features are described of the sampling designs in the two cases introduced in Chapter 4.

Case A. *In this survey, STSI sampling was used with selection of elements in each stratum with replacement and equal probabilities. Stratification was used both to enable separate estimates of the P_{rel} for particular sub-regions (arable versus grass), and to improve the estimate for the whole population. Within strata sampling elements were selected with replacement and equal probability. The sample points were not equally distributed to the strata: to strata defined as 'wet' twice as many sample points were allocated. There were 26 strata and the total sample size was 116. Another definition of the strata would have resulted in an alternative sampling design. Possible classes of designs to achieve operational advantages might have been STSIC sampling or STSI, SI sampling.*

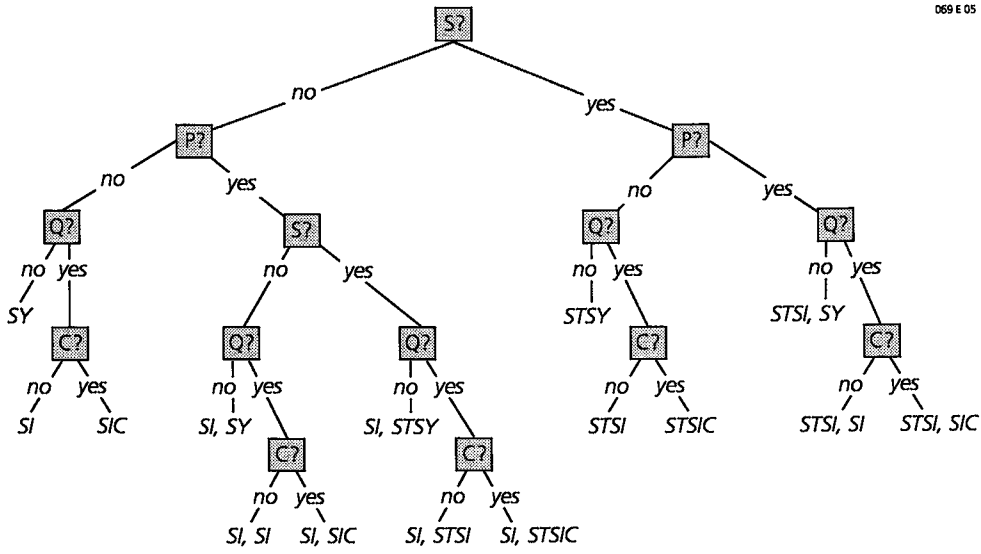
Case B. *The sampling design in case B was STSI, SI sampling. In the first stage per stratum, map delineations were, for operational simplicity, selected with replacement and probabilities proportional to their sizes. So, it was permissable to select a map delineation more than once, and larger areas had more chance of being included in the sample. Within selected map delineations, points were selected with replacement and equal probabilities. If clusters of points had been selected with a fixed distance between them, a minimal distance between sample points could have been achieved, e.g. STSI, SIC sampling or STSIC sampling. This may be an interesting approach for strata with little spatial variation.*

6.3.3 Selecting types of sampling designs

At present, the choice of a type of design is often influenced by the availability and complexity of procedures for sample selection and/or for analysis of the results. If this choice is supported by an automated system, it seems logical to also implement procedures to select samples and to analyse results (not necessarily in the same computer program). So, the procedures for selection and the methods of inference required will no longer be selection criteria. This sub-section focuses on the selection of the main types of sampling designs introduced in the preceding sub-section.

When a design-based approach to sampling is suitable, appropriate classes of sampling designs need to be selected, in which types of designs should be specified. The type of request, the constraints, and the prior information determine which classes of designs are appropriate. Figure 6.5 presents a decision tree to select appropriate classes of designs. At each decision point or node in this tree the system should provide the user with valid options and with information on the main consequences of the various options. A particular sampling procedure may be selected for different purposes, e.g. clusters can be used to reduce survey cost or to improve accuracy. It should be noted that since the criteria for using clusters for each of these two purposes are different, the way in which clusters are defined also differs significantly. If costs need to be reduced, the sample points within a cluster should be close together. However, if there is little variation at short distances, clusters can be defined with larger distances between sample points, in order to increase accuracy. Information on the consequences of a decision regarding efficiency (in terms of accuracy and cost) should always

be provided. These consequences with simple random sampling as the reference procedure are summarized in Table 6.2. Furthermore, the consequences for the sampling frame required and for the information required for prior evaluation (see Chapter 7) should be specified for each option. If such information cannot (easily) be provided, other options should be explored.



Legend:

S?: Can a meaningful partition into strata be made?

P?: Can a meaningful partition into primary units be made?

C?: Can meaningful clusters be defined?

Q?: Should the accuracy be quantifiable from the sample data?

Figure 6.5 Decision tree to select appropriate classes of designs

It should be permissible to select more than one possible class - the user might decide to explore several paths. Thereafter, for each class selected, the type or types of random selection (see Table 6.1) should be chosen. Therefore, similar to the selection of classes of designs, possible options should be presented with their consequences, after which the user should decide.

Table 6.2 The effect of sampling procedures on efficiency (with simple random sampling as the reference procedure)

Procedure	Effect on accuracy	Effect on cost
Simple random	=	=
Stratify	+	≈
Cluster I	+	≈
Cluster II	–	–
Two-stage	–	–
Systematic	+	≈ / –

Legend:
= : equal
≈ : near equal
+ : increase
– : decrease

Cluster I : cluster to improve accuracy
Cluster II : cluster to reduce cost

Rules can be derived to assist in the selection of types of designs for spatial sampling. These rules are partially based on sampling theory and partially on soil survey experience. Examples of important selection criteria and of rules which may help to apply these criteria are:

- type of request:
  IF special interest in sub-regions
  THEN use stratified sampling with sub-regions as strata
- cost reduction:
  IF cost should be reduced
  THEN select groups of sample points instead of individual points (SIC sampling, SY sampling, or two-stage sampling)
- improvement in accuracy:
  IF accuracy should be improved
  THEN select sample points scattered over the survey region (SI sampling, or SY sampling)
  OR use stratified sampling with relatively homogeneous sub-regions as strata
  OR define clusters which ensure that sample points are not located close to each other (cluster sampling).

For the domain of interest in this thesis, it seems difficult to define a set of clear rules which will always lead to the best type of design. Two examples of clear rules in the domain of interest are:

IF the accuracy of the survey result has to be quantified from the sample data only THEN can systematic sampling designs not be applied

IF an estimate of the sampling variance is to be computed from the sampling results per stratum THEN at least two sample points must be drawn from each stratum.

However, the number of clear rules in the domain is limited. In practice, it is hard to define detailed rules for which unambiguous information can be provided and which result in unambiguous outcomes that can be generally accepted. On the one hand there may be disagreement between statisticians on these details, which could hamper the development of a generally accepted rule base. On the other hand, rules at a detailed level may require a great deal of information from the user, which may be rather time-consuming and, furthermore, does not assure that a better solution will be found than when some decisions are left to the user. The system will therefore provide users with information on types of designs and guide them to the most appropriate types for particular requests. The rules related to the decision tree (Fig. 6.5) therefore need to be further elaborated.

6.4 Remarks on the design of outlinear survey schemes

Two of the tasks in which the system should assist (distinguished in Section 5.4) were 'the selection of prior information, methods of determination and types of sampling designs' and 'the generation of outlinear schemes'. The preceding section showed that information on a method of determination influences the selection of a statistical approach (Fig. 6.1), since it influences the sample size allowed. Therefore methods of determination should be selected before appropriate sampling approaches and thus before types of designs can be selected. During these selections the constraints defined earlier should be taken into account, e.g. budgetary restrictions, or limited laboratory capacity. If appropriate types of designs can be selected, this automatically results in outlinear schemes. So, the design of outlinear schemes comprises the following steps:

- select an appropriate method of determination;
- select an appropriate sampling approach;
- if a design-based approach is appropriate, select types of designs.

If more than one method of determination is possible, the selection of a sampling approach and of types of designs should be repeated, resulting in different outlinear schemes. The next chapter deals with the evaluation and optimization of outlinear schemes.

Chapter 7

Evaluation and optimization of survey schemes

Parts of this chapter have been published in:

Domburg, P., Gruijter, J.J. de & Brus, D.J. (1994)
A structured approach to designing soil survey schemes with prediction of sampling error from variograms. Geoderma, 62(1-3), pp. 151-164.

Domburg, P., Gruijter, J.J. de & Beek, P. van (submitted)
Designing Efficient Soil Survey Schemes with a Knowledge-Based System using Dynamic Programming. Environmetrics.

7. Evaluation and optimization of survey schemes

7.1 Scope

The outlinear survey schemes designed need to be refined, taking into account the constraints defined earlier. Therefore, the number of subsets of elements and/or the number of elements to be selected in the sampling design need to be determined, i.e. the type of sampling design needs to be refined into a specific sampling design. The search for an efficient scheme within the outlinear scheme should result, if possible, in a scheme which is optimal under existing conditions. As stated in Chapter 1, the efficiency of survey schemes is related to the predicted accuracy and the predicted cost. The efficiency of a given sampling design p can be defined as the ratio of the sampling variance of a reference design (often *SI* sampling) to the sampling variance of p, at same sample size or at same cost. To be able to objectively compare the efficiency of possible schemes models are required to predict accuracy and cost, i.e. models for prior evaluation (Sections 7.2 and 7.3). These models can be used to search for an optimal scheme within an outlinear scheme.

The objective of optimization is to find a final scheme which maximizes the accuracy for a fixed budget, or alternatively which minimizes the cost for a fixed accuracy constraint. Section 7.4 looks into the possibility of using techniques from the field of OR to support the search for an optimal scheme. The procedure proposed makes it possible to calculate optimal survey schemes within previously designed outlinear schemes. Section 7.5 discusses the procedure for evaluation and optimization of schemes.

To illustrate the theory, a fictitious case, which is related to case A in Chapter 4, is introduced below. This case is used throughout this chapter.

Case. *Let the aim of the survey be to determine the areal proportion of soil saturated with phosphate in the whole of a region. This areal proportion can be estimated by using an indicator variable, indicating whether the soil at a particular point is to be considered as saturated with phosphate.*

Let the hypothetical survey region be 8 km x 8 km, subdivided into four sub-regions of 4 km x 4 km, with different guessed values of the areal proportion saturated, based on prior information on soil type and manuring. The guessed values for the proportion of soil saturated with phosphate were 15, 35, 40, and 50%, respectively.

For this case one method of determination was assumed to be favourite, so no alternative methods of determination had to be compared. Given the aim and the available prior information, two types of sampling designs which seemed appropriate were evaluated: STSI sampling with sampling within strata with replacement and equal probabilities and STSI, SI sampling with sampling of primary units with replacement and probability proportional to size, and sampling within primary units with replacement and equal probabilities. In both designs the sub-regions were used for stratification. In the STSI, SI design the strata were subdivided into primary units of 500 m x 500 m. Within each stratum a number of primary units were selected, in each of which a number of sample points were selected.

The parameter values used in this case were based on historical surveys of phosphate saturation.

7.2 Prediction of accuracy

To support the search for an efficient scheme, a model that predicts the accuracy is required. To be able to predict the accuracy, it first should be specified how the accuracy can be measured (Sub-section 7.2.1). Prior information should be used to evaluate the accuracy in advance (Sub-section 7.2.2). Sub-section 7.2.3 presents a general algorithm for sampling error prediction using prior information on spatial variation, after which Sub-section 7.2.4 deals with the possibility of using specific algorithms for specific types of designs.

7.2.1 Measure of accuracy

In this thesis the term accuracy is deliberately used rather than the more restricted term precision. *Precision* is only related to the error in an unbiased estimator due to an unbiased sampling design, i.e. the sampling variance. This error can usually be reduced by increasing the size of the sample. An estimator is unbiased if it gives the true value on average. In the scope of the KBS proposed a biased estimator or a biased design may sometimes be acceptable in exchange for a greater reduction of variance. In soil surveys bias can also be introduced by other sources of error which are unaffected by the size of the sample, e.g. errors resulting from the choice of the method of determination. This usually causes a systematic error. This type of error is not considered here because it cannot be affected by the type of sampling design. Since both sampling variance and bias are taken into account, the more comprehensive concept of *accuracy*, which is defined as the mean squared error due to sampling, will be used here rather than precision.

7.2.2 Use of prior information

The sampling error or the accuracy of a soil survey scheme is influenced by the variation of the soil property of interest in space. It is therefore desirable to utilize prior information on spatial variation to predict the accuracy. Such information can be quantitatively described using a *variogram*, γ (Journel and Huijbregts, 1978), i.e. a function specifying the relation between the vector $\boldsymbol{h}$ separating any two points in the area and their so-called semi-variance:

$$\gamma(\boldsymbol{h}) = \frac{1}{2} E_{\xi}\{z(\boldsymbol{x}) - z(\boldsymbol{x}+\boldsymbol{h})\}^2 \qquad (1)$$

where E_{ξ} denotes the expectation over realizations from an underlying stochastic model ξ, and $z(\boldsymbol{x})$ is the value of the property of interest at point $\boldsymbol{x}$. Most soil properties vary continuously in space. Variograms originate from the idea that observations at short distances from each other are more similar to each other than observations at larger distances. If the variability is the same in different directions there is isotropy in the region, in which case the semi-variance is only a function of the distance between a pair of sample points, $|\boldsymbol{h}|$, referred to as the lag. In case of anisotropy the variability differs for example in perpendicular directions and the semi-variance depends on both the direction and the length of vector $\boldsymbol{h}$. A clear overview of models for variograms of soil properties is given by McBratney and Webster (1986). They distinguish two main types of models. Firstly, there are transitive models in which the semi-variance increases with increasing lag-distance to some maximum at which it remains with further increase in lag. Beyond this distance (the range) two points are spatially

independent and the semi-variance is maximal (the sill). Secondly, in the unbounded models the semi-variance appears to increase without limit as the area increases.

Variograms based on empirical data are often discontinuous at the origin. Although by definition $\gamma(0) = 0$, the value of γ as the lag approaches 0 is not necessarily zero. This discontinuity is called the nugget effect and can be caused by: i) variability at very short distances, ii) operator error, i.e. the influence of different surveyors, iii) measurement errors. Some examples of variogram models frequently used in soil survey are shown in Table 7.1.

A variogram can be fitted for a soil property in a region, if a data set of observations is available of this property in this region. Problems concerned with choosing functions for variograms and with fitting variograms are frequently dealt with in geostatistical literature (e.g. McBratney & Webster, 1986), and are therefore not considered here. The processing of information on spatial variation falls outside the scope of this thesis.

Table 7.1 Examples of types of variogram models

Type	Model	
Linear with sill	$\gamma(\lvert\boldsymbol{h}\rvert) = 0$	for $\lvert\boldsymbol{h}\rvert = 0$
	$= c_0 + c\,(\lvert\boldsymbol{h}\rvert/a)$	for $0 < \lvert\boldsymbol{h}\rvert \leq a$
	$= c_0 + c$	for $\lvert\boldsymbol{h}\rvert > a$
Isotropic spherical	$\gamma(\lvert\boldsymbol{h}\rvert) = 0$	for $\lvert\boldsymbol{h}\rvert = 0$
	$= c_0 + c\,[1.5\,(\lvert\boldsymbol{h}\rvert/a) - 0.5\,(\lvert\boldsymbol{h}\rvert/a)^3]$	for $0 < \lvert\boldsymbol{h}\rvert \leq a$
	$= c_0 + c$	for $\lvert\boldsymbol{h}\rvert > a$
Exponential	$\gamma(\lvert\boldsymbol{h}\rvert) = 0$	for $\lvert\boldsymbol{h}\rvert = 0$
	$= c_0 + c\,[1 - \exp(-\lvert\boldsymbol{h}\rvert/r)]$	for $\lvert\boldsymbol{h}\rvert > 0$
Anisotropic spherical	$\gamma(\lvert\boldsymbol{h}\rvert) = 0$	for $\lvert\boldsymbol{h}\rvert = 0$
	$= c_0 + c\,[1.5\,(\lvert\boldsymbol{h}\rvert/a) - 0.5\,(\lvert\boldsymbol{h}\rvert/a)^3]$	for $0 < \lvert\boldsymbol{h}\rvert \leq a$
	$= c_0 + c$	for $\lvert\boldsymbol{h}\rvert > a$
	with:	
	$a(\theta) = \{ a_\phi^2 \cdot \cos^2(\theta - \phi) + a_\perp^2 \cdot \sin^2(\theta - \phi) \}^{1/2}$	

c_0 = nugget; c = difference between nugget and sill; $|\boldsymbol{h}|$ = distance between a pair of sample points; a = range; r = parameter defining the spatial scale of the variation in a way analogous to the range of the previous models (in exponential models there is no strict range and the sill is approached asymptotically); ϕ = preferential direction; a_ϕ = range in direction ϕ; $a_\perp$ = range in direction perpendicular to ϕ; θ = direction of lag vector.

At present, information on spatial variation is not always available nor can it easily be brought into the standard form of a variogram in general. However the amount of information on soil spatial variability increases in time: with respect to this the reader can be referred to a project in the Netherlands called 'National Sampling Map Units' (case B) which aims at collecting detailed quantitative information on the spatial variability within map units of the national soil map of the Netherlands on a scale of 1:50 000. If no information on spatial variation is available for a survey region, it may be possible to use information from a comparable region, or the parameters of a variogram model can be guessed from field experience.

Case. *As stated in Section 7.1, an indicator variable can be used to estimate the areal proportion of soil saturated with phosphate. Given a guessed value v for the areal proportion saturated with phosphate, the sill of a sample indicator variogram can be computed by v·(1-v) (Journel & Posa, 1990).*

A historical survey has shown that a spherical variogram model can describe the spatial dependence of phosphate saturation reasonably well, with a nugget of 0.093 and a range of 468 metres. The values for the sills for the four strata are, respectively, 0.128, 0.228, 0.240, and 0.250. Figure 7.1 shows the sample indicator variograms for the four strata.

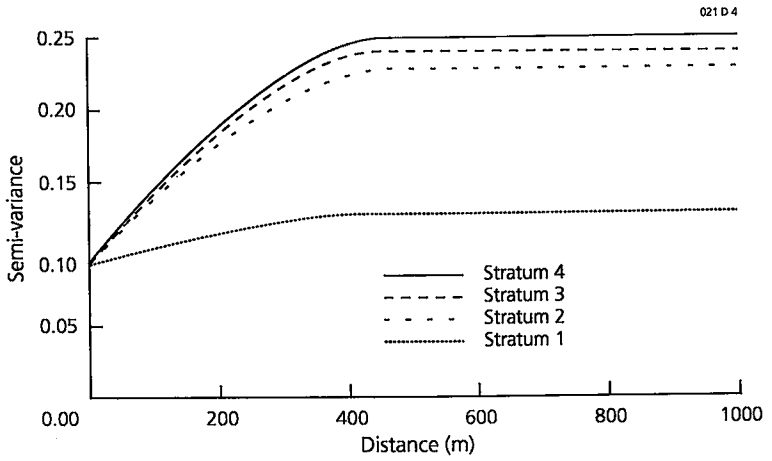


Figure 7.1 Sample indicator variograms for the four strata

7.2.3 General algorithm for sampling-error prediction

The algorithm discussed in this section is general in the sense that it can be used with any sampling design.

Consider the case in which the spatial mean $\bar{z}$ of a property z in a given region A is to be estimated by $\hat{z}$, a weighted mean of a future probability sample from A, and that the mean squared error of $\hat{z}$ due to sampling according to a given design p is to be predicted. Suppose further that:

- the region A has been divided into a large but finite number of sampling elements, which act as possible sample points;
- the estimator is linear, i.e.

$$\hat{z} = \sum_{i=1}^{n} \lambda_i z(\mathbf{x}_i) \qquad (2)$$

where:

n = sample size, i.e. the number of sample points;

$\mathbf{x}_i$ = ith sample point within the region A;

λ_i = weight at $\boldsymbol{x}_i$ which depends on the probability of $\boldsymbol{x}_i$ being included in the sample, governed by the sampling design p; the λ_i ($i = 1, 2, ..., n$) sum to 1;
$z(\boldsymbol{x}_i)$ = value of the property of interest at the location $\boldsymbol{x}_i$.

The measure of accuracy to be established from the prior information is the mean squared error of $\hat{z}$ due to sampling under design p, denoted by r:

$$r = E_p\{(\hat{z}-\bar{z})^2\} \tag{3}$$

where E_p denotes the statistical expectation over repeated sampling under design p.

The following procedure can be used to predict r:

- bring the prior information about the spatial variation of z in A into the standard form of a variogram γ;
- adopt as a predictor of r its statistical expectation $\bar{r}$ over realizations from the stochastic model ξ underlying the variogram:

$$\bar{r} = E_\xi(r) = E_\xi[E_p\{(\hat{z}-\bar{z})^2\}] \tag{4}$$

- evaluate $\bar{r}$ by simulation as discussed below.

This procedure is model-based only in so far as the prediction of sampling error is concerned; it does not violate the design-based character of the sampling strategies themselves, as the weights λ_i used in the estimator $\hat{z}$, Eq. (2), are still determined by design p, and not by the model ξ.

The simulation procedure to predict the value of $\bar{r}$ that follows most directly from Eq. (4) would be:

1. generate from γ a field of z-values in A, and calculate $\bar{z}$;
2. randomly select a sample according to p, and calculate $\hat{z}$;
3. calculate the squared error $(\hat{z}-\bar{z})^2$;
4. repeat steps 2 and 3 a sufficient number of times to estimate the mean squared error r;
5. repeat steps 1 - 4 a sufficient number of times to estimate $\bar{r}$.

Albeit straightforward, this procedure is computationally demanding because not only many samples but also many z-values have to be generated repeatedly. However, generation of z-values can be avoided altogether by making use of the fact that the order of the two expectation operators may be reversed, i.e. $E_\xi(E_p) = E_p(E_\xi)$, hence Eq. (4) can be rewritten as:

$$\bar{r} = E_p[E_\xi\{(\hat{z}-\bar{z})^2\}] \tag{5}$$

For a given sample $E_\xi\{(\hat{z}-\bar{z})^2\}$ can be calculated directly from the variogram, without z-values, using the basic geostatistical equation (Journel & Huijbregts, 1978: p. 305):

$$E_\xi\{(\hat{z}-\bar{z})^2\} = 2\lambda'\cdot\bar{\gamma}_{S,A} - \bar{\gamma}_A - \lambda'\cdot\Gamma_S\cdot\lambda \quad (6)$$

where:

λ = n-vector of sample weights according to design p;
$\bar{\gamma}_{S,A}$ = n-vector of mean semi-variances between each sample point and all points in A;
$\bar{\gamma}_A$ = mean semi-variance between all pairs of points in A;
Γ_S = n x n matrix of semi-variances between sample points.

Substituting Eq. (6) into Eq. (5) gives:

$$\bar{r} = 2E_p(\lambda'\cdot\bar{\gamma}_{S,A}) - \bar{\gamma}_A - E_p(\lambda'\cdot\Gamma_S\cdot\lambda) \quad (7)$$

In the usual case of p-unbiasedness, i.e. $E_p(\hat{z})=\bar{z}$, a great deal of computation can be saved because:

$$E_p(\lambda'\cdot\bar{\gamma}_{S,A}) = E_p(\sum_{i=1}^{n}\lambda_i\cdot\bar{\gamma}_{i,A}) = \bar{\gamma}_A \quad (8)$$

Substitution of Eq. (8) into Eq. (7) gives the formula that is the basis of the final simulation procedure:

$$\bar{r} = \bar{\gamma}_A - E_p(\lambda'\cdot\Gamma_S\cdot\lambda) \quad (9)$$

In soil surveys it is nearly always attempted to improve efficiency by some form of stratification, i.e. by dividing the region into relatively homogeneous sub-regions (strata) and taking samples from each, independently of one another. The estimate for the region as a whole is then calculated as the mean of the stratum means, weighted by the relative areas of the strata. The more homogeneous the strata are relative to the whole, the larger the reduction in sampling variance will be. Stratification is used widely to take advantage of prior information. It is therefore propitious that, as explained below, specializing the general formula of Eq. (9) for stratified sampling reduces computation considerably. The essence of this is that because the strata are sampled independently, $\bar{r}$-values for the strata can be calculated separately to determine the overall value.

The starting point for the derivation is the classical formula for the sampling variance of $\hat{z}$ with *STSI* sampling (Cochran, 1977: p. 92). Due to the unbiasedness property, the sampling variance equals the mean squared error r, hence the classical formula can be written as:

$$r = \sum_{h=1}^{L} w_h^2 r_h \quad (10)$$

where:

L = number of strata;
h = stratum number;
w_h = weight of stratum h, equal to its areal proportion in region A;
r_h = mean squared error of the estimated spatial mean in stratum h.

Taking the ξ-expectation as in Eq. (4) results in:

$$\bar{r} = E_\xi\left(\sum_{h=1}^{L} w_h^2 r_h\right) = \sum_{h=1}^{L} w_h^2 \cdot E_\xi(r_h) = \sum_{h=1}^{L} w_h^2 \bar{r}_h \qquad (11)$$

Applying the same reasoning to $\bar{r}_h$ in Eq. (11) as to $\bar{r}$ in Eq. (5) leads to:

$$\bar{r}_h = \bar{\gamma}_{Ah} - E_p(\lambda_h' \cdot \Gamma_{Sh} \cdot \lambda_h) \qquad (12)$$

where:

$\bar{\gamma}_{Ah}$ = mean semi-variance between all pairs of elements in stratum h of A;

λ_h = n_h-vector of sample weights according to the design applied in stratum h, summing to 1 (n_h denotes the sample size in stratum h);

Γ_{Sh} = n_h x n_h matrix of semi-variances between sample points in stratum h.

Substitution of Eq. (12) into Eq. (11) gives:

$$\bar{r} = \sum_{h=1}^{L} w_h^2 \cdot \bar{\gamma}_{Ah} - \sum_{h=1}^{L} w_h^2 \cdot E_p(\lambda_h' \cdot \Gamma_{Sh} \cdot \lambda_h) \qquad (13)$$

Denoting $\lambda_h' \cdot \Gamma_{Sh} \cdot \lambda_h$ by γ_{Sh} and $E_p(\gamma_{Sh})$ by $\bar{\gamma}_{Sh}$ gives:

$$\bar{r} = \sum_{h=1}^{L} w_h^2 \cdot \bar{\gamma}_{Ah} - \sum_{h=1}^{L} w_h^2 \cdot \bar{\gamma}_{Sh} \qquad (14)$$

The computational advantage of Eq. (14) over Eq. (9) is the much smaller number of semi-variances to be calculated for the L matrices Γ_{Sh} of dimension n_h x n_h than for the n x n matrix Γ_S, since n_h is much smaller than n.

An even more important advantage of Eq. (14) is that it allows different variograms to be specified for the strata. Thus, prior information containing this type of differentiation can be taken fully into account, and this may (theoretically) lead to different designs for the strata. Note that Eq. (14) is sufficiently general even to cover the case of different classes of designs in the strata, e.g. *SI*, *SI* sampling in one stratum and *SIC* sampling in another. Also note that the first term in Eq. (14) depends only on the stratification and hence needs to be calculated only once if various designs with the same stratification are to be compared.

In conclusion, the simulation algorithm is as follows.

1. Estimate $\bar{\gamma}_{Ah}$ for each stratum h (if sampling is not stratified, consider the region formally as a single stratum):
   - 1.1 randomly select two points $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ from stratum h, with equal probabilities of inclusion and independently from each other;
   - 1.2 calculate $\gamma_h(\boldsymbol{x}_1,\boldsymbol{x}_2)$ from the variogram for h;
   - 1.3 repeat steps 1.1 and 1.2 sufficiently often, and calculate the mean of $\gamma_h(\boldsymbol{x}_1,\boldsymbol{x}_2)$ as estimate $\hat{\bar{\gamma}}_{Ah}$ of $\bar{\gamma}_{Ah}$.
2. Estimate $\bar{\gamma}_{Sh}$ for each stratum:
   - 2.1 determine the vector of sample weights λ_h from the design p_h;
   - 2.2 randomly generate a sample according to p_h;
   - 2.3 calculate $\gamma_{Sh}=\lambda_h' \cdot \Gamma_{Sh} \cdot \lambda_h$ from the sample configuration and the variogram for h;

2.4 repeat steps 2.2 and 2.3 a sufficient number of times, and calculate the mean of γ_{Sh} as estimate $\hat{\gamma}_{Sh}$ of $\bar{\gamma}_{Sh}$.

3. Calculate $\hat{r}$ as estimate of $\bar{r}$ by inserting the estimates from steps 1.3 and 2.4 in Eq. (14).

So, the simulation algorithm requires that points are selected in the survey region. In practice, survey regions seldom have a regular shape like a square or a rectangle. When points are drawn by selecting pairs of co-ordinates, it needs to be continuously checked whether a selected point belongs to the survey region. If not, a new point should be selected. In the case of a very irregular survey region, the selection of pairs of points and of samples may be rather time-consuming. To avoid this problem it seems attractive to use a grid map of the survey region for sampling, where the size of the grid should correspond to the required accuracy in localizing sampling elements, and to develop a procedure that selects grid cells at random.

How often the steps in this simulation algorithm need to be repeated to obtain 'sufficiently' accurate results depends on the permissable standard error for the simulation results, $s(\hat{r})$, which should be specified in advance, e.g. as a percentage of the $\hat{r}$ to be calculated. The error in $\hat{r}$ due to limiting the number of simulation runs can be quantified and controlled by calculating variances as well as means in steps (1.3) and (2.4). Because all estimates $\hat{\gamma}_{Ah}$ and $\hat{\gamma}_{Sh}$ are independent, the overall standard error $s(\hat{r})$ can be calculated simply according to:

$$s(\hat{r}) = \sqrt{\sum_{h=1}^{L} w_h^4 \left[\frac{Var\{\gamma_h(\boldsymbol{x}_1,\boldsymbol{x}_2)\}}{m_{1h}} + \frac{Var(\gamma_{Sh})}{m_{2h}} \right]} \qquad (15)$$

where m_{1h} and m_{2h} denote the number of generated values of $\gamma_h(\boldsymbol{x}_1,\boldsymbol{x}_2)$ and γ_{Sh}, respectively. It should, however, be noted that if these variances are small, $\hat{r}$ can still be (slightly) distant from $\bar{r}$, because the simulation runs are not completely stochastic. This risk can be reduced by using a very large word length in the simulation procedure for selecting points. This type of simulation, which uses random (or pseudo-random) numbers to find a solution, is generally referred to as Monte Carlo simulation (Kleijnen, 1974).

The above algorithm can be applied to any type of sampling design. An advantage in cases involving more complex designs is that there is no need to specify the variance function for sampling-error prediction. The main requirement is that efficient subroutines for sample selection are available, but such routines will be needed anyway for the support system to select the actual sample. However, a disadvantage is that for every change, either large or small, in the number of elements (and/or primary units, and/or clusters) in the sample a new simulation is required for error prediction (to estimate $\bar{\gamma}_{Sh}$), which may be time-consuming, especially if it is used in a procedure to search for an optimal scheme, in which the results for various possible sampling designs are calculated. This problem can be avoided by using the specific sample-variance formula for each type of design. The following sub-section deals with this approach.

Finally, it should be noted that any random error in measuring z will be automatically included in the final estimate $\hat{r}$ if the measurement error is included in the variogram(s) used

for simulation.

7.2.4 Specific algorithms for types of designs

If the z-values in A and their mean $\bar{z}$ are considered as fixed (design-based approach) then it follows from the unbiasedness of $\hat{z}$ that r equals the sampling variance. The design p not only determines the sampling variance but also the nature of the constituent variance components, e.g. the variance between clusters of elements or the variance within the primary sampling units of a given stratum. Each type of design has its own variance components. Thus, the prediction of the sampling variance can be constructed from predictions of its specific variance components. For each type of design, another set of variance components needs to be extracted from the prior information. This prior information should again be brought into the form of one or more variograms. The variance formulae for all usual types of designs are given in statistical handbooks (e.g. Cochran, 1977; Krishnaiah & Rao, 1988; Särndal et al., 1992). For some unusual types of designs the variance formulae may have to be derived first. The approach for predicting the sampling error using specific variance formulae is illustrated with three types of designs.

SI sampling. The sampling variance, $V(\hat{z})$, for *SI* sampling without replacement and equal probabilities of selection is given by:

$$V(\hat{z}) = \frac{1-f}{n} \cdot S^2 \tag{16}$$

where:

f = sampling fraction n/N (N denotes the size of the population, i.e. the number of sampling elements in the area);

S^2 = population variance between elements, defined as:

$$S^2 \equiv \frac{1}{N}\sum_{i=1}^{N}(z_i-\bar{z})^2 = \frac{1}{N}\sum_{i=1}^{N} z_i^2 - (\frac{1}{N}\sum_{i=1}^{N} z_i)^2 \tag{17}$$

where z_i denotes $z(\boldsymbol{x}_i)$ for brevity. The value of N can be determined by using a grid map of the survey region, where the size of the grid should correspond to the required accuracy in localizing sampling elements. Such a map can also be used as a sampling frame in sample selection. In soil surveying n is generally very small compared with N, resulting in a very small value for f. Although it is obvious that a smaller grid size may cause an increase in N, the value for f will remain very small.

With a *SI* sampling design only one variance component (S^2) needs to be predicted:

$$\bar{r} = E_\xi\{V(\hat{z})\} = E_\xi(\frac{1-f}{n} \cdot S^2) = \frac{1-f}{n} \cdot E_\xi(S^2) \tag{18}$$

Substitution of Eq. (17) into Eq. (18) gives:

$$
\begin{aligned}
\bar{r} &= \frac{1-f}{n}\cdot E_\xi\{\frac{1}{N}\sum_{i=1}^{N} z_i^2 - \frac{1}{N^2}(\sum_{i=1}^{N} z_i)^2\} \\
&= \frac{1-f}{n}\cdot E_\xi\{\frac{1}{2}(\frac{2}{N}\sum_{i=1}^{N} z_i^2 - \frac{2}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} z_i z_j)\} \\
&= \frac{1-f}{N}\cdot E_\xi[\frac{1}{2}\{\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(z_i^2+z_j^2) - \frac{2}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} z_i z_j\}] \\
&= \frac{1-f}{n}\cdot E_\xi\{\frac{1}{2}\cdot\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(z_i^2-2z_i z_j+z_j^2)\} \\
&= \frac{1-f}{n}\cdot E_\xi\{\frac{1}{2}\cdot\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(z_i-z_j)^2\} \\
&= \frac{1-f}{n}\cdot\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{1}{2}E_\xi(z_i-z_j)^2 \qquad \text{(see Eq.(1))} \\
&= \frac{1-f}{n}\cdot\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\gamma_{ij} = \frac{1-f}{n}\cdot\bar{\gamma}_A
\end{aligned}
\tag{19}
$$

In the case of sampling with replacement $V(\hat{z}) = S^2/n$, hence:

$$
\bar{r} = \frac{1}{n}\cdot\bar{\gamma}_A \tag{20}
$$

It can be easily checked that this is in accordance with the general Eq. (9) in Sub-section 7.2.3. Using $\lambda_i = 1/n$ and $\gamma_i = 0$ for $i = 1, 2, ..., n$ this Eq. (9) simplifies to:

$$
\begin{aligned}
\bar{r} &= \bar{\gamma}_A - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}^{n} E_p(\gamma_{ij}) \\
&= \bar{\gamma}_A - \frac{n^2-n}{n^2}\cdot\bar{\gamma}_A = \frac{1}{n}\bar{\gamma}_A
\end{aligned}
\tag{21}
$$

The unknown parameter $\bar{\gamma}_A$ can be determined using the first part of the simulation algorithm in the previous sub-section.

STSI sampling. The sampling variance for *STSI* sampling with selection of sample points within strata with replacement and equal probabilities is given by:

$$
V(\hat{z}) = \sum_{h=1}^{L}\frac{w_h^2\cdot S_h^2}{n_h} \tag{22}
$$

where:

S_h^2 = variance among elements in stratum h.

With this type of design one variance component (S_h^2) has to be predicted for each stratum:

$$\bar{r} = E_\xi\{V(\hat{z})\} = \sum_{h=1}^{L} \frac{w_h^2}{n_h} \cdot E_\xi(S_h^2) \tag{23}$$

which in analogy with *SI* sampling reduces to:

$$\bar{r} = \sum_{h=1}^{L} \frac{w_h^2}{n_h} \cdot \bar{\gamma}_{Ah} \tag{24}$$

Again, the unknown parameters $\bar{\gamma}_{Ah}$ can be determined using the first part of the simulation algorithm in Sub-section 7.2.3.

STSI, SI sampling. In this case the selection of sampling elements from each stratum h proceeds in the following two stages:

1. select at random n_{1h} primary units, i.e. subsets of elements in stratum h, with replacement and probabilities proportional to their sizes;
2. for each time a primary unit has been selected, select at random n_{2h} sample points from it, with replacement and equal probabilities.

The sampling variance with this type of design is given by:

$$V(\hat{z}) = \sum_{h=1}^{L} w_h^2 \left(\frac{S_{bh}^2}{n_{1h}} + \frac{S_{wh}^2}{n_{1h} n_{2h}}\right) \tag{25}$$

where S_{bh}^2 and S_{wh}^2 denote the variance between and within primary units, respectively, of stratum h. These variance components are defined as:

$$S_{bh}^2 \equiv \frac{1}{N_h} \sum_{u=1}^{N_{1h}} N_{uh}(\bar{z}_{uh} - \bar{z}_h)^2 \tag{26}$$

and

$$S_{wh}^2 \equiv \frac{1}{N_h} \sum_{u=1}^{N_{1h}} \sum_{i=1}^{N_{uh}} (z_{iuh} - \bar{z}_{uh})^2 = \sum_{u=1}^{N_{uh}} \frac{N_{uh}}{N_h} \cdot S_{uh}^2 \tag{27}$$

where:

N_h = total number of sampling elements in stratum h;
N_{1h} = total number of primary units in stratum h;
N_{uh} = total number of sampling elements in unit u of stratum h;
$\bar{z}_h$ = mean value of sample points in stratum h;
$\bar{z}_{uh}$ = mean value of sample points in the uth primary unit of stratum h;
z_{iuh} = value of the ith sample point in the uth primary unit of stratum h;

S_{uh}^2 = variance among elements in primary unit u of stratum h.

As before, the sampling variance is predicted by its ξ-expectation which, using the equality $S_h^2 = S_{bh}^2 + S_{wh}^2$, can be written as:

$$\bar{r} = E_\xi\{V(\hat{z})\} = \sum_{h=1}^{L} w_h^2 \left(\frac{\bar{\gamma}_{Ah} - \bar{\gamma}_{2h}}{n_{1h}} + \frac{\bar{\gamma}_{2h}}{n_{1h} n_{2h}}\right) \tag{28}$$

where $\bar{\gamma}_{2h}$ denotes a weighted mean of semi-variances between all pairs of elements belonging to the same primary unit:

$$\bar{\gamma}_{2h} \equiv \sum_{u=1}^{N_{1h}} \frac{N_{uh}}{N_h} \cdot \bar{\gamma}_{uh} \tag{29}$$

where $\bar{\gamma}_{uh}$ denotes the mean semi-variance between all pairs of elements in unit u of stratum h.

The parameters $\bar{\gamma}_{2h}$ can be determined by the following algorithm:

1. select at random with replacement and equal probabilities one element, say $\boldsymbol{x}_1$, out of all N_h elements in stratum h;
2. select at random with equal probabilities one element, say $\boldsymbol{x}_2$, out of all N_{uh} elements in the primary unit to which $\boldsymbol{x}_1$ belongs;
3. calculate $\gamma_h(\boldsymbol{x}_1,\boldsymbol{x}_2)$;
4. repeat steps 1-3 sufficiently often and calculate the mean of $\gamma_h(\boldsymbol{x}_1,\boldsymbol{x}_2)$ as estimate of $\bar{\gamma}_{2h}$.

In step 4 the variance of $\gamma_h(\boldsymbol{x}_1,\boldsymbol{x}_2)$ can also be calculated. As before, the parameters $\bar{\gamma}_{Ah}$ and their corresponding variances can be determined using the first part of the general simulation algorithm in the previous sub-section. Whether the steps in the simulation are repeated sufficiently often can be determined by calculating the standard error (see Sub-section 7.2.3).

Case. *For both types of sampling designs the corresponding variance components were calculated using the variograms presented in Sub-section 7.2.2. The values of $\hat{\gamma}_{Ah}$ for the four strata were 0.1277, 0.2267, 0.2386 and 0.2485, respectively. For STSI, SI sampling the values of $\hat{\gamma}_{2h}$ were 0.1171, 0.1861, 0.1944 and 0.2013, respectively. These parameter values allowed the models for sampling error prediction to be specified. Figure 7.2 displays $\hat{r}_h$, i.e. the estimated mean squared error in stratum h, as a function of n_h for STSI sampling.*

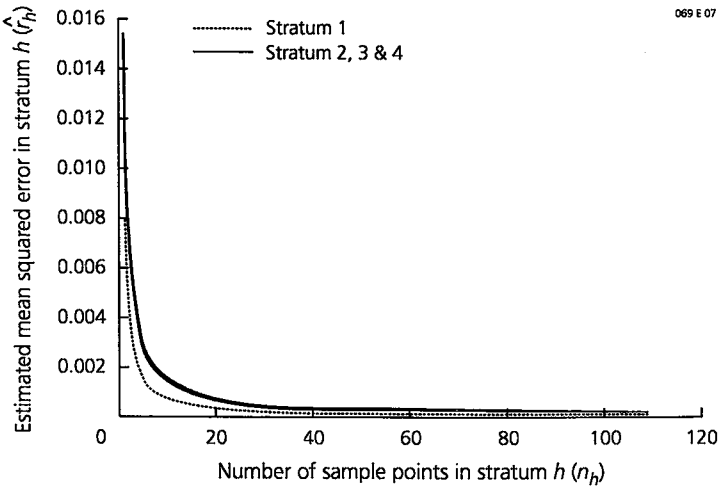


Figure 7.2 Estimated mean squared error, $\hat{r}_h$, as a function of n_h for the four strata

7.3 Prediction of cost

Models are required to predict the cost of realising the spatial inventory for various classes of designs. Sub-section 7.3.1 deals with the cost models available in the literature. These models do not account for differences in spatial patterns of sample points due to different sampling designs, therefore in this thesis an attempt has been made to develop models that do account for differences between classes of sampling designs. Since these models have not yet been tested in practice, they may need to be refined at a later stage. Sub-section 7.3.2 introduces a general cost model. Thereafter, Sub-section 7.3.3 goes into the influence of the time needed for fieldwork, which to a large extent depends on the sampling design. Finally, Sub-section 7.3.4 presents specific cost models for classes of designs.

7.3.1 Cost models in the literature

A general model of the cost of traditional soil mapping is presented by Bie and Beckett (1971). There is no statistical design underlying this soil mapping and the costs are only related to the survey effort of professional staff (in man-days per km^2) which is considered to be largely determined by the intricacy of the soil pattern mapped. The number of augerings needed to produce a proper soil map of a region is related to the number and shape of map units per km^2. The total length of mapped soil boundary, in km per km^2, can be used as a measure of intricacy. Dent and Young (1981: p.97) also state that the survey cost of soil mapping is primarily affected by the salaries of personnel:

"Typically some two-thirds of the total [cost] consists of salaries, the greater part for time spent in the field."

They state that the cost for soil mapping per unit area is mainly influenced by the required map scale: the larger the map scale the more the survey effort (in man-days per km^2).

Bregt et al. (1992) have developed a cost equation in which cost of soil sampling and of analysis of samples are embodied:

$$\text{Total cost} = C_f + \left(\frac{N_o \cdot H}{N_d} \cdot C_s \right) + (N_p \cdot N_s \cdot C_a) \qquad (30)$$

where:

C_f = fixed costs;
C_s = survey cost per day;
C_a = analysis cost per sample;
N_o = number of observations per hectare;
H = the size of the area in hectares;
N_d = number of observations per day;
N_p = number of profiles sampled;
N_s = average number of samples per profile.

Bregt et al. (1992) used a regular grid for sampling, with the aim of producing a map, and the equation is used to predict the costs of surveys at various observation densities.

The cost models referred to above are all related to soil mapping. The influence of the distance between sample points on the cost is included implicitly in these models. These models do not account for the effects of differences in point pattern due to different statistical sampling designs.

In the scope of this thesis, attention is focused on the influence of the sampling design and the method of determination on the inventory cost. In the statistical literature some relatively simple cost models are presented related to sampling designs (e.g. Cochran, 1977) but these are scarcely adjusted to spatial sampling. Here, an attempt is made to model the inventory cost of soil surveys using probability sampling. The influence of the sampling design should be explicit in the cost model, in particular the spatial pattern of sample points forced by the sampling design. For different classes of soil surveys the total inventory cost depends on different cost components or the importance of cost components varies. For example, the distance between sample points has a larger influence in a survey on a regional or national scale than in a survey on a local scale. In accordance with the examples used so far, the modelling of the inventory cost is primarily based on regional soil surveys in the Netherlands.

7.3.2 General cost model

The cost models in this section describe only the cost of spatial inventory, i.e. the cost which - for a specific survey project - can be influenced by the sampling design, and the method of determination. To be able to calculate the total cost of a survey project other cost components also need to be taken into account such as the cost of the design of survey schemes, preparation of field work, data analysis, and report writing. Bregt et al. (1992) added a component to include these fixed costs (Eq. (30)) in the total cost. Since here these fixed costs are assumed to be independent of the outlinear scheme, they are not considered.

The cost of spatial inventory in a soil survey project consists of three main components:

$$C = C_s + C_e + C_\ell \tag{31}$$

where:

C = cost of spatial inventory (in US dollars, $);

C_s = total cost of field work, or survey cost ($); these costs are (mainly) determined by the salaries of personnel;

C_e = total cost of equipment ($); this component is of importance if special equipment has to be hired to execute the survey, which can be very expensive;

C_ℓ = total cost of laboratory analysis ($), including the cost of material used for the samples.

C_e was not distinguished by Bregt et al. (1992); their model was only adapted to their requirements. However, historical surveys at the DLO Winand Staring Centre showed that this component may be considerable, and should therefore be distinguished from the survey cost. Taking into account the relation of cost to the number of sample points and to the time needed for field work, which may differ between strata, this equation can be refined to:

$$C = \sum_{h=1}^{L} t_h \cdot c_s + \sum_{h=1}^{L} t_h \cdot c_e + c_\ell \sum_{h=1}^{L} n_h = (c_s + c_e) \cdot \sum_{h=1}^{L} t_h + n \cdot c_\ell \tag{32}$$

where:

L = number of strata;

h = stratum number;

c_s = survey cost per hour ($ per hour);

c_e = cost of equipment per hour ($ per hour);

c_ℓ = cost of laboratory analyses per sample point ($ per sample point);

t_h = total time needed for field work in stratum h (hours);

n = sample size;

n_h = number of sample points in stratum h.

This cost model assumes the survey cost per hour, the cost of equipment per hour, and the laboratory cost per sample point to be the same for all strata, which is not necessarily the case. However, these assumptions are made provisionally to restrict the required number of parameters. If, at a later stage, it turns out to be desirable to use different values for these parameters per stratum, the cost model can be easily extended.

In Eq. (30) the analysis costs were related to the number of profiles sampled and the average number of samples per profile. For simplicity, in Eq. (32) the total laboratory cost depends on the number of sample points and the value of c_ℓ. In Eq. (32) the total survey cost and the total cost of equipment are determined by the total time needed for field work and the values of c_s and c_e; The total time needed for field work, which is among other things related to the number of sample points, should be modelled more precisely to make the influence of the sampling design on the cost of spatial inventory explicit.

7.3.3 The influence of time

For regional soil surveys in the Netherlands the total (expected) time needed for field work, t_h, depends on the following factors.

- The sampling design, e.g. the location of sample points (e.g. stratified, simple random, clustered), and the total number of sample points.
- The type of sampling element and the method of determination: these two factors determine the sampling equipment to be used, and the depth of the profile to be sampled. Some equipment is easier to handle. If soil samples need to be analysed in a laboratory, sampling in the field usually takes more time and the surveyor has to return more often to his or her car because the number of samples one can carry is limited, and the samples may otherwise get damaged.

  Another point related to taking samples is whether the samples should be delivered to the laboratory every day. This may also influence the number of sample points visited per day.
- The survey region: in the Netherlands a special distinction should be made between peaty regions and sandy or clay regions. The time needed for access to sample points is relatively large in a peaty region due to the many wide ditches.
- The season, and related to this the height of the crops: in the Netherlands autumn is the most suitable period for regional surveys. Then, the crops are harvested, and there is often a period of relatively dry weather. Rain generally obstructs field work. However, not all surveys are sensitive to seasonal influences.
- Time to ask permission from the owner of the land: in regional surveys landowners need to be informed about the purpose of the survey and have to be asked permission for sampling on their property. If owners refuse to grant access or if they set conditions the outcome of the survey may be biased.

Three of these factors cannot be influenced by the design of a survey scheme, although they can significantly influence the total inventory cost: the survey region is defined in the aim, the season for field work is often prescribed, and the willingness of land owners to co-operate is determined by external factors. Only the type of sampling element and the method of determination are established during the design of an outlinear survey scheme (Chapter 6).

The time needed for field work in stratum h can roughly be considered as the sum of two components: the time needed for access to sample points and the total time needed for observation and/or taking samples at the points, i.e. the observation time (Eq. (33)):

$$t_h = t_{ah} + t_{oh} \tag{33}$$

where:

t_{ah} = time for access to sample points in stratum h influenced by: the sampling design (random, clustered, two-stage, and the number of sample points and/or subsets), the method of determination, the survey region, the season, and asking permission (hours);

t_{oh} = observation time in stratum h for a given outlinear scheme, related to: the number of sample points, the type of sampling element, the method of determination, and the survey region (hours).

Substituting Eq. (33) into Eq. (32) gives:

$$C = (c_s + c_e) \cdot \sum_{h=1}^{L} (t_{ah} + t_{oh}) + c_t \cdot \sum_{h=1}^{L} n_h \qquad (34)$$

According to this equation the total inventory cost can be found by adding up the costs of the individual strata. The costs of access to strata are not incorporated separately - they should be included implicitly in the cost per stratum. If there are no strata, L equals 1. Eq. (34) is suitable for most classes of designs considered in this thesis. However, in case of designs starting with a two-stage stratified approach another cost component which is independent of stratification influences the total access time, namely the time for access to primary units. In these cases the access time within primary units needs to be summed over the strata and multiplied with the number of primary units in the sample, and the total access time results from increasing the access time within primary units with the time needed for access to primary units. So, for these types of designs Eq. (34) can be rewritten as:

$$C = (c_s + c_e) \cdot \{t_{a1} + n_1 \cdot \sum_{h=1}^{L} (t_{ah} + t_{oh})\} + c_t \cdot \sum_{h=1}^{L} n_h \qquad (35)$$

where:

t_{a1} = time for access to primary units in the survey region related to: the number of primary units in the sample, the method of determination, the survey region, the season, and asking permission (hours);

n_1 = number of primary units in the sample.

Eq. (35) assumes the number of sample points per primary unit to be fixed.

7.3.4 Specific cost models

In the general models, Eq. (34) and Eq. (35), the components related to the time needed for field work can be worked out more precisely as functions of the number of points and units to be selected. The observation time in a stratum (t_{oh}) is independent of the sampling design and can therefore in general be written as the product of the observation time per sample point in stratum h, $\bar{t}_{oh}$, which can usually be estimated rather precisely, and the number of sample points in this stratum: $t_{oh} = \bar{t}_{oh} \cdot n_h$. The components representing the time needed for access (t_{ah} and t_{a1}) remain complex, since they are related to a number of parameters. Since the system should be able to compare possible designs and types of designs, the influence of the sampling design on access time needs to be made more explicit. This influence is first of all related to the specified numbers of elements (sample points) and subsets of elements (primary units or clusters) to be selected per stratum. Furthermore, it depends on the mean distances between these elements and/or sample points or clusters. The assumption is made that the influence on the survey cost of the way in which the samples are selected can be ignored, since the models should not be too complicated to start with.

Sub-models for the main classes of designs are worked out and depicted in Table 7.2. These sub-models start out with the assumption that in the case of stratification or two-stage sampling the same design will be followed within strata or primary units. In the case of two-stage designs the area of primary units does not need to be fixed but the number of sample

Table 7.2 Models of access time for different classes of sampling designs

class of designs	model of access time related to class of designs
SI	$t_{aA} = t_{a0A} \cdot \sqrt{(A \cdot n)}$
SIC	$t_{aA} = t_{a0A} \cdot \sqrt{(A \cdot n_0)} + (t_{a3A} \cdot n_4) \cdot n_3$
SY	$t_{aA} = t_{a0A} \cdot \sqrt{(A)} + (t_{a3A} \cdot n_4) \cdot n_3$
STSI	$t_{ah} = t_{a0h} \cdot \sqrt{(A_h \cdot n_h)}$
STSIC	$t_{ah} = t_{a0h} \cdot \sqrt{(A_h \cdot n_{3h})} + (t_{a3h} \cdot n_{4h}) \cdot n_{3h}$
STSY	$t_{ah} = t_{a0h} \cdot \sqrt{(A_h)} + (t_{a3h} \cdot n_{4h})$
STSI, SI	$t_{ah} = t_{a1h} \cdot \sqrt{(A_h \cdot n_{1h})} + t_{a2h} \cdot n_{1h} \cdot \sqrt{n_{01h}} \cdot \sum_u \{(A_{uh} / A_h) \cdot \sqrt{A_{uh}}\}$
STSI, SIC	$t_{ah} = t_{a1h} \cdot \sqrt{(A_h \cdot n_{1h})} + [t_{a2h} \cdot \sqrt{n_{31h}} \cdot \sum_u \{(A_{uh} / A_h) \cdot \sqrt{A_{uh}}\} + (t_{a3h} \cdot n_{4h}) \cdot n_{31h}] \cdot n_{1h}$
STSI, SY	$t_{ah} = t_{a1h} \cdot \sqrt{(A_h \cdot n_{1h})} + [t_{a2h} \cdot \sum_u \{(A_{uh} / A_h) \cdot \sqrt{A_{uh}}\} + (t_{a3h} \cdot n_{4h}] \cdot n_{1h}$
SI, SI	$t_{aA} = t_{a1A} \cdot \sqrt{(A \cdot n_1)} + t_{a2A} \cdot \sqrt{n_{01}} \cdot \sum_u \{(A_u / A) \cdot \sqrt{A_u}\} \cdot n_1$
SI, SIC	$t_{aA} = t_{a1A} \cdot \sqrt{(A \cdot n_1)} + [t_{a2A} \cdot \sqrt{n_{41}} \cdot \sum_u \{(A_u / A) \cdot \sqrt{A_u}\} + (t_{a3A} \cdot n_4) \cdot n_{41}] \cdot n_1$
SI, SY	$t_{aA} = t_{a1A} \cdot \sqrt{(A \cdot n_1)} + [t_{a2A} \cdot \sum_u \{(A_u / A) \cdot \sqrt{A_u}\} + (t_{a3A} \cdot n_4)] \cdot n_1$
SI, STSI	$t_{a1} = t_{a1A} \cdot \sqrt{(A \cdot n_1)}$; $t_{ah} = t_{a0h} \cdot \sqrt{(A_h \cdot n_{01h})}$
SI, STSIC	$t_{a1} = t_{a1A} \cdot \sqrt{(A \cdot n_1)}$; $t_{ah} = t_{a0h} \cdot \sqrt{(A_h \cdot n_{3h1})} + (t_{a3h} \cdot n_{4h}) \cdot n_{3h1}$
SI, STSY	$t_{a1} = t_{a1A} \cdot \sqrt{(A \cdot n_1)}$; $t_{ah} = t_{a0h} \cdot \sqrt{(A_h)} + (t_{a3h} \cdot n_{4h})$

variables:

t_{aA} = access time within the survey region A (hours); $t_{aA} = t_{ah}$ for $L = 1$

t_{ah} = access time within stratum h (hours)

t_{a1} = access time to primary units (hours)

t_{a0h}, t_{a0A} = access time per kilometre to randomly selected points in stratum h or region A including location time, influence of the survey region, and time to ask permission (hours/km)

t_{a3h}, t_{a3A} = access time between two successive points in a cluster in stratum h or region A (hours)

t_{a1h}, t_{a1A} = access time per kilometre to selected primary units in stratum h or region A (hours/km)

t_{a2h}, t_{a2A} = access time per kilometre between selected secondary units (random points) within a primary unit in stratum h or in region A (hours/km)

A (A_h) = area of the survey region (or stratum h) (km^2)

A_u (A_{uh}) = area of primary unit u (in stratum h) (km^2)

n = sample size, i.e. the number of sample points

n_0 = number of random points to be selected

n_h = number of sample points in stratum h

n_3 = number of clusters in the sample

n_{3h} = number of clusters to be selected in stratum h

n_{31h} = number of clusters to be selected per selected primary unit in stratum h

n_{3h1} = number of clusters to be selected in stratum h in a selected primary unit

n_4 = number of successive points per cluster (randomly selected starting point excluded)

n_{4h} = number of successive points per cluster in stratum h (randomly selected starting point excluded)

n_1 = number of primary units to be selected

n_{1h} = number of primary units to be selected in stratum h

n_{41} = number of clusters to be selected in selected primary units

n_{01} = number of random points to be selected per selected primary unit

n_{01h} = number of random points to be selected per selected primary unit in stratum h

n_{0h1} = number of random points to be selected in stratum h in a selected primary unit

$\sum_h$ = sum from $h = 1$ to L, where L is the number of strata

$\sum_u$ = sum from $u = 1$ to N_1(or N_{1h}), where N_1(or N_{1h}) is the number of primary units in A (or in stratum h)

points per primary units is considered to be constant. There is also a basic assumption that strata are contiguous areas, which is not necessarily the case in practice. If strata are non-contiguous, the access time to different parts of a stratum should be included in the component for the access time within that stratum.

The models in Table 7.2 enable a better understanding of the effects of designs on the access time and thus on the inventory cost. For example, distinctions are made between independently selected random sample points and grouped sample points (in primary units or clusters), and between the location of randomly selected starting points and that of successive points in a cluster or systematic sample.

The mean distance between two neighbouring points or units in a random configuration is approximated by: $\sqrt{(A/n)}$, where A is the area of the survey region, and n is the number of sample points or units. This approximation is based on the assumption that the average area per sample point A/n is a square region. Then, the shortest distance between two adjacent points can be approximated by the distance between the centres of two adjacent squares of size A/n: $\sqrt{(A/n)}$. This distance should be multiplied by the number of elements or units to be visited, resulting in the total distance to be covered: $n \cdot \sqrt{(A/n)} = \sqrt{(A \cdot n)}$. This relation between area and number of sample points has been found earlier by Beardwood et al. (1959), who prove that the length of the shortest closed path through n points in a bounded region of area A is "*almost always*" asymptotically proportional to $\sqrt{(A \cdot n)}$ for a large n. It is obvious that this may be a rather rough approximation in situations with a small n. Besides, neglecting the shape of the survey region may also cause distortion. However, the objective here is to develop cost models reflecting the relation of sampling design to inventory cost in which the specific approximation of the mean distance between two near units or points is provisionally of secondary importance.

Specification of the components for access time, t_{a0h}, t_{a3h}, t_{a1h} and t_{a2h}, may be difficult, however there are never more than three t_{ah}'s required. The reason for introducing different t_{ah}'s is the objective to explicitly include in these models components which can be influenced by the class of sampling designs. For a given sampling design the t_a-values are fixed and have to be supplied only once, whereas the sample size may change. (However, if a large variation in n is allowed, one value for t_a will probably not suffice.) During the lifetime of the system information on t_{ah}'s will be collected and stored for re-use in new survey projects.

The cost of an outlinear survey scheme can be predicted by: selecting the corresponding sub-model of the access time from Table 7.2, implementing it into Eq. (34) or (35), and inserting the corresponding values. Most models in Table 7.2 can replace t_{ah} in Eq. (34). For example, in the case of *STSI, SIC* sampling Eq. (34) can be rewritten as:

$$C = \sum_{h=1}^{L} [\, t_{a1h} \cdot \sqrt{(A_h \cdot n_{1h})} + \{ t_{a2h} \cdot \sqrt{n_{31h}} \cdot \sum_{u=1}^{N_{1h}} (\frac{A_{uh}}{A_h} \cdot \sqrt{A_{uh}}) + (t_{a3h} \cdot n_{4h}) \cdot n_{31h} \} \cdot n_{1h} + \bar{t}_{oh} \cdot n_h] \cdot (c_s + c_e) + c_\ell \cdot \sum_{h=1}^{L} n_h \qquad (36)$$

In the case of two-stage stratified designs summation over strata is only required within the primary units, not between primary units (Eq. (35)). For these types of designs two variables in Eq. (35), t_{a1} and t_{ah}, have to be replaced with the sub-models shown in Table

7.2. For example, the final cost model for *SI, STSI* sampling can be found by rewriting Eq. (35) as:

$$C = [t_{a1A}\cdot\sqrt{(A\cdot n_1)} + n_1\cdot\sum_{h=1}^{L}\{(t_{a0h}\cdot\sqrt{(A_h\cdot n_{01h})}\} + n_1\cdot\sum_{h=1}^{L}(\bar{t}_{oh}\cdot n_h)]\cdot(c_s + c_e) + c_\ell\cdot\sum_{h=1}^{L} n_h \qquad (37)$$

It should be noted that the actual values of components for the access time and observation time are also influenced by the surveyor: a more experienced surveyor may need less time that a novice. Differences between surveyors always exist. The system should deal with values for the 'average' surveyor.

As pointed out at the beginning of this section, the emphasis here is on regional soil surveys in the Netherlands. In other countries there may be other factors which influence the total inventory cost and the access-time parameters (t_{ah}'s); however, the sub-models describing the relations between classes of designs and the total access time are generally applicable. This is also true for local surveys where total access time appears to be limited; nevertheless it is clear that in the case of local surveys a distinction should also be made between localizing each sample point separately (*SI* sampling) or localizing only a starting point and finding the other points by pacing (*SIC* or *SY* sampling). So these models may also be usable for local surveys, although the values of the access-time parameters and their relative importance in the cost of spatial inventory will differ significantly from the values in regional surveys.

Case.

Values for the cost parameters were derived from the historical case; since this was a survey in the Netherlands, the cost parameters are expressed in Dutch guilders (Dfl.) instead of US dollars: $c_s = 87.5$ *Dfl.,* $c_e = 0$, $c_\ell = 170$ *Dfl. In the historical survey on which case A was based a log-book was kept during field work. This made it possible to estimate values for the time parameters. In the fictitious case used in this chapter, the differences in the time needed for access and the time needed for observation were assumed to be the same in all strata:* $\bar{t}_{oh} = 0.42$ *(for h=1,2,3,4).*

STSI sampling: $t_{a01} = t_{a02} = 3$; $t_{a03} = 3.5$; $t_{a04} = 2.5$.

STSI, SI sampling: $t_{a11} = t_{a12} = t_{a21} = t_{a22} = 3$

$t_{a13} = t_{a23} = 3.5$; $t_{a14} = t_{a24} = 2.5$

Figure 7.3 depicts the cost of spatial inventory per stratum, c_h, *as a function of* n_h *for STSI sampling.*

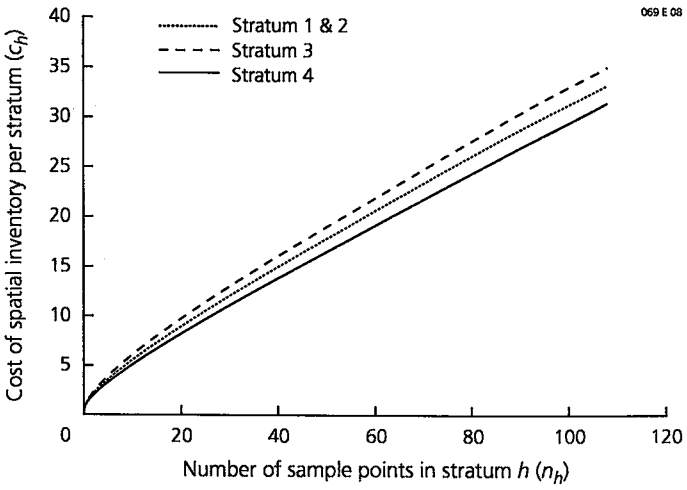


Figure 7.3 Cost of spatial inventory per stratum, c_h, as a function of n_h for STSI sampling:
$c_h = 87.5 \cdot \{t_{a0h} \cdot \sqrt{(16 \cdot n_h)} + 0.42 \cdot n_h\} + 170 \cdot n_h$

7.4 The search for an optimal scheme

In the preceding sections, models for predicting the accuracy and cost of survey schemes have been introduced. These models can be used to search for an efficient and, if possible, an optimal scheme within an outlinear scheme taking into account the constraints defined earlier. It should be noted that the models for predicting accuracy and cost are non-linear, as the relationship between sampling error and inventory cost on the one hand, and the number of sample points on the other is not a linear one. A notable feature of the models for stratified sampling, which is frequently used in soil surveys, is that they are separable functions, i.e. they can be written as a sum of functions of the individual strata (e.g. Eq. (24) and Eq. (34)), whereas parameter values for the strata may differ. These characteristics are relevant to the choice of an optimization technique.

This section deals with the search for an optimal scheme, which starts with formulating the problem (Sub-section 7.4.1). Thereafter, a theoretical exposition of an OR technique which is frequently used for optimizing problems with separable, non-linear functions follows: dynamic programming (Sub-section 7.4.2). These characteristics hold for classes of designs that start with stratification. Since stratification is often used in soil surveying, this is an attractive category to start with when looking into the options for optimizing outlinear survey schemes. The next sub-section (7.4.3) deals with the various approaches to optimization required for the classes of designs considered in this thesis. Finally, Sub-section 7.4.4 presents the mathematical models for two classes of sampling designs and the results of optimizing the fictitious case.

7.4.1 Problem formulation

The objective of optimization is to minimize the predicted sampling error, i.e. to maximize the accuracy of results, for a fixed budget within the space of feasible solutions which is delimited by the constraints. This is referred to as the first problem. It is also possible to search for a scheme which minimizes the survey cost for a given accuracy constraint (the second problem). The same procedure can be used to solve both problems. This section focuses on solving the first problem.

When considering the problem of minimizing the sampling error for a fixed budget, besides the financial constraint there may be a constraint on the maximal allowable number of sample points, e.g. due to limited laboratory capacity (capacity constraint). If there is no strict capacity constraint, the maximal allowable number of sample points depends on the budget available and on the applicable cost model.

In any case, the budget and the possibly limited number of sample points should be allocated to the survey region according to the directions of the outlinear scheme. The numbers of sample points, and/or primary units, and/or clusters which provide a sampling design which is optimal under existing conditions need to be calculated.

7.4.2 Dynamic programming

An introduction to the mathematical theory of multi-stage decision processes, in which the term dynamic programming (DP) is introduced, is given by Bellman (1957). DP provides a procedure for making a number of interrelated decisions which produce an optimal solution. As far as known, DP has not been applied before to optimize schemes for soil surveys using probability sampling. It seems attractive to formulate the optimization problem stated above using a DP approach. Hillier and Lieberman (1990: pp. 398-401) present eight features to characterize DP problems, which are summarized below:

1. the problem can be divided into stages, and at each stage a decision is required;
2. at each stage there are a number of associated states;
3. the decision at a stage transforms the current state into a state associated with the next stage (possibly according to a probability distribution);
4. the solution procedure is used to produce an optimal policy for the overall problem, i.e. a prescription of the optimal decision at each stage for each of the possible states;
5. the principle of optimality holds: given the current state an optimal decision for the remaining stages is independent of the decisions in previous stages;
6. the problem solving starts with identifying the optimal decision for the last stage;
7. a recursive relationship (often called Bellman relationship) is used to find the optimal decision for stage s, given the optimal decision for stage $(s+1)$; there is no standard form for this relationship;
8. the solution procedure finds the optimal decision at each stage, and keeps moving backward until the optimal policy starting at the initial stage is identified.

Figure 7.4 depicts two examples of survey scheme problems with four strata for which an optimal schemes can be found using DP. Again, *STSI* designs are considered and in the examples the survey region has been subdivided into four strata. In the first example, 7.4A, there is only a financial constraint, i.e. the budget is limited; in the second example, 7.4B,

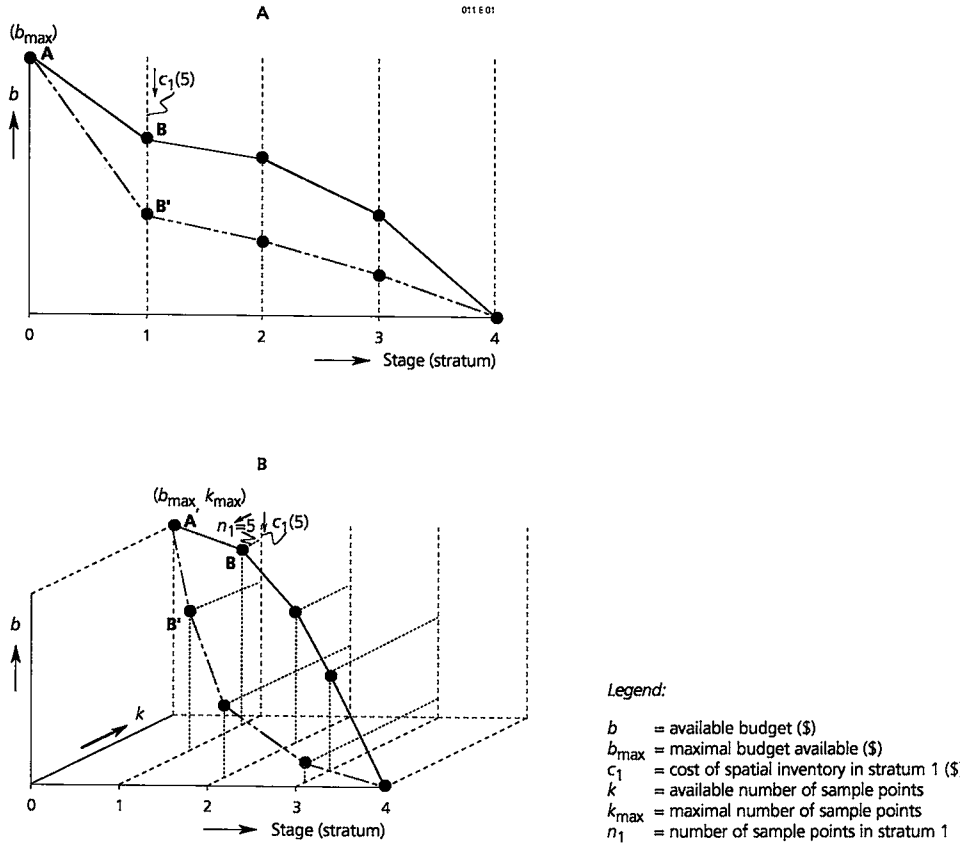


Figure 7.4 Survey scheme problems with four strata: A. One-dimensional dynamic programming, B. Two-dimensional dynamic programming

besides a budgetary constraint there is a capacity constraint, i.e. the maximum number of sample points is fixed, e.g. due to limited laboratory capacity. Figure 7.4A shows a one-dimensional DP problem, and figure 7.4B a two-dimensional DP problem. In the next paragraph these two examples are elucidated with references to the DP characteristics listed above.

In DP problems the stages are related to the number of decisions to be made to find the solution for the whole problem. These stages may be subsequent periods in time with a fixed sequence, but sometimes there is no fixed sequence, e.g. in the case of distributing medical teams to countries, the countries can be considered as stages in the DP formulation (Hillier & Lieberman, 1990).

(1) In the search for an efficient sampling scheme using DP the strata can be considered as stages and a decision applies to the number of sample points allocated to the next stratum and the corresponding amount of budget spent. (2) In Figure 7.4A the state consists of the available budget (*b*) at a stage; in Figure 7.4B it consists of the combination of the available

number of sample points (k) and the available budget (b) at a stage (state vector). (3) The allocation of sample points and budget starts at stage 0, where there is only one state, point A, i.e. a maximal budget, b_{max} (and maximal number of sample points, k_{max}). At stage 1 there are a number of states which can be achieved by the allocation of different numbers of sample points and corresponding budgets to that stage, i.e. stratum 1. For example, if five sample points are allocated to stratum 1, i.e. n_1 equals 5, the available budget for the state at stage 2 will be reduced by the cost of taking five samples in stratum 1: $c_1(5)$ (and in Figure 7.4B the available number of sample points will decrease by five). Then, the state of the system reaches point B; another decision for the first stage leads to point B'. (5) The way in which the budget left in stage 1 is allocated to the remaining stages (2, 3, and 4) is independent of the previous decision. Two possible allocations of sample points and budget are shown for both DP problems. The number of allowable states at a stage can be restricted by constraints such as a minimum number of points per stratum. (4) For each decision which results in an allowable state the contribution to the sampling error can be calculated. The combination of decisions for all stages, i.e. the allocation of budget (and sample points) to all strata, which produces the smallest total prediction of the sampling error, gives the optimal sampling design. (6, 7, and 8) This combination can be found using the backward-moving, recursive relationship or Bellman relationship.

Since the predicted sampling error always decreases with increasing sample size, the entire budget should be used if there are no capacity constraints. At stage 4 in Figure 7.4B there is no budget left, and the sum of the numbers of sample points allocated to the strata equals k_{max}. This is not necessarily true under all circumstances. In the case of a two-dimensional problem, for example, the cost constraint may be more restrictive than the capacity constraint. So at the final stage there may be space left in only one of the directions, i.e. budget or capacity. In the case of one-dimensional optimization there will also be some budget left if it is insufficient for an additional sample point.

The introduction to DP given so far assumed that the problems to be solved are *deterministic*. Then, the state of the next stage is completely determined by the state and the policy decision at the current stage. If the result is influenced by a probability distribution and not just by the state and the policy decision at the current state, *stochastic* DP can be applied. The probability distribution, however, should be completely determined by the state and policy decision at the current stage. If, for example clusters of unequal sizes are selected, then for a fixed number of clusters the sample size is stochastic. For a given definition of clusters the probability distribution of their size can be determined from the sampling frame. Compared to deterministic DP, one-dimensional and two-dimensional stochastic DP problems can be distinguished. Solving stochastic DP problems is obviously more complicated and time-consuming than solving deterministic DP problems.

The aim here is to present a procedure for optimizing (at least) frequently occurring survey schemes within the support system. This procedure will certainly need to be refined or adapted to be applicable to a larger category of schemes. The optimization of more complex designs is beyond the scope of this thesis. The mathematical formulations described in Sub-section 7.4.4 focus on deterministic DP.

7.4.3 Various approaches to optimizing outlinear survey schemes

As stated in the previous sub-sections, the DP approach is particularly appropriate for optimizing stratified designs. DP may not be appropriate for optimizing other classes of designs.

If the type of systematic pattern for a *SY* design has been specified, and the accuracy needs to be maximized for the available budget, then the sample size should be as large as the budget allows. In that case the sample size allowed determines the distance between adjacent points in the systematic pattern, which results in the 'optimal' design unless the period of the grid coincides with periodicity in the field (see Sub-section 6.3.2).

In the taxonomy of classes of designs (Fig. 6.3), a number of designs were mentioned in which there is only one stage, i.e. the one and only 'stratum' encompasses the whole survey region. In these cases DP can also be applied to optimize survey schemes. However, in the cases of *SI* sampling, *SIC* sampling, *SI, SI* sampling with a fixed number of sample points per primary unit, *SI, SIC* sampling with a fixed number of clusters per primary unit, or *SI, SY* sampling, a short cut can be used. This employs the fact that the sampling error decreases when the sample size increases. Therefore, if the aim is to minimize the sampling error, the sample size should be as large as the budget allows. For a given budget the cost model of these designs can simply be used to calculate the largest possible sample size, resulting automatically in the optimal design for the given constraints. There is no need to evaluate other possibilities. If, in the case of two-stage sampling, the number of sample points or clusters per primary unit is not fixed, the best allocation of primary and secondary units under the existing conditions should be calculated, which will require a number of calculations and comparisons.

A special category of designs consists of two-stage stratified designs. These designs are rarely used and, as stated in Sub-section 7.3.3, they require different cost models. Furthermore, prediction of the sampling error of these designs puts heavy demands on the prior information required: a variogram for the survey region as a whole is required besides variograms for the strata. It should be possible to describe both the variance between and within (stratified) primary units. This difficult issue was not considered in Section 7.2. Moreover it seems impossible to optimize these designs using a standard DP approach: both the number of primary units and the number of second-stage units per stratum within primary units need to be optimized. These situations require more complex optimization procedures, which is beyond the scope of this thesis.

7.4.4 Mathematical models

There is no standard mathematical formulation for solving DP problems. The proposed approach to search for an optimal survey scheme requires a specific formulation for various types of sampling designs. The models for sampling-error prediction can be constructed for different types of designs. The cost models which are specified for classes of designs are applicable to all types of designs within a class. In this sub-section the models are elaborated for two types of designs, both with only a financial constraint or a financial and a capacity constraint. The results of optimizing the two outlinear schemes in the fictitious case are also presented.

STSI sampling. First, the problem is considered that corresponds with the one-dimensional situation depicted in Figure 7.4A. The aim is to minimize the sampling error for a given budget, which results in the following objective function:

$$\text{Min } \bar{r} = \text{Min} \sum_{h=1}^{L} g_h(n_h) \tag{38}$$

where:

$g_h(n_h)$ = contribution from stratum h to the sampling-error prediction if n_h sample points are selected in stratum h.

For a *STSI* design $g_h(n_h)$ can be written according to Eq. (24) in section 7.2.4:

$$g_h(n_h) = \frac{w_h^2}{n_h} \cdot \bar{\gamma}_{Ah} \tag{39}$$

The budget is fixed, so there is a financial constraint:

$$\sum_{h=1}^{L} c_h(n_h) \leq b_{max} \tag{40}$$

where:

$c_h(n_h)$ = cost of spatial inventory in stratum h in case of n_h sample points (\$);

b_{max} = maximal budget available (\$).

There may be additional constraints per stratum such as: $n_{min\,h} \leq n_h$, where $n_{min\,h}$ is the minimum number of sample points in stratum h required, to influence the variance of the survey results per stratum. To be able to estimate the variance per stratum, $n_h \geq 2$.

The model for the cost of spatial inventory in stratum h for a *STSI* design can be derived by inserting the corresponding sub-model of the access time (Table 7.2) into Eq. (34) in section 7.3.3 omitting the summations over strata:

$$c_h(n_h) = (c_s + c_e) \cdot (t_{a0h} \cdot \sqrt{(A_h \cdot n_h)} + \bar{t}_{oh} \cdot n_h) + c_l \cdot n_h \tag{41}$$

The recursive relationship, or value function, which can be used to calculate an optimal scheme for *STSI* sampling is:

$$V_h(b_h) = \underset{n_{h+1} \in X_{h+1}}{\text{Min}} \left[g_{h+1}(n_{h+1}) + V_{h+1}\{b_h - c_{h+1}(n_{h+1})\} \right] \tag{42}$$

where:

b_h = available budget, or financial state, at stage h (\$);

$V_h(b_h)$ = contribution of stages h+1, h+2, ... , L to the objective function if the system starts in state b_h at stage h, the immediate decision is n_{h+1}, and optimal decisions are

made thereafter;

and the following set is defined:

$X_{h+1} = \{n_{h+1} \mid n_{h+1} \geq n_{\min h+1} \;\&\; c_{h+1}(n_{h+1}) \leq b_h\}$

In the two-dimensional case (Fig. 7.4B) the aim does not change so the same objective function holds:

$$\text{Min}\,\bar{r} = \text{Min}\sum_{h=1}^{L} g_h(n_h) \tag{38}$$

There are two constraints:

$$\sum_{h=1}^{L} c_h(n_h) \leq b_{\max} \qquad \text{(financial constraint)} \tag{39}$$

and:

$$\sum_{h=1}^{L} n_h \leq k_{\max} \qquad \text{(capacity constraint)} \tag{43}$$

where:

$k_{\max}$ = maximal number of sample points.

In this two-dimensional case the recursive relationship can be formulated as:

$$V_h(b_h, k_h) = \min_{n_{h+1} \in Y_{h+1}} [g_{h+1}(n_{h+1}) + V_{h+1}\{b_h - c_{h+1}(n_{h+1}), k_h - n_{h+1}\}] \tag{44}$$

where:

b_h = available budget, or financial state, at stage h (\$);

k_h = available number of sample points, or capacity state, at stage h;

$V_h(b_h,k_h)$ = contribution of stages h+1, h+2, ... , L to the objective function if the system starts in state (b_h,k_h) at stage h, the immediate decision is n_{h+1}, and optimal decisions are made thereafter;

and the set Y_{h+1} is defined as:

$Y_{h+1} = \{n_{h+1} \mid n_{\min h+1} \leq n_{h+1} \leq k_h \;\&\; c_{h+1}(n_{h+1}) \leq b_h\}$

STSI, SI sampling. The selection of sampling elements from each stratum h proceeds in the following two stages:

1. select at random n_{1h} primary units with replacement and probabilities proportional to their sizes;
2. select at random from primary units each time they have been selected n_{01h} sample points with replacement and equal probabilities.

Both n_{1h} and n_{01h} can be optimized per stratum. Then the following equation can be used for sampling-error prediction per stratum (according to Eq. (28) in Sub-section 7.2.4):

$$g_h(n_{1h}, n_{01h}) = w_h^2 \left(\frac{\bar{\gamma}_{Ah} - \bar{\gamma}_{2h}}{n_{1h}} + \frac{\bar{\gamma}_{2h}}{n_{1h} n_{01h}}\right) \tag{45}$$

and the corresponding cost model is:

$$c_h(n_{1h}, n_{01h}) = \left[t_{a1h} \cdot \sqrt{(A_h \cdot n_{1h})} + t_{a2h} \cdot n_{1h} \cdot \sqrt{n_{01h}} \cdot \sum_{u=1}^{N_{1h}} \left\{\left(\frac{A_{uh}}{A_h}\right) \cdot \sqrt{A_{uh}}\right\} + \bar{t}_{oh} \cdot n_h\right] \cdot (c_s + c_e) + c_t \cdot n_h \tag{46}$$

where: $n_h = n_{1h} \cdot n_{01h}$.
When only a budgetary restriction influences the optimization, the following problem formulation holds. The objective function can be formulated as:

$$\text{Min } \bar{r} = \text{Min } g_h(n_{1h}, n_{01h}) \tag{47}$$

and there is a financial constraint:

$$\sum_{h=1}^{L} c_h(n_{1h}, n_{01h}) \leq b_{\max} \tag{48}$$

There may be additional constraints per stratum such as: $n_{\min 1h} \leq n_{1h}$, where $n_{\min 1h}$ is the minimum number of primary units to be selected in stratum h, or $n_{\min 01h} \leq n_{01h}$, where $n_{\min 01h}$ is the minimum number of random points to be selected per selected primary unit in stratum h. The corresponding value function is as follows:

$$V_h(b_h) = \underset{n_{h+1} \in Q_{h+1}}{\text{Min}} \left[g_{h+1}(n_{1(h+1)}, n_{01(h+1)}) + V_{h+1}\{b_h - c_{h+1}(n_{1(h+1)}, n_{01(h+1)})\}\right] \tag{49}$$

where the set Q_{h+1} is defined as:

$$Q_{h+1} = \{n_{h+1} \mid n_{h+1} = n_{1(h+1)} \cdot n_{01(h+1)} \;\&\; n_{1(h+1)} \geq n_{\min 1(h+1)} \;\&\; n_{01(h+1)} \geq n_{\min 01(h+1)} \;\&\; c_{h+1}(n_{1(h+1)}, n_{01(h+1)}) \leq b_h\}$$

However, it may also occur that, besides a financial constraint, a capacity constraint should also be taken into account:

$$\sum_{h=1}^{L} (n_{1h} \cdot n_{01h}) \leq k_{\max} \tag{50}$$

Again there may be additional constraints such as: $n_{\min 1h} \leq n_{1h}$, or $n_{\min 01h} \leq n_{01h}$ which influence the variance of the survey results.

This results in the following value function:

$$V_h(b_h, k_h) = \underset{n_{h+1} \in R_{h+1}}{\text{Min}} [g_{h+1}(n_{1(h+1)}, n_{01(h+1)}) + V_{h+1}\{b_h - c_{h+1}(n_{1(h+1)}, n_{01(h+1)}), k_h - (n_{1(h+1)} \cdot n_{01(h+1)})\}] \quad (51)$$

where the set R_{h+1} is defined as:

$$R_{h+1} = \{n_{h+1} \mid n_{h+1} = n_{1\,(h+1)} \cdot n_{01\,(h+1)} \;\&\; n_{1\,(h+1)} \geq n_{\min 1\,(h+1)} \;\&\; n_{01\,(h+1)} \geq n_{\min 01\,(h+1)} \;\&\; n_{1\,(h+1)} \cdot n_{01\,(h+1)} \leq k_h \;\&\; c_{h+1}(n_{1\,(h+1)}, n_{01\,(h+1)}) \leq b_h\}$$

Two-stage sampling is generally applied to reduce cost, and therefore the number of secondary units to be measured per primary unit should result in an integer number of primary units to be completed per day. Therefore n_{01h} may be fixed, leaving only the value of n_{1h} to be calculated per stratum. This results in a mathematical formulation comparable to the formulation presented for *STSI* sampling: only one variable per stratum needs to be optimized. This is a simpler problem formulation, which can be optimized for different values of n_{01h} and so also assist in finding an allocation of primary and secondary units over the strata which suits the constraints best. Such an approach would ensure that designs are only optimized when they seem to be operationally advantageous.

The models presented above make it possible to search for an optimal sampling design within a given outlinear scheme. A disadvantage of DP is that there is no standard mathematical formulation. For optimizing different types of sampling designs (slightly) different value functions are required. Solving the second problem, i.e. minimizing the survey cost for an accuracy constraint, also requires different value functions.

Case. *A DP program was written in Fortran to calculate the optimum allocation of sample points over the four strata for STSI sampling and for STSI, SI sampling. The cost models produce discrete values, which can result in a large number of possible states per stage. To restrict the number of calculations and to avoid many irrelevant calculations in the case of a continuous state space, the state space should be subdivided into equal portions. Values of the value function should only be calculated for a limited number of points at a fixed distance. If a state between two known points is reached, an interpolation technique should be used to calculate a value for this intermediate point (Hadley, 1964). The Fortran program used linear interpolation which caused an interpolation error since the value function was non-linear. However, if the distance between two adjacent points is small (e.g. near the mean cost of an sample point) this error will be limited.*

The results of optimization of two schemes with STSI sampling and STSI, SI sampling, respectively, are presented in Table 7.3. The total budget in both situations was Dfl. 105,000 and the state space was subdivided into 50 equal portions. The allocation of sample points over strata for the STSI, SI design was also calculated for 10 and 15 sample points per selected primary unit. As expected, this resulted in larger predictions of the sampling error: 13.87×10^{-4} and 17.53×10^{-4}, respectively. The time needed to obtain results for the two designs in Table 7.3 was a few minutes. If less parts are distinguished in the state space, less states have to be calculated per stage and the time for calculation will decrease. For STSI, SI sampling this time also decreases with an increase in the number of sample points per primary unit.

The solutions for both types of designs show that less sample points are allocated to the first stratum and that the numbers of sample points in strata 2, 3 and 4 are near to each other.

This distribution of observation points is related to the variance components for the four strata presented in the description of the case in Sub-section 7.2.4. The influence of the cost parameters presented in Sub-section 7.3.4 can also be noted: stratum 3 is the most expensive stratum, and stratum 4 the cheapest. Both solutions show that there is some budget left over, as the total costs are less than the available budget. However, the amount of budget left is insufficient to add an additional sample point or primary unit with five sample points.

Table 7.3 Results of optimizing STSI sampling and STSI, SI sampling using DP: available budget is Dfl. 105,000 and the state space is subdivided into 50 equal portions

***STSI* sampling**					
Stratum	1	2	3	4	total
n_h	62	85	87	91	325
$\hat{r}_h$ x 10^{-4}	1.287	1.667	1.714	1.707	6.375
c_h (Dfl.)	21 086	27 254	29 413	27 161	104 914
***STSI, SI* sampling with five sample points per selected primary unit**					
Stratum	1	2	3	4	total
n_{1h}	11	18	19	18	66
n_h	55	90	95	90	330
$\hat{r}_h$ x 10^{-4}	1.933	2.702	2.733	3.037	10.405
c_h (Dfl.)	18 082	28 345	31 486	26 722	104 635

The main objective of the calculations presented above, is to show that an optimal allocation of sample points within outlinear schemes can be obtained by DP. This is a vital part in the KBS aimed at. The algorithm and the way in which it is implemented may be improved in the future.

7.5 Discussion of the procedure for evaluation and optimization

In this chapter a procedure to search for efficient survey schemes was proposed. The procedure proposed consists of models for predicting the sampling error, models for predicting the cost, and optimization models. The models presented are discussed below.

The prediction of the sampling error depends largely on the available prior information on spatial variability. At present the availability of this information may be a bottleneck. If no stored information is available, the user may be asked to guess values based on field experience. It is obvious that this may introduce some uncertainty in the evaluation results. The system should, however, allow the evaluation to be repeated for other guessed values and thus give the user insight into the sensitivity of the procedure for parameter values. At present, the amount of information on spatial variability increases steadily and will increase due to the fact that the proposed system will store information of previous surveys.

An advantage of the general simulation algorithm for sampling error prediction (Sub-section 7.2.3) is that it can be used with any sampling design. A disadvantage is that it is rather time-consuming. This particularly becomes a problem if it is used in combination with DP for

optimization. The algorithm requires efficient subroutines for sample selection to be available. However, such routines will be needed anyway for the support system to select the actual sample. If evaluation of the accuracy is to be used in combination with optimization, specific algorithms for types of designs seem more appropriate since they are less time-consuming. Then the variance formula for each type of sampling design needs to be specified. The variance components in these formulae can be determined using simulation, after which the model for sampling-error prediction can be used in the optimization procedure. An appropriate random number generator should be selected to ensure reliable simulation results.

Prediction of the sampling error may require considerable memory and computing capacity if grid files are used depending on the survey region and the grid size. These technical details fall beyond the scope of this thesis but need attention at a later stage. For a complex sampling design it may be difficult to specify the variance function. Then, the general algorithm for sampling error prediction can be applied to evaluate the accuracy of a particular design, if subroutines for sample selection are available. For such situations there is, however, no optimization procedure.

Cost models are defined for the main classes of sampling designs. The estimation of values for the access-time parameters in the cost models may be difficult in the beginning. However, by continuously storing and adjusting these parameters in the future, this problem will decrease in the course of time. As mentioned above in relation to the prediction of the sampling error, the sensitivity of the system for different cost parameters can also be examined. Since the aim is to use these models to compare possible survey schemes, the emphasis is on the influence of classes of designs. The usefulness of the cost models has to be tested in practice, after which they can be adapted and refined where necessary.

According to the description of the domain in Section 2.2 the case described in this chapter was a single criterion problem. In practice, there often are several variables of interest. Since the total cost of spatial inventory should be taken into account when designing a survey scheme, the parameter values of the cost components should be determined for the combination of these target variables. However, when optimizing a survey scheme, the evaluation of the accuracy should always be related to the most important target variable.

DP seems a suitable technique for mathematically optimizing soil survey schemes and from the preliminary experiments it may be concluded that it is possible to optimize outlinear survey schemes within a reasonable time (a few minutes). The simulation algorithm for estimating the variance components should be implemented as a module preceding this optimization. Advantages of the proposed procedure are that it enables objective comparison of possible schemes and that differences between sub-regions (strata), e.g. concerning spatial variability or access time, can be taken into account. A disadvantage is that there is no standard mathematical formulation for DP and that therefore specific value functions are required for different classes of designs, and for different problem formulations (first versus second problem). However, the DP formulations presented in Sub-section 7.4.4 can easily be adapted for frequently used types of stratified designs (e.g. *STSIC* sampling with or without replacement, and with or without equal probabilities, *STSI, SIC* sampling with or without replacement, and with or without equal probabilities).

Chapter 8

Basic design considerations

8 Basic design considerations

8.1 Background

In Chapter 1, the aim of this study has been stated as identifying basic design considerations for a system to assist in the design of soil survey schemes. These design considerations are based on knowledge of the structure of the domain and the main tasks (Chapter 4 and 5), on knowledge about methods of determination and statistics (Chapter 6), and on knowledge about evaluation and optimization of survey schemes (Chapter 7). This knowledge is structured and generated in the previous chapters. The design considerations are the subject of this chapter.

In Section 4.1 it was noted that the choice of a tool for implementation at an early stage might limit the development of an appropriate KBS. When developing a prototype, a preliminary choice should be made, but such a choice should be based on well-thought out design considerations. The basic design considerations presented here were developed independently of a choice for an implementation technique.

This chapter does not pretend to present a conceptual design sufficiently detailed for implementing a KBS to be used in practice. The design considerations proposed may, up to a point, be incomplete and debatable. However, this study resulted in the structure of the KBS and in a description of the main components. Moreover, the design considerations will serve as a starting point for the development of a prototype.

Section 8.2 discusses the main requirements of the system as proposed in Chapter 1. Thereafter, Section 8.3 describes the intended use of the system. Then, Section 8.4 focuses on the components for an actual knowledge-based system. These components are evaluated in Section 8.5.

8.2 Requirements

In conventional software engineering, producing a software specification starts with defining the requirements. Two categories of requirements are distinguished: functional and non-functional requirements (Sommerville, 1992). Both should be testable. The functional requirements are related to the system services expected by the user. Their subdivision depends on the domain under consideration. The non-functional requirements point out the constraints under which the system must operate and the standards to be met by the final system. The following three classes of non-functional requirements can be distinguished (Sommerville, 1992):

- the product requirements, which are related to users needs, e.g. a requirement for the maximum response time for user commands, or requirements on the usability;
- the process requirements, imposed on the system development process, e.g. implementation requirements;
- the external requirements, which cover all remaining requirements, e.g. requirements for

interaction with other systems, or organizational requirements.

Sommerville (1992) notes that initial versions of software requirements are often incomplete and inconsistent. Therefore, the requirements often need to be corrected and completed during reviews or at later stages.

The definition of requirements for a KBS is analogous to that of conventional software engineering. In addition to the functional and non-functional requirements, KBS designers distinguish explicitly the system specifics, which are related to the structure of the system, i.e. the logical and functional decomposition of the system. For the KADS methodology (see Sub-section 3.2.3), Hesketh and Barrett (1989) introduce the so-called requirements model consisting of:

- the system specifics or the structure of the system;
- the functional requirements;
- the human/computer interface requirements;
- the hardware and software requirements;
- the external requirements.

The latter three can be considered as non-functional requirements. Below they are briefly described in separate sub-sections. Before doing so, insight is given into the structure of a KBS (Sub-section 8.2.1), followed by a description of the functional requirements (Sub-section 8.2.2). This section deals with an initial requirements definition for the KBS; a complete and testable requirement specification should be developed at a later stage.

8.2.1 Structure of a KBS

Section 3.7 characterized the system aimed at as a KBS to assist in the design of soil survey schemes. It was suggested that the intended system should be the result of combining knowledge from various disciplines. In Figure 3.3, a rough structure of the system was presented in which five components were distinguished: a database, a knowledge base, a model base, an inference engine, and a user interface. The results of knowledge acquisition, knowledge generation, and knowledge structuring (described in Chapters 4, 5, 6 and 7) suggested a somewhat different structure of the system, that is a structure in which the strict separation of the knowledge base and the inference engine has been taken as subject to discussion. Here, the interaction problem (Bylander & Chandrasekaran, 1987) is encountered. This sub-section discusses some theoretical aspects and describes the proposed structure.

Theoretical aspects

A main characteristic of a KBS is the separation of domain knowledge from procedures manipulating this knowledge. The procedures for manipulating knowledge constitute an inference engine. An advantage of this separation is that the domain knowledge can be extended and adapted if necessary without a need for changing the inference engine. Moreover, the separation also facilitates systematic knowledge acquisition (Steels, 1990), e.g. through step-wise development and refinement of the knowledge base. Furthermore in principle, a new ES can be constructed by replacing the domain knowledge by new knowledge

corresponding to a related problem domain (Rich & Knight, 1991). The inference engines and their environments which can be used in various domains are called *shells*; a shell also implies an implementation formalism.

In practice, a separation of a knowledge base and an inference engine may be hard to achieve (e.g. Brownston et al., 1985) or may be artificial (Bylander & Chandrasekaran, 1987). According to Bylander and Chandrasekaran (1987: p. 232) this is caused by the *interaction problem*:

"Representing knowledge for the purpose of solving some problem is strongly affected by the nature of the problem and by the inference strategy to be applied."

In other words, the representation of knowledge is related to its use. Somewhat earlier Waterman (1986) had already recognized this problem by stating that the inference engine cannot be characterized in a simple, general way and that the structure of the inference engine is affected by the problem domain and the way in which the knowledge can be represented and organized.

Bylander and Chandrasekaran (1987) propose the identification of *generic tasks* to deal with the interaction problem. These generic tasks can be characterized by information about the type of problem, the representation of knowledge, and the strategy of the inference engine. In a generic task they combine a goal (e.g. 'diagnosis') with the method used to achieve this goal (e.g. 'data abstraction' and 'classification'). Chandrasekaran et al. (1992) describe a task-structure analysis in which tasks (i.e. problem-solving goals), methods and sub-tasks are distinguished separately.

The KADS methodology is based on the assumption that different categories of knowledge can be modelled and represented independently and independent of their use. Wielinga et al. (1992) integrated KADS with other lines of research with the aim *"to span the whole continuum from weak to strong knowledge interaction"* resulting in CommonKADS.

The interaction problem mainly emerged from the idea to re-use inference engines or to re-use approaches to problem solving. The re-use of other components of a KBS also seems attractive, e.g. the knowledge base. Much effort has been spent on building elaborate knowledge bases for specific problems. Gruber (1991) focuses on the ability to share and re-use knowledge bases. Therefore, he proposes a common representation and programming environment for which he lists the main characteristics. He stresses that KBSs will always require application-specific knowledge, but expects that parts of knowledge bases may be re-usable in new KBSs. This point of view seems contradictory to the interaction problem (Bylander & Chandrasekaran, 1987), since knowledge is dependent on its use and therefore related to a type of problem.

The methodologies for developing KBSs described in the literature (Gruber, 1991; Chandrasekaran et al., 1992; Wielinga et al., 1992) aim at facilitating knowledge acquisition and knowledge structuring, and at the re-use of parts of a KBS. Development of common accepted and versatile approaches can facilitate the development of ESs and KBSs. However, to achieve these objectives, the descriptions of the methodologies often remain at a high level of abstraction. Fortunately, there is growing agreement on the main characteristics of the methodologies. At the beginning, it was difficult to see the common element of the generic task methodology (Bylander & Chandrasekaran, 1987) and KADS (Breuker & Wielinga, 1989).

The current task-structure analysis (Chandrasekaran et al., 1992) and the CommonKADS framework (Wielinga et al. 1992) are more detailed and are clearly converging. Therefore, these methodologies are easier to comprehend. Nevertheless, much work needs to be done before the methodology of task-structure analysis and the CommonKADS framework are available for various, specific domains.

To what extent domain knowledge can be separated from the inference engine depends on the type of problem and on the domain. The desirability of such a separation may also be a point of discussion. The relation between domain knowledge and inference engine gives rise to three main types of systems:

- systems in which the two components are completely separated, e.g. resulting in shells for ESs;
- systems in which the two components are interwoven, such as in procedural programs;
- systems in which the structure of the knowledge base and the inference engine are functionally related, depending on the domain.

Below, the last type of systems is explicitly discussed, since this type is felt to be most appropriate for the domain of the design of soil survey schemes.

Proposed structure

The proposed structure for the KBS aimed at consists of the following components: a database, a knowledge base, a model base, a problem-solving model, and a user interface. For the domain of interest it seems difficult and inefficient to separate all knowledge from the inference engine. Again, it is stressed that this project only aims at developing a KBS to assist in the design of soil survey schemes, and not at, for example, providing a tool for the development of various knowledge-based systems. For the domain of interest two types of knowledge are distinguished: fixed knowledge and 'variable' knowledge. The fixed knowledge can be included in a problem-solving model closely related to the inference engine, since the reasoning mechanism makes extensive use of the established facts in this field of research. The 'variable' knowledge, so called since it will be updated and extended, is stored in the database, the knowledge base, and the model base, depending on its specific nature (see the corresponding sub-sections of Section 8.4). The final component in the proposed structure of the KBS is the user interface.

To find a borderline between fixed knowledge and 'variable' knowledge the process of designing a soil survey scheme is investigated. Although the design process is an iterative process (Sections 4.4 and 5.4), the structure of this process is assumed to be fixed (see Section 4.4). This implies that the order in which questions for finding a solution can be posed is also fixed, since the answer to a question immediately indicates the next question to be asked (e.g. Figures 6.1 and 6.5). Moreover, after an answer has been supplied, the system applies its knowledge to confine the solution space (see Sub-section 6.3.3). This fixed knowledge will be included in the problem-solving model.

Usually, to obtain answers to the questions, interaction with the user is required. For example, whether a survey region can be meaningfully partitioned into strata or whether meaningful clusters can be defined depends on case-specific conditions, such as the target variable and the characteristics of the survey region of interest. These conditions may vary

considerably among survey projects. Therefore the knowledge specific for any soil survey project has to be provided by the user. The system will ask questions for which knowledge on the applicability of types of sampling designs is required. Obviously, the answers provided by the user may lead to different search processes in different regions of the search space. Under these circumstances, it seems rather cumbersome to develop a general inference engine or to use a separate general search algorithm.

The problem-solving model serves as an interface between the user and the various pieces of knowledge stored in the database, the knowledge base, and the model base. The 'variable' knowledge is, for example, related to instantiations of concepts for actual soil survey projects, or to (mathematical) models for evaluation and optimization. The problem-solving model will assist the user in designing a soil survey scheme through interaction and selection.

The main characteristics of the proposed KBS are summarized below:

- knowledge of the fixed structure of the design process is included in a problem-solving model;
- the knowledge in the system will comprise data, rules, and (mathematical) models (stored in the database, the knowledge base, and the model base, respectively) to be updated without changes in the problem-solving model;
- a large amount of knowledge is stored from which only a small part is used to assist in the design of a scheme for a specific soil survey project;
- interaction between the system and a user results in the selection of appropriate knowledge for a specific problem.

In general, it should be noted that with the enhancement of tools and environments for developing computer programs, complex computer programs with well-developed user interfaces look like intelligent programs. Examples of such developments are X-Windows and Hypertext. Since these software products are easy to use and flexible, their behaviour looks intelligent to users, yet they are no more than interactive systems with adequate user interfaces.

8.2.2 Functional requirements

The functions required of the system are assumed to coincide with the tasks to be supported. The main tasks have been presented in Section 5.4. In Chapter 6 and Chapter 7 refinements of tasks in the design process are described. Below an updated list of tasks and sub-tasks is provided.

- *Definition of the request for soil survey.*
- *Selection of prior information.*
- *Design of outlinear schemes*:
  - *selection of appropriate methods of determination*;
  - *selection of an appropriate sampling approach*;

  in the case of a design-based approach:
  - *selection of appropriate types of sampling designs.*
- *Evaluation and optimization of outlinear schemes.*
- *Report generation.*

These tasks and sub-tasks are executed during the design of a soil survey scheme. A task which completes a soil survey project is:

- *Evaluation a posteriori.*

Section 8.4.4 describes how these tasks will be executed in interaction with the user.

8.2.3 Human/computer interface requirements

So far, no specific attention has been paid to the user interface of the system proposed, since a great deal of work had to be done on the interior of the system. However, in places some remarks on the user interface have been made, and below the main requirements for the exterior of the system are listed:

- easy to use: the screens should be well-organized, and the system should be usable without the need to refer frequently to a manual;
- flexible: possess an option to return to earlier stages to allow the user to change information supplied earlier or to add information, after which the system should reconsider its conclusions;
- interactive: ask information from the user, or allow the user to rank priorities or to take some decisions; for example, during the selection of types of sampling designs the system will present possible decisions to the user and provide information on their consequences; the user will be allowed to ask the system for explanation where required;
- explanation facility: be able to explain how conclusions are reached, or on which arguments decisions are based, and provide the user with information on possible options, or explanation of the meaning of questions;
- reliable: perform as expected by the users; it is hard to give a useful formal definition of reliability (Sommerville, 1992), but an attempt should always be made to minimize the number of inconsistent answers, software failures or other erroneous results.

8.2.4 Hardware and software requirements

Again, it is noted that this study has not decided on a tool for implementing the system. The following requirements need to be considered when choosing a shell or a programming language:

- the system should be developed on a personal computer to increase the number of potential users;
- the system should be able to use information from GIS databases as sampling frames (see Chapter 7).

Development of software for statistical data analysis for the types of designs supported by the system is desirable, but is not an essential part of the system.

8.2.5 External requirements

At this stage, the main external requirement for organization of the further development of the system is that:

- from the start of the development of a first prototype of the whole system, a group of future users should be involved in developing the system further.

8.3 The intended use

The development of the system started from the need for assistance in the design of schemes for soil surveys (Section 1.5). The system will be consulted in actual soil survey projects, and will be used to assist in negotiating the economical and technical conditions of actual surveys. Besides this use in survey projects, the system may also fulfil a role at a more strategic level. It may, for example, be used as a tool to get insight into the effects of constraints on the accuracy of specific categories of surveys, e.g. surveys on phosphate saturation, and so help to assess the feasibility of regulations on soil surveying.

The system aimed at here will be meaningful to different parties involved in soil survey projects. On the one hand the people who execute the surveys will profit from assistance in the design of soil survey schemes. On the other hand those who commission surveys will profit from the possibility to get better insight into the consequences of various constraints and of possible schemes for achieving the efficiency predicted.

Two categories of users executing surveys are distinguished: researchers at institutes and universities, and environmental consultants. The former are interested in surveys on various scales, e.g. a survey to determine the phosphate-saturated area in a province, or a survey to determine the mean humus content of the top-soil of a parcel. In the Netherlands; the latter are mostly investigating local areas, e.g. industrial sites, or waste disposal sites. The users of the system should at least have a basic notion of the use of sampling strategies in soil surveys. Otherwise they will not be able to communicate with the system and appreciate the advice and the report presented.

Examples of people who commission surveys are local, regional and national authorities, and private land-owners. Although these people may profit from the system, they are not distinguished as intended users because they generally have little statistical knowledge and are rarely involved in the whole process of designing a soil survey scheme. Although the system will possess an explanation facility, it will not be suitable for statistically naive users. Such users not only require a great deal of explanation of basic statistical concepts, but will probably not be able to interpret the meaning of these concepts in the soil survey context.

8.4 Components for an actual KBS

In Sub-section 8.2.1, which dealt with the structure of a KBS, the following components were distinguished for the KBS of interest: a database, a knowledge base, a model base, a problem-solving model, and a user interface. This section goes into the contents of these components for an actual KBS.

8.4.1 Database

Information related to soil survey projects will be stored in the database. This information will serve initially as prior information. The information which is used and the information which is stored during the design of a particular survey scheme has to be represented finally in the report of a survey scheme. This report will contain information on all concepts introduced in the conceptual framework (Fig. 4.4 and Fig. 4.5), except information on the selection technique, i.e. the operational method by which sampling elements are selected according to the sampling design (see Sub-section 4.5.2). A technique for sample selection will be a fixed part of the system, stored in the model base and clarified in the system documents, so that it does not need to be pointed out in every survey report. Information on the other concepts of the conceptual framework will be stored in the database. The structure of the database can be represented as a data model. Figure 8.1 shows a conceptual data model of the information on soil survey projects.

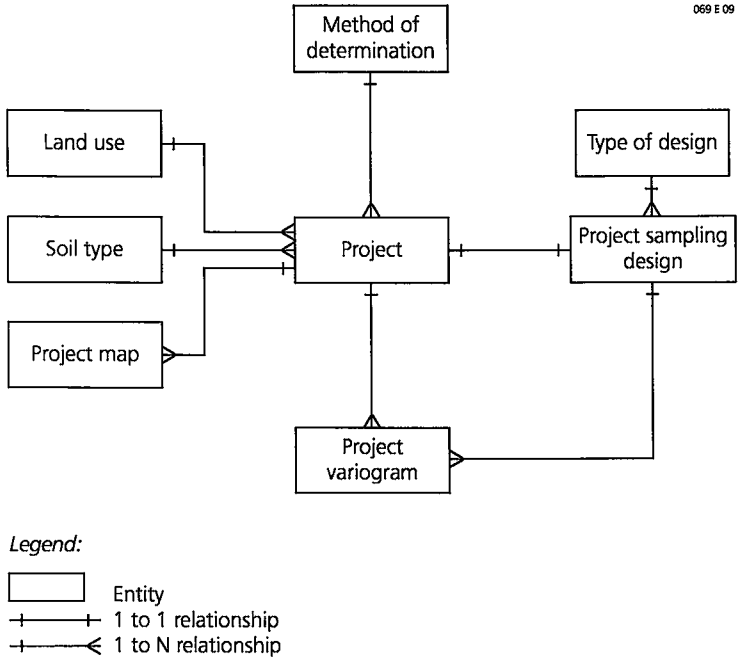


Figure 8.1 Conceptual data model of information on soil survey projects

For each entity in this figure a table will be created in the database. Figure 8.1 shows two types of relationships between these tables: '1 to 1' and '1 to N' relationships. The first type of relationship means that one row or *record* of the first table is connected with one record in the second table, e.g. for each project there is one project sampling design. The second type of relationship refers to the possibility that one record in the first table can be related to several records in the second table, e.g. a project uses one method of determination, but this method may be used in several projects.

Most tables in the conceptual database have a fixed format that suits every survey project.

However, information on the 'project sampling design' is specific for a project, there is no general structure of attributes that is suitable for every project, e.g. the number and the presence of strata, primary units, and clusters will vary among projects. Information on the sampling designs can be stored in a table with variable-length records (e.g. Hansen & Hansen, 1992), which allows the records with information on sampling designs to vary between projects. Another option is to create a maximum table in which for each project only the relevant attributes get values.

Below the content of the database is further discussed, thereafter attention is paid to the collection and storage of information.

Content

For each entity in Figure 8.1, i.e. a table in the database, a number of attributes are distinguished on which information has to be stored, i.e. the content of the table. This determines the content of the database. The entities with the main attributes on which information should be stored are listed in Table 8.1. The contents of the tables are briefly characterized below.

Table 8.1 Entities in the conceptual data model with the main attributes

Entity	Attributes
Project	project_code, name, date, target_quantity, target_variable, land_use_code, soil_type_code, accuracy_constraint, cost_constraint, capacity_constraint, method_of_determination_code, text_file
Method of determination	method_of_determination_code, reference, target_variable, sampling_element, c_e, c_v, $\bar{t}_{on}$, standard_error
Type of design	type_of_sampling_design_code, description
Land use	land_use_code, description
Soil type	soil_type_code, description
Project map	file_name, project_code, function
Project variogram	project_code, type_of_variogram, source, region / stratum, value_file
Project sampling design	project_code, type_of_sampling_design_code, {stratum, GIS_code_ST, primary units, GIS_code-P, number_of_primary_units_selected, number_of_clusters, number_of_elements}, template_name

General information on a project, and information on the aim and constraints will be stored in the table 'project'. If the user wants to add remarks on the decisions during survey design, or if comments have to be stored when the project is evaluated after execution of field work, this information will be stored in a text file to which the database refers in the table 'project'. Since this information is of a descriptive nature, it is difficult to store it in the database. For example, remarks on partitioning criteria used to divide the survey region and the specification of instructions for field work may be stored in a text file. Since the instructions for field work need to be specified for any survey project, the information required on this concept is briefly described below.

To ensure that all relevant instructions for field work are considered in advance, the system will present a checklist of types of instructions to be specified (Fig. 8.2). The user will have to specify these instructions, and this specification can be stored in the text file. The types of instructions distinguished are: instructions on coding, instructions on the use of reserve points, and instructions on the log-book.

Instructions for field work
- instructions on coding:
  - how should observation points be registered that appear to be located in non-soil?
  - how should inaccessible observation points be registered?
  - how should values be registered that are outside the range of measurement?
- instructions on the use of reserve points:
  - how many reserve points need to be selected?
  - under which conditions (e.g. non-soil, inaccessibility) should these points be used?
- instructions re the log-book:
  - note changes in the original scheme, with justifications;
  - report time spent on field work per day, number of observation points visited per day,...

Figure 8.2 Checklist for the instructions for field work

A description of a method of determination and information on parameter values related to these methods is stored in the table 'method of determination'.

The table 'type of design' contains brief characterizations of the main types of sampling designs. Knowledge related to the applicability of these types of designs is stored in the knowledge base, and the related formulae are stored in the model base.

There are separate tables in which types of 'land use' and 'soil types' are described. These characteristics of the survey region may influence the survey design, e.g. the (type of) variogram is generally related to the features of a particular region. They may also be useful to retrieve appropriate information from historical surveys. An example of a type of land use which may be used as a key for retrieving information is 'former gasworks'; sites with this type of (former) land use are often polluted. When a new site is to be surveyed it may be useful to consider the survey schemes and the survey results of past investigations on similar sites.

The names of the files from a GIS database which contain maps used in a survey project will be stored in the table 'project map'. The functions of these maps, i.e. prior information or sampling frame, will be stored with the file names.

In the table 'project variogram' the types of variograms used in a project will be stored to improve the accessibility of information on spatial variability. This table has to refer to the source of the variogram and has to specify the region to which the variogram is related. Furthermore, it will refer to a special file ('value_file'), e.g. a template, in which the parameter values of the variogram are stored (see below).

Finally, the sampling design which results from the design process is stored in the table 'project sampling design'. In this table the partitioning of the survey region and the number of elements or sub-sets of elements to be selected are stored. Furthermore, it refers to a

special file, e.g. a template, in which the parameter values related to the evaluation of survey results are stored. For each survey project, the parameters of which values have to be stored depend on the method of determination selected and the project sampling design. It seems impossible to create a table in which parameter values for all possible survey schemes can be stored, because the number of strata can vary, and because there is significant variation in the parameter values related to types of designs. Furthermore, it seems impractical, since for each survey only some parameters get values, and all other parameters should be given zeros. These difficulties can be circumvented by creating a template for each survey scheme, which is a much more flexible way of storage. Such a template will contain two rows of numbers containing values for the parameters in the cost model, in the accuracy model, and values resulting from the optimization. The first row can contain the values used during survey design, and the second row the values computed after execution of field work, i.e. the results of evaluation *a posteriori*. The order in which these parameter values are stored must be recorded.

Since the number and types of variograms also vary among survey projects and since the number and types of parameter values vary among variograms, parameter values of the project variograms used will also be stored in templates.

Collection and storage of information

In Section 4.3 and 5.4 the need for collection and storage of knowledge and experience from historical surveys was mentioned as a means of improving the use of this knowledge and experience. It was also stated that a simple form of a self-learning mechanism might help to fulfil this need. The system can contribute to its own maintenance if it possesses such a mechanism.

In a first version of the KBS a mechanism will be available to store knowledge of surveys supported by the system, i.e. information on the decisions made and related attribute values, in the database. When information is stored, situations should be prevented where the same information is stored several times, which would result in an inconveniently organized database in the course of time. The system will have to examine new information and check whether similar information is present. If so, the user should be asked to decide whether the present information should be adapted. Furthermore, the system needs to verify parameter values, and present extreme values to the user. This storage of information obtained during problem-solving is related to a the technique of machine learning called rote learning (Rich and Knight, 1991). It will assure that the amount of knowledge on survey projects increases. Although this mechanism may be relatively simple compared to other learning mechanisms, it seems realistic to assume that this type of automatic maintenance can be achieved.

8.4.2 Knowledge base

Two categories of knowledge will be stored in the knowledge base: knowledge about statistics and knowledge related to explanation.

Statistical knowledge

Section 6.3 has discussed the statistical knowledge needed for designing soil survey schemes. Decision trees were presented to guide the selection of applicable statistical approaches and classes of sampling designs (Fig. 6.1 and Fig. 6.5). Besides, some rules were formulated

to assist in the selection of appropriate types of designs (Sub-section 6.3.3). The knowledge base also contains knowledge of the selection of models from the model base related to classes or types of sampling designs. In addition to this knowledge, knowledge concerning the applicability of statistical strategies and of types of variograms in soil survey practice will be stored, e.g. whether a particular type of design is appropriate for surveys on a particular soil property in a particular type of survey region. This knowledge still needs to be collected and structured.

When a structured approach to the use of statistical strategies for soil surveying is regularly used, knowledge about the applicability of various strategies in soil survey practice may increase. However, it may be difficult to generalize from this practical experience. Therefore, it seems advisable to record the whole consultation process of projects in log-files. These log-files of consultations and the corresponding results of the survey projects must be analysed afterwards to see whether there is good reason to update the knowledge base. When the knowledge base is updated, attention to correctness and consistency are of the utmost importance.

Explanation

The knowledge base will also contain knowledge required for explanation. The system needs to be able to explain how a solution or sub-solution has been reached. To provide a satisfactory explanation the reasoning process should proceed in understandable steps and sufficient meta-knowledge about this process should be available (Rich & Knight, 1991). This meta-knowledge will be stored in the knowledge base and during consultations of the system a trace of the program's execution will be stored. This knowledge can be used to explain why a question is asked to the user or how the system derived a conclusion or selection. Rich and Knight (1991) state that it is hard to support all possible explanations a user might want. A minimum approach will be available in a first version of the system using meta-knowledge about the reasoning process and the trace of the reasoning. The explanation facility may be extended at later stages. For example, a more innovative way of explanation is the ability to explain the rationale behind the rules used (Waterman, 1986). This can contribute to a greater confidence of the users in the system. Therefore, more knowledge has to be stored in the system. Indeed, considerable effort is required to develop a more elaborate explanation facility, but such a facility may facilitate knowledge maintenance because the knowledge is more explicit.

When listing features for expert systems in statistics, Hand (1985) mentioned the possibility to explain itself and the ability to explain technical terms, as two separate features. Here, the concepts defined in the conceptual framework (Section 4.5) can be considered as technical terms. The user should be aware of how terms are used in the system. Therefore, the system has to be able to explain these concepts, which are vital in the design process. In addition, there may be other (statistical) terms which the system should also be able to explain. The analysis of the log-files of consultations mentioned above can give insight into the requirements for explanation.

8.4.3 Model base

The model base will contain (mathematical) models required during the design of soil survey schemes. These models can be used for various survey projects. The models required for a particular outlinear scheme are related. The structure of the model base is based on these relations.

Models that can be coupled together to form larger models can be arranged in a hierarchical structure (Zeigler, 1990). Figure 8.3 exhibits the models needed during the design of survey schemes in a so-called *composition tree* (Zeigler, 1990). This figure is clarified below, starting with the branch on the right hand side.

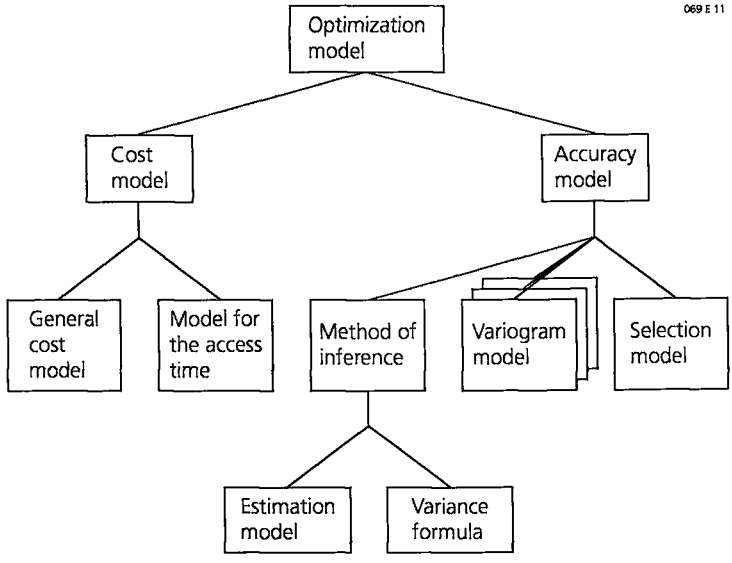


Figure 8.3 A composition tree of models needed for the design of soil survey schemes

For a particular type of sampling design and method of determination, i.e. an outlinear scheme, each model in Figure 8.3 needs to be specified. Initially, the estimation model and the corresponding variance formula, which are determined by the target quantity of interest, result in the method of inference. Besides, the required variogram models and the selection model need to be available. When the parameter values of the variograms are known, the variance components in the accuracy model, i.e. the model for prediction of the sampling error, can be calculated. The accuracy model is a *coupled model* of the method of inference (model), the variogram models, and the selection model. If there are several sub-regions distinguished in the design, which is very common in soil surveying, a variogram model is required for each sub-region. Therefore, Figure 8.3 depicts a number of variogram models. The selection model should make it possible to select random points for the prediction of the variance components (see Sub-sections 7.2.3 and 7.2.4). Furthermore, a selection model will be required to select a sample according to a sampling design, i.e. the technique for sample selection in the conceptual framework (Sub-section 4.5.2). The selection model may also be a coupled model itself, but since specification of this model is beyond the scope of

this thesis, it is represented as a leaf in Figure 8.3. On the left hand side of the tree, the cost model is a represented as a coupled model of the model for the access time (for a given class of sampling designs) and the general cost model (Sub-section 7.3.4). Finally, the root of the tree shows the optimization model which includes cost model and accuracy model.

The following information has to be stored on every model: the name of the model, the formula or algorithm, the input, and the output. Figure 8.4 shows an example of information to be stored on a cost model and how this can be presented to the user.

Model name:	Cost model *STSI* sampling		
formula / algorithm	$c_h = c_s \cdot \{t_{a0h} \cdot \sqrt{(A_h \cdot n_h)} + \bar{t}_{oh} \cdot n_h\} + c_e \cdot n_h$		
input:			
parameter 1	survey cost per hour	:	$c_s = 87.5$
parameter 2	cost of laboratory analysis per sample	:	$c_e = 170$
parameter 3	access time in stratum 1	:	$t_{a01} = 3$
parameter 4	access time in stratum 2	:	$t_{a02} = 3.5$
parameter 5	area of stratum h	:	$A_h = 16$ ($h = 1, 2$)
variable	number of observation points in stratum h :		n_h
output:			
variable	cost of spatial inventory in stratum h	:	c_h

Figure 8.4 Presentation of information from the model base

Zeigler (1986) introduced a framework for knowledge representation in simulation, in which the knowledge base encompasses a system entity structure and a model base. Declarative knowledge needed to select and construct appropriate models is stored in the system entity structure. The system aimed at the knowledge base contains knowledge about the selection of appropriate models as far as this is related to classes or types of sampling designs. Information on the required variogram models will be stored in the database during the design process (see Sub-section 8.4.1).

If required, models may be updated and new models can be added. Experience in practice may, for example, make clear that it is desirable to slightly adjust the cost models introduced in Section 7.3. It may also be desirable to add other optimization models, or variogram models. However, the relations between models always need to be taken into account when changing the contents of the model base.

8.4.4 Problem-solving model

The fixed knowledge of the structure of the design process, of the order to pose questions, and of the way to confine the solution space is integrated in the problem-solving model. Figure 8.5 depicts a flow diagram of this model for the design of soil-survey schemes. The tasks during the design process listed in Sub-section 8.2.2 can be recognized in this model. The task of 'evaluation *a posteriori*' is not included. This task does not need to be performed during the design of a survey scheme, but after the execution of the plan of action as prescribed in the survey scheme. However, this task also has to be incorporated in the problem-solving

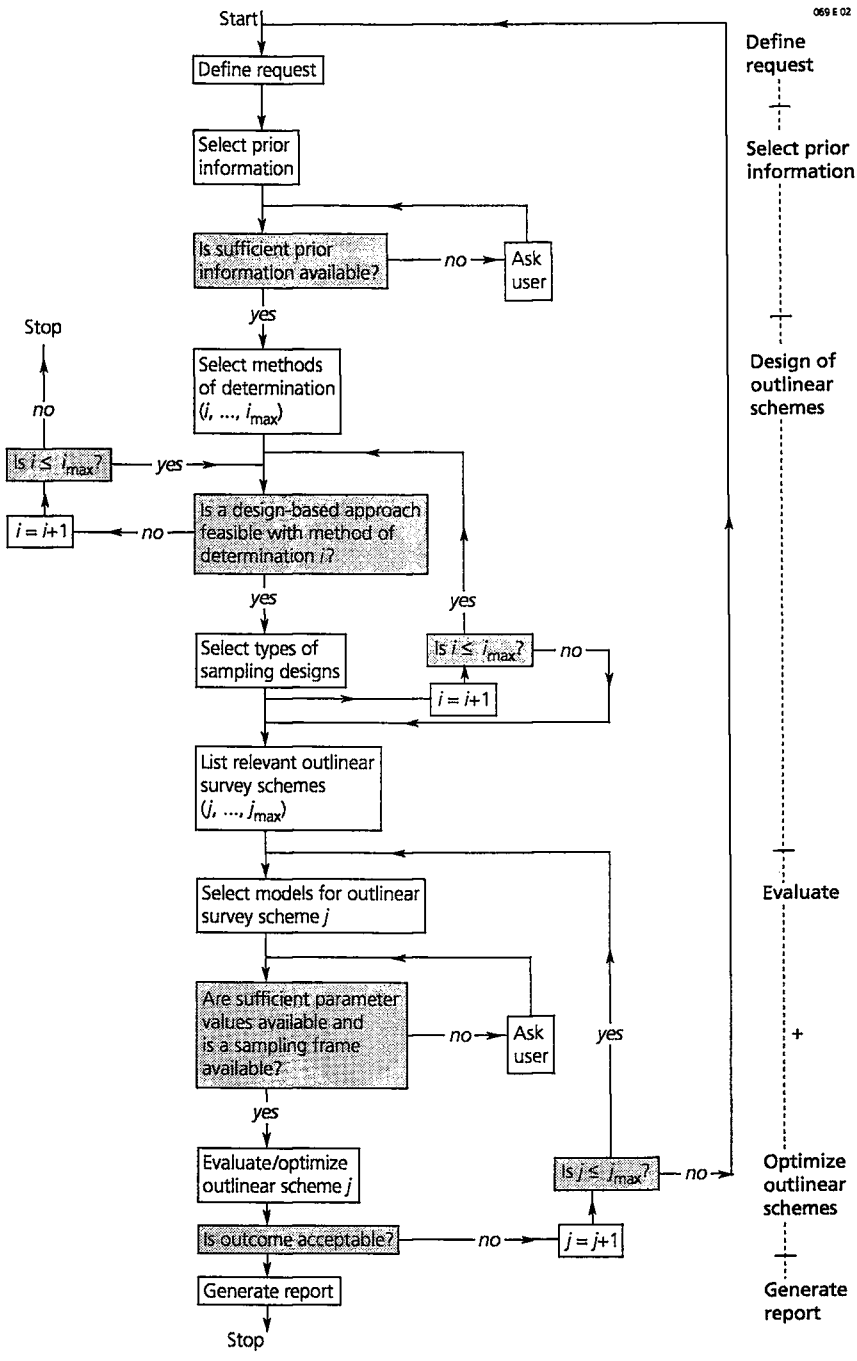


Figure 8.5 Flow diagram of the problem-solving model for the design of soil survey schemes

model, so that the system will be able to assist in the collection and storage of knowledge. The way in which the system operates through the problem-solving model, i.e. the way in which the problem-solving model guides the interaction between the user and the system knowledge, is described below. This description is based on the tasks distinguished.

Define request

The design process starts with defining the request. The system will assist the user in specifying the aim and the constraints of a survey project, e.g. by presenting possible options for the specification of concepts like, for example, a list of possible target quantities. During the definition phase, the type of request which is related to the applicability of statistical approaches has to be determined (see Chapter 1). The system will therefore present information on the types of request distinguished in the domain: *how much*, *where*, and *how much & where*?

At this stage, it will be checked whether it is permissible to continue, i.e. whether the request is part of the limited domain for which the system can provide assistance. The system needs to be aware of the limits of its domain, and has to be able to make this clear to the user, to prevent erroneous use. This issue needs certainly to be given thorough attention. The lack of insight into the limitations of their knowledge often causes ESs to fail (e.g. Bell, 1985), e.g. they interpret information incorrectly to make it fit in their knowledge domain. A similar risk may be present when interacting with a KBS: if the system and the user are unaware of the limitations of the domain, a wrong problem may be solved or a problem may be solved incorrectly. An example of incorrect use of the system that has to be prevented is its use for designing a scheme for surveys that aim at detecting whether a target variable is present or not, e.g. surveys to detect the presence of contaminated spots. For such types of surveys purposive sampling seems most appropriate, which is beyond the present domain of the system.

Select prior information

The system will search in the database for prior information from similar surveys in the past, e.g. prior information in the form of variograms, information on methods of determination, or maps of the survey region. The user will be asked to assess the suitability of the information and to provide additional information, if necessary. For example, since an initial version of the system will not be connected to a GIS, the user will have to provide information about appropriate maps. The user then acts as an interface between the system and a GIS database. The user must be aware of the role of the prior information in the design process. Retrieval of relevant information may take some time.

Design of outlinear schemes

As described in Sub-section 8.2.2 this task can be subdivided in a number of steps (see also Chapter 6). Initially, the user will be asked to provide information on one or more appropriate method of determination. For each method of determination a number of attributes need to be specified, e.g. c_q, and the standard error. When the information on methods of determination in the database accumulates, the system may select an appropriate method when selecting prior information. Then, the user will be asked to decide whether this method should be considered further and whether the stored attribute values are still appropriate.

If not, the information in the database has to be updated, e.g. the cost of laboratory analysis may have changed. For a given method of determination the user will be asked to roughly estimate the sample size allowed. This estimate of the sample size and the information on the spatial variability (in the form of variograms) are needed to select an appropriate sampling approach (Fig. 6.1). This selection requires interaction with the user. The system will ask questions and provide the user with possible options and with information on the main consequences of various options. The user has to take the decisions. During the selection of types of designs the user is allowed to explore several options. At this stage, it is not necessary to select just one type of design for each method of determination. When several methods of determination have been selected, the selection of an appropriate sampling approach may be repeated for each method of determination. If a design-based approach is feasible, the design process will continue and appropriate types of sampling designs will be selected (Sub-section 6.3.2) again through interaction with the user. If the survey request and the method (or methods) of determination selected do not allow a design-based approach to sampling, there will provisionally be no further support.

Combinations of selected methods of determination and types of sampling designs result in outlinear schemes. The outlinear schemes will be listed.

Evaluate and optimize outlinear schemes

The order in which outlinear schemes will be evaluated and optimized and the number of outlinear schemes that have to be explored will be decided by the user. When there is an outlinear scheme in which the observation points are finally selected using *SI* sampling, it seems advisable to explore this scheme first. If after optimization the efficiency predicted is not acceptable, this scheme can be used as a reference to select an outlinear scheme that might be expected to improve the outcome in the direction desired, i.e. to reduce cost or improve accuracy (see Sub-section 6.3.3). The system can provide information on the direction of the consequences on the efficiency of other schemes, e.g. decrease in cost or increase in accuracy, but not on the magnitude of such consequences. Therefore, the decisions on evaluating another outlinear scheme will be left to the user.

For each outlinear scheme that needs to be explored further, the corresponding models are selected from the model base. Thereafter, it is necessary to check whether sufficient parameter values are available, or if the user has to supply additional parameter values. Besides, an appropriate sampling frame needs to be available, which may already have been provided during the selection of prior information. Then, the optimization procedure will start. During this optimization various possible sampling designs are evaluated, i.e. the accuracy and the cost are repeatedly predicted. The result of this optimization is presented to the user, and the user is asked whether the outcome is acceptable or whether another outlinear scheme has to be explored. These steps are repeated until an acceptable outcome is reached. If it is impossible to find an acceptable scheme for the request specified, the system will return to the definition stage and allow the user to adapt, for example, the survey region, or the constraints.

Generate report

When an acceptable solution has been reached a report of the final scheme will be produced, including a justification for the decisions made. Most of the elements in this report will result

from previous tasks. Only the instructions for field work need to be specified at this stage with the help of the user. Furthermore, the system will select the actual sample. Sub-section 8.4.1 has described how information on the concepts in the report will be stored in the database. The explanation facility (see Sub-section 8.4.2) will help to supply information on the justification for the decisions from which the scheme described in the report results.

Evaluate a posteriori
Again, it is noted that evaluation *a posteriori* is not part of the design process. However, it has to be incorporated in the problem-solving model of an actual system. To fulfil this task information about the time spent on field work and on changes in the original scheme should be recorded in a log-book during the execution of the plan of action, as given by a survey scheme. Evaluation *a posteriori* aims at collecting and storing actual parameter values to improve the re-use of knowledge from historical survey projects. The parameter values estimated beforehand and those calculated after execution of a plan of action will be stored and compared. If significant discrepancies are established, which can be checked automatically, the user is alerted and requested to co-operate in finding possible causes. A structured approach is required to performing this task, which makes it possible for the system to contribute to its own maintenance. Evaluation *a posteriori* completes a soil survey project and assistance by the KBS needs to be further investigated.

It is obvious from the above description of the problem-solving model that the system operates interactively. The main reason for allowing so much input from the user is the importance of case-specific conditions during the design process, for which the user has to provide information. According to Hand (1985) in statistical domains there is often no 'right' answer, instead there may be a number of adequate answers. It seems impossible to develop a system that could operate in this domain without substantial input from the user. When a first version of the KBS is introduced, communication between system and user may be time-consuming since the amount of information in the system is limited and the user has to get used to providing information in an adequate form. These problems diminish in the course of time, because the amount of knowledge in the system increases and the user becomes more skilful in using the system.

8.4.5 User interface
Although most of the time in this study has been spent on the internal part of the system, some attention has already been paid to the exterior, i.e. the user interface. During the processes of knowledge acquisition and structuring a first prototype of the user interface has been developed. This exercise has contributed to the development of ideas on the user interface, which are presented in this sub-section. Some screen-dumps of this prototype are used as illustrations. Since this prototype has been developed before most of the relevant knowledge was structured, the screen-dumps do not completely agree with the design considerations of the user interface presented here.

The user interface has to ensure that the system is easy to use. Therefore, a number of screens will be available. These screens do not completely coincide with the tasks distinguished. This will be clarified below. At any time it has to be possible to return to a previous screen to assess the decisions made earlier. Moreover, a help facility should be

available to give information on how the system operates (e.g. how to return to a previous screen, to change parameter values, or to store results). Furthermore, the user will always have access to the explanation facility. These options will be presented in a menu bar, e.g. at the top of each screen. Figure 8.6 presents a conceptual structure for organizing the screens.

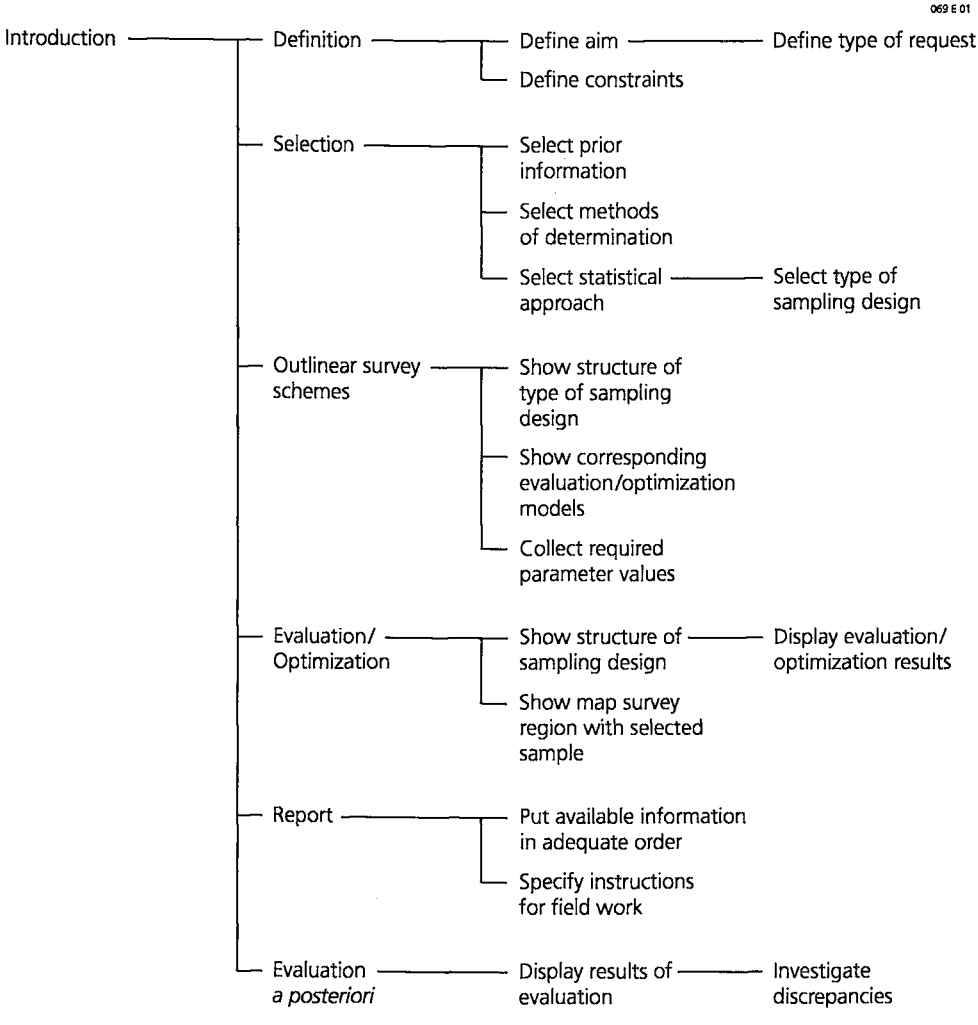


Figure 8.6 Conceptual structure of screens

The system will start with an introduction screen, after which the design of a soil survey scheme for a project can start.

On the definition screen the survey request needs to be specified. If the number of options for instantiation of a concept is limited, the system may present these options in a pull-down menu, from which the appropriate options can be selected with the mouse. Furthermore there

should be a separate window in which background information can be presented. Figure 8.7 shows the definition screen in the prototype system. (In the prototype of the user interface there is no selection screen and the selected prior information is also presented on this screen.) Once the aim has been specified a special window will appear to specify the type of request.

The selection screen will present selected prior information and methods of determination to the user and will allow the user to add or adjust this information. There will be separate windows for the prior information and the method of determination. Besides, there will be a separate screen for the selection of a sampling approach. Selection of applicable types of designs will be related to the selection of the statistical approach. Instead of distinguishing separate screens for the selection of prior information and the design of outlinear schemes, all selections are grouped on one screen, because the selected prior information and the selected methods of determination influence the selection of an appropriate sampling approach and the selection of types of sampling designs.

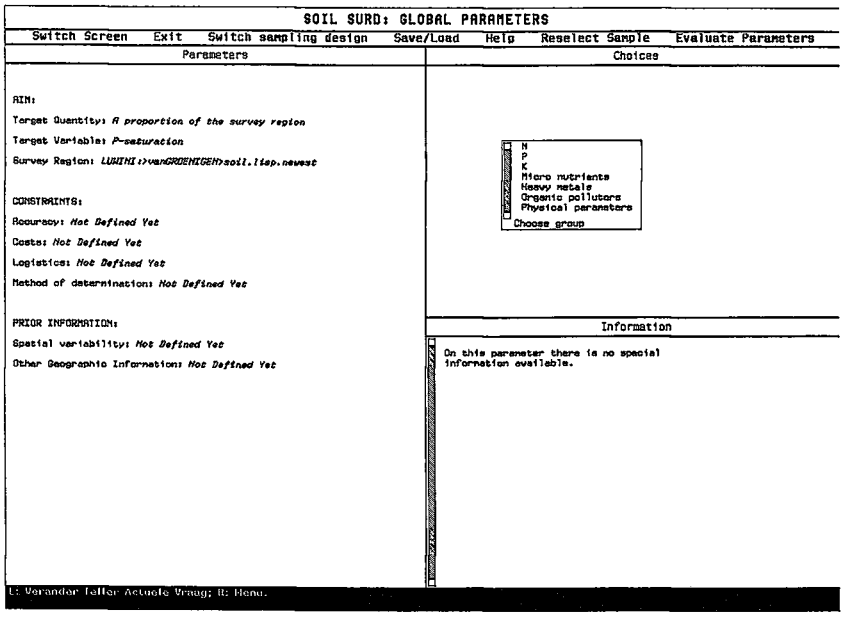


Figure 8.7 Definition screen in the prototype system

On the next screen information related to outlinear survey schemes can be displayed and collected such as the structure of the type of sampling design, the models for evaluation and optimization, and the corresponding parameter values. The results of evaluation and optimization will be presented in detail on the evaluation / optimization screen. To give the user insight into types of sampling designs the structure of types of designs may be presented in a tree structure (Fig. 8.8). It should be possible to supply information at every node of this tree. For example, for the *STSI* design in Figure 8.8, information on the whole survey region will be retrievable at the root, and at the stratum nodes information on each particular stratum

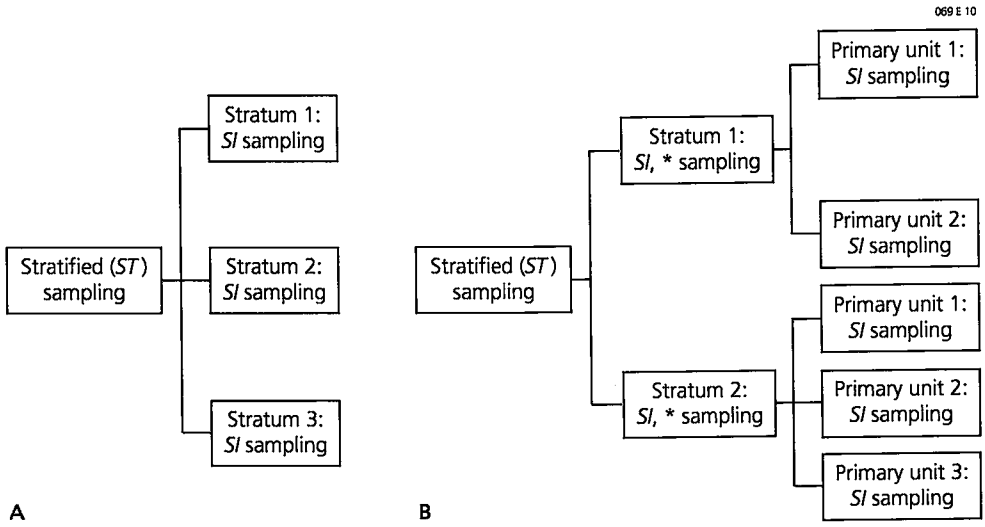


Figure 8.8 Graphical representation of types of sampling designs: A. STSI sampling; B. STSI, SI sampling

will be available, e.g. the size of the stratum, the co-ordinates, and, after optimization, the umber of sample points. In the prototype system, where the sampling designs are presented in this way, the user may ask for information at a node by activating it with the mouse. Then, information on that node appears in a separate window on the screen. So, the user decides on which parts of the sampling design information is required and in which order. This seems a flexible, and therefore attractive approach to provide the user with information on the sampling design. Since the system will have to deal with spatial problems it will also be helpful if the user can get insight into the spatial consequences of various sampling designs. Therefore, it should be possible to present a map of the survey region on which the observation points of a selected sample for an optimized scheme are marked.

Figure 8.9 and 8.10 show how information on spatial consequences of sampling designs can be presented in the prototype system for a fictitious survey project with a square survey region.

Finally, the information to be presented in the report of the soil survey scheme will be shown on the report screen. Here details of the designed scheme will be displayed and there will be a window to specify the instructions for field work.

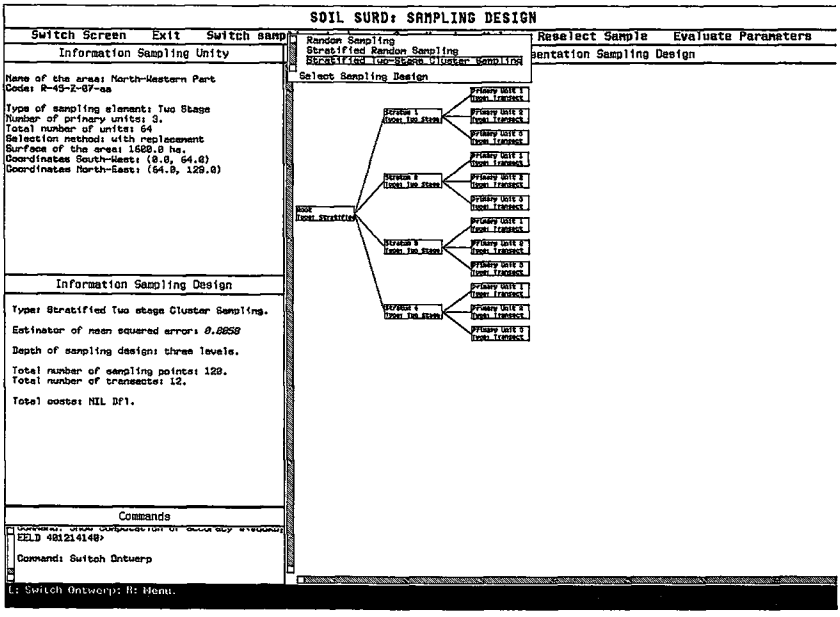


Figure 8.9 Presentation of information on sampling designs in the prototype system

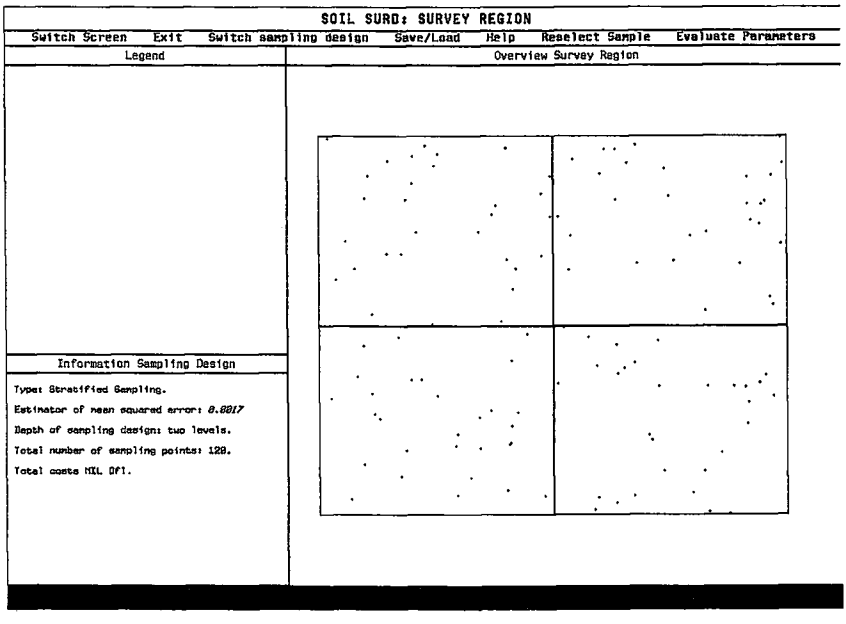


Figure 8.10 Graphical representation of a sample in the survey region in the prototype system

8.5 Evaluation of the components

The preceding section described the components for an actual KBS. These components are not sufficiently detailed for implementing a KBS to be used in practice. Below the components of the KBS are evaluated on their merits and their completeness.

Database
The composition and the contents of the database are well-structured (Sub-section 8.4.1), but the procedure for collection and storage of data has to be elaborated further. Special attention needs to be paid to the way in which consistency can be checked, and the range of parameter values can be controlled.

Knowledge base
Most of the statistical knowledge to be stored in this component of the KBS has been structured in Chapter 6. Further attention needs to be paid to formulating rules to assist in the selection of appropriate sampling designs. Besides, knowledge on the applicability of statistical strategies and of types of variograms in soil survey practice need to be collected and structured to complete the statistical knowledge in the knowledge base.

In this study the required function of the explanation facility has been formulated (Sub-section 8.4.2). However, little attention has been paid to the structure of this facility. The structure of the explanation facility and the knowledge required for explanation need to be specified further.

Model base
The relations between the various mathematical models gave rise to a hierarchical structure of the model base (Sub-section 8.4.3). Most important models for storage in the model base are available, e.g. the evaluation models, the variogram models, and the formulae of methods of inference. In the short term the collection of optimization models might be extended; initially, for frequently used types of designs, at a later stage, for different problem formulations, e.g. to minimize the cost for a given accuracy constraint.

Problem-solving model
The problem-solving model has been described, but not specified in detail in Sub-section 8.4.4. A flow diagram of the problem-solving model for the design of soil survey schemes has been shown in Figure 8.5. This diagram will have to be refined to further specify the problem-solving model. When specifying the problem-solving model, special attention needs to be paid to the specification of *a posteriori* evaluation in the problem-solving model, since this task has only been given limited attention here. It is difficult to indicate which other parts of the problem-solving model need special attention, since this chapter only describes basic design considerations. However, specification of this model needs thorough attention since it will determine how the system operates by serving as the interface between a user and the various pieces of knowledge in the database, the knowledge base and the model base. Some examples of elements that need further attention are preventing improper use (see Sub-section 8.4.4) and handling prior information. Prior information plays an important role in the design of soil survey schemes. Sub-section 8.4.4 showed that while consulting the

system, a significant amount of information is required from the user. The system may have to operate on the basis of uncertain information, e.g. the parameter values for the variogram models and cost models may have to be guessed. This uncertainty will influence the results of evaluation and optimization. The user has to be aware of the quality of the outcome. Therefore, it might be advisable to evaluate and optimize outlinear schemes for various parameter values to give insight into the sensitivity of the procedure for (slightly differing) parameter values.

User interface
Sub-section 8.4.5 has presented a conceptual structure of screens and some ideas on the presentation of information on these screens. Further specification of the user interface can be discussed with the intended users when developing a prototype of the whole system.

Although the components for an actual design have to be worked out further, they may - as parts of the basic design considerations presented in this chapter - suffice for the development of a prototype of the whole system. A prototype will help to refine the requirements and the components for an actual design. Furthermore, a prototype will be used to involve users in the further development of the system, since the knowledge and skills of the intended users also influence further development.

Chapter 9

Concluding remarks

9 Concluding remarks

9.1 Main results and conclusions

The use of statistical approaches to soil surveying makes it possible to quantify the accuracy of the survey results. Before such information can be produced, a soil survey scheme has to be designed specifying which sites are to be sampled, which data are to be recorded and how they are to be analysed statistically. Soil survey results with quantified accuracy are relevant to decisions on land use or on environmental issues. The design of soil survey schemes is often hampered by obstacles related to the re-use of knowledge from past surveys, the availability of prior information and the shortage of information on spatial variability and the accuracy of results. The limited time available for designing a scheme, and the lack of procedures for prior evaluation can also play a role. At present, computerized support during the design of soil survey schemes is rather limited. Sometimes, prior information is retrieved from a database or a GIS database, or the selection of a sample is (partially) automated, but in general most tasks and sub-tasks during the design process are performed manually. This study has investigated the possibilities for providing computerized support for the design of soil survey schemes using statistical approaches.

The results of this study are integrated in the *basic design considerations* (Chapter 8) for a KBS to assist in the design of soil survey schemes. This KBS will assist in the tasks and sub-tasks distinguished in the design process (e.g. Section 5.4 and Sub-section 8.2.2). In Section 1.5 six research questions have been formulated to identify basic design considerations for the KBS. In the following, the answers to those questions and the conclusions arrived at are presented.

1. How can the design of soil survey schemes be structured?
The domain of soil survey projects has been structured to make further investigation of computerized support possible (Chapter 4). The KBS will have to function smoothly in the existing working environment, therefore to start with, an entity structure of a soil survey project has been developed (Section 4.3). Designing a soil survey scheme is only a part of this. Moreover, the actual process of designing soil survey schemes has been modelled to serve as a basic structure for the design of the KBS (Section 4.4). Finally, all important concepts have been structured and defined in the conceptual framework to ensure effective communication (Section 4.5).

The entity structure has clarified the need for a system that learns from its use. To improve the re-use of information from historical surveys, new knowledge has to be continuously collected and stored with the help of the system itself, i.e. what is required is a simple form of a self-learning system (Section 5.4). This collection and storage of knowledge will not only take place during the design of survey schemes but also through evaluation *a posteriori*, i.e. after execution of a plan of action prescribed in a survey scheme.

The model of the design process has been used as a starting point for analysing the problems in designing soil survey schemes (Chapter 5). This resulted in a specification of the tasks to be supported (Section 5.4). After structuring the knowledge about methods of

determination and statistics required to design outlinear survey schemes (Chapter 6), and after developing models for evaluating and optimizing survey schemes (Chapter 7), the list of tasks and sub-tasks has been updated (Sub-section 8.2.2). The model of the design process, the tasks distinguished, and the conceptual framework will already be useful before any system has actually been implemented, because they ensure that all relevant aspects are properly considered and discussed. At present, the lack of a structured approach to designing survey schemes hampers the re-use of knowledge and the quality control of surveys. If survey schemes are developed using the same structured approach, they are comparable, and schemes can be verified by comparison. Furthermore, a structured design process will help to control the quality of the survey results and so facilitate the re-use of knowledge in other survey projects.

The development of the conceptual framework has revealed that the applicability and meaning of concepts is not always clear. The main sources in the relevant literature differ also in terminology, and often no sharp definitions are given at all. The conceptual framework provides unambiguous definitions of the relevant concepts in the soil survey domain and makes clear how they are related. This framework can already be used in consultations on soil survey projects to prevent confusion.

A KBS developed according to the basic design considerations described in Chapter 8 will ensure the use of a structured approach. Such a system will strengthen the effects of a structured approach for designing soil survey schemes on the re-use of knowledge and on the quality control of surveys as mentioned above.

2. What are the main decision problems during the design of soil survey schemes?
Initially, the problems during the design process have been analysed according to the model of the design process (Chapter 5). The resulting specification of the tasks has been updated later (see Sub-section 8.2.2). The tasks finally distinguished, with their main problems, are summarized below. The solutions to some of these problems are briefly mentioned. The answers to the remaining research questions (3, 4, 5, and 6) show further how these problems can be solved.

Define request. At the start of the design process, the survey request, i.e. the aim and constraints, needs to be specified. If the request is not fully and properly specified this may hamper the design of a survey scheme or even result in an inappropriate scheme. Researchers are not always aware of the possibilities with respect to types of results of spatial inventories which may lead to the aim being specified incorrectly. The types of results are related to types of requests, i.e. *how much*, *where*, or both *how much & where* (Section 1.1). These types of requests are related to the aim. The specification of the type of request needs to be given thorough attention to prevent an inappropriate scheme being designed. Figure 1.2 has shown how the three types of request distinguished are related to statistical approaches, i.e. to a design-based or to a model-based approach. However, there are also survey requests that do not belong to one of those types of requests. At the start of the design process, it needs to be determined whether the KBS is able to assist in designing a survey scheme for the specified request.

Select prior information. At present, the main problems related to the selection of prior information are the limited availability of information and the limited amount of information on spatial variability and on the accuracy of survey results (Section 5.2.1). Relevant information will be stored and continuously updated in the system (see answers to research questions 3 and 6). Moreover, when it is used, the amount of information on the spatial variability of soil properties will increase, and the use of statistical approaches to soil surveying will make it possible to quantify the accuracy of survey results in most situations. So, the availability and the quality of prior information will improve.

Design of outlinear schemes. The problems to deal with when designing outlinear survey schemes are:

- how can an appropriate method of determination be selected?
- how can an appropriate sampling approach, i.e. a design-based or a model-based approach, be selected?
- in the case of a design-based approach: how can appropriate types of sampling designs be selected?

The selection of an appropriate method of determination will be left primarily to the user; the system will ask for relevant information on features of a method of determination (Section 6.2). If information on the accuracy and cost of selected methods of determination is available, the system will be an effective tool for comparing various methods. The answer to research question 3 also provides the answers to the two other problems.

Evaluate and optimize outlinear schemes. At the start of this study, there were no models available to evaluate the accuracy and cost of survey schemes and to optimize the sampling design within an outlinear scheme. This lack of models hampered objective comparison of schemes and efficient allocation of available resources. In general, the efficiency of schemes was roughly assessed. Research questions 4 and 5 aimed at generating the knowledge needed to assist in evaluating and optimizing outlinear schemes.

Generate report. The reports of historical surveys are often incomplete and the decisions made are hardly ever justified, since there are no general prescriptions for reporting surveys. This hampers the re-use of knowledge from historical survey projects. The KBS aimed at will assist in generating a report of the final soil survey scheme, including a justification of the decisions made (see research question 6).

Evaluate a posteriori. After execution of a plan of action prescribed in a survey scheme, the scheme should be evaluated, i.e. the prior evaluation should be compared with the actual results. At the moment, evaluation *a posteriori* is not included in soil survey projects, which hampers the collection and storage of new knowledge. The KBS will assist in evaluation *a posteriori* (see research question 6).

3. *How can relevant knowledge and prior information be stored, selected and used to design schemes?*

In Section 1.5 it has been stated that pedological and statistical knowledge should be readily available in the KBS. The description of the components for an actual KBS (Section 8.4) has made clear that the knowledge from various sources will be stored in four components of the KBS: the problem-solving model, the database, the knowledge base, and the model base. Sub-section 8.2.1 has pointed out that within the domain of interest two types of knowledge can be distinguished, namely fixed knowledge and 'variable' knowledge. The fixed knowledge of the structure of the design process, of the order in which questions are to be posed, and of the way in which the solution space may be confined, is integrated in the problem-solving model. Knowledge which will be updated and extended when the KBS is used is called 'variable' knowledge. This knowledge is stored in the database, the knowledge base, and the model base, depending on its specific nature. The storage, selection and use of the pedological and statistical knowledge in the various components of the KBS is briefly described below.

The problem-solving model is related to the tasks distinguished (Sub-section 8.4.4). Appropriate fixed knowledge is selected through interaction with the user. The system will ask questions and will use the answers provided by the user for confining the solution space. The problem-solving model serves as an interface between the user and the components of the KBS where the 'variable' knowledge is stored.

Information on soil survey projects will be stored in the database (Sub-section 8.4.1), e.g. the prior information used, the method of determination, and the final sampling design. Sub-section 8.4.1 has presented a conceptual data model with the main entities and their attributes in relation to soil survey projects. During the design of a survey scheme, information from similar past surveys may be retrieved from the database and information relating to the actual survey can be stored. The information in the database serves as a source of prior information and will be used during the generation of a report. The number and types of parameters needed for the evaluation and optimization of schemes vary between projects, e.g. due to a different type of design, or another method of determination. Therefore, it has been suggested to store parameter values from prior and *a posteriori* evaluation, and parameter values for variograms in templates, which will be specifically created for each survey project.

The knowledge base will contain statistical knowledge and knowledge related to explanation (Sub-section 8.4.2). In Chapter 6 attention has been paid to the statistical knowledge. A decision tree has been presented to guide the selection of an appropriate sampling approach, i.e. a design-based or model-based approach. Given the limitation of the domain to assistance in the use of classical sampling theory (Section 2.2), knowledge on design-based sampling has been further structured. Since there was no suitable classification of sampling designs available, a hierarchical framework of sampling designs has been constructed in which sampling designs are grouped into types of sampling designs, and types are grouped into classes of sampling designs. Furthermore the main classes of sampling designs treated in the literature on sampling have been ordered in a taxonomy. Finally, attention has been paid to the selection of types of designs: a decision tree may be used as a guide when selecting appropriate classes of designs and rules may assist in the selection of types of designs. The knowledge base will also assist in providing an explanation of how solutions or sub-solutions are reached, why a question needs to be answered, what

the consequences of various options are, and what the meaning is of technical terms (Sub-section 8.4.2). The interaction with the user, which is controlled by the problem-solving model, determines which statistical knowledge is appropriate for a survey. Explanation will be provided when the user asks for it.

The (mathematical) models required during the design of soil survey schemes will be stored in the model base (Sub-section 8.4.3), e.g. evaluation models, variogram models, and optimization models. These models are related and have been ordered in a hierarchical structure which serves as the structure of the model base. The next research questions (4 and 5) go into the knowledge required for to evaluate and optimize outlinear schemes.

4. How can schemes be evaluated in advance with respect to accuracy and cost?
To enable objective comparison of survey schemes, models for prior evaluation, i.e. prediction, of accuracy and cost have been developed (Sections 7.2 and 7.3).

The accuracy has been defined as the mean squared error due to sampling. Initially, a general algorithm for sampling-error prediction was developed which can be used with any sampling design. A disadvantage of this algorithm is that it is rather time-consuming. Therefore, specific algorithms for types of designs, based on the variance formulae, have been introduced. Prior information in the form of variograms is needed to predict the accuracy.

Cost models have been defined for various classes of designs. The spatial pattern of sample points as generated by a sampling design is incorporated in these models. The cost models require information on the survey cost (mainly determined by the salaries of personnel), the cost of equipment and of laboratory analysis, and furthermore on the expected access time and observation time in the field.

The models presented for prediction of sampling error and cost enable objective comparison of schemes. These models may give the user insight into the consequences of various decisions in the design. It will not always be easy to provide accurate prior information to determine the parameter values needed for the evaluation. When the information required is uncertain, the sensitivity to this uncertainty can be determined by repeatedly predicting accuracy and cost for various parameter values. It should be noted that actual accuracy and cost may be influenced by causes which are not accounted for in the models and which cannot be controlled by the KBS. For example, inaccuracies in handling during sampling, during transport of samples, and when preparing a sample for laboratory analysis, all influence the actual accuracy and cost. The prior evaluation can only give insight into the expected accuracy and cost.

5. Can an optimal soil survey scheme be found?
The use of dynamic programming (DP) seems suitable for optimizing survey schemes, i.e. calculating an optimal allocation of sample points over the survey region, taking into account constraints on accuracy and cost (Section 7.4). This approach repeatedly uses the models for predicting accuracy and cost. The DP approach is particularly appropriate for optimizing stratified designs, which are frequently used for soil surveying, since it allows differences between sub-regions (strata), e.g. concerning spatial variability or cost parameters, to be taken into account. When optimal allocations have been calculated for a number of possible outlinear schemes, then the scheme which seems best under existing constraints, i.e. the most efficient scheme, may be selected. The procedure can ensure that the available

resources are used optimally in the final scheme (for a given method of determination and type of sampling design). Optimization models have been formulated, aiming at minimizing the sampling error for a fixed budget. Preliminary experiments have shown that an optimal allocation can be calculated within a few minutes.

There are classes of designs for which DP optimization is not appropriate. For example, *SY* samples can be easily optimized by calculating the minimal distance allowed between adjacent points and for classes of designs without strata, a short cut can be used for optimization, i.e. the largest possible sample size, which results in the smallest sampling error, can be calculated directly. DP does not seem to be suitable for optimizing two-stage stratified designs. These types of designs require more complex optimization procedures.

There is no standard DP formulation that suits a large number of sampling designs. Different formulations are required for different classes of designs, and for types of designs within the same class, the corresponding models for sampling error prediction have to be inserted into the mathematical model. Furthermore, different DP formulations are required if the objective of optimization is to minimize the cost for a given accuracy constraint instead of minimizing the sampling error for a fixed budget.

6. How should a system to assist in the design of soil survey schemes be constructed?
The answers to the preceding research questions have been integrated in the basic design considerations of a KBS to assist in the design of soil survey schemes (Chapter 8). These design considerations consist of an initial requirements definition of the KBS, a description of the intended use of the system, and a specification of the components for an actual KBS. The main ideas on the construction of the KBS presented in the basic design considerations are given below.

The KBS will consist of the following five components: a database, a knowledge base, a model base, a problem-solving model, and a user interface. The contents of the first four components have been already discussed (research question 3). The user interface will be an important component of the system, because considerable interaction with the user is required, as in any KBS. In Sub-section 8.4.5 a conceptual structure of screens has been presented. A well-designed user interface will facilitate the use of the system. All components of the KBS need to be elaborated further for implementing a KBS to be used in practice. However, they may be used as a starting point for the development of a prototype of the whole KBS, which may itself be used for elaborating the design considerations.

The main characteristics of the KBS summarized in Sub-section 8.2.1 are listed again below:

- knowledge of the fixed structure of the design process is included in a problem-solving model;
- the knowledge in the system will comprise data, rules, and (mathematical) models (stored in the database, the knowledge base, and the model base, respectively) to be updated without changes in the problem-solving model;
- a large amount of knowledge is stored from which only a small part is used to assist in the design of a scheme for a specific soil survey project;
- interaction between the system and a user results in the selection of appropriate knowledge for a specific problem.

The intended use of the system needs to be considered when constructing a KBS. Here, the use is described briefly. The system will be used for actual soil survey projects, and can also be used as a tool to explore the effects of regulation on soil surveying (Section 8.3). Different parties involved in soil survey projects can profit from the system, but two specific user categories are distinguished: researchers at institutes and universities, and environmental consultants. The intended users must have at least a basic notion of the use of sampling strategies in soil surveys (see Section 8.3).

It should be noted that the system requires a different organization compared to the current practice of soil survey design, e.g. the user will be forced to specify the survey request and to answer a large number of questions, and a survey should always be evaluated after collection and analysis of the data, and therefore a log-book should be kept during field work. The proposed approach to the design of survey schemes is more extensive than current practice, which may result in greater demands on the user, since the system proposed will not replace the whole consultation process but, as already stated, will still require considerable input from the users. However, since the whole process is structured and since there will be an explanation facility, the design of survey schemes can be facilitated. It should be noted that this does not imply that there is no need to start in good time with the design of survey schemes. Some tasks can be executed faster and better, but it may still occur that the required prior information is not directly available, e.g. the values for cost parameters, or that the user needs time to think about a decision, e.g. an appropriate stratification of the survey region. Then, the user must have enough time to take well-founded decisions or to provide appropriate information. In any case the approach proposed will assist in controlling the quality of soil survey projects.

9.2 Applicability to other survey domains

In Section 2.2 the domain of the system was limited to:

- soil surveys for which a design-based approach is appropriate (at a later stage, the system should be broadened to assist in geostatistical sampling);
- sampling of points in a plane;
- survey requests with one (main) target quantity and one (main) target variable.

This study has focused on surveys within this limited domain. Survey requests within this domain may aim at inventories of all kinds of soil properties for various objectives, e.g. for soil science, for environmental issues, or for nature conservation. This section discusses the applicability of the whole system or of parts of the system to surveys in domains that are closely related to the limited domain of interest.

9.2.1 The use of classical sampling theory

Again it is noted that this study has focused on those soil surveys for which a design-based approach, i.e. the use of classical sampling theory, is appropriate. This sub-section outlines the applicability of the proposed system outside the typical case of a new soil survey. Three situations are identified:

- repeated sampling;
- soil sampling within field experiments;
- inventories of non-soil populations.

Repeated sampling

It may be that a survey region has been sampled before, but that the results of this sample are insufficient for the present purpose. So, additional information has to be collected. Then, the available survey results can be used as part of the additional survey, as long as the same type of sampling design is used and the specification of sub-regions is equal. Suppose that in the historical case the type of design was a *STSI* design with selection of sample points with replacement and equal probabilities. Then, the information collected at these sample points may be used in the additional survey, if the same type of design is used and the specification of the strata is equal. When the additional design has to be optimized using DP, the number of sample points in a stratum in the previous survey must be used as a minimum constraint for the number of sample points in that stratum. The data of the previous and the additional survey may be combined for data analysis.

Soil sampling within field experiments

In experimental designs, fields often have to be sampled as part of the experiment. In such cases, the system may assist in the design of a survey scheme with the fields considered as strata and the type of request characterized as a *how much & where* request. So, the system may assist in designing part of a scheme for an experimental design.

Inventories of non-soil populations

Although the cases used for illustration in the preceding chapters were based on regional soil surveys in the Netherlands, the approach to the design of survey schemes may be used for survey projects of different sizes and in other countries. The models for prediction of accuracy and cost are generally applicable. However, it should be noted that the parameter values will depend on the local conditions of the survey project and the survey region; these values will differ among regions and countries, and between small-scale and large-scale surveys.

Classical sampling theory is generally applicable and is not restricted to specific domains. The structuring of sampling designs (Chapter 6) is relevant to all domains were classical sampling theory can be applied, e.g. surveys among the inhabitants of a town, surveys of vegetation in a region, or surveys of industrial products. It may facilitate the selection of appropriate types of designs for survey requests in all these domains. However, the whole approach to survey design arrived at may not be directly applicable to other domains. For other domains adjustments may be required, e.g. related to the availability and form of prior information, or to evaluation models.

An example of a domain which is closely related to soil surveying is sampling of aquatic sediments. A system to advise on sampling strategies for the sediments in lakes, taking into account constraints on accuracy and cost, is discussed by Wehrens et al. (1993). The sampling strategies their system advises are always systematic grids, with different grid sizes for sub-regions. Their system focuses on a very limited domain: strategies for lakes. Although their system aims at answering *how much* requests, the possibility to use classical sampling theory is not addressed; it uses geostatistical techniques (kriging) and autocorrelation techniques (Wehrens, 1993). Differences between sampling terrestrial and aquatic soils cannot be disregarded, e.g. in general much more prior information is available for terrestrial soils, temporal variation is of particular importance to aquatic sediments, and the costs of sampling are much higher for aquatic sediments. However, it might be attractive to develop a system which can assist in a number of sampling strategies for both soil types. The same types of requests are relevant to both domains: *how much*, *where*, and *how much & where*. Assistance in the use of the same statistical approaches therefore seems relevant to both domains.

9.2.2 Not just point sampling in the plane

In practice, there are a great deal of soil survey requests that do not belong to the category of sampling points in a plane. The sample point may, for example, be a composite sample or a point in a three- or four-dimensional space. If appropriate variograms are available and if the cost models can be adapted to these requests, the system may also assist in the design of efficient schemes for these requests.

Composite samples

In the practice of soil surveying *composite samples* are often used to reduce analysis cost, i.e. samples from a number of locations are mixed and analysed as one sample. For such surveys, the configuration and number of locations that make up the composite sample need to be specified and the sample size equals the number of composite samples. If a variogram is available which describes the spatial dependence between these composite samples, the system may optimize outlinear survey schemes with a composite sample as the sampling element. It is also possible to consider taking composite samples as a special type of two-stage sampling with selection of one composite sample within each selected primary unit. Then, outlinear schemes with a two-stage design may be evaluated and optimized using a variogram for population elements (as in the cases described in Chapter 7).

Three- or four-dimensional space

There are also survey request which require the depth of sample points to be randomly selected, which means that points need to be selected in a three-dimensional space. E.g. in the case of soil pollution the aim could be to determine *how much* of a pollutant is present in the upper 1.5 m. Then samples taken at various depths may be separately analysed, or composite samples can be taken over depth. If corresponding variogram models and cost models are available, the system may assist in designing survey schemes for these survey requests.

Sample surveys in space and time (four-dimensional) do also require specific variograms and cost models. Papritz (1993) discusses the use of design-based and model-based methods for monitoring temporal change of soil contamination on field plots. The approach to designing

survey schemes described in this study is in principle also applicable whenever classical sampling theory can be used for monitoring and the target variable is defined as a change per time unit.

9.2.3 Multiple criteria requests

As stated in Sub-section 2.2.3, there is no operational knowledge on developing schemes for multiple target quantities and target variables. In practice, the design of a survey scheme is based on one important quantity and one important variable. However, if there are a number of variables of interest, the system can predict the accuracy for each of the variables for which an appropriate variogram is available. The costs models allow for considering the cost of multiple soil properties at the same time. If one soil property has been indicated as target variable the cost of determining values for the additional soil properties can be taken into account by including the additional time for observation and the additional cost of analysis in the parameter values of the cost model. If values for a number of target variables are collected at the sample points, there still needs to be one target variable of interest for the objective function for the optimization, which aims at minimizing the sampling error for one target variable. When information on multiple variables is included in the cost parameters, the higher cost per sample point will result in a smaller sample size allowed and so also influence the design of outlinear schemes, and the predicted accuracy. So, the system will be able to handle the consequences of multiple target variables to some degree.

9.3 Further developments

This study resulted in a structured approach to the design of soil survey schemes and basic design considerations for a KBS to assist in this design process. These basic design considerations will be used as a starting point for the development of a prototype. Such a prototype can be used to further specify the requirements and to elaborate the design for a KBS to be used in practice. It can also facilitate communication with the intended users.

In Section 8.5 the components for an actual design have been evaluated. There, a number of issues were mentioned that need further attention. These are summarized and complemented below.

- Procedures need to be developed to check the consistency of the knowledge in the system, and to prevent knowledge being stored twice.
- The set of rules in the knowledge base to assist in the selection of types of designs, which may be partially based on experience in soil survey practice, needs to be elaborated further.
- A design of the explanation facility should be developed. In the final system this facility will be rather comprehensive. A hypertext-like approach seems attractive, because it is very flexible and allows each user to select the information he or she requires.
- The sensitivity of the evaluation and optimization models for inaccurate prior information should be analysed (sensitivity analysis).
- The cost models should be tested in practice and slightly adjusted if necessary.
- Optimization models should be developed for other types of designs, e.g. for designs that

require stochastic DP. Available knowledge on optimizing schemes for soil surveys using geostatistics (McBratney et al., 1981; McBratney & Webster, 1981) may be used when further assistance in the design of geostatistical strategies is supplied.

- Efficient procedures for sample selection need to be developed, and therefore the required accuracy for localizing sample points should be determined. If grid maps are used for sample selection, the desired grid size should be determined; the larger the number of grid cells, the more computer memory is required. The accuracy with which sample points are selected should be in accordance with the accuracy with which they are located in the field.
- The problem-solving model needs to be refined.
- The user interface needs to be elaborated. A prototype of the whole system may be helpful in this respect.
- The potential for assisting in the design of geostatistical survey schemes will be further investigated. In the first place the type of request determines which statistical approach is most appropriate (design-based versus model-based), but besides, the constraints determine which approach suits best in practice. Knowledge about geostatistical strategies should be structured and the criteria for selecting an appropriate strategy need to be formalized. A system which can assist in the use of both approaches to soil surveying, and which can evaluate and optimize various survey schemes would be extremely valuable in the practice of soil surveying.

References

References

Adèr, H.J. (1992)
Formalisation Methods for Statistical Domain Knowledge. In: F. Faulbaum (Ed.), SoftStat '91: Advances in Statistical Software. Stuttgart, Gustav Fischer, pp. 3-11.

Amarel, S. (1987)
Problem Solving. In: Shapiro, S.C. (Ed.), Encyclopedia of Artificial Intelligence, Volume 2. New York, John Wiley and Sons, pp. 767-779.

Barcelona, M.J. (1988)
Overview of the Sampling Process. In: Keith, L.H. (Ed.), Principles of Environmental Sampling. U.S.A., American Chemical Society, Professional Reference Book, pp. 3-23.

Barr, A. & Feigenbaum, E.A. (1981)
The Handbook of Artificial Intelligence, Volume I. Los Altos, California, William Kaufmann, Inc.

Barr, A. & Feigenbaum, E.A. (1982)
The Handbook of Artificial Intelligence, Volume II. Los Altos, California, William Kaufmann, Inc.

Beardwood, J., Halton, J.H. & Hammersley, J.M. (1959)
The shortest path through many points. In: Proceedings of the Cambridge Philosophical Society, 55, pp. 299-327.

Beek, P. van & Hendriks, Th. H. B. (1985)
Optimaliseringstechnieken: principes en toepassingen. 2nd edition. Utrecht, Bohn, Scheltema & Holkema. (in Dutch)

Bell, M.Z. (1985)
Why Expert Systems Fail. Journal of the Operational Research Society, 36(7), pp. 613-619.

Bellman, R. (1957)
Dynamic Programming. Princeton, New Jersey, Princeton University Press.

Bennett, J.L. (1983)
Building Decision Support Systems. Reading, Massachusetts, Addison-Wesley Publishing Company.

Berg, G.M van den (1992)
Choosing an analysis method, An empirical study of statisticians' ideas in view of the design of computerized support. Ph.D. thesis. Leiden, DSWO Press, Leiden University.

Berg, G.M. van den & Visser, R.A. (1990)
Knowledge Modelling for Statistical Consultation Systems; Two Empirical Studies. In: Momirović, K. & Moldner, V. (Eds.), Compstat Proceedings in Computational Statistics, Dubrovnic, pp. 75-80.

Bie, S.W. & Beckett, P.H.T. (1971)
Quality control in soil survey, II: The costs of soil survey. Journal of Soil Science, 22(4), pp.453-465.

Bonissone, P. (1987)
Plausible Reasoning. In: Shapiro, S.C. (Ed.), Encyclopedia of Artificial Intelligence, Volume 2. New York, John Wiley and Sons, pp. 854-863.

Breeuwsma, A., Reijerink, J.G.A., Schoumans, O.F., Brus, D.J. & Loo, H. van het (1989)
Fosfaatbelasting van bodem, grond- en oppervlaktewater in het stroomgebied van de Schuitenbeek. Rapport 10. Wageningen, DLO Winand Staring Centre for Integrated Land, Soil and Water Research. (in Dutch)

Breeuwsma, A., Wösten, J.H.M., Vleeshouwer, J.J., Slobbe, A.M. van & Bouma, J. (1986)
Derivation of Land Qualities to Assess Environmental Problems from Soil Surveys. Soil Science Society of America Journal, 50(1), pp. 186-190.

Bregt, A.K., Janssen, J.A.M., Griendt, J.S. van de, Andriesse, W. & Alkasuma (1992)
Optimum observation density for mapping acid sulphate soils in Conoco, Indonesia: accuracy and costs. In: Bregt, A.K., Processing of soil survey data, Ph.D. thesis. Wageningen, Agricultural University Wageningen, pp. 41-53.
Breuker, J. & Wielinga, B. (1989)
Models of expertise in knowledge acquisition. In: Guida, G. & Tasso, C. (Eds.), Topics in Expert System Design. Amsterdam, Elsevier Science Publishers, pp. 265-295.
Brownston, L., Farrell, R., Kant, E. & Martin, N. (1985)
Programming expert systems in OPS5, An Introduction to Rule-Based Programming. Reading, Massachusetts, Addison-Wesley Publishing Company, Inc.
Brus, D.J. (1993)
Incorporating models of spatial variation in sampling strategies for soil. Ph.D. thesis. Wageningen, Agricultural University Wageningen.
Burgess, T.M. & Webster, R. (1984)
Optimal sampling strategies for mapping soil types - I, Distribution of boundary spacings. Journal of Soil Science, 35(4), pp. 641-654.
Burrough, P.A. (1982)
Computer assistance for soil survey and land evaluation. Soil Survey and Land Evaluation, 2(1), pp. 25-36.
Burrough, P.A. (1986)
Principles of geographical information systems for land resources assessment. Oxford, Clarendon Press.
Burrough, P.A. (1993)
The technologic paradox in soil survey: new methods and techniques of data capture and handling. ITC Journal, 1, pp. 15-22.
Bylander, T. & Chandrasekaran, B. (1987)
Generic tasks for knowledge-based reasoning: the "right" level of abstraction for knowledge acquisition. International Journal of Man-Machine Studies, 26, pp. 231-243.
Carbonell, J. & Langley, P. (1987)
Machine learning. In: Shapiro, S.C. (Ed.), Encyclopedia of Artificial Intelligence, Volume 1. New York, John Wiley and Sons, pp. 464-488.
Cassel, C-M., Särndal, C-E. & Wretman, J.H. (1977)
Foundations of Inference in Survey Sampling. New York, John Wiley & Sons, Inc.
Chandrasekaran, B., Johnson, T.R. & Smith, J.W. (1992)
Task-Structure Analysis for Knowledge Modelling. Communications of the ACM, 35(9), pp.124-137.
Cochran, W.G. (1977)
Sampling Techniques, third edition. New York, John Wiley & Sons, Inc.
Cyert, R.M. (1981)
The Future of Operations Research. In: Brans, J.P. (Ed.), Operational Research '81, Ninth IFORS International Conference on Operational Research. Reprints - Part I. Amsterdam, North-Holland, pp. 7-14.
Damoiseaux, J.H., Harbers, P. & Teunissen van Manen, T.C. (1990)
Bodemkaart van Nederland 1:50.000, 61-62 West en Oost, Maastricht-Heerlen. Wageningen, DLO Winand Staring Centre for Integrated Land, Soil and Water Research. (in Dutch)
Dannenbring, D.G. & Starr, M.K. (1981)
Management science: an introduction. Auckland, McGraw-Hill Publishing Company.
Dent, D. & Young, A. (1981)
Soil Survey and Land Evaluation. London, George Allen & Unwin.

Domburg, P. & Elzas, M.S. (1994)
Structuring the Domain of a Complex System: a basis for a knowledge-based system supporting soil survey design. In: Beulens, A.J.M, Doležal, J. & Sebastian, H-J. (Eds.), Optimization-Based Computer-Aided Modelling and Design, Proceedings of the second Working Conference of the IFIP TC 7.6 Working Group, Dagstuhl, Germany, 1992. Leidschendam, Lansa Publishing, pp. 181-195.

Domburg, P. & Gruijter, J.J. de (1992)
A framework of concepts for soil survey using probability sampling. Report 55. Wageningen, DLO Winand Staring Centre for Integrated Land, Soil and Water Research.

Domburg, P., Gruijter, J.J. de & Beek, P. van (submitted)
Designing Efficient Soil Survey Schemes with a Knowledge-Based System using Dynamic Programming. Environmetrics.

Domburg, P., Gruijter, J.J. de & Brus, D.J. (1994)
A structured approach to designing soil survey schemes with prediction of sampling error from variograms. Geoderma, 62(1-3), pp. 151-164.

Dreyfus, H.L. & Dreyfus, S.E. (1988)
Making a Mind Versus Modeling the Brain: Artificial Intelligence Back at a Branchpoint. Dædalus, 117(1), pp. 15-43.

Elzas, M.S. (1986)
The kinship between artificial intelligence, modelling & simulation: an appraisal. In: Elzas, M.S., Ören, T.I. & Zeigler, B.P. (Eds.), Modelling and Simulation in the Artificial Intelligence Era. Amsterdam, Elsevier Science Publishers, pp. 3-13.

Elzas, M.S. (1989)
The role of artificial intelligence/entity structure concept manipulation methods in system modelling and design. In: Elzas, M.S., Ören, T.I. & Zeigler, B.P. (Eds.), Modelling and Simulation Methodology. Amsterdam, North Holland, pp. 3-27.

Englund, E.J. (1990)
A Variance of Geostatisticians. Mathematical Geology, 22(4), pp. 417-455.

Finlay, P.N. (1990)
Decision Support Systems and Expert Systems: a comparison of their components and design methodologies. Computers & Operations Research, 17(6), pp. 535-543.

Gale, W.A. (1986)
REX Review. In: Gale, W.A. (Ed.), Artificial Intelligence and Statistics. Reading, Massachusetts, Addison-Wesley Publishing Company.

Greef, P. de (1991)
Analysis of co-operation for consultation systems. Journal of Applied Statistics, 18(1), pp. 175-184.

Gruber, T. R. (1991)
The Role of Common Ontology in Achieving Sharable, Reusable Knowledge Bases. In: Allen, J., Fikes, R. & Sandewall, E. (Eds.), Principles of Knowledge Representation and Reasoning, Proceedings of the Second International Conference. Cambridge Massachusetts, Morgan Kaufmann, pp. 601-602.

Gruijter, J.J. de & Braak, C.J.F. ter (1990)
Model-free estimation from spatial samples: a reappraisal of classical sampling theory. Mathematical Geology, 22(4), pp. 407-415.

Grünwald, H.J. & Fortuin, L. (1989)
DSS and ES in the 'information organization' - Back to the roots of OR. European Journal of Operational Research, 41, pp. 142-150.

Hack-ten Broeke, M.J.D., Kleijer, H., Breeuwsma, A., Reijerink, J.G.A. & Brus, D.J. (1990)
Fosfaatverzadiging van de bodem in twee gebieden in Overijssel. Rapport 108. Wageningen, DLO Winand Staring Centre for Integrated Land, Soil and Water Research. (in Dutch)
Hadley, G. (1964)
Nonlinear and dynamic programming. Reading, Massachusetts.
Hand, D.J. (1984)
Statistical Expert Systems: Design. The Statistician, 33, pp. 351-369.
Hand, D.J. (1985)
Statistical expert systems: necessary attributes. Journal of Applied Statistics, 12(1), pp. 19-27.
Hansen, G.W. & Hansen, J.V. (1992)
Database Management and Design. Englewood Cliffs, New Jersey, Prentice Hall.
Hayes-Roth, F. (1987)
Expert Systems. In: Shapiro, S.C. (Ed.), Encyclopedia of Artificial Intelligence, Volume 1. New York, John Wiley and Sons, pp. 287-298.
Hayes-Roth, F. & Jacobstein, N. (1994)
The State of Knowledge-Based Systems. Communications of the ACM, 37(3), pp. 27-39.
Hendriks, Th.H.B. (1990)
Operationele Research. Agro-Informatica, 3(2), pp. 7-12. (in Dutch)
Herik, H.J. van den (1988)
Informatica en het menselijk blikveld. Maastricht, Rijksuniversiteit Limburg. (in Dutch)
Hesketh, P. & Barrett, T. (1989)
M1, An Introduction to the KADS Methodology. Harlow, Essex, STC Technology Ltd.
Hillier, F.S. & Lieberman, G.J. (1990)
Introduction to Operations Research, Fifth Edition. New York, McGraw-Hill Publishing Company.
Ignizio, J.P. (1990)
A brief introduction to expert systems. Computers & Operations Research, 17(6), pp. 523-533.
Jöckel, K-H. (1986)
Statistical expert systems and the statistical consultant - considerations about the planning stage of clinical studies. In: Haux, R. (Ed.), Expert Systems in Statistics, Selected papers from a workshop, organized by the working group "Computational Statistics" of the German Region of the International Biometric Society. Stuttgart, Gustav Fischer, pp. 27-43.
Jones, B. (1980)
The computer as a statistical consultant. Bias, 7(2), pp. 168-195.
Journel, A.G. & Huijbregts, Ch.J. (1978)
Mining Geostatistics. London, Academic Press Inc.
Journel, A.G. & Posa, D. (1990)
Characteristic behavior and order relations for indicator variograms. Mathematical Geology, 22, pp. 1011-1025.
Keen, P.G. & Scott Morton, M.S. (1978)
Decision Support Systems: an organizational perspective. Reading, Massachusetts, Addison-Wesley Publishing Company.
Kerschberg, L. (1986)
Expert Database Systems, Proceedings From the First International Workshop. Menlo Park, California, The Benjamin/Cummings Publishing Company.
Kleijnen, J.P.C. (1974)
Statistical techniques in simulation, Part I. New York, Marcel Dekker, Inc.
Kleijnen, J.P.C. & Groenendaal, W.J.H. van (1988)
Simulatie: technieken en toepassingen. Schoonhoven, Academic Service. (in Dutch)

Koop, J.C. (1990)
Systematic sampling of two-dimensional surfaces and related problems. Commun.Statist.- Theory Meth., 19(5), pp. 1701-1750.
Korf, R.E.. (1987)
Heuristics. In: Shapiro, S.C. (Ed.), Encyclopedia of Artificial Intelligence, Volume 1. New York, John Wiley and Sons, pp. 376-380.
Krishnaiah, P.R. & Rao, C.R. (1988)
Handbook of Statistics, Volume 6: Sampling. Amsterdam, Elsevier Science Publishers.
Kuilenburg, J. van, Gruijter, J.J. de, Marsman, B.A. & Bouma, J. (1982)
Accuracy of spatial interpolation between point data on soil moisture supply capacity, compared with estimates from mapping units. Geoderma, 27(4), pp. 311-325.
McBratney, A.B., & Webster, R. (1981)
The design of optimal sampling schemes for local estimation and mapping of regionalized variables - II. Program and examples. Computers & Geosciences, 7(4), pp. 335-365.
McBratney, A.B. & Webster, R. (1986)
Choosing functions for semi-variograms of soil properties and fitting them to sampling estimates. Journal of Soil Science, 37, pp. 617-639.
McBratney, A.B., Webster, R. & Burgess, T.M. (1981)
The design of optimal sampling schemes for local estimation and mapping of regionalized variables - I. Theory and method. Computers & Geosciences, 7(4), pp. 331-334.
Newell, A. (1983)
Intellectual Issues in the History of Artificial Intelligence. In: Machlup, F. & Mansfield, U. (Eds.), The Study of Information: Interdisciplinary Messages. New York, Wiley.
O'Keefe, R.M. (1985)
Expert Systems and Operational Research - Mutual Benefits. Journal of the Operational Research Society, 36(2), pp. 125-129.
O'Keefe, R.M., Belton, V. & Ball, T. (1986)
Experiences with Using Expert Systems in O.R. Journal of the Operational Research Society, 37(7), pp. 657-668.
Papert, S. (1988)
One AI or Many? Dædalus, 117(1), pp. 1-14.
Papritz, A.J. (1993)
Estimating Temporal Change of Soil Properties. Ph.D. thesis. Zürich, Swiss Federal Institute of Technology.
Rich, E. (1983)
Artificial Intelligence. Singapore, McGraw-Hill Book Company.
Rich, E. & Knight, K. (1991)
Artificial Intelligence, Second edition. New York, McGraw-Hill.
Rogoff, M.J. (1982)
Computer display of soil survey interpretations using a geographic information system. Soil Survey and Land Evaluation, 2(1), pp. 37-41.
Rosenblatt, F. (1958)
The Perceptron, a Probabilistic Model for Information Storage and Organization in the Brain. Psychological Review, 62(6), pp. 386-408.
Särndal, C-E. (1978)
Design-based and model-based inference in survey sampling. Scand. J. Statist., 5, pp. 27-52.
Särndal, C.E., Swensson, B. & Wretman, J. (1992)
Model Assisted Survey Sampling. New York, Springer Verlag.

Schach, S. (1986)
Computer support for the design and analysis of survey samples. In: Haux, R. (Ed.), Expert Systems in Statistics, Selected papers from a workshop, organized by the working group "Computational Statistics" of the German Region of the International Biometric Society. Stuttgart, Gustav Fischer, pp. 99-110.
Schoumans, O.F., Marsman, B.A. & Breeuwsma, A. (1989)
Assessment of representative soil data for phosphate leaching. In: Bouma, J. & Bregt, A.K. (Eds.), Land qualities in space and time. Wageningen, Pudoc, pp. 201-204.
Shapiro, S.C. (1987)
Encyclopedia of Artificial Intelligence, Volumes 1 and 2. New York, John Wiley and Sons.
Simon, H.A. (1987)
Two Heads Are Better than One: The Collaboration between AI and OR. Interfaces, 17(4), pp. 8-15.
Sluis, P. van der & Gruijter, J.J. de (1985)
Water table classes: a method to describe seasonal fluctuation and duration of water tables on Dutch soil maps. Agricultural Water Management, (10), pp. 109-125.
Smith, J.M. (1986)
Expert Database Systems: A Database Perspective. In: Kerschberg, L. (Ed.), Expert Database Systems, Proceedings From the First International Workshop. Menlo Park, California, The Benjamin/Cummings Publishing Company, pp. 3-15.
Sommerville, I. (1992)
Software Engineering, Fourth edition. Workingham, Addison-Wesley Publishing Company.
Steels, L. (1990)
Components of expertise. AI Magazine, (summer issue), pp. 28-49.
Stefik, M., Aikins, J., Balzer, R., Benoit, J., Birnbaum, L., Hayes-Roth, F. & Sacerdoti E. (1983)
Chapter 3: Basic Concepts for Building Expert Systems. In: Hayes-Roth, F., Waterman, D.A. & Lenat, D.B. (Eds.), Building Expert Systems. Reading, Massachusetts, Addison-Wesley Publishing Company, pp. 59-86.
Stefik, M., Aikins, J., Balzer, R., Benoit, J., Birnbaum, L., Hayes-Roth, F. & Sacerdoti E. (1983)
Chapter 4: The Architecture of Expert Systems. In: Hayes-Roth, F., Waterman, D.A. & Lenat, D.B. (Eds.), Building Expert Systems. Reading, Massachusetts, Addison-Wesley Publishing Company, pp. 89-126.
Stein, A. (1991)
Spatial Interpolation. Ph.D. thesis. Wageningen, Agricultural University Wageningen.
Stein, A. (1994)
The use of prior information in spatial statistics. Geoderma, 62(1-3), pp. 199-216.
Steube, M.M. & Johnston, D.M. (1990)
Runoff volume estimation using GIS techniques. Water Resources Bulletin, 26(4), pp. 611-620.
Steur, G.G.L. (1961)
Methods of soil surveying in use at the Netherlands Soil Survey Institute. Auger and Spade, XI. Wageningen, H. Veenman & Zonen N.V., pp. 59-77.
Swartout, W.R. (1987)
Explanation. In: Shapiro, S.C. (Ed.), Encyclopedia of Artificial Intelligence, Volume 1. New York, John Wiley and Sons, pp. 298-300.
Vieux, B.E. (1991)
Geographic Information Systems and non-point source water quality and quantity modelling. Hydrological Processes, 5(1), pp. 101-113.

Visschers, R. (1993)
Upgrading van de Bodemkaart van Nederland, schaal 1:50 000, door steekproeven in kaarteenheden van veldpodzolgronden Hn21-V en Hn21-VI. Rapport 186. Wageningen, DLO Winand Staring Centre for Integrated Land, Soil and Water Research. (in Dutch)
Vleeshouwer, J.J. & Damoiseaux, J.H. (1990)
Bodemkaart van Nederland 1:50.000, toelichting bij kaartblad 61-62 West en Oost, Maastricht-Heerlen. Wageningen, DLO Winand Staring Centre for Integrated Land, Soil and Water Research. (in Dutch)
Wagner, H.M. (1975)
Principles of Operations Research, with applications to managerial decisions. London, Prentice Hall.
Waterman, D.A. (1986)
A guide to expert systems. Reading, Massachusetts, Addison-Wesley Publishing Company.
Waterman, D.A., & Hayes-Roth, F. (1983)
An Investigation of Tools for Building Expert Systems. In: Hayes-Roth, F., Waterman, D.A. & Lenat, D.B., Building Expert Systems, Reading, Massachusetts, Addison-Wesley Publishing Company, Inc., pp. 169-215.
Webster, R. & Oliver, M.A. (1990)
Statistical Methods in Soil and Land Resource Survey. New York, Oxford University Press.
Webster, R. & Oliver, M.A. (1992)
Sample adequately to estimate variograms of soil properties. Journal of Soil Science, 43, pp. 177-192.
Wehrens, R., Hoof, P. van, Buydens, L., Kateman, G., Vossen, M., Mulder, W.H. & Bakker, T. (1993)
Sampling of aquatic sediments. The designs of a decision-support system and a case study. Anal. Chim. Acta, 271, pp. 11-24.
Wielinga, B., Velde, W. van de, Schreiber, G. & Akkermans, H. (1992)
The CommonKADS Framework for Knowledge Modelling. Report KADS-II/T1.1/PP/ UvA/35/1.0. Amsterdam, University of Amsterdam.
Winston, W.L. (1991)
Operation Research: applications and algorithms, 2nd edition. Boston, PWS-Kent Publishing Company.
Zeigler, B.P. (1986)
Toward a simulation methodology for variable structure modeling. In: Elzas, M.S., Ören, T.I, and Zeigler, B.P. (Eds.), Modelling and Simulation Methodology in the Artificial Intelligence Era. Amsterdam, North-Holland, pp. 195-210.
Zeigler, B.P. (1990)
Object-Oriented Simulation with Hierarchical, Modular Models, Intelligent Agents and Endomorphic Systems. London, Academic Press.

Samenvatting

Aanleiding en doel

Als gevolg van de toenemende druk van verschillende vormen van landgebruik en als gevolg van de toenemende zorg voor de kwaliteit van het milieu, groeit de behoefte aan kwantitatieve informatie over de bodem, bijvoorbeeld informatie over de concentraties van bepaalde stoffen in de bodem. Om het risico van onjuiste beslissingen over landgebruik en bodembescherming te beperken, is het van belang ook inzicht te hebben in de nauwkeurigheid van deze informatie. De nauwkeurigheid van bodeminventarisaties kan gekwantificeerd worden als bij het verzamelen en verwerken van gegevens een statistische aanpak is gevolgd. Voordat bemonstering kan beginnen moet een bodeminventarisatieplan worden ontworpen waarin is vastgelegd welke plaatsen bemonsterd moeten worden, welke gegevens verzameld moeten worden en hoe de gegevens verwerkt moeten worden. Bij het ontwerpen van zo'n plan wordt bodemkundige en statistische kennis gebruikt. Belangrijke randvoorwaarden zijn meestal de beschikbare financiële middelen en de eisen ten aanzien van de nauwkeurigheid van de resultaten. Het ontwerpen van een inventarisatieplan wordt vaak bemoeilijkt doordat de gewenste kennis en informatie slecht of niet toegankelijk zijn, de kwaliteit van voorinformatie onbekend is en procedures om vooraf inventarisatieplannen te evalueren (met betrekking tot kosten en nauwkeurigheid) ontbreken. Hierdoor kan beschikbare kennis en informatie niet goed worden benut en kunnen de kosten en nauwkeurigheid van mogelijke plannen niet objectief tegen elkaar worden afgewogen.

Het voorgaande was aanleiding om te onderzoeken hoe het ontwerpen van bodeminventarisatieplannen met de computer ondersteund kan worden. Het doel van deze studie was het leggen van een basis voor het ontwerp van een kennissysteem dat het ontwerpen van plannen voor bodeminventarisaties ondersteunt. Het systeem zou bestaande kennis en informatie beter toegankelijk moeten maken en de mogelijkheid moeten bieden plannen vooraf te evalueren en optimaliseren. De aandacht was daarbij vooral gericht op bodeminventarisaties waarvoor gebruik gemaakt kan worden van de klassieke steekproeftheorie.

Structuur van het domein

Het toepassingsgebied, of domein, voor een kennissysteem moet gestructureerd worden. De structuur van het domein van bodeminventarisatie is beschreven in drie lagen. Allereerst is een overzicht gemaakt van de verschillende fasen in een inventarisatie-project; het ontwerpen van een bodeminventarisatieplan is een onderdeel van een inventarisatie-project. Daarna is de samenhang tussen verschillende fasen in het ontwerpproces weergegeven in een model. Tenslotte is een begrippenkader ontwikkeld, waarin belangrijke begrippen die tijdens het ontwerpproces worden gebruikt zijn geordend en gedefiniëerd.

Taken

Het model van het ontwerpproces heeft als uitgangspunt gediend voor de analyse van de (keuze-)problemen tijdens het ontwerpen. Op basis van deze analyse zijn de taken gespecificeerd die het systeem zal ondersteunen:

- *definitie van de vraag voor bodeminventarisatie* (doel en randvoorwaarden): dit is het uitgangspunt voor het verdere ontwerp en beïnvloedt onder andere welke statistische benadering toepasbaar is (klassieke-steekproefbenadering versus geostatistische benadering);
- *selectie van voorinformatie*: voor de gedefiniëerde vraag moet voorinformatie over de ruimtelijke variabiliteit en andere geografische informatie worden geselecteerd;
- *ontwerpen van globale plannen*: eerst moet een methode worden geselecteerd waarmee waarden van het gewenste bodemkenmerk kunnen worden bepaald (bepalingsmethode), daarna kan een geschikte statistische benadering worden gekozen, en tenslotte kunnen - in het geval van de klassieke-steekproefbenadering - geschikte typen steekproefopzetten worden geselecteerd; omdat in eerste instantie slechts enkele elementen van het uiteindelijke plan gespecificeerd worden wordt gesproken van een *globaal* plan;
- *evaluatie (vooraf) en optimalisatie van globale plannen*: de efficiëntie van een bodeminventarisatieplan wordt bepaald door de nauwkeurigheid van de resultaten en de kosten van de uitvoering van het plan; de nauwkeurigheid en de kosten hangen vooral af van de gekozen bepalingsmethode en de steekproefopzet; om globale plannen te kunnen vergelijken moeten de nauwkeurigheid en de kosten vooraf voorspeld worden (de evaluatie vooraf); door voor verschillende steekproefomvangen de nauwkeurigheid en kosten te voorspellen kan binnen een globaal plan gezocht worden naar een optimale steekproefopzet;
- *maken van een rapport*: het definitieve, geoptimaliseerde plan moet worden uitgewerkt in een rapport.

Naast het ondersteunen van deze taken tijdens het ontwerpen van een inventarisatieplan zal het systeem nog een taak ondersteunen na de uitvoering van het werkplan zoals dat in het inventarisatieplan is voorgeschreven:

- *evaluatie achteraf*: deze taak is vooral gericht op het verzamelen van kennis en ervaring van uitgevoerde projecten om daar in de toekomst weer gebruik van te kunnen maken; door evaluatie achteraf kan het systeem bijdragen aan zijn eigen onderhoud.

Statistische kennis

Na de specificatie van de taken is aandacht besteed aan het structureren van de statistische kennis. Er is allereerst een hiërarchische ordening voorgesteld waarbij steekproefopzetten zijn gegroepeerd in typen steekproefopzetten, en typen zijn gegroepeerd in klassen steekproefopzetten. Een voorbeeld van een *klasse steekproefopzet* is een gestratificeerde aselecte steekproef; een *type steekproefopzet* daarbinnen is een gestratificeerde aselecte steekproef met selectie van steekproefpunten binnen strata met teruglegging en gelijke kansen; van een *steekproefopzet* kan worden gesproken als ook het aantal steekproefpunten in gedefiniëerde strata is gespecificeerd. Naast deze classificatie van steekproefopzetten zijn de belangrijkste klassen steekproefopzetten uit de literatuur geordend in een taxonomie. Vervolgens zijn beslisbomen opgesteld om de keus voor een statistische benadering te ondersteunen (klassieke-steekproefbenadering versus geostatistische benadering), en - in het geval van een klassieke-steekproefbenadering - om de keus voor een geschikte klasse steekproefopzetten te begeleiden. Verder zijn de criteria aangegeven voor het selecteren

van een type steekproefopzet binnen een klasse.

Modellen voor evaluatie en optimalisatie
Er zijn modellen ontwikkeld om de nauwkeurigheid en kosten van steekproefopzetten te voorspellen. De kenmerken van steekproefopzetten, m.n. wat betreft de consequenties voor de ligging van steekproefpunten (verspreid of gegroepeerd in clusters of primaire eenheden), zijn in deze modellen terug te vinden. De nauwkeurigheidsmodellen maken gebruik van voorinformatie over ruimtelijke variabiliteit in de vorm van variogrammen. Voor de kostenmodellen zijn waarden voor verschillende parameters vereist, zoals de analyse-kosten per monster, en de benodigde tijd per steekproefpunt. Bij de evaluatie kan rekening worden gehouden met specifieke voorinformatie voor deelgebieden, bijvoorbeeld verschil in ruimtelijke variabiliteit of verschil in kostenparameters.
Het gebruik van dynamisch programmeren is voorgesteld om binnen een globaal plan te zoeken naar een optimale steekproefopzet. Dynamisch programmeren is vooral geschikt voor het optimaliseren van gestratificeerde steekproefopzetten, die veel in het bodemonderzoek worden toegepast. Het doel van de optimalisatie is om uitgaande van het beschikbare budget een steekproefopzet te vinden die de nauwkeurigheid maximaliseert, ofwel de steekproeffout minimaliseert. Voor twee typen steekproefopzetten is deze optimalisatie uitgewerkt. Dezelfde optimalisatie-procedure kan ook gebruikt worden om gegeven een bepaalde nauwkeurigheidseis de kosten te minimaliseren.

Kennissysteem
De gestructureerde kennis en de ontwikkelde procedures zijn uiteindelijk samengebracht om als basis te dienen voor het ontwerp van een kennissysteem. De volgende vijf componenten zijn in het kennissysteem onderscheiden:

- een *gegevensbank*: hierin worden gegevens over bodeminventarisatieprojecten opgeslagen; door tijdens het ontwerpproces informatie op te slaan in de gegevensbank neemt de hoeveelheid (voor-)informatie in het systeem toe; dit kan worden gezien als een eenvoudige vorm van 'leren' waarmee het systeem zichzelf gedeeltelijk onderhoudt;
- een *kennisbank*: hierin wordt kennis over statistiek en kennis vereist voor uitleg door het systeem opgeslagen; het systeem zal kunnen uitleggen hoe een (deel-)oplossing is bereikt, waarom een bepaalde vraag wordt gesteld, en wat de betekenis is van gebruikte begrippen;
- een *modellenbank*: de (wiskundige) modellen die nodig zijn voor het ontwerpen van een bodeminventarisatieplan worden hierin opgeslagen, bijvoorbeeld de modellen voor evaluatie en optimalisatie;
- een *probleem-oplos-model*: de vaste structuur van het ontwerpproces, de volgorde waarin vragen gesteld moeten worden, en manier waarop de oplossingsruimte wordt ingeperkt is vastgelegd in een model;
- een *gebruikersinterface*: er is een conceptuele structuur van de opbouw van schermen gepresenteerd voor het doorlopen van alle onderscheiden taken.

De belangrijkste kenmerken van het kennissysteem zijn:

- de kennis over de vaste structuur van het ontwerpproces is opgenomen in het probleem-oplos-model (er is geen volledige scheiding van kennisbank en redeneermechanisme waarnaar bij kennissystemen vaak wordt gestreefd);
- de kennis in het systeem zal bestaan uit gegevens, regels en (wiskundige) modellen (respectievelijk opgeslagen in de gegevensbank, de kennisbank en de modellenbank), deze kunnen bijgewerkt worden zonder het probleem-oplos-model te wijzigen;
- er is een grote hoeveelheid kennis opgeslagen waarvan slechts een deel wordt gebruikt voor het ontwerpen van een plan voor een specifiek inventarisatieproject;
- interactie tussen het systeem en de gebruiker leidt tot de selectie van toepasbare kennis.

Het kennissysteem kan van belang zijn voor verschillende partijen die bij bodeminventarisaties zijn betrokken. Twee mogelijke gebruikersgroepen zijn onderzoekers bij instituten en universiteiten, en medewerkers van ingenieursbureaus.

Conclusies

Dit onderzoek heeft geresulteerd in een basis voor het ontwerp van een kennissysteem dat het ontwerpen van plannen voor bodeminventarisaties ondersteunt. Het kennissysteem vereist een gestructureerde aanpak van bodeminventarisaties, die al voor een groot deel kan worden gehanteerd voordat het systeem operationeel is. De gestructureerde aanpak maakt het mogelijk de resultaten van verschillende projecten beter te controleren en vergelijken. Hierdoor kan verzamelde kennis en informatie ook beter worden hergebruikt bij nieuwe projecten. Het kennissysteem zal ondersteuning bieden bij het opslaan van deze kennis en informatie. Een belangrijk onderdeel in het ontwerpproces met het kennissysteem is de mogelijkheid om plannen vooraf te evalueren en te optimaliseren. Met behulp van het systeem kan de kwaliteit van bodemonderzoek beter beheerst worden.

Glossary

accuracy constraint	minimum requirement on the accuracy of the survey results, e.g. defined in terms of the mean squared error of estimate
cluster sampling	sampling procedure in which groups or clusters of sampling elements are selected instead of individual elements
cost constraint	limited budget available
design-based approach (also: *probability sampling*)	the use of classical sampling theory
efficiency (statistical)	efficiency of a sampling design p can be defined as the ratio of the sampling variance of a reference design (often simple random sampling) to the sampling variance of p, at same sample size or at same cost
entity structure	hierarchy of entities and aspects used to structure systems; the aspects refer to a process or stage and the entities are characteristics of a particular aspect
estimator	method of estimation of the target quantity
evaluation a posteriori	evaluation of survey projects after execution of field work to ensure collection and storage of new knowledge in the system
how much request	survey concerning how much of a soil property is present, e.g. estimating a mean or an areal proportion
how much & where request	survey which besides an estimate of the soil property for the whole survey region aims at estimates for sub-regions
knowledge acquisition	process of extracting, structuring, and organizing knowledge from different sources, usually including human experts, so that it can be used in a computer program
knowledge-based system	program that operates on previously stored knowledge
method of determination	specification of how the values of the target variable are determined for given sampling elements

method of inference	combination of the method of estimation of the target quantity, i.e. the estimator, and the procedure to quantify the accuracy of the estimator from the sample data
model-based approach	the use of geostatistical techniques
outlinear plan of action	plan of action in which only some of the elements which need to be specified in the final plan are considered
outlinear survey scheme	(soil) survey scheme in which not all element which need to be specified in the final scheme are made explicit
pedological knowledge	knowledge about soil properties and soil survey
population	complete set of elements under study in a particular instance
prior information	information available from previous surveys
prior evaluation	predictions of both accuracy and cost of a soil survey scheme
probability sampling	see: *design-based approach*
purposive sampling	survey in which the sample points are deliberately selected (using prior information) and the results are not analysed statistically
sample	(i) selected set of (locations of) the elements to be observed (ii) a single observation element taken in the field
sample point	element to be sampled
sample size	number of elements in the sample
sample survey	survey using probability sampling
sampling design	mathematical function assigning a probability of selection to every possible sample
sampling element	(possible) object that is identifiable and that is element for the method of determination
sampling frame	list of all sampling elements in the population used to select elements to be sampled
sampling strategy	combination of a sampling design and an estimator

selection technique	operational method by which sampling elements are selected to be included in the sample, with predetermined probabilities according to the sampling design
simple random sampling	sampling procedure in which each possible sample which may result from the sampling design has an equal probability of being selected
soil survey scheme	scheme specifying which sites are to be sampled, which data are to be recorded, and how they are to be analysed statistically
spatial variability	variation of soil properties in space
stratified sampling	sampling procedure in which the survey region is divided into sub-regions or strata in each of which a number of sample points or sets of sample points is selected
survey region	geographical region to be surveyed
survey sampling	use of statistical sampling to collect data for a survey
systematic sample	sampling procedure in which the survey region is covered with a systematic pattern of sample points
target variable	soil property of interest
target quantity	quantity to be estimated or predicted from the sample survey data
two-stage sampling	sampling procedure in which sampling elements are selected in two stages: in the first stage primary units are selected in which in the second stage sample points or sets of sample points are selected
variogram	function specifying the relation between the vector $\boldsymbol{h}$ separating any two points in the area and their so-called semivariance
where request	survey concerning where specific soil properties are present, usually resulting in a map

List of abbreviations

AI	Artificial intelligence.
Aℓ	Aluminium.
DBMS	Database management system.
Dfl.	Dutch guilders
DP	Dynamic programming.
DSS	Decision support system.
EDS	Expert database system.
ES	Expert system.
Eq.	Equation.
Fe	Iron.
GIS	Geographical information system.
ha	Hectare(s).
KBS	Knowledge-based system.
KE	Knowledge engineering.
km	Kilometre(s).
km^2	Square kilometre(s).
LP	Linear programming.
MHW	Mean highest water table.
MLW	Mean lowest water table.
MS	Management science.
OR	Operations research.
P	Areic mass of phosphate sorbed by soil.
P_{max}	Maximum areic mass of phosphate sorbed by soil.
P_{rel}	Relative mass of phosphate sorbed by soil.

classes of sampling designs:

SI	Simple random sampling.
SIC	Simple random cluster sampling.
SY	Systematic sampling.
STSI	Stratified sampling with *SI* sampling in each stratum.
STSIC	Stratified sampling with *SIC* sampling in each stratum.
STSY	Stratified sampling with *SY* sampling in each stratum.
STSI, SI	Stratified two-stage sampling with *SI* sampling in both stages.
STSI, SIC	Stratified two-stage sampling with *SI* sampling in the first stage, and *SIC* sampling in the second stage.
STSI, SY	Stratified two-stage sampling with *SI* sampling in the first stage, and *SY* sampling in the second stage.
SI, SI	Two-stage sampling with *SI* sampling in both stages.
SI, SIC	Two-stage sampling with *SI* sampling in the first stage, and *SIC* sampling in the second stage.

SI, SY	Two-stage sampling with *SI* sampling in the first stage, and *SY* sampling in the second
SI, STSI	Two-stage sampling with *SI* sampling in the first stage, and *STSI* sampling in the second stage.
SI, STSIC	Two-stage sampling with *SI* sampling in the first stage, and *STSIC* sampling in the second stage.
SI, STSY	Two-stage sampling with *SI* sampling in the first stage, and *STSY* sampling in the second stage.

List of symbols

Symbol	Definition / Description	Units
A (A_h)	Area of the survey region (or stratum h).	km^2
A_u (A_{uh})	Area of primary unit u (in stratum h).	km^2
b	Available budget.	$
b_h	Available budget, or financial state, at stage h.	$
b_{max}	Maximal budget available.	$
C	Cost of spatial inventory.	$
c_e	Cost of equipment per hour.	$ / hour
C_e	Total cost of equipment.	$
c_h	Cost of spatial inventory in stratum h.	$
$c_h(n_h)$	Cost of spatial inventory in stratum h in case of n_h sample points.	$
c_l	Cost of laboratory analyses per sample point.	$
C_l	Total cost of laboratory analysis, including the cost of material used for the samples.	$
c_s	Survey cost per hour.	$ / hour
C_s	Total cost of field work, or survey cost.	$
E_p	Statistical expectation over repeated sampling under design p.	
E_ξ	Expectation over realizations from a stochastic model ξ.	
f	Sampling fraction n/N.	-
$g_h(n_h)$	Contribution from stratum h to the sampling-error prediction if n_h sample points are selected in stratum h.	
h	Stratum number or stage in optimization.	-
$\boldsymbol{h}$	Vector separating any two points in an area.	
$\lvert\boldsymbol{h}\rvert$	Distance between a pair of sample points.	km
k	Available number of sample points.	-
k_h	Available number of sample points, or capacity state, at stage h.	-
k_{max}	Maximal number of sample points.	-
L	Number of strata.	-
n	Sample size, i.e. the number of sample points.	-
N	Size of the population, i.e. total number of sampling elements.	-
n_h	Number of sample points in stratum h.	-
N_h	Total number of sampling elements in stratum h.	-
N_{uh}	Total number of sampling elements in unit u of stratum h.	-
n_0	Number of random points to be selected.	-
n_{01}	Number of random points to be selected per selected primary unit.	-
n_{01h}	Number of random points to be selected per selected primary unit in stratum h.	-

Symbol	Definition / Description	Units
n_{0h1}	Number of random points to be selected in stratum h in a selected primary unit.	-
n_1	Number of primary units in the sample.	-
N_1	Number of primary units in region A.	-
n_{1h}	Number of primary units to be selected in stratum h.	-
N_{1h}	Total number of primary units in stratum h.	-
n_{2h}	Number of sample points per selected primary unit.	-
n_3	Number of clusters in the sample.	-
n_{3h}	Number of clusters to be selected in stratum h.	-
n_{31h}	Number of clusters to be selected per selected primary unit in stratum h.	-
n_{3h1}	Number of clusters to be selected in stratum h in a selected primary unit.	-
n_4	Number of successive points per cluster (randomly selected starting point excluded).	-
n_{41}	Number of clusters to be selected in selected primary units.	-
n_{4h}	Number of successive points per cluster in stratum h (randomly selected starting point excluded).	-
$n_{\min h}$	Minimum number of sample points required in stratum h.	-
$n_{\min 1h}$	Minimum number of primary units to be selected in stratum h.	-
$n_{\min 01h}$	Minimum number of random points per selected primary unit in stratum h.	-
m_{1h}	Number of generated values of $\gamma_h(\boldsymbol{x}_1,\boldsymbol{x}_2)$ in the simulation to calculate $\hat{r}$.	-
m_{2h}	Number of generated values of γ_{Sh} in the simulation to calculate $\hat{r}$.	-
p	Sampling design.	-
p_h	Sampling design in stratum h.	-
r	Mean squared error of $\hat{z}$, i.e. the estimator of the spatial mean, due to sampling under design p.	
$\bar{r}$	Statistical expectation of r over realizations from the stochastic model ξ.	
$\hat{r}$	Estimator of the mean squared error of $\hat{z}$.	
r_h	Mean squared error of the estimated spatial mean in stratum h.	
$\bar{r}_h$	Statistical expectation of r_h over realizations from the stochastic model ξ.	
$\hat{r}_h$	Estimator of the mean squared error of the estimated spatial mean in stratum h.	
$s(\hat{r})$	Standard error of the simulation results to calculate $\hat{r}$ as estimate of $\bar{r}$.	
S^2	Population variance between elements.	
S_{bh}^2	Variance between primary units of stratum h.	
S_h^2	Variance among elements in stratum h.	
S_{wh}^2	Variance within primary units of stratum h.	

Symbol	Definition / Description	Units
S_{uh}^2	Variance among elements in primary unit *u* of stratum *h*.	
t_{aA}	Access time (i.e. time for access) within the survey region *A*.	hours
t_{ah}	Access time within stratum *h*.	hours
t_h	Total time needed for field work in stratum *h*.	hours
t_{oh}	Observation time in stratum *h*.	hours
$\bar{t}_{oh}$	Observation time per sample point in stratum *h*.	hours
t_{a1}	Access time to primary units in the survey region.	hours
t_{a0h} , t_{a0A}	Access time per kilometre to randomly selected points in stratum *h* or region *A* including location time, influence of the survey region, and time to ask permission.	hours / km
t_{a3h} , t_{a3A}	Access time between two successive points in a cluster in stratum *h* or region *A*.	hours
t_{a1h} , t_{a1A}	Access time per kilometre to selected primary units in stratum *h* or region *A*.	hours / km
t_{a2h} , t_{a2A}	Access time per kilometre between selected secondary units (random points) within a primary unit in stratum *h* or in region *A*.	hours / km
$V(\hat{z})$	Variance of the estimator of the spatial mean.	
$V_h(b_h)$	Contribution of stages (or strata) *h*+1, *h*+2, ... , *L* to the objective function if the system starts in state b_h at stage *h*, the immediate decision is n_{h+1}, and optimal decisions are made thereafter.	
$V_h(b_h,k_h)$	Contribution of stages (or strata) *h*+1, *h*+2, ... , *L* to the objective function if the system starts in state (b_h,k_h) at stage *h*, the immediate decision is n_{h+1}, and optimal decisions are made thereafter.	
w_h	Weight of stratum *h*, equal to its areal proportion in region *A*.	
$\boldsymbol{x}_i$	Randomly selected point or *i*th sample point within the region *A*.	-
$\bar{z}$	Spatial mean of a property *z* in a given region *A*.	
$\hat{z}$	Estimator of the spatial mean of a property *z* in a region *A*.	
$\bar{z}_h$	Mean value of sample points in stratum *h*.	
z_i	Value at the *i*th sample point.	
z_{iuh}	Value of the *i*th sample point in the *u*th primary unit of stratum *h*.	
$\bar{z}_{uh}$	Mean value of sample points in the *u*th primary unit of stratum *h*.	
$z(\boldsymbol{x}_i)$	Value of the property of interest at the location $\boldsymbol{x}_i$.	
γ	Semi-variance.	
$\bar{\gamma}_A$	Mean semi-variance between all pairs of points in *A*.	
$\bar{\gamma}_{Ah}$	Mean semi-variance between all pairs of elements in stratum *h* of region *A*.	
$\hat{\bar{\gamma}}_{Ah}$	Estimate of the mean semi-variance between all pairs of elements in stratum *h* of region *A*.	
$\gamma_h(\boldsymbol{x}_1, \boldsymbol{x}_2)$	Semi-variance between two randomly selected points in stratum *h*.	
Γ_s	*n* x *n* matrix of semi-variances between sample points.	

Symbol	Definition / Description	Units
$\bar{\gamma}_{S,A}$	n-vector of mean semi-variances between each sample point and all points in A.	
γ_{Sh}	Semi-variance between sample points in stratum h of region A.	
$\bar{\gamma}_{Sh}$	Mean semi-variance between sample points in stratum h of region A.	
$\hat{\gamma}_{Sh}$	Estimate of the mean semi-variance between sample points in stratum h of region A.	
Γ_{Sh}	n_h x n_h matrix of semi-variances between sample points in stratum h.	
$\bar{\gamma}_{uh}$	Mean semi-variance between all pairs of elements in unit u of stratum h.	
$\bar{\gamma}_{2h}$	Weighted mean of semi-variances between all pairs of elements belonging to the same primary unit.	
λ	n-vector of sample weights according to design p.	
λ_h	n_h-vector of sample weights according to the design applied in stratum h, summing to 1.	
λ_i	Weight at point $\boldsymbol{x}_i$ which depends on the probability of $\boldsymbol{x}_i$ being included in the sample, governed by the sampling design p; the λ_i (i = 1, 2, ..., n)sum to 1.	
Σ_h	Sum from h = 1 to L, where L is the number of strata.	
Σ_u	Sum from u = 1 to N_1(or N_{1h}), where N_1(or N_{1h}) is the number of primary units in A (or in stratum h).	

Subject Index

Curriculum vitae

Petronella Domburg werd geboren op 17 december 1965 te Marum. In 1984 behaalde zij het diploma Gymnasium B aan het Ubbo Emmius Lyceum te Stadskanaal. Daarna begon zij met haar studie aan de Landbouwuniversiteit in Wageningen. In 1990 studeerde ze af in de studierichting Cultuurtechniek met een afstudeervak Recreatiekunde en een afstudeervak Planologie. Van maart 1990 tot juli 1994 werkte zij bij het DLO-Staring Centrum in Wageningen, gedetacheerd vanuit de Vakgroep Informatica van de Landbouwuniversiteit, aan het onderzoek dat in dit proefschrift is beschreven.