A framework of concepts for soil survey using probability sampling

P. Domburg
J.J. de Gruijter

Report 55

DLO The Winand Staring Centre, Wageningen (The Netherlands), 1992

ABSTRACT

Domburg, P. and J.J. de Gruijter, 1992. *A framework of concepts for soil survey using probability sampling.* Wageningen (The Netherlands), DLO The Winand Staring Centre. Report 55. 32 pp. ; 4 Figs; 16 Refs.

Available soil survey information is mainly qualitative. However, there is a growing need for information with quantified accuracy. Therefore statistical methodology should be applied to collect and analyse data. This study concentrates on soil survey using probability sampling. Before data are collected a soil survey scheme should be designed. In order to develop a decision support system for this design process, we have started to develop a framework of concepts used during this process. An unambiguous framework can facilitate communication between parties involved in a soil survey. It is developed by applying two types of knowledge acquisition: studying the literature and interviewing a statistician with experience in soil survey.

Keywords: soil survey using probability sampling, sample survey, knowledge acquisition, framework of concepts, soil survey scheme, decision support system

ISSN 0927-4537

DLO The WINAND STARING CENTRE is continuing the research of: Institute for Land and Water Management Research (ICW), Institute for Pesticide Research, Environment Division (IOB), Dorschkamp Research Institute for Forestry and Landscape Planning, Division of Landscape Planning (LB), and Soil Survey Institute (STIBOKA).

Project 436

CONTENTS

page

PREFACE 7

SUMMARY 9

1 INTRODUCTION 11
1.1 Objectives 11
1.2 Requests for soil survey on a statistical basis 12
1.3 Probability sampling 14
1.4 Outline 16

2 PROCEDURE USED TO DEVELOP THE FRAMEWORK 17
2.1 Overview 17
2.2 Knowledge acquisition 17
2.3 Two cases of soil survey using probability sampling 18

3 FRAMEWORK OF CONCEPTS 21
3.1 Overview 21
3.2 Framework 21
3.3 Specification of concepts 22

4 FINAL REMARKS 29

REFERENCES 30

APPENDIX
Terminology: English - Dutch 32

FIGURES
1 Relations between types of results and types of requests. 13
2 Emphasis of design-based and model-based approach on different types of survey requests. 14
3 Factors governing the construction of a soil survey scheme. 21
4 Structure of a soil survey scheme. 21

PREFACE

The framework of concepts for soil survey using probability sampling, described in this report, has been developed within the scope of a project on the "Development of a Decision Support System for spatial inventory of soil characteristics" (from March 1990 to March 1994), which is being funded by the Netherlands Integrated Soil Research Programme.

The initiative for developing the conceptual framework and writing this report was taken by P. Domburg, who is working on the research project mentioned above as a Ph.D. student in the Department of Computer Science at Wageningen Agricultural University, and seconded to the Department of Survey Methods of DLO The Winand Staring Centre under the supervision of J.J. de Gruijter.

The authors would like to thank B.J.A. van der Pouw, D.J. Brus (both employed by DLO The Winand Staring Centre) and S.J.M. de Ree (the Netherlands Central Bureau of Statistics: CBS), for suggesting constructive changes in the way some explanations are presented.

SUMMARY

There is a growing need for soil survey information with quantified accuracy, whereas the available pedological information, i.e. information on properties of soil in situ, is mostly qualitative. Therefore, new requests for pedological information often require (additional) data to be collected in the field. Statistical methodology should be applied to collect and analyse data in order to be able to quantify the accuracy of the survey results. Before data collection starts a soil survey scheme needs to be designed specifying which data are to be collected, how they are to be collected and how they are to be analysed statistically. This design requires the effective use of pedological and statistical knowledge.

This study is part of a project to develop a decision support system for designing schemes for soil survey on a statistical basis. A framework of concepts which are specified unambiguously is a prerequisite for developing such a system. The aim of this study is to develop a conceptual framework for soil survey using probability sampling: to define the essential concepts and list them based on their relationships within the context of the design process. Examining the use of statistical methodology, this study focuses on the use of the classical sampling theory (section 1.3), i.e. soil survey using probability sampling.

Knowledge acquisition as used in the field of computer science known as knowledge technology, 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. For this study knowledge was collected by studying the literature on statistics and soil survey and by interviewing a statistician with experience in designing soil survey schemes. Two historic cases of soil survey using probability sampling are described to clarify the meaning of the conceptual framework in practice.

The framework consists on the one hand of three main factors from where the design of a survey scheme starts: aim, constraints and prior information. On the other hand the framework contains the structure of a soil survey scheme with four main factors: plan of action, method of inference, sample and prior evaluation. At present, not all elements of the framework are always documented in detail, but they are all important during the design process.

The conceptual framework has been successfully applied to describe 23 historic cases of soil surveys. Although the framework concentrates on using the classical sampling theory, some of it may be applicable to soil survey in general or to other spatial sample surveys. The framework may facilitate communication between the parties involved in a soil survey and is a first step towards a decision support system for designing soil survey schemes.

1 INTRODUCTION

1.1 Objectives

In recent decades there has been a growing need for soil survey information with quantified accuracy; researchers and those commissioning the projects are not only interested in information about soil, but also in the accuracy of this information. Answers to requests for pedological information, i.e. information on properties of soil in situ, cannot always be based on available data stored in soil maps, reports, or databases. The amount of relevant information is often too limited, and uncertainty or error due to spatial variability of soil properties is often largely unquantified. Decisions on land use, for example nature conservation, agricultural use or residential use, often largely depend on pedological information. Given a certain aim, one endeavoures to make a satisfactory decision. If more accurate information on soil properties is available, this decision can be improved, and the risk of taking a decision which is less appropriate can be reduced.

Pedological information with quantified accuracy is for example relevant to requests for environmental protection purposes, to requests for land evaluation and to studies on soil physics and soil chemistry. The need for pedological information with quantified accuracy can be fulfilled by applying appropriate statistical methodology. The use of statistical methodology also enables efficient soil survey schemes to be developed. The efficiency of survey schemes is determined by two factors: accuracy of the results and operational costs. Both these factors are related to the statistical methodology used.

The available pedological information is often inadequate when dealing with a new request for information, so that additional data frequently need to be collected in the field. If the survey is to be done on a statistical basis then, before the field work can start, a scheme needs to be developed specifying which data are to be collected, how they are to be collected, and how they are to be analysed statistically. It is essential to make effective use of pedological and statistical knowledge when designing such a scheme, in order to exploit existing resources adequately and to reduce risks of producing inappropriate information.

This study is part of a project to develop a decision support system for designing schemes for soil surveys on a statistical basis. Support of this design process should enable and facilitate proper use of existing pedological and statistical knowledge. The aim of the project is to integrate such knowledge into a coherent computer system, intended to support decisions in the process of designing soil survey schemes. A framework of concepts which are specified explicitly and unambiguously is a principal requisite for developing such a system (see for example: Waterman, 1986). As far as we know, such a framework for soil survey does not exist, although with respect to the statistical part of the framework, descriptions of concepts as used in general can be found in the literature on statistics (e.g. Cochran, 1977; Krishnaiah

& Rao, 1988). In pedology a variety of concepts is used, but these concepts are not always unambiguously defined and consistently used.

Therefore, the objective of the study reported here is to define those concepts which are essential to the design process of soil survey schemes, and list such concepts on the basis of their relationships within the context of this design process. This objective induces the following two research issues:

1. to develop a framework: to describe the concepts of the design process of soil survey schemes, and the relationships that exists between these concepts;
2. to define the concepts explicitly and unambiguously.

This study concentrates on concepts related to a specific type of soil survey: soil survey using probability sampling. This means that sampling theory is used. However, some of the terminology is also relevant to other types of soil survey.
The domain of interest in this study is outlined in the following two sections. In section 1.2 different types of requests for soil survey are delineated, and in section 1.3 the role of probability sampling in soil survey is discussed.

1.2 Requests for soil survey on a statistical basis

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*, used to produce multi-purpose soil maps in the Netherlands and elsewhere, 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 type by sampling at some sites (Steur, 1961). Their descriptions mainly contain qualitative information and only limited information on the variability of soil properties and on the accuracy of the results of their survey. In order to acquire information on these topics, data should be collected and analysed on a statistical basis. Three categories of requests for soil survey on a statistical basis, which influence the choice of the sampling design in different ways (see below), can be distinguished. Distinctions between these categories are related to the type of result required (Figure 1).

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 for statistical parameters, such as mean, variance, or areal proportion, for a given soil property. This is soil *inventory* in the strictest sense; it may be considered as a special type of soil survey. 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 of a single property that is of interest, the result of the study is a single value for the whole survey region indicating the areal proportion, accompanied by its quantified accuracy.

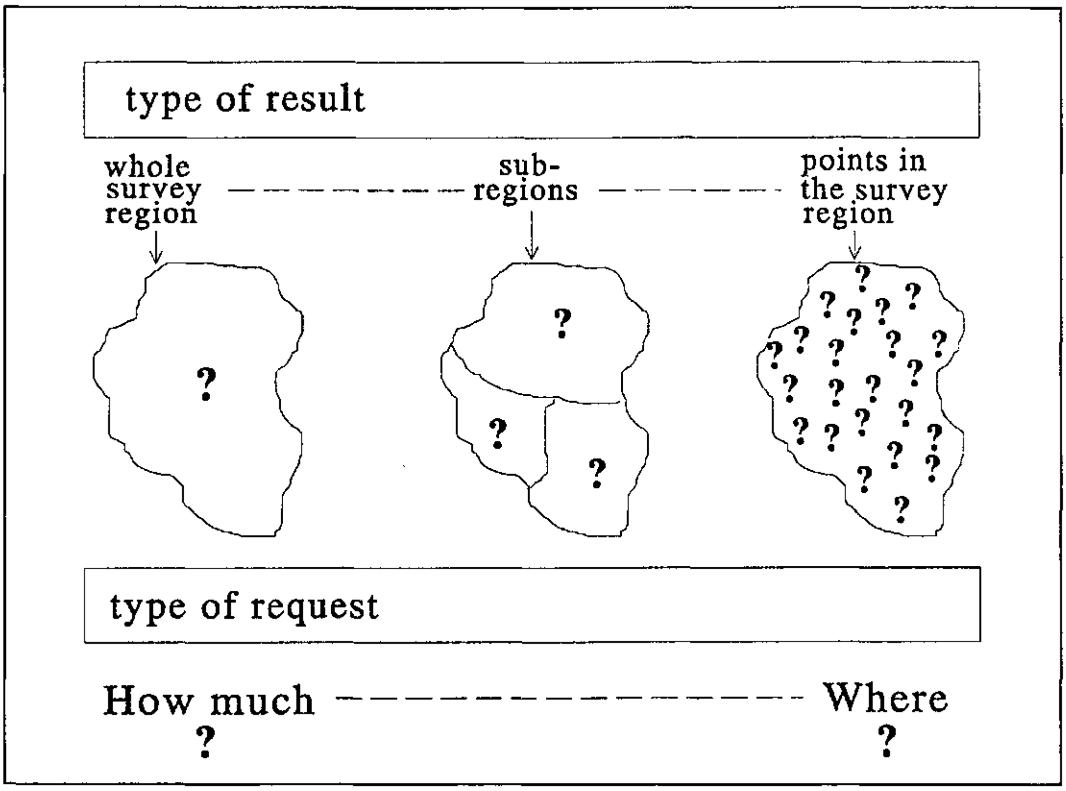

Figure 1. Relations between types of results and types of requests.

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 'organic matter content of the topsoil', or '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.

The relative importance of *how much* and *where* may influence the way in which data should be collected and analysed, i.e. it can suggest which statistical methodology seems most appropriate for the design of a soil survey scheme (see Figure 2). However, it is impossible to divide soil survey requests neatly into *how much* and *where* requests. The distinction is more like a continuum with two extremes. In many cases the emphasis is on one of these aspects, and then one aspect mainly influences the choice of the sampling design.

Requests for "how much" and "where"
Between these two extreme categories exists a third group of requests that deal equally with both *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 design.

1.3 Probability sampling

As it is unfeasible to observe all distinguishable elements in a survey region, because it costs too much time and money, a sample must be taken. Two advantages of taking a sample are: reducing costs, and increasing speed of survey.

Examining the use of statistical methodology for soil survey, two approaches can be distinguished: the use of the classical sampling theory (design-based approach) and the use of geostatistical techniques (model-based approach) (Särndal, 1978; De Gruijter & Ter Braak, 1991). 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. This is roughly schematized in Figure 2.

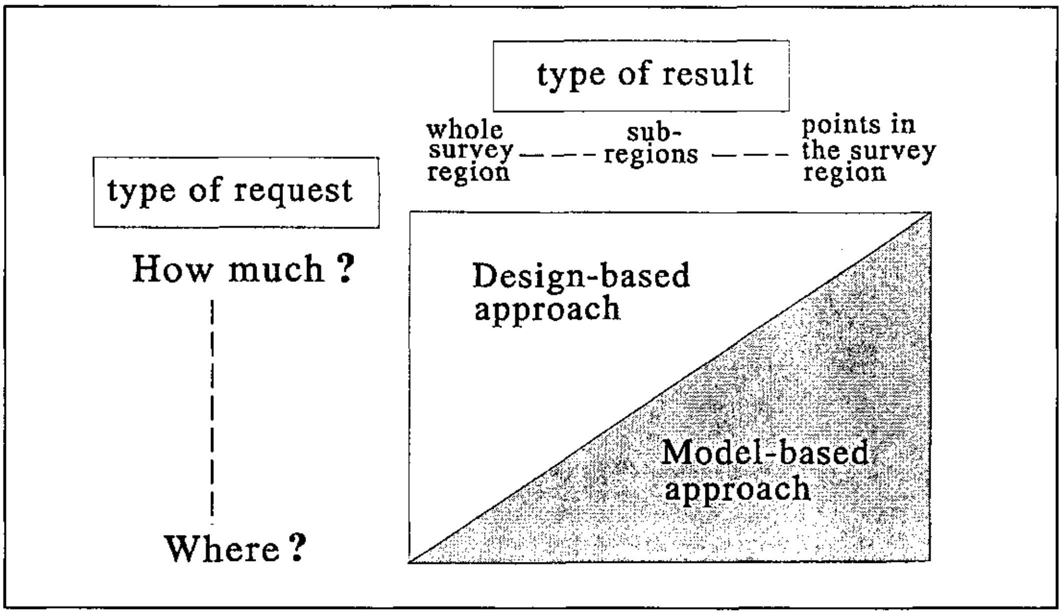

Figure 2. Emphasis of design-based and model-based approach on different types of survey requests.

The classical sampling theory has been used in soil survey for many years. During the last decades the use of geostatistical techniques has increased and knowledge on the usefulness of these techniques is expanding (see for example: Journel & Huijbregts, 1978; Webster & Oliver, 1990). In this study we are concentrating on a design-based approach to soil survey. Although literature on the classical sampling theory focuses on the *how much* type of surveys, this approach is often also

applicable for requests on both *how much* and *where*. A framework of concepts for the model-based approach may partly differ from the one described here.
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 the population may consist of the complete set of possible locations for observation in the survey region. Values of the soil properties at all locations are considered to be unknown but fixed, i.e. not random. A subset of elements of the population is selected for observation and the probability for any subset to be selected is determined by the sampling design, which also determines whether the observations are mutually independent. Estimation of parameters is 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:

1. 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;
2. each possible sample S_i has assigned to it a known probability of selection π_i;
3. one of the S_is is selected by a random process in which each S_i receives its appropriate probability π_i of being selected;
4. 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* (see for example: Cochran, 1977: p. 2-4; Krishnaiah & Rao, 1988: p. 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 set of (locations of) the elements to be observed (see section 3.3). We will also use the term *sample survey* throughout the rest of this report. The use of statistical sampling to collect data for survey is called *survey sampling* (see for example: Krishnaiah & Rao, 1988: p. 16; Cassel et al., 1977: p. 34).

An important distinctive property of the model-based approach compared with the design-based approach is that the sampling elements 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: observations made close to each other are more similar than observations made further apart. In the model-based approach data are therefore often collected at a fixed regular grid, while randomness and independence of observation points are the main characteristics of samples from the design-based approach.

1.4 Outline

In Chapter 2 we will explain how the conceptual framework was developed. Chapter 3 presents the description of the framework and definitions of the concepts. Finally, Chapter 4 explains the advantages and applicability of the framework.
The main terms used in this report and their equivalents in Dutch are summarized in the Appendix.

2 PROCEDURE USED TO DEVELOP THE FRAMEWORK

2.1 Overview

The framework of concepts is based on the literature on statistics concerning the sampling and on practical experience of soil survey. Since, to our knowledge, a clear framework and explicit definitions of the concepts, except for general descriptions of statistical concepts, are not available in the literature, knowledge acquisition has been used to construct the framework (see section 2.2). In section 2.3 two cases are introduced that will serve to illustrate the meaning of the concepts in Chapter 3. These cases are derived from descriptions of historic cases of sample surveys executed at the Winand Staring Centre.

2.2 Knowledge acquisition

Two domains of knowledge relevant in the design of soil survey schemes are: statistical and pedological knowledge. Furthermore, knowledge of constructing soil survey schemes (i.e. knowledge of the *design process*) utilizing statistical and pedological knowledge, is also important. As far as we know, no description of this design process exists; the knowledge of designing soil survey schemes has not been formalized until now. To gain insight into this type of knowledge, a technique of knowledge acquisition has been used.

The term *knowledge acquisition* as used in the specific field of computer science known as knowledge technology, 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, 1985: p. 392). In this project knowledge was collected from the literature, and interview techniques were used to extract knowledge from an expert. The two sources from which knowledge was acquired for the development of the conceptual framework of Chapter 3 is described below.

Literature
General knowledge on the classical sampling theory and definitions of statistical terms are derived from the books of Cassel et al. (1977), Cochran (1977), and Krishnaiah & Rao (1988). Slight differences in terminology exists between these handbooks; sometimes no crisp definition is given or a particular specification is indicated as 'recommended use'. Sometimes we had to chose from alternative definitions or adjust a definition to our framework. Literature on the application of statistical methodology 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 conceptual framework for constructing soil survey schemes.

Interviewing an expert

The development of a soil survey scheme can be seen 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. The statistician conducts the process by helping the researchers to make their aim more explicit and tries to recover all relevant information. He can be considered as an expert in designing soil survey schemes. In order to describe the design process and to discover relevant concepts, a statistician with experience in designing soil survey schemes (an expert) was interviewed for about ten sessions. During these sessions specific questions were asked about the task of designing a soil survey scheme in order to identify key concepts needed to find a solution for any given case, and to identify the path to a survey scheme. The 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 was structured and organized from the interviews and debated with the expert. Through this interaction the knowledge could be formalized, i.e. crisp descriptions of the concepts, the relations between concepts, and the structure of the design process could be made.

After the ten exploratory interviews on the process, we described 23 historic cases of soil survey using probability sampling done at DLO The Winand Staring Centre. These descriptions were used to check and adjust the framework and to evaluate its applicability in practice. In addition we will use the descriptions of the historic cases to further analyse the design process and to describe the domain knowledge from pedology and statistics.

2.3 Two cases of soil survey using probability sampling

Here we discuss the backgrounds and aims of two cases exemplifying soil survey using probability sampling. More details of these cases are given in the following chapter. To avoid irrelevant complications and to fit into the limited domain of the project at large, the cases as presented in Chapter 3 are somewhat schematic versions of the original soil surveys. The original context of the cases, implemented by DLO The Winand Staring Centre, is outlined below.

Case A: Phosphate saturation in the Ootmarsum region

This case is part of a research project commissioned by the Province of Overijssel (the Netherlands), and aims at quantifying the phosphate saturation of cultivated soils in two regions in the province. The purpose of the project was to quantify the phosphate saturation in a region representative of eastern sandy regions of the Netherlands, given a particular definition of saturation. On a higher level, results of this study were used to support decisions on the control of ground and surface water pollution.

In many rural regions in the Netherlands there is a large production of animal manure. If too much manure is applied the soil becomes saturated with phosphate from the manure. Phosphate then leaches and pollutes ground and surface water. The phosphate sorption capacity varies between soils. It is by definition (e.g. Schoumans et al., 1988: p. 201) 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). The degree of phosphate saturation is calculated by dividing the actual phosphate content by the phosphate sorption capacity, both summed over depth to the MHW. If the phosphate sorption capacity and the degree of phosphate saturation are known, this information can be used to support decisions on the control of ground and surface water pollution.

Two regions in Overijssel with contrasting pedological and land use characteristics were selected to be surveyed. One region, the Ootmarsum region, has partly a high elevation and deep groundwater tables. Roughly thirty percent of this region is forest with functions for nature conservation. Many valleys cross the region. The other region, Bentelo-Beckum, has shallower groundwater tables and is intensively used for agriculture. Both regions consist largely of sandy soils, but in the Ootmarsum region there are some clay soils. The spatial inventory was in both regions confined to the parts used for agriculture.

Case A includes only the Ootmarsum region, because in the first instance we prefer to limit ourselves to sampling a single population. The survey region includes 2252 ha of agricultural land near the village of Ootmarsum. There are some specific features that must be taken into account while designing a scheme for a soil survey in this region for the objectives mentioned above. One feature, related to this particular survey region, is the presence of dry and wet sub-regions, the latter of which being relatively small with regard to the whole survey region. Wet regions are more sensitive to phosphate leaching than dryer ones and therefore accurate information is especially required about the former. 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 the phosphate sorption capacity and actual phosphate concentration on the other. Both these features had an important impact on the design of the survey scheme, apart from the usual constraints concerning the available budget and required accuracy. A report on this research project has been written by Hack-ten Broeke et al. (1990).

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, sample 1" implemented by DLO The Winand Staring Centre. The purpose of the project is to upgrade the national soil map, scale 1:50 000, by collecting detailed quantitative information on the spatial variability of soil properties within the map units. The mentioned map is a multi-purpose soil map. The 62 map sheets, mainly subdivided into West and East, were produced by the*free survey* method (see section 1.2). This production was started about 30 years ago and is now nearly finished. The map sheets have extensive legends and memoirs which contain mainly qualitative information and only limited quantitative information on the spatial variability of soil properties; for example Damoiseaux et al. (1990), and Vleeshouwer & Damoiseaux (1990). The project "National Sampling Map Units" aims to satisfy the growing need for soil information with quantified accuracy by upgrading the existing national soil map.

The first sample of this project relates to map unit Hn21-VI (*Veldpodzolgronden* on groundwater class VI); see Visschers et al. (in preparation).

In the original study, data on all soil properties generally relevant in sandy regions were collected, whereas case B is limited to collecting of information on the MHW. We are focusing on inventory studies in which one soil property is of main interest so that only this single property is to be considered when designing the survey scheme. We selected only the MHW as a property because it is highly relevant to many other research projects, particularly to 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 travel costs. Besides this constraint, which is related to both logistics and financial aspects, the available budget for the project is limited and some minimum accuracy of the results is required. These constraints all affect decisions in the design of the survey scheme.

3 FRAMEWORK OF CONCEPTS

3.1 Overview

The framework of concepts is introduced in section 3.2, outlining the concepts and their interrelationships. In section 3.3 the concepts are defined and illustrated with the two example cases of section 2.3.

3.2 Framework

A framework of the factors governing the construction of a soil survey scheme (Figure 3), and of the elements that make up a survey scheme (Figure 4) is a necessary tool to describe and analyse the process of designing survey schemes. The concepts in the framework bear a slight resemblance to the principle steps in a sample survey as described by Cochran (1977: p. 4-8), who writes about sampling techniques in general. As we focus on the application of statistical methodology for soil survey, we need some additional concepts. The order of the concepts is based on mutual correlations, and on the order in which they appear during the design of a sample survey.

AIM
- target quantity
- target variable
- survey region

CONSTRAINTS
- accuracy
- cost
- logistics

PRIOR INFORMATION
- spatial variability
- other geographical information

Figure 3. Factors governing the construction of a soil survey scheme.

SOIL SURVEY SCHEME

PLAN OF ACTION
- sampling element
- population
- method of determination
- sampling design
- sampling frame
- selection technique
- instructions for field work

METHOD OF INFERENCE
- method of estimation
- procedure to quantify the accuracy

SAMPLE

PRIOR EVALUATION
- prediction of the accuracy
- prediction of the cost

Figure 4. Structure of a soil survey scheme.

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 phosphate saturation in a region representative of eastern sandy regions of the Netherlands, in order to support the control of ground and surface water pollution. The aim of the survey is more specifically to determine the areal percentage of the Ootmarsum region, 2252 ha of agricultural land, where the soil should be considered as being saturated with phosphate, given a particular definition of saturation.

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) surveys, for example maps with legends, reports and databases.

The design process starts from the available statistical and pedological knowledge, and from the specifications of the aim, constraints and prior information. The final result of this process is a scheme specifying:

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

3.3 Specification of concepts

The concepts we distinguish are clarified below. Definitions of the concepts are given with examples from the cases introduced in section 2.3. Since no structured approach to designing and describing soil survey schemes is prescribed, there are large distinctions between the available information on historic studies. Some concepts could not be recovered retrospectively due to limited documentation. Descriptions of constraints and prior information used are hardly ever lacking. Information on elements of the plan of action, such as the selection techniques and instructions for field work are also rarely reported. Furthermore, a prior evaluation of a scheme for the survey is never described.

The **aim** of a survey consists basically of the following three elements.

1. *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 report (see section 1.3), the target quantity is a *parameter*.

Case A: *proportion (of the region where the soil should be considered as being saturated with phosphate).*
Case B: *mean (spatial mean of the Mean Highest Water table in map unit Hn21-VI).*

2. *Target variable*
Target variables are soil properties (e.g. highest groundwater, clay content, moisture supply capacity) of which a target quantity is to be determined by the survey. Although their values may be measured, 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 the degree of phosphate saturation, defined as the actual phosphate concentration divided by the phosphate sorption capacity, both averaged over depth to the Mean Highest Water table, exceeds 0.25.*
Case B: *depth of Mean Highest Water table (MHW) in cm.*

3. *Survey region*
The survey region is the geographical region to be surveyed. The boundaries and location of the region are important here. The survey region may be a three dimensional body, but also a plane or a line element and, apart from that, 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 of 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.

1. *Accuracy*
There are two issues to consider with respect to accuracy of the survey result. 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 or may not be required for the accuracy of the result to 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.
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 are relatively small with regard to the whole survey region. Accurate information is especially required on the wet regions, because these are more sensitive to phosphate leaching than dryer regions. The accuracy of the result must be quantifiable from the sample data.*
Case B: *strive for maximum accuracy given the available budget.*

2. *Cost*
   The available budget is almost always limited. The cost of a spatial inventory is mainly determined by the sampling design and method of determination. A limited budget influences the sampling design, and the choice of the method of determination.

3. *Logistics*
   A third category of constraints are those of a logistical nature. Restricted capacity of a laboratory, or a limited period in which the field work may be done, are examples of this category. Such constraints may 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 amount of possible alternatives.

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

1. *Spatial variability*
   Information on the spatial variability of soil properties, i.e. the way in which the properties vary in space, can support the design of an efficient soil survey scheme. Prior information on spatial variability is required in order to predict the accuracy for a given soil survey scheme. If available, information on the target variable in the survey region should be used. Otherwise information on a co-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 target variable (degree of phosphate saturation at points) in a comparable survey region may be useful. For case A information and experience from a comparable study 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 phosphate sorption capacity 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.*

2. *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 geographical information systems, data on vegetation, geomorphology, etc. This information might be used in setting up a sampling design. 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. If information from executed sample surveys is stored, it could be useful in designing schemes for future surveys.

   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 chosen sampling design), the method of inference, a specification of the selected sample, and a prior evaluation of the scheme. Before field work starts all these elements need to be specified.

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

1. *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 sample survey. Examples of sampling elements in a soil survey are: a particular soil pit, an auguring, or a soil sample.

   Case A: *standard auguring to MHW, with a maximum of 1 m.*
   Case B: *standard auguring to Mean Lowest Water table (MLW), with a minimum of 1.5 m.*

2. *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 *non-soil.*

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 auguring locations identifiable as soil in the survey region of this case.*
Case B: *similar to case A.*

3. *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: *the degree of phosphate saturation at sample points is related to the content of oxalate-extractable Fe and Al, the density of the soil, and the depth to MHW. The content of oxalate-extractable Fe and Al is 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 auguring. These values should be compared with those at reference tubes (i.e. with known values of MHW) measured on the same day. Values of MHW in the survey region can be derived from this comparison. A regression model should be used to calculate the degree of phosphate saturation.*
Case B: *values of 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 at reference tubes (with known values of MHW) in the neighbourhood; both measured at the instant of MHW-level in the reference tubes. 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 (see for example: Van der Sluijs & De Gruijter, 1985).*

4. *Sampling design*
The sampling design is a mathematical function determining the probability of inclusion in the sample for every possible subset of sampling elements. The sample size is the number of sampling elements in the sample. If this is fixed and pre-determined then it is implied by the sampling design, as any subset of a different size will be assigned probability zero making up the sample. Different types of sampling designs can be distinguished, for example simple random sampling, stratified sampling, cluster sampling (see: Cochran, 1977). All these types can be subdivided into more specific designs. Each design has characteristics of its own, for example concerning its usability under specific conditions, or its applicability to answer a particular request.

Case A: *stratified random sampling.*
Strata: seven combinations of map units of 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 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.*
Strata: map sheets. Design within strata: two-stage. First stage: random selection of two map delineations with replacement and probabilities proportional to size (i.e. area of delineations). Second stage: four points by simple random sampling and equal probability. Total sample size: 264.

5. *Sampling frame*
A list of all sampling elements in the population used to select elements to be observed is referred to as a sampling frame (Krishnaiah & Rao, 1988: p. 21). 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 geographical information system.
The sampling frame should correspond as well as possible with 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.
A specific sampling design is related to requirements on the sampling frame. Efforts should be made to find or construct a sampling frame which fits the design. Sometimes a design requires more than one frame, for example a two-stage design may require different frames for selections in the first and second stages.

Case A: *an overlay of soil map, scale 1:50 000 and land use map was used to select the sampling elements. 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.*

6. *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 strata, and the coordinates 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.

7. *Instructions for field work*
   As stated under point 5 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 an observation element is inaccessible (e.g. because of crops). The ways to register these elements (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 and the procedure to quantify the accuracy of the estimator from the sample data (Krishnaiah & Rao, 1988: p. 247). Sometimes an estimate can be improved by means of an auxiliary variable (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 formulas for stratified random sampling. The proportion (i.e. the target quantity) of the region where the soil should be considered as being phosphate saturated can be estimated by the information on the phosphate saturation at sample points (see the method of determination). Calculating the proportion is comparable with calculating a spatial mean.*
Case B: *standard formulas for stratified two-stage sampling.*

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

The **prior evaluation** shows the predictions of both accuracy and cost of the scheme proposed. A prior evaluation of the accuracy is based on the sampling design, the method of determination, the method of estimation, and the prior information on spatial variability. The cost of a scheme can be predicted from the sampling design, and the method of determination. It is worthwhile evaluating a scheme before field work starts, so that it is possible to check whether the researchers or those commissioning the project agree with the scheme. If they disagree with the predicted accuracy or with the cost, the plan can be adapted beforehand, for example by revising the original constraints.
In the present situation only limited attention is paid to evaluating soil survey schemes a priori. The constraints are always taken into account during the design of a scheme and a finally proposed scheme is assumed to fulfil these constraints. The time and means to compare alternative schemes are generally lacking. When models of cost and accuracy are available a better comparison of alternative schemes is possible.

4 FINAL REMARKS

In this report a framework for soil survey using probability sampling is described and specified. This conceptual framework has been successfully applied to describe 23 historic cases of soil surveys. These cases are referred to as *historic* because they were executed before this framework was developed.
Although the framework concentrates on soil survey using probability sampling, i.e. using the classical sampling theory, part of it may be applicable to soil survey in general or to other spatial sample surveys. Other methods of soil survey will probably need additional concepts for their formal description.

The framework we have developed may also facilitate negotiations concerning aims and conditions of soil surveys. The absence of clear concepts may cause ambiguity and confusion among researchers, or between researchers and decision-makers (e.g. policy-makers). The use of unambiguously defined concepts may support effective communication between all parties involved in a soil survey.

If this framework is used to report on sample surveys in soil, experience gained from historic studies could be better utilized in the future.

Apart from the advantages of the framework mentioned above, we intend to use it to develop a decision support system for the design of schemes for soil survey on a statistical basis. The availability of a clear framework is one of the principal requirements for the development of such a system.

REFERENCES

BREEUWSMA, A., J.G.A. REIJERINK, O.F. SCHOUMANS, D.J. BRUS & H. VAN HET LOO, 1989. *Fosfaatbelasting van bodem, grond- en oppervlaktewater in het stroomgebied van de Schuitenbeek.* Wageningen, DLO The Winand Staring Centre, rapport 10.

CASSEL, C-M., C-E. SÄRNDAL & J.H. WRETMAN, 1977. *Foundations of Inference in Survey Sampling.* New York, John Wiley & Sons, Inc.

COCHRAN, W.G., 1977. *Sampling Techniques*, third edition. New York, John Wiley & Sons, Inc.

DAMOISEAUX, J.H., P. HARBERS EN T.C. TEUNISSEN VAN MANEN, 1990. *Bodemkaart van Nederland 1:50.000, 61-62 West en Oost, Maastricht-Heerlen.* Wageningen, DLO The Winand Staring Centre.

GRUIJTER, J.J. DE & C.J.F. TER BRAAK, 1990. Model-free estimation from spatial samples: a reappraisal of classical sampling theory. *Mathematical Geology*, 22/4, pp. 407-415.

HACK-TEN BROEKE, M.J.D., H. KLEIJER, A. BREEUWSMA, J.G.A. REIJERINK & D.J. BRUS, 1990. *Fosfaatverzadiging van de bodem in twee gebieden in Overijssel.* Wageningen, DLO The Winand Staring Centre, rapport 108.

JOURNEL, A.G. & CH.J. HUIJBREGTS, 1978. *Mining Geostatistics.* London, Academic Press Inc.

KRISHNAIAH, P.R. & C.R. RAO, 1988. *Handbook of Statistics, volume 6: Sampling.* Amsterdam, Elsevier Science Publishers b.v.

SÄRNDAL, C-E., 1978. Design-based and model-based inference in survey sampling. *Scand. J. Statist.*, no. 5, pp. 27-52.

SCHOUMANS, O.F., B.A. MARSMAN & A. BREEUWSMA, 1988. Assessment of representative soil data for phosphate leaching. In: *Land qualities in space and time*, J. Bouma & A.K. Bregt (eds.), pp. 201-204.

SLUIJS, P. VAN DER & J.J. DE GRUIJTER, 1985. Water table classes: a method to describe seasonal fluctuation and duration of water tables on Dutch soil maps. *Agricultural Water Management*, volume 10, pp. 109-125.

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.

VISSCHERS, R., 1992. *Upgrading van de bodemkaart van Nederland door steekproeven in kaarteenheden.* Wageningen, DLO The Winand Staring Centre, (in preparation).

VLEESHOUWER, J.J. & J.H. DAMOISEAUX, 1990. *Bodemkaart van Nederland 1:50.000, toelichting bij kaartblad 61-62 West en Oost, Maastricht-Heerlen.* Wageningen, DLO The Winand Staring Centre.

WATERMAN, D.A., 1986. *A guide to expert systems.* Reading, Massachusetts, Addison-Wesley Publishing Company.

WEBSTER, R., & M.A. OLIVER, 1990. *Statistical Methods in Soil and Land Resource Survey.* New York, Oxford University Press.

APPENDIX

Terminology: English - Dutch

accuracy	*nauwkeurigheid*
aim	*doel*
auxiliary variable	*hulpvariabele*
constraint	*randvoorwaarde*
co-variable	*covariabele*
estimation	*schatting*
free survey	*vrije kartering*
geographical information	*geografische informatie*
logistics	*logistiek*
method of determination	*bepalingsmethode*
method of estimation	*schattingsmethode*
method of inference	*statistische verwerkingsmethode*
pedology	*bodemkunde*
plan of action	*werkplan*
population	*populatie*
prior evaluation	*evaluatie vooraf*
prior information	*voorinformatie*
probability sampling	*het nemen van een kanssteekproef*
random	*aselect*
sample	*1. steekproef; 2. (grond-)monster*
sample size	*steekproefomvang*
sample survey	*steekproefsgewijze inventarisatie*
sampling design	*steekproefopzet*
sampling element	*steekproefelement*
sampling frame	*steekproefkader*
selection technique	*selectie-techniek*
soil inventory	*bodeminventarisatie (strikte betekenis)*
soil survey	*bodeminventarisatie (in ruime zin)*
soil survey scheme	*bodeminventarisatieplan*
stratum	*stratum*
sub-region	*deelgebied*
survey region	*onderzoeksgebied*
survey sampling	*steekproefname ten behoeve van een inventarisatie*
target variable	*doelvariabele*
target quantity	*doelgrootheid*