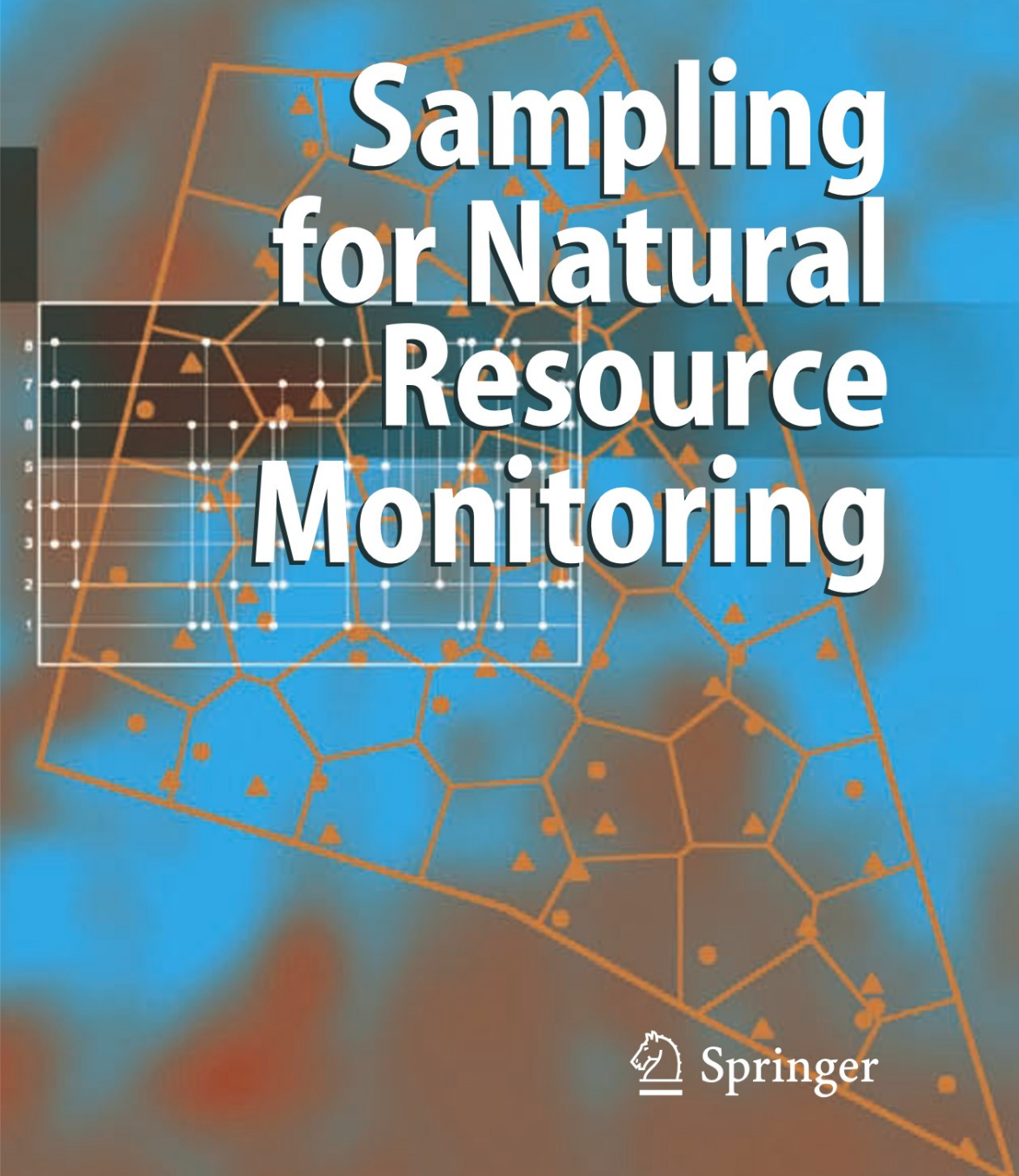
Jaap de Gruijter
Dick Brus
Marc Bierkens
Martin Knotters

Sampling for Natural Resource Monitoring



Springer

Jaap de Gruijter
Dick Brus
Marc Bierkens
Martin Knotters

Sampling for Natural Resource Monitoring

Jaap de Gruijter
Dick Brus
Marc Bierkens
Martin Knotters

Sampling for Natural Resource Monitoring

With 91 Figures

Springer

Jaap J. de Gruijter
Wageningen University
and Research Centre
Alterra
P.O. Box 47
6700 AA Wageningen
The Netherlands

Dick J. Brus
Wageningen University
and Research Centre
Alterra
P.O. Box 47
6700 AA Wageningen
The Netherlands

Marc F.P. Bierkens
Department of Physical Geography
Faculty of Geosciences
Utrecht University
P.O. Box 80115
3508 TC Utrecht
The Netherlands

Martin Knotters
Wageningen University
and Research Centre
Alterra
P.O. Box 47
6700 AA Wageningen
The Netherlands

Library of Congress Control Number: 2005937930

ISBN-10 540-22486-6 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-22486-0 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media
springer.com

Printed in The Netherlands

Cover design: E. Kirchner, Heidelberg
Production: Almas Schimmel
Printing: Krips bv, Meppel
Binding: Stürtz AG, Würzburg

Printed on acid-free paper 32/3141/as 5 4 3 2 1 0

To

Jaap's wife Marianne
Dick's wife Nelleke
Marc's family
Martin's parents
the memory of our colleague Wim te Riele

Preface

"Can you give me a reference to a book on monitoring?" is the kind of question we are frequently asked by colleagues who are planning to design a monitoring scheme. Although this is a perfectly reasonable question, we became increasingly uneasy, because we had to admit that, to our knowledge, a book meeting these researchers' needs did not exist. More frustration built up as monitoring projects grew in number, diversity and importance in our research environment. Therefore, after much hesitation, we decided to try and write such a book ourselves. But why were we not satisfied with the existing literature? Let us spell this out in some detail, because this explains largely why we wrote this book the way we did.

There are several books and chapters of books dealing with the subject of monitoring, i.e., sampling in space, time or space–time. These books focus on applications within a particular discipline, for instance hydrology or vegetation science, and of course treat only the methodology relevant to that discipline. In doing so, they normally use the scientific jargon common to that field. However, scientists working in other fields may need different monitoring methods and may also profit from a more generic presentation. Furthermore, we disagree with some of the existing texts on statistical grounds. Our criticism relates to misconceptions about the differences between the design-based and the model-based approach to sampling.

Broadly speaking, monitoring draws on three distinct statistical methodologies: sampling theory, time-series analysis and geostatistics. Many books are available on each of these fields, but they lack the methodological scope that is generally needed in monitoring. Furthermore, they rarely give enough help to practitioners who want to know how to make a proper selection of methods from these fields, and how to integrate them into a monitoring scheme. Books on sampling theory are usually written in the abstract and generic terms of sampling from finite populations, while it is not always clear how these methods should be adapted and applied to continuous space, time or space–time. These books are usually written for statisticians. There are several good, practical handbooks on geostatistics, but they do not cover classical

sampling methodology, nor sampling in time, whereas books on time-series analysis fail to discuss classical sampling methodology and sampling in space.

There is a host of research articles on statistical methods for monitoring, scattered over a wide range of journals and proceedings, many of which are highly specialized or theoretical. It would hardly be possible for a practitioner to extract the relevant knowledge from this type of literature alone. Review articles come closer to our purpose, but limitations of space prevent them from offering full coverage of methods, and they often focus on a particular field of application. Thus we identified a need for a handbook on the statistical methodology of monitoring that gives applied scientists sufficient guidance in how to design a monitoring scheme, rather than presenting a large collection of methods.

Although we focus on practitioners in the field of natural resource monitoring rather than statisticians, basic statistical knowledge is required for a proper understanding of the methodologies described. If one decides, on the basis of this book, to apply an advanced method of which the theoretical basis is not fully fathomed, we advise to consult an applied statistician. In general, such a consult is useful in the very beginning of a monitoring project. This book may offer the kind of knowledge helpful in an efficient and fruitful communication on statistical aspects of monitoring.

The following persons commented on an earlier version of the manuscript: Prof. Jelke Bethlehem (Statistics Netherlands), Dr. Frans van Geer (Netherlands Institute of Applied Geoscience TNO), Dr. Gerard Heuvelink (Wageningen University, Soil Science Centre), Prof. Alex McBratney (The University of Sydney, Agricultural Chemistry and Soil Science Dept.), Dr. Tony Olsen (US Environmental Protection Agency's Western Ecology Division), and Prof. Scott Urquhart (Colorado State University, Statistics Dept.). We are grateful for their constructive criticism and suggestions. Of course, all shortcomings of this book remain the responsibility of the authors. We also thank Martin Jansen (Alterra) for helping us to produce the figures, and Wybo Dekker (Dutch TeX Users Group) for assistance with LaTeX.

Financing was received from the Dutch Ministry of Agriculture, Nature Management and Fisheries via research programs 317, 325, 328, 329, 358, 382, 394 and 395, and by Alterra (Wageningen University and Research Centre).

Last but not least we thank all colleagues who presented their sampling problems to us in such a great variety, and who were patient enough to explain why our initial perceptions and ideas were too simplistic. This helped us much to deepen and widen our understanding of sampling.

Wageningen
August 2005

Jaap de Gruijter
Dick Brus
Marc Bierkens
Martin Knotters

Contents

1

Introduction

1.1 A View on Survey and Monitoring

Survey and monitoring of natural resources are becoming increasingly important worldwide. Although we have no reliable figures, it is safe to say that survey and monitoring activities have grown rapidly during the past 50 years, both in number and diversity, and continue to do so. Two reasons for this are apparent: human pressures on natural resources are increasing dramatically, and public awareness of their vulnerability and the necessity of sustainable use has risen sharply. Exhaustion of non-renewable resources (e.g., fossil fuels), and deterioration due to pollution, erosion, desertification and loss of biodiversity are now widely recognized as threats to human life. The rational management, protection, and use of natural resources at various decision levels all need reliable information on these resources, and this information can only be gathered by observation.

Definitions

Survey means collecting information on an object with a spatial extent through observation, such that possible changes of the object during the observation are negligible. The result of a spatial survey may consist of one or more statistics of the object as a whole, so-called 'global quantities', or it may relate to the spatial distribution of properties within the object: so-called 'local quantities'. Examples of a survey aimed at global quantities are: (a) establishing the mean clay content of the topsoil of an arable field, and (b) establishing the areal fraction of a region occupied by a given vegetation type. Examples of a survey aimed at local quantities are: (a) establishing the mean clay content of the topsoil of sections of an arable field, and (b) mapping the vegetation types in a region. Broadly speaking, the reason for survey is to provide a factual basis for management (e.g., in the context of precision agriculture or countryside stewardship), or for scientific research.

Survey of the same object may or may not be repeated in time. If it is repeated, and the reason for this lies in the need to keep track of possible changes in the object, i.e., to *update* the existing information, this represents an instance of 'monitoring'.

According to Webster's dictionary, monitoring is: "to watch, observe or check for special purposes". The good part of this definition is that it links monitoring explicitly with "special purposes", to the extent that without a special purpose we should not even speak of monitoring. Nevertheless, for the purpose of this book, two essential elements are missing in this definition. First, monitoring is repeated and continued for a shorter or longer period of time. Second, observation is being done in a more-or-less systematic way. Generally speaking, monitoring of natural resources should provide the information that is necessary for taking proper decisions on natural resources management.

Hence our definition: monitoring is *collecting information on an object through repeated or continuous observation in order to determine possible changes in the object*. Note that the object of monitoring may or may not have a spatial extent. If it has, observation can proceed via sampling in space–time. An example of an object without spatial extent is a point in a river where water levels are measured repeatedly: a case of sampling in time.

Not every sequence of surveys of the same object is regarded as monitoring. For instance, the soil in a region may be surveyed at successively larger cartographic scales (changing spatial resolution), starting with a 'reconnaissance' and then mapping at semi-detailed and detailed scales. In this case, the reason for returning to the same object is not to update the information (as the properties concerned are virtually constant in time), but rather the need to *upgrade* the existing information. Apart from greater precision, the aim of upgrading may also be to provide information on new properties. In practice, combinations of updating and upgrading also occur. All in all, we distinguish five situations with respect to observation:

1. Single observation on an object with spatial extent: 'survey';
2. Sequence of surveys for upgrading: collectively called 'survey';
3. Sequence of surveys for updating: collectively called 'monitoring';
4. Sequence of surveys for updating and upgrading: collectively called 'monitoring';
5. Repeated observation on an object without spatial extent: 'monitoring'.

In this section, we have made a distinction between survey and monitoring, which necessarily stressed the difference between them. We made the distinction because it has important implications for the design of schemes and the choice of methods. On the other hand, the reason why we treat these two kinds of activity in one book lies not in their differences, but in their similarities. The main similarity is that survey and monitoring are both based on sampling: sampling in space for spatial survey and sampling in time or space–time for monitoring. Many sampling-related methods and techniques are generally

applicable: in space, in time and in space–time. Therefore it is economical to cover both survey and monitoring in the same book. From the user's point of view, this is appropriate as well, because a common approach to monitoring is to perform it as a more or less coordinated sequence of surveys.

Objects and Techniques

The object of survey or monitoring can be any natural or semi-natural system. Examples include a river, a forest, the soil of a farm, the groundwater in a region, the vegetation in a conservation area, the rainfall on a continent, the atmosphere of the world. These examples illustrate the wide variety of possible objects, in terms of both their nature and their spatial extent. The temporal extent of monitoring also varies greatly: from a few months, e.g., a growing season of a crop, to more than a century in the case of meteorological monitoring.

With regard to soil, a traditional type of monitoring is for the nutrient status of agricultural fields as a whole, while recently schemes are being developed to monitor variations within fields, to support precision agriculture. Also, in the last decades schemes have been set up to monitor soil quality and soil pollution at a regional or national scale.

Monitoring in hydrology shows a large variety of aims and scales. As for soil, monitoring of open water as well as groundwater may be directed to quality and pollution, or otherwise to quantity, with water level as an important aspect. Monitoring in ecology has still a wider scope than in soil science and hydrology; important objectives are evaluation of effects of environmental changes on species abundances and occurrence of vegetation types.

A huge variety of observation techniques are used for survey and monitoring, ranging from remote sensing to proximal sensing techniques, from simple field observations to highly advanced laboratory analyses, from a single observation to hundreds of different observations made per sampling unit. The observations can be made on a nominal, ordinal, interval or ratio scale. Observations can be made directly on the object, or indirectly, by pre-processing the results of one or more direct observations.

Aims

With a view on designing monitoring schemes, it is useful to distinguish three categories of monitoring according to its aim (Dixon and Chiswell, 1996; Loaiciga et al., 1992):

- *status monitoring* for quantitative description of the universe as it changes with time;
- *trend monitoring* to decide whether temporal trends are present in the universe;
- *regulatory* or *compliance monitoring* to decide whether the universe satisfies regulatory conditions.

In *status* or *ambient* monitoring, the aim is to characterize the status of the object, and to follow this over time. Examples are: the timber volume of a forest, the presence of indicator plant species in an ecosystem, the leaching of nitrate to the groundwater from the soil of a farm, the emission of greenhouse gases by a country, the climate in a given part of the world. The main reasons for this type of monitoring are that information about the system is needed for management, administration, regulation or scientific research.

In *trend* or *effect* monitoring, the aim is to study the possible effects of a natural event or a human activity on the object, for instance the effect of drinking water extraction by a pumping station on the water tables in a region, the effect of a hydrologic measure against desiccation of an ecosystem in a conservation area, the effect of a change in agricultural policy on the land use in a country. Thus, the aim of effect monitoring is not only to find out whether there has been a change, as in status monitoring, but also to establish whether the change was caused by a specified event or measure. The reasons for effect monitoring are similar to those for status monitoring.

In *compliance* or *regulatory* monitoring, the aim is to decide whether the object complies with a given regulatory standard, e.g., to check whether a system of obligatory crop rotation is actually being applied in an agricultural region, or whether heavy metal concentrations in soil used for crop production remain below specified maximum levels. The reason for compliance monitoring is generally law enforcement.

The above broad distinction of aims is relevant to sampling, because for status monitoring the sampling should allow efficient estimation or prediction of descriptive parameters repeatedly, while for trend and regulatory monitoring it should provide statistical validity and sufficient power of hypothesis testing or acceptable error rates in classifying the object into categories. (See Sect. 2.2 for these modes of statistical inference.)

1.2 Aim and Scope

The aim of this book is to present to practitioners the statistical knowledge and methodology needed for survey and monitoring of natural resources. We intend to omit all theory not essential for applications or for basic understanding. Where possible, we refer to the sampling and monitoring literature for specific topics. In one respect, however, this presentation is broader than standard statistical texts: we pay much attention to how statistical methodology can be employed and embedded in real-life survey and monitoring projects. Thus, we discuss in detail how efficient schemes for survey and monitoring can be designed in view of the aims and constraints of a project.

Our scope is limited to *statistical methods*, because these methods allow the quality of results to be quantified, which is a prerequisite for optimization of survey or monitoring schemes, as well as for risk management and quality control. A further limitation is imposed by the assumption that complete

observation of the object is not feasible, so that the survey or the monitoring must be conducted by observations on one or more *samples* from the object. Since budgets for survey and monitoring are and will remain limited, it is important to know how to design cost-effective schemes.

The book presents statistical methods of *sampling* and *inference* from sample data. From a statistical point of view, the core of survey and monitoring is first of all *sampling*, either in space (at a single instant but at multiple locations), or in time (at a single location but at several instants), or in space–time (at multiple locations and times). Sampling is therefore the main focus of this book. However, not all sampling in space and/or time is directly aimed at survey or monitoring purposes. For instance, in scientific research, the purpose of sampling may be to generate a hypothesis about a physical or ecological process, to calibrate a model, or to describe relationships via multivariate analysis. Sampling for such purposes is not covered by this book. Obvious exceptions are sampling for variogram modelling (Chap. 9) and time-series modelling (Chap. 13), because such models are essential for model-based survey and monitoring. A borderline case is where a sample is taken to build a regression model, which is then used to make predictions about the target variable at grid nodes in space, as a method of survey. In this case, the sample is used only indirectly for survey and, from a statistical point of view, sampling for regression modelling is quite different from sampling directly for survey or monitoring. This is why we do not address this case.

We present methodologies that we consider to be generally useful for survey and monitoring of natural resources. We do not, however, cover highly specialized methods of geologic, meteorologic and faunistic survey and monitoring, nor do we treat sampling of lots of natural products or sampling for detection of local critical conditions (Sect. 2.2.6).

The spatial scale varies from a single agricultural field, as in precision agriculture, to continental, as in monitoring the water quality of large rivers. The temporal extent in monitoring varies from, say, a growing season, to many decades in long-term monitoring of variables such as water tables.

Although the methodology presented in this book is widely applicable in natural resource monitoring, the examples that we present to illustrate these methods are mostly taken from our own background knowledge: soil, groundwater, land use, landscape and, to a lesser extent, vegetation.

1.3 Basic Concepts and Terminology

The natural resources about which survey or monitoring in a given application intends to provide information are referred to as the *universe of interest*, or briefly the *universe*. Universes in this context are biotic or a-biotic systems varying in space and/or time. Some universes may, for the purpose of sampling, be regarded as a physical continuum, e.g., the soil in a region at some moment, the water of a river passing a point during a period, the crop on a field at some

time[1], or the natural vegetation in an area during a period[2]. The number of dimensions of continuous universes may be 1, 2 or 3 for spatial universes, 1 for temporal universes, and 2, 3 or 4 for spatio-temporal universes. Discrete universes are populations of individual entities, such as the trees in a forest, the lakes in a province, the arable fields in a region, or the growing seasons of a crop in a region. Although the individuals of a discrete population have dimensions and positions in space and time, contrary to continuous universes, a discrete universe itself has no measure of size other then the number of individuals in it.

We use the term *sampling* in the usual broad sense of selecting parts from a universe with the purpose of taking observations on them. The selected parts may be observed in situ, or material may be taken out from them for measurement in a laboratory. The collection [3] of selected parts is referred to as the *sample*. To avoid confusion, a single part that is or could be selected is called a *sampling unit*. The number of sampling units in a sample is referred to as the *sample size*. The material possibly taken from a sampling unit is referred to as an *aliquot*. Aliquots from different sampling units bulked together form a *composite aliquot* or briefly a *composite*.

The individuals of discrete universes naturally act as sampling units. Sampling units from a continuous universe have to be defined more or less arbitrarily. However defined, any sampling unit has a shape and a size. Within the universe it has an orientation (if the shape is not round) and a position. Shape, size and orientation together are referred to as the *sample support*. The dimensions of a sampling unit may be so small compared with the universe, that they can be neglected. In that case the unit can be identified by a point in space and/or time, and we speak of *point support*. In ecological monitoring the sampling units are usually two-dimensional in space, and referred to as *quadrats*.

We refer to the position of a sampling unit in space-time as a *sampling event*, with a *sampling location* and a *sampling time*. Of course, when sampling from a spatial universe, the sampling time is irrelevant from a statistical point of view, and we shall speak about the sampling locations only. Similarly, when sampling from a temporal universe, the sampling location is irrelevant, and we shall speak about the sampling times only. However, a sampling event in a spatio-temporal universe must identified by both location and time. For instance, an observation taken at such an event could be the water-table elevation at a given location and a given time. The same location and a different sampling time would make a different sampling event, as would a different

[1] A 'continuum view' of crop is appropriate if the interest lies in crop properties per areal unit of the field. However, if the interest lies in properties per plant, then the universe is to be regarded as a discrete population.

[2] A monitoring period may or may not have a pre-defined end.

[3] If each part occurs only once in the collection, as is usually the case, then it is a set. In probability sampling, however, there are two forms of collection: *sets* in sampling without replacement, and *sequences* in sampling with replacement

location and the same sampling time. Also, when dealing with positions of sampling units in a universe in general, without specifying whether this is spatial, temporal or spatio-temporal, we shall use the generic term 'sampling event'.

The positions of all sampling units of a sample together form a pattern in space and/or time, referred to as a *sampling pattern*. Selecting a sample can thus be done by selecting a sampling pattern. Very often, the actual selection of sampling units is not done in the field but in the office, using some representation of the universe, such as a list or a map. This representation is called a *sampling frame*.

More terms are explained elsewhere in this book, where they are first introduced. A subject index is provided at the end.

1.4 Structure and Use of this Book

We tried to structure this book such that a scientist who has to set up a survey or monitoring project, can find ample guidance in how to analyze his/her particular monitoring or survey problem, and how to select the methods to solve it. Part I is entirely devoted to this. Those who are looking for methods which optimally fit their specific purpose, especially beginners in this field, are advised to read this part before going to the methodological parts.

Part I is composed of four chapters. The preparatory Chap. 2 recapitulates the various modes of sampling unit selection and of statistical inference from sample data. Chapter 3 presents seven principles that we consider essential for scheme design. Chapter 4 discusses three major design decisions: (1) the choice between design-based and model-based inference, (2) the choice of sample support, and (3) choices on composite sampling. Finally, Chap. 5 deals with optimization of sample selection.

After the part on scheme design follows the presentation of sampling methods in Parts II, III and IV. We have structured this presentation broadly in a problem-oriented rather than a method-oriented way, to make it as easy as possible to find a suitable method, given the survey or monitoring problem at hand. Therefore, we adopted the following hierarchy to structure the material:

1. at part level, a division according to the dimensions of universe: *space*, *time* or *space–time*;
2. at chapter level, a division according to the kind of target quantity: *global* or *local*;
3. at section level, a division according to the approach to sampling: *design-based* or *model-based*.

Although the latter distinction is method-oriented, it is strongly related to the kind of results that are requested from survey or monitoring. The division into design-based and model-based methods at this high level is also warranted because the choice between them has major consequences, both for the sampling and the inference stages.

Many methods dealt with in this book can be applied to spatial, as well as temporal or spatio-temporal universes. Rather then re-iterating such methods in all three parts, we treat them only in Part II on sampling in space. This part therefore includes many generally applicable methods, thus forming a large portion of the book. For example, Chap. 9 about methods for variograms is placed in Part II, but these methods are equally relevant to survey and monitoring. Similarly, time-series models are only presented in Part III on sampling in time, but they are equally relevant to space–time monitoring.

The price for the above mentioned conciseness is, of course, that one may have to go back to one or even two earlier parts. Thus the reference to parts is as follows:

- sampling for survey: go to Part II on sampling in space;
- sampling for monitoring on a single location: go to Part III on sampling in time, and go back to Part II when needed;
- sampling for monitoring in space and time: go to Part IV on sampling in space–time, and go back to Part II and/or Part III when needed.

Referencing at a more detailed level is provided in the introductory chapter at the beginning of each part.

1.5 Notation

The typographic conventions in this book are as follows.

- Variables: small or capital italic. Target variables, transformed target variables, and ancillary variables are generically denoted with the small letters z, y and x, respectively, if they are deterministic. If they are stochastic, then they are denoted with the capitals Z, Y and X.
- Errors or residuals: e if they are deterministic, and ϵ if they are stochastic.
- Vectors: bold upright small letters ($\mathbf{s}, \boldsymbol{\lambda}$)
- Matrices: bold upright capitals ($\mathbf{S}$, $\mathbf{C}$)
- Transposed vectors and matrices are denoted with a prime ($\mathbf{s}', \boldsymbol{\lambda}'$)
- Sets: calligraphic capitals ($\mathcal{S}$,$\mathcal{U}$), but the usual $\mathbb{R}$ for the set of real numbers
- Size of sets: $|\mathcal{U}|$, etc.

The following general symbolism is used throughout this book.

- Means over spatial, temporal and space–time universes are indicated by bars, for instance: $\bar{z}_{\mathcal{U}}$ is the mean of deterministic variable z in universe $\mathcal{U}$.
- Estimators defined on the basis of sample data are indicated by a hat, for instance: $\hat{\bar{z}}_{\mathcal{U}}$ is an estimator of the mean of deterministic variable z in universe $\mathcal{U}$.

- Predictors defined on the basis of sample data are indicated by a tilde, for instance: $\widetilde{\overline{Z}}_{\mathcal{U}}$ is a predictor of the mean of stochastic variable Z in universe $\mathcal{U}$.
- Prior estimates or 'guesses' are indicated by a breve, for instance: $\breve{\breve{z}}_{\mathcal{U}}$ is a prior estimate of the mean of deterministic variable z in universe $\mathcal{U}$.
- Variances are either denoted with $V(\cdot)$ or $S^2(\cdot)$ or σ^2, depending on the kind of variance. Variances between realizations of a stochastic process are denoted with $V(\cdot)$. Such a process may be repeated sampling (as in the design-based approach), or it may be hypothesized through a stochastic model of spatial or temporal variation (as in the model-based approach). For instance: $V(Z(\mathbf{s}))$ is the variance of stochastic variable Z at location $\mathbf{s}$, as determined by a stochastic model.
  Variances as a measure of dispersion in space, time or space–time are denoted with $S^2(\cdot)$. For instance, $S^2(z_{\mathcal{D}})$ is the variance between all values of variable z within domain $\mathcal{D}$. Variance as a parameter in stochastic models of the variation in space and/or time is denoted with σ^2, also used as short-hand for prediction-error variance in kriging and Kalman filtering.

Part I

DESIGNING SCHEMES FOR SURVEY AND MONITORING

2

Modes of Sampling and Statistical Inference

As explained in Sect. 1.4 the methods for estimating global or local quantities in space, time or space–time, are grouped in design-based and model-based methods. With 'method' we mean a combination of a method for selecting sampling units and a method for statistical inference, for instance estimation of a spatial mean or prediction of the values at points. A design-based method is defined as a method in which sampling units are selected by probability sampling and in which statistical inference is based on the sampling design, i.e., design-based inference, see Table 2.1. A model-based method is defined as a method in which the statistical inference is based on the model. There are no requirements on the selection of the sampling units, but in general probability sampling is sub-optimal for model-based inference, and purposive sampling of units is more efficient. Typical examples of purposive sampling for model-based inference are Centred Grid Sampling, Spatial Coverage Sampling and Geostatistical Sampling. In the following sections we elaborate on these modes of sampling unit selection and on modes of statistical inference.

2.1 Modes of Sampling Unit Selection

Three possible modes of sampling unit selection can be distinguished: *convenience* sampling, *purposive* sampling and *probability* or *random* sampling. The concept of convenience sampling is self-explanatory. An obvious example is when sampling is limited to roadsides or other easily accessible spots. The advantage of this mode is that it saves time and costs. The disadvantage is that the statistical properties are inferior to those of the other modes. For instance, estimates from a convenience sample have to be regarded as biased unless one is willing to accept specific assumptions about the sampling process and the spatial and temporal variation. These assumptions are often debatable, and this may or may not be acceptable, depending on the context of the project.

Table 2.1. Definition of design-based and model-based method as a combination of a method for selection of sampling units and a method for statistical inference

Type of method	Selection method	Inference method
Design-based method	Probability sampling	Design-based
Model-based method	Purposive sampling	Model-based

Purposive sampling tries to select the sampling units such that a given purpose is served best. An example is the 'free survey' method of mapping soil classes, whereby the surveyor selects the sampling locations that are expected to be most informative with respect to soil class delineation. In this example, the locations are selected in a subjective manner, using experience, visible landscape features and pedogenetic hypotheses, such that the surveyor expects the most useful information from his observations. Another example is where a centred regular grid or a zigzag transect is projected on a field in order to obtain a 'representative' sample.

Purposive sampling can also be done by formally optimizing an objective function related to the purpose. For instance, if the purpose is to map a spatial distribution by kriging and if geographical boundary effects are disregarded, it can be shown that the prediction-error variance is minimized by a centred triangular grid of sampling locations, under assumptions of stationarity and isotropy (McBratney et al., 1981).

If prior to the sampling a statistical model for the variation in the universe can be postulated and the prediction-error variance is a function of the sampling events only, i.e., independent of the sample data, then one could use this model to optimize the spacing of a regular grid, given a quality requirement on the mean or maximum kriging variance. The model can also be used to optimize the sampling events in the universe given the sample size. Such samples are referred to as model-based samples, and more specific as geostatistical samples when the postulated model is a geostatistical model. These methods are discussed in Sects. 8.3.2 and 8.3.4, respectively. A technique of intermediate complexity is that of Spatial Coverage Sampling, which optimizes an objective function of distances only (Sect. 8.3.3).

Probability sampling, unlike the other modes, selects sampling units at random. If this is done properly (according to a well-defined sampling design) the probabilities of selecting the units are known, and these probabilities provide the basis for statistical inference from the data. As discussed in Sects. 7.2 and 15.2, there are many techniques for random selection of sampling units. Collectively, this approach to sampling is referred to as the design-based approach, as opposed to the model-based approach, where the sampling units are fixed instead of random, and statistical inference is based on a model of the variation in the universe. The difference between design-based and model-

based inference is further explained in Sect. 2.2.1. The choice between these two approaches is an important statistical issue, which is dealt with separately in Sect. 4.1.

2.2 Modes of Statistical Inference from Sample Data

2.2.1 Design-Based versus Model-Based Inference

There are two fundamentally different approaches to sampling: the *design-based* approach, followed in classical survey sampling, and the *model-based* approach, followed in geostatistics and in time series analysis. Differences and relationships between these two approaches are extensively addressed in Särndal et al. (1992). De Gruijter and ter Braak (1990) discuss the issue in the spatial context, but the distinction also holds, and is equally relevant, for sampling in time and in space–time. The difference between the two approaches is illustrated in Fig. 2.1 by a simple example (Brus and de Gruijter, 1997): a square area is sampled at 25 locations and a 0/1 indicator variable z is measured to estimate the fraction of the area with value 1. Figure 2.1**a** shows a spatial distribution of z and 25 sampling locations. Averaging the observed values at these locations yields an estimate of the fraction.

Both approaches quantify the uncertainty of such an estimate by considering what would happen if sampling were repeated many times in a hypothetical experiment. Obviously, if neither the pattern of values nor the sampling locations were changed in this experiment, there would be no variation, so one or the other has to be varied. The two approaches differ as to which one is varied.

The design-based approach evaluates the uncertainty by repeated sampling with different sets of sampling locations, while regarding the pattern of values in the area as unknown but fixed. The sets of sampling locations are generated according to a chosen random sampling design. The *row* of figures (**a**, **b** and **c**) represents three possible outcomes.

By contrast, the model-based approach evaluates the uncertainty by repeated sampling with a fixed set of sampling locations, while varying the pattern of values in the area according to a chosen random model of the spatial variation. In this approach, the *column* of figures (**a**, **d** and **e**) represents three possible outcomes.

The experiment can remain truly hypothetical in most instances, because probability calculus enables one to determine what happens on average over all possible realizations. In more intricate situations, however, this is infeasible and repeated sampling has to be simulated numerically, varying either the sampling locations or the pattern of values, or both.

The fact that the two approaches use a different source of randomness has several important practical as well as theoretical consequences. Here we briefly discuss three of them. The main consequence is that the statistical inference

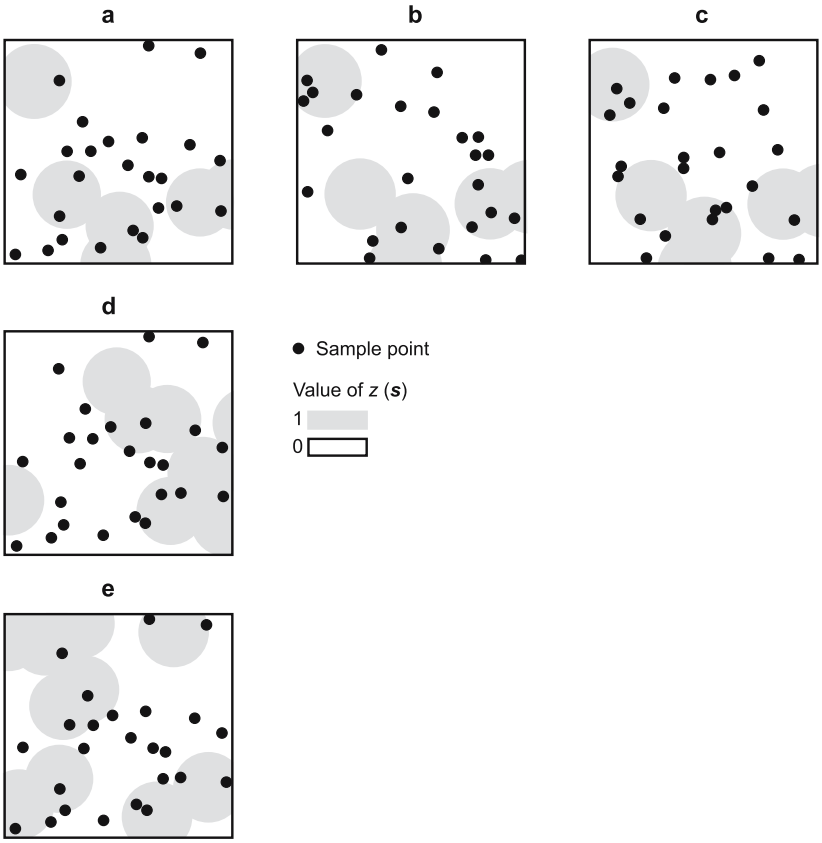


Fig. 2.1. Repeated sampling in the design-based approach (**a**, **b**, **c**) and in the model-based approach (**a**, **d**, **e**). In the design-based approach, the pattern of values is regarded as fixed and the sampling locations are random. In the model-based approach, the sampling locations are fixed and the pattern of values is regarded as random. (from Brus and de Gruijter, 1997)

from sample data is entirely different. In the design-based approach, inference is based on the selection probabilities of sampling locations as determined by the random sampling design. This means that in calculating weighted averages, for instance, the data are assigned weights that are determined by the selection probabilities of the sampling locations, not by their geographical coordinates. In the model-based approach, inference is based on a stochastic model of the variation in the universe. Here the weights of the data are determined by the covariances between the observations, which are given by the model as a function of the coordinates of the sampling locations.

Another important difference between design-based and model-based inference, apart from weighting, is the way they consider the quantity about which inference is made. This 'target quantity', a concept discussed in more detail in Sect. 3.1, is regarded as unknown but fixed in design-based inference. See, for instance, the constant areal fraction in Figs. 2.1**a**, **b** and **c**. In the model-based approach, however, there are two kinds of target quantities that can be chosen for inference: functions of parameters of the adopted stochastic model, and functions of realizations from that model. Model parameters and functions thereof are fixed by definition, while quantities of realizations are random.

A common and practically relevant example is that the model-based approach allows inference to be made about the model mean, denoted as μ, or about the mean defined by summation or integration over space and/or time, denoted by $\overline{Z}$. The former is fixed while the latter is random. For instance, the realizations in Figs. 2.1**a**, **d** and **e** have different areal fractions but the same underlying model mean. The difference between a fixed and a random target quantity is not merely a theoretical subtlety, but has direct consequences for the quantification and even the definition of the precision and reliability of results, as will be explained later in this section.

Finally, the choice between design-based and model-based inference (discussed in Sect. 4.1) has major consequences for sampling. Design-based inference requires some form of probability (random) sampling, while model-based inference typically builds on purposive sampling, and needs a model of the spatial variation. Acquisition of such a model will usually require an extra sampling effort. Furthermore, random sampling optimization is methodologically different from purposive sampling optimization, as discussed in Chap. 5.

It should also be realized that optimal sampling for model-based inference on model parameters differs from optimal sampling for model-based inference on realization quantities. For instance, inference about a model mean often requires a considerably larger sample than inference about a spatial, temporal or spatio–temporal mean with the same level of precision. Also, these two target quantities generally require different sampling patterns. Hence, different equations should be used to calculate sample sizes and different methods used for sample optimization.

It is useful to distinguish between inference modes not only in terms of whether the target quantity is considered as fixed or random, but also whether the inference result is quantitative (number or interval) or qualitative, for instance, that the target quantity exceeds a given level. Quantitative and qualitative results need their own methods of inference, and have their own quality measures for precision and reliability, so the sample size needed to meet a quality requirement has to be determined in different ways.

Combining the distinctions 'random versus fixed target quantity' with 'qualitative versus quantitative inference result' yields four different modes

Table 2.2. Summary of inference modes for different combinations of target quantities and types of results

Basis of inference	Target quantities	Type of result	Inference mode
Sampling design	Statistics defined on a *fixed* field of z-values, (e.g., the deterministic spatial mean $\bar{z}$)	Quantitative Qualitative	Estimation Testing
Stochastic model	Statistics defined on a *random* field of Z-values, (e.g., the stochastic spatial mean $\overline{Z}$), or:	Quantitative Qualitative	Prediction Classification
	Model parameters or functions thereof (*fixed* by definition)	Quantitative Qualitative	Estimation Testing

of inference: *estimation*, *prediction*[1], *testing* and *classification*. These modes are summarized in Table 2.2. A fifth mode, *detection*, should be applied when the question is whether at any point in the universe some critical condition is present, without asking where or when. The answer to this type of question can be coded as a 0/1 indicator variable. As this relates to the universe as a whole, it is a global quantity.

It is important to choose carefully between estimation and prediction on the one hand, and hypothesis testing, classification and detection on the other. This choice should be made prior to the actual design of the sampling scheme, because it is determined by the kind of information that is required. This issue is therefore discussed in Sect. 3.1 as part of the design information. Here we repeat as a guideline that whenever a choice must be made between two alternatives, e.g., whether or not to take a particular action or to draw a particular conclusion, and when this choice is to be made on the basis of sample data, hypothesis testing, classification or detection is in order. In all other cases, the original problem can be formulated as an estimation or prediction problem.

The statistical literature provides a huge variety of inference methods. In the following sections we only recapitulate the concepts that are most relevant to sampling. We illustrate the four modes of inference with the mean as target quantity, and the case depicted in Fig. 2.1 as an example. The

[1] Prediction should not be confused with forecasting. Prediction is quantitative inference on a random target quantity that may be spatial, temporal or spatio-temporal. Forecasting is a special case of prediction, i.e., when a (spatio-)temporal quantity is related to the future.

target variable in Fig. 2.1 is a 0/1 variable, indicating the presence or absence of some condition, e.g., the occurrence of a given pathogen in the soil. So in this example the mean can be interpreted as the areal fraction of infected soil.

2.2.2 Estimation

Estimation is quantitative inference on a fixed target quantity, however, as explained in the previous section, there are two kinds of fixed quantities: statistics defined on a fixed field of z-values, and (functions of) model parameters (Table 2.2). So, in the example of the areal fraction of infected soil, one first has to decide what is to be estimated: the *spatial* mean or a *model* mean. If the interest is focused on the infection actually present in the particular study region, then one would estimate the spatial mean. If, on the other hand, the interest is broader and relates to infection that may be present on average in a hypothetical ensemble of similar regions (of which the study region is only one example), then one would estimate a model mean.

Estimation of the Spatial Mean

When estimating the spatial mean, we consider this as fixed, as in the row of Figs. 2.1**a**, **b**, and **c**. The true value of the mean, denoted by $\bar{z}$, equals 0.30, but is unknown in practice. The design-based *estimator* of $\bar{z}$, denoted by $\hat{\bar{z}}$, is generally a function of the sample data without their coordinates in space and/or time. If the sample is drawn according to Simple Random Sampling (see Sect. 7.2.3), the usual estimator is the unweighted sample mean (unweighted because the sampling locations were selected with equal probability, not because of a model assumption):

$$\hat{\bar{z}} = \frac{1}{n} \sum_{i=1}^{n} z_i \ , \tag{2.1}$$

where n denotes the sample size (here 25), and z_i denotes the value measured at the i-th sampling location.

The data of the random sample depicted in Fig. 2.1**a** are: 8 'infected' out of 25, giving an *estimate* $\hat{\bar{z}} = 0.32$, with *estimation error* $\hat{\bar{z}} - \bar{z} = 0.02$. The other, equally probable samples depicted in Figs. 2.1**b** and **c** would yield 0.40 and 0.32, respectively, with estimation errors 0.10 and 0.02. In practice, the estimation error is unknown for any specific sample (otherwise there would be no need for sampling), but probability theory enables us to make statements about the estimation error in terms of averages.

First, it can be shown that the estimator, averaged over samples, equals the true value. In other words, $\hat{\bar{z}}$ is *p-unbiased*:

$$E_p\big(\hat{\bar{z}}\big) = \bar{z} \ , \tag{2.2}$$

where $E_p(\cdot)$ denotes the *p-expectation*, i.e., the average over a large (strictly: infinite) number of samples drawn according to the random sampling design p, here Simple Random Sampling with $n = 25$.

Second, the p-expectation of the squared estimation error equals the *sampling variance* of the estimator:

$$V_p\big(\hat{\bar{z}}\big) = E_p\big\{\hat{\bar{z}} - E_p\big(\hat{\bar{z}}\big)\big\}^2 = E_p\big(\hat{\bar{z}} - \bar{z}\big)^2 \ . \tag{2.3}$$

Because the sampling locations were selected independently from each other, the number of 'infected' in the data follows a binomial distribution, and the sampling variance equals

$$V_p\big(\hat{\bar{z}}\big) = \frac{\bar{z}\,(1-\bar{z})}{n} \ . \tag{2.4}$$

The *standard error* of the estimator is a usual statistical quality measure. It equals the square root of the sampling variance:

$$S_p\big(\hat{\bar{z}}\big) = \sqrt{\frac{\bar{z}\,(1-\bar{z})}{n}} \ . \tag{2.5}$$

In our example V_p equals $0.30 \times 0.70/25 = 0.0084$, and S_p equals 0.0917.

In practice these true values would be unknown because $\bar{z}$ is unknown but, from data obtained by Simple Random Sampling, the sampling variance can be estimated (again p-unbiased) by

$$\widehat{V}_p\big(\hat{\bar{z}}\big) = \frac{1}{n\,(n-1)} \sum_{i=1}^{n} \big(z_i - \hat{\bar{z}}\big)^2 \ . \tag{2.6}$$

In our example, with 0/1 data, this results in:

$$\widehat{V}_p\big(\hat{\bar{z}}\big) = \frac{\hat{\bar{z}}\,\big(1-\hat{\bar{z}}\big)}{n-1} \ . \tag{2.7}$$

The estimated standard error follows as its square root:

$$\widehat{S}_p\big(\hat{\bar{z}}\big) = \sqrt{\frac{\hat{\bar{z}}\,\big(1-\hat{\bar{z}}\big)}{n-1}} \ . \tag{2.8}$$

From the sample data in Fig. 2.1**a** we calculate $\widehat{V}_p = 0.00907$, and $\widehat{S}_p = 0.0952$.

Apart from *point* estimation as discussed above, giving a single number or point on the scale of possible values, one can also perform *interval* estimation. The result is a *confidence interval*. Confidence intervals are constructed in such a way that they contain the true value of the target quantity with probability $(1-\alpha)$, referred to as the *confidence level*. For instance, the data from the sample in Fig. 2.1**a**, 8 'infected' out of 25, result in (0.14; 0.54) as a 95% confidence interval for the areal fraction of infected soil. (This interval is based on the fact that the number of 'infected' follows a binomial distribution, but a Normal or Student distribution is usually applicable, especially with quantitative target variables and medium size or large samples.)

Estimation of the Model Mean

In practical applications one usually has to choose a model for the inference on the basis of sample data and possibly ancillary data. In the present example, however, we happen to have exact knowledge a priori of the model underlying the patterns of Figs. 2.1**a**, **b** and **c**, simply because we used this model to generate those patterns.

The model was discussed by Matérn (1986) in the context of the so-called 'bombing problem': bombs are dropped at random over a region and each bomb devastates the area within a given distance r from the point of impact. It can be shown (Matérn, 1986, Eq. 3.4.2) that under this model the expectation of the 0/1 variable Z at location $\mathbf{s}$ equals:

$$E_\xi\{Z(\mathbf{s})\} = 1 - \mathrm{e}^{-\lambda\pi r^2} \;, \tag{2.9}$$

where $E_\xi(\cdot)$ denotes the ξ-*expectation*, i.e., the average over a large (strictly: infinite) number of random realizations from the chosen model, and where λ is the intensity of the assumed stationary Poisson process by which the centres of the infection circles are spread over the region.

Because of the spatial stationarity, meaning that the expectation at locations is constant over the region, the model mean μ equals $E_\xi[Z(\mathbf{s})]$. For our example we have chosen $\lambda = 6$ and $r = 0.15$. From (2.9) we calculate for the true value of the model mean: $\mu = 0.346$, which is the areal fraction 'infected' averaged over a large (strictly infinite) number of realizations.

To estimate the model mean, we need the covariance between any two observations on Z. The covariance between two observations at locations h apart equals (Matérn, 1986, Eq. 3.4.3):

$$C(h) = \mathrm{e}^{-2\lambda\pi r^2}\left(\mathrm{e}^{\lambda\, A(h,r)} - 1\right) \;, \tag{2.10}$$

where $A(h,r)$ is the area of the intersection common to two circles with radius r and centres h apart. This area equals:

$$A(h,r) = r^2\left[2\arccos\left(\frac{h}{2r}\right) - \sin\left\{2\arccos\left(\frac{h}{2r}\right)\right\}\right] \;. \tag{2.11}$$

The model mean can be estimated with the so-called Best Linear Unbiased Estimator (Searle, 1997):

$$\hat{\mu} = \left(\mathbf{1}'\mathbf{C}^{-1}\mathbf{1}\right)^{-1}\mathbf{1}'\mathbf{C}^{-1}\mathbf{z} \;, \tag{2.12}$$

where $\mathbf{z}$ is a vector of n observations, $Z(\mathbf{s}_i)$, $\mathbf{C}$ is the matrix of their variances and covariances, and $\mathbf{1}$ is the n-vector of ones. This estimator is ξ-unbiased, a property defined by averaging over model realizations (given the sampling locations), not over the sampling process, i.e., $E_\xi(\hat{\mu}) = \mu$. The variance of $\hat{\mu}$ equals:

$$V_\xi(\hat{\mu}) = \left(\mathbf{1}'\mathbf{C}^{-1}\mathbf{1}\right)^{-1} . \tag{2.13}$$

From the sample depicted in Fig. 2.1**a**, and Eqs. 2.10 – 2.13, we calculate $\hat{\mu} = 0.353$ with standard error $S_\xi(\hat{\mu}) = \sqrt{V_\xi(\hat{\mu})} = 0.132$. Note that, although the estimation error in this case is only 0.007, the standard error of estimation of the model mean is 44% larger than the standard error of estimation of the spatial mean.

2.2.3 Prediction

With the example of Section 2.2.2 we illustrated design-based estimation of the spatial mean $\bar{z}$ and model-based estimation of the model mean μ, both means considered as fixed. Using the same data and the same model, we now illustrate model-based prediction of the spatial mean $\overline{Z}$, considered as random. Analogous to the Best Linear Unbiased Estimator of μ, one can predict $\overline{Z}$ with the Best Linear Unbiased Predictor:

$$\widetilde{\overline{Z}} = \boldsymbol{\lambda}'\mathbf{z} , \tag{2.14}$$

where $\boldsymbol{\lambda}$ is a vector of optimal weights, and $\mathbf{z}$ is again the vector of observations. These weights are chosen such that the prediction is ξ-unbiased, i.e., $E_\xi\left(\widetilde{\overline{Z}} - \overline{Z}\right) = 0$, and the variance of the prediction error, $E_\xi\left(\widetilde{\overline{Z}} - \overline{Z}\right)^2$, is minimized. These optimal weights can be calculated by:

$$\boldsymbol{\lambda} = \mathbf{C}^{-1}\mathbf{r} - \mathbf{C}^{-1}\mathbf{1}\left(\mathbf{1}'\mathbf{C}^{-1}\mathbf{1}\right)^{-1}\mathbf{1}'\mathbf{C}^{-1}\mathbf{r} + \left(\mathbf{1}'\mathbf{C}^{-1}\mathbf{1}\right)^{-1}\mathbf{C}^{-1}\mathbf{1} , \tag{2.15}$$

where $\mathbf{r}$ is the vector of mean covariances between each of the actual sampling locations and all potential locations in the region. This can be approximated numerically by the mean covariances between the sampling locations and the nodes of a fine grid.

The variance of the prediction error equals

$$V_\xi\left(\widetilde{\overline{Z}} - \overline{Z}\right) = E_\xi\left(\widetilde{\overline{Z}} - \overline{Z}\right)^2 = \overline{C}_{G,G} + \boldsymbol{\lambda}'\mathbf{C}\boldsymbol{\lambda} - 2\boldsymbol{\lambda}'\mathbf{r} , \tag{2.16}$$

where $\overline{C}_{G,G}$ is the mean covariance between all potential sampling locations of the region, which can be approximated numerically by the mean covariance between the nodes of a fine grid.

From the sample depicted in Fig. 2.1**a** and Eqs. 2.14 – 2.16 we calculate $\widetilde{\overline{Z}} = 0.386$, with standard error (square root of the prediction-error variance) 0.078. Note that this standard error of the spatial mean is only 59% of the standard error of the model mean, using the same data and the same model. (A considerably larger sample would be needed to estimate the model mean with the same precision as for predicting the spatial mean.)

Apart from *point* prediction as discussed above, one can also perform *interval* prediction. Just as an estimation interval (Sect. 2.2.2), a prediction

interval is constructed in such a way that it contains the true value of the target quantity with probability $(1 - \alpha)$. If on the basis of the Central Limit Theorem it can be assumed that the prediction error is approximately normally distributed, then for the boundaries of a 95% prediction interval, for instance, one can simply take the prediction plus and minus 1.96 times the standard error. The data from the sample in Fig. 2.1**a**, 8 'infected' out of 25, then give (0.233; 0.538) as 95% prediction interval for the areal fraction of infected soil. Note that this model-based prediction interval is 25% narrower than the design-based estimation interval calculated in Sect. 2.2.2 for the same target quantity, using the same sample data and the same confidence level. (One should realize, however, that in practical applications the variogram is not known and must be estimated, which is an additional source of error.)

In cases where the assumption of approximate normality does not hold, the distribution of $\overline{Z}$ may be evaluated numerically by Monte Carlo simulation of realizations from the chosen model. Deutsch and Journel (1998) and Pebesma (2004) provide simulation software.

2.2.4 Testing

When decisions are based on an estimate of the global mean, the areal fraction or a percentile (e.g., median or P95), one must take account of the uncertainty in the estimate. For instance, in environmental studies a threshold value for the concentration of a pollutant is often defined on the basis of regulatory requirements, risk assessments or a reference standard for deciding whether or not to take action. The threshold value is referred to as the Action Level. Due to sampling and measurement error, there is always some degree of uncertainty whether the true value is above or below the Action Level, which means that there is a chance of taking a wrong decision. A statistical tool for decision making in the presence of uncertainty is statistical testing of hypotheses. A distinction can be made between one-sample tests and two-sample tests. In one-sample tests, sample data from a target area are compared with an absolute criterion such as a regulatory threshold. In two-sample tests sample data from a target area are compared with other sample data, for instance from another region or from another period for the same target area. In the two-sample case, the target quantity is usually the difference between the two (spatial or temporal) means, and the Action Level is often zero.

For an example of a one-sample test we go back to Fig. 2.1 and suppose that some remediation action is to be taken if the areal fraction of infected soil, $\bar{z}$, is larger than Action Level 0.2. The null-hypothesis, H_0, is then: $\bar{z} \leq 0.2$, which is tested against the alternative hypothesis H_1: $\bar{z} > 0.2$. Acceptance or rejection of H_0 is determined by the value of a *test statistic*, in this case the sample mean $\hat{\bar{z}}$, in the following way. If the value of the test statistic falls in a specific interval, called the *critical region*, then H_0 is rejected, otherwise it is accepted. This critical region is constructed such that the probability of falsely rejecting H_0 is limited to a value α, chosen by the user.

Table 2.3. The four possible combinations of the state of nature and test results.

	Test result	
State of nature	H_0	H_1
H_0	No error Prob= $1-\alpha$ Confidence	Type I error Prob= α
H_1	Type II error Prob= β	No error Prob= $1-\beta$ Power

Table 2.3 shows the four possible combinations of the state of nature and test results. Wrongly accepting H_1 is called a type I error, the probability of which is α. The probability of accepting H_0 when H_0 is true is called the confidence level of the test, $(1-\alpha)$. Wrongly accepting H_0 is called a type II error, the probability of which is denoted by β. The probability of accepting H_1 when H_1 is true is called the power of the test, $1-\beta$. Given the sampling design, the power is a function of the sample size, the confidence level $(1-\alpha)$, and the H_1 hypothesis itself.

When we use the data from the sample in Fig. 2.1**a** for a one-sided binomial test of H_0: $\bar{z} \leq 0.2$ at confidence level 0.95, we find as critical region [0.36; 1], which does not include the value of the test statistic (0.32), so that H_0 is not rejected.

There is a close connection between estimation and testing: a confidence interval contains all possible values of the target quantity that would not be rejected on testing at the same confidence level. However, the statistical quality measures differ (standard error or half-width of confidence interval versus confidence level and power), and so do methods to determine the sample size needed to reach a given quality level.

2.2.5 Classification

The term classification is used here for assigning an object to a class under uncertainty about the properties of the object. In the present context the object is the universe of interest or a part of it, and the uncertainty about its properties is represented by a stochastic model. For a simple illustration we take once more the case of Fig. 2.1**a**. Suppose that the region is to be classified in one of two classes: 'low infection' defined as $\overline{Z} \leq 0.5$, and 'high infection' defined as $\overline{Z} > 0.5$. After sampling and measuring one could decide to classify the region simply by comparing the predicted spatial mean with the class boundary 0.5. However, due to the uncertainty about the true value

of the spatial mean, there will be a risk of false 'low' classification and a risk of false 'high' classification. So, just as in testing, there are two types of possible errors, and in view of the consequences of either one it may be appropriate to classify in such a way that the probability of the most important error type, for instance false 'low', is not too high. This can be done by classifying according to a rule in terms of error probability instead of the prediction itself.

For example, the rule could be that the region will only be classified as 'low' if the probability of false 'low' is smaller than 0.05, and otherwise as 'high'. Assuming that the prediction error is normally distributed, the probability of false 'low' can be calculated from the prediction (0.386) and its standard error (0.078; see Sect. 2.2.3). This turns out to be 0.07, which is larger than the chosen threshold 0.05, hence the region is classified as 'high', although the predicted value falls in class 'low'.

As with prediction, if the assumption of normality does not hold, the distribution of $\overline{Z}$ must be evaluated numerically by Monte Carlo (geostatistical) simulation of realizations from the chosen model (Deutsch and Journel, 1998; Pebesma, 2004).

2.2.6 Detection

An example of a detection problem in the case of Fig. 2.1**a** is the question whether there is soil infection *at all* in the area, regardless where. The assessment method is often designed such that if the critical condition is observed in any of the sampling units, then it surely exists. In that case the probability of a false positive is zero, and inference only needs to quantify the probability of a false negative or its complement, the detection probability. Sampling for detection should therefore aim at maximizing the detection probability for a given budget, or minimizing the sampling costs under the condition of a given minimal detection probability. This will naturally lead to grid sampling or spatial coverage sampling.

The detection probability can in principle be calculated from the geometry of the sampling pattern used, and certain assumptions about the occurrence and extent of the condition in space and/or time. These assumptions may be captured in a stochastic model of the variation, e.g., the second-order stationary model with covariance function (2.10) underlying Fig. 2.1. The detection probability for any sampling pattern can then be calculated by geostatistical simulation (Deutsch and Journel, 1998; Pebesma, 2004). Alternatively, knowledge about the occurrence and extent of the critical condition in space and/or time may be embodied in a physical process model.

Detection of critical conditions sometimes asks for highly specialized and application-dependent sampling and inference methods. Such methods are not treated in this book.

3

Basic Design Principles

The aim of this chapter is to offer help in designing schemes for survey or monitoring. To this end, we present in the following sections seven principles that we consider essential for good design. These principles are:

1. Develop a complete scheme (Sect. 3.1).
2. Structure the design process (Sect. 3.2).
3. Pay ample attention to practical issues (Sect. 3.3).
4. Employ prior information on variation (Sect. 3.4).
5. Balance the various sources of error (Sect. 3.5).
6. Anticipate changing conditions during monitoring (Sect. 3.6).
7. Calculate the sample size appropriately (Sect. 3.7).

3.1 Develop a Complete Scheme

Survey and monitoring of natural resources often involves the following activities.

- Planning field activities: given the purpose of the project, the budget and possible logistical constraints, it is decided how many samples and/or field measurements are to be taken, as well as where, how and when.
- Field activities: taking samples and/or field measurements.
- Laboratory work: sample preparation and analyses.
- Data recording.
- Data processing.
- Reporting.

Roughly speaking, these activities can be thought of as the consecutive stages of a survey project. In the case of monitoring, field and laboratory work, as well as data recording and processing are obviously done in some cyclic or continuous fashion. The activities mentioned above may overlap in time. For instance, data recording and field work are often done simultaneously.

Also, the process may involve switching back and forth between activities. For instance, if some deficiency is discovered during data processing, additional field work may be needed. Laboratory work is optional, as measurements may be taken in the field.

The main purpose of this section is to argue that, although the above sequence of activities may seem reasonable, it does not make good sampling practice. The reason is that an essential element is missing at the beginning of the sequence: the element of planning the whole chain of activities, including the statistical procedure to be used in data processing. Careful planning of the entire project is a prerequisite of good sampling practice and should precede any other activity. Whereas researchers usually invest enough effort and ingenuity in deciding how, where and when to take samples and measurements, their ideas about how to analyze the data very often remain rather vague until the data are there and crisp decisions must be made about what to do with them. In that case, more likely than not, data analysis and data acquisition will not be properly attuned to each other. Due to this mismatch, the potential qualities that a data acquisition plan might have are not fully exploited, and sub-optimal results are obtained. One example is where a stratified random sample has been taken, but this sample is analyzed as if it were a simple random sample. Another example is where the data are to be analyzed by some form of kriging, but it is found that the variogram needed for this cannot be reliably estimated from the data. Finally, a common situation is where the conclusions to be drawn from the sample data can only be based on questionable assumptions because the sample was not properly randomized, like most legacy soil survey data. These examples will be discussed in greater detail in the next sections.

Apart from attuning data acquisition to data processing and vice versa, there is a more general reason why the project should be planned as a whole rather than by optimizing parts of it in isolation: the consequences of a decision about a particular issue, in terms of quality and costs, depend on the decisions taken on other issues. A simple example is where two assessment methods are available for the target variable: an inexpensive but inaccurate method and an expensive but accurate method. The choice between the two affects both the costs and the accuracy of the final result, and these effects depend on the sample size. Given a fixed budget, choosing the inexpensive method implies that a larger sample size can be used, which may or may not lead to a better result.

In summary, we recommend planning not only the data acquisition but the entire project, paying special attention to the agreement between data acquisition and data processing. Proper planning of the entire project will most likely pay itself back by increased efficacy and efficiency. We want to emphasize this principle by referring to the entire plan as the *scheme*. Our broad concept of scheme covers much more than just how, where and when to sample and measure. A scheme captures explicitly all the decisions and

information pertinent to data acquisition, data recording and data processing. It consists of the following items.

1. Detailed analysis and specification of the *objective*.
   a) *Target universe*: a precise definition of the universe of interest, with boundaries in space and/or time, and possibly a specification of exclusions. (Note that, for various reasons, the target universe may differ from the actually *sampled universe*.) In case of ecological populations, a decision on how the universe will be treated, as continuous or as discrete (Sect. 4.2).
   b) *Domain(s)* of interest: a specification of the part(s) of the universe for which a separate result is required. At one extreme (in terms of extent), this is the entire universe, at the other extreme it is a point or a set of points in the universe (e.g., grid nodes used to prepare a contour map). In between these extremes, there may be a number of domains with smaller or larger extents in space and/or time. Examples include sub-areas within a spatial universe, or spatial cross-sections through a space–time universe (i.e., the space at given points in time).
   c) *Target variable(s)*: a precise definition of the variable(s) to be determined for each of the sampling units. (Note that target variables are generally not identical with actually *measured variables*, because of measurement errors or transformations of measured variables prior to statistical inference.)
   d) *Target parameter*: the type of statistic for which a result is needed, given the target variable(s) and the domain(s). Examples include total, mean, fraction, median, standard deviation or trend parameter.
   e) *Target quantity*: the combination of a domain, target variable and target parameter is referred to in this book as a target quantity. An example is the mean (parameter) phosphate content in the topsoil (target variable) of the agricultural soils in the Netherlands (domain). A target quantity that is related to the entire universe is referred to as a 'global quantity'; in all other cases it is referred to as a 'local quantity'.
   f) *Type of result*: quantitative or qualitative. If a quantitative result is desired, then the mode of inference has to be *estimation* or *prediction*. If a qualitative result is needed, e.g., an answer to the question whether or not the target quantity exceeds a given level, then the mode of inference has to be *testing*, *classification* or *detection* (Sect. 2.2).
2. *Quality measure*: the quantity used to express numerically the statistical quality of the survey or monitoring result (Sect. 5.1). Examples include the half-width of a 95% confidence interval in estimation, the error variance in prediction, the power in hypothesis testing, the error rate in classification.
3. *Constraints*: the allocated budget and/or minimum quality of the result, fieldwork (optional), transport (optional) and laboratory capacity (optional).

4. *Prior information* (Sect. 3.4).
   a) *Sampling frame*: the list, file or map identifying the sampling units from which the sample is selected.
   b) *Miscellaneous information*: general knowledge, experience and information from comparable projects, existing sample data, maps or GIS files.
   c) *Model of the variation* of the target variable within the universe (eventually needed if model-based inference is chosen; see item 8). Examples include a variogram adopted from a comparable case or estimated from a preliminary sampling round (Chap. 9).
5. *Sample support*, in the case of a continuous universe (Sect. 4.2); physical sampling devices for taking aliquots (optional).
6. *Assessment method*: field and/or laboratory measurement procedures (reference to existing protocols where possible); method of data pre-processing to calculate the target variable from measured variables (optional).
7. Whether and how to use *composite sampling*, i.e., bulking aliquots (Sect. 4.3).
8. Choice between *design-based* and *model-based inference* from sample data (Sect. 4.1).
9. For design-based inference: choice of random sampling *design type* and *attributes* of the chosen design type (Sect. 5).
10. For model-based inference: choice of *sampling pattern type* and *optimization* algorithm and restrictions (optional), (Sect. 5).
11. Identification of the actually selected *sample*: a list of sampling unit labels, a map with sampling locations, a table of sampling times or coordinates of sampling events in space–time.
12. *Protocols* on data recording and field work (Sect. 3.3).
13. Method to be used for *statistical inference*.
14. *Prediction* of operational *costs* and *quality* of results: ex-ante evaluation (Sect. 3.3).

The scheme item 'target parameter' deserves more attention before we continue with examples of a scheme. Parameters may be related to either a frequency distribution or a probability distribution. Frequency distributions are integrals over space and/or time, which of course applies only to non-point domains, i.e., domains with an extent. Parameters related to frequency distributions are the total, mean, mode, standard deviation, percentiles (e.g., the median) and fractions. (A fraction is the relative size of that part of the domain where a given state is present, such as exceedance of a threshold value). Examples include an areal fraction, a temporal mean and a spatio-temporal total. We refer to such parameters as 'frequential parameters'. Probability distributions, on the other hand, are integrals of probabilities defined by a stochastic model. Parameters usually related to probability distributions are the expected value, mode, standard deviation, percentiles (e.g., the median) and probabilities of exceedance. Such parameters are referred to as 'probabilistic parameters'. They may be defined for the target variable at a given

point of the universe but also for frequential parameters, which unfortunately complicates matters. For instance, the chosen parameter may be the probability that the spatio–temporal total of the target variable over a given domain exceeds a given threshold value.

To illustrate our concept of a survey or monitoring scheme, we give two hypothetical examples. In order to avoid lengthy descriptions and details not required for a correct understanding, we present the examples in a concise form. In a real project, they should of course include the full details.

Example I, 2D survey

1. *Objective.*
   a) *Target universe*: the topsoil of all arable fields of a specified class in region R at time T, a continuous universe.
   b) *Domain*: the entire region.
   c) *Target variable*: a variable indicating whether or not the (true) mean cadmium concentration in the topsoil at a location in the region exceeds critical concentration level C (0 means 'no'; 1 means 'yes'). This 'indicator variable' is needed because the corresponding areal fraction is to be used as the target quantity (see item 1e).
   d) *Target parameter*: the spatial mean.
   e) *Target quantity*: the areal fraction of the arable fields in region R where at time T the (true) cadmium concentration in the topsoil exceeds level C. (Note: sampling is done over a relatively short period, during which the concentrations are assumed to be constant.)
   f) *Type of result*: qualitative result, accepting or rejecting the null-hypothesis that the areal fraction is below a given critical level F, by testing at the 95% confidence level.
2. *Quality measure*: the power of the test (the probability of rightly rejecting the null-hypothesis), when the actual areal fraction exceeds the critical level F by a given amount.
3. *Constraints*: the power must not be less than a given value (a quality requirement). This implies that the scheme should aim to minimize the costs while satisfying the quality requirement.
4. *Prior information*:
   a) *Sampling frame*: a GIS file containing the boundaries of the arable fields in region R. (Note that this may not be an error-free representation of the target universe.)
   b) *Miscellaneous information*: a map showing expected cadmium concentrations, compiled from local knowledge of pollution sources.
   c) *Model of the spatial variation*: not available.
5. *Sample support*: standard auger core from the topsoil.
6. *Assessment method*: a specified laboratory method for concentration (the measured variable), followed by 0/1 transformation of the cadmium con-

centrations, with 0 for no exceedance and 1 for exceedance of critical level C.

7. *Composite sampling?* No Composite sampling. (Note: bulking aliquots would in this case lead to biased estimates of the target variable; see Sect. 4.3.)
8. *Design-based or model-based inference?* Design-based inference.
9. *Design type* and *design attributes*: Stratified Simple Random Sampling (design type), with specified strata (attribute 1) derived from the map of expected concentrations, and specified sample sizes in the strata (attribute 2) chosen to ensure that costs are minimized while the quality requirement is satisfied.
10. (For model-based inference: not applicable)
11. *Sample*: reference to a map indicating the sampling locations, selected according to the chosen design type and design attributes.
12. *Protocols*: reference to documents.
13. *Method of inference*: standard one-sided test of the null-hypothesis, with the assumption that the estimated areal fraction is approximately normally distributed (the sample size should be large enough).
14. *Ex-ante evaluation*: reference to a document.

Example II, monitoring in 2D space and time

1. *Objective.*
   a) *Target universe*: the water passing through a cross-section S of a river during year Y, a continuous universe.
   b) *Domains*: the water passing through the cross-section during the summer and during the winter.
   c) *Target variable*: the (true) amount of salt passing through 1 m^2 of the cross-section during 1 min.
   d) *Target parameter*: spatio–temporal total.
   e) *Target quantities*: the total salt loads passing through cross-section S during the summer and the winter of year Y.
   f) *Type of result*: quantitative, as 95% prediction intervals.
2. *Quality measure*: half-width of the widest of the two prediction intervals.
3. *Constraints*: a limited budget is available, which implies that the scheme should aim at minimizing the half-width of the widest prediction interval while keeping the costs within budget; due to shipping traffic on the river, sampling is confined to a specified acceptable zone of the cross-section.
4. *Prior information*:
   a) *Sampling frame*: a map of cross-section S, conceptually combined with the continuous time scale of a year with 365 days.
   b) *Miscellaneous information*: sample data from the same river but for previous years.
   c) *Model of the variation*: a space–time variogram of the salt flux was developed from the available sample data.

5. *Sample support*: implicitly defined by the assessment method.
6. *Assessment method*: a specified sensing method to measure salt concentration and a specified sensing method to measure water flux (two measured variables), followed by multiplication of the salt concentrations with the water fluxes to calculate the target variable.
7. *Composite sampling?* No composite sampling (measurement in situ; no aliquots are taken).
8. *Design-based or model-based inference?* Model-based inference.
9. (For design-based inference: not applicable)
10. *Sampling pattern type and optimization*: a number of fixed sampling locations where measurements are taken simultaneously at equidistant sampling times (sampling pattern type: regular grid in space–time); optimization through evaluation of costs and quality measure for all combinations of eligible sampling locations (e.g., from a $1 \times 1\,\mathrm{m}^2$ grid on the acceptable zone of the cross-section) and eligible sampling time intervals (e.g., in a series of one day, one week, two weeks, four weeks, 13 weeks).
11. *Sample*: a map indicating the sampling locations in the cross-section and a table of sampling times.
12. *Protocols*: reference to documents.
13. *Method of inference*: calculation of the 95% prediction interval for the salt loads from space–time block-kriging predictions and associated kriging variances.
14. *Ex-ante evaluation*: reference to a document.

3.2 Structure the Design Process

If designing a scheme is regarded as an instance of problem solving, items 1–4 can be seen as the information which is used to find a solution: the 'design information'. From this strict perspective, items 5–13 together constitute the selected solution, and the final item (14) is an ex-ante evaluation of that solution. From a broader perspective, the design information, especially the items 'objective', 'quality measure' and 'constraints' will already be the result of a 'translation' of an initial, more general description of the aim of the project. This translation typically settles various details that were left undecided until then. This can usually be done in different ways, the alternatives having a potential effect on the costs and quality of the results. Therefore, we consider it to be a fundamental part of the design process as a whole. Clearly, the translation requires close interaction with and agreement from the stakeholders. It is probably not unusual that, at some point in the design process, parts of the translation will have to be reconsidered and repeated.

A safe way to obtain a good scheme is based on the following principle: '*Start at the end, then reason backwards*'. This means that one should first determine precisely what information is needed. Only when the information need has been defined it does become useful to search for a scheme that

satisfies this need in an efficient way. The reason for this is that different information needs generally ask for different schemes. Although this is one of the most important facts in sampling, it does not seem to be always clearly acknowledged. We shall therefore discuss this in more detail.

Information needs in the context of survey or monitoring can be divided into two broad groups. In the first group, the purpose may be to estimate a global quantity, i.e., a parameter of the *cumulative distribution function* of the target variable over the entire universe. Examples are 'location'[1] parameters such as the mean, quantiles (e.g., the median) and the mode, and 'dispersion' parameters such as the standard deviation, range and tolerance intervals.

In the second group, the purpose may be some kind of description of the *spatial and/or temporal distribution* of the target variable within the universe. Examples are prediction of values at specific points within the universe, estimation of means within parts of the universe and construction of contour maps.

In general, different types of results ask for different sampling designs, because a given scheme may not yield the type of result that is required, or it may do so in an inefficient way. For instance, estimating the spatial mean of a region (a global target quantity) requires other, less expensive schemes than the prediction of the values at grid nodes (local quantities), as is done for mapping. In conclusion, a good way of designing a scheme is by reasoning backwards through the following steps:

1. Decide precisely what information is needed. Examples include a map of a given variable, at a given scale and with a given accuracy, or testing a given hypothesis, at a given confidence level and with a given statistical power.
2. Identify the constraints that apply to the production of the required information.
3. Identify what useful information is already available.
4. Determine what kind of data analysis leads to the required type of result.
5. Identify the data needs for this analysis and search for a strategy to obtain these data in the most efficient way.

This sequence of steps implies that, prior to making any design decision, one should first collect all design information. The reason for this is that the design decisions are otherwise likely to be premature and need to be reconsidered once the design information has been made more complete or explicit.

After all design information has been collected, the question remains how to organize the rest of the design process. More specifically, in what order should the various other items of the scheme be decided? It would be unfeasible to provide a detailed design protocol that is suitable for all possible circumstances. Nevertheless, the following global guidelines seem to be useful:

[1] In this context 'location' does not refer to position in geographical space, but to position on the measurement scale.

1. Make a provisional decision on the assessment method (scheme item 6).
2. Choose a quality measure (scheme item 2, see Sect. 5.1).
3. Make a rough estimate of the sample size affordable at a given budget or needed to meet a given quality requirement.
4. Make provisional decisions on the following major design issues:
   a) Choice between design-based and model-based inference (scheme item 8; see Sect. 4.1).
   b) Choice of sample support (scheme item 5; see Sect. 4.2).
   c) Choice of whether and how to bulk aliquots (scheme item 7; see Sect. 4.3).
5. In the case of design-based inference: search for an optimal random sampling design (scheme item 9; Sect. 5.2.1).
6. In the case of model-based inference: choose a sampling pattern type and optimization technique (scheme item 10), such as simulated annealing and genetic algorithms (Appendix A), and optimize the sample.
7. Make a prediction of operational costs and quality of results (scheme item 14). If the expected costs are too high or the quality too low, then revise one or more of the provisional design decisions. Otherwise, the decisions are final.
8. Draw a random sample according to the chosen sampling design, or optimize a sample (scheme item 11).
9. Work out the method of statistical inference (scheme item 13).
10. Develop protocols on field work and data recording (scheme item 12).

The reasons for making the major design decisions (mode of inference, sample support and aliquot bulking) early in the process is that these decisions tend to have dominant effects on both costs and quality, and that most other decisions depend on them.

The scheme above assumes a single target quantity. In practice, especially in monitoring situations, one has multiple target quantities, which makes optimization a more complex problem, further discussed in Sect. 5.2.

Design processes are seldom linear and one-way. There are often good reasons to loop back to earlier provisional design decisions, or even to the design information, e.g., to switch to a less demanding aim, to relax a constraint, or to search for other prior information. Our advice is to keep track of the process to prevent it from becoming haphazard or chaotic, and also to enable reporting of the reasons for the various choices that are finally made.

3.3 Pay Sufficient Attention to Practical Issues

Designing a scheme for survey or monitoring is not just a matter of statistical methodology. On the contrary, if the practical issues are disregarded, there is a high risk that the project will be unsuccessful. Therefore, without pretending to be exhaustive, we discuss what would seem to be the main practical issues.

Avoid Undue Complexity

Researchers often know a great deal about the physical processes that generate spatial patterns or time series of properties in the universe of interest. They may be tempted to express all this knowledge in detail in the form of a highly complex sampling design. Although understandable, this attitude entails three risks which are easily underestimated.

First, due to unforeseen operational difficulties during field work, it may prove impossible to carry out the design in all its complexity. The field work must then be adjourned until the design has been adjusted. This may be time-consuming and is likely to cause undesirable delay.

Second, complexities are introduced to increase the efficiency, but they may make the statistical analysis much more intricate and time consuming than expected. It is therefore usually wise to avoid highly complex sampling designs, because the theoretical gain in efficiency compared with simpler solutions is easily outweighed by the practical difficulties. Also, multiple target variables may be of interest, and one may face the problem that an efficient design for one target variable can be inefficient for another.

Third, complex sampling designs can be efficient for one target variable, but inefficient for another variable. For instance, stratification of the target area may lead to increased precision for a target variable related with the stratification variable, but for target variables that are not related, there may be no gain in precision or even a loss of precision. Therefore, for surveys and monitoring with multiple target variables we recommend keeping the sampling design as simple as possible, and using instead the ancillary information at the estimation stage, for example by using the post-stratification estimator (Sect. 7.2.11).

Allow for Unexpected Delay in Field Work

Even if one is familiar with the circumstances in the terrain, there may be factors beyond one's control that prevent the field work from being completed within the available time. Clearly, unfinished field work may seriously harm the statistical potential of the design. It is therefore prudent to allow some extra time in the scheme for contretemps, say 20 % of the total time for field work, and to include a number of optional sampling locations to be visited as the extra time allows.

Include a Test Phase if Necessary

If there is significant uncertainty about the logistics of the field work or the spatial or temporal variability, a preliminary test phase is always worth the extra effort. The data from even a small sample, collected prior to the main sample, enables the latter to be optimized more precisely and reduces the risk

that the project will not meet its goal at all. In the final statistical analysis, the data from the test phase can often be combined with those for the main sample, so that the additional effort is limited to extra travel time and statistical analysis.

Evaluate the Scheme Beforehand

It is good practice to quantitatively predict the operational costs of the scheme, and the accuracy of the result, prior to starting the field work. Predicting cost and accuracy can be done in sophisticated ways, using mathematical models (Domburg et al., 1994), or more generally, using experience from similar projects, rules-of-thumb and approximations. A test phase will of course improve the prediction of costs and accuracy.

Explicit ex-ante evaluation in terms of costs and accuracy is not only a final check of whether the scheme can be trusted to lead to the goal, it also enables comparison with ex-post evaluation, i.e., after the project has been completed. If this reveals significant discrepancies, the causes should be analyzed. This may provide a basis for better planning of future projects.

Protocol for Field Work

Rules for field work will usually concern the physical act of taking samples and/or measurements in the field, but they should also indicate what should be done if a sampling location is inaccessible or if it falls outside the universe. An example of the latter in soil sampling is where, on inspection in the field, it turns out that at the given location there is no 'soil' according to an intended definition.

A poor protocol may seriously affect the quality of the results. Obvious requirements for a protocol are that it is complete, unambiguous, practically feasible and scientifically sound. The scientific aspect plays a role, for instance, when a rule says that an inaccessible sampling location is to be shifted to a nearby location in a certain way. In principle, this leads to over-representation of boundary zones and, depending on the kind of design and the statistical analysis, may result in biased estimates.

Protocol for Data Recording

Just as for field work, there should be sound rules for data recording. These rules should not only cover regular recording but also prescribe different codes for when a sampling location falls outside the universe, when it is inaccessible, when a variable cannot be measured because its value is too large or too small ('censoring' in the statistical sense), and when a variable cannot be measured for other reasons.

3.4 Employ Prior Information on Variation

Any prior information about the variation in the universe should be utilized as good as possible in the search for an efficient sampling design. Examples of prior information are satellite images, aerial photographs, thematic maps (e.g., groundwater, soil or vegetation maps) and theories about the spatial and/or temporal patterns of variation in the universe. Such theories may be available in a verbal, qualitative form, or in the quantitative form of a mathematical model.

There are many ways in which prior information can be exploited in schemes. Two modes can be distinguished. The first mode is to use the prior information in the sampling design, i.e., in the data acquisition stage. The second mode is to use it in the statistical analysis of the sample data, i.e., in the data processing stage. In the following we give examples of each mode.

An example of the first mode is when images, photographs or maps are used to stratify the universe. In this case, the universe is split into a number of relatively homogeneous sub-universes (called 'strata'), which are then sampled independently (Sect. 7.2.4). Another example is when genetic theory enables intelligent guesses about spatial and/or temporal correlations. For instance, in the case of a universe consisting of the soil in a given area, aeolian deposition of parent material in that area may be known to have resulted in little short-range variation of texture. If the target variable is closely related to texture, it will be then important for the sake of efficiency to avoid sampling at locations in close proximity. A final example of the first mode is when a variogram (Chap. 9) is used to optimize the sampling density or sampling frequency.

An example of the second mode is when prior point data are used to design a space-filling sample, and prior and new data are used in spatial interpolation. Another example is the use of ancillary data in regression estimators (Sect. 7.2.11). Brus and de Gruijter (2003) present a method for using prior point data from non-probability sampling in design-based estimation of spatial means.

If prior information on the variation is captured in the form of variograms, these functions can be used to predict the sampling variance for a given design (Sect. 7.2.15). If in addition a model of the costs is available, then it is possible to optimize the sampling design in a fully quantitative manner (Domburg et al., 1997).

3.5 Balance the Sources of Error

It is important to realize that the accuracy of the final result is not only determined by the sampling error, i.e., the error due to the fact that sampling is limited to a finite number of units. Examples of other sources of error are sample treatment, observation, model of the variation, 'censoring' and 'non-response'. Censoring means that no quantitative measurement is possible on

a particular sampling unit, because its true value falls outside the measurable range of the measuring device used for measuring. An example would be a situation in which the water table is to be measured in an auger hole with a depth of 1.5 m depth and no groundwater is observed in the hole. This indicates that the level is deeper than 1.5 m ('right censoring'), and a quantitative assessment can only be produced by guessing or by statistical estimation based on an assumed distribution function. (See Knotters et al. (1995) for an example of the latter.) Another example is where the true value of a concentration is below the detection limit of a chemical analysis ('left censoring').

Non-response is a term used in the general statistical literature to indicate the situation where for some reason no data can be obtained from a sampling unit. In groundwater, soil and vegetation sampling this occurs when a location in the field cannot be visited or when measurement is impossible for a different reason than censoring, e.g., loss of an aliquot.

When the inference from the sample data is based on a model of the spatial variation, this model will generally be a source of error, because the underlying assumptions deviate from reality (see Sect. 4.1).

In many cases the target variable can not be observed without error. Examples are measurements of chemical and physical properties on soil- and water-aliquots. Also, in surveys of elusive populations of plants or animals the observer is typically unable to detect every individual in the neighbourhood of the sampling location or line-transect.

It may be tempting to adopt a cheap-to-measure target variable at the cost, however, of large bias in the final results. Suppose, for instance, that the objective is to estimate the total emission of a pollutant from the soil to the groundwater in a given area during a given period. One possible strategy would be to measure the concentration of the pollutant in the soil moisture at the sampling locations, to estimate the mean concentration from these data, and to multiply the mean concentration with the total groundwater recharge taken from a water balance for the area. The advantage of this strategy is that only concentrations need to be measured. However, the disadvantage is that the estimate of the total emission is possibly seriously biased. The cause of this bias is that the strategy assumes implicitly that concentration and recharge are independent variables, whereas in reality this will not be true; for instance, there may be a tendency for high concentrations at times and at places with low recharge to the groundwater. A solution is measuring not only the concentration at the sampling locations but also the flux to the groundwater, and to take the product of these two as the target variable.

Although any reduction of the sampling error will lead to a smaller total error, there is little point in investing all efforts in further reduction of the sampling error if another source of error has a higher order of magnitude. Therefore, in devising a scheme, the relative importance of all error sources should be taken into consideration. For instance, if the spatial variation within a sampling unit (plot) is small compared to that within the domain, it does

not pay to take many aliquots in the selected plots to estimate the means of the plots. The optimal number of aliquots in a plot also depends on the time needed for taking the aliquots.

See Gy (1979) for a comprehensive theory of error sources in sampling, especially sampling of particulate materials.

3.6 Anticipate Changing Conditions During Monitoring

An extremely important difference between survey and monitoring with respect to scheme design is that survey takes place within a relatively short period of time, during which neither the universe is supposed to change in any relevant way, nor the operational, organisational and budgetary conditions alter. With monitoring, on the other hand, not only the universe may undergo large, unexpected changes, but especially in long-term monitoring the conditions will often alter in a way that makes adaptation of the scheme inevitable or at least desirable. Budgets may vary from one year to another, and operational constraints that were originally present may be relaxed, or new unforeseen ones may come into force.

It is also quite common that the focus is shifted, or that new objectives are defined, e.g., through the introduction of new target variables or domains of interest. Better measurement techniques may become available and, last but not least, spatial variation patterns often change in time. For instance, the spatial variation within strata, as originally defined, may increase to a level that makes stratified sampling on the basis of these strata no longer efficient.

One important condition that will always change during monitoring is the amount of available data: more and more data about the universe will become available through monitoring itself. Thus, the available knowledge about the variation in space and/or time will accumulate to a higher level than the prior information that was used to design the scheme. This in itself may be a good reason for fine-tuning or redesigning the scheme.

All changes mentioned above may call for specific adaptations of the scheme, but the point is that some schemes do not lend themselves well to particular adaptations. For instance, suppose that the target area has been divided into small strata, with only one (permanent) sampling location in each, that the statistical inference will be design-based, and that after some years the total number of locations must be reduced due to budget cuts. One then has to choose between (a) leaving some strata unsampled, which leads to biased results, or (b) switching to a new stratification with fewer strata and newly selected sampling locations within them. This may lead to extra costs, and to less precise estimates of change over time.

Clearly, both options are undesirable. The chosen type of sampling design does not survive any reduction of the sample size, i.e., it has no flexibility in this respect. This trap might be avoided, for instance, by using a non-stratified type of design or by defining fewer strata, allowing for more sampling locations

in each. Such a choice may yield less precise results at the original budget, but the expected loss of initial precision may be less important than greater adaptability to changing budgets. As for surveys with multiple target variables (see Sect.3.3), we recommend strongly to avoid complex sampling designs for selecting the sampling locations of a monitoring scheme. See Overton and Stehman (1996) for a discussion of desirable design characteristics for long-term monitoring.

More generally, it would be unwise to limit the ex-ante evaluation of long-term monitoring schemes to cost and quality based on the initial design information. Different 'what–if' scenarios in terms of changes in conditions and possible adaptations to such changes should be worked out before final decisions are made.

The fact that monitoring, especially long-term monitoring, is bound to face changing conditions calls for flexibility of the scheme. This flexibility is largely determined by the installation costs of the monitoring locations. When these costs constitute a considerable part of the total costs of monitoring, one will be reluctant to move at the next sampling round to other locations, leading to a *static* or a *static-synchronous* sampling pattern, see Sects. 14.1 and 14.2 for a discussion of these pattern types.

3.7 Calculate the Sample Size Appropriately

It is perfectly obvious that in scheme design a correct formula for the sample size must be used, so why should we stress this by presenting it as a design principle? The reason is that we have repeatedly encountered cases in the literature where an incorrect formula was used, sometimes leading to a much larger sample than necessary and a waste of time and money. We discuss three different kinds of mistake in calculating the sample size.

Design Effect Disregarded

A mistake sometimes made in design-based sampling is to use a sample size formula intended for Simple Random Sampling (Eq. 7.8 or 7.9), when a different sampling design will be used, such as Stratified Simple Random Sampling (Sect. 7.2.4). By doing this, the effect of the chosen sampling design on the sampling variance, compared with Simple Random Sampling, is disregarded and as a result the calculated sample size may be either too large or too small. For instance, if stratification is applied, the design effect is usually a reduction of the sampling variance. Disregarding this effect by using a sample size formula for Simple Random Sampling would then lead to a sample that is larger than necessary.

Sample size calculation specifically for designs other than Simple Random Sampling may be problematic in practice, because it needs prior information that is difficult to obtain. In that case one can deliberately choose to calculate

the sample size as if Simple Random Sampling would be applied. If one expects a positive design effect (reducing the sampling variance compared with Simple Random Sampling), one can either accept the calculated sample size as conservative estimate (a number on the safe side), or one can correct it with a reduction factor equal to a prior estimate of the design effect based on experience in comparable projects. If a negative design effect is to be expected, for instance when Cluster Random Sampling or Two-Stage Random Sampling is adopted for operational reasons, then it is especially important to correct the calculated sample size with a prior estimate of the design effect, in order to avoid undersized sampling.

Autocorrelation Disregarded

When sample data are collected not by random but by purposive selection, they should be analyzed by model-based inference, such as block-kriging for prediction of the spatial mean. From a model-based point of view the observations will usually be auto-correlated, which makes the prediction error variance smaller than when no autocorrelation exists. In that sense, kriging takes advantage of auto-correlation. However, if one calculates the sample size from the assumption that there is no autocorrelation (in other words: assuming a pure nugget variogram), while in reality there is, then this advantage is not accounted for. The result is an oversized sample.

Estimation of Model Mean Instead of Spatial or Temporal Mean

A pitfall also worth mentioning here is using a formula that is appropriate for estimating a model mean but not for a spatial or temporal mean. To explain this, consider the variance of the predicted mean of some target variable Z over a universe. Suppose we have n observations on Z, where z satisfies a model with mean μ plus a random component ϵ with variance σ^2:

$$Z_i = \mu + \epsilon_i \qquad (i = 1, \ldots, n) \tag{3.1}$$

If we take the unweighted sample mean as estimator of μ:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} Z_i \ , \tag{3.2}$$

then, if the observations are independent, the variance of this estimator is given by the well-known formula:

$$V_{\text{ind}}(\hat{\mu}) = \frac{\sigma^2}{n} \ . \tag{3.3}$$

However, it was realized long ago (Bayley and Hammersley, 1946) that, if the observations are not independent, then this formula needs adjustment by taking account of the covariances between the observations:

$$V_{\text{dep}}(\hat{\mu}) = \frac{1}{n^2}\left\{\sum_{i=1}^{n}\sigma^2 + 2\sum_{i=1}^{n}\sum_{j=1}^{n}C(z_i, z_j)\right\} = \frac{\sigma^2}{n}\left\{1 + (n-1)\bar{\rho}\right\} \ , \quad (3.4)$$

where $\bar{\rho}$ denotes the average correlation between the observations. So, an equivalent sample size was defined, equal to the nominal sample size divided by the correction factor in (3.4):

$$n_{\text{eq}} = \frac{n}{\{1 + (n-1)\bar{\rho}\}} \ . \quad (3.5)$$

This formula for equivalent sample size has become rather popular and is applied in time-series analysis (Lettenmaier, 1976; Matalas and Langbein, 1962; Zhou, 1996) as well as in spatial statistics, for instance in Gilbert's book on ecological monitoring (Gilbert, 1987). The formula is entirely correct, but if one looks at what happens with the variance of the mean when the sample size is increased, an odd behaviour can be noticed. Take as a simple example an equidistant time series with the exponential autocorrelation function $\rho(t) = \mathrm{e}^{-3t}$ (see Fig. 3.1). Furthermore, take both σ^2 and the monitoring period equal to 1, and increase the sample size by increasing the sampling frequency.

Using (3.3) and (3.4), respectively, for independent and dependent observations we obtain the variance of the estimated mean ($\hat{\mu}$) as a function of sample size, depicted in Fig. 3.2. This figure shows that, with independent observations, the variance decreases continuously with increasing sample size, however, with dependent observations the variance first drops, but not lower than a certain level, and after that it stays nearly constant. In other words, according to (3.4) one cannot reach a precision beyond a certain level no matter how many observations one takes. This counters the intuition that the larger the sample, the more one knows about the universe. The reason for this is not that (3.4) is incorrect, but that it is intended for estimating the model mean, not the integral mean, i.e., the average of z over the universe of interest:

$$\bar{z} = \frac{1}{|\mathcal{U}|}\int_{u\in\mathcal{U}} z \, \mathrm{d}u \ . \quad (3.6)$$

If the integral mean (spatial, temporal or spatio-temporal) is to be estimated or predicted, then (3.5) is not applicable and a different formula should be applied, which depends on how the sample will be taken. Following the design-based approach, with some form of random sampling to estimate the integral mean, the correct sample size formula depends on the chosen type of sampling design (see Sect. 7.2). For instance, if Simple Random Sampling is chosen as type of design, then (7.8) or (7.9) should be applied.

Using the model-based approach, the variance of the prediction error of the Best Linear Unbiased Predictor of the integral mean is given by (2.16), which can be used to calculate the required sample size via some optimization routine (see Sect. 7.3 and Appendix A).

It follows from the above example that it is important to choose the target quantity carefully. Just 'the mean' is not enough; the kind of mean is what

counts for the sample size. We expect that for surveys the integral mean rather than the model mean would usually be relevant, because it reflects directly the actual state of the universe. The same applies for status and compliance monitoring. For effect and trend monitoring, on the other hand, the model mean may be more relevant.

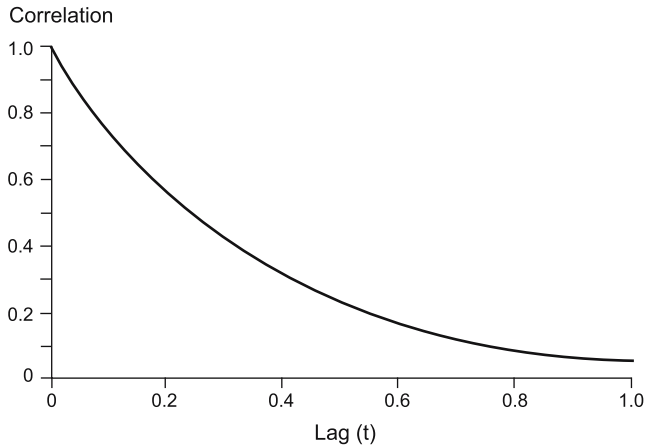


Fig. 3.1. Autocorrelation function used to calculate the variance of the estimated mean in Fig. 3.2

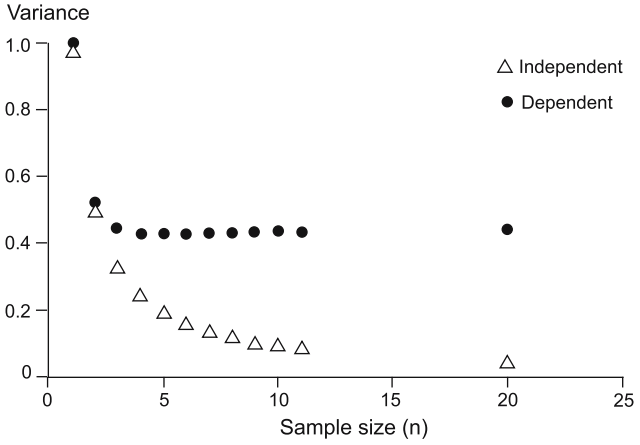


Fig. 3.2. Variance of the estimated mean as a function of sample size, for independent and dependent observations (see text)

4

Major Design Decisions

4.1 Choice Between Design-Based and Model-Based Inference

Before deciding on the details of a sampling design, a choice should be made between design-based and model-based inference, because design-based inference requires probability sampling, whereas for model-based inference probability sampling generally is sub-optimal. It is beyond the scope of this book to discuss this issue in detail, so only an outline is given here. An extensive discussion is presented in Brus and de Gruijter (1997). The 'ideal' circumstances for application of the design-based approach are as follows:

i. The required result is an estimate of the frequency distribution of the target variable in the universe as a whole, or a parameter of this distribution, such as the mean, the standard deviation or a quantile.
ii. A minimum sample size of, say, 5 or 10 units can be afforded, depending on the variation in the universe, to have at least a rough idea of the sampling error.
iii. It is practically feasible to sample at randomly selected locations and/or times.
iv. It is important to obtain an estimate that is unbiased in the sense that, averaged over all possible samples of the applied design, the estimate equals the true value of the target parameter.
v. It is important to obtain an objective assessment of the uncertainty of the estimate.

Around this 'ideal' there is a range of circumstances in which the design-based approach is still preferable to the model-based approach.

The 'ideal' circumstances for application of the model-based approach are as follows:

i. The required result is a prediction of values at individual points in the universe, as with forecasting, or a distribution of values over the entire universe, as with mapping.
ii. At least a medium sample size can be afforded, depending on the spatial and temporal variation. The model usually implies stationarity assumptions and a variogram, which should be estimated from about 100 to 150 sampling locations (Webster and Oliver, 1992).
iii. A reliable model of the variation is available.
iv. Strong autocorrelations exist in the universe.

As before, around this 'ideal' there is a range of circumstances in which the model-based approach is still preferable to the design-based approach. A typical intermediate situation is where averages are required for a number of sub-universes or 'blocks', in which only sparse sampling can be done. Brus and de Gruijter (1997) explored this in a case study. Also, prior data of non-probability samples favour the use of a model-based strategy, although there are various possibilities of using the prior data at the sampling or the inference stage of a design-based strategy (Brus and de Gruijter, 2003). A more extensive discussion on the choice between design-based and model-based methods in the case of sampling in space is given in Chap. 6.

4.2 Choice of Sample Support

As discussed in Sect. 1.3, the universe of interest is conceptualized as a set of sampling units and these units can be selected for measurement, either in situ or ex situ. The geometry of the sampling units, i.e., their shape, size and orientation within the universe, is referred to as 'sample support'. The sample support may be constant or variable. Of course, with sampling in time, the concept of sample support simplifies to points in time or periods, and similarly for sampling in 1D space.

There is not always a choice with respect to sample support. The universe may consist of discrete objects of a certain kind, such as the trees in a forest stand, the fields in an administrative unit, or the rivers of a continent. In that case we may speak of a discontinuous universe. In such cases the objects serve as sampling units and measurements are taken from them, because the target quantity is defined in terms of one or more properties of the objects, e.g., the average height of the trees in a forest stand. In many situations, however, the target quantity is defined in terms of one or more properties of a continuum, such as the soil in a field or the water in a lake. Then we speak of a continuous universe.

The distinction between discontinuous and continuous universes is not always as clear-cut as one may think. It should be realized that continuity or discontinuity of a universe is not an inherent property of reality, rather it is determined by the purpose of the survey or monitoring, in particular the target

quantity. Consider the above example of a forest stand. If the target quantity is defined in terms of a property of the trees, such as the average height, then indeed the universe is the set of trees in the stand and the sampling units will be trees. However, if the target quantity relates to a terrain attribute like the average number of trees per ha, then the universe is the area occupied by the stand and the sampling units will be areal units or 'plots', also referred to as quadrats. Finally, if the total timber volume of the stand is the target quantity, then both plots and trees can be used as sampling units.

In the case of a continuous universe, one is in principle free to define the sampling units as seems fit, and to choose the sample support accordingly. This choice is a complex problem and to our knowledge there are no simple rules that will always lead to an optimal solution. Instead we put forward a number of considerations that may be relevant in making a choice. Finally, in surveys of plants or animals, the probability that an individual is detected typically decreases with distance from the observer (Sect. 7.2.13). In this case the support of the sampling units is not clearly defined, and it makes more sense to specify the detectability function.

Effects on the Informativeness of Sample Data

The choice of sample support is rather fundamental since it defines the cumulative distribution function of the target variable in the universe. In case of a *quantitative* target variable, the larger the support, the more small-scale variation averages out within the sampling units. This will generally shorten the range of values within the universe and make the cumulative distribution function steeper. This in turn will affect parameters of that distribution (e.g., quantiles), except for the mean and the total. Therefore, presenting an estimate of (parameters of) the without specification of the sample support is useless.

In case of a *qualitative* target variable, increasing the support size has similar effects. The larger the support, the more small-scale variation averages out within the sampling units and the more they will be similar in terms of composition. This will generally affect measures of diversity, while in the special case of a presence/absence indicator the frequency of 'present' will increase. (Note that at this point we consider true values, which we would obtain by error-free observation.)

Given the potential impact on the informativeness of the sample data, the first question to be answered is: is there a unique 'ideal' sample support resulting in data relevant to the purpose and if so, what is it? The answer should be given by how the target quantity is defined. For instance, if the purpose is to estimate a mean per location in space, such as the mean cadmium concentration of the soil in a given region, then there is no such 'ideal' support. Any support will give data from which this mean can be estimated. In case of variable supports, the sizes of the supports must be known in order to weight the sample data.

On the other hand, if in a study on cadmium uptake by wheat plants the interest lies in the cumulative distribution function of soil cadmium in a field, then an ideal support may exist, e.g., the size and shape of the soil bodies as usually rooted by individual wheat plants. A larger support would give a cumulative distribution that is too steep and a smaller support would give one that is too flat.

The purpose may also set an upper bound on the sample support: the support size should not exceed the size of the domain(s) of interest (Splitstone, 2001). For instance, if a decision is to be made on the remediation of polluted soil for blocks of $50 \times 50 \times 1$ m^3 , then the sample support should not be larger than this block. A lower bound on the support size may be implied by the nature of the target variable. For instance, hydraulic conductivity and porosity in groundwater flow are only defined for volumes of some minimum size, referred to as the 'representative elementary volume' (Bear, 1972).

Finally, with large plot sizes, the problem that one cannot maintain the plot shape near the boundaries of the universe may become serious (Stevens and Urquhart, 2000).

Effects on the Efficiency

Increasing the support size will generally give more accurate results with the same sample size, if also the measurement-error variance would remain the same. This is because the larger the support, the more variation is averaged out by measuring the sampling units, and the smaller the variation between measurements. From this it follows that the support affects the precision of estimates of parameters of the cumulative distribution function, for instance the sampling variance of the estimated mean. In model-based prediction of local quantities, the prediction-error variance is also affected by the sample support. For instance, if the values at locations in space are predicted by kriging, the sample support affects the sill and nugget of the variogram and, as a consequence, the prediction-error variance (Zhang et al., 1999).

In practice, when the support size is increased, the accuracy of the measurements will generally deteriorate. Furthermore, the costs per sampling unit will rise, so that a smaller sample size will be available for the same budget. Therefore, as far as there is freedom of choice, one should try to identify the support that is most efficient, i.e., the support by which the best result is produced for the given budget, or by which the required quality is achieved for the lowest costs.

Composite Sampling to Measure Larger Sampling Units

When a sampling device is used to take aliquots for ex situ measurements, then in principle this device determines the support. For instance, when soil is sampled with a ring, the support is determined by the geometry of this ring. In theory the support size can be enlarged to any level by increasing the

diameter of the ring. However, in practice one will soon encounter limits of practicability.

This example brings us to a technique that can be used for ex situ measurements on large sampling units. Instead of increasing the volume of a single aliquot, one might also take more than one aliquot of the original volume and mix these to obtain a composite aliquot. If this is done properly, the measurement on the composite can be used as a measurement of the average of the unit from which the composite was taken. For instance, when $n \times n$ aliquots are taken on the nodes of a square grid with a grid distance d, then the sample support can be taken to be a square with a side of length $n \times d$. This technique assumes additivity of the target variable, an issue discussed in Sect. 4.3 on composite sampling.

In general, the larger the grid distance d, the larger the variance between the aliquots taken at the grid nodes, the more variation is averaged out by the compositing, and the smaller the variance between composites. However, reducing the variance between composites by increasing the support size is achieved at the expense of the precision of the 'measurements' of the sampling unit means. This is because the composite means differ from the sampling unit means and, given n, the mean squared difference between them generally increases with increasing grid distance.

The within-unit sampling error can be regarded as a pseudo-measurement error. Of course, given the support size, this sampling error can be reduced by taking more aliquots for the composite, but in general this is at the expense of a larger subsampling error, i.e., the error introduced by taking a subsample from the composite (after mixing) for analysis in laboratory.

We conclude that decisions on the sample support (in 2D space: the plot size), the number of aliquots used in a composite from a sampling unit (plot) and the sample size (the number of plots), should not be made separately, but simultaneously on the basis of prior knowledge of short-distance variation (or variation at short time lags) and variation in the universe, the subsampling error and the measurement error.

Implications of Measuring Technique

When a measuring device is used for in situ measurements, then in principle this device determines the sample support. However, just as composite sampling for *ex situ* measurements, multiple *in situ* measurements within sampling units can be done to increase the support size, simply by averaging the values measured at a number of points within the sampling unit. The same considerations apply as for composite sampling within sampling units, except that subsampling error does not play a role.

The measurement method for a given target variable may imply a lower bound of the support size. For instance, a minimum amount of soil may be needed for laboratory analyses, or a pH-electrode must be held in the water for a sufficiently long time to obtain a stable measurement.

Finally, when different variables must be measured on the same sampling unit, or the same variable repeatedly measured as in monitoring, special care is needed to avoid disturbance through the extraction of aliquots or otherwise destructive or disturbing measurements. To that order the support size should be chosen large enough to allow sufficient separation in space and/or time of the various sampling and measurement actions.

Different Target Quantities Requiring Different Sample Supports

A problem often arising in monitoring practice is when different target quantities are defined which ask for different sample supports. We give two examples.

The first example is found in the national soil quality monitoring network of the Netherlands. Here entire management units are taken as the sampling units, and as a result the areas of the sampling units differ considerably. This is not a problem if the aim is to estimate parameters of the cumulative distribution function of soil properties of management units (e.g., the proportion of management units in the Netherlands with an average phosphate content in the topsoil larger than a given threshold value). However, we do have a problem if the aim is to estimate parameters of the cumulative distribution function for a different, constant support, e.g., the areal fraction with a phosphate content larger than the threshold value, which is defined at point support. The use of data from variable supports in model-based (geostatistical) prediction is also problematic, for instance in estimating the model of spatial variation (variogram) for a given constant support from these data.

The second example may arise in sampling of 3D spatial universes. Suppose, for instance, the mean carbon content of the topsoil of an area is to be estimated. To this end, an auger core is collected by a single augering at each of a number of sampling locations. At all locations, the entire topsoil is sampled. Suppose further that the depth of the topsoil varies in space, so the depth of the augering also varies. If the auger cores are measured integrally, then unbiased estimation is only possible for the spatial mean carbon content per areal unit. It should be realized that this mean is generally different from the spatial mean per volume or mass unit. (Suppose that the mean content of shallow topsoils tends to be smaller than that of deep topsoils. In that case the spatial mean per areal unit is smaller than the spatial mean per volume or mass unit.) If an unbiased estimate of the mean per volume or mass unit is required, then one should measure the content at various depths separately.

4.3 Choices on Composite Sampling

Composite sampling is the technique of putting the materials of individual aliquots together and to mix and analyze the composite sample. As only the composite aliquots are analyzed, the number of analyses is greatly reduced. The technique is often used in soil sampling because of its great advantage

in saving laboratory costs. A vast amount of literature exists on this subject, both theoretical and applied, but a well-founded and generally applicable methodology of composite sampling does not seem to be available. Therefore, some general guidelines are given here.

The basic assumption in its most general form is that analysing a composite aliquot gives the same result as *analysing the individual aliquots* used to form the composite. Three applications can be mentioned here. The first application is where the interest lies in the presence or absence of, for instance, a species of soil microbe or a chemical substance. If the method used to determine its presence or absence has a detection limit that is low enough, then a composite aliquot could be analyzed instead of individual aliquots. This application is often referred to as group screening or group testing.

The second application, related to the first, is in the detection of 'hot spots', i.e., small areas with much higher values than in the rest of the area.

This application is discussed in detail in Sect. 8.3.5.

The third application is where interest lies in the average value of a quantitative variable, for instance, phosphate content in the topsoil. Here the assumption is that, apart from mixing, subsampling and measurement errors, analysing a composite aliquot gives the same result as *averaging the values measured on individual aliquots*. In other words: arithmetic averaging can be replaced by physical averaging. Of course, this assumes that averaging is meaningful, and that it is needed given the purpose of the project. We discuss these assumptions briefly.

Averaging of Values is Meaningful

This requires that the target *variable* is a quantitative attribute, which precludes composite sampling if the target variable is measured on a nominal or ordinal scale.

Averaging of Values is Needed

Taking a non-composite scheme as the point of departure, this assumption implies that, without compositing, the sample statistic for inference on the target *quantity* would be a function of one or more unweighted means of individual aliquot values. The simplest example of such a sample statistic in design-based sampling is the unweighted sample mean, used to estimate the global mean in the case of, for instance, Simple Random Sampling (Sect. 7.2.3), Systematic Random Sampling (Sect. 7.2.7), Systematic Unaligned Sampling or Markov Chain Sampling (Sect. 7.2.8). In these cases, all aliquots could in principle be combined into one composite. Other examples, involving multiple unweighted means, are the estimators used in Stratified Simple Random Sampling (Sect. 7.2.4), Two-Stage Random Sampling (Sect. 7.2.5) and Cluster Random Sampling (Sect. 7.2.6). In these cases, all aliquots belonging to the same stratum, primary unit or cluster could in principle be combined.

The present requirement precludes compositing when the purpose is, for instance, to estimate a quantile, because its estimator would be an order statistic of individual aliquot values, or to predict the value at an unsampled location or the global mean via kriging, because the predictors would be the weighted means of individual aliquot values.

At first sight, one might expect that the present requirement also precludes compositing when the purpose is to estimate the standard deviation between the elements of the universe, because the usual estimator of this quantity is not the unweighted sample mean or a function thereof. However, this problem can be circumvented by using multiple composite values rather than a single one. If one takes a simple random sample of size n, forms a composite aliquot, and repeats this process of sampling and bulking m times, the variation between the m composite values contains information about the standard deviation. Section 7.2.3 discusses how in this case the standard deviation can be estimated from the composite values.

Another instance where compositing would not seem to work is in locating 'hot spots', i.e., places with exceptionally high values. This kind of problem typically requires very large sample sizes and high laboratory costs when working with individual aliquots only. However, methods have been developed that allow a compromise in the sense that the individual aliquots are all divided into a portion used to form composites and a portion stored for possible later analysis. The idea is then to analyze all composites, and only those individual aliquots that relate to high composite values. See Sect. 8.3.5 for a discussion of these methods.

Arithmetic Averaging Can Be Replaced by Physical Averaging

In order to make this basic assumption valid, three requirements must be met. First, the target variable must be directly measured in the aliquots, or be defined as a linear transformation of one or more measured variables. Otherwise, if the target variable is a non-linear transformation of one or more measured variables, the transformation of the mean value(s) for a composite aliquot is not equal to the mean of the transformed values for individual aliquots. Neglecting this fact can lead to an unacceptable systematic error. Examples of a target variable defined as a non-linear transformation are: the indicator variable indicating whether or not the phosphate content in the topsoil exceeds a given threshold, the available soil moisture content calculated with a non-linear model from inputs measured at sampling locations, and pH as a logarithmic transformation of hydrogen ion activity.

Second, after the aliquots have been combined and mixed, no physical, chemical or biological interactions between the increments should take place that influence the value of the target variable. This precludes, for instance, compositing when the target variable depends on pH and some aliquots contain calcium carbonate while others do not. Also, many soil physical measuring techniques require aliquots to be undisturbed, which is usually compromised by compositing.

Third, compositing reduces laboratory costs, but it introduces two interrelated sources of error: error by imperfect mixing of the composites and error by subsampling the mixed composite. Also, random measurement errors will cancel out less well in the case of composite sampling than with non-composite sampling, because fewer measured values are averaged. The additional error due to compositing should not increase the total error too much, and this imposes a limit on the number of aliquots that can be bulked. The increase in the contribution of measurement error to the total error could be counteracted by taking multiple measurements from each composite while still preserving a cost advantage. See Brus et al. (1999) for an example. Also, if mixing and subsampling are important error sources, one could produce a number of smaller composites from random subsets of aliquots, instead of one large composite.

Some influential theoretical publications on composite sampling are Duncan (1962), Brown and Fisher (1972), Rohde (1976) and Elder et al. (1980). Boswell et al. (1996) provide an annotated bibliography. Papers on composite soil sampling have been presented by, e.g., Baker et al. (1981), Brus et al. (1999), Cameron et al. (1971), Carter and Lowe (1986), Courtin et al. (1983), Ruark et al. (1982), Reed and Rigney (1947), Webster and Burgess (1984) and Williams et al. (1989).

5

Optimization of Sample Selection

Decisions on sample selection are often taken on the basis of experiences in comparable cases, practical considerations, prescribed protocols or convention. This may be inevitable, and it may work satisfactorily. However, in principle one would like to optimize the selection of the sample in terms of costs and quality, employing the available information about the universe in question. This is especially true for large projects requiring relatively large investments, and when good prior information is available.

In this chapter we discuss how sample selection can be optimized in terms of costs and the quality of the result. Our discussion focuses on point sampling from a continuous universe and on two modes of optimization:

1. *Quality maximization*, under the constraint that the costs must not exceed a given budget;
2. *Cost minimization*, under the constraint that the quality of the result must meet a given minimum requirement.

As we have seen in previous chapters, the sample is not the only scheme item that determines costs and quality. The assessment method, sample support and bulking of aliquots are other items that affect costs and quality. Hence, the sample selection can only be optimized after decisions have been made (at least provisionally) on these other issues.

In the following sections we discuss various options with respect to quality measures as the basis of optimization, and approaches to the optimization process itself.

5.1 Quality Measures

Defining a quality measure is a prerequisite for ex-ante and ex-post evaluation of the quality of the result, as well as for optimization of the sample selection. As indicated in Sect. 3.2, this definition is usually the result of a 'translation' of an initial, more general formulation of the aim of the survey or monitoring

project. This translation generally leaves room for choice, and the designer should be aware of this, because this choice has direct consequences for the design process. We distinguish three types of quality measures.

Utility Measures

These are functions of the error distribution, which specify the expected economic losses due to given errors in the result. An error in the resulting data may cause a loss when it causes the user to make a sub-optimal or wrong decision. Hence, functions of this kind are derived from a detailed quantitative analysis of the consequences of errors. This involves analysing how the error in the results propagate to the outcome of the economic analysis. See Bie and Ulph (1972) for an early example of this approach in the context of landuse planning on the basis of an imperfect soilmap. In many cases such an uncertainty analysis is unfeasible, but if a realistic utility function can be defined, then this is to be preferred over statistical or geometric functions, because it offers a more accurate representation of the quality. This type of measure requires the availability of a stochastic model of the variation as prior information.

Statistical Measures

These are also functions of the error distribution, but they are generic and do not specify expected economic losses. Instead, they reflect the accuracy, precision or reliability of the result. Common measures for estimates and predictions are the standard error, the error variance and the half–width of confidence or prediction intervals. These measures are appropriate for more-or-less symmetrically distributed errors. For non-symmetrically distributed errors, and in case one is able to estimate the conditional probability density of these errors one could choose to minimize the entropy (Bueso et al., 1998).

For qualitative results, as produced by hypothesis testing or classification, the power of the test and error rates are common measures. Statistical measures are easier to apply than utility measures, but less closely related to the actual use of the result. As with utility measures, statistical measures need a stochastic model of the variation, either in the form of prior estimates of one or more variance components, or in the form of a geostatistical or time-series model.

Geometric Measures

These measures can be viewed as substitutes for statistical and utility measures. They can be used to facilitate the optimization of spatial sampling locations when the application of a statistical or utility measure is unfeasible, e.g., because a realistic model of the variation is not available. An example is the Mean Squared Shortest Distance, a measure used to create ‘spatial coverage’ samples (Sect. 8.3.3).

5.2 Approaches to Optimization

The aim of this section is to give a broad overview of approaches to optimization of sample selection. More details are given in various sections on sampling methods. Section 5.2.1 discusses optimization for a single target quantity. It appears that optimization in the design-based approach is quite different from optimization in the model-based approach. Section 5.2.2 discusses the more complicated case of optimization for multiple target quantities. Here, the choice of an optimization method will be largely governed by the reason why there are multiple target quantities.

5.2.1 Optimization for a Single Target Quantity

As explained below, optimization of sample selection is different in the design-based approach and the model-based approach (Sect. 2.2.1). Optimization in model-based sampling tries to find the best sampling pattern within the target universe. This works roughly as follows.

If the optimization mode is 'quality maximization', the affordable sample size is first determined from the given budget. An iterative search algorithm is then used to find the optimal sample of that size. To keep computation time within reasonable limits, the search is typically confined to some subset of possible positions, e.g., selections from a user-specified discretization grid. During the search, the chosen quality measure is evaluated for a large number of candidate samples, but the algorithm generally does not guarantee that a global optimum will be reached. Therefore the search should be repeated with different initial solutions.

If the optimization mode is 'costs minimization', the problem is more complex because now the best combination of both sample size *and* sampling pattern has to be found. A practical approach is to conduct a 'quality maximization' for each of a series of eligible sample sizes, and to retain the combination with the smallest sample size that still meets the given quality requirement.

Optimization in design-based sampling is different from that in model-based sampling, because design-based inference implies that a probability sample is drawn, i.e., the sampling pattern is stochastic and cannot be optimized as such. However, the randomized selection takes place within certain randomization restrictions, which characterize the sampling design and vary in nature and complexity. Thus, optimization in design-based sampling tries to find the best sampling design rather than the best sampling pattern.

In order to explain the optimization of sampling designs, we must have a closer look at the randomization restrictions that they impose. A very basic kind of restriction is imposed by the design 'Simple Random Sampling (Sect. 7.2.3) with sample size n', which is that all samples of a size other than n are excluded from selection. A slightly more complex restriction is imposed by the design 'Stratified Simple Random Sampling (Sect. 7.2.4) with sam-

ple sizes $n_1, \ldots, n_k$ in the strata $\mathcal{U}_1, \ldots, \mathcal{U}_k$', allowing only samples with the pre-specified sample sizes in the strata.

An even more complex combination of restrictions is imposed, for instance, by 'Cluster Random Sampling (Sect. 7.2.6) by n random transects in a north-south direction and equidistant sampling locations d metres apart'. While these restrictions determine the number and direction of the transects and the inter-point distance, the only randomness left is in the starting points of the transects. These examples illustrate that randomization restrictions can be quantitative (e.g., numbers, distances) or qualitative (e.g., type of cluster, shape of strata).

Optimization of qualitative randomization restrictions is computationally more cumbersome in general, so in practice optimization will often be limited to quantitative restrictions, given a provisional decision on possible qualitative restrictions. The quantitative restrictions are related to total sample size, sample size per stratum, number of primary and secondary units (as in Two-Stage Random Sampling (Sect. 7.2.5)) and number of clusters.

Optimization methods for these parameters work with a stochastic model of the variation, as with optimization in model-based sampling, according to a utility or statistical measure. The use of a model in the design-based approach may be confusing at first sight. Keep in mind, however, that the model is only used to optimize the sampling design, not for inference from the sample data.

What was said above about costs minimization versus quality maximization in model-based sampling applies to design-based sampling as well. Specific details on optimization in design-based sampling in space are presented in Sect. 7.2, along with the various types of random sampling designs.

5.2.2 Optimization for Multiple Target Quantities

A predominant factor which complicates optimization is that survey and monitoring often have more than one purpose, in the sense that more than one target quantity is defined. This may be because there is more than one domain, or more than one target variable, or more than one target parameter. Application of a single quality measure to multiple target quantities yields multiple qualities (or a multivariate quality), and the outline given above is then no longer directly applicable.

The choice of a suitable approach to optimization with multiple qualities depends on the reason why there are multiple qualities. Therefore we discuss these situations separately.

More Than One Domain

We discuss two cases, one where the domains have extensions in space and/or time, and one where the domains are prediction points.

Non-Point Domains

In this case, the universe is divided into a number of disjoint parts for which separate results are required.

First let us consider costs minimization. This optimization mode assumes that a quality requirement is given for each domain individually, e.g., the error variance for each domain must be smaller than a user-specified maximum. If the overall costs can be reasonably approximated by the sum of the costs per domain, and if sampling and inference is done in such a way that the qualities of the results for the domains are approximately independent of each other, then *overall* costs minimization can simply be done by costs minimization *per domain*, along the line indicated for the case of a single target quantity.

In the design-based approach, the condition of quality independence between domains can be met by using the domains as strata and sampling them independently (Sect. 7.2.4). In the model-based approach, this independence can be introduced by assuming – if only to simplify the optimization – that the inference will be limited to the sample data from within each domain. This may not be unreasonable, because it leads to conservative predictions of the qualities if, in reality, data from other domains are also used.

For quality maximization, given a total budget for all domains together, there are two options. The first option is to assign weights to the qualities in the domains, and then to define a single overall quality measure as the *weighted average* of the qualities per domain (see Sect. 8.2.2). The other option is to define the overall quality measure as the *minimum* of the qualities per domain. Whatever the definition, this single overall quality can in principle be maximized through a search algorithm dividing the total budget among the domains.

Point Domains

A common instance is a regular grid of prediction points created for the construction of a map by means of model-based inference (e.g., kriging, Sect. 8.3).

Quality maximization, given a total budget and a maximum sample size derived from it, also offers the same two options as in the case of domains with an extension. One option is to assign weights to the qualities at the prediction points, and then to define a single overall quality measure as the *weighted average* of the qualities per point. The other option is to define the overall quality measure as the *minimum* of the qualities per point. However it is defined, this single overall quality can be maximized through a search algorithm such as simulated annealing (Sect. 8.3.4).

Costs minimization assumes that a quality requirement is given for each prediction point, and must be implemented over the combination of sample size and sampling pattern. The difference with costs minimization in the case of domains with an extension is that here the assumption of quality independence is untenable; a change in sample size or pattern will affect the qualities

at various prediction points simultaneously. This makes optimization more complicated.

Similar to the case with a single target quantity (Sect. 5.2.1), a practical approach would be to optimize[1] the pattern for each of a series of eligible sample sizes, and to retain the combination with the smallest sample size that still meets the given quality requirements for the prediction points.

More than one target variable

This situation occurs when multiple target variables are to be determined at the same sampling events. We briefly discuss three typical cases.

Spatial Cumulative Distribution Function (SCDF)

In order to estimate or predict the SCDF of some variable, a series of indicator variables is defined that correspond with increasing threshold values of the variable. The indicator variables are treated as the target variables. The spatial means of the indicators are interpreted as fractions, which are the target quantities.

In principle, there are as many qualities as there are threshold values, but in this case it may be appropriate to reduce these to a single quality on prior grounds. For instance, the fraction closest to 0 or 1 will have the smallest relative precision, and this may be selected as the most *critical* single quality, on the basis of which either quality maximization or costs minimization is performed.

Spatial Coverages of Plant Species

In the context of survey of vegetation, information may be required about the spatial coverages of different plant species. The presence or absence of the species is represented by indicator variables which, once again, are treated as the target variables, and their spatial means as target quantities.

Just as in the previous case, there are as many qualities as there are plant species, and to simplify the optimization one might select a single quality on prior grounds. In this case, however, it may be more appropriate to select the quality for the ecologically most *relevant* species as the one for which either quality maximization or costs minimization is performed.

The underlying assumption of this approach is that, while the sample selection is optimized for a single species, the results for the other species will still be satisfactory. If this assumption is deemed unrealistic, it may be necessary to have recourse to the approach outlined for the case of prediction points as domains.

[1] The pattern could be optimized by simulated annealing (Sect. 8.3.4 and Appendix A) with a penalty function accounting for differences between the expected qualities and the required qualities

Concentrations of Pollutants

In the context of environmental monitoring, information is often required about the concentrations of a set of pollutants. The concentration of each pollutant is treated as a target variable, and the spatio-temporal means of these concentrations are the target quantities. This case resembles the previous one as far as optimization is concerned, and sample selection may be optimized in the same way.

There may be one difference, however, namely when the option is available not to measure all concentrations at all sampling events. This is the case, for instance, if the aliquots taken on the sampling events can be analyzed for varying sets of pollutants. It may then be cost-efficient to limit the more expensive analyses to a subsample and to use possible correlations in the inference. This can be done by Two-Phase Random Sampling and regression estimators (Sect. 7.2.11).

More than one target parameter

A typical example is when an estimate is required of both the mean and the standard deviation of a quantitative target variable. Exact optimization in such a case would generally require stochastic simulation.

A rough approach to costs minimization, given a quality requirement for each parameter, would be to assume a parametric distribution on the basis of the available prior information, e.g., a normal distribution with a prior estimate of the standard deviation. Assuming Simple Random Sampling and using standard formulae one can predict the precision of estimates of both parameters for each of a series of eligible sample sizes. The smallest sample size that still satisfies the quality requirements is then chosen. This leaves a safety margin if a more efficient sampling design is applied, as will normally be the case.

Part II

SAMPLING IN SPACE

6

Introduction to Sampling in Space

As regards 'space', it is assumed that in the structured approach to designing survey and monitoring schemes (Chap. 3.2), the design information specifies that the universe of interest is purely spatial, i.e., no time dimension is involved. This part therefore deals with the situation in which a once-only survey can deliver the required information. Of course, the methods presented here can be applied more than once in the same area. That would, however, constitute a form of monitoring, the implications of which are dealt with in Part IV 'Sampling in Space–Time'.

Sampling for survey of natural resources can be done in 1D, 2D or 3D space. Although the spatial universe of interest is often a three-dimensional body, sampling is mostly carried out in the horizontal plane, i.e., in 2D space, so that the sampling locations have only two coordinates. Therefore we present the methods in terms of 2D sampling; for instance, we will use the term 'area' rather than 'spatial universe'. Sampling in 1D or 3D space is discussed separately, in Sect. 7.2.16.

In Sect. 4.1 we already stressed the importance of the choice between design-based or model-based inference because design-based inference requires probability sampling, whereas for model-based inference non-probability sampling is most appropriate. We shall discuss the pros and cons of these two approaches to sampling and inference now in more detail in the context of spatial survey. Broadly speaking, the suitability of design-based methods, relative to model-based methods, is greatest for global quantities such as the spatial mean, and diminishes as one moves to smaller and smaller sub-areas, and finally to estimation at specific points. Neither of the two approaches has a monopoly, not even at the extremes of the spatial resolution continuum, viz. the area as a whole and individual point locations. This broad picture is illustrated in Fig. 6.1. It should be noted that the relative suitability functions depicted in this figure only reflect our global expectations of suitabilities, 'averaged' over a broad class of different cases that could be encountered in practice.

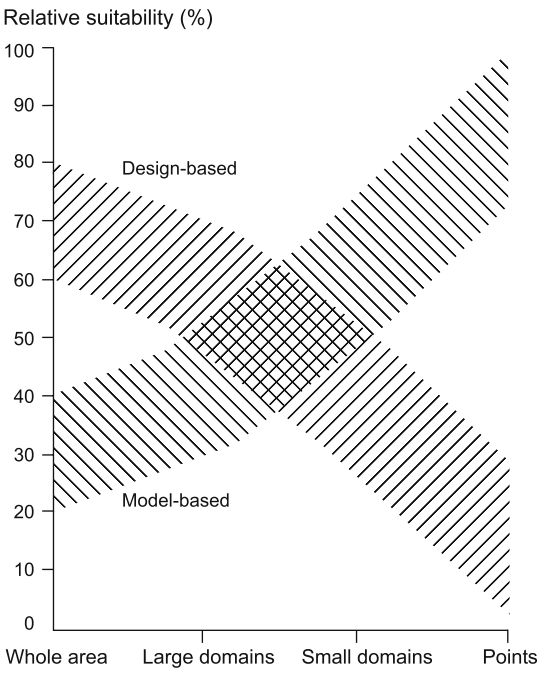


Fig. 6.1. Relative suitability of the design-based and model-based approaches to sampling, as a function of the spatial resolution at which estimates are required

In reality there are more factors that determine the suitability than spatial resolution alone. These factors relate to the following questions.

1. Should the estimation or the test of the global quantity be 'design-unbiased', i.e., correct on average over repetitions of the sampling process using the same sampling design? Design-unbiasedness can be regarded as a strict guarantee against bias in sampling, such as may arise in convenience or purposive sampling. If this guarantee is required, then a design-based method is the only option.
2. Should the accuracy of the estimate or the power of the test be quantified objectively, i.e., without recourse to assumptions on the spatial variation? A positive answer to this question rules out model-based methods.
3. Is random sampling in the field practically feasible? If not, then some form of convenience or purposive sampling combined with model-based inference is the obvious choice.
4. Is a reliable model of the spatial variation available? Only if this is the case, can the model-based approach be sensibly applied.

5. Do substantial spatial autocorrelations exist between sampling locations and prediction locations? If not, then the computational effort involved in model-based inference will be fruitless.
6. Is composite sampling acceptable and would it reduce costs significantly? The conditions under which composite sampling is acceptable are discussed in Sect. 4.3. In principle, costs can be significantly reduced by compositing if laboratory analyses of individual aliquots would consume a considerable portion of the total budget. If composite sampling is an attractive option, this could be a reason to prefer a design-based method over a model-based one. The reason is that design-based methods allow compositing of aliquots taken at large mutual distances, possibly across the entire area, whereas with model-based methods, compositing is in practice always limited to aliquots from within small neighbourhoods. In general, compositing of aliquots that are wider apart yields a greater reduction of sampling variances, hence greater precision of the final estimates.
7. Are multiple realizations of a random field needed for the inference about the target quantity? Such realizations are to generated by simulation with a stochastic model of the variation, hence a model-based method must be used. A condition that makes simulation inevitable is when the target quantity is a nonlinear function of multiple values of the target variable. This is the case, for instance, with detection problems (the target quantity being the maximum of a 0/1 indicator variable), and with target quantities defined by neighbourhood operations, such as the (surface) area of a watershed.

As there are several misconceptions in the literature on this issue, we repeat from Sect. 4.1 that the design-based methods presented in Sect. 7.2 are valid, regardless of the structure of the spatial variation, because they do not make any assumption about that structure.

A typical application of design-based sampling strategies is to estimate the areal mean of a directly measured quantitative variable. However, the scope of these strategies is much wider than this, and can be expanded in three directions: derived variables, other parameters and smaller areas or sub-areas.

First, the target variable need neither be quantitative, nor directly measured. If the target variable is measured on a nominal or ordinal scale, the sample data consist of class labels, and these can be analyzed statistically by first transforming them into 0/1 indicator variables. The presence and absence of a given class are thereby re-coded as 1 and 0, respectively. Of course, if there are k mutually exclusive classes, only $k-1$ indicator variables are needed. The mean of an indicator variable can be interpreted as the fraction of the area in which the class occurs. Transformation into indicator variables can also be applied to quantitative variables in order to estimate the areal fraction in which the variable exceeds a given threshold. This technique can be extended to estimate the entire Spatial Cumulative Distribution Function

(SCDF) of a quantitative variable. In that case, areal fractions are estimated for a series of threshold values.

Apart from the simple 0/1 transformations, the target variable may be the output of a more or less complicated model for which the input data are collected at the sampling locations. Another important case of indirect determination is in validation studies, where the target variable represents an error, i.e., the difference between a measured value and a value predicted by a process model or a spatial distribution model, such as a thematic map. A common example is the error resulting from a classification algorithm applied to remotely sensed images. The errors determined at the sampling locations can be used to estimate their spatial mean (which equals the bias), the mean absolute error, the mean squared error or the entire SCDF of the errors.

Second, the target parameter does not need be the spatial mean. For instance, it may also be a quantile, such as the 90^{th} percentile, the spatial variance, a tolerance interval or a parameter of a model relating one or more predictor variables to a variable of interest. See Krishnaiah and Rao (1988) and Patil and Rao (1994) for design-based statistical inference on these and other target parameters.

Third, the region for which estimation or hypothesis testing is required need not be the entire area sampled; interest may also focus on one or more sub-areas, or in estimation at points. This subject is dealt with in Sect. 8.2.

Traditionally, the design-based method focused on discrete populations, and therefore representation of the universe is discrete in this approach. For instance, the mean is defined as an average over all N population elements. In this book we adhere to this usage in most cases, even when the universe is continuous. The continuous universe is first discretized by a fine grid of which the nodes represent the possible sampling locations. These methods are thus presented in a finite population mode, whereby the size of the universe is a dimensionless quantity (the number of nodes). In the model-based approach, on the other hand, the universe consists of an infinite number of possible sampling locations, and its size is measured in units of length, (surface) area or volume.

6.1 Contents

This part is divided into three chapters, according to what the aim of the sampling is: sampling for *global* quantities in space (Chap. 7), for *local* quantities in space (Chap. 8), or for *variograms* to model the spatial variation (Chap. 9). The chapters 7 and 8 form the main body of this part. They are each divided into a section on design-based methods and a section on model-based methods.

The section on *design-based* methods for *global* quantities (7.2) is the largest section of this part. It contains not only subsections on basic and advanced types of sampling designs and on how to choose from them (7.2.2–7.2.8), but also subsections on special sampling techniques like Probabilities-

Proportional-to-Size Sampling (7.2.9), Sequential Random Sampling (7.2.10), Line-Transect Random Sampling (7.2.13), and Line-Intercept Random Sampling (7.2.14). Two subsections deal explicitly with the use of ancillary information in sampling and in inference from sample data (7.2.11 and 7.2.12). Finally, there is a special subsection on model-based optimization of sample sizes for design-based sampling (7.2.15), and one on sampling in 1D or 3D space (7.2.16).

The section on *model-based* methods for *global* quantities (7.3) treats Centred Grid Sampling (7.3.2) and Geostatistical Sampling (7.3.3), i.e., optimizing the sampling pattern with the aid of a geostatistical model. The section on *model-based* methods for *local* quantities (8.3) in addition contains a subsection on Spatial Coverage Sampling (8.3.3). Both sections (7.3) and (8.3) contain a separate subsection on sampling of hot spots. Sampling for answering the question 'Is there a hot spot?' is treated in Sect. (7.3.4), while the question 'Where is the critical threshold exceeded?' is dealt with in Sect. 8.3.5).

The section on *design-based* methods for *local* quantities (8.2) deals with probability sampling for quantities defined on sub-areas (8.2.2) and for estimation of values at points (8.2.3).

Finally, Chap. 9 presents sampling and inference methods for variogram estimation. The sampling methods entail regular patterns (9.2) and optimized patterns (9.3). The inference methods are the method-of-moments (9.4.1) and maximum likelihood estimation (9.4.2).

7

Global Quantities in Space

7.1 Introduction to Methods for Global Quantities in Space

This chapter presents sampling and statistical inference methods leading to statistical results about global quantities in space. By 'global quantities' we mean the Spatial Cumulative Distribution Function(SCDF) of the target variable in the entire area, or quantities that can be derived from the SCDF, such as the mean, the median and other quantiles, the spatial variance, standard deviation and coefficient of variation, and tolerance intervals. Examples of global quantities in space are the mean phosphate content in the topsoil of an agricultural field, the areal fraction of a nature conservation area covered by some plant community, the SCDF of the nitrate concentration in the upper groundwater at the start of the growing season in an agricultural area, etc.

This chapter focuses on methods for the mean, which is the most relevant global quantity in practice, and on the SCDF, the basic function from which all other global quantities can be derived. In this context, areal fractions and percentages may be regarded as special cases of the mean. The *spatial* mean should not be confused with the concept of *model* mean, which is a model parameter representing the statistical expectation over realizations from a stochastic process, see Sect. 3.7. Estimation of the mean μ of a geostatistical model of the spatial variation is briefly dealt with in Sect. 2.2.

An important choice to be made is that of methodology: should one follow the design-based or the model-based approach? As this choice has important consequences for sampling and inference, this chapter is divided into a section on design-based methods (Sect. 7.2) and one on model-based methods (Sect. 7.3). The difference between these two approaches is explained in Sect. 2.2.1 A general discussion on the choice between them is given in Sect. 4.1. For sampling in space the issue is discussed in more detail in Chap. 6, from which we repeat that design-based methods are generally more suitable for global quantities than model-based methods.

Estimation of global quantities is the 'homeland' of the design-based approach to sampling, and adaptation of these methods from the general non-spatial context in which they were originally developed to the spatial context is straightforward. This is precisely the reason why most of the design-based methodology is presented in this chapter.

7.2 Design-Based Methods for Global Quantities in Space

7.2.1 Introduction

Basic Concepts of Design-Based Sampling

This section discusses how design-based sampling strategies work and how they can be applied in surveys. The aim is to promote understanding of the basic principles at an intuitive level, i.e., the section is not meant as an exposé of sampling theory. A somewhat practically oriented handbook on design-based sampling strategies is Cochran (1977), from which most of the discussion presented here was derived. A comprehensive reference work on sampling theory is Särndal et al. (1992).

The general pattern in the development of sampling strategies is to take the simplest random selection method (Simple Random Sampling, Sect. 7.2.3) as a starting point, with complete random selection of all sampling locations. This is followed by the identification of restrictions on randomization which reduce the sampling variance or the operational costs, or both. Different types of restrictions can be distinguished, each giving rise to a different type of sampling design.

Before discussing the various types of design, the statistical concept of *sampling design* itself needs to be defined more precisely. It is formally defined as a function that assigns a probability of selection to any set or sequence of sampling units in the universe. If units are drawn 'without replacement', a unit can occur only once in the sample, resulting in a random *set*. If they are drawn 'with replacement', a unit can occur more than once in the sample, yielding a random *sequence*. For instance, the sampling design for Simple Random Sampling, with point support, drawing without replacement, and sample size 25, assigns equal selection probabilities to every possible *set* of 25 points in the area and zero probability to any set with less or more than 25 points. Note that a design assigns probabilities to sets or sequences of units, not to individual units.

This book focuses on point sampling from continuous universes in space and/or time. The target universe therefore theoretically represents an infinitely large population of possible sampling locations, and the distinction between sampling with replacement and sampling without replacement is immaterial. Even if the sample is drawn from a fine discretization grid, the number of possible sampling locations should still be very large compared with the

sample size. To avoid undue complexity, we formally assume that point samples are drawn without replacement, thus considering them as sets. Although in principle non-replacement causes correlations between observations, these are negligible in this context, as are finite population corrections.

A *sampling strategy* is defined as a combination (p, t) of a sampling design (p) and an estimator (t) for a given target parameter (T), such as the mean of the area. For instance, Simple Random Sampling with sample size (n) 25 and the unweighted sample mean as an estimator is a sampling strategy for the areal mean. Statistical quality measures, like bias and variance, can only be defined and evaluated for combinations of design and estimator, not for a design or an estimator on its own.

In the example given, n=25 is a design attribute assigned to the selection method of Simple Random Sampling. Another example of a design attribute is the particular stratification applied in Stratified Simple Random Sampling. A selection method without specification of the relevant attributes is referred to below as a *design type*. Within each design type, numerous designs are possible by varying the attributes. Thus, Simple Random Sampling is an example of a design type, and another example is Stratified Simple Random Sampling. Two attributes have to be assigned to the latter type: stratification (a division of the area into sub-areas) and the sample sizes in each of the strata.

Before discussing how to choose a sampling strategy in Sect. 7.2.2, we give an overview of the design types that are most relevant to survey.

Overview of Sampling Design Types

The design types which we present in this section can be divided into five main groups:

1. basic design types;
2. compound design types;
3. spatial design types;
4. two-phase design types;
5. sequential design types.

Basic design types are:

- Simple Random Sampling (Sect. 7.2.3): locations are drawn from the universe, at random and mutually independent.
- Stratified Simple Random Sampling (Sect. 7.2.4): the locations in the universe are divided into groups, here called 'strata', and Simple Random Sampling is applied to each stratum.
- Two-Stage Random Sampling (Sect. 7.2.5): the locations in the universe are divided into groups, here called 'primary units' (PUs), a random sample of PUs is drawn in the first stage, and Simple Random Sampling is applied to each selected PU.

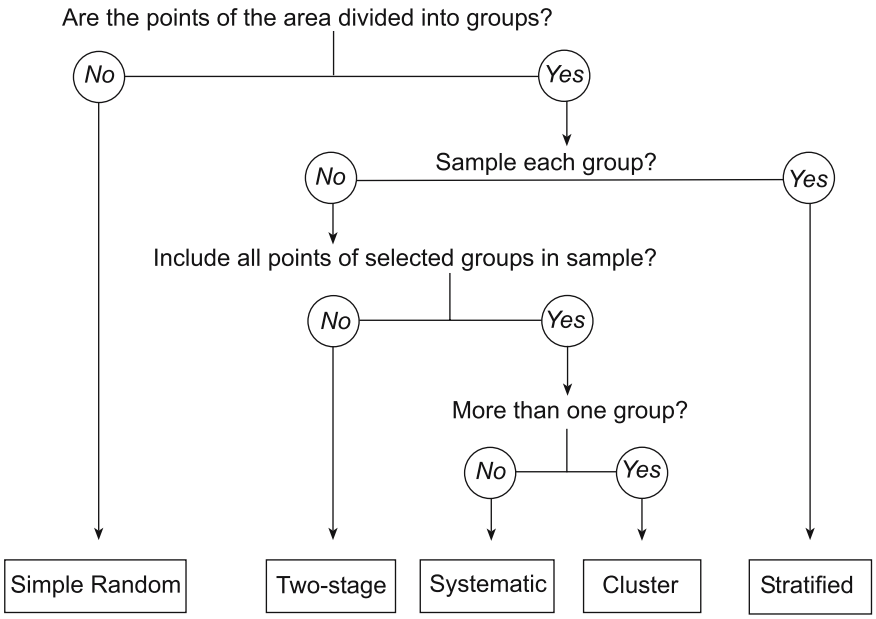


Fig. 7.1. Similarities and differences between the basic design types

- Cluster Random Sampling (Sect. 7.2.6): the locations in the universe are divided into groups, here called 'clusters', after which a number of clusters (> 1) are randomly selected, and all locations of the selected clusters are included in the sample. In the spatial context, the clusters are typically defined so as to form regular spatial patterns, e.g., equidistant locations on a line.
- Systematic Random Sampling (Sect. 7.2.7): similar to Cluster Random Sampling, except that only one cluster is selected. Again, in the spatial context this single cluster is typically defined so as to form a regular spatial pattern of locations, e.g., a square grid.
- Probabilities-Proportional-to-Size (pps-)Sampling (Sect. 7.2.9): Sampling units are selected with probabilities proportional to their size or to an ancillary variable that is correlated with the target variable.

The differences and similarities between these types are illustrated by the logical tree in Fig. 7.1, except for pps-Sampling, which can be regarded as a variety of Simple Random Sampling with unequal instead of equal selection probabilities.

Compound design types (Sect. 7.2.8) are combinations or nested structures of basic design types; they represent more advanced sampling methods. For

instance, in the second stage of Two-Stage Random Sampling one could apply Cluster Random Sampling instead of Simple Random Sampling.

Spatial design types (Sect. 7.2.8) sample on the basis of the spatial coordinates of the possible sampling locations in the area. Sampling by a basic or compound type of design, on the other hand, could in principle be done from a list-type of sampling frame, with all possible sampling locations in any order, regardless of their spatial position. (Spatial varieties of Cluster Random Sampling and Systematic Random Sampling, with their regular point patterns, obviously use coordinates too, but only in defining the clusters). Like compound design types, spatial design types represent more advanced sampling methods.

Two-Phase design types (Sect. 7.2.12) are sampling methods exploiting the correlation between a cheap-to-measure ancillary variable and the target variable. In the first phase, a relatively large sample is taken, in which only the ancillary variable is measured. In the second phase, a subsample is taken from the large sample, and the target variable is measured only in this subsample. If the ancillary variable is quantitative, then a 'regression estimator' (see Sect. 7.2.11) is used to estimate means or fractions. In case of a qualitative ancillary variable one can use 'poststratification' (Sect. 7.2.11).

Sequential design types (Sect. 7.2.10) proceed by taking samples one-at-a-time or batch-by-batch, during which a statistic is calculated to determine whether or not to continue.

7.2.2 Choosing a Design-Based Strategy for Global Quantities in Space

As explained in the previous section, sampling strategies consist of three major components: design type, design attributes and estimator. With few exceptions, these components have to be chosen in this order, because the attributes depend on the type, and the estimator depends on both the type and the attributes (see Fig. 7.2). Only the first choice, that of the design type, is treated in this section; the other two are discussed in the various sections on the design types.

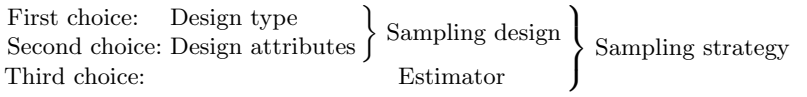


Fig. 7.2. Major choices in deciding on a design-based sampling strategy

Choosing a design type involves many aspects. In order to structure the design process in a manageable form, we have condensed and schematized these considerations in the decision tree presented in Figs. 7.3, 7.4, and 7.5.

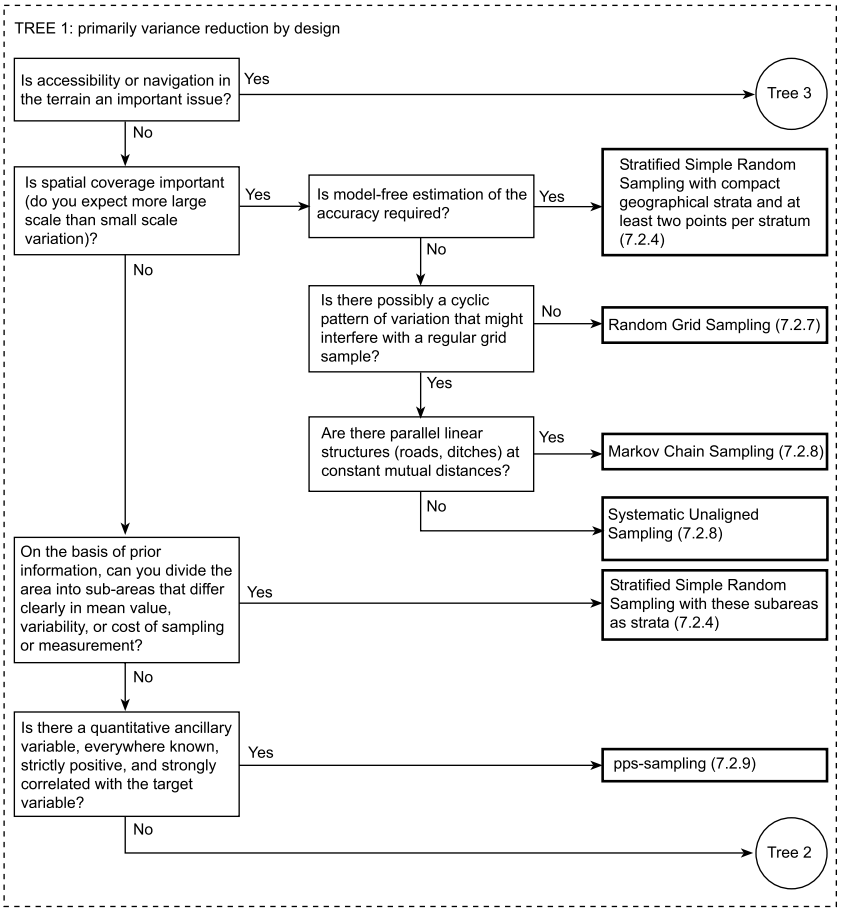


Fig. 7.3. Decision tree to aid the choice of a design type for global quantities in space

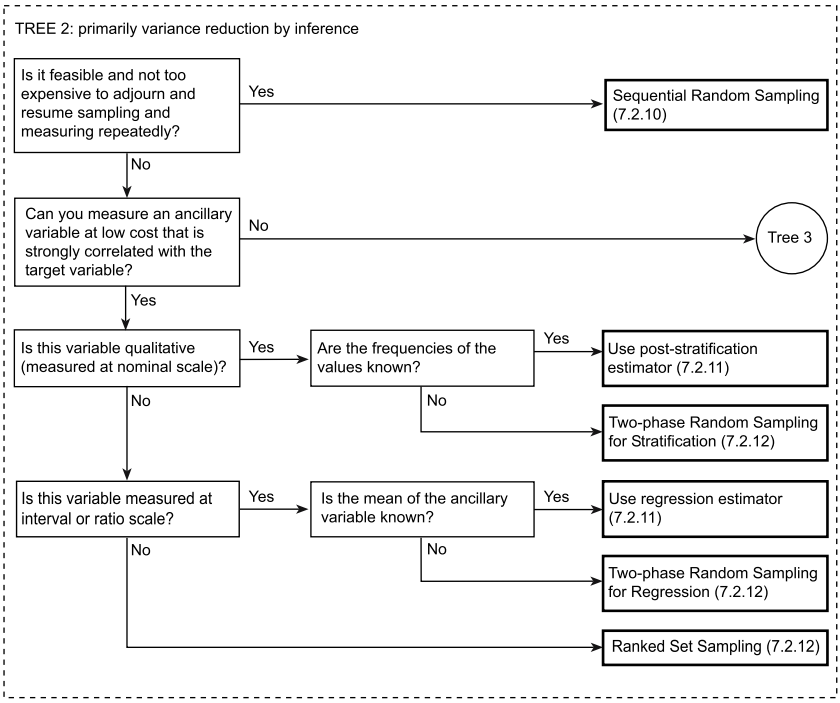


Fig. 7.4. Continuation of Fig. 7.3

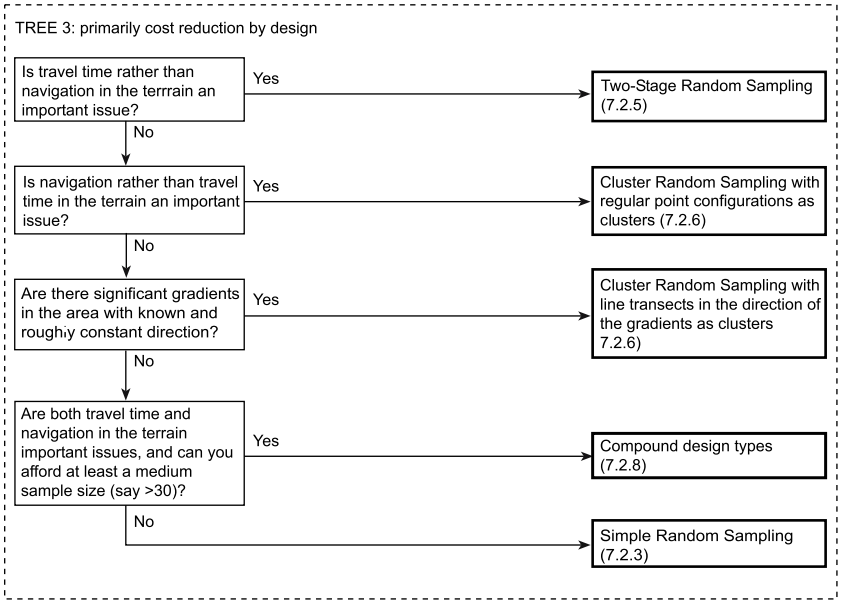


Fig. 7.5. Continuation of Fig. 7.3

Some comments must be made on this tree. Firstly, it is the result of many simplifications, so the outcomes should be interpreted as suggestions rather than inescapable conclusions. Second, we have kept some questions more or less vague because we believe that the underlying logical structure is inherently vague. Thirdly, while this tree assumes clear-cut answers (yes or no), these are often not unequivocally possible. If one has considerable doubt about the answer to a question, we suggest that both branches of the tree are followed.

For surveys with several target variables the answers to the questions can be conflicting. We repeat here our recommendation of Sect. 3.3 to avoid complex sampling designs for surveys with multiple target variables. For instance, stratified random sampling may be efficient for surveys of a single target variable, but inefficient for surveys of several target variables.

Although the decision tree has been developed for survey sampling, it can in principle also be used for choosing a spatial design in monitoring. However, in the case of monitoring two considerations follow from the need for flexibility as argued in Sect. 3.6. First, the spatial sampling design should be kept simple, because simple designs can be adapted most easily to changing conditions during the monitoring (see Overton and Stehman (1996) for a detailed discussion of this issue). Second, designs leading to good spatial coverage are often preferable, because the sub-areas for which separate results are required may change during the monitoring.

7.2.3 Simple Random Sampling

Restriction on Random Selection

No restrictions on random selection are imposed other than that the sample size is fixed and chosen beforehand. All sampling locations are selected with equal probability and independently from each other. The sample size is the only attribute to choose for this type of design; see 'Sample Size for Estimation' or 'Sample Size for Hypothesis Testing'.

Selection Technique

The following algorithm for Simple Random Sampling with sample size n is applicable to irregularly shaped areas.

(a) Determine the minimum and maximum s_1 and s_2 coordinates of the area: $s_{1,\min}$, $s_{1,\max}$, $s_{2,\min}$ and $s_{2,\max}$.
(b) Generate two independent (pseudo-)random coordinates, $s_{1,\mathrm{ran}}$ and $s_{2,\mathrm{ran}}$, from the uniform distribution on the interval ($s_{1,\min}$, $s_{1,\max}$) and ($s_{2,\min}$, $s_{2,\max}$), respectively.
(c) Use a point-in-polygon routine to determine whether the location ($s_{1,\mathrm{ran}}$, $s_{2,\mathrm{ran}}$) falls within the area. Accept the location if it does; skip the location if it does not.

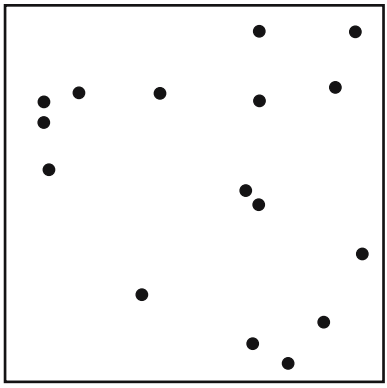


Fig. 7.6. Notional example of a simple random sample

(d) Repeat steps (b) and (c) until n locations have been selected.

Step (c) assumes that a perfect sampling frame is available in the form of a GIS representation. In practice, however, imperfections of the frame may appear after inspection of the projected sampling locations in the field. If projected sampling locations are for some reason judged not to belong to the target universe, then shifting to nearby locations would cause bias, as boundary zones become oversampled. The correct remedy is to delete such locations from the sample and to replace them by sampling locations from a reserve list in pre-determined order.

Example

Figures 2.1**a**, **b** and **c** show three realizations of Simple Random Sampling with 25 locations; Fig. 7.6 is an example with $n = 16$. Notice the irregularity, the clustering and the empty spaces between the sampling locations, which are typical of Simple Random Sampling.

Advantages and Disadvantages

The simplicity of this type of design enables relatively simple and straightforward statistical analyses of the sample data, even for non-standard estimation and testing problems. One disadvantage is that the sampling variance is usually larger than with most other types of design at the same costs, for two possible reasons: (*i*) spatial coverage by the sample may be poor, (*ii*) visiting sampling locations that are irregularly distributed may be more time-consuming for logistical reasons, and higher per-sample costs result in a smaller sample size. Another disadvantage is that estimation in domains (Sect. 8.2) may be impossible because large empty spaces may occur between the sampling locations.

Statistical Inference

The spatial mean of the area, $\bar{z}$, for a quantitative target variable z is estimated by:

$$\hat{\bar{z}}_{\mathrm{SI}} = \frac{1}{n} \sum_{i=1}^{n} z_i \ , \tag{7.1}$$

where n is the sample size and z_i is the value at sampling location i. The subscript SI is added to stress that this estimator is intended for Simple Random Sampling designs.

The strategy (SI, $\hat{\bar{z}}_{\mathrm{SI}}$) is '$p$-unbiased'; this is a quality measure defined as: $E_p\left(\hat{\bar{z}}_{\mathrm{SI}}\right) = \bar{z}$, where $E_p(\cdot)$ denotes the statistical expectation over all possible sample realizations from a design p (in this case Simple Random Sampling). This means that if one were to repeat sampling, measuring and calculating $\hat{\bar{z}}_{\mathrm{SI}}$ in the same way again and again, one would find on average the true value $\bar{z}$. If measurement errors are present, the unbiasedness still holds if the errors are purely random, i.e., zero on average.

The *sampling* variance of the estimated mean, $V\left(\hat{\bar{z}}_{\mathrm{SI}}\right)$, is estimated by:

$$\widehat{V}\left(\hat{\bar{z}}_{\mathrm{SI}}\right) = \frac{1}{n\,(n-1)} \sum_{i=1}^{n} \left(z_i - \hat{\bar{z}}_{\mathrm{SI}}\right)^2 \ , \tag{7.2}$$

and the standard error is estimated by the square root of $\widehat{V}\left(\hat{\bar{z}}_{\mathrm{SI}}\right)$. For sampling without replacement of finite populations the estimated variance must be multiplied by $\left(1 - \frac{n}{N}\right)$, the finite-population correction.

If the data contain random and mutually independent measurement errors, their contribution to the total estimation error is automatically included in the estimated variance $\widehat{V}\left(\hat{\bar{z}}_{\mathrm{SI}}\right)$.

The $100(1-\alpha)\%$ confidence interval for $\bar{z}$ is given by:

$$\hat{\bar{z}}_{\mathrm{SI}} \pm t_{1-\alpha/2} \cdot \sqrt{\widehat{V}\left(\hat{\bar{z}}_{\mathrm{SI}}\right)} \ , \tag{7.3}$$

where $t_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of the Student distribution with $(n-1)$ degrees of freedom. This confidence interval is based on the assumption that z, and as a consequence $\hat{\bar{z}}_{\mathrm{SI}}$, is normally distributed. If the distribution deviates clearly from normality, the data should be first transformed to normality, for instance by taking the logarithm. The interval boundaries thus found are then back-transformed to the original scale. Transformation is not necessary if n is large, because $\hat{\bar{z}}_{\mathrm{SI}}$ is then approximately normally distributed according to the Central Limit Theorem.

In the case of composite sampling (Sect. 4.3), a single composite aliquot may be formed by bulking all n individual aliquots from a simple random sample. The measured z-value of the composite is then a p-unbiased estimate of the spatial mean, with sampling variance $S^2(z)/n$. However, the spatial

variance $S^2(z)$ is generally unknown and cannot be estimated from a single measurement.

An unbiased estimate of the standard error can be obtained by repeating the process of sampling, bulking and measuring. If m composite aliquots are formed, each from a simple random sample of size n, the spatial mean is estimated by:

$$\hat{\bar{z}}_{\mathrm{c}} = \frac{1}{m}\sum_{j=1}^{m} z_{\mathrm{c}j} \ , \tag{7.4}$$

where $z_{\mathrm{c}j}$ is the value measured on composite j. The subscript c indicates that this estimator applies to composite simple random samples.

The sampling variance of $\hat{\bar{z}}_{\mathrm{c}}$ is estimated by:

$$\widehat{V}\left(\hat{\bar{z}}_{\mathrm{c}}\right) = \frac{1}{m\,(m-1)}\sum_{j=1}^{m}\left(z_j - \hat{\bar{z}}_{\mathrm{c}}\right)^2 \ , \tag{7.5}$$

and the standard error is estimated by the square root $\sqrt{\widehat{V}\left(\hat{\bar{z}}_{\mathrm{c}}\right)}$. Confidence intervals are calculated according to (7.3), except that now there are m degrees of freedom.

The formulas for estimating means can also be used to estimate *fractions* of the area where a given condition in terms of the target variable is met, such as the fraction of the area where a qualitative target variable has a given value, for instance 'very suitable', or where a quantitative target variable exceeds a given threshold. Areal fractions can be estimated by first generating a 0/1 indicator variable from the sample data, with value 1 if the condition is met (in the example above, if z= 'very suitable') and 0 otherwise. The above formulas are then simply applied to this indicator variable. The only exception is the calculation of confidence intervals, because the indicator variable is clearly not normally distributed. The sample fraction has a binomial distribution, and with small samples ($n < 20$) this distribution should be used to construct confidence intervals. For larger samples, the distribution is close enough to normality and (7.3) will be sufficiently accurate for most practical applications.

The *Spatial Cumulative Distribution Function* (SCDF) of z can be estimated through repeated application of the indicator technique described before. The measured variable z is first transformed to a series of indicator variables corresponding to a number of increasing thresholds. The areal fractions estimated from these indicators, together with the thresholds, form an estimate of the SCDF.

Estimation of areal fractions or SCDFs on the basis of predictions from a regression model deserves special attention. Suppose one wants to estimate the fraction of the area where some variable z exceeds a given threshold z_{t}, and that z is not measured directly at the sampling locations, but predicted by a regression model. A naive approach would then be to apply the indicator

technique described above to the regression predictions. However, this may lead to serious bias in the estimated fraction, especially if the prediction errors are large relative to the difference between the predictions and the threshold. A simple way to avoid this is to calculate from the regression model the probabilities of exceeding the threshold at the sampling locations, and to apply the usual estimation formulas (7.1), (7.2) and (7.3) to these probabilities instead of the indicator variable. It can be shown that the resulting estimate of the areal fraction as well as its standard error are p-unbiased.

The *spatial* variance of z between locations in the area, $S^2(z)$, is estimated by:

$$\widehat{S^2}(z) = \frac{1}{(n-1)} \sum_{i=1}^{n} \left(z_i - \hat{\bar{z}}_{\mathrm{SI}}\right)^2 \ , \tag{7.6}$$

and the standard deviation by the square root: $\widehat{S}(z) = \sqrt{\widehat{S^2}(z)}$.

The variance between locations in the area can still be estimated in the case of m composite simple random samples, namely by:

$$\widehat{S^2}(z) = \frac{n}{(m-1)} \sum_{j=1}^{m} \left(z_j - \hat{\bar{z}}_{\mathrm{c}}\right)^2 \ . \tag{7.7}$$

Sample Size for Estimation

The sample size needed to estimate a mean such that, with a specified large probability $(1-\alpha)$, the relative error $|(\hat{\bar{z}} - \bar{z})/\bar{z}|$ is smaller than a particular limit r, can be calculated by:

$$n = \left(\frac{u_{1-\alpha/2} \cdot \breve{S}(z)}{r \, \breve{\bar{z}}} \right)^2 \ , \tag{7.8}$$

where $u_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution, $\breve{S}(z)$ is a prior estimate of the standard deviation of z in the area, and $\breve{\bar{z}}$ is a prior estimate of the mean. In this equation, $\breve{S}(z)/\breve{\bar{z}}$ is a prior estimate of the coefficient of variation of z in the area, which should be obtained from a pilot or previous sampling in the same area, from sampling in a similar area, or from general knowledge of the spatial variation.

If instead of the relative error, the absolute error $|\hat{\bar{z}} - \bar{z}|$ should be smaller than a specified limit d, the required sample size is given by:

$$n = \left(\frac{u_{1-\alpha/2} \cdot \breve{S}(z)}{d} \right)^2 \ . \tag{7.9}$$

The sample size needed to estimate a fraction P can be calculated in the same way as with a quantitative variable. In that case the prior estimate of

the standard deviation of the corresponding indicator variable is derived from a prior estimate of the fraction $\breve{P}$ by: $\breve{S}(z) = \sqrt{\breve{P}\,(1-\breve{P})}$.

The above equations for sample size generally do not render integer numbers, so some rounding will be needed. In doing so, caution should be taken because rounding to the nearest integer does not always give the optimal solution. We advise to evaluate the quality measure for some integer sample sizes around the calculated sample size.

Sample Size for Hypothesis Testing

The basics of hypothesis testing as a mode of inference are outlined in Sect. 2.2.4. The sample size needed for testing a hypothesis can be determined through the procedure described below (see EPA (2000) for a comprehensive description). This procedure starts with making the following choices:

- select the baseline condition, i.e., the *de facto* decision outcome when there is insufficient evidence to refute it;
- specify a range of possible values of the target quantity (grey area) where the consequences of accepting the baseline condition while the alternative condition is true are considered tolerable;
- choose tolerable probabilities for the decision error made when the baseline condition is rejected while it is in fact true, and for the decision error made when the baseline condition is accepted while the alternative condition is in fact true.

These choices are explained below.

1. The baseline condition is referred to as the null-hypothesis (H_0). The choice of the baseline condition is important because the same sample data may lead to different decisions depending on the choice of the baseline condition. If sample data are tested against a regulatory threshold, the choice of the baseline condition should be based on where to put the burden of proof. Two situations can be distinguished: law enforcement and licence application. In the first case, the government is the initiator of the investigation, and the baseline condition is generally that the target quantity (statistical parameter of interest) is below the Action Level (no measures). In the second case, a private person or company is the initiator, who must show that the situation complies with the regulations for obtaining a licence. If the choice of the null-hypothesis cannot be based on regulatory considerations, it may be based on the relative consequences of decision errors. If the consequences of deciding that the target quantity is below the Action Level while in fact it is above that level (e.g., a risk to human health) are more serious than those of the reverse error (e.g., the risk of unnecessary clean-up), then the null hypothesis should be that the target quantity is above the Action Level.

2. The width of the grey area is referred to as the 'minimum detectable difference'. At the boundary of the grey area, the consequences of a false acceptance error are considered large enough to set a low probability on this error. In general, the smaller the minimum detectable difference, the larger the required sample size (given the probability of a false acceptance error).
3. In the third step, an upper limit is chosen for the probability that the null-hypothesis is rejected while it is in fact true (false rejection; type I error). This maximum tolerable probability is denoted by α. If the consequences of a false rejection are serious, then a small probability must be chosen, for instance 1%; if they are less serious, a larger probability may be acceptable (e.g., 5% or even 10%). An upper limit is also chosen for the probability of a false acceptance (type II error), β, when the true value of the target quantity equals the Action Level minus the minimum detectable difference (H_0: 'target quantity $\geq$ Action Level'), or Action Level plus the minimum detectable difference (H_0: 'target quantity $\leq$ Action Level'). The probability $1 - \beta$ is referred to as the power of the test. Besides these two probabilities at the boundaries of the grey area, one is free to specify additional probabilities at other values of the target quantity.

The sample size required to meet the chosen probabilities can be calculated as follows. We consider first the situation that the aim is to test the mean of an approximately normally distributed variable. First one calculates the critical value for the mean beyond which H_0 is rejected. For 'H_0: $\bar{z} \geq m_0$', this critical value is given by:

$$m_{\text{crit}} = \Phi^{-1}\left(\alpha; m_0; \frac{\breve{S}^2(z)}{n}\right) , \tag{7.10}$$

where Φ is the cumulative normal distribution, and $\breve{S}^2(z)$ is a prior estimate of the spatial variance of z in the target area. For H_0: $\bar{z} \leq m_0$, the critical value is calculated by substituting $(1 - \alpha)$ for α in (7.10).

For H_0: $\bar{z} \geq m_0$, the probability of false acceptance error, β, can then be calculated by:

$$\beta = 1 - \Phi\left(m_{\text{crit}}; m_0; \frac{\breve{S}^2(z)}{n}\right) . \tag{7.11}$$

For the reverse H_0, this probability is calculated by $\Phi\left(m_{\text{crit}}; m_0; \breve{S}^2(z)/n\right)$. The sample size needed to attain a given power of a one-sided test, as a function of the minimum detectable difference Δ (normalized by dividing by a prior estimate of the standard deviation $\breve{S}(z)$), is given in Figs. 7.7 and 7.8, respectively, for $\alpha = 0.05$ and $\alpha = 0.1$.

If the decision is based on an estimate of some percentile of the cumulative distribution function or an estimated areal fraction, then the estimation procedure is more or less similar. The target variable is now a 0/1 indicator variable. For large samples, the mean can still be assumed to follow an

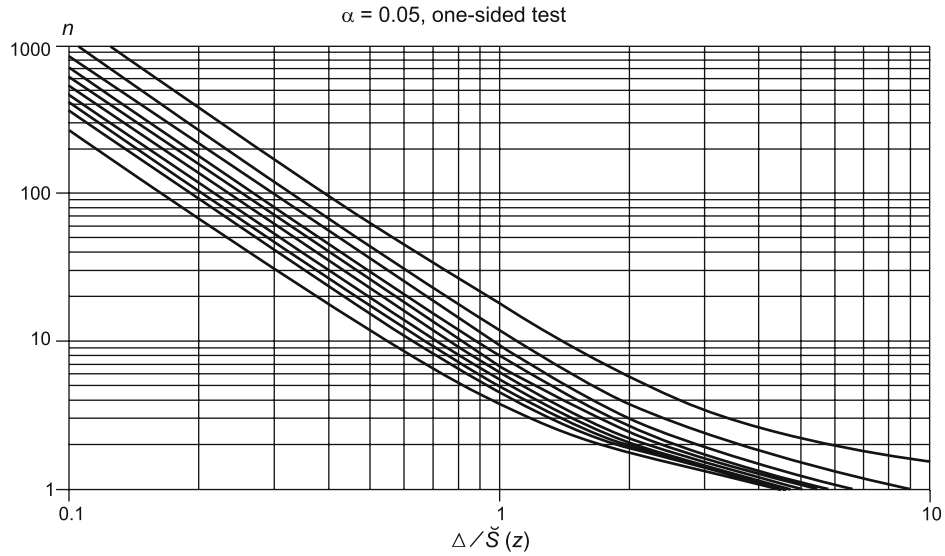


Fig. 7.7. Required sample size versus normalized minimum detectable difference for $\beta = 0.5, 0.4, 0.35, 0.3, 0.25, 0.2, 0.15, 0.1, 0.05, 0.01$ (from left to right) and $\alpha = 0.05$. $\breve{S}(z)$: prior estimate of the standard deviation (z normally distributed)

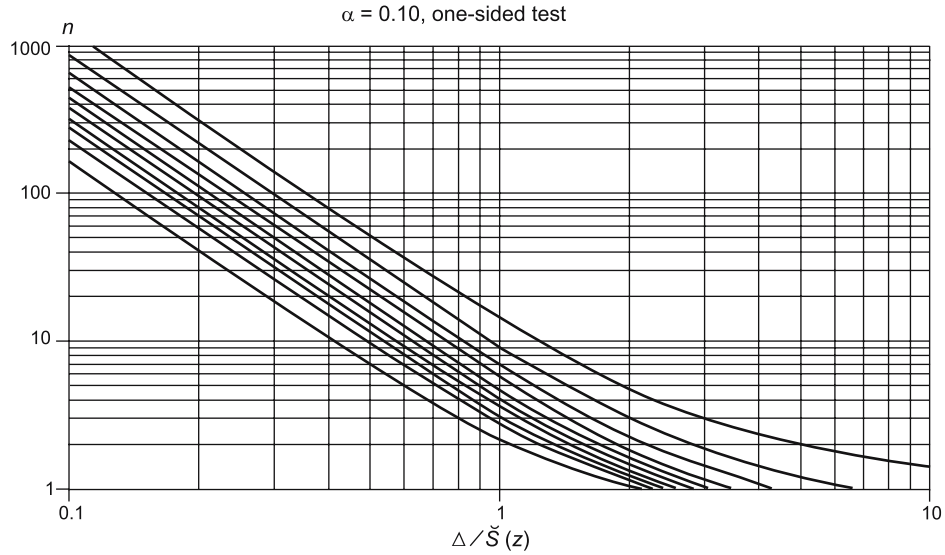


Fig. 7.8. Required sample size versus normalized minimum detectable difference for $\beta = 0.5, 0.4, 0.35, 0.3, 0.25, 0.2, 0.15, 0.1, 0.05, 0.01$ (from left to right) and $\alpha = 0.1$. $\breve{S}(z)$: prior estimate of the standard deviation (z normally distributed)

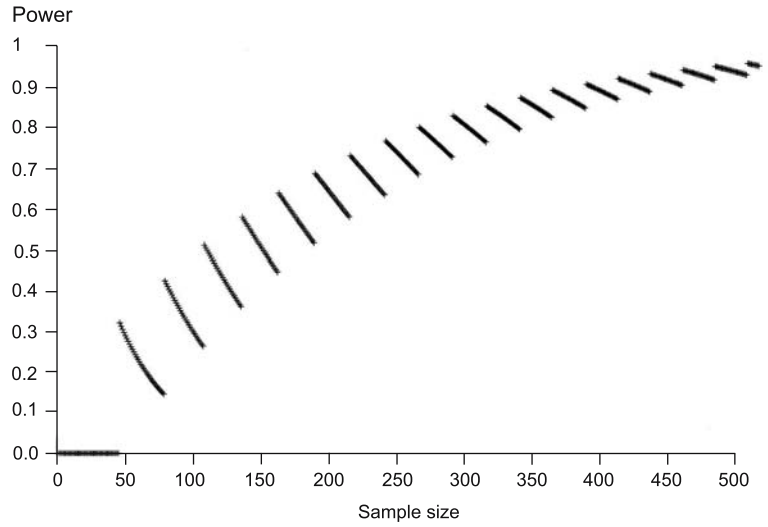


Fig. 7.9. Power versus sample size for 'H_0: $P \geq 0.05$', $\alpha = 0.10$, and a minimum detectable difference of 0.025

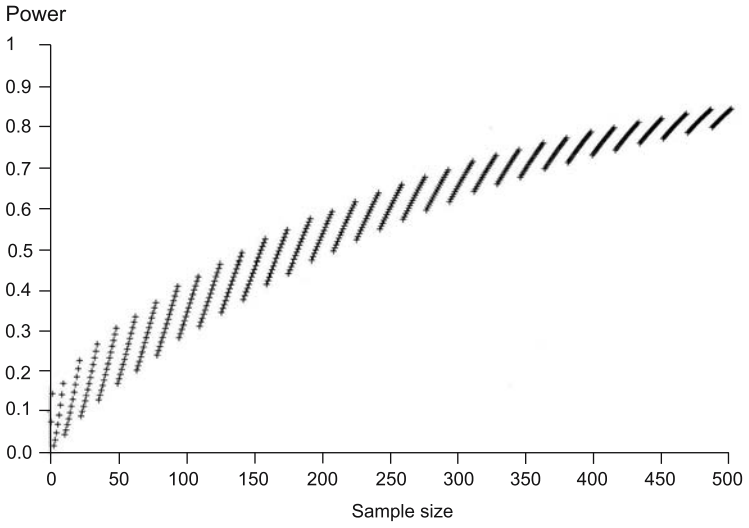


Fig. 7.10. Power versus sample size for 'H_0: $P \leq 0.05$', $\alpha = 0.10$, and a minimum detectable difference of 0.025

approximately normal distribution. For small or medium size samples, however, the normal distribution in (7.10) and (7.11) must be replaced by the binomial distribution. For instance, suppose that the aim is to test the P95 against an Action Level of 75 ('H_0: P95 $\geq$ 75'). This is equivalent to testing the areal fraction with values larger than 75 against an Action Level of 0.05 ('H_0: $P < 0.05$').

Figure 7.9 shows the power as a function of the sample size for 'H_0: $P \geq 0.05$', $\alpha = 0.10$, and a minimum detectable difference of 0.025. First note that very large sample sizes are required for commonly used power values such as 90% ($n = 377$) or 95% ($n = 492$). The required sample size increases even further when smaller probabilities of false rejection, e.g., $\alpha = 0.05$, are tolerated, or smaller differences, e.g., 0.01, must be detectable. Note also that the power does not increase monotonically with the sample size. This is because the binomial distribution is discrete. When the graph indicates two different sample sizes for the same power, the smallest one may be chosen. Figure 7.10 shows the power as a function of the sample size for 'H_0: $P \leq 0.05$', $\alpha = 0.10$, and a minimum detectable difference of 0.025. The function now shows downward jumps.

7.2.4 Stratified Simple Random Sampling

Restriction on Random Selection

The area is divided into sub-areas, called 'strata', in each of which Simple Random Sampling is applied with sample sizes chosen beforehand. Three attributes are to be chosen for this type of design:

1. the definition of the strata, see 'Stratification';
2. the total sample size, see 'Sample Size';
3. the allocation of sample sizes to the strata, see 'Sample Size'

Selection Technique

The algorithm for Simple Random Sampling is applied to each stratum separately.

Example

Figure 7.11 shows an example with 16 square strata and one location in each stratum. Notice the more even spreading compared with Simple Random Sampling in Fig. 7.6.

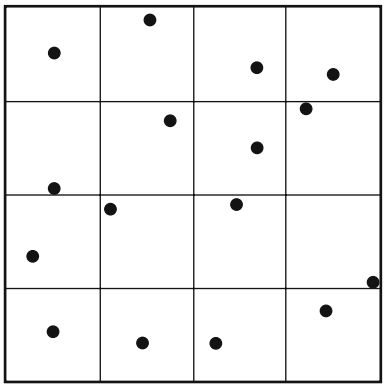


Fig. 7.11. Notional example of a stratified simple random sample

Advantages and Disadvantages

Stratified Simple Random Sampling can have two advantages compared with Simple Random Sampling:

1. with appropriate stratification and allocation of sample sizes to the strata, it is more efficient, i.e., leading to higher accuracy or lower costs, and
2. it allows better control of accuracy for possible sub-areas of interest.

A disadvantage may be that, with inappropriate stratification or sub-optimal allocation of sample sizes, there could be loss rather than gain in efficiency.

Stratification

There are different possible reasons for stratification. The most important one is that the efficiency can be increased compared with Simple Random Sampling, i.e., smaller sampling variance at the same cost, or lower cost with the same variance. In this case, the stratification is chosen so as to maximize the expected gain in efficiency. In practice, this can be achieved by forming strata that are as homogeneous as possible. Also, if the cost per sampling location varies strongly within the area, for instance with the distance from roads, it is efficient to stratify accordingly and to sample the 'inexpensive' strata more densely. (Different conditions in sub-areas may even ask for different types of design.)

Another reason for stratification may be that separate estimates are required for given sub-areas. If the strata coincide with these sub-areas of interest, the accuracy of the estimates can be controlled by allocating sufficient sample sizes to the strata, unlike the situation with Simple Random Sampling. Finally, in large-scale surveys it can make sense for administrative reasons to operate sub-areas (provinces, countries) independently and to treat them as strata.

With inappropriate stratification or sub-optimal allocation of sample sizes, there could be loss rather than gain in efficiency. This may occur if the stratum means differ little or if the sample sizes are strongly disproportional to the surface areas of the strata. If, for instance, one has many small strata with unequal areas and a small sample in each, then these sample sizes are bound to be strongly disproportional because they must be integer numbers.

In order to achieve optimal efficiency, the available prior information should be employed to define strata that are as homogeneous as possible, i.e., having minimal spatial variance of the target variable. (If there is more than one target variable, then one must find some compromise.) Two approaches can be followed: stratification by ancillary variables, or compact geographical stratification.

Stratification by Ancillary Variables

The strata are defined in terms of one or more ancillary variable that are known everywhere in the target area, and that are known or expected to be correlated with the target variable. The strata can be defined in two ways:

1. as the classes of an a priori classification based on knowledge about the relationships between the ancillary variables and the target variable, or
2. if the ancillary variables are quantitative, such as obtained by remote sensing, then the strata can be calculated by cluster analysis.

A suitable cluster method for classification will be 'k-means', a well-established algorithm used to find compact clusters of objects in multivariate attribute space (Hartigan, 1975). The clusters are represented by their multivariate means, referred to as centroids, and the method aims at minimizing the mean squared distance between the objects and their nearest centroid. When applied for stratification, the attribute space is spanned by the ancillary variables, and the objects are the grid nodes of a fine discretization grid. The resulting clusters of grid nodes form the strata.

Compact Geographical Stratification

If no suitable ancillary variables are available for stratification, one may consider stratification on the basis of spatial coordinates. In general, the precision of sampling strategies can be increased by spreading the sampling locations over the study region. A simple way to achieve this spreading is Systematic Random Sampling on regular grids (squares, triangles); see Sect. 7.2.7. Another way to ensure that the randomly selected locations cover the target area as fully as possible is random sampling from compact geographical strata. Just as strata based ancillary variables, such compact geographical strata can also be formed by the k-means clustering algorithm (Brus et al., 1999). For this application, the multivariate attribute space is replaced by the geographical space, with only the spatial coordinates as properties, and again the grid nodes of a fine discretization grid acting as objects. Figure 7.12 shows the

result for a field of 2.8 ha, split up into 40 blocks. In this case we took care that the blocks all have the same area, so that the stratified random sample is self-weighting[1] Section 8.3.3 proposes the use of k-means to optimize spatial coverage samples for spatial mapping.

Statistical Inference

The mean for the area is estimated by:

$$\hat{\bar{z}}_{\mathrm{St}} = \frac{1}{N} \sum_{h=1}^{H} N_h \, \hat{\bar{z}}_h \ , \tag{7.12}$$

where H is the number of strata; N is the number of sampling units in the area; N_h is number of sampling units in stratum h and $\hat{\bar{z}}_h$ is the sample mean of stratum h. Replacing the ratios N_h/N by the relative areas of the strata, a_h, gives the more practical formula:

$$\hat{\bar{z}}_{\mathrm{St}} = \sum_{h=1}^{H} a_h \, \hat{\bar{z}}_h \ . \tag{7.13}$$

The same indicator technique as presented with Simple Random Sampling can be used to estimate areal fractions and the Spatial Cumulative Distribution Function of z.

The strategy (StS, $\hat{\bar{z}}_{\mathrm{St}}$) is p-unbiased. Provided all sample sizes are larger than 1, the variance of $\hat{\bar{z}}_{\mathrm{St}}$ can be estimated by:

$$\widehat{V}\left(\hat{\bar{z}}_{\mathrm{St}}\right) = \sum_{h=1}^{H} a_h^2 \, \widehat{V}\left(\hat{\bar{z}}_h\right) \ , \tag{7.14}$$

where $\widehat{V}\left(\hat{\bar{z}}_h\right)$ is the estimated variance of $\hat{\bar{z}}_h$:

$$\widehat{V}\left(\hat{\bar{z}}_h\right) = \frac{1}{n_h \left(n_h - 1\right)} \sum_{i=1}^{n_h} \left(z_{hi} - \hat{\bar{z}}_h\right)^2 \ , \tag{7.15}$$

where n_h is the sample size in stratum h. The standard error of the estimated mean is estimated by $\sqrt{\widehat{V}\left(\hat{\bar{z}}_{\mathrm{St}}\right)}$. Confidence intervals are calculated in the same way as with Simple Random Sampling (7.3).

An unbiased estimator of the spatial variance $S^2(z)$ is:

$$\widehat{S^2}(z) = \widehat{\overline{z^2}}_{\mathrm{St}} - \left(\hat{\bar{z}}_{\mathrm{St}}\right)^2 + \widehat{V}\left(\hat{\bar{z}}_{\mathrm{St}}\right) \ , \tag{7.16}$$

[1] A self-weighting sample is a sample generated by a design such that weighting of the sample data is unnecessary, i.e., the unweighted sample mean is p-unbiased.

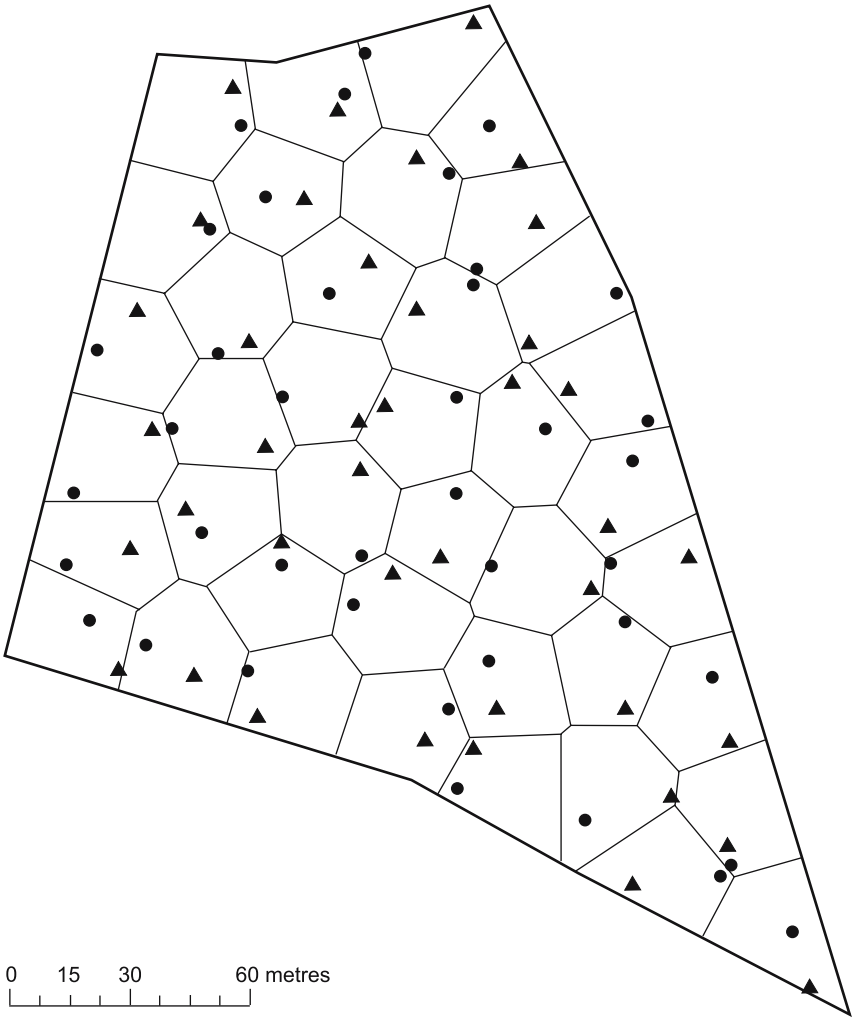


Fig. 7.12. Example of compact geographical stratification of a field. The strata all have the same area, and in each stratum two locations are selected by Simple Random Sampling. Two composite samples can be formed by grouping all samples selected in the first draw (circles), and those selected in the second draw (triangles)

where $\widehat{\overline{z^2}}_{\mathrm{St}}$ denotes the estimated mean of the target variable squared (z^2), obtained in the same way as $\hat{\bar{z}}_{\mathrm{St}}$ (7.13), but using squared values.

It may be asked what gain in precision the stratification has yielded, compared to Simple Random Sampling with the same sample size. The sampling variance that would have been obtained if Simple Random Sampling had been

applied with the same n can be predicted by dividing the estimated spatial variance by the sample size:

$$\widetilde{V}\big(\hat{\bar{z}}_{\mathrm{SI}}\big) = \frac{\widehat{S^2}(z)}{n} \; . \tag{7.17}$$

The efficiency of the applied stratified design, relative to the reference design (Simple Random Sampling with the same n), is defined as the ratio of the two sampling variances:

$$V_{\mathrm{r}} = \frac{\widetilde{V}\big(\hat{\bar{z}}_{\mathrm{SI}}\big)}{\widehat{V}\big(\hat{\bar{z}}_{\mathrm{St}}\big)} \; . \tag{7.18}$$

Multiplied by the actual sample size, this efficiency factor gives the 'equivalent sample size', which would yield the same precision if Simple Random Sampling were used:

$$n_{\mathrm{eq}} = V_{\mathrm{r}} \cdot n \; . \tag{7.19}$$

Brus (1994) used this procedure to quantify the effect of stratification by soil map and land-use map units on the quality of estimated spatial means of phosphate sorption properties of the soil.

Sample Size

The sample sizes in the strata may be chosen so as to minimize the variance $V\big(\hat{\bar{z}}_{\mathrm{St}}\big)$ for a given maximum allowable cost, or to minimize the cost for a given maximum allowable variance. A simple linear cost function is:

$$C = c_o + \sum_{h=1}^{H} c_h \, n_h \; , \tag{7.20}$$

where c_o is the overhead cost and c_h is the cost per sampling location in stratum h.

Adopting this function, the optimal *ratios* of the sample sizes to the total sample size n are:

$$\frac{n_h}{n} = \frac{\frac{a_h \breve{S}_h}{\sqrt{c_h}}}{\sum\limits_{h=1}^{H} \frac{a_h \breve{S}_h}{\sqrt{c_h}}} \; , \tag{7.21}$$

where the $\breve{S}_h$ are prior estimates of the standard deviations in the strata. This formula implies that a stratum gets a larger sample if it is larger or more variable or less expensive to sample.

The total sample size affordable at a fixed cost C, assuming that optimal allocation to the strata is applied, is:

$$n = \frac{(C - c_o) \sum\limits_{h=1}^{H} \frac{a_h \breve{S}_h}{\sqrt{c_h}}}{\sum\limits_{h=1}^{H} a_h \breve{S}_h \sqrt{c_h}} \; . \tag{7.22}$$

The total sample size needed to keep the variance below a maximum value $V_{\max}$, again assuming that optimal allocation to the strata is applied, is:

$$n = \frac{1}{V_{\max}} \cdot \sum_{h=1}^{H} \left(a_h \breve{S}_h \sqrt{c_h} \right) \cdot \sum_{h=1}^{H} \frac{a_h \breve{S}_h}{\sqrt{c_h}} \ . \tag{7.23}$$

If the cost per location is equal for the strata, this reduces to:

$$n = \frac{1}{V_{\max}} \cdot \left(\sum_{h=1}^{H} a_h \breve{S}_h \right)^2 \ . \tag{7.24}$$

If, instead of $V_{\max}$, an absolute error d has been specified with an allowed probability of exceedance α, then $V_{\max}$ can be derived from d and α, according to

$$V_{\max} = \left(\frac{d}{u_{1-\alpha/2}} \right)^2 \ , \tag{7.25}$$

where $u_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution.

When estimating areal fractions rather than means of quantitative variables, the above formulas for sample sizes can still be applied if $\breve{S}_h$ is replaced by $\sqrt{\breve{P}_h \left(1 - \breve{P}_h\right)}$, where $\breve{P}_h$ is a prior estimate of the fraction in stratum h.

If no prior estimates of the standard deviations in the strata are available, and the cost of sampling is constant for the strata, then proportional allocation is recommended:

$$\frac{n_h}{n} = \frac{a_h}{\sum a_h}. \tag{7.26}$$

Also, for surveys with more than one target variable, proportional allocation can be a good choice.

The above equations for sample size generally do not render integer numbers, so some rounding will be needed. In doing so, caution should be taken because rounding to the nearest integer does not always give the optimal solution. We advise to evaluate the quality measure for some integer sample sizes around the calculated sample size.

7.2.5 Two-Stage Random Sampling

Restriction on Random Selection

As with Stratified Simple Random Sampling, the area is divided into a number of sub-areas. Sampling is then restricted to a number of randomly selected sub-areas, in this case called primary units. Note the difference with Stratified Simple Random Sampling, where all sub-areas (strata) are sampled. In large-scale surveys, this principle is often generalized to multistage sampling. (Three-stage crop sampling, for instance, could use sub-areas from remotely

sensed images as primary units, fields as secondary units, and sample plots as tertiary units.)

Attributes to choose for this type of design are the following.

1. The definition of the primary units (PUs). As with stratification, the user is free to define the PUs as seems fit; see 'Advantages and Disadvantages'.
2. The mode of selection of the PUs: with replacement (a PU can be selected more than once) or without replacement; see 'Selection Technique'.
3. The selection probabilities for the PUs: equal or proportional to size; see 'Selection Technique'.
4. The number of PUs to select in the first stage (without replacement) or the number of PU selections (with replacement); see 'Sample Size'.
5. The number of sampling locations in the PUs; see 'Sample Size'.

Selection Technique

Because of its simplicity in statistical inference we present a version of Two-Stage Random Sampling by which the primary units (PUs) are selected with replacement and with probabilities proportional to their area. The following algorithm can be used to make n such selections from all N PUs in the area.

(a) Determine the relative areas of all PUs, $a_1, \ldots, a_N$, and their cumulative sums, $a_1, \ldots, S_k, \ldots, 1$, with $S_k = \sum_{i=1}^{k} a_i$.
(b) Generate a random number u from the uniform distribution on the interval $(0, 1)$.
(c) Select the PU whose corresponding S_k is the first in the series that exceeds u.
(d) Repeat steps b and c until n selections have been made.

An alternative, sometimes more efficient algorithm works with a geographical representation of the area and its PUs:

(a) Select a random point in the area, as in Simple Random Sampling.
(b) Use a point-in-polygon routine to determine in which PU the point falls, and select this PU.
(c) Repeat steps a and b until n selections have been made.

In the second stage, a pre-determined number of sampling locations, m_i, is selected within each of the PUs selected in the first stage. This is done in the same way as in Simple Random Sampling. If the geographical algorithm is applied, the random points used to select the PUs may also be used as sampling locations. If a PU has been selected more than once, an independent sample of locations must be selected for each time the PU was selected.

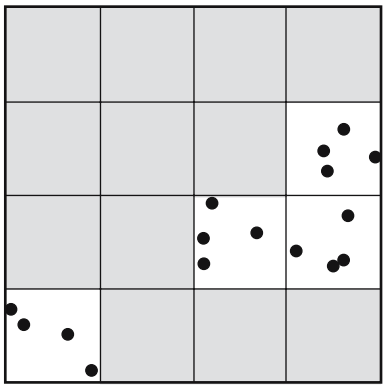


Fig. 7.13. Notional example of a two-stage random sample

Example

Figure 7.13 shows four square PUs selected in the first stage, and four locations in each in the second stage. Notice the greater spatial clustering compared with Simple Random Sampling in Fig. 7.6. This is just a simple, notional example. It should be noted, however, that the PUs may be defined in any way that seems appropriate, and that the number of sampling locations may vary among units.

Advantages and Disadvantages

As with Cluster Random Sampling (Sect. 7.2.6), the spatial clustering of sampling locations created by Two-Stage Random Sampling has the operational advantage of reducing the travel time between locations in the field. Of course, the importance of this advantage depends on the extent of the area relative to the sample size and the accessibility of the terrain. The advantage may be amplified by defining the PUs such that they reflect dominant accessibility features like roads. PUs may also be defined by land ownership, especially if getting permission for fieldwork constitutes an important cost component.

The spatial clustering generally leads to lower precision for a given sample size. However, as with Cluster Random Sampling, the rationale of its application is that the operational advantage allows a larger sample size for the same budget, so that the initial loss of precision is outweighed. This can be expected if much spatial variation occurs *within* the PUs, as compared with the variation *between* the PUs. A disadvantage of Two-Stage Random Sampling may be that estimates from the sample data will turn out to be much less accurate than expected, if the variation within the PUs is much smaller than assumed when deciding on the number of sampling locations in them.

Statistical Inference

The mean of the area is estimated by the simple estimator:

$$\hat{\bar{z}}_{\mathrm{Ts}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\bar{z}}_i \ , \tag{7.27}$$

where n is the number of PU selections, and $\hat{\bar{z}}_i$ is the sample mean of the PU from selection i. The same indicator technique as presented with Simple Random Sampling can be used to estimate areal fractions and the Spatial Cumulative Distribution Function of z.

The strategy (TsS, $\hat{\bar{z}}_{\mathrm{Ts}}$) is p-unbiased. The variance is simply estimated by:

$$\widehat{V}\big(\hat{\bar{z}}_{\mathrm{Ts}}\big) = \frac{1}{n\,(n-1)}\sum_{i=1}^{n} \big(\hat{\bar{z}}_i - \hat{\bar{z}}_{\mathrm{Ts}}\big)^2 \ . \tag{7.28}$$

Notice that neither the areas of the PUs, a_i, nor the secondary sample sizes m_i occur in these formulas. This simplicity is due to the fact that the PUs are selected with replacement and with probabilities proportional to size. The effect of the secondary sample sizes on the variance is implicitly accounted for. (To understand this, remember that the larger m_i, the less variable $\hat{\bar{z}}_i$, and the smaller its contribution to the variance.)

The standard error is estimated by $\sqrt{\widehat{V}\big(\hat{\bar{z}}_{\mathrm{Ts}}\big)}$. Confidence intervals are calculated in the same way as with Simple Random Sampling (7.3). The spatial variance in the area, the efficiency of the sampling strategy and the equivalent sample size can be estimated from the sample data, similar to the procedure presented for Stratified Simple Random Sampling with formulas (7.16), (7.18) and (7.19).

Sample Size

The primary and secondary samples sizes n and m_i can be optimized via dynamic programming, given a budget or variance requirement, any cost function and prior estimates of the within- and between-unit variances; see Domburg et al. (1997). A simple approximation is obtained by assuming the m_i to be constant, say $m_i = m$. This is reasonable if the PUs have roughly the same area and internal variation. The variance of the mean is now

$$V\big(\hat{\bar{z}}_{\mathrm{Ts}}\big) = \frac{1}{n}\left(S_{\mathrm{b}}^2 + \frac{1}{m}S_{\mathrm{w}}^2\right) \ , \tag{7.29}$$

where S_{b}^2 and S_{w}^2 are the between-unit and the pooled within-unit variance, respectively. Given the linear cost function $C = c_1\,n + c_2\,n\,m$, the sample sizes minimizing the variance under the constraint that the cost does not exceed a budget $C_{\max}$ can be found using the Lagrange multiplier method:

$$n = \frac{C_{\max}\, \breve{S}_{\mathrm{b}}}{\breve{S}_{\mathrm{w}}\sqrt{c_1\, c_2} + \breve{S}_{\mathrm{b}}\, c_1} \tag{7.30}$$

and

$$m = \frac{\breve{S}_{\mathrm{w}}}{\breve{S}_{\mathrm{b}}}\sqrt{\frac{c_1}{c_2}}\,, \tag{7.31}$$

where $\breve{S}_{\mathrm{b}}$ and $\breve{S}_{\mathrm{w}}$ are prior estimates of S_{b} and S_{w}, respectively.

Conversely, minimizing the cost under the constraint that the variance does not exceed a maximum $V_{\max}$ can be achieved by:

$$n = \frac{1}{V_{\max}}\left(\breve{S}_{\mathrm{w}}\, \breve{S}_{\mathrm{b}}\sqrt{\frac{c_2}{c_1}} + \breve{S}_{\mathrm{b}}^2\right) \tag{7.32}$$

and m as above.

If, instead of $V_{\max}$, an absolute error d has been specified with an allowed probability of exceedance α, then $V_{\max}$ can be derived from d and α, according to $\sqrt{V_{\max}} = d/u_{1-\alpha/2}$, where $u_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution.

The above equations for sample size generally do not render integer numbers, so some rounding will be needed. In doing so, caution should be taken because rounding to the nearest integer does not always give the optimal solution. We advise to evaluate the quality measure for some integer sample sizes around the calculated sample size. (The need for this becomes obvious in the extreme case of $\breve{S}_B = 0$, for which (7.30) gives $n = 0$, while obviously n should be one.)

When estimating areal fractions rather than means of quantitative variables, the above formulas for sample sizes can still be applied if $\breve{S}_{\mathrm{b}}^2$ is interpreted as a prior estimate of the variance between the fractions in the units P_i, and $\breve{S}_{\mathrm{w}}^2$ is replaced by a prior estimate of the mean of $P_i\,(1-P_i)$ over the units.

7.2.6 Cluster Random Sampling

Restriction on Random Selection

Pre-defined sets of locations are selected, instead of individual locations as in Simple Random Sampling, Stratified Simple Random Sampling and Two-Stage Random Sampling. These sets are referred to as 'clusters'. Attributes to choose for this type of design are:

1. the definition of the clusters, see 'Definition of Clusters';
2. the selection probabilities of the clusters, equal or proportional to size (the number of locations in them), see 'Selection technique';
3. the number of clusters to select, see 'Sample Size'.

Definition of Clusters

In defining the clusters one has to consider the following aspects:

1. Shape: the type of pattern of the locations, e.g., linear with equidistant locations on a line (so-called 'transects'). As the reason for Cluster Random Sampling lies in facilitation of the fieldwork, only regular patterns are applied in practice.
2. Size: the number of locations.
3. Direction: the geographical orientation, e.g., random directions or the direction of a gradient known by prior knowledge.

As with strata in Stratified Simple Random Sampling and primary units in Two-Stage Random Sampling, one is free to define the clusters as seems most appropriate. The rationale of Cluster Random Sampling is that the operational advantage in fieldwork expectedly outweighs any loss of precision due to poor spatial coverage of the area by the sample. So, the cluster definition should aim at maximal operational advantage as well as maximal within-cluster variation. This is similar to the definition of primary units in Two-Stage Random Sampling.

It should be noted that, for the inference method as presented here to be valid, it is not necessary to randomize the directions of the clusters. Indeed we see little use in randomized directions. In the case of a trend with a known direction, random directions are clearly sub-optimal. Even if no trend is known to exist, it seems safer to choose some different directions purposively, e.g., two directions perpendicular to each other, or three at 60 degrees, or four at 45 degrees. If the clusters are given random directions, too many of them may happen to fall perpendicular to the direction of a possible unknown trend, thus producing little information.

Selection Technique

If the area is not discretized beforehand, the number of possible sampling locations is infinite and so is the number of clusters. It would thus be impossible to identify all possible clusters beforehand and to sample from this collection. That could only be done if the area is discretized. However, there is no need to identify all possible clusters beforehand. Only those that are selected need to be identified, and selection of a cluster can take place via selection of one of its locations. Hence the following algorithm:

(a) Select a random location in the area as in Simple Random Sampling; use this location as a 'starting point'.
(b) Find the other locations of the cluster to which the starting location belongs, by applying predetermined geometric rules corresponding with the chosen cluster definition.
(c) Repeat steps a and b until n clusters have been selected.

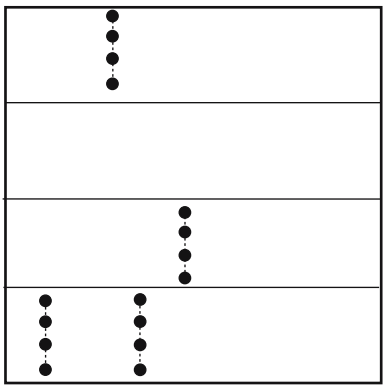


Fig. 7.14. Notional example of a cluster random sample

A condition for this algorithm to be valid is that the geometric rules are such that the same cluster is created, regardless of which of its locations is used as the starting point. A well-known technique satisfying this condition is Random Transect Sampling with equidistant sampling locations on straight lines with a fixed direction. Given this direction, the random starting point determines the line of the transect. The other sampling locations are found by taking a predetermined distance in both directions from the starting point, until the line crosses the boundary of the area. Clusters thus formed will generally consist of a variable number of locations, and the probability of selecting a cluster is proportional to the number of locations in it. This is taken into account in the method of statistical inference as presented.

The above algorithm assumes that a perfect sampling frame is available. In practice, however, imperfections of the frame may appear after inspection of the projected sampling locations in the field. If projected sampling locations are for some reason judged not to belong to the target universe, then shifting to nearby locations would cause bias, as boundary zones become oversampled.

The correct remedy with this type of design is to delete the entire cluster from the sample if its starting point is rejected, and to only delete other locations as far as they are rejected, while maintaining the rest of their cluster. Rejected starting points are to be replaced by sampling locations from a reserve list in pre-determined order.

Example

Figure 7.14 shows four transects, each with four equidistant locations. To limit the length of the transects, the area has first been dissected with internal boundaries perpendicular to the transects. Notice the spatial clustering and the regularity compared with Simple Random Sampling, Stratified Simple Random Sampling and Two-Stage Random Sampling (Figs. 7.6, 7.11

and 7.13). This is just a simple, notional example. It should be noted, however, that the clusters may be defined in any way that seems appropriate.

Advantages and Disadvantages

As with Two-Stage Random Sampling (Sect. 7.2.5), the spatial clustering of sampling locations has the operational advantage of reducing the travel time between locations in the field. In addition, the regularity may reduce the time needed to find consecutive locations in the cluster. Of course, the importance of these advantages depends on the extent of the area relative to the sample size, the trafficability of the terrain and the possibilities and limitations for navigation.

As with Two-Stage Random Sampling, the spatial clustering generally leads to lower precision, given the sample size. Here too, the rationale for using it is that the operational advantages allow a larger sample size for the same budget, so that the initial loss of precision is outweighed. If there is a marked trend with a known direction, the precision can be optimized by defining transects in the direction of the trend.

A disadvantage is that the sample size, i.e., the total number of locations in the clusters which happen to be selected, is generally random. This may be undesirable for budgetary or logistical reasons. The variation in sample size can be reduced by defining clusters of roughly equal size.

Statistical Inference

For this type of design, the same formulas are used as for Two-Stage Random Sampling, with clusters assuming the role of primary units. For clarity, the inference is presented again, together with the 'cluster interpretation' of the quantities.

The means of the area is estimated by the estimator:

$$\hat{\bar{z}}_{\mathrm{Cl}} = \frac{1}{n} \sum_{i=1}^{n} \hat{\bar{z}}_i \ , \tag{7.33}$$

where n is the number of clusters and $\hat{\bar{z}}_i$ is the sample mean of cluster i. The same indicator technique as presented with Simple Random Sampling can be used to estimate areal fractions and the Spatial Cumulative Distribution Function of z.

The strategy (ClS, $\hat{\bar{z}}_{\mathrm{Cl}}$) is p-unbiased. The variance is estimated by:

$$\widehat{V}\left(\hat{\bar{z}}_{\mathrm{Cl}}\right) = \frac{1}{n\,(n-1)} \sum_{i=1}^{n} \left(\hat{\bar{z}}_i - \hat{\bar{z}}_{\mathrm{Cl}}\right)^2 \ . \tag{7.34}$$

Notice that the size of the clusters (number of locations) does not appear in these formulas. This simplicity is due to the fact that the clusters are selected

with probabilities proportional to size. The effect of the cluster size on the variance is implicitly accounted for. (To understand this, consider that larger clusters result in smaller variance among their means.)

The standard error is estimated by $\sqrt{\widehat{V}\left(\hat{\bar{z}}_{\mathrm{Ts}}\right)}$. Confidence intervals are calculated in the same way as with Simple Random Sampling (7.3). The spatial variance in the area, the efficiency of the sampling strategy and the equivalent sample size can be estimated from the sample data, similar to the procedure presented for Stratified Simple Random Sampling, with formulas (7.16), (7.18) and (7.19).

Sample Size

The number of clusters needed to keep the variance of the estimated mean below a given maximum $V_{\max}$ is given by $n = \breve{S}_{\mathrm{b}}^2 / V_{\max}$, where $\breve{S}_{\mathrm{b}}^2$ is a prior estimate of the variance between cluster means. Clearly, this variance depends on the number of locations in the clusters and their spatial pattern. If prior information on the spatial variability is available in the form of a variogram, the method described in Sect. 7.2.15 can be used to estimate S_{b}^2 for a given cluster definition.

If, instead of $V_{\max}$, an absolute error d has been specified with an allowed probability of exceedance α, then $V_{\max}$ can be derived from d and α, according to $\sqrt{V_{\max}} = d/u_{1-\alpha/2}$, where $u_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution.

When estimating areal fractions rather than means of quantitative variables, the above formula for n can still be applied if $\breve{S}_{\mathrm{b}}^2$ is interpreted as a prior estimate of the variance between cluster fractions.

The above equation for sample size generally do not render integer numbers, so some rounding will be needed. In doing so, caution should be taken because rounding to the nearest integer does not always give the optimal solution. We advise to evaluate the quality measure for some integer sample sizes around the calculated sample size.

7.2.7 Systematic Random Sampling

Restriction on Random Selection

As with Cluster Random Sampling (Sect. 7.2.6), random selection is applied to pre-defined *sets* of locations, instead of individual locations as in Simple Random Sampling, Stratified Simple Random Sampling and Two-Stage Random Sampling. The difference with Cluster Random Sampling is that only one cluster is selected. In this sense, Systematic Random Sampling is a special case of Cluster Random Sampling.

The term 'cluster' as used here does not refer to geographical compactness, but to the fact that if one location of a cluster is included in the sample,

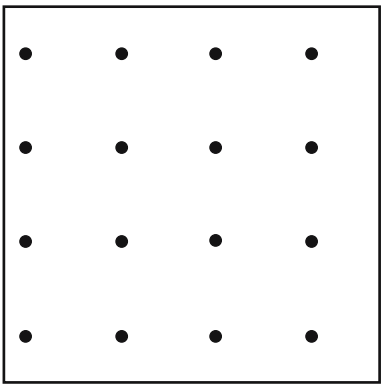


Fig. 7.15. Notional example of a systematic random sample

all other locations are included too. On the contrary, as only one cluster is selected, the clusters are defined such as to allow each of them to cover the area as fully as possible. This is achieved with clusters in the form of regular grids (Random Grid Sampling). Thus the only attribute to be chosen with this type of design is the definition of the grid (see 'Grid Definition').

Selection Technique

The selection algorithm for Cluster Random Sampling is used with $n = 1$.

Example

Figure 7.15 shows a random square grid. Notice the more even spatial spreading and the greater regularity compared with all other types of design (Figs. 7.6 – 7.14).

Grid Definition

Three aspects are to be decided on: the shape of the grid cells, the size of the grid cells, and the direction of the grid. With respect to shape, there are three options: square, triangular or hexagonal. An extensive study by Matérn (1986) showed that triangular grids will generally be slightly more efficient in estimating the spatial mean than square or hexagonal ones with the same location density. In practice, however, a square grid may have an operational advantage over a triangular one that outweighs the small difference in precision.

The size of the grid cells (i.e., the distance between adjacent grid nodes, referred to as the grid spacing), determines the sample size; see 'Sample Size'. With regard to the direction of the grid, it should be noted that, for the

inference method as presented here to be valid, it is not necessary to randomize the direction of the grid. As with Cluster Random Sampling, we see little use in a randomized direction. For instance, if the target area is rectangular, a grid parallel to the boundaries will often make fieldwork easier.

Advantages and Disadvantages

The main advantage is that, due to the optimized spatial coverage, Systematic Random Sampling will often give more accurate results than any other random design. In addition, it has the same operational advantage as Cluster Random Sampling: the regularity of the grid may reduce the time needed to locate consecutive locations in the field. Here too, the importance of this advantage depends on the scale, the accessibility of the terrain and the navigation technique used.

Because this type of design does not produce any random repetition, no unbiased estimate of the sampling variance is available. If the spatial variation in the area is pseudo-cyclic, the variance may be severely underestimated, thus yielding a false impression of accuracy. An operational disadvantage may be that the total travel distance between sampling locations is relatively long, due to their even spreading. Finally, Systematic Random Sampling has the same disadvantage as Cluster Random Sampling: the sample size (i.e., the number of grid nodes that happen to fall inside the area) is generally random, which may be undesirable for budgetary or logistic reasons. The possible variation in sample size will often be larger than with Cluster Random Sampling, and it will be more difficult to reduce this variation.

Statistical Inference

The mean of the area is simply estimated by the sample mean $\hat{\bar{z}}$, as with Simple Random Sampling. The same indicator technique as presented with Simple Random Sampling can be used to estimate areal fractions and the Spatial Cumulative Distribution Function of z.

The strategy (SyS, $\hat{\bar{z}}$) is p-unbiased. This condition holds only if the grid is randomly selected, as is prescribed by the selection technique given above. In 'Centred Grid Sampling', on the other hand, the grid is purposively placed around the centre of the area, so that the boundary zones are avoided. This is a typical model-based strategy (see Sect. 7.3), which is p-biased.

Unfortunately, no p-unbiased variance estimator exists for this type of design. Many variance estimators have been proposed in the literature, but all are based on assumptions about the spatial variation. A well-known procedure is Yates' method of balanced differences (Yates, 1981). An overview of variance estimation is given by Cochran (1977). A simple, often applied procedure is to calculate the variance as if the sample had been obtained by Simple Random Sampling. If there is no pseudo-cyclic variation, this over-estimates the variance, so in that case the accuracy assessment will have a safety margin.

Sample Size

As indicated above, the sample size is generally random. The average sample size is determined by the choice of the cell size, i.e., the grid spacing. A rough approach to this choice is to determine the sample size in the same way as for Simple Random Sampling (Sect. 7.2.3) and to reduce this by an empirical factor (for instance 2) to account for the greater precision of Systematic Random Sampling relative to Simple Random Sampling. The average required grid size for a square grid is then $\sqrt{A/m}$, where A denotes the surface area. However, if an estimated variogram is available, greater accuracy is achieved by applying the method described in Sect. 7.2.15.

7.2.8 Advanced Design-Based Strategies

Apart from the basic strategies outlined in the previous sections, a large number of more advanced strategies have been developed. This section outlines some of the major options.

Compound Strategies

The basic strategies discussed in the previous sections can be combined in many ways to form compound strategies. One example is given in Fig. 7.16, where Two-Stage Random Sampling has been applied while using Systematic Random Sampling instead of Simple Random Sampling in both stages. In this case, a square grid of 2×2 PUs was selected, and then a square grid of 2×2 locations in each of the selected PUs. Notice that the total between-location distance is reduced relative to Systematic Random Sampling in Fig. 7.15, that the risk of interference with possible cyclic variation has practically vanished, and that the operational advantage of regularity in the pattern still largely exists.

Figure 7.17 shows another example of a compound strategy: Stratified Cluster Random Sampling with four strata and two clusters in each stratum. The clusters are perpendicular transects, each with two locations at a fixed distance. Notice that, due to the stratification, a more even spread is obtained than with Cluster Random Sampling in Fig. 7.14, while the operational advantage of regularity still exists. See de Gruijter and Marsman (1984) for an account of perpendicular Random Transect Sampling and an application to the quality assessment of soil maps.

The reason for combining two or more basic strategies is always the enhancement of advantages or mitigation of disadvantages of the basic strategies. As a final example, consider the situation in which the high precision and the operational advantage of regularity in Systematic Random Sampling is required, but it is desirable that the precision can be quantified from the data, without recourse to assumptions about the spatial variability. A possible solution is to adapt the Two-Stage/Systematic compound strategy of Fig. 7.16. In

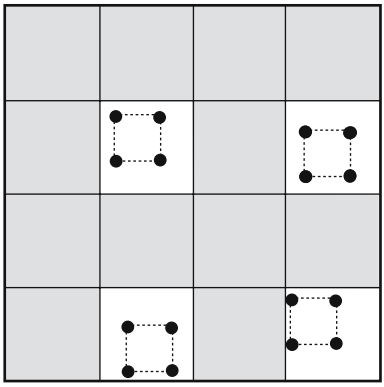


Fig. 7.16. Notional example of a two-stage random sample with Systematic Random Sampling in both stages

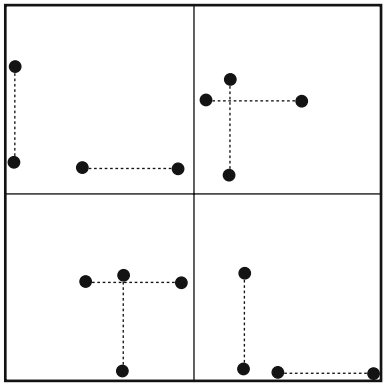


Fig. 7.17. Notional example of a stratified cluster random sample

order to enable model-free variance estimation, the PUs could be selected at random instead of systematically, while maintaining Random Grid Sampling in the second stage. In that case, the variance can be estimated in the same way as with basic Two-Stage Random Sampling.

In devising a compound strategy, there are often good reasons to stratify the area first, and then to decide which designs will be applied in the strata. It is not necessary to have the same type of design in each stratum. As long as the stratum means and their variances are estimated without bias, these estimates can be combined into unbiased overall mean and variance estimates using the formulas given in Sect. 7.2.4.

If a variogram for the area is available, the variance of a compound strategy can be predicted prior to sampling, using the Monte-Carlo simulation tech-

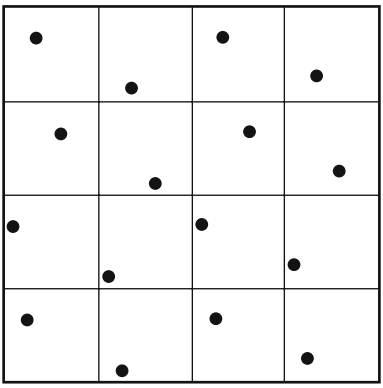


Fig. 7.18. Notional example of a systematic unaligned sample

nique presented in Sect. 7.2.15. In the case of stratification, this technique can be applied to each stratum separately, using different variograms if necessary.

Spatial Systematic Strategies

Most strategies discussed so far are spatial in the sense that primary units and clusters are defined on the basis of geographical coordinates. Strata are also usually defined this way. Given these definitions, however, the random selection restrictions do not refer to the coordinates of sampling locations. A category of more inherently spatial strategies exists whose random selection restrictions make explicit use of coordinates or distances in geographical space. Two examples are given.

Figure 7.18 shows a 'systematic unaligned' sample. This technique was proposed by Quenouille (1949). The area is first divided into square strata and one location is selected in each stratum, although not independently. A random s_1 coordinate is generated for each row of strata, and a random s_2 coordinate for each column. The sampling location in a stratum is then found by combining the coordinates of its row and column. Notice in Fig. 7.18 the irregular, but still fairly even spread of the locations.

Figure 7.19 shows a 'Markov Chain' sample, a technique discussed by Breidt (1995). Again, notice the irregular but fairly even spread of the locations. The underlying principle is that the differences between the coordinates of consecutive locations are not fixed, as with systematic unaligned samples, but stochastic. These differences have a variance which is determined by a parameter ϕ, chosen by the user. Thus Markov Chain designs form a class in which one-per-stratum Stratified Simple Random Sampling and Systematic Unaligned Sampling are special cases, with $\phi = 0$ and $\phi = 1$, respectively. The example in Fig. 7.19 was generated with $\phi = 0.75$. The purpose of this type of strategies is to allow enough randomness to avoid the risk of interfer-

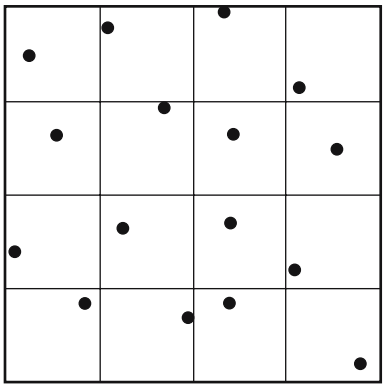


Fig. 7.19. Notional example of a Markov chain sample

ence with periodic variations and linear artefacts like roads, ditches, cables and pipelines, while still maintaining an even spread of the locations over the area as much as possible.

Random Geographical Stratification

In Sect. 7.2.4 a method is described for creating a probability sample with good spatial coverage based on geographical stratification. However, if only one location is selected per stratum, design-based estimation of the sampling variance of the estimator is problematic. To overcome this problem, scientists of the Environmental Monitoring and Assessment Program (EMAP) of the U.S.A. proposed a method in which not only the locations are selected at random, but also the stratification itself is random (Stevens, 1997; Stevens and Olsen, 2003a,b). Figure 7.20 illustrates this method. A lattice of squares is placed over the study region. Then a location is selected fully randomly from the lower-left square, and the lattice is shifted so that the origin of the lattice coincides with this location. Finally, one location is selected at random from each square of the shifted lattice. Note that in this case the original lattice must have one extra column at the left and one extra row at the bottom to ensure that the shifted lattice covers the entire region. The polygons need not be squares, but they must be radially symmetric about a centre point and translation congruent. In EMAP, hexagons are used formed by tessellation of a triangular grid with a spacing of about 27 km. This method can also be used to select discrete objects such as lakes (Stevens and Olsen, 2003a). Each object is represented by a point in geographical space (for instance the centroid or the label point in GIS-files is taken as the reference point), and each object is assigned to the polygon covering its reference point. A simple selection method is to select randomly one object from each randomly placed polygon with at least one object. Unfortunately, the usual (Horvitz-Thompson and

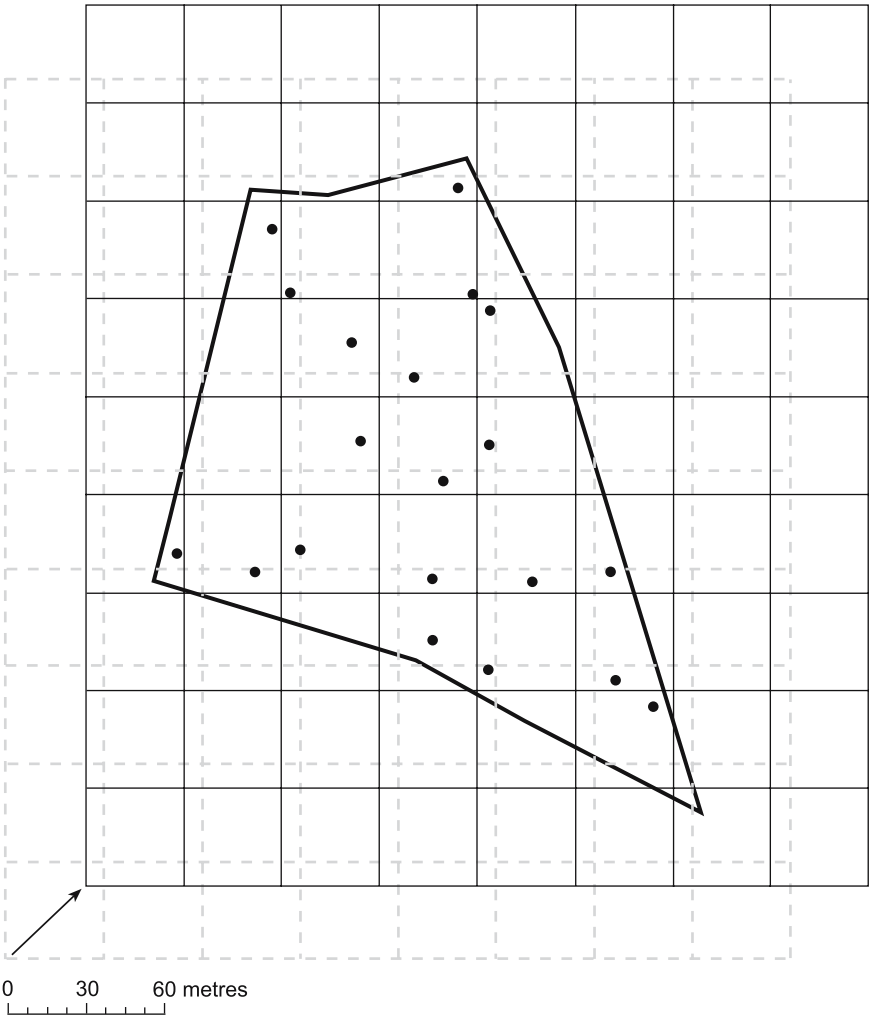


Fig. 7.20. Random stratification of a field followed by Simple Random Sampling of one location per stratum.

Yates-Grundy) variance estimators are unstable. Alternative, contrast-based variance estimators have been proposed by (Stevens and Olsen, 2003b).

7.2.9 Probabilities-Proportional-to-Size Sampling

A straightforward method for using an exhaustively known ancillary variable in sampling is to select units with probabilities proportional to size (pps), i.e., proportional to an ancillary variable x that must be strictly positive. If x is

proportional to the target variable z, then considerable gain in precision will be achieved. However, if this proportionality does not hold, then pps-sampling will be less precise. For finite universes, the ancillary variable is cumulated, i.e. set $T_0 = 0$, and compute $T_k = T_{k-1} + x_k$ for $k = 1, \ldots, N$. Then one number r is drawn from the uniform (0,1) distribution. The selected unit is the one for which $T_{k-1} < rT_N \leq T_k$, where $T_N = \sum_N x_k$, the cumulative total. The selected unit is replaced, and the process is repeated. In this case, the probability that unit k is selected in one draw p_k equals $\frac{x_k}{T_N}$, and the mean can be estimated by

$$\hat{\bar{z}}_{\text{pps}} = \frac{1}{Nn} \sum_{i=1}^{n} \frac{z_i}{p_i}, \tag{7.35}$$

where n is the number of draws (which may be larger than the number of selected units). The sampling variance can be estimated by

$$\widehat{V}\left(\hat{\bar{z}}_{\text{pps}}\right) = \frac{1}{N^2\, n\,(n-1)} \sum_{i=1}^{n} \left(\frac{z_i}{p_i} - N\, \hat{\bar{z}}_{\text{pps}} \right)^2. \tag{7.36}$$

The above estimators are very simple thanks to the replacement of selected units. However, simplicity can be at the cost of efficiency, and therefore we may prefer sampling without replacement. If we do not replace selected units, we must be aware that it is very difficult to accomplish inclusion probabilities (probability that a unit is selected in the sample exactly proportional to x_k. Sunter (1977) describes a selection scheme leading to exactly proportional inclusion probabilities for most of the sampling units, not for all units, but in practice this may not bother us. An alternative is to partition the sampling units in n random groups, and to select from each group one unit with probabilities equal to x_k/T_G, where T_G is the cumulative total for the group (Särndal et al., 1992, p.99). For infinite (spatial) universes, the probability that a unit is selected more than once is 0, and there is no reason for not using the above simple estimators. In practice, the universe will often be discretized, and one must make this discretization grid fine enough.

Systematic pps-Sampling

In Systematic Random Sampling (Sect. 7.2.7) the inclusion probabilities are equal for all units. Units can be selected systematically with probabilities proportional to an ancillary variable by the following procedure (Särndal et al., 1992, p. 96):

1. cumulate the values of the ancillary variable x_k (for simplicity we assume that every x_k is an integer);
2. choose a sampling interval, an integer, say a;
3. select with equal probability an integer, say r, between 1 and a (1 and a included).

The first unit selected is the one for which $T_{k-1} < r \leq T_k$, the second unit is the one for which $T_{k-1} < r+a \leq T_k$, for the third we have $T_{k-1} < r+2a \leq T_k$ etc. Let n be the integer part of T_N/a, and $c = T_N - na$. If $c > 0$, then the sample size is either n or $n+1$. The mean can be estimated by

$$\hat{\bar{z}}_{\mathrm{SYpps}} = \frac{1}{N} \sum_{i=1}^{n} \frac{z_k}{\pi_k}, \tag{7.37}$$

where π_k is the inclusion probability of unit k:

$$\pi_k = \frac{n\, x_k}{T_N - c}. \tag{7.38}$$

As in Systematic Random Sampling with equal probabilities, estimation of the variance of the mean is cumbersome.

7.2.10 Sequential Random Sampling

In sections 7.2.3, 7.2.4, 7.2.5 and 7.2.6 prior estimates of the variance components are used to determine the sample size needed to estimate the mean (fraction) with a particular precision. In situations where no reliable prior estimates of this variance are available, one may sample sequentially, i.e., in two or more steps, and use the estimated variance based on the sample data of the previous step to calculate the required sample size. This is done to achieve better control of the precision of the estimated mean (fraction) based on the final sample. Another important application of Sequential Random Sampling is hypothesis testing.

The sample collected in the first step is used to estimate the variance and, hence, the sample size required to estimate the mean (fraction) with the prescribed precision, or to test the hypothesis with the prescribed power at a given confidence level. If this required sample size is larger than the sample size of the first step, then sampling is continued, otherwise it is stopped. Many types of Sequential Random Sampling design have been described in the literature, differing in the stopping rule, the number of steps, and the sample sizes in these steps. A distinction can be made into classical sequential designs and group sequential designs (Whitehead, 1997). In classical sequential designs after a group (batch) of sampling units has been observed, sampling units are added and observed one-at-a-time, and the precision of the estimated mean (fraction) or power of the test is re-calculated continuously. In group sequential designs, groups of sampling units are collected and observed. If the aim is to estimate the mean with variance V from two groups of data sequentially collected by Simple Random Sampling, the required sample size can be calculated from (Cochran, 1977, section 4.7)

$$n = \frac{\widehat{s_1^2}}{V} \left(1 + \frac{2}{n_1}\right) . \tag{7.39}$$

The sample size required to estimate the fraction with variance V equals

$$n = \frac{\widehat{P}_1\left(1-\widehat{P}_1\right)}{V} + \frac{3-8\widehat{P}_1\left(1-\widehat{P}_1\right)}{\widehat{P}_1\left(1-\widehat{P}_1\right)} + \frac{1-3\widehat{P}_1\left(1-\widehat{P}_1\right)}{V\,n_1}\,, \tag{7.40}$$

where $\widehat{P}_1$ is the fraction estimated from the first sample.

The usual binomial estimate of P based on the n data of the final sample is biased. An unbiased estimate can be obtained from

$$\widehat{P}_{\text{unbiased}} = \widehat{P} + \frac{V\left(1-2\widehat{P}\right)}{\widehat{P}\left(1-\widehat{P}\right)}\,. \tag{7.41}$$

Cochran (1977, section 4.7) also provides formulas for calculating the final sample size if we want to estimate the mean or fraction with a prescribed coefficient of variation with two-step Sequential Random Sampling (with Simple Random Sampling in both steps).

A transition type in between Classical Sequential Random Sampling and Group Sequential Random Sampling is Accelerated Sequential Random Sampling, in which after an initial group of sampling units, more units are added one-at-a-time until a fraction of the required sample size is reached, followed by a final group of sampling units (Mukhopadhyay et al., 1992).

Adaptive Cluster Sampling

In ecological surveys of rare species, it is worth considering the option of selecting sampling units with an Adaptive Cluster Sampling design (Thompson and Seber, 1996). In Adaptive Cluster Sampling, whenever the observed value of a selected sampling unit satisfies a certain condition, sampling units are added to the sample from the neighbourhood of that unit. Adaptive Cluster Sampling is therefore a type of sequential sampling. The difference with sequential sampling described in the previous section is that in Adaptive Cluster Sampling the observations plus the locations of these observations are used to determine the locations of observations in successive steps, rather than only the number of new observations (sampling units).

Figure 7.21 illustrates the basic idea. The aim is to estimate the abundance (total number of individuals: plants, animals, mineral deposits, etc.), or alternatively the density of individuals, i.e., the average number of individuals per unit of area. In the first phase nine cells were selected by Systematic Random Sampling. For each of these cells with at least one individual, the four adjacent cells have been added to the sample. This procedure was repeated until none of the new cells contained an individual. We could have used a different definition of the neighbourhood, for instance the eight cells that share a corner with a cell, leading to larger sample sizes, or a different condition for additional sampling, for instance $z \geq 4$, leading to smaller sample sizes.

To understand the estimators given below, the meaning of the terms *cluster, network* and *edge unit* must be established. The collection of all units that are selected by the design as a result of the initial selection of unit i is referred to as a *cluster*. A *network* is a subset of units within a cluster with the property that selection of any unit within the network would lead to the inclusion of every other unit in that network. In Fig. 7.21 the units within a given cluster that contain at least one point-object form a network. Any unit not satisfying the condition but located in the neighbourhood of one that does is termed an *edge unit*. For the sake of convenience, a unit not satisfying the condition is regarded as a network including only itself, so that the population can be uniquely partitioned into networks.

The initial sample can be selected by various designs. To keep matters simple, we shall present the estimators for the mean and its variance for Simple Random Sampling without replacement, while for other designs we refer to Thompson and Seber (1996).

For sampling designs in which the probability π_i that unit i is included in the sample is known for every unit $i = 1, \ldots, N$, the Horvitz-Thompson estimator (π-estimator) $\bar{z} = \frac{1}{N} \sum z_i/\pi_i$ can be used as an unbiased estimator of the spatial mean. For Simple Random Sampling without replacement

$$\pi_i = 1 - \frac{\binom{N-m_i-a_i}{n}}{\binom{N}{n}} , \tag{7.42}$$

where m_i is the number of units in the network that includes unit i and a_i is the number of units in the networks of which unit i is an edge unit. The numerator in the quotient of (7.42) is the number of distinct (unordered) samples of size n not including unit i that can be formed by selecting n units out of N units by Simple Random Sampling without replacement, while the denominator is the total number of distinct (unordered) samples. Hence, the quotient of (7.42) is the probability that unit i is not included. We now face the problem that the inclusion probabilities are not known for all units in the sample because some of the a_i may be unknown (a unit in the sample may be an edge unit of a cluster not represented in the sample). Therefore, the π-estimator is modified by making use of units not satisfying the condition only when they are included in the initial sample.

For Simple Random Sampling without replacement, the mean can be estimated by

$$\hat{\bar{z}}_{\mathrm{ACl}} = \frac{1}{N} \sum_{i=1}^{\nu} \frac{z_i \, I_i}{\alpha_i} , \tag{7.43}$$

where ν is the number of distinct units in the final sample and I_i is an indicator variable having a value of 1 if unit i is selected in the initial sample or if unit i satisfies the condition; otherwise $i_i = 0$, and π_i' is the probability that unit i is included in the initial sample:

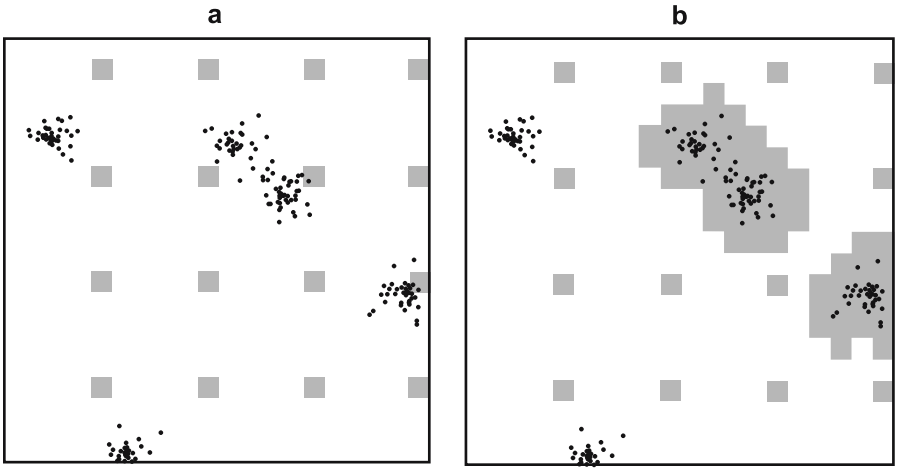


Fig. 7.21. Adaptive cluster sample. The initial sample is a systematic random sample of 16 squares (**a**). In the successive stages the four neighbouring cells of a selected cell are added to the sample if this cell contains at least one point (**b**)

$$\pi_i' = 1 - \frac{\binom{N-m_i}{n_1}}{\binom{N}{n_1}} \,, \tag{7.44}$$

where n_1 is the size of the initial sample and m_i is the number of units in the network that includes unit i. Note that for a unit in the initial sample not satisfying the condition, $m_i = 1$. The variance of $\hat{\bar{z}}$ can be estimated by

$$\widehat{V}\left(\hat{\bar{z}}_{\mathrm{ACl}}\right) = \frac{1}{N^2} \sum_{k=1}^{\kappa} \sum_{m=1}^{\kappa} z_k^* \, z_m^* \, \frac{\pi_{km} - \pi_k \, \pi_m}{\pi_k \, \pi_m \, \pi_{km}} \,, \tag{7.45}$$

where κ is the number of distinct networks represented in the initial sample, z_k^* is the total of the z-values in network k, and π_{km} is the probability that the initial sample contains at least one unit in each of the networks k and m, which for $k \neq m$ can be calculated by

$$\pi_{km} = 1 - \frac{\binom{N-m_k}{n_1} + \binom{N-m_m}{n_1} - \binom{N-m_k-m_m}{n_1}}{\binom{N}{n_1}} \tag{7.46}$$

and for $k = m$

$$\pi_{kk} = \pi_k = 1 - \frac{\binom{N-m_k}{n_1}}{\binom{N}{n_1}} \,. \tag{7.47}$$

7.2.11 Using Ancillary Information in Estimation

In many cases ancillary information is available that could be useful to increase the precision of the estimated mean. The ancillary variable(s) can be qualita-

tive (i.e., classifications) or quantitative. Both types of ancillary variable can be used at the selection stage to improve the performance of the sampling strategy, for instance by stratification (Stratified Simple Random Sampling, see Sect. 7.2.4) or, in case of a quantitative variable, by selecting locations with probabilities proportional to the ancillary variable (pps-Sampling). It may happen that one has 'forgotten' to do so, regrets this, and now wonders whether one can still use this ancillary information to increase the precision of the estimates of the mean. In other cases it may be deliberately decided not to use the ancillary information in the selection. For instance, one may have decided not to use a variable for stratification, because the spatial boundaries of the strata in the area are not known. An example is land use, which may not be known for all parts of the area, whereas it can easily be recorded at the sampling locations during the fieldwork. Another reason for not stratifying the area may be that there are many target variables, which would lead to different stratifications. Multiple stratification would lead to numerous strata. On the other hand, stratification for one target variable may improve estimates of this and correlated target variables, but can be inefficient for other variables. A flexible alternative in this case is to use the ancillary information at the estimation stage.

Post-Stratification

In many cases, the area can be split up into several more or less homogeneous sub-areas. Such sub-areas can be thought of as groups of possible sampling locations that have less variation within them than in the area as a whole. In sampling terminology, these sub-areas are called 'groups'. Even if groups are not used as strata at the selection stage, however, they can still be used at the estimation stage. The only thing that needs to be done is to classify the sampling locations, i.e., determine the group of each sampling location. The mean or the areal fraction can then be estimated by the post-stratification estimator:

$$\hat{\bar{z}}_{\text{pos}} = \sum_{g=1}^{G} a_g \frac{\sum_{i=1}^{n_g} \frac{z_i}{\pi_i}}{\sum_{i=1}^{n_g} \frac{1}{\pi_i}} = \sum_{g=1}^{G} a_g \frac{\sum_{i=1}^{n_g} \frac{z_i}{\pi_i}}{\widehat{N}_g} \; . \tag{7.48}$$

where a_g is the relative size of group g, n_g is the number of sampling locations in group g and π_i is the probability of sampling location i being drawn. In this formula, the estimated group means are weighted by their relative sizes a_g, which are assumed to be known. In spite of this, the group means are estimated by dividing the estimated group totals by their *estimated* size, $\widehat{N}_g$, because this ratio estimator is more precise than the group sample mean.

For Simple Random Sampling, the post-stratification estimator reduces to:

$$\hat{\bar{z}}_{\text{pos}} = \sum_{g=1}^{G} a_g \, \bar{z}_{\text{s}_g} \ , \tag{7.49}$$

where $\bar{z}_{\text{s}_g}$ is the sample mean of group g. In the case of a stratified simple random sample, for which one wants to use a second grouping at the estimation stage, the mean can be estimated by:

$$\hat{\bar{z}}_{\text{pos}} = \sum_{g=1}^{G} a_g \, \hat{\bar{z}}_g = \sum_{g=1}^{G} a_g \sum_{h=1}^{H_g} \frac{\widehat{N}_{gh}}{\widehat{N}_g} \, \bar{z}_{\text{s}_{gh}} \ , \tag{7.50}$$

where H_g is the number of strata in group g, and $\widehat{N}_g$ and $\bar{z}_{\text{s}_{gh}}$ are the estimated size and the sample mean of group g in stratum h, respectively.

For Simple Random Sampling, the variance of the post-stratification estimator for the mean can be estimated by the formula proposed by Särndal et al. (1992, p. 287).

$$\widehat{V}\left(\hat{\bar{z}}_{\text{pos}}\right) = \sum_{g=1}^{G} a_g^2 \frac{\widehat{S_g^2}}{n_g} \ , \tag{7.51}$$

where $\widehat{S_g^2}$ is the estimated spatial variance of z in group g, which for Simple Random Sampling can be estimated by:

$$\widehat{S_g^2} = \frac{1}{n_g - 1} \sum_{i=1}^{n_g} (z_i - \bar{z}_{\text{s}_g})^2 \ . \tag{7.52}$$

Equation (7.51) is an estimator of the conditional sampling variance, i.e., of the variance of the post-stratification estimator over only those possible samples that have group sample sizes equal to the sizes in the sample actually drawn. The post-stratification estimator requires that the sizes (area) of the strata are known. If these sizes are unknown, then these sizes can first be estimated from a relatively large sample, see Sect. 7.2.12 (Two-Phase Random Sampling for Stratification).

Regression Estimators

This section describes how *quantitative* ancillary information can be used at the estimation stage. If the ancillary information is known everywhere, for instance from remote sensing or a digital terrain model, the true means of the ancillary variables are known. The spatial mean of the target variable, $\bar{z}$, can then be estimated by the general regression estimator :

$$\hat{\bar{z}}_{\text{gr}} = \hat{\bar{z}}_{\pi} + \sum_{q=1}^{Q} b_q \left(\bar{x}_q - \hat{\bar{x}}_{q\pi}\right) \ , \tag{7.53}$$

where
$\hat{\bar{z}}_{\pi}$ is the mean of the target variable estimated from the measurements of the

target variable z in the probability sample only;
$\hat{\bar{x}}_{q\pi}$ is the estimated mean of the q-th ancillary variable;
$\bar{x}_q$ is true mean of the q-th ancillary variable; and
b_q is the estimated regression coefficient (slope) for the q-th ancillary variable.

The estimators $\hat{\bar{z}}_\pi$ and $\hat{\bar{x}}_{q\pi}$ are the design-specific estimators for the mean, presented in the previous sections. For instance, for Simple Random Sampling, $\hat{\bar{z}}_\pi$ is the unweighted sample mean $\hat{\bar{z}}_{\mathrm{SI}}$, while for Stratified Simple Random Sampling, it is the weighted mean of the sample means per stratum $\hat{\bar{z}}_{\mathrm{St}}$. Hence, (7.53) is general in the sense that it can be used for any probability design. It can also be used for a single ancillary variable ($Q = 1$), leading to the simple regression estimator, or for two or more ancillary variables (multiple regression estimator). As with spatial means, the sampling design must be taken into account when estimating the regression coefficients. For Simple Random Sampling, the regression coefficients can be estimated by the well-known least squares estimator (Draper and Smith, 1981):

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{z}) \ , \tag{7.54}$$

where
$\mathbf{b}$ is the p-vector of regression coefficients;
$\mathbf{X}$ is the $(n \times p)$ matrix with the values of the ancillary variables at the sampling locations (the first column is a vector with ones); and
$\mathbf{z}$ is the n-vector with values of the target variable at the sampling locations.
The first regression coefficient is the intercept, which is not used in the regression estimator of the mean, but is needed to estimate the variance.

For Stratified Simple Random Sampling, (7.54) is used to estimate the regression coefficients per stratum. If the number of sampling locations selected in a stratum is small, say $n_h < 10$ to 20 (depending on the number of regression coefficients), this stratum must be combined with others to obtain valid estimates of the sampling variance. For a combination of strata, the regression coefficients can be estimated by the weighted means of the coefficients per stratum, using the relative areas as weights (7.13). In the case of Two-Stage Random Sampling, (7.54) is used to estimate the coefficients per primary unit selection, after which the unweighted mean of the coefficients is calculated (7.27). Similarly, in the case of Cluster Random Sampling, the regression coefficients are estimated by the unweighted mean of the estimated regression coefficients per cluster. Finally, in the case of Systematic Random Sampling the vector can be simply estimated by (7.54).

Derivation of the sampling variance of the regression estimator is not straightforward because the regression coefficients are also estimated from the sample, and as a result the regression estimator is non-linear. However, for moderate to large samples (say, for the simple regression estimator $n > 25$) the variance of the regression estimator can be approximated by the variance of the π-estimator for the mean of the residuals, defined as:

$$\mathbf{e} = \mathbf{z} - \mathbf{X}\mathbf{b} \ . \tag{7.55}$$

The variance of the π-estimator for the mean of the residuals can be estimated by the usual sampling design specific estimators of the previous chapters, substituting the target variable z by the residual e. For instance, the approximate variance of the regression estimator for Stratified Simple Random Sampling can be estimated by (compare (7.14)):

$$\widehat{V}\left(\hat{\bar{z}}_{\mathrm{gr}}\right) = \sum_{h=1}^{H} a_h^2 \, \widehat{V}(\bar{e}_{\mathrm{s}_h}) \; , \tag{7.56}$$

where $\widehat{V}(\bar{e}_{\mathrm{s}_h})$ is the estimated variance of the sample mean of the residuals in stratum h:

$$\widehat{V}(\bar{e}_{\mathrm{s}_h}) = \frac{1}{n_h \, (n_h - p_h)} \sum_{i=1}^{n_h} (e_{hi} - \bar{e}_{\mathrm{s}_h})^2 \; . \tag{7.57}$$

where p_h is the number of regression coefficients for stratum h. Note that in (7.57) $n_h \, (n_h - 1)$ is replaced by $n_h \, (n_h - p_h)$ because we have lost p_h degrees of freedom to estimate the mean of the residuals for stratum h. If strata are combined, n_h^2 is substituted in the denominator of (7.57) to estimate the variance per stratum. To correct for the loss of degrees of freedom, the weighted mean of the estimated variances per stratum (7.56) is multiplied by a factor $n/(n-p)$, where p is the number of regression coefficients for the combined strata.

A special case of the regression estimator is the ratio estimator. It follows from the assumption that the intercept is 0 and that the variance of the target variable is proportional to the ancillary variable[2]. The ratio estimator equals:

$$\hat{\bar{z}}_{\mathrm{ra}} = \bar{x} \, \frac{\hat{\bar{z}}_{\pi}}{\hat{\bar{x}}_{\pi}} \; . \tag{7.58}$$

In words: the spatial mean of the target variable is estimated by multiplying the known spatial mean of the ancillary variable by the ratio of the estimated spatial means of the target and ancillary variables. Again, this ratio estimator can be used for any probability design, substituting the design-specific estimators for the mean in (7.58). For small samples, one may prefer the ratio estimator over the regression estimator even in situations where it is unlikely that the intercept is 0, because with small samples the quality of the estimated sampling variance of the ratio estimator outperforms that of the regression estimator. Unlike the model-based approach, the quality of the estimated sampling variance and confidence intervals does not depend on the quality of the model assumptions. This is because the roles of the regression model in the design-based approach and of the spatial variation model in the model-based

[2] In (7.54) it is assumed that the variance of the target variable is constant (homoscedasticity). If heteroscedasticity is assumed, the regression coefficients can be estimated by the Weighted Least Squares estimator using the inverse of the variance as weights.

approach differ fundamentally. The regression model merely serves to choose the type of regression estimator. The statistical inference is still based on the selection probabilities of the samples as determined by the sampling design. In the model-based approach, the model introduces the source of randomness, and inference is based on this stochastic model (see Sect. 4.1). For an extensive discussion of this, we refer to Hansen et al. (1983) and Särndal et al. (1992). Brus (2000) discussed this issue in the context of soil sampling.

7.2.12 Two-Phase Random Sampling

The post-stratification estimator of section 7.2.11 requires that the sizes of the strata are known and, similarly, the regression estimator of Sect. 7.2.11 requires that the means of the ancillary variables are known. If these are unknown, but the ancillary variable can be measured cheaply, one may decide to estimate the spatial means of the ancillary variables (sizes of the strata) from a large sample. The target variable is measured in a random subsample of this large sample only. This technique is known in the sampling literature as 'Two-Phase Random Sampling' or 'Double Sampling'. Three Double Sampling strategies are presented hereafter: Two-Phase Random Sampling for Stratification, Ranked Set Sampling, and Two-Phase Sampling for Regression. The requirements on the ancillary variable increase in this order. For the stratification a nominal variable suffices. To rank the first-phase sample one must have measurements on at least an ordinal scale. The regression estimator requires a quantitative variable. For a comparison of the three Two-Phase Random Sampling techniques described below, we refer to Patil et al. (1993) and Mode et al. (2002).

Two-Phase Random Sampling for Stratification

When the sizes of the strata are unknown, one can take a large sample and classify the selected sampling units. The classes thus formed are then used as strata in the second sampling phase, and the target variable is measured on the units in the resulting subsample. See Fig. 7.22 for a notional example. With Simple Random Sampling in the first phase and Stratified Simple Random Sampling in the second phase, the mean can be estimated by:

$$\hat{\bar{z}}'_{\pi} = \sum_{h=1}^{H_{\mathrm{s}_1}} \frac{n_{1h}}{n_1} \bar{z}_{\mathrm{s}_h} \ , \tag{7.59}$$

where H_{s_1} is the number of strata of the first-phase sample, n_{1h} is the number of locations in the first-phase sample that form stratum h, n_1 is the total number of locations of the first phase sample and $\bar{z}_{\mathrm{s}_h}$ is the mean of the subsample from stratum h. The variance can be estimated by the formula proposed by Särndal et al. (1992, p. 353):

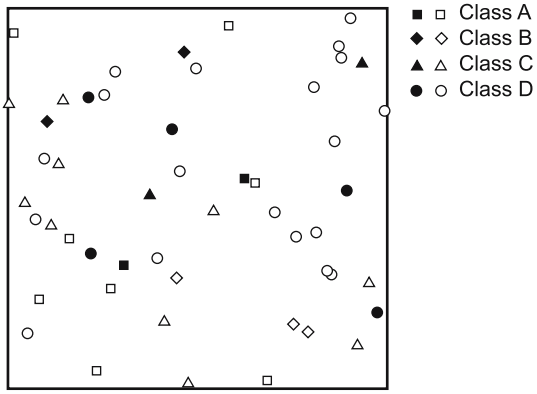


Fig. 7.22. Notional example of a two-phase random sample for stratification, the first-phase sample being selected by Simple Random Sampling. The solid symbols represent the second-phase sample.

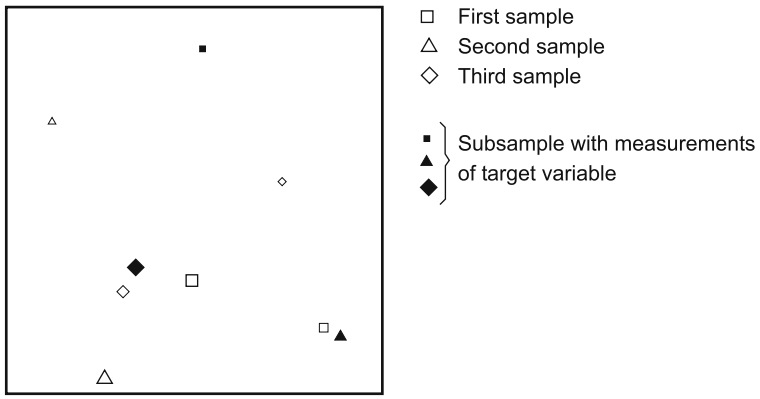


Fig. 7.23. Ranked set sample (one replication only). First-phase sample is selected by Simple Random Sampling. The ranking is represented by the size of the symbols. The solid symbols represent the second-phase sample with measurements of target variable.

$$\widehat{V}(\hat{\bar{z}}'_{\pi}) = \sum_{h=1}^{H_{s_1}} \left(\frac{n_{1h}}{n_1}\right)^2 \frac{S^2_{\mathrm{s}_h}}{n_{2h}} + \frac{1}{n_1} \sum_{h=1}^{H_{s_1}} \frac{n_{1h}}{n_1} \left(\bar{z}_{\mathrm{s}_h} - \hat{\bar{z}}\right)^2 \ . \tag{7.60}$$

Ranked Set Sampling

Ranked Set Sampling is a way of increasing precision of estimates of the mean or total, by ranking randomly selected sampling units. After this ranking, a subset of sampling units is selected for measurement of the target variable. When measurement of the target variable is expensive, and the ranking of the sampling units is cheap and effective, considerable gains in precision can be achieved compared to Simple Random Sampling. The ranking can be done directly by eye, or with the help of some cheap measurement of a covariate, such as electromagnetic readings, remotely sensed data, chemically-responsive papers etc. In the simplest version m^2 sampling units are selected, partitioned into m simple random samples of size m (Fig. 7.23). Next, the m units of each partition are ranked. Finally, the unit with the smallest rank in the first partition, the unit with the second smallest rank in the second partition until the unit with the largest rank in the final partition, are selected for measurement of the target variable. This process is replicated r times.

The number of sampling units per partition must not be too large, otherwise ranking may become difficult and less effective. Typical values for m are two, three or four. For instance, if one wants to measure the target variable on 12 sampling units, then a simple random sample of size nine (partitioned into three groups of size three) may be selected four times (m=3, r=4). An alternative would be to select two times ($r = 2$) a sample of size 36, partitioned into six groups of size six ($m = 6$), but in general ranking six units will be more difficult and less precise than ranking three sampling units.

The mean can simply be estimated by the unweighted mean of the target variable in the subsample

$$\hat{\bar{z}}_{\mathrm{RSS}} = \frac{1}{r}\sum_{j=1}^{r}\frac{1}{m}\sum_{i=1}^{m} z_{(i)j} = \frac{1}{r}\sum_{j=1}^{r}\hat{\bar{z}}_{\mathrm{RSS}j} = \frac{1}{m}\sum_{i=1}^{m}\hat{\bar{z}}_{(i)} \;, \tag{7.61}$$

where $z_{(i)j}$ is the measurement of the target variable on the sampling unit with ranking i in the ith partition, in the jth cycle. The variance can be estimated by (Chen et al., 2004)

$$\widehat{V}\big(\hat{\bar{z}}_{\mathrm{RSS}}\big) = \frac{\widehat{V}\big(\hat{\bar{z}}_{\mathrm{RSS}j}\big)}{r} = \frac{\frac{1}{r-1}\sum_{j=1}^{r}\big(\hat{\bar{z}}_{\mathrm{RSS}j} - \hat{\bar{z}}_{\mathrm{RSS}}\big)^2}{r} \;. \tag{7.62}$$

An alternative variance estimator is

$$\widehat{V}\left(\hat{\bar{z}}_{\mathrm{RSS}}\right) = \frac{1}{r\,m^2}\sum_{i=1}^{m}\widehat{S^2_{(i)}} \;, \tag{7.63}$$

with

$$\widehat{S^2_{(i)}} = \frac{1}{r-1}\sum_{j=1}^{r}\big(z_{(i)j} - \hat{\bar{z}}_{(i)}\big)^2 \;. \tag{7.64}$$

Note that for both variance estimators at least two replications are needed. For details on theory, methods and applications we refer to Chen et al. (2004) and Patil et al. (1994).

Two-Phase Random Sampling for Regression

For Two-Phase Random Sampling, the regression estimator is:

$$\hat{\bar{z}}'_{\mathrm{gr}} = \hat{\bar{z}}'_{\pi} + \sum_{q=1}^{Q} b_q \left(\hat{\bar{x}}_{q\pi} - \hat{\bar{x}}'_{q\pi} \right) \ , \tag{7.65}$$

where
$\hat{\bar{z}}'_{\pi}$ is the mean of target variable estimated from the measurements of the target variable z in the subsample;
$\hat{\bar{x}}'_{q\pi}$ is the mean of the q-th ancillary variable estimated from the subsample;
$\hat{\bar{x}}_{q\pi}$ is the mean of the q-th ancillary variable estimated from the first phase sample;
b_q is the regression coefficient for the q-th ancillary variable estimated from the subsample.

The estimation of the spatial means of the target variable and the ancillary variables from the subsample needs further explanation. A general p-unbiased estimator of the mean for Two-Phase Random Sampling is:

$$\hat{\bar{z}}'_{\pi} = \frac{1}{N} \sum_{i=1}^{n_2} \frac{z_i}{\pi_{1i}\, \pi_{2i}} \ , \tag{7.66}$$

where
N is the total number of possible locations in the area;
n_2 is the number of locations selected in the second phase;
π_{1i} is the probability that location i is included in the first phase sample; and
π_{2i} is the conditional probability that location i is included in the second-phase sample given the first phase sample.

The inclusion probabilities are determined by the sampling designs in the two phases. Table 7.1 presents these probabilities for some combinations of designs. The inclusion probabilities in Table 7.1 assume that, in the case of Simple Random Sampling in phase 2, the locations are selected without replacement, and in the case of Stratified Simple Random Sampling in phase 2, the strata coincide with the strata, PU drawings or clusters used in phase 1.

The regression coefficients must be estimated from the subsample as the target variable is known for the locations of the subsample only. These can be estimated in a similar way as the spatial means, using the inverse of the product of the inclusion probabilities as weights in the Weighted Least Squares estimator:

$$\mathbf{b} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{W}\mathbf{z}) \ , \tag{7.67}$$

Table 7.1. Inclusion probabilities for different combinations of phase 1 and phase 2 designs.

Phase 1 design	Phase 2 design	π_1	π_2
Simple Random Sampling (SI)	SI	n_1/N	n_2/n_1
Stratified Simple Random Sampling (StS)	StS	n_{1h}/N_h	n_{2h}/n_{1h}
Two-Stage Random Sampling (TsS)	SI	nm_{1i}/N	m_2/m_1
Two-Stage Random Sampling	StS	nm_{1i}/N	m_{2i}/m_{1i}
Cluster Random Sampling (ClS)	SI	nm_{1i}/N	m_2/m_1
Cluster Random Sampling	StS	nm_{1i}/N	m_{2i}/m_{1i}
Systematic Random Sampling (SyS)	SI	m_1/N	m_2/m_1

The symbols used in this table have the following meaning:
n_1 is the number of locations selected by SI in phase 1;
n_2 is the number of locations selected by SI in phase 2;
n_{1h} is the number of locations selected in phase 1 from stratum h;
n_{2h} is the number of locations selected in phase 2 from stratum h;
n is the number of PU drawings or selected clusters in phase 1;
m_{1i} is the number of locations (selected in the second stage of TsS in phase 1) in the PU containing the i-th subsampling location, or the number of locations in the cluster (selected by ClS in phase 1) containing the i-th subsampling location;
m_1 is the total number of locations selected in phase 1;
m_{2i} is the number of locations selected in phase 2 in stratum i; and
m_2 is the number of locations selected in phase 2.

Table 7.2. Variance components for different combinations of phase 1 and phase 2 designs (see Table 7.1 for the meaning of abbreviations).

Phase 1	Phase 2	Variance component 1	Variance component 2
SI	SI	$\frac{\widehat{S^2}(z)}{n_1}$	$\left(1-\frac{n_2}{n_1}\right)\frac{\widehat{S^2}(e)}{n_2}$
StS	StS	$\sum_{h=1}^{L} a_h^2 \frac{\widehat{S^2}(z_h)}{n_{1h}}$	$\sum_{h=1}^{L} a_h^2 \left(1-\frac{n_{2h}}{n_{1h}}\right)\frac{\widehat{S^2}(e_h)}{n_{2h}}$
TsS	SI	$\frac{1}{n(n-1)}\sum_{i=1}^{n}(\bar{z}_i'-\bar{z}')^2$ $-\frac{1}{n^2}\sum_{i=1}^{n}(1-\frac{m_2}{m_1})\frac{\widehat{S^2_{\mathrm{W}i}}}{m_2}$	$(1-\frac{m_2}{m_1})\frac{\widehat{S^2}(e)}{m_2}$
TsS	StS	$\frac{1}{n(n-1)}\sum_{i=1}^{n}(\bar{z}_i'-\bar{z}')^2$ $-\frac{1}{n^2}\sum_{i=1}^{n}(1-\frac{m_{2i}}{m_{1i}})\frac{\widehat{S^2_{\mathrm{W}i}}}{m_{2i}}$	$\frac{1}{n^2}\sum_{i=1}^{n}(1-\frac{m_{2i}}{m_{1i}})\frac{\widehat{V}(e_i)}{m_{2i}}$
ClS	SI	as TsS/SI	as TsS/SI
ClS	StS	as TsS/StS	as TsS/StS

where $\mathbf{W}$ is the $n \times n$ matrix with $1/\ (\pi_{1i}\pi_{2i})$ as weights on its diagonal and zeros as off-diagonal elements.

The variance of the regression estimator for Two-Phase Random Sampling equals:

$$V\big(\hat{\bar{z}}'_{\mathrm{gr}}\big) = V_1\big\{E_2\big(\hat{\bar{z}}'_{\pi} \mid s_1\big)\big\} + E_1\big\{V_2\big(\hat{\bar{z}}'_{\mathrm{gr}} \mid s_1\big)\big\}\ , \tag{7.68}$$

where
$V_1(\cdot)$ is the variance over repeated sampling with the design of phase 1;
$E_2(\cdot|s_1)$ is the conditional expectation over repeated sampling with the design of phase 2 given the first-phase sample;
$E_1(\cdot)$ is the expectation over repeated sampling with the design of phase 1;
$V_2(\cdot|s_1)$ is the conditional variance over repeated sampling with the design of phase 2 given the first-phase sample.

The first variance component of (7.68) equals the variance of the estimated mean if the values of the target variable were known for all locations of the first-phase sample. As the values are known for the subsample only, this (weighted) mean of the first-phase sample must be estimated from the subsample, which introduces an extra error, with variance equal to the second variance component of (7.68). Table 7.2 shows estimators for the two variance components for the sampling design combinations of Table 7.1, except for the last one. The following symbols in Table 7.2 need to be defined:
$\widehat{S^2}(e)$ is the estimated residual variance;
$\widehat{S^2}(e_h)$ is the estimated residual variance in stratum h;
$\bar{z}'_i$: subsample mean of the PU or cluster of the i-th selection;
$\bar{z}'$: mean of $\bar{z}'_i$, $i = 1, \ldots, n$.

For the combination of TsS/SI in Table 7.2, it is assumed that the number of locations selected in the second stage of phase 1, each time a PU is selected, is constant (m_{1i}: constant). For the combination ClS/SI it is assumed that the clusters are of equal size, and that the same number of subsampling locations is drawn from each cluster. Brus and te Riele (2001) give an example from soil science of Two-Phase Random Sampling for Regression, with Stratified Simple Random Sampling in both phases.

7.2.13 Line-Transect Random Sampling

Line-Transect Random Sampling is commonly used in surveys of biotic populations (vegetation, animals) to estimate the abundance[3] or the density[4] of individuals, for instance of a specific plant or animal species. An observer moves along one or more randomly selected lines, and records the location or perpendicular distance of every individual of the species detected. In general, more individuals are detected close to a transect than far away from it, simply because individuals at large distance are more difficult to detect. In order to

[3] Total number.
[4] Total number divided by the area

select the transects, a baseline is established across or below the target area. The length of this line is equal to the projection of the area onto the line. Then n points are selected on this baseline, for instance by Simple Random Sampling or Systematic Random Sampling. These points determine the locations of the transects perpendicular to the baseline.

In the simplest version of Line-Transect Random Sampling only the detected individuals in a narrow strip along the line are used in the estimation. This is referred to as the narrow-strip method. It is assumed that the detectability in the narrow strip is perfect, i.e., every individual in the strip is detected with probability 1. Individuals detected outside the narrow strip are ignored, because the detectability of these individuals is < 1. The sampling units generally have different (surface) areas and therefore the density is estimated via the total number of individuals. For Simple Random Sampling of n transects the narrow-strip estimator of the total is

$$\widehat{t}_{\mathrm{ns}}(z) = \sum_{i=1}^{n} z_i \frac{B}{n2w} = \frac{B}{2w}\overline{z}_{\mathrm{s}} \,, \tag{7.69}$$

where z_i is the number of observed individuals in strip i, B is the length of the baseline, w is half the width of the strips, and $\bar{z}_s$ is the sample mean of the numbers. (The term $\frac{B}{n2w}$ in (7.69) is the inverse of the transect inclusion probability density integrated over the width of the strip). The density is then estimated by

$$\widehat{\overline{z}}_{\mathrm{ns}} = \frac{1}{A}\widehat{t}_{\mathrm{ns}}(z) \,, \tag{7.70}$$

where A is the (surface) area of the region.

An alternative estimator, preferable in case of large differences in strip length, is the ratio estimator

$$\widehat{\overline{z}}_{\mathrm{R}} = \frac{\frac{B}{2w}\sum_{i=1}^{n} z_i}{\sum_{i=1}^{n} L_i} \,. \tag{7.71}$$

The variance of (7.69) can be estimated by

$$\widehat{V}\left(\widehat{t}_{\mathrm{ns}}(z)\right) = \left(\frac{B}{2w}\right)^2 \frac{\widehat{S^2}(z_i)}{n} \tag{7.72}$$

with $\widehat{S^2}(z_i)$ the estimated spatial variance of the counts per strip:

$$\widehat{S^2}(z_i) = \frac{1}{n-1}\sum_{i=1}^{n} (z_i - \overline{z}_{\mathrm{s}})^2 \tag{7.73}$$

The variance of the density can be estimated by dividing the estimated variance of the estimated total by A^2.

The narrow-strip estimator is not entirely satisfying, because the detected individuals outside the narrow strip with perfect detectability are not used.

Also, in practice the half-width of the strips, w, is often not known and must be estimated, i.e., one must estimate the maximum distance to which detectability is perfect. This can be done with the help of a histogram of the counts for distance-intervals from the transect, see Thompson (2002).

To get more out of the data, an alternative estimator has been developed. This estimator makes use of the effective half-width of the transects. The effective half-width of a transect with a detectability that decreases with distance is the width of a narrow strip (with perfect detectability) in which on average the same number of individuals would be observed as seen from that transect. This effective half-width can be derived from the above-mentioned histogram. To this end the histogram is scaled to a probability function by dividing the counts in a distance-interval through the total number of counts multiplied by the width of the intervals. The areas of the bars in the probability plot sum to 1. The effective half-width is now the reciprocal value of the estimated probability for zero distance $\hat{f}(0)$. This probability at zero distance can be estimated by fitting a smooth line through the bar-shaped probability plot.

For Simple Random Sampling of n transects, the abundance and density can now estimated by substituting the effective width for the width in (7.69) and (7.70). The variance of the abundance can be estimated by (7.72).

The transects can also be selected with probabilities proportional to their length (pps-Sampling). This can be implemented by selecting n locations simple randomly from the whole target area, and drawing lines through these points perpendicular to the baseline. Statistical inference of the density is now much simpler because larger transects now also have larger selection probabilities. The density can now be estimated by

$$\widehat{\bar{z}}_{\text{pps}} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\bar{z}}_i = \frac{1}{n}\sum_{i=1}^{n}\frac{z_i}{2wL_i}, \tag{7.74}$$

and its variance by

$$\widehat{V}(\widehat{\bar{z}}_{\text{pps}}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(z_i - \widehat{\bar{z}}_{\text{pps}}\right)^2 \tag{7.75}$$

A modification of Line-Transect Random Sampling is Variable-Circular-Plots Random Sampling. By this method observation locations are randomly selected from the area, and from each location every detected individual and its distance to the location is recorded. When the locations are equidistant points on transects selected by Systematic Random Sampling, this method is referred to as Point-Transect Random Sampling. For further details we refer to Thompson (2002).

7.2.14 Line-Intercept Random Sampling

This section deals with the situation where interest is in properties of the individuals (plants, animals), such as the average diameter of trees at breast

height, rather than properties of the terrain, such as the density of trees in the area. In this situation the sampling units are the individuals. A common problem with discrete universes (populations) is the absence of a sampling frame, i.e., a list or a map of all individuals in the universe of interest. If the individuals occupy considerable area, then Line-Intercept Sampling can be used to select individuals in the field. With this type of design a sample of lines is selected, and whenever an individual is intersected by one or more lines, it is measured. Here we present the simplest version of Line-Intercept Random Sampling in which transects, perpendicular to a baseline, are selected through Simple Random Sampling of points on this baseline. For any transect, the probability of selecting an individual k is

$$p_k = \frac{w_k}{B} \,, \tag{7.76}$$

where w_k and B are the length of the projection of individual k and the area, respectively, on the baseline. An unbiased estimator of the mean of the target variable z is

$$\widehat{\overline{z}}_i = \frac{1}{m_i} \sum_{k=1}^{m_i} \frac{z_{ik}}{p_{ik}} \,, \tag{7.77}$$

where m_i is the number of individuals intersected by transect i. For n transects, the population mean can be estimated by the average of the $\widehat{\overline{z}}_i$-values:

$$\widehat{\overline{z}}_{\mathrm{li}} = \frac{1}{n} \sum_{i=1}^{n} \widehat{\overline{z}}_i \,. \tag{7.78}$$

The variance of this estimator can be estimated by

$$\widehat{V}(\hat{\overline{z}}_{\mathrm{li}}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\widehat{\overline{z}}_i - \widehat{\overline{z}}_{\mathrm{li}} \right)^2 . \tag{7.79}$$

Line-Intercept Random Sampling can also be used to estimate the abundance, density or the coverage of individuals. The coverage is defined as the proportion of the total area covered by the individuals, such as vegetation patches. In this case the sampling units are lines (transects), and not individuals. For Simple Random Sampling of transects the total covered length of transects, i.e., the covered area, can be estimated by (compare (7.69))

$$\widehat{t}_{\mathrm{li}}(z) = \sum_{i=1}^{n} z_i \frac{B}{n} = B\overline{z}_{\mathrm{s}} \,, \tag{7.80}$$

where z_i is the covered length of transect i. The coverage is then estimated by

$$\widehat{\overline{z}}_{\mathrm{li}} = \frac{1}{A} \widehat{t}_{\mathrm{li}}(z) \,, \tag{7.81}$$

The sampling variance of (7.80) can be estimated by

$$\widehat{V}\left(\widehat{t}_{\mathrm{li}}(z)\right) = B^2 \frac{\widehat{S^2}(z_i)}{n} \tag{7.82}$$

with $\widehat{S^2}(z_i)$ the estimated spatial variance of the covered transect lengths:

$$\widehat{S^2}(z_i) = \frac{1}{n-1} \sum_{i=1}^{n} (z_i - \overline{z}_{\mathrm{s}})^2 \tag{7.83}$$

The variance of the coverage can be estimated by dividing (7.82) by A^2.

For pps-sampling of transects, an unbiased estimator of the coverage is

$$\widehat{\overline{z}}_{\mathrm{pps}} = \frac{1}{n} \sum_{i=1}^{n} \widehat{\overline{z}}_i = \frac{1}{n} \sum_{i=1}^{n} \frac{z_i}{L_i}. \tag{7.84}$$

The variance can be estimated by (7.75).

7.2.15 Model-Based Optimization of Sample Sizes

If prior information on the spatial variability is available in the form of a variogram, the following method can be used to predict the sampling variance of any design-based strategy. Such predictions can then be used for optimizing sample sizes. The core of the method is the general equation for predicting the variance of a design-based estimated mean from a variogram (Domburg et al., 1994):

$$E_\xi\left\{V_p\left(\hat{\bar{z}}\right)\right\} = \bar{\gamma} - E_p\left(\boldsymbol{\lambda}' \boldsymbol{\Gamma}_{\mathrm{s}} \boldsymbol{\lambda}\right) \ , \tag{7.85}$$

where
$E_\xi(\cdot)$ is the statistical expectation over realizations from the model ξ underlying the chosen variogram;
$E_p(\cdot)$ is the statistical expectation over realizations from the design p;
$V_p(\cdot)$ is the variance over realizations from the design p (the usual sampling variance in the design-based approach);
$\bar{\gamma}$ is the mean semi-variance between two random points in the area;
$\boldsymbol{\lambda}$ is the vector of design-based weights of the locations of a sample selected according to design p (For instance, if one cluster of 3 locations and one of 2 locations were selected, the weights in calculating the mean would be (7.33): 1/6, 1/6, 1/6, 1/4, 1/4);
$\boldsymbol{\Gamma}_{\mathrm{s}}$ is the matrix of semi-variances between the locations of a sample selected according to design p.

The first term, $\bar{\gamma}$, is calculated by numerical integration or by Monte-Carlo simulation, repeatedly selecting a pair of random locations, calculating its semivariance, and averaging. The second term can also be evaluated by Monte-Carlo simulation, repeatedly selecting a sample according to design p, calculating its mean semi-variance $\boldsymbol{\lambda}' \boldsymbol{\Gamma}_{\mathrm{s}} \boldsymbol{\lambda}$, and averaging. This generic procedure is computationally demanding but it is the only option for compound

and spatial systematic strategies (Sect. 7.2.8). For the basic strategies, however, much more efficient algorithms are possible, making use of the structure of the design types. The following special prediction equations can be derived from the general equation (7.85).

Simple Random Sampling

In the case of Simple Random Sampling, (7.85) simplifies to:

$$E_\xi\{V_p(\hat{\bar{z}}_{\mathrm{SI}})\} = \frac{1}{n}\bar{\gamma}\ . \tag{7.86}$$

Stratified Simple Random Sampling

For Stratified Simple Random Sampling, (7.85) becomes:

$$E_\xi\{V_p(\hat{\bar{z}}_{\mathrm{St}})\} = \sum_{h=1}^{H}\frac{\bar{\gamma}_h}{n_h}\ , \tag{7.87}$$

where $\bar{\gamma}_h$ is the mean semi-variance between two random locations in stratum h. Different variograms can be used for the strata.

Two-Stage Random Sampling

For Two-Stage Random Sampling and constant m_i, the sampling variance is given by (7.29). The variance components in this equation are the between-unit and the pooled within-unit variances, S_{b}^2 and S_{w}^2. These components can be predicted from the two terms in (7.85). The first term predicts the total variance, $S_{\mathrm{T}}^2 = S_{\mathrm{b}}^2 + S_{\mathrm{w}}^2$, while the second term predicts $S_{\mathrm{w}}^2/2$ if $n = 1$ and $m = 2$. In other words, the second term is calculated by repeatedly selecting one unit and two random points in it. The result is the mean semi-variance between pairs of random points within units, denoted by $\bar{\gamma}_{\mathrm{u}}$. The sampling variance is then predicted by:

$$E_\xi\{V_p(\hat{\bar{z}}_{\mathrm{Ts}})\} = \frac{1}{n}\left(\bar{\gamma} - \frac{m-1}{m}\cdot\bar{\gamma}_{\mathrm{u}}\right)\ . \tag{7.88}$$

Brus et al. (2002) illustrate the prediction of the sampling variance for two- and three-stage designs, and the optimization of the sample sizes by simulated annealing.

Cluster Random Sampling

The sampling variance with Cluster Random Sampling equals the between-cluster variance, S_{b}^2, divided by the number of clusters, n. To predict S_{b}^2 for a given cluster definition, we apply (7.85) to Cluster Random Sampling with $n = 1$. In other words, the second term is calculated by repeatedly selecting

only one cluster. Within each cluster the locations have equal weight $(1/m_i)$, so that $\boldsymbol{\lambda}' \boldsymbol{\Gamma}_s \boldsymbol{\lambda}$ simplifies to the unweighted mean:

$$\boldsymbol{\lambda}' \boldsymbol{\Gamma}_s \boldsymbol{\lambda} = \frac{1}{m_i^2} \sum_{k=1}^{m_i} \sum_{l=1}^{m_i} \gamma_{kl} = \frac{2}{m_i^2} \sum_{k=1}^{m_i-1} \sum_{l=k+1}^{m_i} \gamma_{kl} \ , \tag{7.89}$$

because $\boldsymbol{\Gamma}_s$ is symmetric with a zero diagonal. The result is the mean semivariance between pairs of locations within clusters, denoted by $\bar{\gamma}_c$. The sampling variance is then predicted by:

$$E_\xi \{ V_p(\hat{\bar{z}}_{Cl}) \} = \frac{1}{n} (\bar{\gamma} - \bar{\gamma}_c) \ . \tag{7.90}$$

Of course, in the special case that all clusters have the same size and shape, $\boldsymbol{\lambda}' \boldsymbol{\Gamma}_s \boldsymbol{\lambda}$ needs to be calculated only once.

Systematic Random Sampling

As Systematic Random Sampling is Cluster Random Sampling with $n = 1$, the sampling variance can be predicted by:

$$E_\xi \{ V_p(\hat{\bar{z}}_{Cl}) \} = \bar{\gamma} - \bar{\gamma}_c \ . \tag{7.91}$$

Again, in the special case that all clusters have the same size and shape, $\boldsymbol{\lambda}' \boldsymbol{\Gamma}_s \boldsymbol{\lambda}$ needs to be calculated only once.

7.2.16 Design-Based Sampling in 1D or 3D Space

The previous sections focused on sampling in 2D space, but in practice 1D or 3D spatial dimensions are often involved. In general, the methodology presented for 2D space can be easily transferred or adapted to these dimensions, as is outlined below.

One-dimensional spatial universes can have a horizontal or a vertical orientation. Examples of horizontal 1D universes are projected trajectories of roads or pipelines. The methodology presented for 2D is directly transferable to this situation. Sampling in vertical 1D space, i.e., sampling at depth, is practically always done at more than one location, hence it is part of sampling in 3D space.

The universe of interest is very often embedded in 3D space. Sampling locations would then have two horizontal coordinates (s_1 and s_2) and one vertical coordinate (s_3). Theoretically, all three could be determined independently, similar to s_1 and s_2 in 2D sampling. This would typically lead to sampling at a single variable depth at each location. However, this is hardly ever done in practice. There are two main reasons to treat the vertical coordinate differently and to decompose the 3D sampling problem into a 2D (horizontal) problem and a 1D (vertical) problem.

The first reason applies when the target variable is defined as a function of soil properties at various depths, as is usually the case in the context of, for instance, plant growth and leaching. It is then logical to sample at these depths at each location. The second reason applies when the target *variable* is defined at points in 3D space, e.g., the concentration of a contaminant, and the target *quantity* is defined over 3D space, e.g., the 3D spatial mean. In this case, although not a necessity, it is usually efficient to take samples at various depths at the same location. The sample is designed and analyzed as a two-stage sample, with locations as the primary units and depths as the secondary units (see Sect. 7.2.5).

The methodology of sampling at depth is, in principle, the same as that for 2D space. However, cluster and two-stage sampling will usually be inefficient, because their operational advantages in 2D space do not hold for sampling at depth. The two dominant techniques in practice are purposive sampling at fixed depths and stratified systematic sampling, with soil horizons as strata and compositing of samples from the same horizon.

7.3 Model-Based Methods for Global Quantities in Space

7.3.1 Introduction

Although model-based methods are typically applied to predict local quantities, such as the values at points or the means of small sub-areas (Sect. 8.3), they can be used for predicting global quantities such as the mean or the cumulative distribution function for the entire study region. In general, a design-based sampling strategy is the most appropriate for estimating global quantities (Sect. 7.2). Broadly speaking, a model-based sampling strategy may be appropriate when the target variable is strongly autocorrelated at the scale of the study region and the sampling density is large enough to profit from this autocorrelation (Sect. 4.1). Another prerequisite is that there must be enough sample data from which to estimate the model. If the quantified uncertainty about the estimated global quantity is important, and one is rather uncertain about the model, a model-based strategy may be refrained from, because the variance of the prediction error is particularly sensitive to the model assumptions.

When a model-based strategy is followed to predict the global spatial mean, the sampling locations need not be selected at random, because randomness is already introduced via the stochastic model of spatial variation (Sect. 4.1). In general, model-based predictions from probability samples are less precise (have larger block-kriging variance) than those from purposive samples, which is why purposive sampling is preferable. If prior information is available on the spatial variation and a plausible model (variogram) can be

postulated, this geostatistical model can be used to search for the sampling locations with minimum block-kriging variance. Samples optimized with a geostatistical model are referred to as geostatistical samples (Sect. 7.3.3). The search for the optimal sample may be rather difficult and time-consuming, and the closeness of the optimum thus found to the true optimum depends on the quality of the postulated variogram. Therefore, in some situations one may decide to follow a much simpler approach for designing a spatial sample to be used in a model-based sampling strategy. Simple alternatives are regular grids and spatial coverage samples. For a discussion of methods to design spatial coverage samples, we refer to Sect. 8.3.3. Nychka et al. (1997) found that in predicting the global mean, spatial coverage samples perform only slightly worse than geostatistical samples optimized with the true (but in practice unknown) variogram.

7.3.2 Centred Grid Sampling

Figure 7.24 shows the block-kriging variance for centred square grids in a square area. The spacing of all grids equals the side of the square divided by the square root of the sample size, $d = \sqrt{A/n}$, where A is the area of the square. For example, for $n = 25$, the grid spacing equals one fifth of the side of the square. For linear variograms, this spacing is optimal, while for spherical variograms, the optimal grid spacing is slightly larger (Webster and Oliver, 1990). In each graph there are four lines, representing different relative ranges (range divided by the side of the square). The block-kriging variance is expressed as a percentage of the sill of the spherical variogram, i.e., the maximum value. These graphs can be used to choose the required sample size, given the constraint on the precision of the predicted global mean, or vice versa, to predict the precision given the sample size. Strictly speaking, regular grids with optimized spacing can be considered as geostatistical samples.

The target area will usually not be square. To explore the effect of its shape on the precision of the predicted global mean, we have calculated the standard error of the predicted mean of a rectangle with a length/width ratio of 100, and divided it by the standard error for a square of the same area. An experiment showed that for non-square rectangles the standard errors differed only marginally, except for small sample sizes (< 20), combined with a large relative range (> 0.5), and a small nugget-to-sill ratio, (< 0.5). Therefore Fig. 7.24 can also be used to obtain an estimate of the required sample size or precision for non-square rectangles. To this end, the range of the postulated variogram must be divided by $\sqrt{A}$ to select the appropriate line in Fig. 7.24.

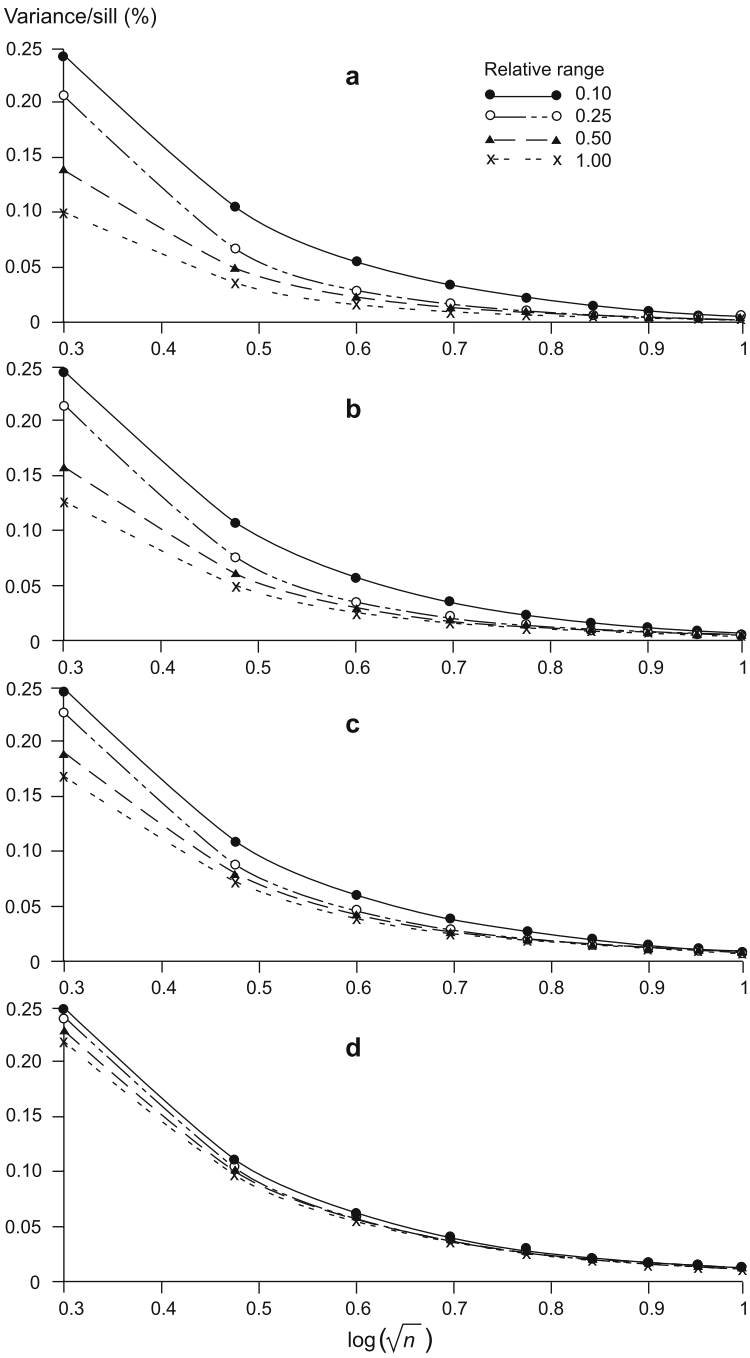


Fig. 7.24. Block-kriging variance of the global mean of a square block, predicted from centred square grids with grid spacing $\Delta s = \sqrt{A/n}$, as a function of the sample size, for spherical variograms. The variance is expressed as a percentage of the sill of the variogram. Nugget-to-sill ratios: **a**: 0.10; **b**: 0.25; **c**: 0.50; **d**: 0.80

7.3.3 Geostatistical Sampling

If a variogram can be postulated, it can be used to optimize sampling pattern, i.e., the sampling pattern with minimum block-kriging variance of the predicted global mean. Section 8.3.4 describes ways to optimize sampling patterns for predicting the values at points. In most cases, however, the sampling pattern with a minimum value for the maximum or mean kriging variance of local predictions will be sub-optimal for predicting the global mean, and vice versa (Sacks and Schiller, 1988; Nychka et al., 1997). To illustrate this, Fig. 7.25 shows the optimized pattern of 23 points in a square, as calculated by simulated annealing (Appendix A), for ordinary block-kriging of the global mean with a spherical variogram with a range equal to half the side of the square, and a nugget-to-sill ratio of 0.10. There is a striking difference with the optimized patterns obtained with the mean and maximum point-kriging variance as quality measure (Fig. 8.4). The latter are distorted triangular grids, alternating between three rows of 5 points and two rows of 4 points. The two rows of 4 points are sandwiched between the rows of 5 points. In Fig. 7.25 the two rows of 4 points are outside the three rows of 5 points, near the top and bottom of the square. One can imagine that this pattern will be sub-optimal for the maximum and mean point-kriging variance, because the point-kriging variance rises sharply near the edges. However, locating points near the edges to counteract this boundary effect seems to be sub-optimal for predicting the global mean.

Finally, Fig. 7.26 shows the optimized pattern of 32 points added to 6 prior points to predict the global mean for the province of Gelderland in the Netherlands, again calculated by simulated annealing. If one compares the pattern with those obtained with the mean and maximum kriging variance as quality measure (Figs. 8.6 and 8.7), the most striking is once again the difference with the maximum kriging variance pattern. Compared to this pattern, points are moved to the interior of the province. There are also fewer points close to the boundaries compared with the mean kriging variance pattern.

Here and in Sect. 7.3.2, the spatial mean is predicted with the block-kriging predictor, which is a linear combination of the values at the observed points. For multivariate normal distributions, this linear predictor is also the best predictor, i.e., has smallest prediction-error variance. For other distributions, however, non-linear predictors may perform better than the linear predictor. For instance, for multivariate lognormal distributions, the precision of predictions can be increased by first taking the logarithm of the values at the sampling locations, and then block-kriging the log-transformed values. In this case, the block-kriging variance is not merely a function of the locations of the observations, but also depends on the values observed. This implies that if one has prior points, then the optimal pattern of the added points depends on the values observed at the prior points (see Chao and Thompson (2001) for an illustration).

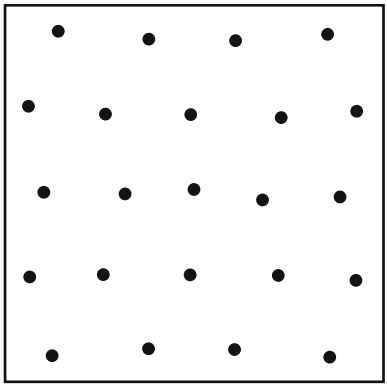


Fig. 7.25. Optimized pattern of 23 points in a square for predicting the global mean by block-kriging with a spherical variogram with a range half the side of the square and a nugget-to-sill ratio of 0.10

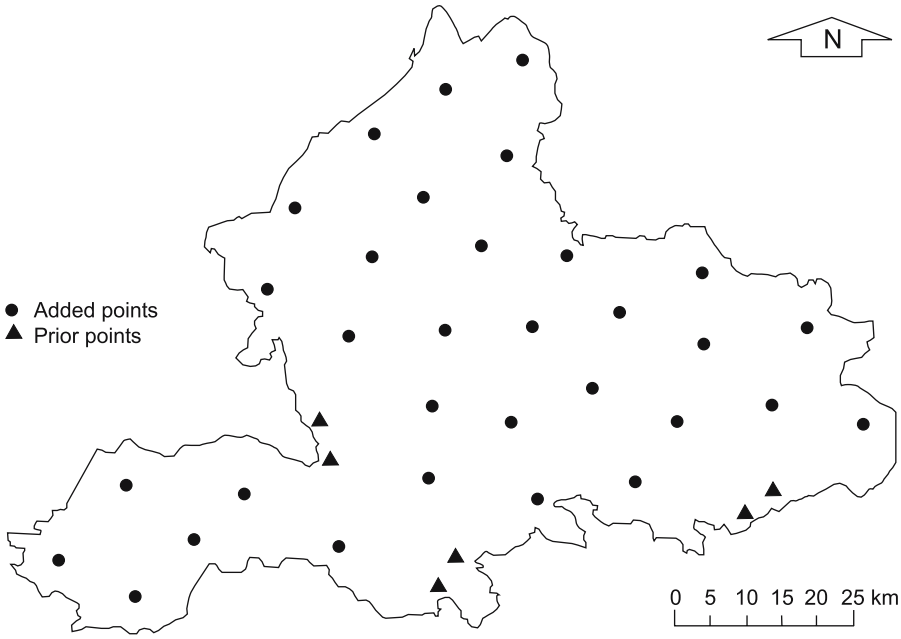


Fig. 7.26. Optimized pattern of 32 points added to 6 prior points for predicting the global mean for the province of Gelderland, the Netherlands, by block-kriging with a spherical variogram with a range of 50 km and a nugget-to-sill ratio of 0.40

7.3.4 Detecting Hot Spots

A hot spot is defined here as a relatively small area with a concentration of some compound or abundance of some object (e.g., a plant or animal species)

that exceeds a certain threshold. Examples are pieces of land with soil or groundwater contaminated by some point-source, anomalous sites with high concentrations of natural geochemical elements, niches inhabited by some rare plant or animal species, or patches of crops infected by some disease. The threshold concentration may be a constant value that does not vary in space, or a spatially varying concentration. An example of the latter is a soil standard defined in terms of soil organic-matter content, clay percentage and pH, thus taking into account the bio-availability of the contaminant. A hot spot may also have sharp boundaries separating them from background values, or more gradual transitions.

In *detecting* hot spots, the aim of the survey is to find out whether at any point in the study area the critical threshold is exceeded, without asking where (Sect. 2.2.6). The answer to this type of question can be coded as a 0/1 indicator variable, which is related to the area as a whole, and consequently is a global quantity. For *delineating* hot spots we refer to Sect. 8.3.5.

Detection of hot spots can be achieved better with purposive sampling than with probability sampling. If one has no prior information on the location of the hot spots, samples are typically taken on a centred, regular grid. Gilbert (1987) worked out a method for calculating the required grid spacing from the consumer's risk, β, i.e., the probability of not hitting a hot spot *if it exists*, and the geometry (size and shape) of the hot spot. This method is implemented in the Visual Sample Plan program (Gilbert et al., 2002). The probability of hitting a hot spot if it exists is calculated by summing the zones of coverage for the sampling locations, excluding overlaps. The zone of coverage for any sampling location can be obtained by drawing the contour of a hot spot with its centre at the sampling location. If the centre of the hot spot is in the zone of coverage, it will be detected from the sampling location. Note that this method can be used only if one has prior information on the size and shape of the hot spot. Moreover, it does not deal with the situation in which there is more than one hot spot, and the intention is to detect all of them.

So far, it has been assumed that a hot spot exists. In other words, it is assumed that the probability that a hot spot exists is 1. If the existence of a hot spot is uncertain, the probability that a hot spot exists and is detected can be estimated by

$$P(A, B) = P(B|A)\, P(A)\ , \tag{7.92}$$

where $P(B|A)$ is the probability that the hot spot is hit, conditional on its existence, and $P(A)$ is the probability that the hot spot exists. Given the grid spacing and geometry of the hot spot one can calculate $P(B|A)$ and simply multiply this by the a priori probability that the hot spot exists to obtain the probability that the hot spots exists and is detected by the sample.

In some situations, information on land use, or a walkover survey of visual or organoleptic indicators of high concentrations can be used to subdivide the target area into subregions with different probabilities of containing the hot spot. The grid spacing can then be adapted to these a priori probabilities as

follows. For all subregions, there must be an equal probability that a hot spot exists when none is detected by the sample. This probability is referred to as the a posteriori probability and denoted by $P(A|\overline{B})$. Bayes' formula can now be used to calculate from the a priori and a posteriori probabilities for each subregion the probability of not hitting the hot spot if it exists, the consumer's risk β:

$$\beta = \frac{1 - P(A)}{P(A)\left\{\frac{1}{P(A|\overline{B})} - 1\right\}} \,. \tag{7.93}$$

Hence, when $P(A)$ differs between subregions, and given a constant $P(A|\overline{B})$ for all subregions, for instance 0.05, β differs between subregions, and this leads to different grid spacings. Subregions with large a priori probabilities will be given smaller grid spacings than subregions with small a priori probabilities. An alternative to Centred Grid Sampling is to optimize the sampling pattern by minimizing the sum of the a priori probabilities outside the zones of coverage of the sampling locations (Tucker et al., 1996).

A different aim, related to detection of hot spots, is estimation of the fraction of the area with values exceeding the threshold or the total number (density) of objects. For this aim the design-based Adaptive Cluster Sampling strategy may be a good choice (Sect. 7.2.10). An unbiased estimate of the sampling variance of the estimated areal fraction (number of objects) can then be obtained from the sample.

8

Local Quantities in Space

8.1 Introduction to Methods for Local Quantities in Space

The region for which estimation or testing of hypotheses is required, need not be the entire area sampled. Interest may also be in one or more sub-areas (domains), and this is treated in this chapter. Examples of local quantities are the mean phosphate content in the topsoils of all agricultural fields in an area, the abundances of some plant species in the cells of a raster covering a nature conservation area, and the nitrate concentration at the nodes of a fine grid covering an agricultural area, at the start of the growing season.

Means and parameters of the Spatial Cumulative Distribution Function of sub-areas can be estimated by a design-based method (Sect. 8.2) or a model-based method (Sect. 8.3). In choosing between these two approaches the sample sizes per sub-area should play an important role. If one can afford a reasonable sample size for each sub-area (domain), and valid interval-estimates are important, then a design-based approach is worth considering. If one has many sub-areas and the average number of sampling locations per sub-area is small, say less than 5 to 10, then the model-based approach comes into scope. A design-based alternative is the regression estimator, which also uses sampling locations outside the domain (Sect. 8.2.2).

At the extreme one may want to estimate or test hypotheses about the values at points. In this case we recommend a model-based method (Sect. 8.3).

8.2 Design-Based Methods for Local Quantities in Space

8.2.1 Introduction

This section mainly deals with design-based estimation of means, or other parameters of the SCDF, for several sub-areas. If the sub-areas can be delineated beforehand, one may select a sample from each of them separately. In that

case the sub-areas act as 'strata' in a stratified sampling design (Sect. 8.2.2, subsection 'Sampling per Sub-Area'). However, even when the sub-areas can be delineated prior to the sampling, this is not always done in practice. When there are numerous sub-areas, independent sampling of all the sub-areas may be too costly. Sometimes it is impossible to delineate the sub-areas, simply because one does not know where they are. An example of the latter is where the spatial mean of a target variable within a given soil type is to be estimated, and no map of the soil types at an appropriate scale is available. This mean can be estimated if, in addition to the target variable, the soil type is recorded at the sampling locations. When it is impractical or impossible to single out the sub-areas as strata, one may select a sample from the total area independent from any division into sub-areas. The sample data are then sorted afterwards according to the sub-areas in which the sampling locations happen to fall. The sample sizes of the sub-areas are random rather than controlled at the selection stage. Estimators for the mean and its sampling variance are given in Sect. 8.2.2. In a separate subsection 'Sampling Across Small Sub-Areas' estimators are presented for means of small sub-areas, and consequently few points per sub-area.

The final section of this chapter deals with design-based estimation of values at points. Although in principle, one would rather prefer a model-based method for this aim, there is a design-based alternative if an estimate of the average accuracy for all points of a sub-area suffices (Sect. 8.2.3).

8.2.2 Quantities Defined on Sub-Areas

Sampling per Sub-Area

When the sub-areas are known at the beginning of the survey and one can afford a reasonable sample size for each domain, we recommend to sample each sub-area separately, as an independent unit. Thus one can control the sample sizes in the sub-areas, and one may choose a suitable type of sampling design for each sub-area. This type of design need not be similar for all sub-areas, For instance, for a sub-area that consists of several disjoint polygons (e.g., all polygons of a map unit) a two-stage design can be efficient, whereas for compact sub-areas this can be rather inefficient. Once a sampling design is chosen, the estimators for the mean (areal fraction) and for the sampling variance can be found in Sect. 7.2. If there is one budget for sampling all sub-areas, one has to decide on the budgets per sub-area. Optimization under the constraint that the total costs of the survey do not exceed the overall budget, means minimizing an objective function that is defined in terms of the sampling variances per sub-area, for instance:

$$J(n_1, \ldots, n_D) = \sum_{d=1}^{D} w_d \widehat{V}\left(\hat{\bar{z}}_d\right) \ . \tag{8.1}$$

The weight w_d for a sub-area d reflects the importance of obtaining a precise estimate of that sub-area. For instance, if it is more important to obtain precise estimates for large sub-areas than for small sub-areas, one can take the relative (surface) area of the sub-areas as weights. Alternatively, one may want estimates that are approximately equally precise for all sub-areas. This can be reached by setting all weights to 1 and adding a penalty for differences between the estimated sampling variances per domain, for instance the variance of the estimated sampling variances per domain:

$$J(n_1,\ldots,n_D) = \alpha\cdot\frac{1}{D}\sum_{d=1}^{D}\widehat{V}\left(\hat{\bar{z}}_d\right)+(1-\alpha)\sqrt{\frac{1}{D-1}\sum_{d=1}^{D}\left\{\widehat{V}\left(\hat{\bar{z}}_d\right)-\frac{\sum\widehat{V}\left(\hat{\bar{z}}_d\right)}{D}\right\}^2}, \tag{8.2}$$

where α is a weighting factor to be specified by the user.

Sampling Across Large Sub-Areas

Frequently, the sub-areas for which an estimate is required are not used in selecting the sampling locations. Sometimes a sample is taken independently from any subdivision into sub-areas, in other situations the sub-areas for which estimates are required differ from those used in selecting the sampling locations, i.e., they cut across the strata. In both cases the sample sizes in the domains are uncontrolled and vary between samples.

Although perfect control is impossible, strong under- and overrepresentation of domains can be avoided by selecting locations by a design type that leads to good spatial coverage, such as Systematic Random Sampling. Design types that lead to spatial clustering such as Two-Stage Random Sampling, Cluster Random Sampling and, to a lesser degree, Simple Random Sampling are less appropriate. Despite this, we give estimators for these designs, because one may want to use the sample data of one of these designs for estimating the means of domains. Finally, if one can subdivide the target area into 'homogeneous' sub-areas, one may use the sub-areas as strata in stratified random sampling.

When there are only a few large domains, the sample sizes in the domains may be large enough, say 10 to 20, to estimate the mean of a given domain accurately enough from the sampling locations that happen to fall in that domain only. In that case the mean of a domain could be estimated by the direct estimator:

$$\hat{\bar{z}}_d = \frac{1}{N_d}\sum_{i=1}^{n_d}\frac{z_{di}}{\pi_{di}}\,, \tag{8.3}$$

where N_d is the size of the domain, n_d is the number of sampling locations in domain d, z_{di} is the value at the point selected in the ith draw from domain d, and π_{di} is the inclusion probability of this point. When the size of the domain is unknown, one can estimate it by:

$$\widehat{N}_d = \sum_{i=1}^{n_d} \frac{1}{\pi_i} \,. \tag{8.4}$$

This results into the general ratio-estimator:

$$\hat{\bar{z}}_{\mathrm{R}d} = \frac{\sum\limits_{i=1}^{n_d} \frac{z_{di}}{\pi_{di}}}{\sum\limits_{i=1}^{n_d} \frac{1}{\pi_{di}}} \,. \tag{8.5}$$

This ratio-estimator is more precise than the direct estimator and is therefore recommended even in situations where the size of the domain is known. The variance of the ratio-estimator can be approximated only, but for expected sample sizes in the domains greater than 10 to 20, this approximation suffices. We shall now work out the general ratio-estimator and its variance for the basic types of design.

Simple Random Sampling

Inserting $\pi_{di} = n/N$ in (8.5) gives:

$$\hat{\bar{z}}_{\mathrm{R}d} = \frac{1}{n_d} \sum_{i=1}^{n_d} z_{di} \,. \tag{8.6}$$

So in this case the mean of the domain is simply estimated by the mean of the z-values observed in the domain. The variance of this estimator can be estimated by:

$$\widehat{V}\left(\hat{\bar{z}}_{\mathrm{R}d}\right) = \frac{1}{\hat{a}_d^2} \cdot \frac{1}{n\,(n-1)} \sum_{i=1}^{n_d} (z_{di} - \bar{z}_{\mathrm{s}_d})^2 \,, \tag{8.7}$$

where $\bar{z}_{\mathrm{s}_d}$ is the sample mean in domain d, and $\hat{a}_d$ is the estimated relative size of domain d:

$$\hat{a}_d = \frac{n_d}{n} \,. \tag{8.8}$$

Stratified Simple Random Sampling

Consider the following situation. The aim of the survey is to estimate the mean of each field in a region. To estimate these means the units of a soil map are used as strata in Stratified Simple Random Sampling. The soil map units cross the fields. In each field the sample size is large enough to estimate its mean with sufficient accuracy. The mean of a domain (field) can be estimated by:

$$\hat{\bar{z}}_{\mathrm{R}d} = \frac{\sum\limits_{h=1}^{H} \frac{a_h}{n_h} \sum\limits_{i=1}^{n_{hd}} z_{hdi}}{\sum\limits_{h=1}^{H} a_h \frac{n_{hd}}{n_h}} \,, \tag{8.9}$$

where n_{hd} is the number of sampling locations falling in the intersection of domain d and stratum h. The sampling variance is estimated by:

$$\widehat{V}\left(\hat{\bar{z}}_{\mathrm{R}d}\right) = \frac{1}{\hat{a}_d^2} \sum_{h=1}^{H} \frac{a_h^2}{n_h\left(n_h-1\right)} \left\{ \sum_i (z_{hdi} - \bar{z}_{\mathrm{s}_{hd}})^2 + n_h \left(1 - \frac{n_{hd}}{n_h}\right) \left(\bar{z}_{\mathrm{s}_{hd}} - \hat{\bar{z}}_{\mathrm{R}d}\right)^2 \right\} , \tag{8.10}$$

where $\bar{z}_{\mathrm{s}_{hd}}$ is the sample mean in the intersection of domain d and stratum h, and $\hat{a}_d$ is the estimated relative size of domain d:

$$\hat{a}_d = \sum_{h=1}^{H} a_h \frac{n_{hd}}{n_h} . \tag{8.11}$$

Two-Stage Random Sampling

We first define a new variable z', which equals z everywhere in the domain, but is zero elsewhere. The mean for a domain can now be estimated by:

$$\hat{\bar{z}}_{\mathrm{R}d} = \frac{\sum_{i=1}^{n} \widehat{\bar{z'}}_i}{\sum_{i=1}^{n} \frac{m_{di}}{m_i}} , \tag{8.12}$$

where n is the number of primary unit (PU) selections, m_i is the number of points in PU selection i, m_{di} is the number of points in PU selection i and domain d, and $\widehat{\bar{z'}}_i$ is the estimated mean of the transformed variable z' for PU selection i. When PUs are selected by Simple Random Sampling, $\widehat{\bar{z'}}_i$ equals the sample mean of z' for PU selection i. The variance can be estimated by:

$$\widehat{V}\left(\hat{\bar{z}}_{\mathrm{R}d}\right) = \frac{1}{\hat{a}_d^2} \cdot \frac{1}{n\left(n-1\right)} \sum_{i=1}^{n} \left(\widehat{\bar{z'}}_i - \hat{\bar{z}}_{\mathrm{R}d} \frac{m_{di}}{m_i} \right)^2 , \tag{8.13}$$

where the relative size of the domain is estimated by:

$$\hat{a}_d = \frac{1}{n} \sum_{i=1}^{n} \frac{m_{di}}{m_i} . \tag{8.14}$$

Cluster Random Sampling

The formulas for the ratio-estimator and its variance are the same as for Two-Stage Random Sampling when primary units are replaced by clusters.

Systematic Random Sampling

Similar to Simple Random Sampling, the mean of a domain d can be estimated by the mean of the z-values observed in the domain:

$$\hat{\bar{z}}_{\mathrm{R}d} = \frac{1}{m_d} \sum_{i=1}^{m_d} z_{di} \ , \tag{8.15}$$

where m_d is the number of grid nodes falling in domain d. If one estimates the variance as if the sample was obtained by Simple Random Sampling, then a conservative estimate of its variance is obtained, provided that the sampling pattern does not interfere with a cyclic variation.

Sampling Across Small Sub-Areas

When the domains are numerous and are not well represented in the sample, the direct estimators from the previous section may lead to sampling variances that are too large. Extensive literature exists about how the means of such small domains can be estimated; see Chauduri (1994) for a review. We shall describe two estimators, the synthetic estimator and the generalized regression estimator. Both estimators use, as opposed to the estimators in the previous section, sampling locations outside the domain. The intention is that, by doing so, the precision will increase. On the other hand, in general some bias will be introduced.

For small sub-areas, it is even more important to spread the sampling locations over the target area, otherwise some sub-areas may be missed. Systematic Random Sampling therefore may be a good choice, despite its difficulties with variance estimation. For the synthetic estimator stratified random sampling is sensible, but again, one should take care of good spatial coverage.

Synthetic Estimator

With the synthetic estimator the area is divided into sub-areas, here referred to as groups, that are as homogeneous as possible. In general these groups cut across the domains, i.e., they do not coincide with the domains. The mean of a domain is estimated as a weighted mean of the estimated means per group:

$$\hat{\bar{z}}_{\mathrm{SY}d} = \sum_{g=1}^{G} a_{gd}\, \hat{\bar{z}}_{\mathrm{R}g} \ , \tag{8.16}$$

where a_{gd} is the size of group g in domain d, relative to the size of the domain ($a_{gd} = N_{gd}/N_d$). The group means are estimated by the ratio-estimator, $\hat{\bar{z}}_{\mathrm{R}g}$:

$$\hat{\bar{z}}_{\mathrm{R}g} = \frac{\sum_{i=1}^{n_g} \frac{z_i}{\pi_i}}{\sum_{i=1}^{n_g} \frac{1}{\pi_i}} = \frac{\sum_{i=1}^{n_g} \frac{z_i}{\pi_i}}{\widehat{N}_g} \ . \tag{8.17}$$

In general this synthetic estimator is not p-unbiased. It is approximately unbiased only in the hypothetical situation where the true means of all domains within a group equal the true mean of that group, i.e., $\bar{z}_{gd} = \bar{z}_g$ for all d. However, the intention is that the reduction of the sampling variance due to the use of the sample data from similar parts of the area outside the domain, outweighs the extra error due to the introduction of design-bias.

The sampling variance of the synthetic estimator can be approximated by:

$$\widehat{V}\left(\hat{\bar{z}}_{\mathrm{SY}d}\right) = \sum_{g=1}^{G} a_{gd}^2 \widehat{V}\left(\hat{\bar{z}}_{\mathrm{R}g}\right) \ , \tag{8.18}$$

where $\widehat{V}\left(\hat{\bar{z}}_{\mathrm{R}g}\right)$ is the estimated sampling variance of the ratio-estimator for the group mean (see the previous section on estimation for large domains).

When the groups are used as strata the synthetic estimator becomes:

$$\hat{\bar{z}}_{\mathrm{SY}d} = \sum_{h=1}^{H} a_{hd}\, \hat{\bar{z}}_h \ , \tag{8.19}$$

where $\hat{\bar{z}}_h$ is the π-estimator for the mean of stratum h. The sampling variance can then be estimated by:

$$\widehat{V}\left(\hat{\bar{z}}_{\mathrm{SY}d}\right) = \sum_{h=1}^{H} a_{hd}^2 \widehat{V}\left(\hat{\bar{z}}_h\right) \ , \tag{8.20}$$

where $\widehat{V}\left(\hat{\bar{z}}_h\right)$ is the estimated variance of the π-estimator for the mean of stratum h.

Generalized Regression Estimator

To eliminate the design-bias, Hidiroglou and Särndal (1985) proposed the generalized regression estimator for small domain estimation:

$$\hat{\bar{z}}_{\mathrm{GR}d} = \frac{1}{N_d} \sum_{k=1}^{N_d} \hat{z}_k + \frac{1}{\widehat{N}_d} \sum_{i=1}^{n_d} \frac{\epsilon_i}{\pi_i}, \tag{8.21}$$

where $\hat{z}_k = \sum_{q=1}^{Q} b_q x_{qk}$ is the predicted value of the target variable z at point k from measurements of Q ancillary variables x_q at point k, $\epsilon_i = z_i - \hat{z}_i$ is the residual at the point selected in the ith draw, π_i is the inclusion probability of this point, and $\widehat{N}_d = \sum_{i=1}^{n_d} 1/\pi_i$ is the estimated size (area) of the domain. Note that the first sum is over all points in the domain d, so the ancillary variables must be known for all points in the domain. The first part of this estimator, the mean of the predicted values $\frac{1}{N_d} \sum_{k=1}^{N_d} \hat{z}_k$, is the synthetic component of the estimator, because points outside the domain are used in estimating the regression coefficients b_q. In Sect. 7.2.11 estimators for

these regression coefficients are presented. To eliminate the bias, the mean of the predicted values is adjusted by the estimated mean of the residuals in the domain d, the second term of the estimator. Note that the estimated size of the domain is used in estimating this mean, i.e., the domain mean of the residuals is estimated by the ratio-estimator.

We shall now work out the generalized regression estimator for the one-way ANOVA model and the Simple Linear Regression model. The one-way ANOVA model assumes that the mean and the variance of z is constant within the groups, i.e., for all points in group g:

$$\begin{aligned} E_\xi(Z_k) &= \beta_g \\ V_\xi(Z_k) &= \sigma_g^2 \,. \end{aligned} \tag{8.22}$$

For this model the generalized regression estimator turns out to be:

$$\hat{\bar{z}}_{\mathrm{AN}d} = \sum_{g=1}^{G} a_{gd}\, \hat{\bar{z}}_{\mathrm{R}g} + \frac{1}{\widehat{N}_d} \sum_{g=1}^{G} \widehat{N}_{gd} \left(\hat{\bar{z}}_{\mathrm{R}gd} - \hat{\bar{z}}_{\mathrm{R}g} \right) \,, \tag{8.23}$$

with:

$$\hat{\bar{z}}_{\mathrm{R}gd} = \frac{\sum_{i=1}^{n_{gd}} \frac{z_i}{\pi_i}}{\widehat{N}_{gd}} = \frac{\sum_{i=1}^{n_{gd}} \frac{z_i}{\pi_i}}{\sum_{i=1}^{n_{gd}} \frac{1}{\pi_i}} \,. \tag{8.24}$$

When the groups are used as strata and these are sampled by Simple Random Sampling (Stratified Simple Random Sampling), the estimator becomes:

$$\hat{\bar{z}}_{\mathrm{AN}d} = \sum_{h=1}^{H} a_{hd}\, \bar{z}_{\mathrm{s}_h} + \frac{1}{\widehat{N}_d} \sum_{h=1}^{H} \widehat{N}_{hd} \left(\bar{z}_{\mathrm{s}_{hd}} - \bar{z}_{\mathrm{s}_h} \right) \,, \tag{8.25}$$

where a_{hd} is the size of stratum h in domain d, relative to the size of the domain ($a_{hd} = N_{hd}/N_d$), $\bar{z}_{\mathrm{s}_h}$ is the sample mean in stratum h, and $\bar{z}_{\mathrm{s}_{hd}}$ is the sample mean in the intersection of stratum h and domain d.

This estimator is similar to the synthetic estimator (8.19), plus a correction term. The variance of this estimator can be estimated with:

$$\widehat{V}\left(\hat{\bar{z}}_{\mathrm{AN}d}\right) = \sum_{h=1}^{H} a_{hd}^2 \widehat{V}\left(\hat{\bar{z}}_{\mathrm{R}hd}\right) \,, \tag{8.26}$$

where $\widehat{V}(\hat{\bar{z}}_{\mathrm{R}hd})$ is the estimated variance of the ratio-estimator for the mean of domain d in stratum h for Simple Random Sampling (8.7).

When the Simple Linear Regression model is adopted, one assumes that

$$\begin{aligned} E_\xi(Z_k) &= \beta_0 + \beta_1 x_k \\ V_\xi(Z_k) &= \sigma^2. \end{aligned} \tag{8.27}$$

For this model the regression estimator equals:

$$\hat{\bar{z}}_{\mathrm{SR}d} = \frac{1}{N_d}\sum_{k=1}^{N_d}(b_0 + b_1 x_k) + \frac{1}{\widehat{N}_d}\sum_{i=1}^{n_d}\frac{\epsilon_i}{\pi_i}\ , \tag{8.28}$$

where b_0 and b_1 are estimates of the regression coefficients. For Simple Random Sampling the variance of this simple regression estimator can be approximated by:

$$\widehat{V}\ \left(\hat{\bar{z}}_{\mathrm{SR}d}\right) = \frac{1}{a_d^2}\cdot\frac{1}{n\,(n-2)}\sum_{i=1}^{n}\lambda_i^2\,\epsilon_i^2\ , \tag{8.29}$$

where a_d is the relative size of domain d ($a_d = N_d/N$), $\bar{x}_{\mathrm{s}_d}$ is the sample mean of x in domain d, $\bar{x}_d$ is the true mean of x in domain d, $\bar{x}_{\mathrm{s}}$ is the overall sample mean of x, and λ_i are the regression weights, to be calculated by:

$$\lambda_i = \frac{N_d}{\widehat{N}_d}I_i + \frac{a_d\,(\bar{x}_d - \bar{x}_{\mathrm{s}_d})\,(x_i - \bar{x}_s)}{\overline{x_s^2} - (\bar{x}_s)^2}\ , \tag{8.30}$$

with:

$$I_i = \begin{cases} 1 & \text{if } \mathbf{s}_i \in \mathcal{D} \\ 0 & \text{otherwise} \end{cases} \tag{8.31}$$

where $\mathbf{s}_i$ is the ith sampling location and $\mathcal{D}$ is the domain in question.

A general estimator for the variance of the regression estimator for small domains that can be used for any model in combination with any sampling design with known inclusion probabilities of points and of pairs of points is Särndal et al. (1992, p. 401):

$$\widehat{V}\big(\hat{\bar{z}}_{\mathrm{GR}d}\big) = \frac{1}{N_d^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\pi_{ij} - \pi_i\,\pi_j}{\pi_{ij}}\cdot\frac{\lambda_i\,\epsilon_i}{\pi_i}\cdot\frac{\lambda_j\,\epsilon_j}{\pi_j}\ , \tag{8.32}$$

where π_{ij} is the probability that both points i and j are included in the sample, and λ_i is the regression weight attached to point i:

$$\lambda_i = \frac{N_d}{\widehat{N}_d}I_i + \left(\sum_{k=1}^{N_d}\mathbf{x}_k - \frac{N_d}{\widehat{N}_d}\sum_{i=1}^{n_d}\frac{\mathbf{x}_i}{\pi_i}\right)'\cdot\left(\sum_{i=1}^{n}\frac{\mathbf{x}_i\,\mathbf{x}_i'}{\sigma_i^2\,\pi_i}\right)^{-1}\cdot\frac{\mathbf{x}_i}{\sigma_i^2}\ , \tag{8.33}$$

where $\mathbf{x}_i$ is the vector of predictor variables at point i, with value one as the first element.

8.2.3 Values at Points

The focus of design-based methods is on estimating the means or other parameters of the cumulative distribution function for the entire area or for subareas. However, one may use the estimated mean of a 'homogeneous' sub-area (group) as an estimate of the value at all points in that group. For instance, to

estimate the value of a soil variable at all nodes of a fine grid, one can estimate the spatial means of the units of a soil map, and assign the estimated mean of a given soil map unit to all grid nodes in that unit. To define the estimation error, we imagine that the sampling is repeated, each sample resulting into an estimate of the spatial mean of the group, which on its turn is used as an estimate of the values at the points in that group. So the variance of the estimator for the value at a point is defined as a sampling variance, which makes the method design-based. We can estimate this sampling variance of the estimator for the value at a point (it equals the sampling variance of the estimator for the mean of the group), however we do not know its bias because the true value at the point is unknown. Therefore the average accuracy is estimated, or more precisely the spatial mean of the expected squared error of the estimates at all points in the group:

$$MSE\,(\hat{z}_g) = \frac{1}{N_g}\sum_{k=1}^{N_g}\left\{E_p(\hat{z}_k - z_k)^2\right\}\ . \tag{8.34}$$

This mean squared error of the estimates at points can be rewritten as the sum of the mean squared error of the estimated mean of the group and the spatial variance of the target variable within the group:

$$MSE\,(\hat{z}_g) = E_p\big(\hat{\bar{z}}_g - \bar{z}_g\big)^2 + \frac{1}{N_g}\sum_{k=1}^{N_g}(z_k - \bar{z}_g)^2\ . \tag{8.35}$$

With p-unbiased sampling strategies the first term on the right hand side of (8.35) equals the sampling variance of the estimator for the mean of the group. The mean squared error of the estimates at points can be estimated by estimating the two terms of (8.35). The sampling variance of the estimated mean depends on the sampling design, and estimators for this sampling variance can be found in previous sections. The spatial variance is a property of the group, and is independent of the sampling design. When the groups are used as strata and these are sampled by Simple Random Sampling (Stratified Simple Random Sampling), the spatial variance of the z-values in a group can simply be estimated by:

$$\widehat{S}^2(z_h) = \frac{1}{n_h - 1}\sum_{i=1}^{n_h}(z_{hi} - \bar{z}_{\mathrm{s}_h})^2\,. \tag{8.36}$$

Finally, we would like to stress that the difference between design-based and model-based estimation of values at points is in some situations not as large as one might think. Suppose that an area is sampled regularly and sparsely, such that the distance between neighbouring points is always large compared to the range of the variogram. When the ordinary kriging model is adopted, the surface of the model-based predictions is a horizontal plane with spikes at the sampling locations. The values are predicted by the unweighted

sample mean everywhere except close to the sampling locations. One does obtain a separate estimate of the variance at each prediction point, however these estimates are equal for all points except near the sampling locations.

8.3 Model-Based Methods for Local Quantities in Space

8.3.1 Introduction

In general, the most appropriate sampling strategies for mapping target variables and predicting the means of many, relatively small sub-areas are model-based. To profit from the spatial variation model in predicting the values at particular points or the block means, the following conditions should be met: (1) the target variable must be spatially autocorrelated at the scale of the study region; (2) the sampling density must be large enough so that the distances between the sampling locations are small compared to the range of the variogram; (3) there must be enough data to estimate the model from.

If, prior to the sampling, one can postulate a tentative model, then this model can be used either to optimize the spacing of a grid (Sect. 8.3.2), or to optimize an irregular sampling pattern (Sect. 8.3.4). So, in that case a model is used at the sampling stage and the inference (prediction) stage. However, in many situations it is not feasible to postulate a model prior to the sampling. In these cases, one can sample on a centred grid, with a spacing derived from the affordable sample size (Sect. 8.3.2). Sampling on grids may be sub-optimal, for instance in the case of small, irregularly shaped areas or when measurements are already available for locations that cannot be matched with the grid. An alternative is then a 'spatial coverage sample', i.e., a sample that covers the space as uniformly as possible (Sect. 8.3.3).

The final section (Sect. 8.3.5) deals with sampling to delineate areas where the target variable exceeds a critical threshold, such as 'hot spots'. In this case, the mode of inference is not prediction but classification.

8.3.2 Centred Grid Sampling

Sampling on a regular grid is attractive because of its simplicity. Contrary to design-based Systematic Random Sampling (Sect. 7.2.7) the starting point is not selected at random, but the grid is purposively placed such that it covers the area as good as possible, because this maximizes the precision of model-based predictions. Grid Sampling is generally a good choice when the area is large (relative to the grid spacing) and has a fairly regular shape, and there are no prior measurements that can be used to map the target variable. However, in situations were there may be a cyclic pattern of variation, one needs to ensure that the grid does not interfere with this cyclic pattern, as this may lead to grossly misleading results. With model-based Grid Sampling, it is necessary to choose the pattern and orientation of the grid, as well as the grid spacing.

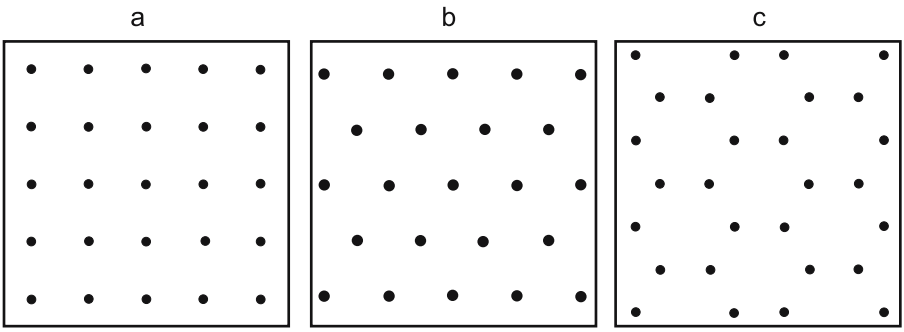


Fig. 8.1. Commonly used grid patterns. **a**: square grid; **b**: triangular grid; **c**: hexagonal grid

Pattern and Orientation

Figure 8.1 shows three commonly used patterns. Which sample pattern is optimal depends on the predictor, the variogram, and the quality measure. Assume that the values at specific points are predicted by ordinary kriging and that we want to minimize the maximum of the kriging variance. This maximum occurs at the centres of gravity of the squares, triangles or hexagons. Which pattern is optimal depends on the variogram. For variograms with a small nugget-to-sill ratio, the maximum kriging variance is minimal for the triangular grid at the same number of measurements (Yfantis et al., 1987). When the nugget is large, say > 0.9 times the sill of the variogram, and the distance between neighbouring points approaches the range of the variogram, the hexagonal pattern is optimal. When anisotropic spatial variation is expected, a rectangular pattern may be a good choice. The short sides of the rectangles must then be aligned in the direction of the shortest range. In practice, a square pattern is often chosen for convenience.

Ideally, the grid should be oriented so as to minimize the boundary effects. For rectangular target areas, this can be achieved by orienting one side of the squares, triangles or hexagons, parallel to one side of the target area. For irregularly shaped areas, it will be hard to decide what orientation of the regular grid is optimal with respect to the boundary effect. However, if the boundary effect is expected to be an important issue, Spatial Coverage Sampling or Geostatistical Sampling is more appropriate than Grid Sampling.

Grid Spacing

If there is no specific requirement for the precision of the predictions but a cost constraint only, the grid spacing can be calculated by equating the area of a unit grid cell to the total area of the study region divided by the affordable sample size. For instance, the grid spacing Δs for a square grid can

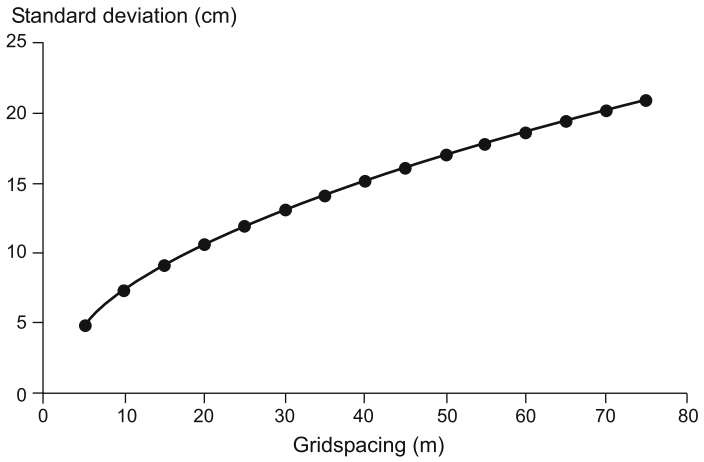


Fig. 8.2. Standard deviation of the predicted mean altitude of small blocks (10 m x 10 m), centred at the midpoints of square grid cells as a function of the grid spacing

be calculated by $\Delta s = \sqrt{A/n}$, where A is the area of the study region, and n is the affordable sample size. This procedure can also be used if no model of spatial variation is available.

If a minimum precision is required and a reasonable model of spatial variation can be postulated, this model can be used to calculate the grid spacing required to achieve this minimum precision. For instance, if a variogram to be used in ordinary kriging is postulated, one can calculate the variance of the prediction error (ordinary kriging variance) at any prediction point in the area, for any sampling pattern. This is because the error variance is independent of the values at the data points, and depends only on the variogram and on the locations of the data points and the prediction point. Conversely, for a given variogram and grid pattern, it is possible to calculate the grid spacing required to achieve a mean or maximum error variance that is smaller than a specified maximum. The OSSFIM program developed by McBratney and Webster (1981) can be used for the maximum error variance situation. This program can also be used to calculate the required grid spacing when the aim is to predict the spatial means of small blocks. Figure 8.2 is an example taken from archeology, where the aim is to map the altitude of the pleistocene surface, rich in artefacts, beneath a holocene layer of clays. The variogram used is a spherical model without nugget (range 338 m, sill 2354 cm^2). Figure 8.2 shows the standard deviation of the mean altitude of small $10 \times 10\,m$ blocks centred at the midpoints of the square grid cells (where this standard deviation is largest) as a function of the grid spacing.

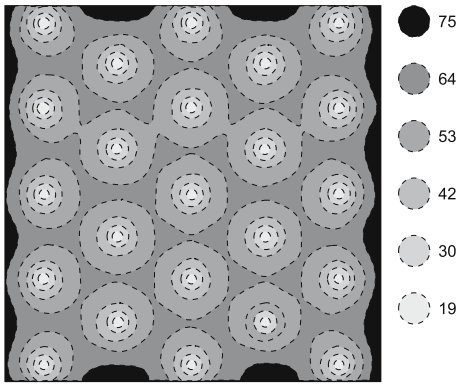


Fig. 8.3. Contourplot of the kriging variance for a triangular grid, showing a boundary effect

McBratney et al. (1981) assume that the data are interpolated by ordinary (block) kriging, which is appropriate in situations without trend. If there is a spatial trend, it can be accounted for by universal kriging or IRF-k kriging. For a given pattern and spacing of the grid, the variances of the universal kriging predictor and IRF-k kriging predictor will generally differ from the ordinary (block) kriging predictor. Conversely, for a given mean or maximum kriging variance, the required grid spacings will be different for these predictors. As for the ordinary kriging variance, the universal kriging variance and IRF-k kriging variance at a prediction location is independent of the sample values, and therefore one can calculate the mean or maximum universal kriging or IRF-k kriging variance for any grid spacing, and proceed as described above to select the required grid spacing.

8.3.3 Spatial Coverage Sampling

In practice, Centred Grid Sampling may be sub-optimal for several reasons:

- there may be prior measurements that cannot be matched with the regular grid;
- sampling may be hampered by enclosures that are inaccessible for sampling;
- irregular shape of the universe;
- the dimensions of the area may be small compared to the grid spacing that can be afforded. As a result, the prediction-error variances are relatively large near the boundaries of the target area (see Fig. 8.3 for an illustration of this boundary effect).

In these situations an alternative to a regular grid is a spatial coverage sample, also referred to as a space filling sample. For such samples, a pattern is calculated that fills in the space as uniformly as possible. This is achieved by minimizing a quality measure that is defined in terms of the distances between the prediction points (usually the nodes of a fine grid) and the sampling locations. Royle and Nychka (1998) proposed the general quality measure

$$J_{\mathrm{MD}} = \left\{ \sum_{i=1}^{N} \left(\sum_{j=1}^{n} D_{ij}^{p} \right)^{\frac{q}{p}} \right\}^{\frac{1}{q}} , \tag{8.37}$$

where D_{ij} is the distance between the ith grid node and the jth sampling location, and p and q are parameters to be chosen by the user. The inner sum is the sum of the distances raised to the power p of the ith node to all sampling locations. This sum can be thought of as a measure of how well the sample covers the ith node. The outer sum is a sum over all N nodes. Royle and Nychka (1998) used a point-swapping algorithm, starting from a random pattern, to find an optimal pattern. In this algorithm, a given point in the current sample is replaced by a candidate point, i.e., a grid node not yet included in the sample. If the quality measure is reduced by this swap, then the swap is accepted, i.e., the sampling location is replaced by the tested node. Since the process may end in a local minimum, depending on the starting pattern, the algorithm must be repeated a sufficient number of times with different starting patterns.

Brus et al. (2003) proposed to minimize the mean of the squared shortest distances (MSSD) of the grid nodes to the sampling locations,

$$J_{\mathrm{MSSD}} = \frac{1}{N} \sum_{i=1}^{N} \min_{j} \left(D_{ij}^{2} \right) . \tag{8.38}$$

An advantage of this measure is that it can be minimized by the fast k-means algorithm. In Sect. 7.2.4, k-means is proposed to construct compact geographical strata for design-based Stratified Simple Random Sampling, and we refer to this section for a brief description. The difference between Sect. 7.2.4 and the present section is that Sect. 7.2.4 addresses the clustering of the pixels (nodes of a discretization grid), whereas the present section focuses on the centroids. These centroids are taken as sampling locations. K-means proceeds in an iterative way, starting with an initial solution (e.g., a random sampling pattern), and repeatedly alternating two steps: re-allocation of grid nodes from one cluster to the other and re-calculation of the coordinates of the sampling locations (centroids of clusters). In the re-allocation step, each grid node is allocated to the nearest sampling location. In the re-calculation step, each sampling location is shifted to the mean of the coordinates of the grid nodes allocated to that sampling location. The iteration process is stopped when the J_{MSSD} cannot be reduced any further, or when the improvements become

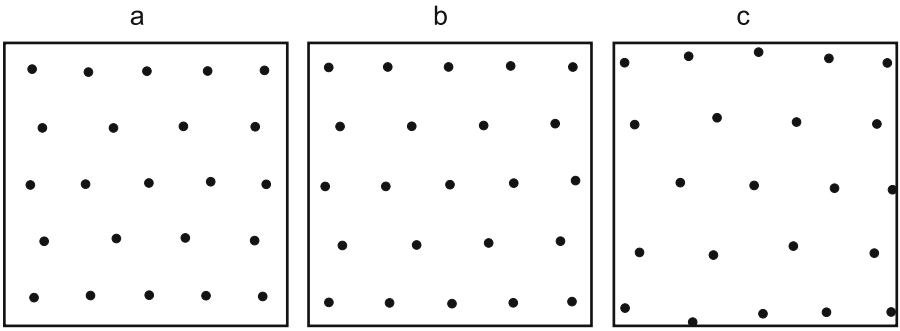


Fig. 8.4. Optimized pattern of 23 sampling locations in a square obtained by minimizing (**a**) the Mean Squared Shortest Distance with k-means; (**b**) the mean kriging variance with simulated annealing; (**c**) the maximum kriging variance with simulated annealing. A spherical variogram with a range of half the side of the square, and a nugget-to-sill ratio of 0.10 is postulated

smaller than a user-specified threshold. Re-calculation can be done as soon as a grid node has been re-allocated to a different sampling location, or it may be postponed and done when all grid nodes have been re-allocated. Figure 8.4**a** shows 23 sampling locations in a square, optimized by k-means. At the centre of the square, the points form a triangular grid, but at the sides of the square, this grid is clearly distorted. Four points have been drawn nearer to the boundaries. Brus et al. (2003) showed that the mean kriging variance for this spatial coverage sample is very close to the mean kriging variance of the geostatistical sample obtained by directly minimizing the mean kriging variance.

There are two shortcomings associated with k-means as it is described above. The first is that it does not account for data points that already exist in the area. However, existing data points can be easily accommodated in the k-means algorithm, simply by re-allocating the grid nodes to both the existing and the new sampling locations, but limiting the re-calculation step to the new sampling locations. This allows the existing points to remain in their place, while only the new points move. The other shortcoming is that with areas of a non-convex shape, one or more sampling locations may end up outside the area. If this is not acceptable, the problem can be solved by shifting such points to the nearest grid point after each re-calculation. This forces the sampling locations to stay within the area during the process. By way of illustration, we have optimized the pattern of 32 sampling locations, added to a sample of 6 prior points in the province of Gelderland, the Netherlands (Fig. 8.5).

8.3.4 Geostatistical Sampling

Geostatistical samples are samples obtained by minimizing a quality measure defined in terms of the variances of the geostatistical predictions errors. Two

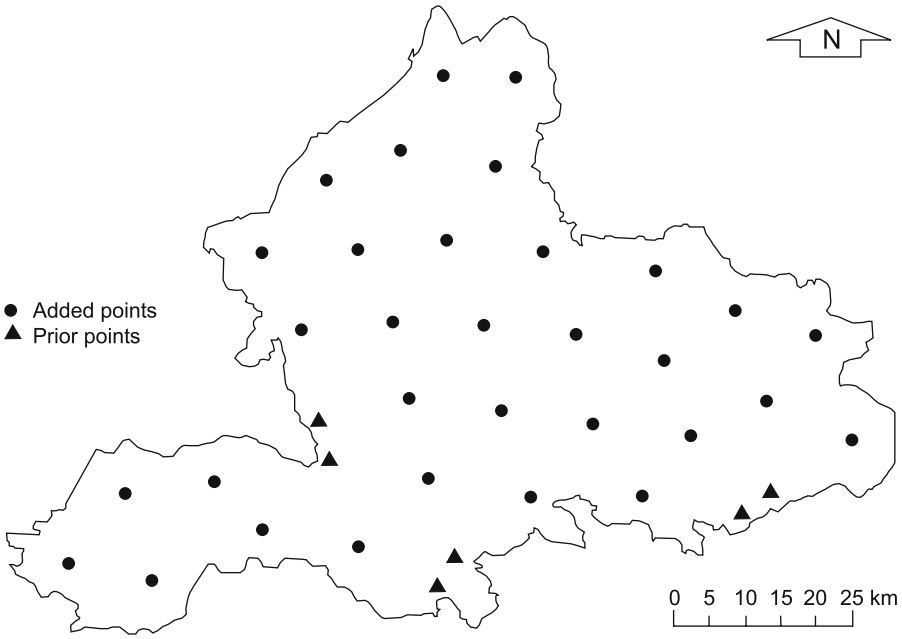


Fig. 8.5. Optimized pattern of 32 sampling locations added to a prior sample of 6 locations, obtained by minimizing the Mean Squared Shortest Distance with k-means, for the province of Gelderland, the Netherlands

commonly used quality measures for constructing geostatistical samples are the mean of the error variances (Sacks and Schiller, 1988; van Groenigen et al., 1999),

$$J_{\mathrm{MeanV}} = \frac{1}{N}\sum_{i=1}^{N}\left\{V\left(\widetilde{Z}_i - Z_i\right)\right\} \ , \tag{8.39}$$

and the maximum of the error variances,

$$J_{\mathrm{MaxV}} = \max_i \left\{V\left(\widetilde{Z}_i - Z_i\right)\right\} \ . \tag{8.40}$$

Figures 8.4**b** and 8.4**c** show the result when these two quality measures are used to optimize 23 locations in a square, and Figs. 8.6 and 8.7 for the 32 additional locations in the province of Gelderland. Whereas the differences between the J_{MSSD} sample and the J_{MeanV} sample are small, the locations for the J_{MaxV} sample are much closer to the boundaries.

From this we may tentatively conclude that:

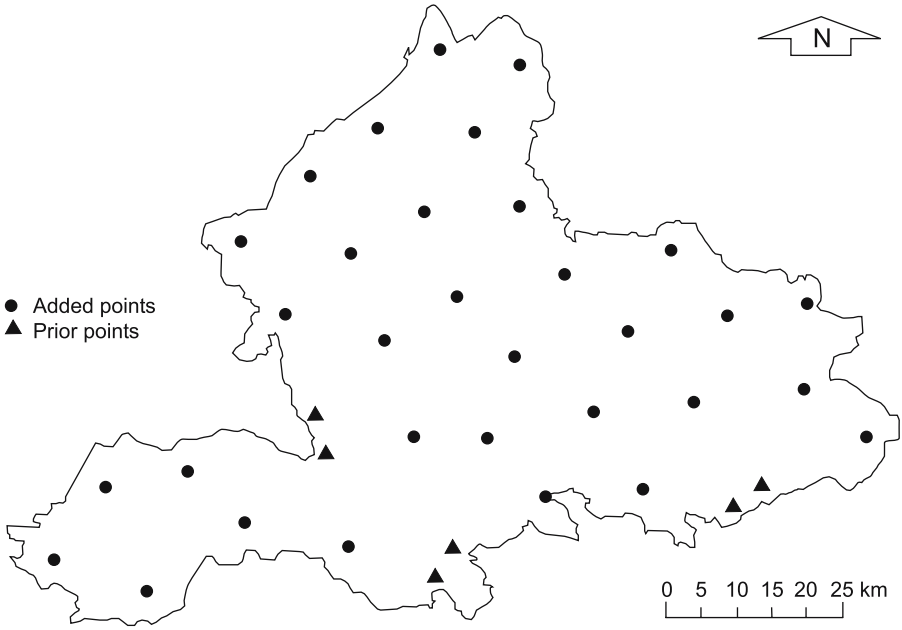


Fig. 8.6. Optimized pattern of 32 locations added to 6 prior locations, obtained by minimizing the mean kriging variance of predicted values at locations in the province of Gelderland, the Netherlands. A spherical variogram with a range of 50 km and a nugget-to-sill ratio of 0.40 is postulated

- if the aim is to minimize the mean kriging variance, then optimization with the J_{MSSD} quality measure gives satisfactory results, and computing time for the J_{MSSD} quality measure minimized by k-means is substantially less than for the J_{MeanV} quality measure minimized by simulated annealing (Appendix A);
- if the aim is to minimize the maximum kriging variance, then it is worth considering calculating a J_{MaxV} sample instead of a J_{MSSD} sample.

In practice, one is always to some extent uncertain about the variogram. This matters because the optimal geostatistical sample depends on the variogram model. Sacks and Schiller (1988) showed that the J_{MeanV} samples obtained with an exponential and a Gaussian model differed considerably. Van Groenigen et al. (1999) found that a linear, exponential and spherical model with comparable parameters resulted in very similar J_{MeanV} samples, and that a Gaussian model led to a very different sample. There is also an effect of the range and of the nugget-to-sill ratio. Van Groenigen et al. (1999)

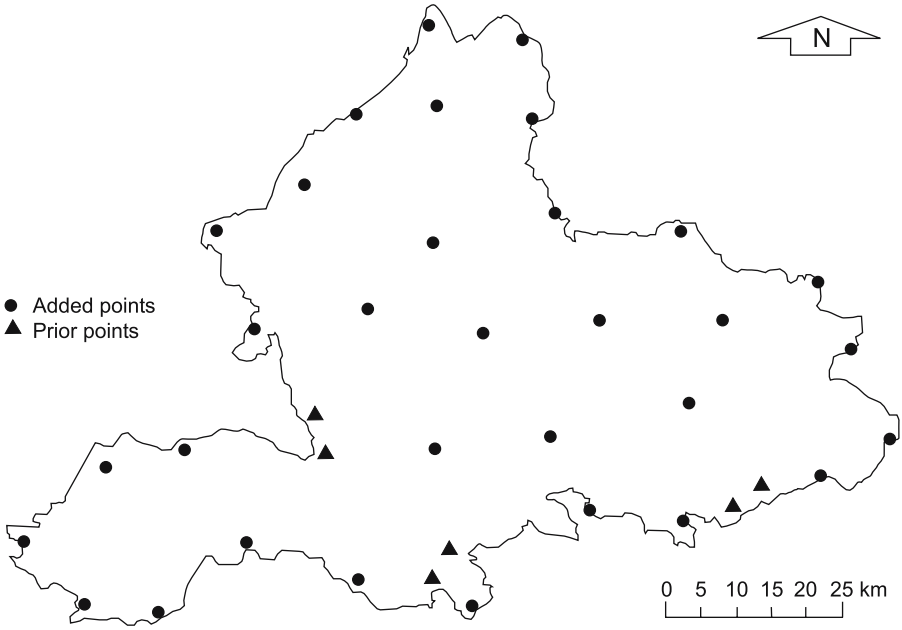


Fig. 8.7. Optimized pattern of 32 locations added to 6 prior locations obtained by minimizing the maximum kriging variance of predicted values at locations in the province of Gelderland, the Netherlands. A spherical variogram with a range of 50 km and a nugget-to-sill ratio of 0.40 is postulated

found that a spherical model with a range equal to 1/20 of the side of the square resulted in a very irregular sampling pattern. Neighbouring points were located at distances larger than twice the range, but the spatial coverage of the geostatistical sample was very poor. In practice, one would hesitate to collect such a sample, because it is never certain that the assumption of stationarity holds. The effect of the nugget-to-sill ratio on the sample showed up at the large value of 0.75 only, and even for this value the sample differed only marginally from those with smaller nugget-to-sill ratios.

As there is uncertainty about the variogram, one may decide to sample in two or more phases. Sampling in the first phase is done on a regular grid or according to a spatial coverage sample, supplemented with points at short distances of these grid nodes. These first-phase points are then used to estimate the experimental variogram and to fit a variogram model. For efficient sampling designs to estimate the variogram, we refer to Chap. 9. The fitted variogram is then used to design a second-phase sample, using either the J_{MaxV} or J_{MeanV} quality measure. Geostatistical samples also come into scope

when data is already available that can be used to estimate the variogram, and the aim is to add more locations. This is referred to as infill sampling.

Depending on the size of the area, the optimization algorithm, and the type of processor, designing a geostatistical sampling pattern may take much computing time, which may even become prohibitive. Therefore it is important to use efficient optimization algorithms. However, it will in many situations be unfeasible to calculate the quality measure for all patterns that can be formed by selecting m additional locations out of the N possible sampling locations. Benedetti and Palma (1995) compared several optimization algorithms (albeit for predicting the global mean), and concluded that simulated annealing (Appendix A) is a suitable solution.

Optimized Samples for Mapping with Universal Kriging

We now consider the situation where the target variable is not stationary, but shows a trend. This trend can be a spatial trend or a relation with other ancillary variables that are known everywhere in the target area. The relation between the target variable and the spatial coordinates or ancillary variables can be described by a (multiple) linear regression model. If one can postulate, prior to sampling, a regression model for the trend up to the predictors the regression coefficients need not be known), and a variogram for the regression residuals, then one can optimize the sampling locations by simulated annealing, using the mean or maximum universal kriging variance as a quality measure. citetheu05 showed that in this case the locations are spread in geographical space and in feature space. This is because estimation of the relation between the target variable and the ancillary variable profits from a large spread of the observations in feature space, while spatial interpolation of the residuals gains from a uniform spreading of the observations in geographic space.

8.3.5 Delineating Hot Spots

A hot spot is defined here as a relatively small area with a concentration of some compound or abundance of some object (e.g., a plant or animal species) that exceeds a certain threshold. This section deals with the question where the critical threshold is exceeded. Sampling for finding out whether at any point in the study area the critical threshold is exceeded, without asking where, is dealt with in Sect. 7.3.4.

When no prior measurements on the target variable are available, and one has no idea of the location of the hot spots, the best option is sampling on a regular grid (Sect. 8.3.2), or alternatively according to some spatial coverage (Sect. 8.3.3) or geostatistical sample (Sect. 8.3.4). There are two ways to increase the efficiency of these samples in order to delineate hot spots: sampling in two or more phases and composite sampling.

Phased Sampling

The efficiency of sampling can be increased by sampling in two, or even more phases (batches). For instance, the first phase could involve sampling the area on a centred, square grid. The sample data are then used to design an additional sample, by predicting the concentration at the nodes of the discretization grid, or alternatively the mean values of blocks, for instance remediation units. Based on these predictions, the nodes are classified as above the threshold (hot spot) or not. If the cost of a false positive (a node is classified as a hot spot whereas in reality it is not) equals the cost of a false negative (node classified as 'outside a hot spot', whereas in reality it is inside), it is natural to classify a node (or block) as a hot spot if the predicted value exceeds the threshold, $\widehat{Z}(\mathbf{s}_0) > z_\mathrm{t}$, or, alternatively, if the probability that the concentration exceeds the threshold (conditional on the measurements of the first batch) exceeds 0.50. Aspie and Barnes (1990) considered the case that the cost of false positives differs from that of false negatives. They show that for the optimal classification cut-off $z_{\mathrm{t}'}$

$$\Pr\left\{Z(\mathbf{s}_0) > z_\mathrm{t} \mid \widetilde{Z}(\mathbf{s}_0) = z_{\mathrm{t}'}\right\} = \frac{C_\mathrm{o}}{C_\mathrm{o} + C_\mathrm{u}}\ , \tag{8.41}$$

where C_o is the cost of false positives per unit area, and C_u is the cost of false negatives per unit area. This implies that if the conditional probability of exceeding the threshold exceeds the ratio of the cost of false positives and the sum of the two cost components, then the point is classified as a hot spot. Thus, if the cost of false negatives is larger than the cost of false positives, then the conditional probability of (8.41) must be smaller than 0.5. For a multivariate Gaussian stochastic function (SF), this is the case when $z_{\mathrm{t}'} < z_\mathrm{t}$.

When adding new sampling locations one would like to select these locations in such a way that, when the measurements of the first *and second* batches are used to predict the values at the grid nodes and to classify the nodes, the expected costs of misclassification are minimal. In formula, additional sampling locations are selected by minimizing the expected cost of misclassification (loss):

$$\begin{aligned} E(C) = \textstyle\sum_{i=1}^{N} \Big[& C_\mathrm{o} \Pr\left\{Z(\mathbf{s}_i) < z_\mathrm{t} \cap \widetilde{Z}(\mathbf{s}_i) > z_{\mathrm{t}'}\right\} + \\ & C_\mathrm{u} \Pr\left\{Z(\mathbf{s}_i) > z_\mathrm{t} \cap \widetilde{Z}(\mathbf{s}_i) < z_{\mathrm{t}'}\right\}\Big]\ , \end{aligned} \tag{8.42}$$

where N is the number of discretization nodes in the target area. Note that the probabilities in (8.42) are defined on bivariate distributions: the true value $Z(\mathbf{s}_i)$ is unknown, while the predicted value $\widetilde{Z}(\mathbf{s}_i)$ is also unknown, because the values at the locations of the second batch are unknown.

Christakos and Killam (1993) modified (8.42) for the situation in which the two cost components are not constant, but are linearly related to the magnitude of the error. For a multivariate Gaussian SF, the probabilities of (8.42)

can be calculated from the simple kriging prediction using the measurements of the first batch, the simple kriging variance using the first batch measurements, and the simple kriging variance using the first and second batch measurements (updated simple kriging variance). Even for small N ($N < 100$), an exhaustive search for the optimal pattern of additional sampling locations may be impracticable because of computing time (there are $\binom{N}{n}$ possible ways of selecting n additional sampling locations out of N possible locations). In this case, a random (non-exhaustive) search may be more efficient, for instance by simulated annealing (Appendix A).

A much simpler, more practical method for selecting additional locations in Phased Sampling is described by Englund and Heravi (1994). The measurements of the first-phase sample are used to predict the values at unsampled locations (nodes of a fine discretization grid) or the mean values of blocks (for instance remediation units) and the kriging variance of these predictions. A triangular probability distribution on the interval $[\widetilde{Z} - 3V(\widetilde{Z}),\ \widetilde{Z} + 3V(\widetilde{Z})]$ is then assumed to calculate for each prediction point (block) the probability of decision errors and the expected loss. The point (block) with the highest expected loss is selected as the first additional sampling unit. This procedure is repeated (the kriging variances must be updated in each iteration) until the predetermined number of additional locations has been selected.

A special case is when none of the prior measurements exceeds the threshold, and one wants to know whether the threshold is exceeded anywhere in the area. For this situation, Watson and Barnes (1995) proposed to select additional locations by maximizing the joint conditional probability that at least one of the new observations exceeds the threshold, which is equivalent to minimizing the joint conditional probability that all the observations are smaller than the threshold. Minimization of this quality measure requires substantial computing time, limiting its potentials for practical situations. For a multivariate (i.e., multipoint) Gaussian random field, adding a single location boils down to selecting the location with the minimum value for

$$\zeta(\mathbf{s}_0) = \frac{z_\mathrm{t} - \widetilde{Z}(\mathbf{s}_0)}{\sqrt{V\{\widetilde{Z}(\mathbf{s}_0) - Z(\mathbf{s}_0)\}}}\,, \tag{8.43}$$

where $\widetilde{Z}(\mathbf{s}_0)$ is the predicted value at the new sampling location obtained by ordinary kriging, and $V\{\widetilde{Z}(\mathbf{s}_0) - Z(\mathbf{s}_0)\}$ is the variance of the prediction error. Equation (8.43) shows that, depending on the threshold value, a location is selected either near a data point with a large value ($\widetilde{Z}(\mathbf{s}_0)$ is large), or in the empty space, where $V\{\widetilde{Z}(\mathbf{s}_0) - Z(\mathbf{s}_0)\}$ is large. A 'quick and dirty' method for adding a single location is to transform the data so that the univariate distribution is approximately Gaussian, then krige the transformed data, and select the location with the minimum value for ζ calculated from the transformed data. An even more 'quick and dirty' method to select more

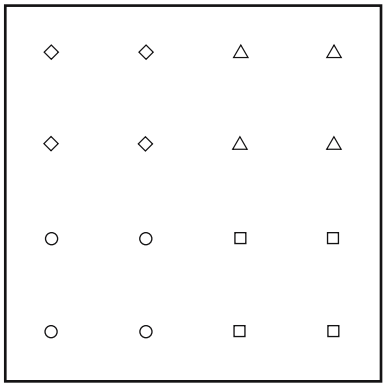


Fig. 8.8. Example of the formation of composites from aliquots taken on a grid. Aliquots from locations with the same symbol form a composite

than 1 location is to repeat this selection procedure, i.e., add one location at a time.

So far, nothing has been said about the sample sizes, i.e., the total sample size and the distribution of this total over the phases. If the costs of sampling and measurement per sampling unit are substantial compared to the costs of decision errors, then the optimal sample size can be calculated by including this cost component in the loss function. Englund and Heravi (1994) found that the optimal sample size is insensitive to the number of phases, and they therefore recommend determining this optimal sample size for one-phase sampling. They also found that to profit from Phased Sampling, the sample size in the first phase must be substantial: they recommend approximately 75% of the total sample size.

Hsiao et al. (2000) describe a method, based on Bayes' theorem, for calculating the size of the second batch required to detect the remaining blocks (remediation units) with mean concentrations above the threshold not detected in the first batch. The method requires as input an a priori probability that all hot blocks are detected in the first batch, and a parameter that determines the a priori probabilities that there is one undetected hot block, two undetected hot blocks, etc. Their method assumes that sampling locations are selected by Simple Random Sampling, whereas in practice locations will be selected purposively, for instance by the method described above.

Composite Sampling

Composite sampling comes into scope when the measurement costs per aliquot are high. In composite sampling, individual aliquots taken at the sampling locations are bulked (Sect. 4.3). Figure 8.8 shows an example where the four

aliquots at the corners of square cells are combined to form a composite. Analysing the composite instead of the individual aliquots reduces the measurement costs, so that a larger sample size can be afforded for the same budget. Due to the larger sample size, the sample has a better spatial coverage, so that the probability of hitting a hot spot is higher. The problem is that mixing the aliquots implies a loss of information on the concentration (or density) in the individual aliquots. The concentration of an individual aliquot may exceed the threshold, while that of the composite does not.

Several methods have been developed to identify the individual aliquots (further briefly 'individuals') with the highest values or with values above a threshold, by measuring some or all individuals of some composites, assuming that some of the material of the individuals is kept for later analysis (Gore and Patil, 1994; Gore et al., 1996). In these methods, the composites are sorted on the basis of their concentration in descending order. For each composite, an upper bound of the maximum value is calculated by assuming that the concentration of all except one of the individuals is zero. Under this assumption, the maximum equals the total of the composite, i.e., the composite concentration multiplied by the number of individuals in that composite.

In the simple sweep-out method (Gore and Patil, 1994), all individuals of the first ordered composite are measured, and the largest value in this composite is identified and recorded as the tentative maximum of the individuals of all composites (global maximum). If the upper bound of the maximum of the next ordered composite is smaller than this tentative global maximum, then it is clear that this is the true global maximum, and no further composites need to be measured. If the upper bound of the maximum for the next ordered composite is larger than the tentative global maximum, then this composite may contain an individual with a value larger than the tentative global maximum, and the individuals of this composite have to be measured. If the maximum identified in this second composite is larger than the maximum identified in the first composite, then the tentative global maximum is updated. This procedure is continued until the tentative global maximum is larger than the upper bound of the maximum in the next ordered composite.

If the aim is to identify the two largest values, the second largest value is also identified each time the individuals constituting a composite are measured, and the procedure is continued until the tentative second largest value is smaller than the upper bound of the maximum value in the next ordered composite. If the aim is to identify all individuals with values larger than a threshold, then the procedure goes on until the individual with the largest value below the threshold has been identified. It is also possible to identify $p\%$ of the total number of selected individuals with the largest values. For a self-weighting design, the smallest of these largest values can be used as an estimate of the pth quantile of the spatial cumulative distribution function.

Contrary to the simple sweep-out method, where, once a composite is selected, all individuals except one constituting this composite are measured, the sequential sweep-out methods (Gore et al., 1996) require measuring indi-

viduals of a composite as long as there is a possibility that there is a larger value among the unmeasured individuals of that composite. After each analysis of an individual, the total of the remaining, unmeasured individual samples is updated by subtracting the concentration of the measured individual sample from the previous total. This updated total is used as the upper bound of the maximum of the 'residual' composite, i.e., the composite that could be formed by mixing the remaining individuals. The procedure stops when the tentative global maximum exceeds the updated upper bound of the maximum of all (residual) composites.

There are two versions of the sequential sweep-out method, the locally and globally sequential sweep-out methods. In the locally sequential sweep-out method, once a composite has been selected, individuals of this composite are measured as long as there is a possibility that the remaining individuals have a maximum larger than the tentative global maximum, before another composite is considered for breaking down. In the globally sequential sweep-out method, after each analysis and updating of the total, the composite with the largest probability of containing the maximum is selected, i.e., the composite with the largest (residual) mean concentration. Gore et al. (1996) showed that the number of measurements needed to identify the largest value with these sweep-out methods can be much smaller than the total number of individuals, especially for the sequential sweep-out methods. For this purpose, therefore, the cost of measurement can be reduced considerably by compositing. The cost savings depend on aspects such as the probability distribution and the number of individuals per composite (composite size). The more extreme the largest values (the threshold), the larger the potential savings. The optimal composite size depends on the type of distribution. For a lognormal distribution, the optimal size increases with the coefficient of variation, from 3 (c.v. = 0.5) to 8 (c.v. = 2.0).

In the sweep-out methods described above, the individual to be measured in a selected composite is selected randomly, which is reasonable because there is no information available on the individual values. When composites are formed by a two-way compositing design some information is available, and it is possible to select the individual with the highest probability of having the maximum (Gore et al., 2001). Fig. 8.9 shows a two-way composite sampling design. From every individual aliquot, two portions are taken, one of which is combined with the portions from the same *column* to form a column composite, the other is combined with the portions from the same *row* to form a row composite. The individual that contributes to the column composite with the largest value *and* to the row composite with the largest value has the largest probability of having the maximum value, and will therefore be selected for measurement. For a sweep-out method for this two-way compositing design, we refer to Gore et al. (2001).

In the above methods, all values above a threshold are identified with certainty. In general, after all values above the threshold have been identified, some additional measurements are needed to confirm that all values above the

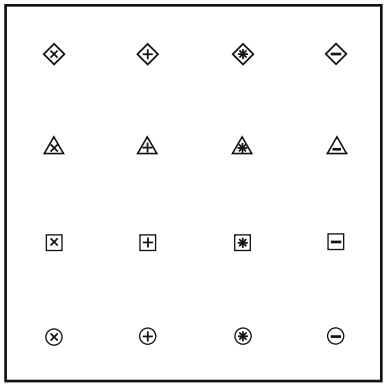


Fig. 8.9. Example of two-way composite sampling on a square grid. At the node of the grid, two samples are used, one for a row composite, one for a column composite

threshold have been identified. The alternative is to accept some uncertainty, and to continue measuring individuals until, for all (residual) composites, the probability that they contain an individual with a value above the threshold is smaller than some chosen limit. Carson (2001) describes a method to calculate from the concentration of the composite the probability that all individuals have concentrations less than the threshold. This method assumes that all combinations of individual concentrations that lead to the concentration measured on the composite have equal probability. Correll (2001) proposed to measure the individuals of all composites having a concentration higher than $z_t/\sqrt{k}$, where z_t is the threshold for the individual values, and k is the composite size. This modified threshold gave few false negatives (i.e., cases where the composite is below the modified threshold, while at least one of the individuals exceeds the original threshold), and few false positives (i.e., cases where the composite exceeds the modified threshold, while none of the individuals exceeds the original threshold). The modified threshold could be site-specific. False positives lead to unnecessary laboratory costs, while false negatives lead to risks to human health or the environment. However, these negative effects may be more than compensated for by the positive effects on costs and risks due to the increased sampling density.

9

Variograms

9.1 Introduction to Methods for Variograms

In model-based methods of spatial prediction (kriging) the values of the target variable at the sampling locations are considered to be realizations of a Stochastic Function (see Appendix B). In a spatial context a Stochastic Function (SF) is a field of spatially dependent random variables (RVs), and therefore is also referred to as a Random Field (RF). In many cases it is assumed that the mean difference of the RVs at two locations is zero (constant mean), and the variance of this difference depends on the spatial separation vector (lag) $\mathbf{h}$ only, not on the locations themselves**:**

$$E\{Z(\mathbf{s}_2) - Z(\mathbf{s}_1)\} = 0 \tag{9.1}$$

$$E\{Z(\mathbf{s}_2) - Z(\mathbf{s}_1)\}^2 = 2\gamma(\mathbf{s}_2 - \mathbf{s}_1) = 2\gamma(\mathbf{h}) \; . \tag{9.2}$$

A SF that meets these requirements is an intrinsic SF. The function $\gamma(\mathbf{h})$ is referred to as the (semi-)variogram. If this variogram is known, then one can obtain Best Linear Unbiased Predictions (BLUP) of the values at points or the means of blocks from sample data. One may also use the variogram for geostatistical simulation of fields (realizations) that are used, for instance, as input in a process-based simulation model. In this chapter we present sampling designs appropriate for estimating the variogram. In many situations the collected sample data are used both for estimating the variogram and for geostatistical interpolation or simulation. In Sect. 8.3 appropriate designs for sampling in one-phase are described. If the available time and budget allow for sampling in two phases, then we recommend to focus the first phase sample on estimating the variogram, and the second-phase sample on interpolation. In general, this two-phase sampling is more efficient, because the variogram estimated from the first phase sample can be used to optimize the sample for the second phase. This chapter describes principles for designing efficient samples for the first phase of such a phased sampling approach.

The first choice to be made is the size of the first phase sample. According to Webster and Oliver (1992), 150 locations might suffice in many situations,

and 225 locations would be almost certainly adequate in most applications where spatial variation is equal in all directions (isotropic). These sample sizes already make clear that for model-based prediction quite a few locations must be sampled. This is usually only realistic when many local means or the values at points must be predicted. Once one has decided on the number of locations to estimate the variogram from, one must choose the locations themselves. We distinguish two approaches for selecting the locations. In the first approach clusters of locations are selected. The locations within a cluster have a more or less regular, predetermined pattern. In this approach, one must choose a pattern, and a method or principle for selecting the clusters. In the second approach the sampling locations are optimized by minimizing a quality measure. In general this approach will result in an irregular pattern of locations. In this approach one must choose a quality measure and an algorithm for minimizing it.

9.2 Regular Patterns

9.2.1 Transect and Grid Sampling

Due to its operational advantages, a commonly used cluster type for estimating variograms is the transect. When sampling on transects, one has to choose the number of locations per transect, the distances between the locations, and the location and orientation of the transects. The distance between neighbouring locations on a given transect, i.e., the sampling interval, can be chosen constant or varying. Pettit and McBratney (1993) recommend transects in three directions, with the sampling locations in a given transect exponentially spaced. For instance, one may select 25 transects of 6 locations with interpoint distances of 0.2, 1, 5, 25 and 125 meter, i.e., each time the interval increases by a factor five. The transects must be evenly spread over the target area, for instance by dividing the area into squares or compact geographical strata of equal area (Sect. 7.2.4), and selecting one transect from each square (Fig. 9.1).

Transect sampling is appropriate when the average distance across the study area is large compared to the range of the variogram. For such areas sampling on a regular grid would give insufficient information on the semivariance at lags smaller than the range. If the distance across the area is small compared to the range of the variogram, then an alternative for transect sampling is sampling on a regular grid, supplemented by locations at a short distance of some grid nodes. To account for anisotropy, triangular grids are more suitable than square grids (Yfantis et al., 1987).

The 'short distance locations' are used to estimate the semivariance at lags smaller than the grid distance. Accurate estimates of the semivariance at small lags are important for estimating the nugget of the variogram, and for choosing between alternative models, for instance a spherical or a Gaussian model. It is

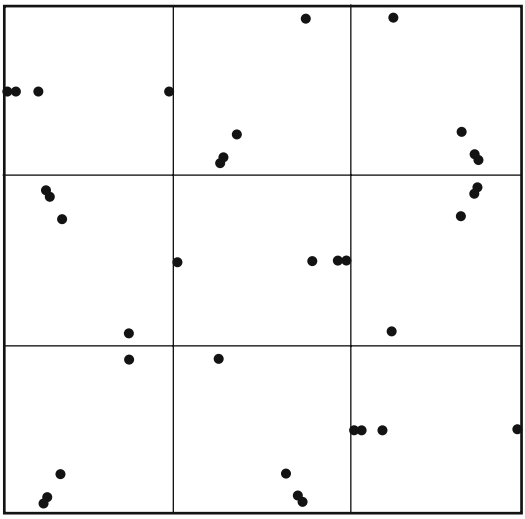


Fig. 9.1. Transect sampling from squares. After Pettit and McBratney (1993)

well known that if the variogram is used for geostatistical interpolation, then it is especially important to have reliable information on the variogram near the origin. We recommend to select at least 30 to 50 additional locations per lag, because this results into 30 to 50 disjoint pairs of locations, which is the minimum number of pairs mentioned by Journel and Huijbregts (1978).

To avoid spatial clustering of short distance locations in certain parts of the area, we recommend to select the grid nodes that will receive a short distance location purposively and not at random, for instance by subsampling the regular grid systematically (see Fig. 9.2). Also, we recommend to locate the additional locations on the sides of the grid cells, so that the directions for the smallest lag coincide with those of the larger lags.

9.2.2 Nested Sampling

In Nested Sampling the sampling locations are selected in stages (batches) in such a way that the distance between a randomly selected location of a given stage to a location of a previous stage is controlled. Figure 9.3 shows a nested design with four stages: in the first stage three locations with a mutual distance of h_1 m are randomly selected from the area. In the second stage at each of these locations in a random direction a location is selected at a distance of $h_2 = h_1/3$ m. We now have $3 \times 2 = 6$ locations. In the third stage at each of these six locations in a random direction a locations is selected at a distance of $h_2/3$ m, which makes the total number of locations $3 \times 2 \times 2 = 12$ locations.

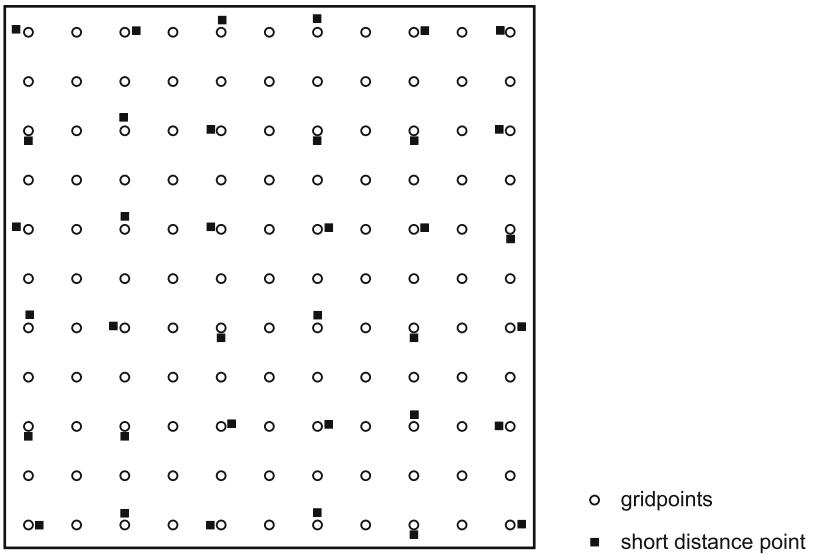


Fig. 9.2. Square grid sample with additional short distance locations for variogram estimation

This procedure is repeated once more, leading to a total of 24 locations. Note that in two dimensions it is impossible to select more than three mutually equidistant locations.

In practice, in the first stage often more than three locations are selected, for instance at the nodes of a square grid with a grid distance of h_1 m (Oliver, 1984). However, in that case one must be aware that the largest lag at which the semivariance is estimated is larger than h_1 m. Note that in the design of Fig. 9.3 the sample size doubles at each stage. So, if one has five stages with 9 locations in the first stage, the size of the nested sample becomes $9 \times 2 \times 2 \times 2 \times 2 = 144$. Adding a sixth stage implies an increase of the sample size of 144 locations.

Clearly, with nested designs one generally cannot afford many stages, because that would imply a too large sample size. However, full replication at each stage is unnecessary because with this design the variogram at the smaller lags is estimated much more precisely than at the larger distances. Therefore one may decide to replicate for the lower stages at only a proportion of the units, leading to so-called unbalanced designs. For the nested sample of Fig. 9.3 a 50% replication at the fourth stage leads to a total of 18 locations. For a nested design with five levels, 9 locations in the first stage, and a 50% replication at the fifth stage the total sample size becomes 108. One can now afford a sixth stage of 36 locations for the same budget, leading to an estimate of the variogram at one extra lag near the origin.

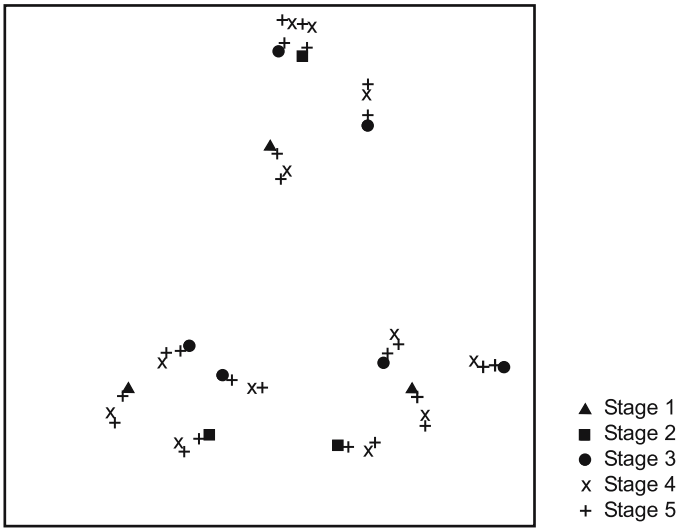


Fig. 9.3. Balanced nested sample with four stages. In the first stage three locations are selected. In subsequent stages at each of the sampling locations of the previous stages one additional location is selected

9.2.3 Independent Sampling of Pairs of Locations

If the method-of-moments is used to estimate the variogram (see Sect. 9.4.1), then one must be aware that the estimated semivariances for the M lags are logically correlated. These correlations should be taken into account when fitting the variogram model, but calculating the correlations is cumbersome. Moreover, from time-series analysis it is well known that due to these correlations one may be misled as to the type of model. For instance, the experimental variogram (the estimated semivariances for several lags) may show a hole effect which does not exist in reality.

These problems can be avoided by selecting pairs of locations independently as proposed by Brus and de Gruijter (1994). They proposed a design-based approach for estimating local (non-ergodic) variograms (Isaaks and Srivastava, 1988). In this approach, first a number of lags and the numbers of pairs per lag are chosen. To choose these lags and numbers the theory of experimental design can be used. Then for each lag the pairs of locations are selected by some type of design. For instance, in Simple Random Sampling (Sect. 7.2.3) of $M(h)$ pairs of locations with lag h, $M(h)$ locations are selected at random, with equal probability and independently from each other. Then for each location a counterpart is selected in a random or fixed direction at distance h from the starting location. If this counterpart is outside

the target area, also the starting location is omitted, and a new starting location is selected. Due to the independent selection of the pairs of locations, the estimated semivariances at the selected lags are uncorrelated, and their sampling variances can be estimated simply. The same holds for the sampling variance and covariance of the parameters of the fitted model. These advantages concerning the quantification of the uncertainty about the variogram are at the expense of the precision, because only $n/2$ (n is total number of locations) pairs of locations are used in the estimation, and the remaining $n\,(n-2)/2$ pairs are neglected. It depends on the situation how large this loss of information is and whether it is outweighed by the mentioned advantages.

9.3 Optimized Sampling Patterns

One may calculate the sampling pattern that explicitly has a minimum value for some objective function. In this case a quality measure and an algorithm to optimize it has to be selected. In the first papers on this subject a quality measure is proposed that quantifies how close the numbers of location pairs per lag class are to prespecified numbers (Warrick and Myers, 1987). However, the problem then shifts to the choice of the numbers of location pairs, and the question becomes what distribution of numbers of location pairs is best. Müller and Zimmerman (1999) and Lark (2002) have shown that a uniform distribution is sub-optimal; see also Müller (2001).

In subsequent papers it was proposed to base the quality measure on the variance-covariance matrix of the estimated parameters of a variogram model (Zimmerman and Homer, 1991; Bogaert and Russo, 1999; Müller and Zimmerman, 1999). For variograms that are non-linear functions of the parameters this is not straightforward. If one approximates the non-linear function by a first-order Taylor expansion, then the variance-covariance matrix of the parameters fitted by Generalized Least Squares to the estimated semivariances can be approximated by

$$\mathbf{V}_{\hat{\theta}} \approx \left(\mathbf{G}'_{\theta}\, \mathbf{V}_{\hat{\gamma}}^{-1}\, \mathbf{G}_{\theta} \right)^{-1} , \tag{9.3}$$

where $\mathbf{G}_{\theta}$ is the $M \times p$ matrix with the partial derivatives of the variogram, evaluated at the true (but unknown) values of the parameters:

$$\mathbf{G}_{\theta} = \begin{bmatrix} \frac{\partial \gamma(\mathbf{h}_1;\theta)}{\partial \theta_1} & \cdots & \frac{\partial \gamma(\mathbf{h}_M;\theta)}{\partial \theta_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial \gamma(\mathbf{h}_M;\theta)}{\partial \theta_1} & \cdots & \frac{\partial \gamma(\mathbf{h}_M;\theta)}{\partial \theta_p} \end{bmatrix} , \tag{9.4}$$

and $\mathbf{V}_{\hat{\gamma}}$ is the $M \times M$ variance–covariance matrix of the estimated semivariances. Cressie (1993) shows how $\mathbf{V}_{\hat{\gamma}}$ can be calculated when the SF is assumed

to be second-order stationary (Appendix B.1) and multivariate Gaussian. Bogaert and Russo (1999) and Müller and Zimmerman (1999) proposed to optimize the locations by minimizing the determinant of $\mathbf{V}_{\hat{\theta}}$, which is equivalent to maximizing the determinant of the so-called information matrix $\mathbf{G}'_{\theta}\,\mathbf{V}^{-1}_{\hat{\gamma}}\,\mathbf{G}_{\theta}$ in (9.3). Such designs are referred to as D-optimal designs. Note that both $\mathbf{G}_{\theta}$ and $\mathbf{V}_{\hat{\gamma}}$ depend on the parameters of the variogram, and therefore to minimize $\det(\mathbf{V}_{\hat{\theta}})$ with respect to the sampling locations, one must know the variogram. So, there is a circular problem.

A way out is to sample in two phases, and to use estimates of the parameters from the relatively small first phase sample to optimize the sample of the second phase. Bogaert and Russo (1999) optimized the pattern of 100 locations for an exponential variogram without nugget (sill = 1; effective range = 1/2 of side of square) and for an exponential variogram with nugget (nugget = 0.5; sill = 1; effective range = 1/2 of side of square). For the exponential variogram without nugget the gain in precision of the estimated parameters compared to Simple Random Sampling was limited. For the exponential variogram with nugget the optimized sample had many locations at very short distance, and as a result the estimated nugget was considerably more precise than for Simple Random Sampling. There was also gain for the sill, however for the range parameter the gain was again limited. Müller and Zimmerman (1999) and Boer et al. (2001) studied the effect of ignoring the correlations between the location pairs. This implies that the variogram is estimated by Weighted Least Squares (weights equal to $n(\mathbf{h})/\gamma^2(\mathbf{h};\theta)$) instead of Generalized Least Squares (Cressie, 1985). The quality measure then slightly changes: in the matrix $\mathbf{V}_{\hat{\gamma}}$ of (9.3) the off-diagonal elements are substituted by zeroes. Müller and Zimmerman (1999) found that this simplification led to a very similar pattern which was only slightly inferior to the optimal pattern. Boer et al. (2001) found that ignoring correlations led to strong clustering of locations, even for a variogram without nugget. Boer et al. (2001) also found that the value of the quality measure, $\det(\mathbf{V}_{\hat{\gamma}})$, and the pattern of locations was rather insensitive to changes in the parameters of the variogram, although there was some influence of the range and (for large ranges) of the nugget.

Finally, Lark (2002) proposed a quality measure based on the suitability of the estimated variogram for geostatistical interpolation. This seems rational because in many cases the ultimate goal is not the variogram itself, but a map obtained by kriging with the variogram. It is well-known that the kriging variance is more sensitive to changes in the variogram than the kriging prediction itself, and this is the main reason why Lark proposed to look at the kriging variance (Appendix B, equation B.19). Due to uncertainty about the variogram, there is also uncertainty about this kriging variance, and Lark (2002) proposed to use as a quality measure the variance of the kriging variance. This variance is approximated by a first-order Taylor expansion:

$$V(V_{\mathrm{K}}) \approx \mathbf{g}'_{V}\,\mathbf{V}_{\hat{\theta}}\,\mathbf{g}_{V}\ , \tag{9.5}$$

where $\mathbf{g}_V$ is the p-vector with partial derivatives of the kriging variance to the variogram parameters:

$$\mathbf{g}_V = \begin{bmatrix} \frac{\partial V_{\mathrm{K}}}{\partial \theta_1} \\ \vdots \\ \frac{\partial V_{\mathrm{K}}}{\partial \theta_p} \end{bmatrix} , \tag{9.6}$$

and $\mathbf{V}_{\hat{\theta}}$ is the $p \times p$ variance–covariance matrix of the estimated variogram parameters. Lark (2002) optimized 49 sampling locations assuming an exponential variogram with nugget, $\gamma(h) = c_0 + c_1\{1 - \exp(-h/a)\}$, at three levels for the distance parameter a, and for the ratio of spatial dependence $c_1/(c_0+c_1)$, resulting into nine combinations. Lark (2002) considered the kriging variance at the centre of a square grid of 5 units. Figure 9.4 shows the optimized sampling locations. For a small ratio of spatial dependence (large nugget-to-sill ratio) and/or a small range the optimized sample showed several clusters of locations. For the intermediate ratio of spatial dependence combined with the two largest ranges the optimized sample showed a more or less regular distribution with some of the locations supplemented by an additional location at short distance. For the largest ratio of spatial dependence and the two largest ranges the optimized sample has a more or less regular pattern with several chains of locations. Lark (2002) compared the optimized samples with 7 randomly selected transects of 7 locations with a regular spacing. He found comparable values for the quality measure, and therefore concluded that when one is ignorant about the variogram, then the most robust approach is to sample on transects. In an experiment where a first phase sample of 7 randomly selected transects of 7 locations each was supplemented by an optimized sample of 31 locations, there was a benefit from the optimization compared to sampling entirely on transects.

9.4 Estimating the Variogram

We shall now describe how the variogram can be estimated from the sample data. The most widely used method for variogram estimation is the method-of-moments. Alternatively, the variogram can be estimated by the maximum likelihood method. In principle for all sampling designs described above both estimation methods can be used. For nested designs the natural way of estimating the variogram is ANOVA. By summing the variance components associated with the stages we get the variogram for the chosen lags (Webster and Oliver, 1990). Miesch (1975) and Corsten and Stein (1994) have shown that for balanced designs ANOVA and the method-of-moments are equivalent and lead to identical experimental variograms. For unbalanced designs these two methods lead to different estimates.

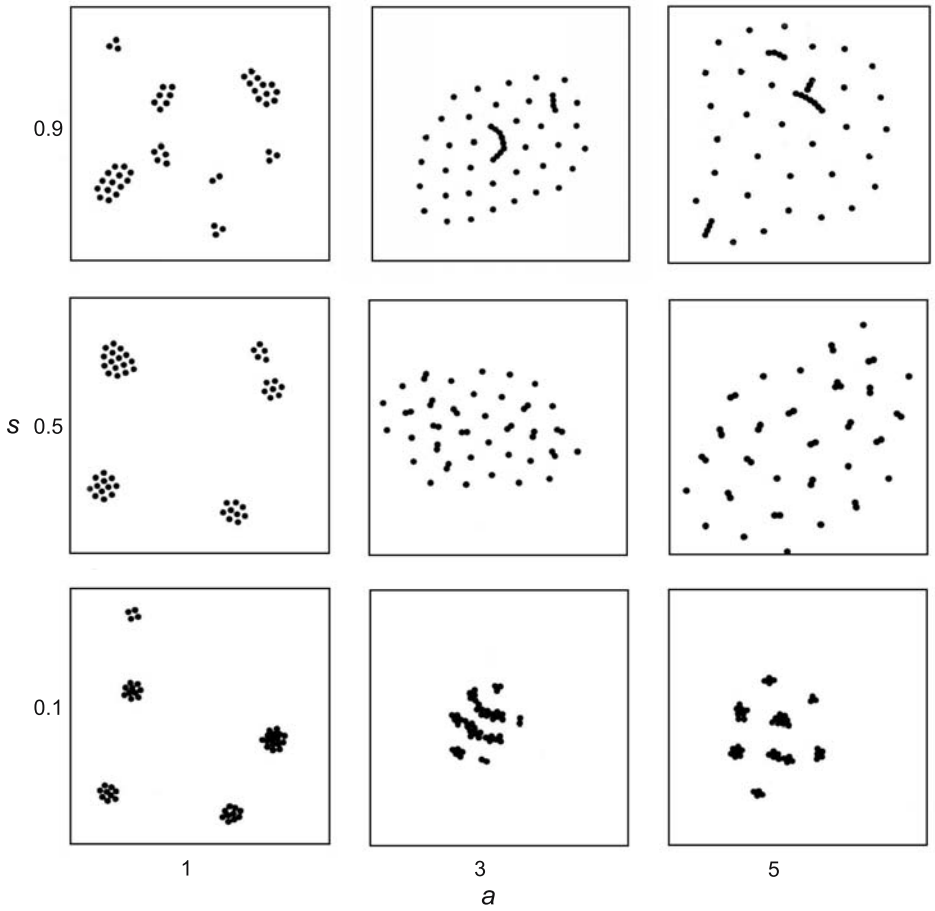


Fig. 9.4. 49 sampling locations optimized for the kriging variance of predictions at the centre of the square grid cells with a spacing of 5 distance units. An exponential variogram is assumed, with varying distance parameters, and ratios of spatial dependence. (Reproduced from Lark (2002, p. 69) with permission from Elsevier.)

9.4.1 Method-of-Moments

With the method-of-moments, the variogram is estimated in two steps. In the first step, the data are used to form pairs of locations. Then we estimate the variogram for a given lag $\mathbf{h}$ by selecting all pairs of sampling locations $\mathbf{h}$ apart, and calculating:

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2M(\mathbf{h})} \sum_{i=1}^{M(\mathbf{h})} \{z(\mathbf{s}_i) - z(\mathbf{s}_i + \mathbf{h})\}^2 \ , \tag{9.7}$$

where $M(\mathbf{h})$ is the number of pairs separated by lag $\mathbf{h}$. Unless locations are chosen at regular intervals along transects or on grids, it is necessary to group the lags by distance and direction, i.e., to use tolerance intervals both for the length and for the direction of the vector. If it is assumed that the variogram is independent of direction, i.e., an omnidirectional or isotropic variogram, then one may group all lags with (approximately) the same length, and estimate the experimental variogram for these lags. For kriging one needs a continuous function, and therefore in the second step a continuous model is fitted to the experimental variogram. Only models are permitted which ensure that the variances of the prediction errors are non-negative. Permissible models that are commonly used are the spherical, the exponential, and the Gaussian model (Appendix B). The estimates $\hat{\gamma}(\mathbf{h})$ are correlated, and therefore the parameters of the model must be estimated by taking into account the variances and covariances of the values in the experimental variogram. In statistical literature this is referred to as Generalized Least Squares estimation. This implies that the parameters must be estimated iteratively, because the variances and covariances depend on the variogram itself. Cressie (1985) recommended to neglect the covariances, and to fit the model by Weighted Least Squares, using the number of pairs divided by the squared semivariance in the previous iteration as weights.

9.4.2 Maximum Likelihood Estimation

Contrary to the method-of-moments, with the maximum likelihood method the data are not paired into couples, and the variogram is estimated in one step. To apply this method one typically assumes that the n sample data are a realization of a second-order stationary n-variate Gaussian Stochastic Function. Second-order stationarity is a slightly stronger assumption than the intrinsic hypothesis because the variance of the process is assumed finite (Appendix B). Especially the assumption that the process is multivariate Gaussian is a rather strong assumption. When the Spatial Cumulative Distribution Function is clearly non-Gaussian, we recommend transforming the data first, for instance by taking logarithms or square roots, and to estimate the variogram of the transformed data.

One needs the assumption of a multivariate Gaussian process, because then the joint probability density of the sample data can be calculated by:

$$P(\mathbf{z}, \boldsymbol{\mu}, \mathbf{p}) = (2\pi)^{-\frac{n}{2}} |\mathbf{V}|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} (\mathbf{z} - \boldsymbol{\mu})' \, \mathbf{V}^{-1} \, (\mathbf{z} - \boldsymbol{\mu}) \right\} , \tag{9.8}$$

where $\mathbf{z}$ is the vector with the n sample data, $\boldsymbol{\mu}$ is the vector with means (all values are equal), $\mathbf{p}$ is the vector with parameters of the covariance function, and $\mathbf{V}$ is the $n \times n$ matrix with variances and covariances of the sample data. This equation may not be familiar to the reader, however rewriting this equation for $n = 1$ gives

$$P(z, \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{z-\mu}{\sigma}\right)^2\right\}, \tag{9.9}$$

the univariate normal density function. Note that there is a matrix with covariances in the equation, and not a matrix with semivariances, however for second-order stationary Stochastic Functions the variogram can be obtained from the covariance function (Appendix B, equation B.15). If the values at the sampling locations are considered as fixed, then (9.8) can be used to calculate the probability of finding these values for any combination of values for μ and for the parameters of the covariance function. The parameters of the covariance function can now be estimated by maximizing this probability. The estimates thus obtained are referred to as maximum likelihood estimates. Usually the logarithm of the probability is taken and multiplied by -1, and this negative log-likelihood is minimized, which is equivalent to maximizing the likelihood itself. Lark (2000) compared the method-of-moments and the maximum likelihood method. In general the maximum likelihood estimates were better than the method-of-moments estimates, especially for small sample sizes (say $n < 75$) and when spatial structure is moderate to strong (small nugget, large range). Lark (2000) also found that, although the method-of-moments does not make the multivariate Gaussian assumption, this method is equally sensitive to skewness of data as the maximum likelihood method. For larger sample sizes ($n > 150$) ML estimation becomes impractical because a huge number of computations are then required. For more information on this method we refer to Pardo-Igúzquiza and Dowd (1998).

Part III

SAMPLING IN TIME

10

Introduction to Sampling in Time

From the definition in Sect. 1.1 follows that monitoring concerns repeated or continued sampling in time, with or without a spatial extent. This part of the book addresses sampling in time of objects without a spatial extent, i.e., repeated or continued sampling at one location. It should be mentioned that the methods described here can also be applied in a space–time context, e.g., with multiple time series distributed in space, or a time series of spatial means of an area. Sampling in space–time, however, is treated in Part IV.

In contrast to Part II on sampling in space, a substantial portion of this Part on sampling in time is devoted to estimation of model parameters such as the temporal trend. The reason for this is that in monitoring of processes, the interest is often in model parameters, because these parameters tell us what can be expected on average, and not in target quantities only related to a bounded monitoring period.

Chapter 11 describes sampling methods for estimation of global quantities in time. Many methods for spatial sampling given in Chaps. 7 and 8 apply to temporal sampling as well: the methods can be adapted straightforwardly to the 1D context. Notwithstanding the similarities, temporal sampling differs in two principal aspects from spatial sampling.

First, time is unidirectional and sampling in past time is impossible, hence designs such as Two-Phase Random Sampling are inappropriate for sampling in time. Second, while spatial sampling always takes place in a bounded universe, the temporal universe to be monitored may have an end-point that is still undetermined when the monitoring scheme is designed. This has several consequences. One practical consequence is that, instead of the total sample size, the average sample size per unit of time or the sampling frequency becomes the major parameter of the monitoring sample.

Chapter 13 discusses the sampling aspects of time-series modelling. In Sect. 13.2 the sampling aspects of estimating process parameters are discussed. These parameters characterize the dynamic behaviour of a process. Next, in Sect. 13.3 the estimation of means, such as annual means, is elaborated upon.

In contrast to Chap. 11 we focus on *model* means here. Finally, Sect. 13.4 discusses the sampling aspects in the context of detecting trends in time series.

11

Global Quantities in Time

11.1 Introduction to Methods for Global Quantities in Time

Sampling in time is done, for instance, to monitor the quality of surface water at some critical location in a river or at the outlet of a water-management unit, or to monitor the soil quality at a control site. It is similar to sampling in space in the sense that, although the practical aspects may differ, the same principles, theory and problems of choice play a role. Therefore, much of what has been said about spatial sampling in Chap. 7 is applicable to temporal sampling as well. In particular, the distinction and the choice between the design-based and the model-based approach is again of paramount importance and is taken here too as the main sub-division of the methodology. See Sect. 4.1 and Chap. 6 for a general discussion of how to choose between these two approaches.

In the case of sampling in time, it should be added that cyclic variations seem to be more common in time than in space. If this is true, then for sampling in time more caution is needed with systematic (random) sampling, i.e., at constant time-intervals, because of a greater risk that the sampling interval interferes with some cyclic pattern of variation (see also Sect. 7.2.7). On the other hand, taking samples at constant intervals is often more convenient. Of course this advantage vanishes when a programmable automatic sampling device can be installed.

As with sampling in space, design-based methods are usually most appropriate for global quantities in time. They have the advantage of greater simplicity and more robustness in the sense that the statistical inference from the sample data does not rely on the validity of a time-series model. This is especially important for regulatory monitoring. We repeat that when interest is in the model-mean, one should ship this chapter and proceed to Chap. 13.

11.2 Design-Based Methods for Global Quantities in Time

11.2.1 Introduction

Random sampling is less common in time than in space, Nevertheless, most of the methods as presented in Sect. 7.2 for design-based spatial sampling can be used for sampling in time. Clearly, in 2D spatial sampling, populations, domains, strata and primary units are all areas, in temporal sampling they are periods of time. The advantages and disadvantages indicated for the various spatial sampling strategies in Sect. 7.2 hold, mutatis mutandis, for sampling in time. If the end of the monitoring period is pre-determined, the selection techniques presented in Sect. 7.2 for the 2D spatial context only need obvious adaptations to the 1D temporal context. Long-term monitoring projects often have no pre-determined end, but budgets tend to be allocated periodically, often annually. In that case it is practical to take the budgetary years as strata, and to determine the sample size for each successive year from the available budget.

11.2.2 Practical Issues

A point which deserves special attention in the application of design-based strategies in time is non-response. Non-response generally indicates the situation where for some reason no data can be obtained from a sampling unit. If non-response occurs in spatial sampling, data are often collected from additional sampling units which were kept in reserve, in order to meet the required sample size. In temporal sampling, however, this would lead to overrepresentation of the end of the monitoring period since sampling in the past is not possible. To avoid bias as a result of this overrepresentation, temporal stratification is recommended.

There are some exceptions to the rule that the 2D spatial sampling designs of Sect. 7.2 are straightforwardly applicable in time. In Cluster Random Sampling (Sect. 7.2.6) the configuration of the clusters is restricted to 'transects' with equidistant points in one 'direction', namely the time scale. For instance, the cluster might be a period of a day during which is measured hourly. Cluster Random Sampling in time has the operational advantage of reducing the travel time between, for example, the field spot and the laboratory.

Systematic Random Sampling (Sect. 7.2.7) is straightforwardly applicable in time. However, the warning is emphasized that if the temporal variation is cyclic or pseudo-cyclic, the variance may be severely underestimated. Note that cyclic or pseudo-cyclic variations, such as daily or yearly fluctuations, are often found in environmental temporal variables.

The systematic unaligned type of design in Sect. 7.2.8 needs two dimensions and is therefore not appropriate for sampling in time. In place of that, the Markov Chain design discussed in Sect. 7.2.8 is well suited to achieve a

fairly even spreading of sampling times, while still avoiding the risk of interference with cyclic or pseudo-cyclic variations.

Sampling with probabilities proportional to size (pps, Sect. 7.2.9) cannot be applied in sampling in time, because the required information on an ancillary variable is not available at the time that the sampling strategy is designed. For the same reason Sequential Random Sampling (Sect. 7.2.10) may not be appropriate for sampling in time. However, in some cases it might be reasonable to extrapolate ancillary variables to the future, in particular those which are dominated by a seasonal fluctuation.

Adaptive Cluster Sampling (Sect. 7.2.10) is not appropriate for sampling in time, because the method would require sampling in the past which is impossible. For the same reason Two-Phase Sampling strategies (Sect. 7.2.12) are not straightforwardly applicable. A possible way to apply Two-Phase Sampling is to collect a large number of samples and to store them. After the monitoring period the samples are analysed using an inexpensive, indicative technique in the first phase. Next, in the second phase, a subsample is analysed by using a more precise methods.

11.2.3 Estimating Temporal Means by Using an Ancillary Variable

If the temporal variation of the target variable differs between periods of time, then a constant interval between sampling times can be inefficient. If one has a time series of an ancillary variable correlated with the target variable, then one can use this ancillary variable for selecting a systematic sample with varying intervals between sampling times. The time series of the ancillary variable is used to calculate cumulative totals for $t = t_\mathrm{b}, t_{\mathrm{b}+1}, \ldots, t_\mathrm{e}$, where t_b and t_e are the beginning and end time of the monitoring. Then a systematic random sample with constant interval is drawn between 0 and the cumulative total at t_e. Note that the starting point is selected randomly. The points in time at which the cumulative total equals the randomly selected numbers are taken as sampling times. Figure 11.1 illustrates this principle for estimating the annual mean concentration of, e.g., nitrate in soil pore water at a given depth, using the precipitation surplus (precipitation minus evapotranspiration) as an ancillary variable (when the precipitation surplus is negative, it is taken as zero, so that the cumulative precipitation surplus is a non-decreasing function). Note that the sampling frequency in summer, when the precipitation surplus is small, is much lower than in winter.

11.2.4 Testing Step Trends

One important purpose in temporal sampling is not covered by the design-based methods presented for spatial sampling in Sect. 7.2: estimation or testing of a step trend. If interest lies in possible effects of a sudden natural or human-induced change that starts at a given point in time, then a relevant

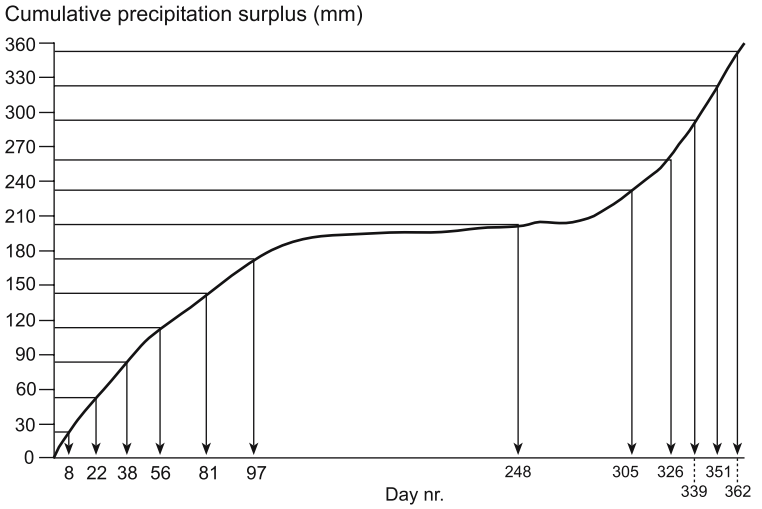


Fig. 11.1. Systematic Random Sampling in time, with sampling times equidistant in cumulative precipitation surplus

quantity to estimate may be the difference between the temporal means before and after the change:

$$D = \bar{z}_{\mathrm{a}} - \bar{z}_{\mathrm{b}} = \frac{1}{t_{\mathrm{e}} - t_{\mathrm{c}}} \int_{t_{\mathrm{c}}}^{t_{\mathrm{e}}} z \, \mathrm{d}t - \frac{1}{t_{\mathrm{c}} - t_{\mathrm{b}}} \int_{t_{\mathrm{b}}}^{t_{\mathrm{c}}} z \, \mathrm{d}t \ , \tag{11.1}$$

where $\bar{z}_{\mathrm{a}}$ and $\bar{z}_{\mathrm{b}}$ are the temporal means after and before the change, respectively, t_{b} and t_{e} are the beginning and end time of the monitoring, and t_{c} is the time at which the change happens. This effect is simply estimated by:

$$\widehat{D} = \hat{\bar{z}}_{\mathrm{a}} - \hat{\bar{z}}_{\mathrm{b}} \ , \tag{11.2}$$

where $\hat{\bar{z}}_{\mathrm{a}}$ and $\hat{\bar{z}}_{\mathrm{b}}$ are estimators of the temporal means, depending on the type of sampling design (see Sect. 7.2). If the samples taken before and after the change are taken independently from each other, then the variance of $\widehat{D}$ equals:

$$V\left(\widehat{D}\right) = V\left(\hat{\bar{z}}_{\mathrm{a}}\right) + V\left(\hat{\bar{z}}_{\mathrm{b}}\right) \ , \tag{11.3}$$

where $V\left(\hat{\bar{z}}_{\mathrm{a}}\right)$ and $V\left(\hat{\bar{z}}_{\mathrm{b}}\right)$ are the true sampling variances of the estimated means. An estimate $\widehat{V}\left(\widehat{D}\right)$ of $V\left(\widehat{D}\right)$ can simply be obtained by inserting the estimates of $V\left(\hat{\bar{z}}_{\mathrm{a}}\right)$ and $V\left(\hat{\bar{z}}_{\mathrm{b}}\right)$, as given in Sect. 7.2 for the various designs:

$$\widehat{V}\left(\widehat{D}\right) = \widehat{V}\left(\hat{\bar{z}}_{\mathrm{a}}\right) + \widehat{V}\left(\hat{\bar{z}}_{\mathrm{b}}\right) \ . \tag{11.4}$$

A two-sided $100(1-\alpha)$ % confidence interval for D is given by:

$$\widehat{D} \pm t_{1-\alpha/2} \cdot \sqrt{\widehat{V}\left(\widehat{D}\right)}\,, \tag{11.5}$$

where $t_{1-\alpha/2}$ is the $(1-\frac{\alpha}{2})$ quantile of the Student distribution with $(\eta_\mathrm{a}+\eta_\mathrm{b})$ degrees of freedom; η_a and η_b denoting the degrees of freedom on which the estimates $\widehat{V}(\hat{\bar{z}}_\mathrm{a})$ and $\widehat{V}(\hat{\bar{z}}_\mathrm{b})$ are based. The null-hypothesis of no effect ($D=0$) can be tested against the alternative $D \neq 0$ with the two-sided two-sample t-test. The null-hypothesis is rejected if the confidence interval of (11.5) does not contain zero. It is emphasized that this method of trend testing is based on temporal means, in contrast to the methods described in Sect. 13.4, which are based on model means.

11.3 Model-Based Methods for Global Quantities in Time

11.3.1 Introduction

The simplest and most common model-based method of sampling in time is systematic sampling, i.e., at constant time-intervals, and to use a time-series model for statistical inference from the sample data. Time-series analysis is a broad subject on its own, and a vast literature exists on its methodology. A practical textbook is Box and Jenkins (1976); see Hipel and McLeod (1994) for applications in natural resources. These books discuss in detail how models of the temporal variation may be selected and fitted to the data, and how these models can be used to test, estimate and forecast quantities of interest. We repeat the warning in Sect. 3.7 that some formulas for sample size imply that the *model* mean is to be estimated, and therefore render sample sizes that are larger than needed for estimating the often more relevant *temporal* mean, the average of the target variable over the monitoring period.

In systematic sampling, an appropriate sampling frequency or interval length must be chosen. Many time-series models are based on equidistant time series, as is explained in Appendix C. The tendency to sample at constant time-intervals is obviously caused by operational advantages but probably enforced by the fact that equidistant series can be analyzed by methods which are mathematically relatively simple. However, taking spatial sampling as an analogy, it can be conjectured that systematic sampling is not always the best option even in the model-based approach. As explained in Sect. 7.2.4, if in spatial sampling sub-areas can be outlined beforehand that are expected to be more variable than others, then it is efficient to stratify accordingly and to sample more densely in the more variable strata. Similarly, if the temporal variation varies with time then it should be efficient to sample more densely in periods with larger variation. Another example can be borrowed

from Sect. 7.3 on model-based sampling in space, where it was demonstrated that when optimizing a sample for an area in which certain parts are not accessible for sampling, the sampling locations are attracted or repelled by the boundaries of the inaccessible parts, depending on whether they belong to the target area or not. Similarly, if sampling is impossible in one or more sub-periods, then a sampling design adapted in the same way should be more efficient than an equidistant one.

11.3.2 Kriging the Temporal Mean

As an alternative to the estimation of model means by the methods described in Sect. 13.3, the mean value of a variable over a bounded period of time can be predicted. This can be done by block-kriging. It is important to note that with block-kriging the temporal mean of a single realization is predicted, whereas in Sect. 13.3 methods are described that estimate *model* means, μ. Temporal means might be relevant in compliance monitoring, when one is interested in the departure of the true mean from a critical value during a certain bounded period, for instance a year. Besides this, estimation of temporal means by block-kriging is recommendable if only an unevenly spaced time series is available. Block-kriging for prediction of spatial means is described in Appendix B. Basically, the block-kriging system is straightforwardly restricted to one dimension when it is applied to periods of time. Figure 11.2 shows the ordinary block-kriging variance as a function of the sample size, for a centred systematic sample and an exponential variogram. The block-kriging variance is expressed as a ratio of the sill (horizontal asymptote) of the variogram. Figure 11.2 shows that the larger the nugget-to-sill ratio and the smaller the relative range, the more sampling times are needed to achieve a given precision (block-kriging variance). The four lines in Fig. 11.2**d** are very close, showing that the effect of the range on the block-kriging variance is small for large nugget-to-sill ratios.

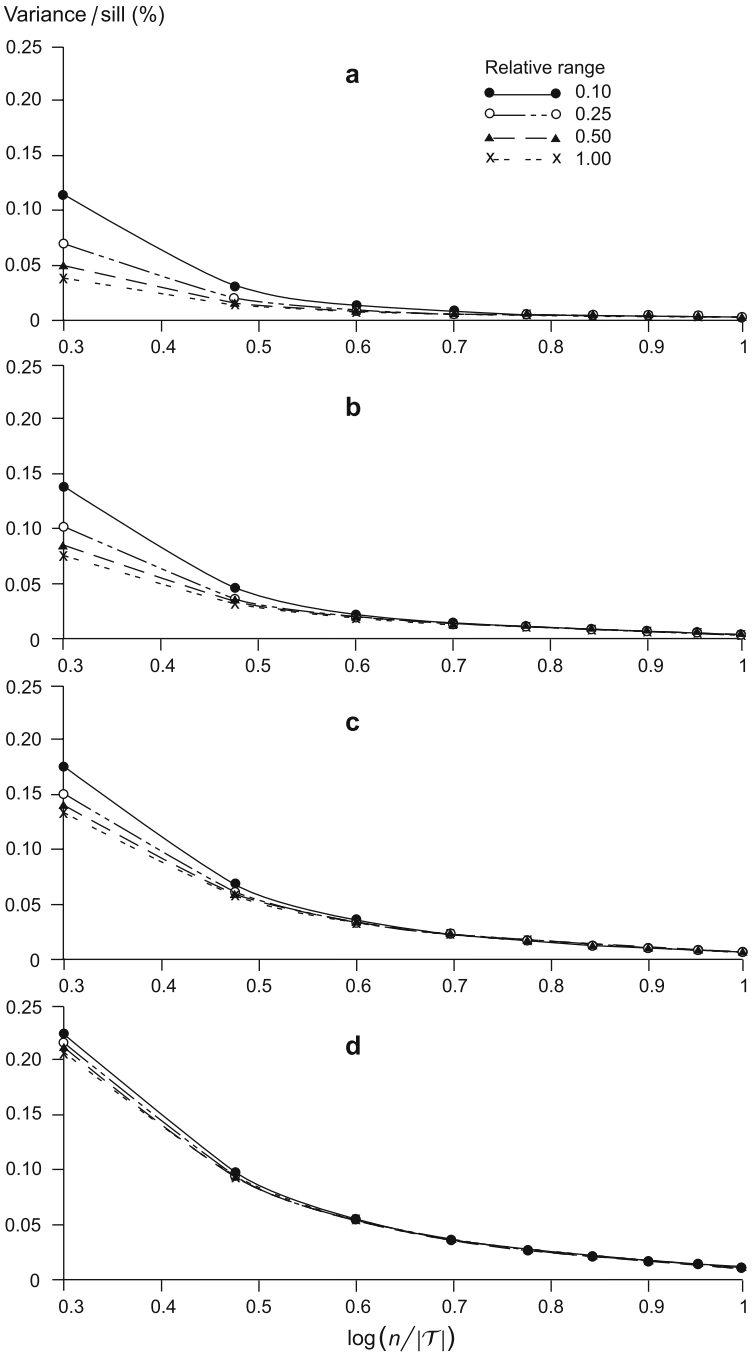


Fig. 11.2. Block-kriging variance of the temporal mean predicted from a centred systematic sample with interval length $\Delta s = |T|/n$, as a function of the sample size, for exponential variograms. The variance is expressed as a percentage of the sill of the variogram. Nugget-to-sill ratios: **a**: 0.10; **b**: 0.25; **c**: 0.50; **d**: 0.80

12

Local Quantities in Time

12.1 Model-Based Methods for Local Quantities in Time

12.1.1 Introduction

This chapter deals with retrospective prediction of the temporal means for many short periods of time or of the values at points in time. An example is the reconstruction of a continuous time series of water table depth, to be used as input of a model for the leaching of heavy metals in the soil. The non-linear relation between model-output and model-input makes that the global, temporal mean is insufficient as input. For local prediction in time model-based methods are most appropriate. Kriging is a flexible method because it does not require observations at regular intervals.

12.1.2 Kriging Values at Points in Time

When prior information can be used to postulate a tentative variogram for the random variable $Z(t)$, then this variogram can be used to decide on the interval length for systematic sampling in time, given a quality requirement on the mean or maximum error-variance of predictions at intermediate times.

Figures 12.1 and 12.2 show the mean and maximum ordinary kriging variance as a function of the interval length, for exponential variograms with different nugget-to-sill ratios. The kriging variances are expressed as a ratio of the sill of the variogram, and the interval length as a ratio of the range of the variogram (see Appendix B). The kriging variances are based on 20 observations used in kriging, 10 on each side of the prediction-point in time. When fewer data are used, the kriging variance will be slightly larger. The larger the nugget-to-sill ratio, the larger the mean and maximum kriging variance. Also, given the range of the variogram, the mean and maximum kriging variance increase with the interval length. The effect of the interval length is largest for small nugget-to-sill ratios. Setting a limit to the maximum kriging variance requires smaller interval lengths (higher sampling frequencies) than setting

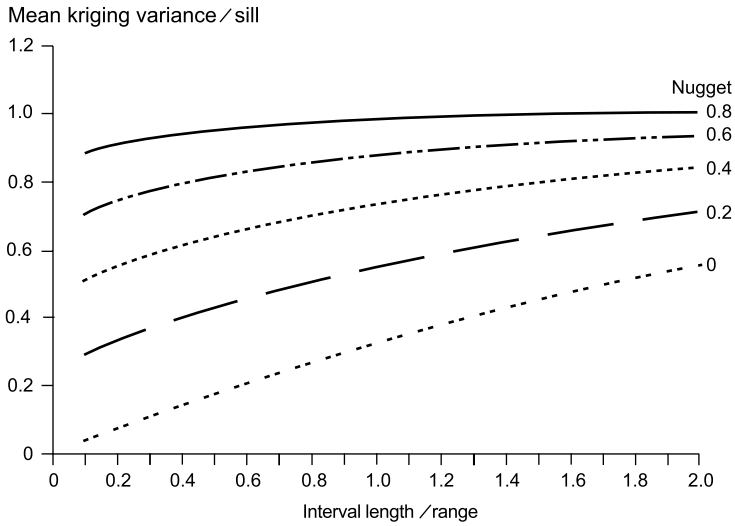


Fig. 12.1. Ratio of the mean kriging variance to the sill as a function of the ratio of the interval length to the range.

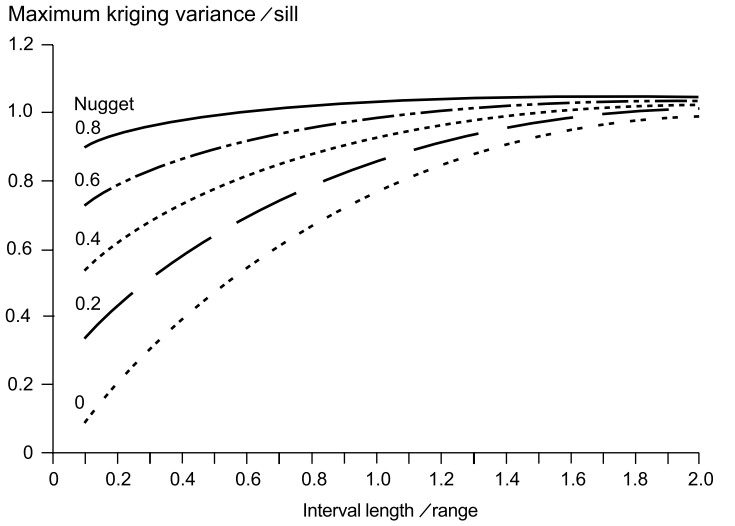


Fig. 12.2. Ratio of the maximum kriging variance to the sill as a function of the ratio of the interval length to the range.

the same limit to the mean kriging variance, especially for small nugget-to-sill ratios.

13

Time-Series Models

13.1 Introduction

Appendix C recapitulates the most commonly assumed linear stochastic processes. In this chapter the sampling aspects of time-series modelling are dealt with. Section 13.2 focuses on the sampling aspects of estimating process parameters which characterize the dynamic behaviour of the process. In Sect. 13.3 the estimation of model means, such as annual model means, is elaborated upon. Finally, Sect. 13.4 discusses the sampling aspects in the perspective of detecting trends in time series.

13.2 Estimation of Process Parameters

This section focuses on the development of monitoring strategies to obtain parameters which give a *general* description of the temporal behaviour of a physical phenomenon, i.e., status monitoring (Sect. 1.1). Statistical inference is made by assuming that the temporal behaviour results from a stochastic process which can be specified by a time-series model. Therefore, the main question to be answered in this section is: *How should samples be distributed in time, in order to derive time-series models that describe the dynamic behaviour of a physical phenomenon adequately?* The answer to this question starts with making assumptions about the stochastic processes, the most applied of which are recapitulated in Appendix C. As is mentioned in Sect. C.1, stochastic processes are data-based. The discrete-time stochastic processes described in Sects. C.2 to C.4 reflect the dynamic behaviour of physical phenomena, given the sampling frequency and the length of the monitoring period. Decisions must be taken on the sampling frequency and the length of the monitoring period. These decisions are discussed in this section.

We restrict ourselves here to discrete-time stochastic processes with discrete, constant time-intervals . Statistical inference for this type of processes is possible by using relatively simple mathematical methods, as compared with continuous-time processes. Box and Jenkins (1976) give a procedure of model identification, fitting (calibration) and diagnostic checking (verification). In Fig. 13.1 this procedure is summarized and extended for decisions on the length of the monitoring period and the sampling frequency.

Important tools in both model identification and diagnostic checking are the sample autocorrelation function (sample ACF), the sample partial autocorrelation function (sample PACF) and the residual cross-correlation function (residual CCF), see Appendix C. As an alternative to the identification procedure, automatic model selection can be applied, using a selection criterion (e.g., Akaike's Information Criterion, AIC, or Bayes Information Criterion, BIC). The procedures of either model identification or automatic model selection, fitting and diagnostic checking result in a model that describes the *data* adequately. The next step is to analyze the extent to which *the underlying physical processes* are described adequately by the model. This can be done by physical interpretation of the modelling results. Besides this, whenever possible validation is advised, which means that the model performance is tested using independent validation data. Both physical interpretation and validation results may not only give rise to further model improvements, but also to extension of the monitoring period, and adjustment of the sampling frequency. This is illustrated by the following two situations, which are given without aiming to be complete:

1. A large value of the autoregressive parameter of a first-order autoregressive model (Sect. C.2.1) is found. The validation results show large systematic errors. In this case the monitoring period may not fully cover the correlation length, i.e., the time lag at which the autocorrelation is (approximately) zero. In the case of a dynamic relationship between two variables, the monitoring period may not cover the response time. The monitoring should be continued at least until the correlation length or the response time is completely covered;
2. Although autoregressive relationships were expected on the basis of physical insights, no significant autoregressive relationships were found. The validation results show large random errors. In this case the sampling interval may be larger than the correlation length or, in the case of a dynamic relationship between two variables, than the response time.

13.3 Sampling Frequency for Estimation of Model Means

As is mentioned in Sect. 1.1 in status monitoring the status of a system is characterized and followed in time. A general statistic of a system is the mean, for instance the annual mean. If monitoring is restricted to systematic sampling

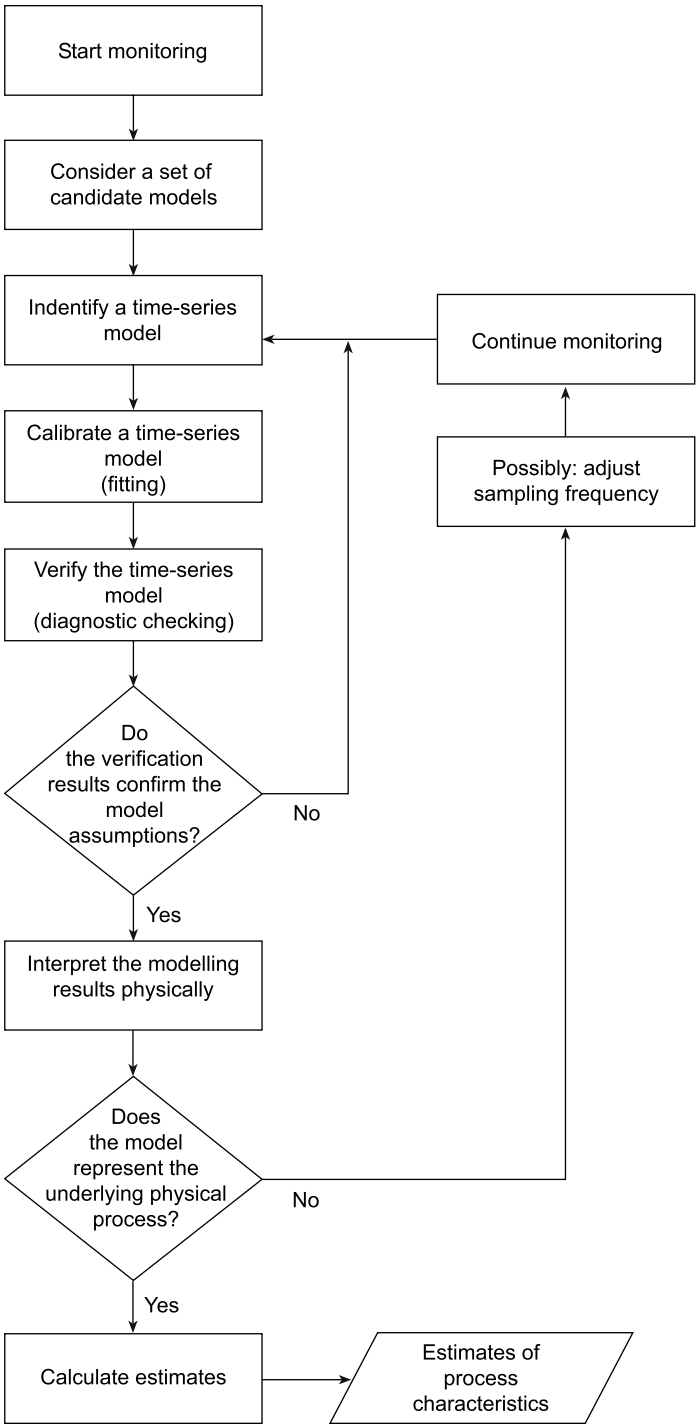


Fig. 13.1. Flow chart for the procedure of time-series model identification, fitting and diagnostic checking, extended with sampling aspects

in time, a sampling frequency needs to be chosen which enables estimates of means which are sufficiently accurate given the purposes of monitoring.

An observed series $\{z_i\}, i = 1 \ldots n$ is considered to be a realization of a second-order stationary stochastic process with a random fluctuation ϵ_i with variance σ_ϵ^2 around a deterministic mean μ:

$$Z_i = \mu + \epsilon_i \; . \tag{13.1}$$

Note that μ is a model parameter, and not the average of z over the universe of interest. Sanders and Adrian (1978) presented a sampling frequency criterion based on the relationship between sampling frequency and the magnitude of half the confidence interval of the model mean. Their approach will be followed here.

Suppose that n second-order stationary, independent and identically distributed observations on z are available. To obey the stationarity assumption, it may be necessary to remove seasonal nonstationarity first, for instance by fitting a deterministic annual cycle to the data and to use the residuals in further analysis. The variance σ_ϵ^2 can be estimated by:

$$\widehat{\sigma_\epsilon^2} = \frac{1}{n-1} \sum_{i=1}^{n} (z_i - \bar{z})^2 \, , \tag{13.2}$$

where $\bar{z}$ is the calculated mean of z_i, $i = 1, \ldots, m$. Confidence intervals for estimates of μ are estimated using the Student's t statistic:

$$t = \frac{\bar{z} - \mu}{\sigma_\epsilon / \sqrt{n}} \; . \tag{13.3}$$

The probability that t is within the confidence interval is given by the confidence level $1 - \alpha$:

$$\Pr\left(t_{\alpha/2} < \frac{\bar{z} - \mu}{\sigma_\epsilon / \sqrt{n}} \right) \; , \tag{13.4}$$

where $t_{\alpha/2}$ and $t_{1-\alpha/2}$ are constants from the Student's t distribution for a corresponding number of observations and confidence level. Since $t_{1-\alpha/2} = -t_{\alpha/2}$, the confidence interval of the model mean μ is given by

$$\bar{z} - \frac{t_{\alpha/2}\, \sigma_\epsilon}{\sqrt{n}} < \mu < \bar{z} + \frac{t_{\alpha/2}\, \sigma_\epsilon}{\sqrt{n}} \; . \tag{13.5}$$

If σ_ϵ has been estimated from prior information, and a decision on the confidence level $1 - \alpha$ has been taken, the width of the confidence interval can be plotted against values of n. If annual means are considered, n is the yearly number of observations.

Until now it was assumed that the values of z are mutually independent. However, the presence of autocorrelation is often indicated in time series of environmental variables. Dependency can be dealt with in two ways: (a) it is avoided, or (b) it is accounted for.

Avoiding dependency

To avoid dependency, the interval length should be larger than the length of serial correlation. Thus, prior information is needed on the length of serial correlation, but this is not always available. Alternatively, one could start to measure with a high frequency to estimate the correlation length, and then to continue at intervals longer than the estimated correlation length. Another possibility is to take average values over sufficiently long periods. For instance, if daily values appear to be correlated, one could calculate a series of monthly averages which may possibly be uncorrelated.

The correlation length can be estimated from observed time series using (C.8), provided that the interval length is shorter than the correlation length and the series amply covers the correlation length. Once a sampling frequency is found which provides serially uncorrelated observations on a target variable z_t, the maximum sampling frequency is known. Equation (13.5) is applied to investigate lower frequencies.

Accounting for dependency

A first way to account for serial correlation in estimating model means μ is by using the relationship between the actual number of observations and the equivalent number of independent observations for an autocorrelated time series, given by Bayley and Hammersley (1946):

$$V(\hat{\mu}) = \frac{\sigma_\epsilon^2}{n_\mathrm{b}^*} \,, \tag{13.6}$$

where n_b^* is the equivalent number of independent observations, $\hat{\mu}$ is the estimated mean of a process $\{Z_t\}$ and σ^2 is the variance for the random fluctuation ϵ_i in (13.1). For second-order stationary stochastic processes with n observations, n_b^* can be calculated by

$$\frac{1}{n_\mathrm{b}^*} = \frac{1}{n} + \frac{2}{n^2} \sum_{j=1}^{n-1} (n-j)\, \rho_{j\Delta t} \,, \tag{13.7}$$

where Δt is the observation interval and $\rho(j\Delta t)$ is the correlation coefficient for lag $j\Delta t$. For first-order autoregressive processes (AR(1), see Sect. C.2.1), (13.7) reduces to

$$\frac{1}{n_\mathrm{b}^*} = \frac{1}{n} + \frac{2}{n^2} \cdot \frac{\rho^{(n+1)\,\Delta t} - n\,\rho^{2\,\Delta t} + (n-1)\,\rho^{\Delta t}}{(\rho^{\Delta t} - 1)^2} \,, \tag{13.8}$$

(Matalas and Langbein, 1962), where ρ is the lag 1 correlation coefficient for a selected base lag period. If the base lag period equals the observation interval the AR(1) process is given by

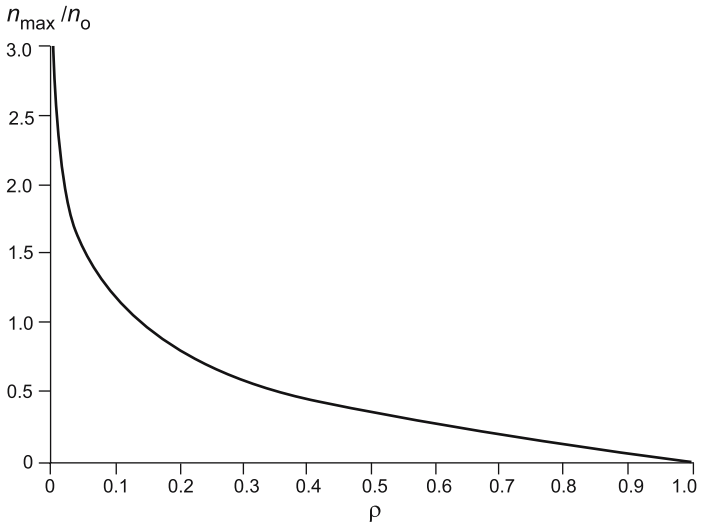


Fig. 13.2. Maximum equivalent number of independent observations as a function of the daily lag 1 correlation coefficient ρ for a Markov process; $n_o = 365$ days. (Reproduced from Lettenmaier (1976, Fig. 6) with permission of American Geophysical Union.)

$$Z_i - \mu = \phi\,(Z_{i-1} - \mu) + \epsilon_i\ , \tag{13.9}$$

and $\rho^{\Delta t}$ in (13.8) can be replaced by ϕ. The width of the confidence interval can be estimated by (13.5), replacing n by n_b^*.

Within a specified period n_o a maximum number of equivalent independent observations may be collected:

$$n_{max} = \frac{n_o}{2} \cdot \frac{(\ln \rho)^2}{\rho^{n_o} - n_o \ln \rho - 1}\ , \tag{13.10}$$

(Lettenmaier, 1976), where n_o is the specified period, for example 365 days, and ρ is the daily lag 1 correlation coefficient. Thus, if the number of observations n goes to infinity within the period n_o, the equivalent number of independent observations will not exceed a certain n_{max}.

In Fig. 13.2 n_{max} is given as a function of the lag 1 correlation coefficient ρ. Figure 13.2 shows that n_{max} decreases with an increasing lag 1 correlation coefficient. If the number of actual samples were infinite and ρ were 0.14, then from (13.10) and Fig. 13.2 it follows that the equivalent number of independent observations equals 365.

Figure 13.3 gives the ratio n_b^*/n_{max} as a function of the normalized sampling frequency n/n_o, for various values of the daily lag 1 correlation coefficient ρ. It can be seen that n_{max} is approached quite rapidly for large values of ρ. If observations are taken weekly ($n/n_o = 0.14$) and $\rho = 0.85$, then from Fig. 13.3

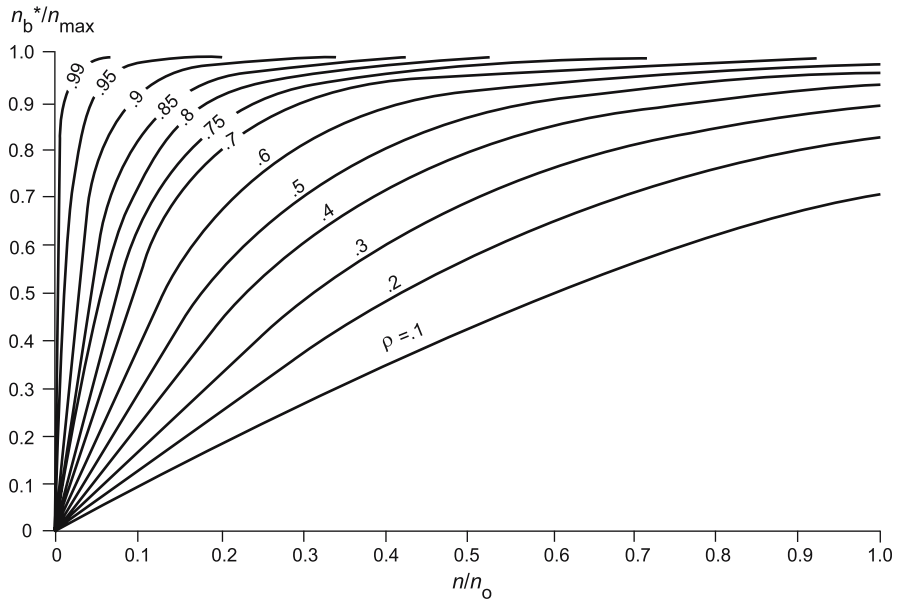


Fig. 13.3. Ratio of equivalent and maximum equivalent number of independent observations, as a function of the normalized sampling frequency; $n/n_o = 1$ corresponds to daily sampling. (Reproduced from Lettenmaier (1976, Fig. 7) with permission of American Geophysical Union.)

follows that n_b^*/n_{max} equals 0.9. Thus, 90 % of the information that may be collected in a period of 365 days is already provided by weekly sampling.

13.4 Sampling Frequency for Detecting Trends

In the previous section second-order stationarity of Z in (13.1) was assumed, which implies a time invariant model mean μ. This section focuses on the detection of temporal changes in the model mean μ. In Sect. 13.4.1 sampling aspects of tests for step trends and linear trends are discussed. Section 13.4.2 deals with sampling aspects of intervention analysis. Note that the methods described here are based on model means, in contrast to the method for trend testing described in Sect. 11.2 which is based on temporal means.

13.4.1 Tests for Step Trends and Linear Trends

If temporal sampling is restricted to systematic sampling, then the choice of an appropriate sampling design reduces to choosing an appropriate sampling frequency. Analogous to Sect. 13.3, serial correlation in regularly spaced series is either prevented for or accounted for.

Avoiding dependency

In Sect. 13.3 it was discussed how the sampling frequency can be found below which observations are practically serially uncorrelated. Once a sampling frequency is found which provides serially uncorrelated observations on a target variable Z, the minimum sample size and thus the minimum length of the monitoring period must be decided on. Lettenmaier (1976) gives criteria for the minimum sample size if trend detection is the purpose of the monitoring. The trend can concern either a step trend or a linear trend. Lettenmaier (1976) considers trend detection as a testing problem, with H_0: a trend is not present in the underlying process and H_1: a trend is present. Following Lettenmaier (1976), it is explained below how the minimum length of series can be determined for testing on step trends and linear trends, respectively.

Sampling Frequency for Detecting Step Trends

A step trend is defined here as a sudden change in the mean level of a process. Suppose that this sudden change occurs halfway a series with an even number of n measurements. Furthermore, let μ_1 and μ_2 be the mean levels of the series before and after the abrupt change. The underlying process is now defined as

$$Z_i = \begin{cases} \mu_1 + \epsilon_i & \text{if } i \le n/2 \\ \mu_2 + \epsilon_i & \text{if } i > n/2 \,, \end{cases} \tag{13.11}$$

where $i = 1, \ldots, n$ indicates the ith element of an equidistant series of length n. In (13.11) the ϵ's are independent identically distributed random variables with zero mean and variance σ_ϵ^2.

The test chooses between

$$\mathrm{H}_0 : \mu_1 = \mu_2$$
$$\mathrm{H}_1 : \mu_1 \neq \mu 2 \,.$$

The test statistic is

$$T = \frac{|\bar{z}_1 - \bar{z}_2|\sqrt{n}}{2\,\hat{\sigma}_\epsilon} - t_{1-\alpha/2,\,\nu} \,, \tag{13.12}$$

with

$$\bar{z}_1 = \frac{1}{n/2} \sum_{i=1}^{n/2} z_i, \quad \bar{z}_2 = \frac{1}{n/2} \sum_{i=n/2+1}^{n} z_i \,, \tag{13.13}$$

being estimators for μ_1 and μ_2, respectively, $t_{1-\alpha/2,\nu}$ is the quantile of the Student's t distribution at confidence level $1-\alpha/2$ and for $\nu = n-2$ degrees of freedom, and with $\hat{\sigma}_\epsilon$ being the sample standard deviation,

$$\hat{\sigma}_\epsilon^2 = \frac{1}{n-2} \left(\sum_{i=1}^{n/2} (z_i - \bar{z}_1)^2 + \sum_{i=n/2+1}^{n} (z_i - \bar{z}_2)^2 \right) \,. \tag{13.14}$$

If $T \leq 0$ then H_0 is accepted, otherwise H_0 is rejected.

Now, for the purpose of sample design, the absolute value of the true difference between μ_1 and μ_2 is assumed to be known: $T_r = |\mu_1 - \mu_2|$, as well as the variance σ_ϵ^2. Then, the following population statistic can be formed:

$$N_T = \frac{\sqrt{n}}{2\,\sigma_\epsilon} T_r \; . \tag{13.15}$$

The power of the test is now given by

$$1 - \beta = F(N_T - t_{1-\alpha/2,\,\nu}) \; , \tag{13.16}$$

where F is the cumulative distribution of a standard Student's t distribution with $\nu = n - 2$ degrees of freedom.

If prior information on the variance σ_ϵ^2 is available, a guess can be made of the minimum length of the series needed for the detection of a step trend which occurs halfway this series, for a given α and β. Figure 13.4 shows a diagram for the relationship between the normalized magnitude of the step trend (T_r/σ_ϵ) and the minimum length of the series needed for given values of β and with $\alpha = 0.05$. The step trend is assumed to occur halfway the series, and $n = n_1 + n_2$, where $n_1 = n_2$ are the numbers of observations before and after the step change, respectively. The effect on the sample size of a decision for a lower confidence level $1 - \alpha$ is illustrated by Fig. 13.5, where $\alpha = 0.10$.

Minimum Sample Size for Detecting Linear Trends

According to Lettenmaier (1976, p. 1038), a linear trend is parameterized as

$$Z_i = \epsilon_i + i\,\tau + \gamma \; , \tag{13.17}$$

where ϵ_i is a series of random disturbances from a normal distribution with mean zero and variance σ_ϵ^2, τ is the trend magnitude, and γ is the process base level. The parameter τ is estimated by

$$\hat{\tau} = \frac{\sum\limits_{i=1}^{n} i'\, z_i'}{\sum\limits_{i=1}^{n} i'^2} \; , \tag{13.18}$$

where

$$i' = i - \frac{n+1}{2}, \qquad z_i' = z_i - \frac{1}{n}\sum_{i=1}^{n} z_i \; . \tag{13.19}$$

The estimate $\hat{\tau}$ has variance

$$V(\hat{\tau}) = \frac{\sigma_\epsilon^2}{\sum\limits_{i=1}^{n} i'^2} \; . \tag{13.20}$$

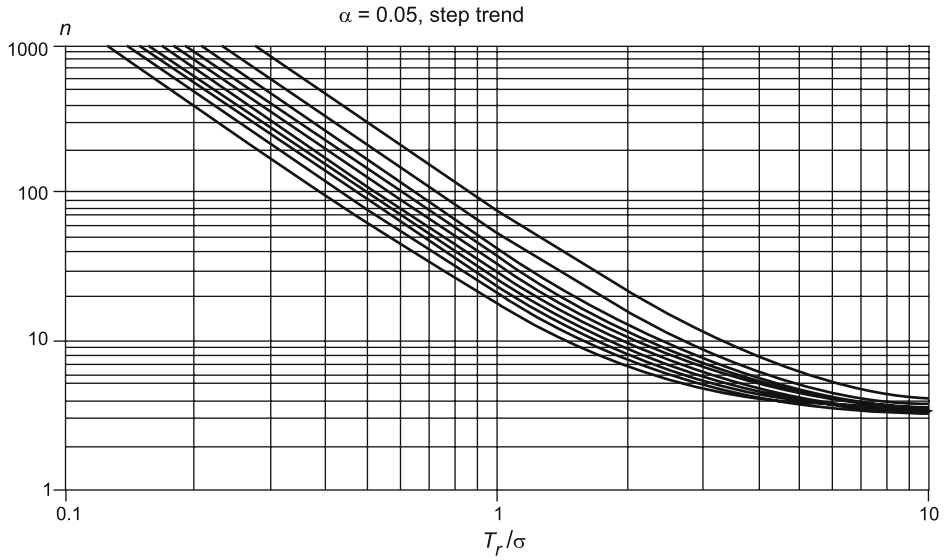


Fig. 13.4. Sample size against normalized step trend for $\beta = 0.5$, 0.4, 0.35, 0.3, 0.25, 0.2, 0.15, 0.1, 0.05, 0.01 (from left to right) and $\alpha = 0.05$. T_r: the magnitude of the step trend, σ: the standard deviation, $n = n_1 + n_2$, with $n_1 = n_2$: the number of equidistant observations before and after the step change, respectively.

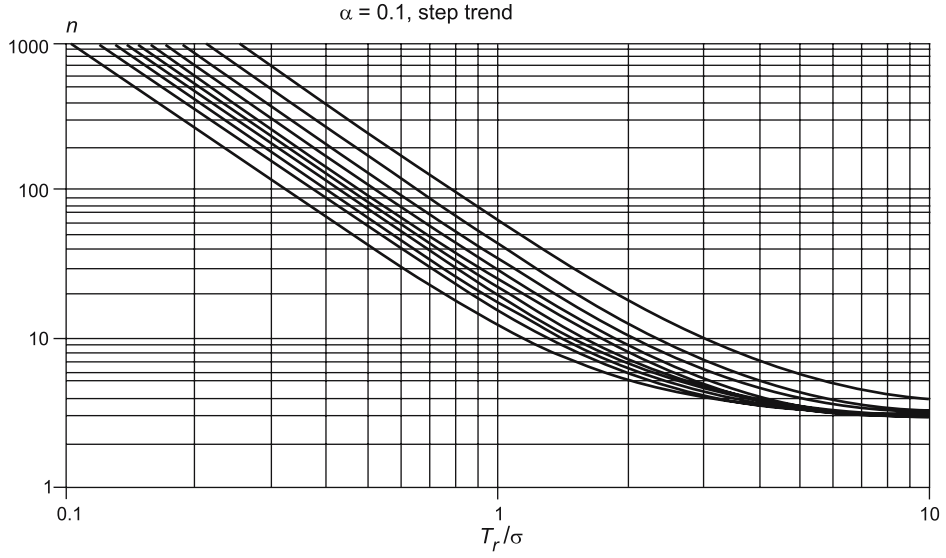


Fig. 13.5. Sample size against normalized step trend for given values of β, and $\alpha = 0.1$. See Fig. 13.4

Assuming normality of $\hat{\tau}$, a test statistic T is given by

$$T = |\hat{\tau}| - \frac{t_{1-\alpha/2,\,\nu} \cdot \hat{\sigma}_\epsilon}{\sqrt{\sum_{i=1}^{n} i'^2}} \,, \tag{13.21}$$

with $\nu = n - 2$ degrees of freedom. The sample estimate $\hat{\sigma}_\epsilon$ is calculated by

$$\hat{\sigma}_\epsilon = \frac{1}{n-2} \sum_{i=1}^{n} (z_i - i\,\hat{\tau} - \hat{\gamma})^2 \,, \tag{13.22}$$

and the sample estimate $\hat{\gamma}$ is calculated by

$$\hat{\gamma} = \frac{1}{n} \sum_{i=1}^{n} z_i - \hat{\tau} \frac{n+1}{2} \,. \tag{13.23}$$

The test statistic T in (13.21) can be normalized as follows:

$$T' = \frac{|\hat{\tau}| \left(\sum_{i=1}^{n} i^2\right)^{1/2}}{\hat{\sigma}_\epsilon} - t_{1-\alpha/2,\,\nu} \,. \tag{13.24}$$

Given the identity

$$\sum_{i=1}^{n} i^2 = \frac{1}{6} n\,(n+1)\,(2n+1) \,, \tag{13.25}$$

Lettenmaier (1976) derives the following dimensionless statistic for the existence of a linear trend, assuming that the population trend magnitude τ is known:

$$N'_T = \frac{\{n\,(n+1)\,(n-1)\}^{1/2}\,\tau}{\sqrt{12}\sigma_\epsilon} \,. \tag{13.26}$$

If $n\,\tau$ is substituted by T'_{r}, then N'_T becomes

$$N'_T = \frac{\{n\,(n+1)\,(n-1)\}^{1/2}\,T'_{\mathrm{r}}}{n\sqrt{12}\sigma_\epsilon} \,. \tag{13.27}$$

The power of the test for linear trend can be calculated by (13.16), with N_T replaced by N'_T. Figures 13.6 and 13.7 give minimum series lengths needed for given values of β and for $\alpha = 0.05$ and 0.1, respectively, to detect linear trends with normalized magnitudes T'/σ_ϵ.

Lettenmaier (1976) gives power curves for nonparametric trend tests. These tests are appropriate if the assumptions of parametric tests are violated. Lettenmaier (1976) shows that Mann–Witney's test and Spearman's rho test are adequate for testing against a step trend and a linear trend, respectively.

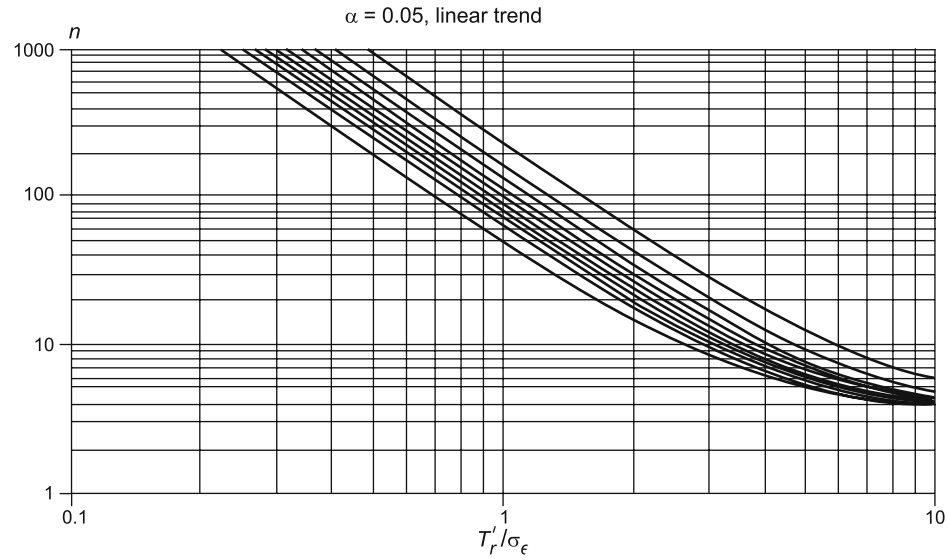


Fig. 13.6. Sample size against normalized linear trend for given values of β (See Fig. 13.4), and $\alpha = 0.05$. T_{r}': magnitude of the linear trend, σ_ϵ: the residual standard deviation of the linear trend model, n: the number of equidistant observations

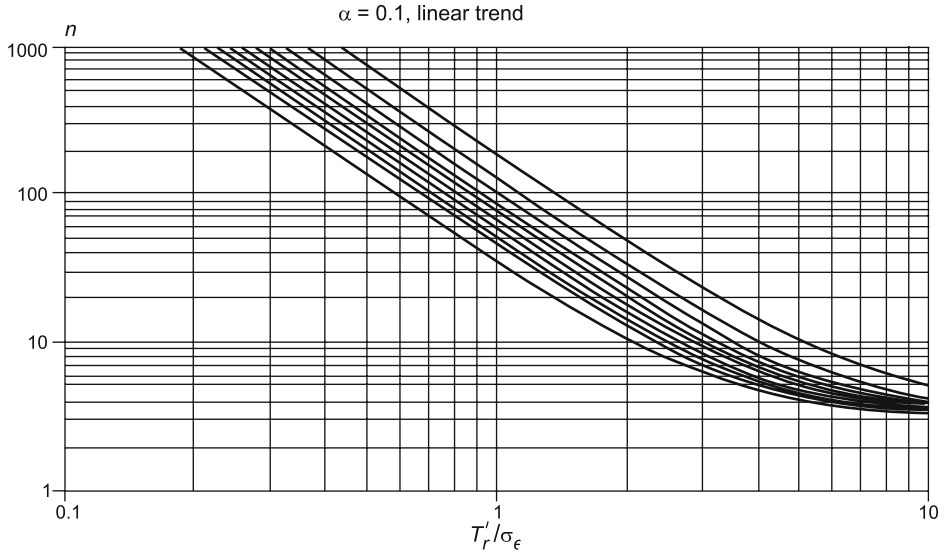


Fig. 13.7. Sample size against normalized linear trend for given values of β (See Fig. 13.4), and $\alpha = 0.1$. T_{r}': magnitude of the linear trend, σ_ϵ: the residual standard deviation of the linear trend model, n: the number of equidistant observations

Accounting for dependency

The trend tests described above are based on the assumption that the observations are mutually independent. This is the case if the interval length is larger than the correlation length. However, in many cases the assumption of mutual independence is not very useful. The required minimum interval length implies extensive monitoring periods in order to obtain a sufficiently large number of observations to execute powerful tests. Furthermore, if serial correlation is removed by extending the interval length or by calculating averages over sufficiently long periods, information gets lost. Therefore, it may be attractive to use methods for trend detection that account for serial correlation.

Serial correlation can be accounted for in trend tests by using the relationship between the actual number of observations and the equivalent number of independent observations for an autocorrelated time series, given in (13.6) to (13.8). Tests for step trends or linear trends in autocorrelated series can be performed using the equations given before for independent series, replacing n by n_{b}^{*}.

Within a specified period n_{o} a maximum number of equivalent independent observations may be collected, see (13.10). This is important in deciding on the sampling frequency and the length of the observation period: the power of trend tests may increase more by observing longer than by observing more frequently. Lettenmaier (1976) made an extension of nonparametric trend tests to dependent time series, and gives diagrams of maximum power and power to maximum power ratio for Mann–Whitney's test against a step trend. Similar diagrams for parametric t-tests are given in Figs. 13.8 and 13.9. As compared to the diagrams for Mann–Whitney's test against step trend, the diagrams for t-tests given in Fig. 13.8 indicate that t-tests have smaller maximum power for given daily lag 1 correlation coefficient and trend to standard deviation ratio. For the application of nonparametric trend tests to hydrological time series we refer to Hirsch et al. (1982), van Belle and Hughes (1984), Hirsch and Slack (1984) and Yue et al. (2002).

Lettenmaier (1978) compared the statistical power of trend tests for uniformly collected time series (i.e., equidistant observation times) and for 'stratified' or unequally spaced time series, resulting from rotational monitoring designs (Sect. 14.1). In the latter case observations are taken, for instance, during one year in three, which may be more travel-economical than collecting equidistant time series in a monitoring network. It was concluded that equidistant time series are preferred over unequally spaced time series in trend detection. In the rotational design 2–3 times as many samples need to be taken to achieve the same power as in trend tests for equidistant time series.

13.4.2 Intervention Analysis

Methods accounting for serial correlation include the intervention models, described by Hipel et al. (1975) and Hipel and McLeod (1994). Intervention

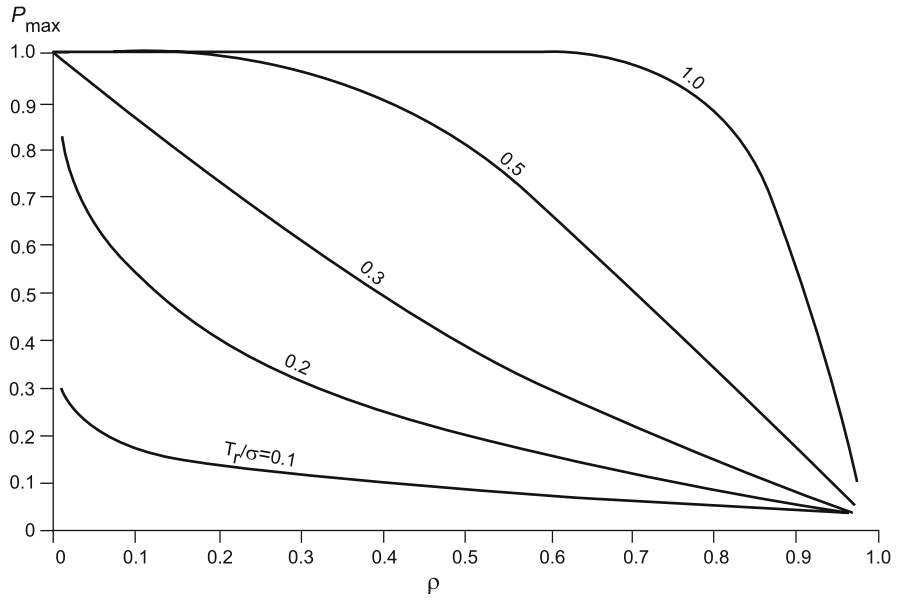


Fig. 13.8. Maximum power of a t-test against a step trend, as a function of the daily lag 1 correlation coefficient ρ and the trend to standard deviation ratio T_r/σ for a Markov process. Analogous to Lettenmaier (1976)

models form a special class of transfer function-noise models, see Sect. C.4. The intervention model for a step trend is given by

$$Z_t = I_t + N_t \,, \tag{13.28}$$

where $t = 1, \ldots, n$ indicates the t-th element of a series of length n, Z_t is the process of interest, I_t is the trend component and N_t is a noise component describing the part of Z_t that cannot be explained from the trend. The noise component is usually taken as an ARMA model, see (C.23). The trend component I_t is a transfer function with the following general form:

$$I_t = \delta_1 I_{t-1} + \delta_2 I_{t-2} + \cdots + \delta_r I_{t-r} + \omega_0 S^{(T)}_{t-b} - \omega_1 S^{(T)}_{t-1-b} - \cdots - \omega_m S^{(T)}_{t-m-b} \,, \tag{13.29}$$

where $\delta_1 \ldots \delta_r$ are autoregressive parameters up to order r, $\omega_0 \ldots \omega_m$ are moving average parameters up to order m, b is a pure delay parameter. Using the backward shift operator B, (13.29) can be written as

$$I_t = \frac{\omega(B)}{\delta(B)} B^b S^{(T)}_t \,, \tag{13.30}$$

with $B^k z_t = z_{t-k}$ and k is a positive integer.

$S^{(T)}_t$ is an input series indicating the step intervention:

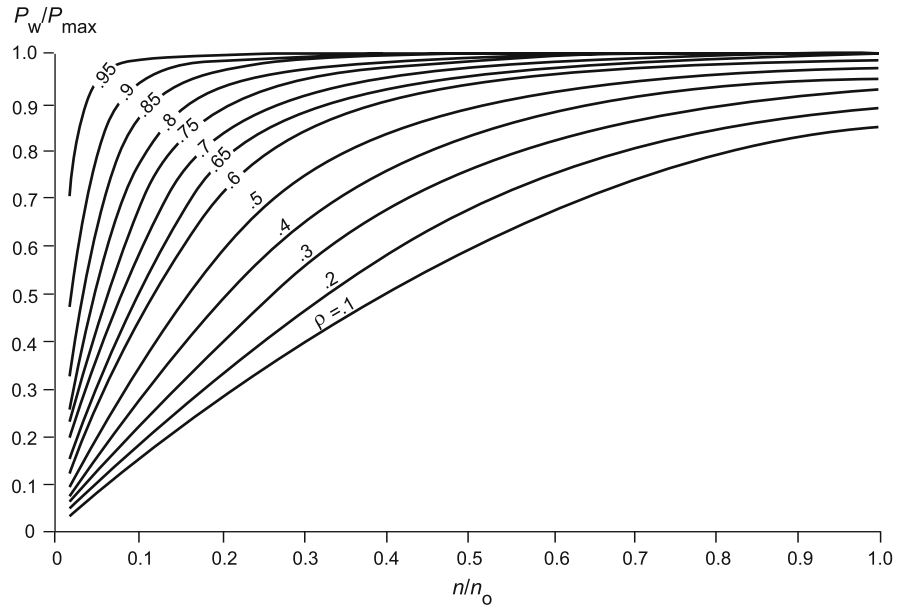


Fig. 13.9. Power to maximum power ratio as a function of relative sampling frequency for t-tests against a step trend in time series following a Markov process. ρ is the daily lag 1 correlation coefficient. $n_o = 365$. Analogous to Lettenmaier (1976).

$$\begin{aligned} S_t^{(T)} &= 0 \quad \text{if } t < T, \\ S_t^{(T)} &= 1 \quad \text{if } t \geq T\,. \end{aligned} \tag{13.31}$$

Step interventions influence processes in different ways, which can be expressed by different forms of the transfer function, see Fig. 13.10. As compared to the testing procedures described before, intervention modelling has the advantage that the effect of interventions can be separated from other independent influences. The model in (13.28) can be extended with other transfer components besides the intervention:

$$Z_t = I_t + X_{i,t} + N_t, \;\; \imath = 1, \ldots, m\,, \tag{13.32}$$

where $X_{i,t}$, $i = 1, \ldots, m$ are m transfer components of m independent inputs. Lettenmaier et al. (1978) discussed the sampling aspects of intervention analysis. They considered intervention analysis as a hypothesis test with H_0: no intervention has taken place, which means that $\omega B/\delta B = 0$ in (13.30). Based on knowledge of the covariance matrix of the model parameters, they constructed power functions for several specified intervention models. These models do not include models containing other inputs besides the intervention, as in (13.32). Nevertheless, the power curves presented by Lettenmaier et al. (1978) provide an indication of required sample sizes and the ratio of

the number of observations collected before and after the intervention took place. In summary, their conclusions on the sampling aspects of intervention analysis are:

1. For the step decay model, the linear model and the impulse decay model (Figs. 13.10**c**, **d** and **e**, respectively), relatively small pre-intervention series lengths are required. For the step model (Fig. 13.10**a**, **b**) it is indicated that equal pre- and post-intervention series lengths are optimal;
2. For the impulse decay model, it is important that data are collected frequently during the period that the intervention response is non-constant;
3. It is indicated that the minimum detectable intervention effect depends on the complexity of the intervention model: more complex models require a larger number of observations. Let ω be the intervention response magnitude. Furthermore, let γ be the pre-intervention series length relative to the total series length. For a step model and an impulse decay model ω equals ω_0. For a step decay model ω equals $\omega_0/(1-\delta_1)$. For a linear model $\omega = \mathrm{m}\omega_0/(1-\gamma)$ where m is the number of post-intervention observations. It is indicated that a minimum level of change, relative to the process standard deviation, ω/σ_Z, of about 0.5 can be detected for the step model. For the linear model the minimum level of ω/σ_Z that can be detected is at about 0.75, and 1.0 for the impulse decay model. For the step decay model a higher minimum level is indicated. Below these values intervention effects cannot be detected with reasonable sample sizes.

Additional to these conclusions Lettenmaier et al. (1978) suggest:

1. If the fitted parameter values for a hypothesized intervention model are not significant, a simpler model should be fitted (e.g., a step model instead of a step–decay model);
2. Seasonality should be removed from the data. However, seasonal differencing will lead to substantial loss of power. Therefore alternative methods of deseasonalization are recommended (for instance differencing against a control time series, or removing seasonal means);
3. If the process variance varies seasonally, homoscedasticity (i.e., constant variance) should be achieved by an appropriate Box–Cox transformation;
4. Data should be collected uniformly, i.e., at constant time-intervals. However, Kalman filter methods for modelling irregularly spaced time series, as proposed by Bierkens et al. (1999), could be extended to intervention analysis.

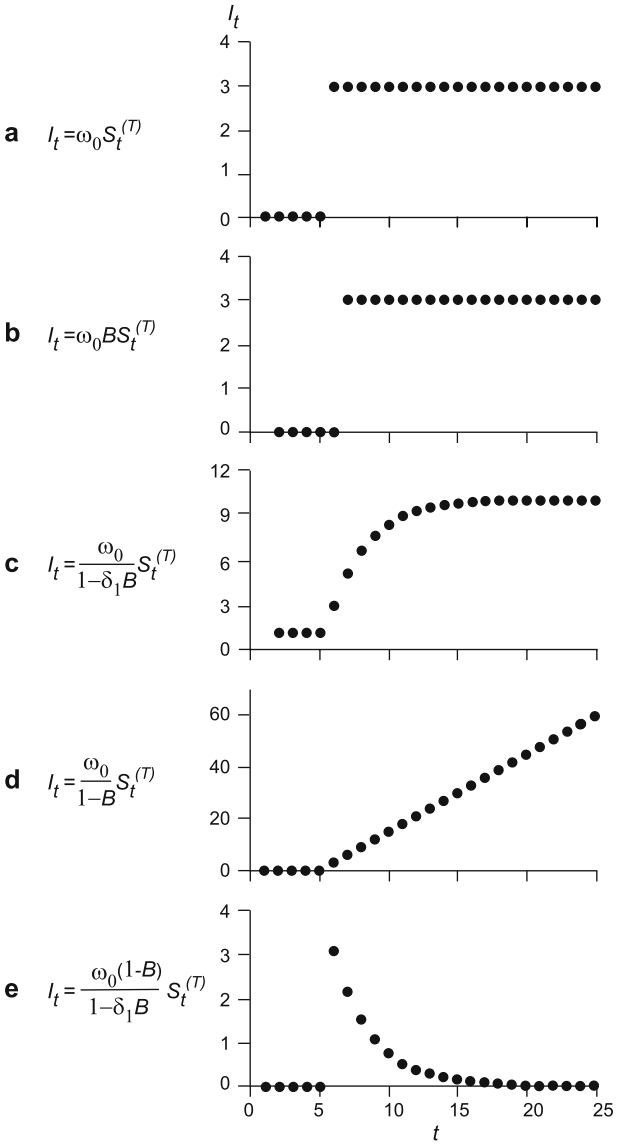


Fig. 13.10. Responses to a step intervention. $T = 5$, $S_t^{(T)} = 0$ for $t < T$, $S_t^{(T)} = 1$ for $t \geq T$, $\omega_0 = 3$, $\delta_1 = 0.7$, $b = 1$. **a**: step model, **b**: delayed step model, **c**: step decay model, **d**: linear model, **e**: impulse decay model.

Part IV

SAMPLING IN SPACE–TIME

14

Introduction to Sampling in Space–Time

This part deals with monitoring of objects with a spatial extent. Examples are monitoring the quality of surface-water or the level of the groundwater in a watershed, monitoring the soil quality in an agricultural area, and monitoring the biodiversity in a nature conservation area. As will be shown, the sampling events can be distributed in many ways in the space–time universe. Monitoring implies that observations are repeated or continued in time. A survey, on the other hand, is defined as the collection of information on an object with a spatial extent through observation, such that possible changes of the object during the observation are negligible (Sect. 1.1). By keeping the observation period of a given survey as short as possible, temporal variation is eliminated, and the sampling error is related to spatial variation only. By repeating these surveys, information is obtained on the temporal variation of the target variable. In this part we neither give details on sampling designs for the separate surveys, nor on sampling designs for longitudinal studies at a single location. For this we refer to Parts II and III, respectively. This part focuses on the choice of sampling patterns in both space and time.

An important question is whether at all sampling times the same locations must be observed, or whether this restriction should be relaxed and all or part of the sampling locations is to be replaced by new locations. This issue is dealt with in Sects. 14.1 and 14.2. Also, for estimating a spatio-temporal mean or a temporal trend of the spatial mean, an important question is the optimal number of sampling locations in relation to the number of sampling times or the sampling frequency. This is discussed in the sections on these quantities.

14.1 Types of Sampling Pattern for Monitoring

The design types for probability sampling in *space* (Sect. 7.2) are derived from Simple Random Sampling, by imposing restrictions on the randomization in order to increase the efficiency. Similarly, pattern types in *space–time* can be conceived as derived from the irregular kind of pattern resulting from

Simple Random Sampling, by imposing regularity restrictions. In model-based sampling the patterns are deterministic, in design-based sampling they are selected at random within the restrictions of the chosen type of pattern.

Fully unrestricted patters in space–time are not further discussed here, because they are clearly inefficient in nearly all practical situations. The reason is that the number of sampling locations as well as sampling times would equal the sample size (i.e., the number of sampling events), which is uneconomical in view of the fixed costs of sampling locations and sampling times. Depending on logistics and techniques, the average costs of sampling events generally decreases when at a given sampling time more than one location is sampled, or when a given location is sampled more than once.

Two kinds of restrictions are imposed on space–time patterns to improve the efficiency: sampling multiple times at the same location (stationarity), and sampling at multiple locations at the same time (synchronicity). Four basic types of sampling pattern arise by imposing these restrictions: *static, synchronous, static-synchronous* and *rotational* patterns. In *Static Sampling* all sampling takes place at a fixed set of locations. See Fig. 14.1 for a notional example. If continuous measuring takes place at each location, then sampling is in fact only spatial and the methods of sampling in space (Part II) are applicable. Sampling at the various locations may or may not follow the same pattern in time.

In *Synchronous Sampling*, also referred to as repeated or dynamic sampling, a different set of sampling locations is selected for each sampling time, i.e., the sampling locations are not revisited. See Fig. 14.2 for a notional example. If one measures exhaustively in space at each sampling time, e.g., by remote sensing, then sampling is in fact only temporal and the methods of sampling in time (Part III) are applicable. The spatial patterns used at different times may or may nor be the same. If they are the same, then they do not coincide spatially, because otherwise the pattern would be static-synchronous.

An interesting special type of synchronous pattern is the *interpenetrating space–time grid*. In that case the same grid is used at each sampling time, however, as opposed to Space–Time Grid Sampling (see hereafter), the grids do not coincide spatially. Instead of this, the grids are shifted so that the grid points at a given sampling time are farthest from those at the previous time. See Fig. 14.3 for three notional examples.

When Static Sampling and Synchronous Sampling are combined with each other, we speak of *Static-Synchronous Sampling*. Figure 14.4 is an illustration of such a pattern, also referred to as a pure panel[1] (Fuller, 1999). If in addition the spatial pattern is a grid and temporal sampling is systematic, then this is referred to as *Space–Time Grid Sampling*.

[1] Note that the pattern of Fig. 14.4, although generated by Simple Random Sampling in both space and time, is not equivalent with Simple Random Sampling in space–time!

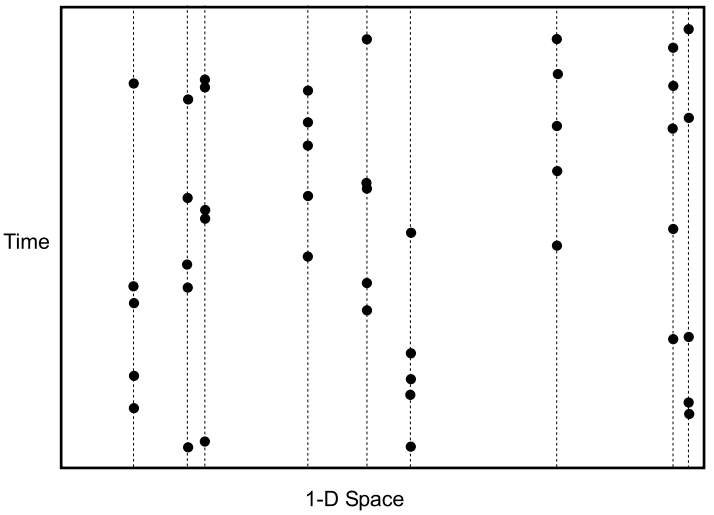


Fig. 14.1. Notional example of a static sample, with Simple Random Sampling in both space and time

Rotational Sampling is a compromise between Static Sampling and Synchronous Sampling, in the sense that the locations of the previous sampling time are partially replaced by new ones. See Fig. 14.5 for a notional example.

Synchronous Sampling implies that the locations are not resampled. With a variation of this type of pattern, after some fixed number of sampling times, no new sampling locations are selected anymore, but the existing locations are resampled in the same order (Fig. 14.6). Such a pattern is referred to as a serially alternating pattern (Urquhart and Kincaid, 1999). To distinguish it from other periodic sampling patterns, we refer to it as an r-period synchronous sampling pattern. Similarly, in Rotational Sampling one may decide at any sampling time to stop selecting new locations and to resample locations of previous sampling times (Fig. 14.7). Such a sampling pattern is referred to as an r-period rotational pattern or r-period rotating panel (Fuller, 1999).

The pattern types as described above may be combined. For instance, a static sample (pure panel) can be augmented with a synchronous sample or a rotation panel. Such samples are referred to as split panels or supplemented panels (Duncan and Kalton, 1987).

A term often used in connection with monitoring is 'monitoring network', defined by Loaiciga et al. (1992) as a fixed set of sampling locations and a sampling frequency. Depending on whether or not the locations are sampled simultaneously, this represents what we call a *static* or a *static-synchronous sampling pattern*.

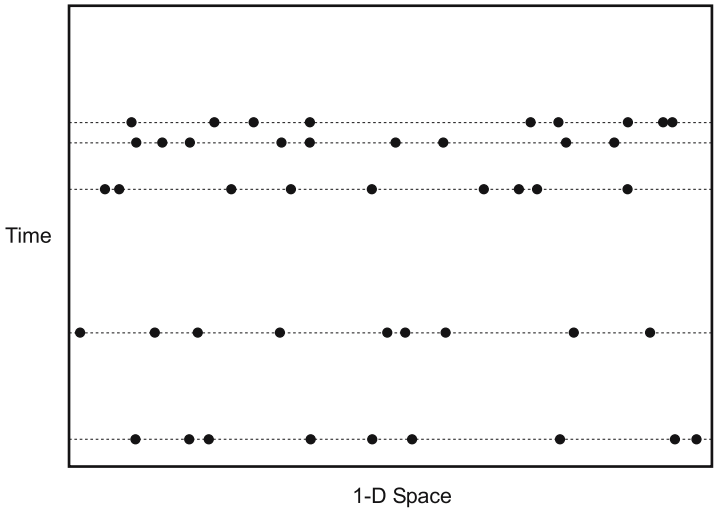


Fig. 14.2. Notional example of a synchronous sample, with Simple Random Sampling in both space and time

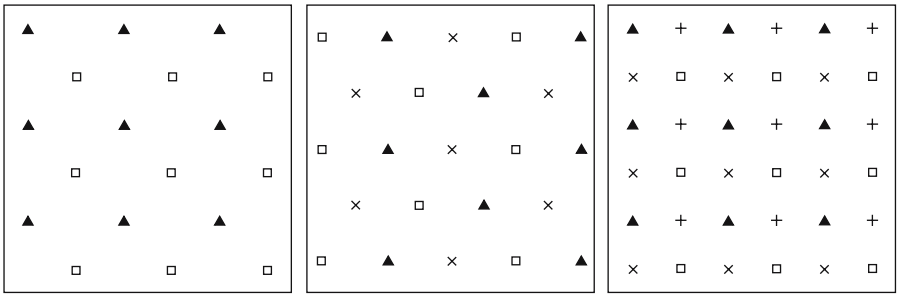


Fig. 14.3. Notional examples of interpenetrating space–time grids for two, three and four sampling times

14.2 Statistical and Operational Aspects

The choice of a type of sampling pattern for monitoring should be guided by operational as well as statistical considerations. Obviously, the static and the static-synchronous type of pattern have a financial advantage if the costs of repeated sampling at the same location are lower than for sampling at different locations with the same total sample size. This will be the case when retrieval of sampling locations in the field can be made easier by marking them, or when sampling, measuring or recording equipment is installed at fixed locations in the field on a semi-permanent basis. A statistical disadvantage of these patterns compared with synchronous and rotational ones is

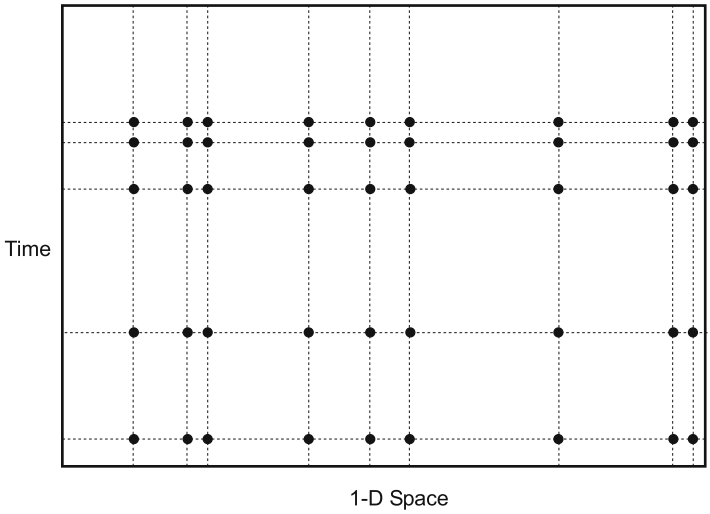


Fig. 14.4. Notional example of a static-synchronous sample, with Simple Random Sampling in both space and time

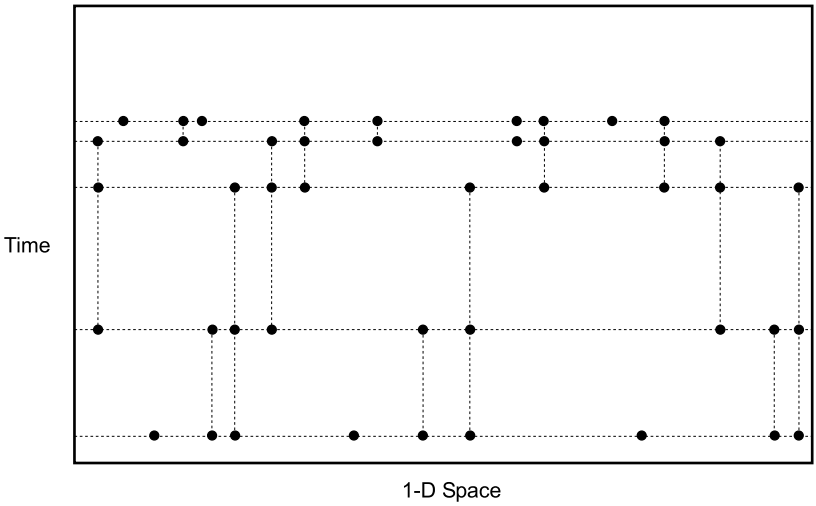


Fig. 14.5. Notional example of a rotational sample, with Simple Random Sampling in both space and time

that while monitoring goes on, only the information on temporal variation increases, not that on spatial variation. Also, in environmental and ecological studies, disturbances of the environment in previous rounds may cause bias in

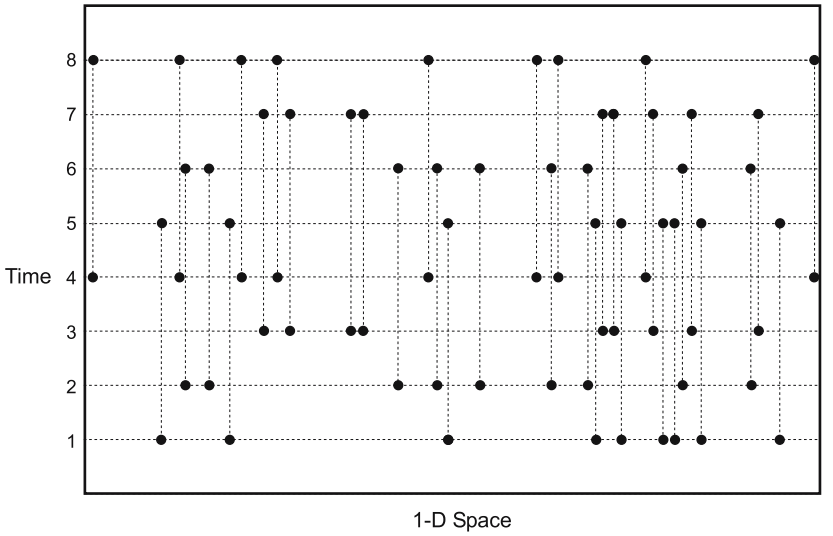


Fig. 14.6. Notional example of a 4-period synchronous sample, with Simple Random Sampling in space and Systematic Random Sampling in time

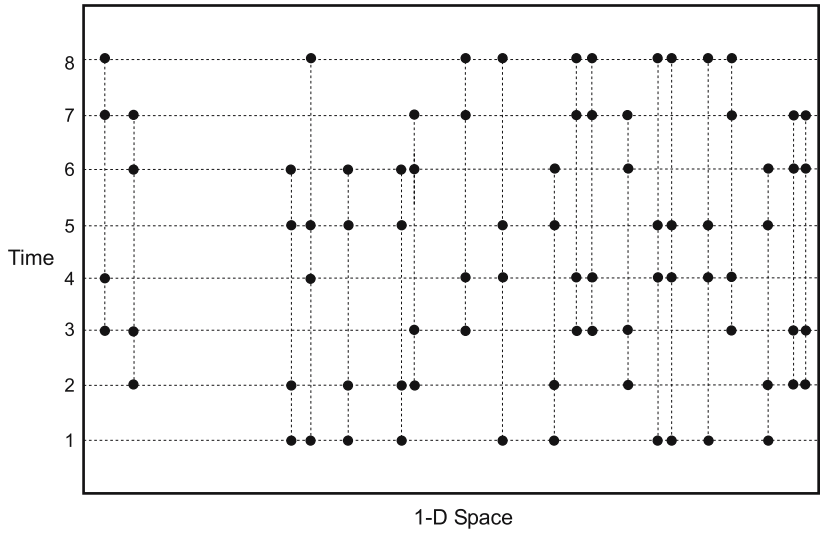


Fig. 14.7. Notional example of a 4-period rotational sample, with Simple Random Sampling in space and Systematic Random Sampling in time

estimated target quantities such as the estimated change of the spatial mean or total.

An advantage of static over static-synchronous patterns is that the selection of sampling times may be adapted to local circumstances, e.g., differences in costs of sampling or measuring, or differences in temporal variation. On the other hand, static-synchronous patterns may have an important advantage over static ones because, given the sample size, they reduce the number of sampling times. If the fixed costs of sampling times are high relative to the variable costs (i.e., the costs of sampling an additional location at the same sampling time), then reducing the number of sampling times enables more locations to be sampled for the same budget. This in turn yields more accurate estimates of spatio-temporal global quantities like means, fractions and totals, as long as the number of sampling times is not too small and the sampling events remain sufficiently well spread over the space–time universe.

Even more important is that synchronous patterns are much more flexible than static and static-synchronous ones. This is because at each sampling time the locations can be adapted to altered circumstances with respect to the spatial or temporal variation existing in the universe, the accumulating amount of information on both these sources of variation, the information needs or the available budget. As argued by Overton and Stehman (1996), the importance of flexibility in sample size and locations for long-term monitoring can hardly be overrated (see also Sect. 3.6).

Statistical inference from synchronous samples is straightforward and relatively simple compared to static, static-synchronous and rotational samples. For instance, design-based statistical testing of the change of the spatial mean is simple when the locations at different sampling times are selected independently.

The choice between the pattern types should also be guided by the target quantity. For instance, for the current mean (total, areal fraction) rotational patterns are more efficient than synchronous patterns (no overlap between samples at successive sampling times) and static-synchronous patterns (100 percent overlap), whereas for the change of mean (total, areal fraction) static-synchronous pattern are optimal, i.e. have the smallest prediction-error variance.

Advantages of rotational patterns over static and static-synchronous ones are greater flexibility and better spatial coverage. A drawback is that design-based inference of the sampling variance may be cumbersome. If there is a fair amount of correlation between observations at consecutive sampling times, then the advantage of rotational patterns compared with synchronous samples is higher efficiency in estimating temporal trends as well as current means and totals; see Sect. 15.2.1.

14.3 Contents

For both space and time a choice has to be made between the design-based and the model-based approach, so there are four possible combinations: design-

based in space and time, design-based in space plus model-based in time, model-based in space plus design-based in time, and model-based in space and time.

Chapter 15 deals with sampling methods for global quantities in space–time such as the spatio-temporal mean and the temporal trend of the spatial mean (also referred to as the spatial mean temporal trend). Section 15.2 describes purely design-based methods, i.e., the sampling events (locations and times) are selected by probability sampling. There is one exception, viz. sampling for the temporal trend of the spatial mean, which assumes that sampling times are selected by systematic sampling, not necessarily Systematic Random Sampling, and the temporal trend is estimated by model-based inference. A separate subsection (Sect. 15.2.6) deals with the situation where one is interested in the difference between the spatio-temporal means before and after an intervention (BACI designs).

Section 15.3 describes model-based methods, i.e., methods using a model for statistical inference. In this case, purposive sampling in space and time is recommendable. If prior to the sampling, a model for the variation in space–time can be postulated, then this model can be used to optimize the sampling events. This will be shown for various geostatistical models.

Chapter 16 deals with sampling methods for local quantities in space–time. Only model-based methods are presented in this chapter. Design-based methods for large space–time domains follow directly from Sect. 8.2.2. For sampling on a space–time grid, the effects of the interval length and the grid spacing on the prediction-error variance is explored for two prediction methods, viz. space-time kriging, which is a purely statistical method, and for Kalman Filtering, a data-assimilation method that uses a dynamic–mechanistic model to describe temporal variation.

In this part, the spatio-temporal target universe is denoted by $\mathcal{U}$, and assumed to be the Cartesian product of a two- or three-dimensional spatial universe $\mathcal{S}$ and a temporal universe $\mathcal{T}$: $\mathcal{U} = \mathcal{S} \times \mathcal{T}$. A sampling event in $\mathcal{U}$ is denoted by the vector $\mathbf{u} = (\mathbf{s}', t)'$, where $\mathbf{s}$ is the vector of spatial coordinates, and t is the temporal coordinate.

15

Global Quantities in Space–Time

15.1 Introduction to Methods for Global Quantities in Space–Time

In monitoring studies with a temporal and a spatial dimension, a large variety of global target quantities can be defined. Commonly estimated target quantities are:

- the current mean, i.e., the spatial mean at the most recent sampling time;
- the change of the spatial mean from one sampling time to the other;
- the temporal trend of the spatial mean;
- the spatial mean of the temporal trend
- the spatio-temporal mean;
- the difference between the spatio-temporal means before and after an intervention.

In some cases one is interested in totals rather than means, and means can be interpreted as fractions when the target variable is a 0/1 indicator variable. Also, one may be more interested in a change of the spatial mean from one sampling time to the other than in the current level of the spatial mean. This is because the change tells more about environmental processes than the status of the environment. For instance, in a study on the greenhouse effect, one may want to estimate the change in carbon stocks in soil between now and 10 years hence. Or, to calculate the water balance of a watershed, one may want to estimate the difference in groundwater storage at the beginning and at the end of a year. The change in the spatial mean from one sampling time to the other is defined as:

$$\bar{d}_{2,1} = \frac{1}{|\mathcal{S}|}\int_{\mathbf{s}\in\mathcal{S}} z(\mathbf{s},t_2)\,\mathrm{d}\mathbf{s} - \frac{1}{|\mathcal{S}|}\int_{\mathbf{s}\in\mathcal{S}} z(\mathbf{s},t_1)\,\mathrm{d}\mathbf{s} = \frac{1}{|\mathcal{S}|}\int_{\mathbf{s}\in\mathcal{S}} \{z(\mathbf{s},t_2) - z(\mathbf{s},t_1)\}\,\mathrm{d}\mathbf{s}\ . \tag{15.1}$$

(For model-based inference, the definition is of course the same, but the target variable $z(\cdot)$ and the target quantity $\bar{d}_{2,1}$ are random instead of deterministic; see (15.23).)

When sampling has been done at more than two sampling times, one may be interested in the mean change per time unit, i.e., the temporal trend. This trend may differ considerably between locations, and in that case one may want to estimate the Spatial Cumulative Distribution Function of the temporal trend, or one or more parameters of this SCDF. The spatial mean temporal trend is defined as the spatial mean of the trend parameter $\beta(\mathbf{s})$ of a (linear) time-series model for the target variable at location $\mathbf{s}$ at time t, $Z(\mathbf{s}, t)$

$$Z(\mathbf{s}, t) = \alpha(\mathbf{s}) + \beta(\mathbf{s}) \cdot (t - t_0) + \epsilon(\mathbf{s}, t) \; , \tag{15.2}$$

where t_0 is the first time the target variable is measured.

It can be shown that this spatial mean temporal trend is equal to the temporal trend of the spatial mean, i.e., the trend parameter β of a time-series model for the spatial mean of Z at time t, $\overline{Z}(t)$:

$$\overline{Z}(t) = \alpha + \beta \cdot (t - t_0) + \epsilon(t) \; . \tag{15.3}$$

A target quantity related to the spatial mean temporal trend is the temporal trend of the areal fraction where the target variable meets certain conditions, for instance the areal fraction where a quantitative target variable exceeds a given threshold. Clearly, a static type of pattern is inappropriate for estimating this quantity.

This target quantity should be distinguished from the areal fraction where the temporal trend meets a given condition. In this case the condition is in terms of the trend, whereas in the former it is in terms of the target variable. This areal fraction can be estimated by the indicator technique described before (Sect. 7.2.3). The Spatial Cumulative Distribution Function of the temporal trend is the most informative target quantity. This SCDF can simply be estimated by repeated application of the indicator technique.

The spatio-temporal mean is defined as

$$\overline{z}_{\mathcal{U}} = \frac{1}{|\mathcal{U}|} \int_{\mathbf{u} \in \mathcal{U}} z(\mathbf{u}) \, \mathrm{d}\mathbf{u} \; , \tag{15.4}$$

and similarly the fraction in space and time that $z(\mathbf{u})$ exceeds some critical threshold) is defined as

$$F_{\mathcal{U}}(z) \equiv \frac{1}{|\mathcal{U}|} \int_{\mathbf{u} \in \mathcal{U}} i(\mathbf{u}; z) \, \mathrm{d}\mathbf{u} \; . \tag{15.5}$$

with

$$i(\mathbf{u}; z) = \begin{cases} 1 & \text{if } \; z(\mathbf{u}) \leq z \\ 0 & \text{if } \; z(\mathbf{u}) > z \end{cases} \tag{15.6}$$

If a change in the environmental conditions is foreseen, for instance due to the implementation of measures that have a positive or negative effect

on the quality of the environment, then one may want to assess this effect. One may choose one sampling time before and one after the intervention, and estimate the change in the spatial mean, but the difference found may strongly depend on the chosen times. Therefore this difference can be a very imprecise estimate of the difference between the spatio-temporal means before and after the intervention. Repeated measurements in time, both before and after the intervention, will increase precision. When the putatively 'disturbed' area has such an extent that it is unreasonable to assume that the effect is equal everywhere, it is also recommended to repeat the measurements in space. For example, assume that in an agricultural area measures are planned to reduce the leaching of nitrate to the ground and surface water. To assess the effect of these measures one may want to estimate the total amount of leached nitrate in the area where the measures are planned in the year before the intervention $\mathcal{B}$, and in the year after the intervention $\mathcal{A}$. The target quantity to be estimated is

$$\begin{aligned} d_{\mathcal{A},\mathcal{B}} &= \int_{t\in\mathcal{A}} \int_{s\in\mathcal{S}} z(\mathbf{s},t)\,\mathrm{d}\mathbf{s}\,\mathrm{d}t - \int_{t\in\mathcal{B}} \int_{s\in\mathcal{S}} z(\mathbf{s},t)\,\mathrm{d}\mathbf{s}\,\mathrm{d}t \\ &= \int_{s\in\mathcal{S}} \left\{ \int_{t\in\mathcal{A}} z(\mathbf{s},t)\,\mathrm{d}t - \int_{t\in\mathcal{B}} z(\mathbf{s},t)\,\mathrm{d}t \right\} \mathrm{d}\mathbf{s}\,. \end{aligned} \tag{15.7}$$

As for global quantities in space and global quantities in time, design-based methods are the most appropriate. Especially for regulatory monitoring, objectivity of the method and validity of the results are of great importance. For instance, if a regulation specifies a threshold value (Action Level) for the spatio-temporal mean, then a valid interval estimate of this quantity is important for statistical testing.

15.2 Design-Based Methods for Global Quantities in Space–Time

15.2.1 Introduction

The typology of sampling patterns for monitoring presented in Sect. 14.1 is equally relevant for design-based and model-based methods. In model-based methods, however, the patterns are deterministic and in design-based methods they are random, i.e., selected by probability sampling. A static-synchronous sampling design, for instance, generates random static-synchronous patterns. This section deals with sampling and estimation for spatio-temporal and current global quantities, change of global quantities and spatial mean temporal trends, by synchronous, static, static-synchronous and rotational designs.

With synchronous designs, at each sampling time one is free to choose a spatial sampling design from Sect. 7.2 that seems most appropriate given

the circumstances at that time. So one may adapt the sample size, possible stratification, clusters and/or primary units, and even the very type of design.

Synchronous Sampling can be considered as a special case of Two-Stage Random Sampling in space–time, using spatial sections of the universe at given times as primary units, and sampling locations as secondary units (Vos, 1964). Therefore, the methods of inference for Two-Stage Random Sampling in space, given in Sect. 7.2.5, can be applied. For instance, inference about the spatio-temporal mean proceeds by first estimating the spatial mean at each sampling time (using the method associated with the spatial design at each time), and then estimating the spatio-temporal mean from these means as 'observations' (using the method associated with the temporal design). Inference on totals and trend parameters is similar.

With static designs the order of space and time in the two stages is reversed: sampling locations are selected as primary units and sampling times as secondary units. Now the set of sampling locations remains fixed through time, as with static-synchronous designs, which brings similar operational advantages. The difference with static-synchronous designs is that sampling is not synchronized, so that correlation due to synchronized sampling is avoided. Another difference is that the temporal design may be adapted to local circumstances. Static designs are attractive when considerable spatial variation between time series is known to exist, and when the operational advantages of fixed locations are real.

A static-synchronous design can be considered as a combination of a spatial sampling design and a temporal sampling design, so that at each sampling time all locations are sampled (see Fig. 14.4). The sampling locations can be selected by the same designs as described in Sect. 7.2 on design-based sampling in space, while the sampling times can be selected by the methods discussed in Sect. 11.2 on design-based sampling in time. The inference for static-synchronous designs depends primarily on these two constituting partial designs.

Rotational Sampling or 'sampling with partial replacement' represents a compromise between static and synchronous designs. The rationale is to avoid on the one hand the unbalancedness of static designs that accumulate more data only in time. On the other hand, the relative inefficiency of synchronous designs for estimating temporal trends is partially avoided because repeated measurements are made at the same locations.

The principle of Rotational Sampling is to divide the locations of an initial spatial sample into different rotational groups, and to replace each time one group by a new set of locations (see Fig. 14.5). Many different strategies of Rotational Sampling have been developed, including improved estimation procedures. In some strategies a set of locations would be re-introduced into the sample after having been rotated out for some time. See Binder and Hidiroglou (1988) for a review on Rotational Sampling.

The suitabilities, from a statistical point of view, of the four design types for estimating global quantities are summarized in Table 15.1.

Table 15.1. Suitability of the four main types of design for estimating spatio-temporal global quantities (StGQ), current global quantities (CuGQ), change of global quantities (ChGQ), and spatial mean temporal trends (SMTT). A question mark means that estimation of the standard error may be problematic.

Type of design	StGQ	CuGQ	ChGQ	SMTT
Synchronous	+	+	+	+
Static	+	−	−	++
Static-Synchronous	+?	+	++	+?
Rotational	+?	++	+	+?

15.2.2 Spatio-Temporal Global Quantities

Spatio-temporal global quantities most relevant in practice are spatio-temporal means, fractions and totals, and (parameters of) spatio-temporal Cumulative Frequency Distributions. An example of a spatio-temporal total is the total emission of a pollutant in a target area during a target period. The temporal mean of the spatial fraction of the area where the emission rate exceeds a given threshold is an example of a spatio-temporal fraction.

Synchronous Designs

Synchronous designs can be considered as two-stage designs, and therefore the formulas of Sect. 7.2.5 can be used to calculate the number of sampling locations and sampling times. The primary units are then spatial sections of the universe at given times, and sampling locations are secondary units. The (pooled) within-unit variance in this case is the (time-averaged) spatial variance of the target variable at a given time, and the between-unit variance is the variance of the spatial means over time. Note that (7.30 – 7.32) hold for Simple Random Sampling in space and Simple Random Sampling in time, an equal number of sampling locations at all sampling times, and a linear cost function $C = c_0 + c_1 n_\mathrm{t} + c_2 n_\mathrm{t} n_\mathrm{s}$, where c_0 is the fixed costs of installing the monitoring design, for instance costs of preparing the sampling frame, c_1 is the variable costs per sampling time and c_2 is the variable costs per sampling location, and n_t and n_s, are the number of sampling times and locations, respectively.

Usually, more efficient types of design than Simple Random Sampling will be chosen for space and time, for instance, systematic in time and stratified in space. In that case, the above formulas can still be used, either by adopting the resulting sample sizes n_t and n_s, as conservative (safe) estimates, or by dividing them by a prior estimate of the design-effects (accounting for the higher efficiency), e.g., 1.1 or 1.2. Of course, after one or more sampling rounds

the data then collected can be used as prior information for adapting parts of the design that are still to be carried out.

Estimating the spatio-temporal mean (total, fraction) proceeds by first estimating the spatial mean at each sampling time (using the method associated with the spatial design at that time), and then estimating the spatio-temporal mean and its standard error from these means as 'observations' (using the method associated with the temporal design). The spatio-temporal total is obtained by multiplying the mean with the size of spatio-temporal universe ($|\mathcal{S}| \cdot |\mathcal{T}|$), and similarly for the standard error.

Static Designs

Like synchronous designs, static designs can be considered as two-stage designs, but the role of space and time are interchanged. The primary units are now temporal sections at given locations, and sampling times are secondary units. The same formulas for the number of locations and times can be used. However, now the (pooled) within-unit variance is the (space-averaged) temporal variance at a given location, and the between-unit variance is the variance of the temporal means over space. The cost function is now $C = c_0 + c_1 n_\mathrm{s} + c_2 n_\mathrm{t} n_\mathrm{s}$, where c_1 is the variable costs per sampling location and c_2 is the variable costs per sampling time. The remark about using more efficient designs than Simple Random Sampling, made for synchronous designs, applies to static designs as well. Just as with synchronous designs, after some time the sample data then collected can be used as prior information in adapting parts of the design still to be carried out.

Inference about the spatio-temporal mean (total, fraction) proceeds by first estimating the temporal mean at each sampling location (using the method associated with the temporal design at that location), and then estimating the spatio-temporal mean and its standard error from these means as 'observations' (using the method associated with the spatial design). The spatio-temporal total is obtained by multiplying the mean with the size of spatio-temporal universe ($|\mathcal{S}| \cdot |\mathcal{T}|$), and similarly for the standard error.

Static-Synchronous Designs

Due to the two-fold alignment of the sampling events, sample optimization for static-synchronous designs is more complicated than for synchronous and static designs. With synchronous and static designs there are two variance components to take into account: the variance between and the (pooled) variance within primary units, i.e., spatial and temporal sections, respectively. With static-synchronous designs it appears that there are three variance components. Building on the early work of Quenouille (1949), Koop (1990) worked out the sampling variance in estimating (surface) areas for different combinations of two designs of point sampling in the plane, one along the X-axis and one along the Y-axis, with or without alignment of the sampling points in

either direction. Taking time for the Y-axis and space for the X-axis (or vice versa), one of Koop's designs types, 'random sampling with alignment in both directions', is analogous to static-synchronous sampling with Simple Random Sampling in both space and time. Translating Koop's variance formula for this type of design to estimation of the spatio-temporal mean gives, cf. Koop (1990, eq. 3.3.5):

$$V\left(\hat{\bar{z}}\right) = \frac{S^2(z)}{n} + \left(\frac{1}{n_\mathrm{t}} - \frac{1}{n}\right) S_\mathrm{t}^2(\bar{z}_\mathrm{s}) + \left(\frac{1}{n_\mathrm{s}} - \frac{1}{n}\right) S_\mathrm{s}^2(\bar{z}_\mathrm{t}) \;, \tag{15.8}$$

where $S^2(z)$ is the spatio-temporal variance of z over space and time, $S_\mathrm{t}^2(\bar{z}_\mathrm{s})$ is the variance over time of the spatial mean, and $S_\mathrm{s}^2(\bar{z}_\mathrm{t})$ is the variance over space of the temporal mean. This formula can be used for sample optimization as follows.

1. Make prior estimates of the three variance components.
2. Make estimates of the cost components in a linear cost function such as those mentioned under synchronous and static designs.
3. Choose relevant ranges for n_s and n_t and calculate for each combination of n_s and n_t the expected sampling variance and costs.
4. In case of quality maximization under a given budget constraint, select n_s and n_t for which the expected sampling variance is smallest and the expected costs are still within the budget.
5. In case of costs minimization under a given quality requirement, select n_s and n_t for which the expected costs is smallest and the expected sampling variance still meets the requirement.

Estimation can be done in two steps, the order of which may be reversed. First, for each sampling location the quantity over time is estimated from the data at that location, using the method associated with the temporal design. Then the spatio-temporal quantity and its standard error are estimated using these temporal values as 'observations', using the method associated with the spatial design. This standard error accounts automatically for errors due to sampling in space and sampling in time, but not for possible spatial correlations between the estimated temporal quantities due to synchronized sampling at the locations. This will generally lead to underestimation of the standard error. Due to the two-fold alignment of the sampling events, there is no unbiased estimator of the sampling variance available (Koop, 1990). One option is to substitute posterior estimates of the variance components in (15.8). Another possibility is to form, by random partioning, a number of smaller static-synchronous subsamples. The variance between the means of these subsamples could be then used as an estimate of the variance for the original sample.

15.2.3 Current Global Quantities

Synchronous Designs

The choice of the sampling design and the inference are as for global quantities in space (see Sect. 7.2). The inference depends only on the current type of spatial design. There is no overlap between the spatial samples at different sampling times, and as a result, there is no simple way of exploiting the information in the previous samples to estimate the current quantity. This is a drawback of synchronous designs compared to rotational designs, which do create such overlap, see Sect. 15.2.1.

Static-Synchronous Designs

Inference on a current global quantity, such as the spatial mean, fraction, total or (a parameter of) the Spatial Cumulative Distribution Function at any given sampling time, can be done by applying the appropriate method from Sect. 7.2 on the data collected at that time. To estimate the current quantity, only measurements taken at the current sampling time need to be used. There is no additional information in the measurements from the previous sampling times, because the locations coincide.

Rotational Designs

In Rotational Sampling there is partial overlap between samples of successive sampling times, and consequently the sample of the previous sampling time can be used in estimating a current spatial mean, fraction or total. We present the procedure for the mean; fractions are estimated by applying the same procedure to indicator variables, and totals are estimated by multiplying the estimated mean and its standard error with the size of the spatial universe (surface area in case of 2D).

To start with, two sampling times are considered. The sample of the first sampling time is subsampled, and on the locations of this subsample the target variable is also measured at the second sampling time. This subsample with measurements at both sampling times is referred to as the matched sample; the unmatched sample consists of the locations with measurements at the first sampling time only. At the second sampling time, the target variable is also measured on a set of new locations. The spatial mean at the second sampling time $\bar{z}_2$, is estimated by the composite estimator (Cochran, 1977, p. 346)

$$\hat{\bar{z}}_{2c} = \hat{w}_1 \, \hat{\bar{z}}_{2\mathrm{gr}}^{(m)} + \hat{w}_2 \, \hat{\bar{z}}_{2\pi}^{(u)} \,, \tag{15.9}$$

where $\hat{w}_1$ and $\hat{w}_2$ are weights summing to 1, $\hat{\bar{z}}_{2\pi}^{(u)}$ is the π-estimator for the mean of z_2 estimated from the unmatched sample, and $\hat{\bar{z}}_{2\mathrm{gr}}^{(m)}$ is the Two-Phase

Random Sampling regression estimator for the mean of z_2 estimated from the matched (remeasured) sample (see Sect. 7.2.12)

$$\hat{\bar{z}}_{2\mathrm{gr}}^{(m)} = \hat{\bar{z}}_{2\pi}^{(m)} + b\left(\hat{\bar{z}}_{1\pi} - \hat{\bar{z}}_{1\pi}^{(m)}\right) , \tag{15.10}$$

where $\hat{\bar{z}}_{2\pi}^{(m)}$ is the second sampling time mean estimated from the matched sample, b is the estimated slope coefficient, $\hat{\bar{z}}_{1\pi}$ is the first sampling time mean estimated from the entire first-phase sample (matched plus unmatched), and $\hat{\bar{z}}_{1\pi}^{(m)}$ is the first sampling time mean estimated from the matched sample. The estimated optimal weights $\hat{w}_1$ and $\hat{w}_2$ equal

$$\hat{w}_1 = 1 - \hat{w}_2 = \frac{\widehat{V}\left(\hat{\bar{z}}_{2\pi}^{(u)}\right)}{\widehat{V}\left(\hat{\bar{z}}_{2\mathrm{gr}}^{(m)}\right) + \widehat{V}\left(\hat{\bar{z}}_{2\pi}^{(u)}\right)} , \tag{15.11}$$

where $\widehat{V}\left(\hat{\bar{z}}_{2\mathrm{gr}}^{(m)}\right)$ is the estimated variance of the regression estimator, and $\widehat{V}\left(\hat{\bar{z}}_{2\pi}^{(u)}\right)$ is the estimated variance of the π-estimator for the mean of z_2 for the unmatched sample. For Simple Random Sampling of n locations at both sampling times and m matched (remeasured) locations this variance is given by

$$\widehat{V}\left(\hat{\bar{z}}_{2\mathrm{gr}}^{(m)}\right) = \frac{\widehat{S^2}(e)}{m} + \frac{\widehat{S^2}(z_2) - \widehat{S^2}(e)}{n} , \tag{15.12}$$

where $\widehat{S^2}(e)$ is the estimated variance of the residuals $e = z_2 - z_1 b$, and

$$\widehat{V}\left(\hat{\bar{z}}_{2\pi}^{(u)}\right) = \frac{\widehat{S^2}(z_2)}{n-m} . \tag{15.13}$$

The variance of the composite estimator can be estimated by (Schreuder et al., 1987):

$$\widehat{V}\left(\hat{\bar{z}}_{2c}\right) = \frac{1 + 4\,\hat{w}_1\,\hat{w}_2\left(\frac{1}{m-1} + \frac{1}{n-m-1}\right)}{\hat{w}_1 + \hat{w}_2} . \tag{15.14}$$

The variance depends on the proportion of matched sampling locations. The optimal matching proportion can be calculated with (Cochran, 1977)

$$\frac{m}{n} = \frac{\sqrt{1-\rho^2}}{1+\sqrt{1-\rho^2}} , \tag{15.15}$$

where ρ is the correlation coefficient between z_1 and z_2. For $\rho = 0.9$, 0.8 and 0.5 the optimum matching proportion equals 0.30, 0.38 and 0.48 respectively. When ρ goes to 0, m/n approaches 0.5. Given these correlation coefficients the gain in precision, calculated as the ratio of the variance with no matching (Synchronous Sampling) and the variance with the optimum matching

proportion, equals 1.39, 1.25 and 1.07, respectively. When the costs of remeasuring a location are lower than the costs of measuring a new location, the optimum matching proportion increases. When choosing a matching proportion one must take care that the number of matched locations is large enough to obtain reliable estimates of the regression coefficient and the variance of the regression estimator, say $m > 10$.

With three or more sampling times, the current mean $\bar{z}_0$ is estimated by substituting the composite estimator for the mean at the previous sampling time, $\hat{\bar{z}}'_{-1}$ (15.9), for the π-estimator for the previous mean, $\hat{\bar{z}}_{1\pi}$, in the regression estimator (15.10):

$$\hat{\bar{z}}^{(m)}_{0\mathrm{gr}} = \hat{\bar{z}}^{(m)}_{0\pi} + b\left(\hat{\bar{z}}_{-1c} - \hat{\bar{z}}^{(m)}_{-1\pi}\right) \ , \tag{15.16}$$

and then weighting this regression estimator and the π-estimator for the current mean inversely proportional to the variance (15.11). Cochran (1977) shows that the optimal matching proportion increases rapidly with the sampling time. For the fifth sampling time the optimal matching proportion is close to 0.5 for a correlation coefficient ≤ 0.95.

Once the matching proportion is chosen, one can calculate the sample size needed for a given precision (15.14). Prior estimates of the residual variance and the variance of the target variable at the current time are needed to calculate the weights.

15.2.4 Change of Global Quantities

Synchronous Designs

Change of the spatial mean (total, fraction) can be estimated as with static-synchronous designs (Eq. 15.19). Because the samples taken at different times are mutually independent, the estimated means $\hat{\bar{z}}(t_1)$ and $\hat{\bar{z}}(t_2)$ are uncorrelated. The sampling variance of $\hat{\bar{d}}_{2,1}$ equals

$$V\left(\hat{\bar{d}}_{2,1}\right) = V\left(\hat{\bar{z}}(t_2)\right) + V\left(\hat{\bar{z}}(t_1)\right) \ , \tag{15.17}$$

which can be simply estimated by:

$$\widehat{V}\left(\hat{\bar{d}}_{2,1}\right) = \widehat{V}\left(\hat{\bar{z}}(t_2)\right) + \widehat{V}\left(\hat{\bar{z}}(t_1)\right) \ . \tag{15.18}$$

Note that, contrary to (15.20), there is no covariance-term, which makes synchronous designs in general less efficient than static-synchronous designs. In the case of classical testing, this procedure leads to the common two-sample t-test. Change of spatial fractions and totals can be estimated in the same way as change of spatial means.

If both sampling rounds are still to be designed, one has to decide on the sampling design type and the sample size at both sampling times. In general

there will be no reason for choosing different design types for the two sampling times. Also, in general prior estimates of the spatial variance components for the target variable will be equal for the two sampling times. In that case the optimal ratio of sizes of the first and second sample will be 0.5. The sample size per sampling time required to estimate the change with prescribed precision can then be calculated by the formulas of Sect. 7.2, substituting half the maximum allowed variance of the estimated change for the variance of the estimated mean of the target variable.

After the first sampling time, one has new information that can be used to redesign the sample of the second time. The estimated sampling variance of the mean at the first sampling time can be subtracted from the variance of the estimated change to obtain the variance of the mean at the second sampling time. Also, estimates of the spatial variance components at the first sampling time can be used as prior estimates to calculate the sample size needed to meet this updated constraint on the sampling variance of the estimated mean at the second sampling time.

Static-Synchronous Designs

The change of the spatial mean or fraction from one sampling time to the other, $\bar{d}_{2,1}$ (15.1), can be estimated straightforwardly by

$$\hat{\bar{d}}_{2,1} = \hat{\bar{z}}(t_2) - \hat{\bar{z}}(t_1) \; . \tag{15.19}$$

In static-synchronous samples, the locations of the first and the second sampling time coincide. This implies that in estimating the sampling variance of the change, a possible temporal correlation between the estimated means $\hat{\bar{z}}(t_1)$ and $\hat{\bar{z}}(t_2)$ must be taken into account. The true sampling variance equals

$$V\left(\hat{\bar{d}}_{2,1}\right) = V\left(\hat{\bar{z}}(t_2)\right) + V\left(\hat{\bar{z}}(t_1)\right) - 2C\left(\hat{\bar{z}}(t_2), \hat{\bar{z}}(t_1)\right) \; . \tag{15.20}$$

So, the stronger (more positive) the temporal correlation between the two estimated spatial means, the smaller the sampling variance of the change. In general this correlation will be largest when the sampling locations at the first and second sampling time coincide, as is the case with static designs and nondestructive sampling. With destructive sampling, the shifts should be kept as small as possible. Also, if a spatial trend is suspected, then the direction of the separation vector must be randomized to avoid bias (Papritz and Flühler, 1994). A simple way to estimate the variance (15.20) is first calculating the difference $d_i = z_i(t_2) - z_i(t_1)$ at each sampling location i, and then applying the appropriate method of inference from Sect. 7.2 to those differences. (If change of a fraction is to be estimated, z is an indicator variable, and d can take the values -1, 0 or 1.) In the case of classical testing, this procedure leads to the common t-test for paired observations.

The change of a spatial total from one sampling time to the other can be estimated by multiplying the estimated change of the spatial mean with the

size of the target universe (area in case of 2D), and similarly for the standard error.

The required sample size can be calculated with the formulas from Sect. 7.2, substituting spatial variances of the differences for the spatial variances of the target variable. From the third sampling time onwards, these variance components can be estimated from the data of the previous sampling times.

Rotational Designs

The change of the spatial mean from the previous to the latest sampling time can be estimated by:

$$\hat{\bar{d}} = \hat{\bar{z}}_{0c} - \hat{\bar{z}}_{-1c} \,, \tag{15.21}$$

where $\hat{\bar{z}}_{0c}$ and $\hat{\bar{z}}_{-1c}$ are the composite estimators at the latest and the previous sampling time, respectively (see (15.9)). An alternative, more precise but more complicated estimator of the change is to combine two estimators of change, one built on the matched sample and one built on the unmatched sample, into a composite estimator with optimized weights (Schreuder et al., 1993, p. 180). Change fractions can be estimated by applying this method to indicator variables, and change of totals are obtained by multiplying estimated change of means with the size of the spatial universe.

15.2.5 Spatial Mean Temporal Trend

Synchronous Designs

With synchronous designs the spatial mean temporal trend (temporal trend of spatial mean) is estimated by first estimating the spatial means at time t, $\bar{z}(t)$, and then estimating the model parameter β in (15.3) and the variance of $\hat{\beta}$ by Weighted Least Squares fitting, with weights inversely proportional to the variances of spatial means. The variance accounts for uncertainty about β due to the residual term $\epsilon(t)$, and for uncertainty about the spatial means due to sampling errors.

Static Designs

Compared with synchronous designs, the inference proceeds in reversed order. First the temporal trend parameters are estimated for each sampling location separately, then these estimates are averaged to a spatial mean, using the method associated with the spatial design. To estimate the temporal trend of an areal fraction static designs are inappropriate.

Static-Synchronous Designs

To estimate the spatial mean temporal trend, first the model parameter $\beta(\mathbf{s})$ is estimated at all sampling locations, and then these estimates are used to estimate the spatial mean of the model parameter, $\bar{\beta}$. The variance of the estimated spatial mean temporal trend can be estimated straightforwardly by the estimators of Sect. 7.2. This variance accounts automatically for uncertainty about $\bar{\beta}$ due to sampling error, and for uncertainty about the model parameters due to the residual term $\epsilon(\mathbf{s}, t_i)$, but not for spatial correlations (due to synchronized sampling) between these two error-components.

If systematic sampling in time, i.e., sampling at constant time-intervals is applied, then the sampling frequency to be optimized, see Fig. 15.4. For the required number of sampling locations the formulas from Sect. 7.2 can be used. A prior estimate of the spatial variance of the temporal trend at locations (within the target area, strata, primary units or clusters) is needed. The (averaged) estimation variance of the temporal trend at locations must be added to this spatial variance, see (13.20).

Rotational Designs

Similarly to static designs, the spatial mean temporal trend can be estimated by first estimating the temporal trend $\beta(\mathbf{s})$ at the sampling locations, and then estimating the spatial mean of the model parameter, $\bar{\beta}$. However, compared to static designs, considerable time elapses before all sampling locations have been observed repeatedly. For instance, in the 4-period rotational sample of Fig. 14.7 one must wait for the seventh sampling time until all locations have been sampled three times. The alternative is to estimate the spatial means (areal fractions, totals) first, and then the trend of the spatial mean (areal fraction, total). With this procedure an estimate of the spatial mean temporal trend (trend of areal fraction or total) can already be obtained after the third sampling time. Successive estimates of the current mean (areal faction, total) estimated by the composite estimator evidently are correlated because measurements of the previous time are used to estimate the current global quantity. Therefore, it is recommendable not to use the measurements of the previous sampling time to estimate the current status of the global quantity, i.e., use the π-estimator. Due to overlap of successive samples the estimated global quantities at successive times still can be correlated, but this correlation will be much less serious a problem.

15.2.6 Estimating Effects with BACI designs

Figure 15.1 shows a monitoring design composed of two independent synchronous designs, one before and one after an intervention. The difference in spatio-temporal means before and after the intervention $\bar{d}_{\mathcal{A},\mathcal{B}}$ can be estimated by estimating the two space–time means with the design-based estimators mentioned in the previous sections. The sampling variance of the

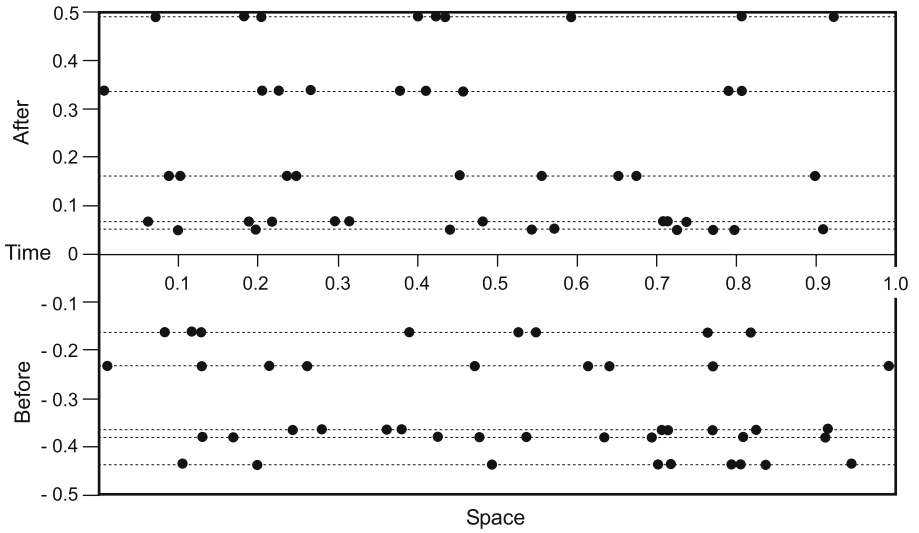


Fig. 15.1. Notional example of a synchronous sample before and after an intervention

estimated difference can be estimated simply by the sum of the variances of the two estimated space–time means. As stated above, Synchronous Sampling can be considered as a special case of Two-Stage Random Sampling in space–time, using spatial sections of the universe at given times as primary units, and sampling locations as secondary units. Assuming that the variance within and between primary units before and after the intervention are equal, the formulas of Sect. 7.2.5 can be used to determine the optimal number of sampling times and sampling locations before and after the intervention.

Suppose that the target variable shows a linear trend in time due to processes working in an area that is much larger than the area where the measures are planned. Then one will find a difference between the space–time means before and after the intervention which has nothing to do with the intervention. To overcome this problem, one can measure the target variable at one or more reference (control) sites, i.e., locations more or less similar to the impact sites but outside the area with the planned measures. The target quantity is now defined as

$$\bar{\delta}_{\mathcal{A},\mathcal{B}} = \frac{1}{|\mathcal{S} \times \mathcal{A}|} \int_{t \in \mathcal{A}} \int_{\mathbf{s} \in \mathcal{S}} \delta(\mathbf{s},t)\, \mathrm{d}\mathbf{s}\, \mathrm{d}t - \frac{1}{|\mathcal{S} \times \mathcal{B}|} \int_{t \in \mathcal{B}} \int_{\mathbf{s} \in \mathcal{S}} \delta(\mathbf{s},t)\, \mathrm{d}\mathbf{s}\, \mathrm{d}t \;, \quad (15.22)$$

where $\delta(\mathbf{s},t) = z(\mathbf{s},t) - \bar{z}_{\mathcal{C}}(t)$, with $\bar{z}_{\mathcal{C}}(t)$ being equal to the mean of z at the control sites at time t. The control sites can be restricted to one or more purposively selected locations outside the impact area. In that case the mean

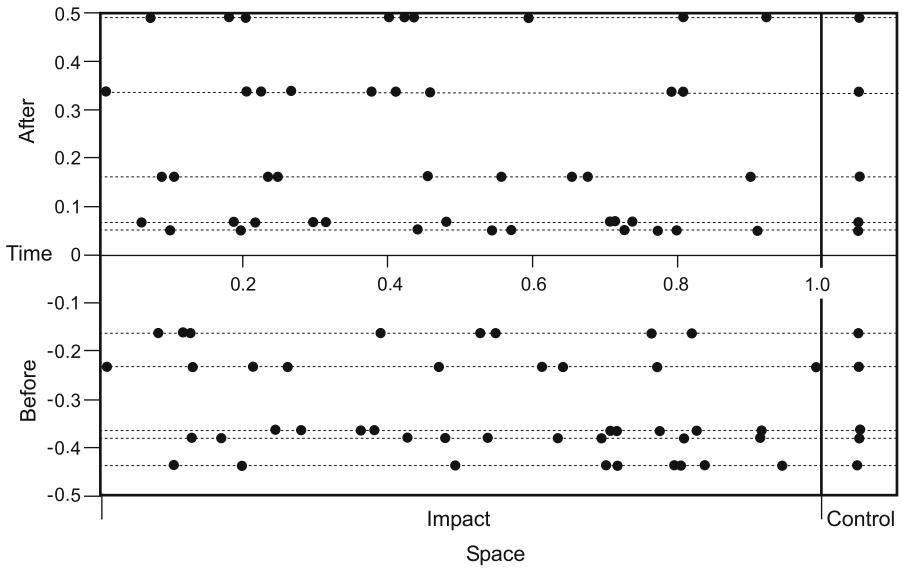


Fig. 15.2. Notional example of a synchronous sample before and after an intervention, with simultaneous sampling of a control site (a BACI design)

$\bar{z}_{\mathcal{C}}(t)$ is known without error. It is convenient to measure the impact and control sites simultaneously, because then estimating or testing the quantity (15.22) is analogous to that of (15.7). Figure 15.2 shows an example of a synchronous sampling pattern in the impact area, and one purposively selected control site. One may also select the control sites randomly from a bounded control area. In that case one can account for uncertainty in the spatial means $\bar{z}_{\mathcal{C}}(t)$. Random sampling in the control area can be synchronized with the sampling in the impact-area. Besides possible operational advantages, this also leads to higher precision when the estimated spatial means at time t in the control area and in the impact area are positively correlated.

Note that if the estimator of (15.22) differs significantly from zero, then still one cannot conclude that this is caused by the intervention. The treatment levels (impact versus control) are not randomly allocated to the sites as in experimental design, and as a consequence one must be careful to interpret the estimated target quantity as the effect of the treatment (intervention).

Underwood (1994) considers the case of a single impact location that is not randomly selected, but predetermined by the source of the disturbance. To introduce randomness, Underwood proposes selecting the control sites randomly. However, Stewart-Oaten and Bence (2001) pointed out that in this case the estimation of the effect of the intervention at the impact site must be based necessarily on a geostatistical model. The measurements after the intervention at the control sites are used to predict the target variable at the

impact site. These predictions are used as possible values if no intervention would have occurred at the impact site. Note that this strategy can be used if before-measurements at the impact site are unavailable. If one does have measurements before the intervention at the impact site, the alternative is to describe the temporal variation at the impact site with a time-series model and use the synchronized measurements at the control sites as covariates.

15.3 Model-Based Methods for Global Quantities in Space–Time

15.3.1 Introduction

This section describes sampling in space and time for prediction of global quantities (e.g., spatio-temporal means or spatio-temporal cumulative distribution functions), where a model is used for statistical inference. If prior to sampling a reasonable model can be postulated, then this model can also be used to guide the sampling, i.e., to optimize the pattern of sampling events. Although design-based methods generally are well suited for global quantities, there may be reasons to prefer model-based methods. An example is where one has prior monitoring data from purposive samples in space–time that need to be extended with additional data. Another example is where the global target quantity is related to a detection problem (Sect. 2.2.6).

The first two sections deal with optimization of the sampling locations at two given sampling times. The target quantities considered are the change of the mean between two successive sampling times (Sect. 15.3.2), and the current mean, i.e., the mean at the latest sampling time (Sect. 15.3.3). Both target quantities are predicted by co-kriging. The optimal patterns for the two target quantities will generally be different.

The following two sections deal with the situation where more than two sampling times are to be considered. In that case co-kriging becomes cumbersome, because a co-regionalization model for more than two co-variables is hard to obtain. The alternative is then to postulate a relatively simple geostatistical model for the variation in space–time. In Sect. 15.3.4 such a model is used to optimize the spacing and interval length of a space–time grid for predicting the spatio-temporal mean. Sect. 15.3.5 elaborates on optimization of the sample pattern for the current mean with this model.

Finally, Sect. 15.3.6 deals with sampling for predicting the spatial mean temporal trend. A simple model is postulated for the residuals of the temporal trend, which is used to optimize the spacing and interval length of a space–time grid.

In some situations one may have knowledge about the dynamics of the target variable, described by a process model. If this model can be used to predict the spatio-temporal evolution of the target variable, then the predicted spatio-temporal images can be used to direct sampling effort in space and

time. For instance, a groundwater flow and transport model could be used to describe the development of a contaminant plume by a series of concentration maps that are subsequently used to determine where and when concentrations can best be measured. Examples of the use of a process model for sampling in space–time can be found in Meyer and Bril (1988), Cieniawski et al. (1995) and Bierkens (2002). When a realistic process model is available, its use in sampling is recommendable because this is likely to increase the efficiency. The reason that we do not treat this type of model-based sampling is that it is highly application-specific and therefore beyond the scope of this book.

15.3.2 Co-Kriging the Change of Spatial Mean

This section deals with sampling for predicting the change of the spatial mean between two sampling times. What should be optimized are the sampling locations at these two times. The target quantity is predicted by block co-kriging. In co-kriging the change of the mean, there is not a primary variable and a secondary variable, but the cross-correlation between the random variable at sampling time 1, $Z(\mathbf{s}, t_1)$, and at sampling time 2, $Z(\mathbf{s}, t_2)$ is used to improve the predicted change of the spatial means.

Hereafter, we denote $Z_1(\mathbf{s}) = Z(\mathbf{s}, t_1)$ and $Z_2(\mathbf{s}) = Z(\mathbf{s}, t_2)$, and assume that $Z_1(\mathbf{s})$ and $Z_2(\mathbf{s})$ are two second-order stationary functions with unknown means μ_1 and μ_2 and with the following covariance functions: $C_{11}(\mathbf{h}) = \mathrm{Cov}[Z_1(\mathbf{s}), Z_1(\mathbf{s}+\mathbf{h})], C_{22}(\mathbf{h}) = \mathrm{Cov}[Z_2(\mathbf{s}), Z_2(\mathbf{s}+\mathbf{h})]$ and cross-covariance function $C_{12}(\mathbf{h}) = \mathrm{Cov}[Z_1(\mathbf{s}), Z_2(\mathbf{s}+\mathbf{h})]$.

Here we use covariances rather than semivariances, because the estimation of the cross-variogram requires that data at two sampling times are observed at the same location[1] (see Goovaerts, 1997).

The change of the spatial mean

$$\overline{D}_{2,1} = \frac{1}{|\mathcal{S}|}\int_{\mathbf{s}\in\mathcal{S}} Z_2(\mathbf{s})\,\mathrm{d}\mathbf{s} - \frac{1}{|\mathcal{S}|}\int_{\mathbf{s}\in\mathcal{S}} Z_1(\mathbf{s})\,\mathrm{d}\mathbf{s}\,, \tag{15.23}$$

is predicted by (Papritz and Flühler, 1994)

$$\widetilde{\overline{D}}_{2,1} = \sum_{i=1}^{n_2} \lambda_{2i}\, Z_2(\mathbf{s}_{2i}) - \sum_{i=1}^{n_1} \lambda_{1i}\, Z_1(\mathbf{s}_{1i})\,, \tag{15.24}$$

where n_1 and n_2 are the number of sampling locations at t_1 and t_2, respectively. The co-kriging weights λ_{1i} and λ_{2i} are obtained by solving the following sets of linear equations:

[1] An alternative formulation in terms of so-called pseudo cross-variograms is possible, which is also suitable for intrinsic Stochastic Functions. This however yields much more complicated expressions (see Papritz and Flühler, 1994).

$$\begin{aligned}
&\sum_{j=1}^{n_1} \lambda_{1j}\, C_{11}(\mathbf{s}_{1j} - \mathbf{s}_{1i}) - \sum_{j=1}^{n_2} \lambda_{2j}\, C_{21}(\mathbf{s}_{2j} - \mathbf{s}_{1i}) - \nu_1 \\
&\quad = C_{11}(\mathcal{S}, \mathbf{s}_{1i}) - C_{21}(\mathcal{S}, \mathbf{s}_{1i}) \qquad i = 1, \ldots, n_1 \\
&\sum_{j=1}^{n_2} \lambda_{2j}\, C_{22}(\mathbf{s}_{2i} - \mathbf{s}_{2j}) - \sum_{j=1}^{n_1} \lambda_{1j}\, C_{21}(\mathbf{s}_{2i} - \mathbf{s}_{1j}) - \nu_2 \\
&\quad = C_{22}(\mathbf{s}_{2i}, \mathcal{S}) - C_{21}(\mathbf{s}_{2i}, \mathcal{S}) \qquad i = 1, \ldots, n_2 \\
&\sum_{i=1}^{n_1} \lambda_{1i} = 1 \qquad \sum_{i=1}^{n_2} \lambda_{2i} = 1\,,
\end{aligned} \tag{15.25}$$

where ν_1 and ν_2 are Lagrange multipliers, and where two unbiasedness constraints are included to assure that the predictor (15.24) is unbiased. $C_{11}(\mathcal{S}, \mathbf{s}_{1i})$, $C_{21}(\mathcal{S}, \mathbf{s}_{1i})$, $C_{21}(\mathbf{s}_{2i}, \mathcal{S})$ and $C_{22}(\mathbf{s}_{2i}, \mathcal{S})$ are point-to-block averaged covariances. With $C_{11}(\mathcal{S}, \mathcal{S})$, $C_{21}(\mathcal{S}, \mathcal{S})$ and $C_{22}(\mathcal{S}, \mathcal{S})$, the within-block ($\mathcal{S}$-averaged) (cross-)covariances, the variance of the prediction error (block co-kriging variance) can be calculated as

$$\begin{aligned}
V(\widetilde{\overline{D}}_{2,1} - \overline{D}_{2,1}) &= C_{11}(\mathcal{S}, \mathcal{S}) + C_{22}(\mathcal{S}, \mathcal{S}) - 2C_{21}(\mathcal{S}, \mathcal{S}) + \nu_1 + \nu_2 \\
&\quad - \sum_{i=1}^{n_1} \lambda_{1i}[C_{11}(\mathcal{S}, \mathbf{s}_{1i}) - C_{21}(\mathcal{S}, \mathbf{s}_{1i})] - \sum_{i=1}^{n_2} \lambda_{2i}[C_{22}(\mathbf{s}_{2i}, \mathcal{S}) - C_{21}(\mathbf{s}_{2i}, \mathcal{S})]\,.
\end{aligned} \tag{15.26}$$

As can be seen from (15.26), the prediction-error variance depends only on the sampling locations at the sampling times t_2 and t_1 and can thus be used for optimization of the sampling locations. Papritz and Webster (1995) have shown that if the observations at the two times are positively correlated, then the prediction-error variance is minimal when the sampling locations at the two times coincide. When sampling is destructive (e.g., soil sampling), it is impossible to exactly sample the same location. In that case it is advisable to sample at sampling time t_2 as closely as possible to the sampling locations at sampling time t_1.

Some additional remarks about co-kriging of differences are in order. First, if the sampling locations at the two sampling times do not coincide, then co-kriging always yields more accurate predictions than first ordinary kriging separately at both sampling times and then subtracting the two predicted means. If the observations at the two sampling times coincide and the cross-covariance structure is intrinsic, i.e., $C_{11}(\mathbf{h}) = \alpha\, C_{22}(\mathbf{h}) = \beta\, C_{12}(\mathbf{h})$, where α and β are positive real valued constants for all lags $\mathbf{h}$, then co-kriging yields the same results as kriging for each sampling time first and then obtaining differences. The system is called 'autokrigeable'. If the system is autokrigeable, the kriging weights will also be the same for each sampling time. In this case, $\overline{D}_{2,1}$ can simply be estimated by direct ordinary block-kriging of differences:

$$\widetilde{\overline{D}}_{2,1} = \sum_{i=1}^{n_s} \lambda_i\, [Z_2(\mathbf{s}_i) - Z_1(\mathbf{s}_i)] = \sum_{i=1}^{n_s} \lambda_i\, D_{2,1}(\mathbf{s}_i)\,, \tag{15.27}$$

with λ_i obtained from solving the ordinary block-kriging equations (see Appendix B)

$$\begin{aligned} &\sum_{j=1}^{n_\mathrm{s}} \lambda_j \, \gamma_D(\mathbf{h}_{ij}) + \nu = \gamma_D(\mathbf{s}_i, \mathcal{S}) \qquad\qquad i = 1, \ldots, n_\mathrm{s} \\ &\sum_{i=1}^{n_\mathrm{s}} \lambda_i = 1 \; , \end{aligned} \tag{15.28}$$

with $\gamma_D(\mathbf{h}_{ij})$ the variogram of $D_{2,1}$, which should be estimated directly from the differences. The block-kriging variance of the predicted mean difference is given by:

$$V\left(\widetilde{\overline{D}}_{2,1} - \overline{D}_{2,1}\right) = \sum_{i=1}^{n_\mathrm{s}} \lambda_i \, \gamma_D(\mathbf{s}_i, \mathcal{S}) + \nu - \gamma_D(\mathcal{S}, \mathcal{S}) \; . \tag{15.29}$$

In conclusion, model-based sampling for predicting the change of the mean can be treated as a special case of model-based sampling in space because it is optimal to sample the same locations at the two times. The pattern with minimum block co-kriging variance (15.26) can be searched for by simulated annealing, see Sect. 7.3.3 for further details. Optimization becomes even more simple when an intrinsic covariance model is postulated. In that case block co-kriging is equivalent to block-kriging the differences, and the pattern can be optimized by minimization of the block-kriging variance of the predicted mean difference (15.29). A simple alternative for situations where one is not able to postulate a model for the variation in space–time, is to design a spatial coverage sample (Sect. 8.3.3) or a regular grid (Sect. 7.3.2).

15.3.3 Co-Kriging Current Means

In co-kriging the current mean, the measurement of the target variable at the previous sampling time is used as a secondary variable, i.e., a co-variable. The current mean is predicted by the ordinary block co-kriging predictor:

$$\widetilde{\overline{Z}}_2 = \sum_{i=1}^{n_2} \lambda_{2i} \, Z_2(\mathbf{s}_{2i}) + \sum_{i=1}^{n_1} \lambda_{1i} \, Z_1(\mathbf{s}_{1i}) \; , \tag{15.30}$$

The co-kriging weights λ_{1i} and λ_{2i} are obtained by solving the following sets of linear equations:

$$\begin{aligned} &\sum_{j=1}^{n_2} \lambda_{2j} \, C_{22}(\mathbf{s}_{2i} - \mathbf{s}_{2j}) + \sum_{j=1}^{n_1} \lambda_{1j} \, C_{21}(\mathbf{s}_{2i} - \mathbf{s}_{1j}) + \nu_1 \\ &\quad = C_{22}(\mathbf{s}_{2i}, \mathcal{S}) \qquad\qquad i = 1, \ldots, n_2 \\ &\sum_{j=1}^{n_2} \lambda_{2j} \, C_{12}(\mathbf{s}_{1i} - \mathbf{s}_{2j}) + \sum_{j=1}^{n_1} \lambda_{1j} \, C_{11}(\mathbf{s}_{1i} - \mathbf{s}_{1j}) + \nu_2 \\ &\quad = C_{12}(\mathbf{s}_{1i}, \mathcal{S}) \qquad\qquad i = 1, \ldots, n_1 \\ &\sum_{j=1}^{n_2} \lambda_{2j} = 1 \qquad\qquad \sum_{j=1}^{n_1} \lambda_{1j} = 0 \; , \end{aligned} \tag{15.31}$$

Finally, the block co-kriging variance of the predicted current mean equals

$$V\left(\widetilde{\overline{Z}}_2\right) = C_{22}(\mathcal{S},\mathcal{S}) - \nu_1 - \sum_{i=1}^{n_2} \lambda_{2i} C_{22}(\mathbf{s}_{2i},\mathcal{S}) - \sum_{i=1}^{n_1} \lambda_{1i} C_{12}(\mathbf{s}_{1i},\mathcal{S}) \; . \tag{15.32}$$

The optimal sampling pattern for the current mean may differ from the change of the mean. Whereas for the change of mean it is optimal to sample the same locations at the two times, for the current mean this will be optimal only when the spatial autocorrelation strongly dominates the temporal autocorrelation.

In the reverse case, it is optimal to sample at time t_2 at locations farthest from those at time t_1. This is because at these intermediate locations one has the least information on the current values, whereas at or near to a location sampled at t_1 a more precise estimate of the current value could be obtained with the observation at time t_1. In this case a simple solution is two interpenetrating grids, one for each sampling time. An alternative for irregularly shaped areas is to optimize the locations at time t_2 with k-means using the locations of time t_1 as prior locations, leading to a spatial infill sample (Sect. 8.3.3).

If neither the temporal nor the spatial autocorrelation is dominant, then the pattern of the locations might be optimized with simulated annealing, using the locations at the previous sampling time as prior data. The quality measure to be minimized is the block co-kriging variance of the predicted current mean.

15.3.4 Space–Time Kriging the Spatio-Temporal Mean

This section deals with the design of a space–time sample for the whole monitoring period, to predict the spatio-temporal mean. The sampling pattern will be optimized for the space–time block-kriging predictor.

Space–time kriging is a simple extension to spatial kriging, treating time as an extra dimension (e.g., Heuvelink et al., 1997). The spatio-temporal variation is modelled with a Stochastic Function $Z(\mathbf{s},t)$, $\mathbf{s} \in \mathcal{S}$ and $t \in \mathcal{T}$, which is assumed to be second-order stationary in both space and time. We model the space–time semivariance between $Z(\mathbf{s}_i,t_i)$ and $Z(\mathbf{s}_j,t_j)$ with the following variogram model, assuming isotropy in space and space–time geometric anisotropy (see Heuvelink et al., 1997):

$$\gamma(\mathbf{u}_i,\mathbf{u}_j) = \gamma(\mathbf{s}_i - \mathbf{s}_j,\, t_i - t_j) = \gamma(\mathbf{h}_{ij},\tau_{ij}) = \gamma\left(\sqrt{\frac{|\mathbf{h}_{ij}|^2}{a_\mathrm{s}^2} + \frac{\tau_{ij}^2}{a_\mathrm{t}^2}}\right) \; , \tag{15.33}$$

with $|\mathbf{h}_{ij}| = |\mathbf{s}_i - \mathbf{s}_j|$ and $\tau_{ij} = |t_i - t_j|$ the Euclidian distances in space and time, respectively, and a_s and a_t the variogram range parameters in space and time, respectively. The target quantity is the spatio-temporal mean $\overline{Z}_\mathcal{U}$ of $Z(\mathbf{s},t)$ over $\mathcal{U}$ (15.4).

If the mean value $\mu = E[Z(\mathbf{u})]$ is not known, $\overline{Z}_\mathcal{U}$ can be predicted with ordinary block-kriging:

$$\widetilde{\overline{Z}}_{\mathcal{U}} = \sum_{i=1}^{n} \lambda_i \, Z(\mathbf{u}_i) \; , \tag{15.34}$$

where the weights λ_i are obtained by solving the following set of equations

$$\begin{aligned} &\sum_{j=1}^{n} \lambda_j \, \gamma(\mathbf{u}_i, \mathbf{u}_j) + \nu = \gamma(\mathbf{u}_i, \mathcal{U}) \qquad\qquad i = 1, \ldots, n \\ &\sum_{i=1}^{n} \lambda_i = 1 \end{aligned} \tag{15.35}$$

and the variance of the prediction error is given by

$$V[\widetilde{\overline{Z}}_{\mathcal{U}} - \overline{Z}_{\mathcal{U}}] = \sigma^2_{\mathrm{obk}} = \sum_{i=1}^{n} \lambda_i \, \gamma(\mathbf{u}_i, \mathcal{U}) + \nu - \gamma(\mathcal{U}, \mathcal{U}) \; , \tag{15.36}$$

with

$$\gamma(\mathbf{u}_i, \mathcal{U}) = \frac{1}{|\mathcal{U}|} \int_{\mathbf{u} \in \mathcal{U}} \gamma(\mathbf{u}_i, \mathbf{u}) \, \mathrm{d}\mathbf{u} \tag{15.37}$$

$$\gamma(\mathcal{U}, \mathcal{U}) = \frac{1}{|\mathcal{U}|^2} \int_{\mathbf{u}_2 \in \mathcal{U}} \int_{\mathbf{u}_1 \in \mathcal{U}} \gamma(\mathbf{u}_1, \mathbf{u}_2) \, \mathrm{d}\mathbf{u}_1 \, \mathrm{d}\mathbf{u}_2 \; . \tag{15.38}$$

In practice, the integrals (15.37) and (15.38) are approximated by discretizing $\mathcal{U}$ with a grid and averaging semivariances between locations on the grid (see Appendix B). The prediction-error variance (15.36) can be used as a quality measure to be minimized through sampling. It can be seen that this depends only on the projected n sampling locations. Thus, it can be used for sample optimization when new sampling locations are projected.

A simple and practical type of sampling pattern for space–time kriging the spatio-temporal mean is a space–time grid. The trade-off between the sampling effort in space and in time, and the effect of grid spacing and the interval length on the prediction-error variance will be evaluated for a square grid pattern in space, in a block-shaped universe $\mathcal{S} \times \mathcal{T}$.

From analysis of the prediction-error variance (15.36) it follows that for grid sampling the ratio $\sigma^2_{\mathrm{obk}}/\sigma^2$ with $\sigma^2 = V[Z(\mathbf{s}, t)]$ can be represented by the following function $r(\cdot)$:

$$\frac{\sigma^2_{\mathrm{obk}}}{\sigma^2} = r\left(n_{\mathrm{s}}, n_{\mathrm{t}}, \frac{a_{\mathrm{s}}}{\sqrt{|\mathcal{S}|}}, \frac{a_{\mathrm{t}}}{|\mathcal{T}|}\right) \; , \tag{15.39}$$

where n_{s} and n_{t} are the number of sampling locations and sampling times, respectively. Figure 15.3 shows isolines of the ratio $\sigma^2_{\mathrm{obk}}/\sigma^2$ as a function of n_{s} and n_{t} for $a_{\mathrm{s}}/\sqrt{|\mathcal{S}|} = 1$ and $a_{\mathrm{t}}/|\mathcal{T}| = 1$, using a spherical model (with zero nugget) for the variogram (see Appendix B). Appendix D shows similar figures for other combinations of $a_{\mathrm{s}}^2/\sqrt{|\mathcal{S}|}$ and $a_{\mathrm{t}}/|\mathcal{T}|$. These figures can be used to determine the required grid spacing and interval length for grid sampling in

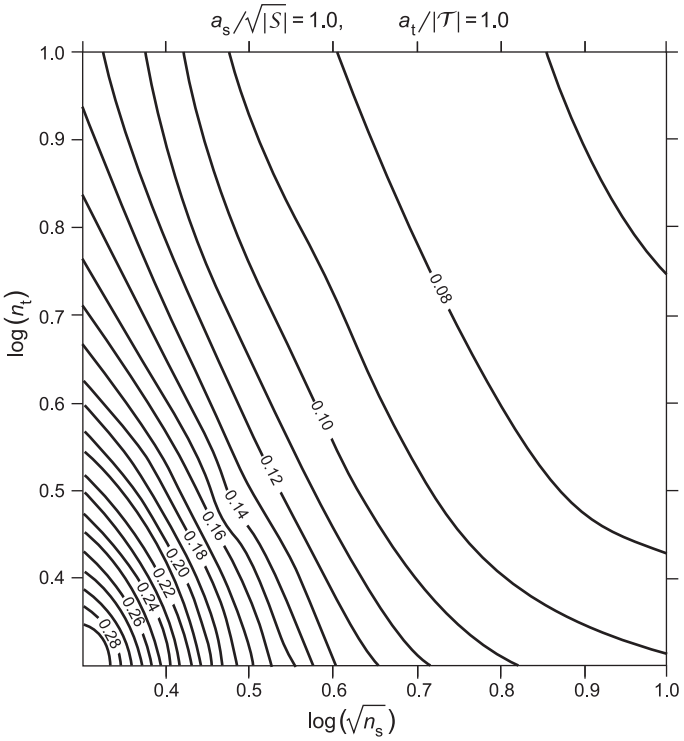


Fig. 15.3. Sampling on a centred space–time grid (square grid pattern in space) for predicting the spatio-temporal mean by space–time kriging. The figure shows the variance ratio $\sigma^2_{\text{obk}}/\sigma^2$ for $a_s/\sqrt{|\mathcal{S}|} = 1$ and $a_t/|\mathcal{T}| = 1$ as a function of the number of sampling locations n_s and sampling times n_t. Figures for other combinations of $a_s/\sqrt{|\mathcal{S}|}$ and $a_t/|\mathcal{T}|$ are given in Appendix D.

space and time, given values of area $|\mathcal{S}|$, length of monitoring period $|\mathcal{T}|$ and the statistical parameters σ^2, a_s and a_t.

Suppose that the aim is to predict the spatio-temporal mean for a block-shaped universe, square in space, and that a variance reduction of 90 percent is required, i.e., $\sigma^2_{\text{obk}}/\sigma^2 = 0.1$. Further, suppose that $a_s/\sqrt{|\mathcal{S}|} = 1$ and $a_t/|\mathcal{T}| = 1$, so that we can use Fig. 15.3. One possible combination of sample sizes in space and time is $\log(\sqrt{n_s}) = 0.47$ and $\log(n_t) = 0.82$. Rounding fractions upwards to integers, this leads to 9 sampling locations and 7 sampling times. Alternatively, one could choose the combination $\log(\sqrt{n_s}) = 0.70$ and $\log(n_t) = 0.40$, leading to 25 sampling locations and 3 sampling times. The number of sampling events is 63 for the first combination and 75 for the second. To determine which combination is preferable, a costs model can be used. If the total costs of sampling (c) is dominated by the costs per event (c_o), for instance due to lengthy observation times or an expensive method of

determination, then the total sample size is leading. This implies that taking a number of observations in space at a single time costs the same as taking the same number of observations at a single location at multiple times, i.e., $c = n_s n_t c_o$. Given this costs model, the first combination is least expensive and therefore preferable.

Repeated sampling is often more expensive than taking the same number of observations in one sampling round. A linear costs model accounting for this effect is $c = n_s n_t c_o + n_t c_t$, where c_t is the fixed costs per sampling round. Given a required variance reduction, the optimal combination can be determined by evaluating this cost function for all combinations on the corresponding isoline in Fig. 15.3, and selecting the least expensive combination.

Alternatively, for quality optimization the aim is to find the sample size combination that results in the smallest prediction error variance for a given budget B, i.e., $n_s n_t c_o + n_t c_t \leq B$. In this case the optimal combination can be found by plotting the line $n_s n_t c_o + n_t c_t = B$ in Fig. 15.3. The point on this line for which the variance ratio is minimal is the optimal combination.

Figures D.1 to D.4 in Appendix D show that if $a_s/\sqrt{|\mathcal{S}|}$ and $a_t/|\mathcal{T}|$ are similar in magnitude, then the lines run roughly diagonal with a slope of approximately -2. Note that if we would have plotted $\log(n_s)$ instead of $\log(\sqrt{n_s})$ the slope would have been -1, indicating that the effect on the prediction error variance of adding one sampling location with n_t observations is equal to the effect of adding one sampling time at which $n_s = n_t$ locations are observed. In case $a_s/\sqrt{|\mathcal{S}|} \ll a_t/|\mathcal{T}|$, the lines run roughly vertical (bottom diagrams in Fig. D.1), showing that much more can be gained by adding sampling locations, while for $a_s/\sqrt{|\mathcal{S}|} \gg a_t/|\mathcal{T}|$ (upper left diagram in Fig. D.4) adding sampling times is much more efficient.

Although Figs. D.1 to D.4 are based on a square area, these figures can also be used to obtain rough estimates for irregularly shaped areas. Note that for such areas the number of sampling locations is not restricted to squares of integers (4, 9, 16, 25 etc.), but can be any integer. For irregularly shaped areas a regular grid can be too restrictive. Alternatives are a spatial coverage pattern type (Sect. 8.3.3) or a geostatistical pattern type (Sect. 8.3.4).

Figures D.1 to D.4 are based on variograms without nugget. For variograms with nugget a different set of figures is required. First, substituting part of the structured variance by unstructured variance (nugget variance) leads to smaller variance ratios $\sigma^2_{\text{obk}}/\sigma^2$. The larger the nugget-to-sill ratio, the smaller the ratio $\sigma^2_{\text{obk}}/\sigma^2$, i.e., the stronger the variance reduction. This implies that less observations are required to achieve the same variance reduction as depicted in these figures.

Second, it turns out that the nugget has an effect on the optimal sample-size combination, i.e., the optimal grid spacing and interval length. The larger the nugget-to-sill ratio, the smaller the difference between the number of sampling locations per spatial correlation length and the number of sampling times per temporal correlation length.

In case the universe $\mathcal{S} \times \mathcal{T}$ is irregular or observations have already been made, other spatial patterns such as those of spatial coverage samples may be preferable. Even a different type of space–time pattern, such as interpenetrating space–time grids, could be in order.

A standard reference in model-based sampling design for spatio-temporal means is the paper by Rodríguez-Iturbe and Meíja (1974) on the design of rainfall networks. Here, the target is the long term spatial mean ($|\mathcal{T}| \to \infty$) and the sampling is exhaustive in time (rainfall is measured as cumulative amounts) such that the length of the monitoring period and the number of rain gauges are the variables to be optimized.

15.3.5 Space–Time Kriging Current Means

The previous section treats the design of space–time samples for the whole monitoring period, simultaneously for all sampling times. This section deals with the situation where a spatial sample is designed for the next sampling time only, in other words *spatial* samples are designed sequentially.

As with co-kriging, static-synchronous patterns such as space–time grids is a good choice only when the spatial autocorrelation strongly dominates the temporal autocorrelation. In this case the sampling problem can be treated as one of sampling in space, see Sect. 7.3.

In the reverse case, there are two simple solutions. The first solution is an interpenetrating space–time grid. The required spacing of the grids at each sampling time might be approximated by calculating the space–time block-kriging variance for a range of grid spacings. The second solution is to design a spatial infill sample with k-means, using the locations of all sampling times that are temporally autocorrelated as prior locations (Sect. 8.3.3).

If neither the temporal nor the spatial autocorrelation is dominant, then the pattern of the locations might be optimized with simulated annealing, using the previous sampling events as prior data. The quality measure to be minimized is the block-kriging variance of the predicted current mean.

15.3.6 Kriging the Spatial Mean Temporal Trend

A question such as 'has the surface temperature increased over the last 30 years?' is quite common in environmental research. Usually such a question has to be answered based on a small number of time series of the variable involved (e.g. temperature) scattered around the area of interest. If the time series are long enough it is possible to estimate a trend at each sampling location, see Sect. 13.4.1. Of course, the magnitude and sign of the trend may be different at different locations, so that the question whether the average temperature has increased in a certain area cannot be answered by looking at time series only. The real question to be answered is therefore whether the spatial mean of the temporal trend in temperature is positive and significantly different from zero. Consequently, the goal of this section is sampling for

predicting spatial mean temporal trends. A general space–time model for this purpose has been developed by Sølna and Switzer (1996). Here we will use a much simpler model for designing a sample in space–time.

This model has the following form:

$$Z(\mathbf{s},t) = \alpha(\mathbf{s}) + \beta(\mathbf{s}) \cdot (t - t_0) + \epsilon(\mathbf{s},t) \;, \tag{15.40}$$

where $\alpha(\mathbf{s})$ and $\beta(\mathbf{s})$ are level and trend coefficients respectively, that are Stochastic Functions of location in space and, for a given location $\mathbf{s}$, parameters in time, t_0 is the initial time and $\epsilon(\mathbf{s},t)$ is a zero-mean residual which is assumed to have the following properties:

$$E[\epsilon(\mathbf{s}_1,t_1) \cdot \epsilon(\mathbf{s}_2,t_2)] = \begin{cases} \sigma_\epsilon^2 \exp(-|t_2 - t_1|/a_\mathrm{t}) & \text{if } \mathbf{s}_1 = \mathbf{s}_2 \\ 0 & \text{if } \mathbf{s}_1 \neq \mathbf{s}_2 \end{cases} \tag{15.41}$$

In words, we assume that the residuals are correlated in time, but are uncorrelated in space.

Equation (15.40) can be reformulated in matrix–vector form as:

$$\mathbf{z}(\mathbf{s}) = \begin{bmatrix} z(\mathbf{s},t_1) \\ z(\mathbf{s},t_2) \\ \vdots \\ z(\mathbf{s},t_{n_\mathrm{t}}) \end{bmatrix} \qquad \boldsymbol{\beta}(\mathbf{s}) = \begin{bmatrix} \alpha(\mathbf{s}) \\ \beta(\mathbf{s}) \end{bmatrix}$$

$$\mathbf{T}(\mathbf{s}) = \begin{bmatrix} 1 & t_1 - t_0 \\ 1 & t_2 - t_0 \\ \vdots & \vdots \\ 1 & t_{n_\mathrm{t}} - t_0 \end{bmatrix} \quad \boldsymbol{\epsilon}(\mathbf{s}) = \begin{bmatrix} \epsilon(\mathbf{s},t_1) \\ \epsilon(\mathbf{s},t_2) \\ \vdots \\ \epsilon(\mathbf{s},t_{n_\mathrm{t}}) \end{bmatrix} \;,$$

so that

$$\mathbf{z}(\mathbf{s}) = \mathbf{T}(\mathbf{s}) \cdot \boldsymbol{\beta}(\mathbf{s}) + \boldsymbol{\epsilon}(\mathbf{s}) \;. \tag{15.42}$$

Using (15.41) the covariance matrix $\mathbf{C}(\mathbf{s}) = \boldsymbol{\epsilon}(\mathbf{s}) \cdot \boldsymbol{\epsilon}(\mathbf{s})'$ can be constructed. With the help of this covariance matrix and the above matrix–vector definitions the Generalized Least Squares estimate of $\boldsymbol{\beta}(\mathbf{s})$ can be obtained as (Cressie, 1993):

$$\widehat{\boldsymbol{\beta}}(\mathbf{s}) = \left[\mathbf{T}'(\mathbf{s}) \cdot \mathbf{C}^{-1}(\mathbf{s}) \cdot \mathbf{T}(\mathbf{s})\right]^{-1} \cdot \mathbf{T}'(\mathbf{s}) \cdot \mathbf{z}(\mathbf{s}) \;, \tag{15.43}$$

and the estimation covariance matrix as

$$\mathbf{V}[\widehat{\boldsymbol{\beta}}(\mathbf{s})] = \left[\mathbf{T}'(\mathbf{s}) \cdot \mathbf{C}^{-1}(\mathbf{s}) \cdot \mathbf{T}(\mathbf{s})\right]^{-1} \;. \tag{15.44}$$

From application of (15.43) and (15.44) to all n_s locations one obtains estimates of trends $\widehat{\beta}(\mathbf{s}_i)$ and the variances of the estimation errors $V[\widehat{\beta}(\mathbf{s}_i)]$.

Next, the spatial average $\overline{\beta}$ can be predicted using block-kriging of the $\widehat{\beta}(\mathbf{s}_i)$, where the estimation errors $\widehat{\beta}(\mathbf{s}_i) - \beta(\mathbf{s}_i)$ of the temporal estimation

problem are now treated as 'observation' errors in a spatial context. Thus, ordinary block-kriging with uncertain data is used (de Marsily, 1986). The prediction and prediction-error variance have the same form as with the regular ordinary kriging system:

$$\widetilde{\overline{\beta}} = \sum_{i=1}^{n_\mathrm{s}} \lambda_i \, \widehat{\beta}(\mathbf{s}_i) \tag{15.45}$$

$$V\left(\widetilde{\overline{\beta}} - \overline{\beta}\right) = \sum_{i=1}^{n_\mathrm{s}} \lambda_i \, \gamma_\beta(\mathbf{s}_i, \mathcal{S}) + \nu - \gamma_\beta(\mathcal{U}, \, \mathcal{U}) \; , \tag{15.46}$$

but the normal equations to obtain the weights and the value of the Lagrange multiplier have additional terms containing the estimation variances: $V[\widehat{\beta}(\mathbf{s}_i)]$:

$$\begin{aligned} &\sum_{j=1}^{n_\mathrm{s}} \lambda_j \, \gamma_\beta(\mathbf{h}_{ij}) - \lambda_i V[\widehat{\beta}(\mathbf{s}_i)] + \nu = \gamma_\beta(\mathbf{s}_i, \mathcal{S}) \qquad\qquad i = 1, \ldots, n_\mathrm{s} \\ &\sum_{i=1}^{n_\mathrm{s}} \lambda_i = 1 \end{aligned} \tag{15.47}$$

The function $\gamma_\beta(\mathbf{h}_{ij})$ is the variogram of the real trend coefficients β. Of course this is unknown. What can be estimated from the estimates $\widehat{\beta}(\mathbf{s}_i)$ at the sampling locations is the variogram $\gamma_{\widehat{\beta}}(\mathbf{h}_{ij})$. An approximation of the true variogram $\gamma_\beta(\mathbf{h}_{ij})$ may be obtained as follows ($n_{\mathrm{t}i}$ is the number of sampling times at sampling location i; the sampling interval length is assumed to be constant and equal for all sampling locations):

$$\gamma_\beta(\mathbf{h}_{ij}) \approx \gamma_{\widehat{\beta}}(\mathbf{h}_{ij}) - \frac{\sum\limits_{i=1}^{n_\mathrm{s}} n_{\mathrm{t}i} V[\widehat{\beta}(\mathbf{s}_i)]}{\sum\limits_{i=1}^{n_\mathrm{s}} n_{\mathrm{t}i}} \; . \tag{15.48}$$

In practice, the prediction of the spatial mean temporal trend consists of the following steps:

1. perform a Generalized Least Squares estimate of the trend parameter at each location with a time series. This entails:
   a) start with an Ordinary Least Squares regression of $\alpha + \beta(t - t_0)$ to the time series;
   b) calculate the residuals ϵ_t;
   c) estimate the covariance of the residuals (using the variogram estimator if observations are not equally spaced in time; see Chap. 9);
   d) fit relation (15.41) to the estimated covariance function;
   e) build the covariance matrix $\mathbf{C}$ with (15.42) and perform the Generalized Least Squares estimate with (15.43);
   f) repeat steps b to e until the estimate $\widehat{\beta}(\mathbf{s}_i)$ converges;

g) evaluate (15.44) to obtain the estimation variance $V[\widehat{\beta}(\mathbf{s}_i)]$;

2. estimate the variogram $\gamma_{\widehat{\beta}}(\mathbf{s}_i - \mathbf{s}_j)$ from the estimated trend coefficients $\widehat{\beta}(\mathbf{s}_i)$ at the locations and fit a permissible variogram function (see Chap. 9);
3. approximate the true variogram $\gamma_\beta(\mathbf{h}_{ij})$ with (15.48), making sure that $\gamma_\beta(\mathbf{h}_{ij})$ is positive for all lags;
4. solve (15.47) and evaluate (15.45) and (15.46) to obtain the prediction $\widetilde{\overline{\beta}}$ and the prediction error variance $V(\widetilde{\overline{\beta}} - \overline{\beta})$.

The prediction and the prediction-error variance can then be used to calculate a prediction interval for the spatial mean temporal trend $\overline{\beta}$. Assuming normality and a confidence level of 0.95, the interval equals $\widetilde{\overline{\beta}} \pm 1.96\sqrt{V(\widetilde{\overline{\beta}} - \overline{\beta})}$. If this interval does not include zero, one can conclude that a spatial mean temporal trend exists.

In this context, the smallest relevant trend $\beta_{\min}$ that can still be detected, can be used to assess the sample size needed. To this end, a quality requirement related to the half-width of the 95% prediction interval can be used:

$$V(\widetilde{\overline{\beta}} - \overline{\beta}) \leq \left(\frac{\beta_{\min}}{1.96}\right)^2 . \tag{15.49}$$

The sampling problem is then to assess the required length of the time series, and the number and locations of these time series to make sure that the quality requirement (15.49) is met.

In the following we consider the case of sampling on a space–time grid, with a square grid pattern in space. Some figures are presented that can be used to determine the optimal interval length and grid spacing required to estimate the spatial mean temporal trend with prescribed precision. The assumptions are that the temporal covariance parameters in (15.41) are equal for all locations, and that the variogram $\gamma_\beta(\mathbf{h}_{ij})$ is of spherical form with zero nugget. Given these assumptions, the parameters that must be known are: length of the time series $|\mathcal{T}|$, size of the area $|\mathcal{S}|$, temporal covariance parameters σ_ϵ^2 and a_t and semivariance parameters σ_β^2 and a_s.

Figure 15.4 shows the relation between the ratio $V(\widehat{\beta})|\mathcal{T}|^2/\sigma_\epsilon^2$ and the number of sampling times n_t, for several ratios $a_\mathrm{t}/|\mathcal{T}|$. Note that the trend estimation variance has been normalized both by the residual variance as well as by the length of the monitoring period to obtain a dimensionless parameter. Given the residual variance σ_ϵ^2 and the length of monitoring period $|\mathcal{T}|$, the estimation variance $V(\widehat{\beta})$ decreases with the number of sampling times n_t. The smaller temporal autocorrelation length, the stronger this sampling-frequency effect is. Figure 15.5 shows isolines of the variance ratio $V(\widetilde{\overline{\beta}} - \overline{\beta})/\sigma_\beta^2$ for combinations of the ratio $V(\widehat{\beta})/\sigma_\beta^2$ and the number of sampling locations n_s, for $a_\mathrm{s}/\sqrt{|\mathcal{S}|} = 0.1,\ 0.5,\ 1.0,\ 2.0$ (σ_β^2 is the sill of the variogram of the real

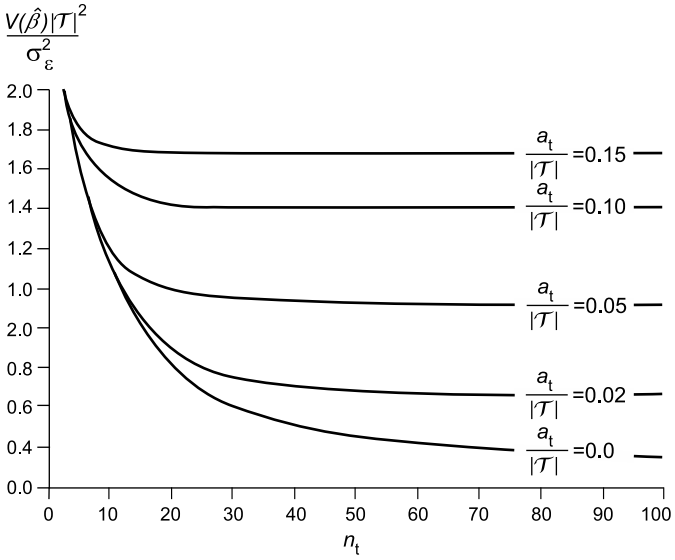


Fig. 15.4. Systematic sampling for estimating the temporal trend at a given location with Generalized Least Squares, assuming an exponential temporal covariance of the residuals. The figure shows the relationship between the dimensionless estimation variance $V(\widehat{\beta})|\mathcal{T}|^2/\sigma_\epsilon^2$ and the number of sampling times n_t, for five ratios $a_t/|\mathcal{T}|$.

trend coefficient). Figures 15.4 and 15.5 can be used to evaluate the effect of the number of sampling times and the number of sampling locations on the prediction-error variance $V(\widetilde{\overline{\beta}} - \overline{\beta})$. First, for a proposed n_t the value of $V(\widehat{\beta})$ is determined from Fig. 15.4. Next, using the value of $V(\widehat{\beta})$ to determine the ratio $V(\widehat{\beta})/\sigma_{\overline{\beta}}^2$, the variance ratio $V(\widetilde{\overline{\beta}} - \overline{\beta})/\sigma_{\overline{\beta}}^2$, and thus $V(\widetilde{\overline{\beta}} - \overline{\beta})$ can be obtained from Fig. 15.5 for a given number of sampling locations n_s. This way, combinations of numbers of sampling locations and sampling times can be sought that are in accordance with quality requirement (15.49).

Although space–time grids clearly have operational advantages, we would like to stress that for predicting spatial mean temporal trend this type of pattern will not always be optimal. In situations with a large temporal range of the variogram of the residuals, an r-period synchronous pattern may be more efficient.

The assumption that the spatio-temporal residual $\epsilon(\mathbf{s}, t)$ is spatially independent, leading to spatially independent estimation errors $\widehat{\beta}(\mathbf{s}) - \beta(\mathbf{s})$ is rather strong. However, it leads to relatively simple equations and is therefore suitable for sampling design. If, after data have been collected, it turns out that these assumptions are not supported by the data, then a more general model for inference and prediction may be in order. One option is the statisti-

cal space–time model and prediction method described by Sølna and Switzer (1996).

A special case of a temporal trend is a step trend, a sudden change in the model mean due to an intervention, see Sect. 13.4.1. If one wants to account for changes in the model mean not related to the intervention, then one may also sample synchronously one or more purposively selected control sites outside the intervention area, and postulate a time model for the pairwise differences between the impact and control sites, see Sect. 15.2.6.

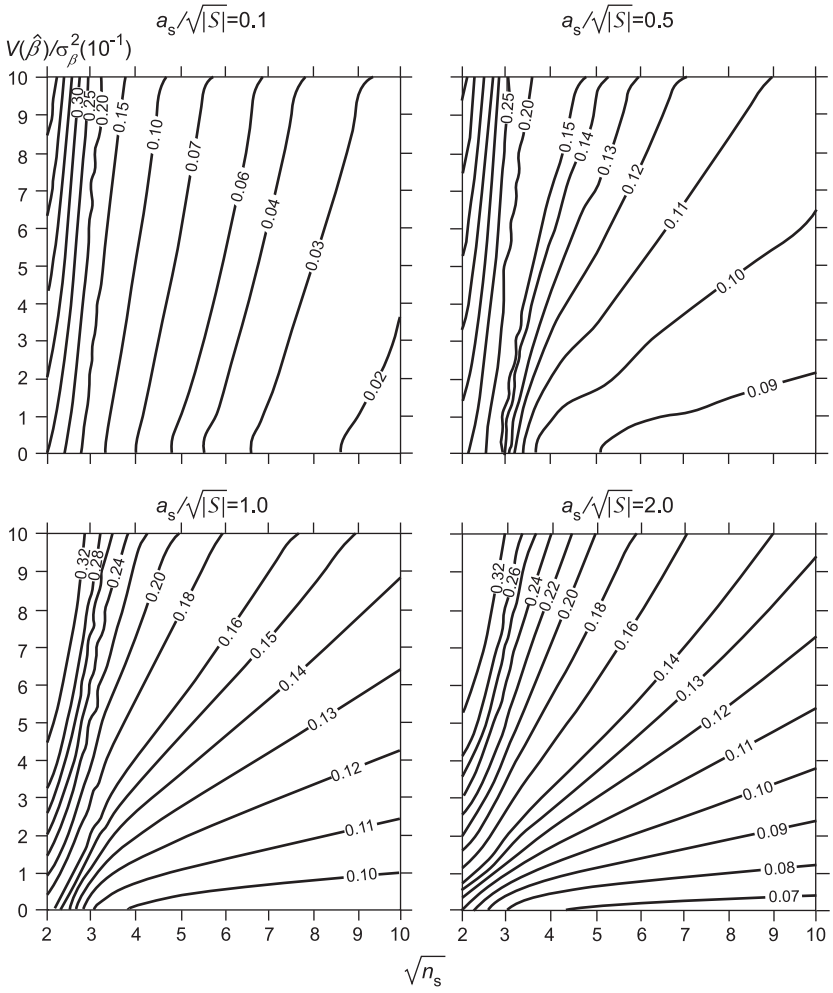


Fig. 15.5. Sampling on a centred space–time grid (square grid pattern in space) for predicting the spatial mean temporal trend by kriging with uncertain data, assuming a spherical spatial covariance model. The figure shows the ratio $V(\widehat{\overline{\beta}} - \overline{\beta})/\sigma^2_\beta$ as a function of the ratio $V(\widehat{\beta})/\sigma^2_\beta$ and the number of sampling locations n_s, for four different ratios $a_s/\sqrt{|\mathcal{S}|}$.

16

Local Quantities in Space–Time

16.1 Introduction to Methods for Local Quantities in Space–Time

With local quantities we mean the target variable at multiple locations or blocks and at multiple time steps or time intervals. Typical goals in this context are spatio-temporal mapping, i.e., mapping some variable over the entire space–time universe, and updating (spatial) maps. We make this distinction because it leads to different sampling problems. For spatio-temporal mapping a space–time sample is designed for the whole monitoring period, whereas in updating maps a spatial sample is designed for the next sampling time only, in other words *spatial* samples are designed sequentially.

In most of the literature on spatio-temporal mapping and updating maps model-based methods are used. Application of design-based methods for local space–time quantities seem to be rare in natural resource monitoring. Design-based methods come into scope when one has few domains, and as a result the number of sampling events per domain is considerable, and validity of results is important, as in regulatory monitoring. Examples are estimation of spatio-temporal means, totals or fractions for a limited number of sub-areas and no subdivision of the monitoring period, and estimation of the spatial mean temporal trend for several sub-areas.

When the domains are known at the beginning of the monitoring and one can afford a reasonable sample size for each domain, we recommend to sample each domain separately and independently, i.e., the domains are used as strata in stratified random sampling. Thus one can control the sample sizes, and one can choose for each stratum a suitable basic design type for monitoring (static, synchronous or rotational design), along with suitable partial designs for sampling in space and sampling in time. We refer to Sect. 14.1 for deciding on the basic design types for monitoring, and Sect. 7.2.2 for deciding on the spatial design. For sampling in time we refer to Part III.

When there are numerous domains, independent sampling of all the domains may be too costly. Also, sometimes it is impossible to delineate the

domains simply because one does not know where they are. When it is impractical or impossible to single out the domains as strata, one may select a sample from the total space–time universum, independent from any division into domains. The sample data are then sorted afterwards according to the domains in which the sampling events happen to fall. The sample sizes of the domains are random rather than controlled at the selection stage. For estimators for the mean and its sampling variance we refer to Sect. 8.2.2 (subsection 'Sampling Across Large Sub-Areas'). If one has only a few sampling events per domain, one might think of design-based estimators that use observations in similar domains, such as the synthetic estimator and the generalized regression estimator for small domains, see Sect. 8.2.2 (subsection 'Sampling Across Small Sub-Areas').

16.2 Model-Based Methods for Local Quantities in Space–Time

16.2.1 Introduction

As stated in Sect. 16.1, two primary goals are treated: space–time mapping and mapping current values. With regard to the model, two different approaches can be distinguished: the geostatistical and the data assimilation approach. For both modelling approaches, a relatively simple exponent is used in designing samples for the two goals. This leads to four combinations of goals and modelling approaches, treated in separate sections.

The geostatistical approach is conceptually simple and therefore seems attractive. However, since most environmental variables show trends, periodicity or changing spatial variation with time, a stationary Stochastic Function is often a poor model for environmental variables, making ordinary and simple space–time kriging of limited use.

The inclusion of temporal trends in space–time kriging has resulted in many complicated space–time models, with separate procedures for trend estimation and the prediction of residuals (e.g., Rouhani and Hall, 1989; Angulo et al., 1998). However, apart from being complicated, such models are difficult to infer from limited data and they are not always realistic. A further problem with such space–time models is that they do not take explicit account of direction and causality in time. For instance, Snepvangers et al. (2003) found, while mapping soil moisture content with space–time kriging, that the predicted moisture content had started to rise before the actual rainfall event occurred. Thus, space–time models using the autoregressive nature of time series (Wikle and Royle, 1999) and exogenous variables such as precipitation to explain variation and trend (Knotters and Bierkens, 2001) are more suitable for modelling space–time phenomena. Also, such models often have a simpler structure.

If there exists a dynamic-mechanistic model (e.g. a partial differential equation) to describe the spatio-temporal variation of an environmental variable, it is advisable to use the model in the sampling problem. One way is to describe the principal spatio-temporal variation (say the trend) with a mechanistic model and to predict the residuals with kriging. Another way is to merge observations and results of the mechanistic model in a physically consistent manner, making sure that can the predictions obey certain physical laws. Such methods are called data assimilation methods. One of these methods, called the 'Kalman filter' is treated in Sect. 16.2.4.

An overview of geostatistical and data-assimilation methods is provided in Sect. 16.2.6.

16.2.2 Space–Time Kriging the Space–Time Map

Space–time kriging is a extension to spatial kriging where time is simply treated as an extra dimension. Good examples of this technique are given by Rouhani and Hall (1989) and Heuvelink et al. (1997). Recent advances in space–time kriging can for instance be found in De Cesare et al. (2001); Gneiting (2002); Myers (2002); De Iaco et al. (2003); Hartfield and Gunst (2003); Huerta et al. (2004) and Lophaven et al. (2004). Let us assume that the target quantity is the value of some target variable z, varying in a spatio-temporal universe $\mathcal{U}$. The unknown spatio-temporal variation is modelled with a Stochastic Function $Z(\mathbf{u})$, which is assumed to be second-order stationary in both space and time. We model the space–time semivariance between $Z(\mathbf{u}_i)$ and $Z(\mathbf{u}_j)$ with the following variogram model, assuming isotropy in space and space–time geometric anisotropy (see Heuvelink et al., 1997):

$$\gamma(\mathbf{u}_i, \mathbf{u}_j) = \gamma(\mathbf{s}_i - \mathbf{s}_j,\, t_i - t_j) = \gamma(\mathbf{h}, \tau) = \gamma\left(\sqrt{\frac{|\mathbf{h}|^2}{a_\mathrm{s}^2} + \frac{\tau^2}{a_\mathrm{t}^2}}\right) , \qquad (16.1)$$

with $|\mathbf{h_{ij}}| = |\mathbf{s}_i - \mathbf{s}_j|$ and $\tau = |t_i - t_j|$ the Euclidean distances in space and time, respectively, and a_s and a_t the range parameters in space and time, respectively. If the mean value $\mu = E[Z(\mathbf{u})]$ is not known, $Z(\mathbf{s}_0, t_0)$ at space–time coordinate $(\mathbf{s}_0, t_0)$ can be predicted by ordinary kriging as a weighted mean of the sample data:

$$\widetilde{Z}(\mathbf{s}_0, t_0) = \sum_{i=1}^{n} \lambda_i\, Z(\mathbf{s}_i, t_i) , \qquad (16.2)$$

where the weights λ_i are obtained by solving the following set of equations

$$\begin{aligned} &\sum_{j=1}^{n} \lambda_j\, \gamma(\mathbf{s}_i - \mathbf{s}_j,\, t_i - t_j) + \nu = \gamma(\mathbf{s}_i - \mathbf{s}_0,\, t_i - t_0) \qquad i = 1, \ldots, n \\ &\sum_{i=1}^{n} \lambda_i = 1 , \end{aligned} \qquad (16.3)$$

and the variance of the prediction error is given by

$$V[\widetilde{Z}(\mathbf{s}_0, t_0) - Z(\mathbf{s}_0, t_0)] = \sigma^2_{\mathrm{ok}} = \sum_{i=1}^{n} \lambda_i \, \gamma(\mathbf{s}_i - \mathbf{s}_0, \, t_i - t_0) + \nu \, . \qquad (16.4)$$

If predictions are to be made on a grid of size N in space–time with coordinates $\mathbf{u}_k$ $(k = 1, \ldots, N)$, to make spatio-temporal maps, then a possible quality measure to minimize with regard to the sampling events $\mathbf{u}_i$ $(i = 1, \ldots, n)$ is given by:

$$J[(\mathbf{u}_i), \; i = 1, \ldots, n] = \overline{\sigma^2_{\mathrm{ok}}} = \frac{1}{N} \sum_{k=1}^{N} V[\widetilde{Z}(\mathbf{u}_k) - Z(\mathbf{u}_k)] \, . \qquad (16.5)$$

It can be seen from combining (16.4) and (16.5) that the quality measure depends only on the coordinates of the n sampling events.

With regard to the type of sampling pattern for space–time mapping, we recommend to first consider a space–time grid. This is a practical type of pattern when the installation costs of sample plots are high. If the budget is given, and one can derive from this the total sample size, then still one must decide on the grid spacing (number of locations) and the sampling interval length (number of sampling times).

Figures 16.1 and 16.2 show the effect of the grid spacing and interval length on the kriging variance for ordinary space–time kriging. The distances between sampling events on the grid in space Δs and time Δt were varied relative to the correlation scales. This was achieved by changing the variogram ranges a_{s} and a_{t} of a spherical variogram of form (16.1), which has the same effect as changing Δs and Δt while keeping the ranges constant. We calculated, for various combinations of $\Delta s/a_{\mathrm{s}}$ and $\Delta t/a_{\mathrm{t}}$, the average prediction-error variance $\overline{\sigma^2_{\mathrm{ok}}}$ (16.5) for a block-shaped spatio-temporal universe discretized with $10 \times 10 \times 10 = 1000$ grid nodes and a distance of 10 grid nodes between sampling events (so the spatio-temporal universe is bordered by 8 sampling events). The universe for which $\overline{\sigma^2_{\mathrm{ok}}}$ was calculated was placed in the centre of a larger spatio-temporal universe of $30 \times 30 \times 30 = 27000$ grid nodes with the same sampling density (i.e., a total of 64 sampling events) to avoid edge effects in the calculation of $\overline{\sigma^2_{\mathrm{ok}}}$. Figure 16.1 shows for $\Delta s/a_{\mathrm{s}} = \Delta t/a_{\mathrm{t}} = 1$ the spatial variation of $\sigma^2_{\mathrm{ok}}/\sigma^2$ (with $\sigma^2 = V[Z(\mathbf{s}, t)]$) for the large universe at two time steps, at a sampling time and between sampling times, and the temporal variation at two locations, one close to a sampling location and one further away. The effect of observations reducing the uncertainty in both space and time is clearly visible. Figure 16.2 shows isolines of the average variance ratio $\overline{\sigma^2_{\mathrm{ok}}}/\sigma^2$ as a function of the ratios $\Delta s/a_{\mathrm{s}}$ and $\Delta t/a_{\mathrm{t}}$. This figure clearly shows the trade-off between sampling in time and space.

The spatio-temporal variogram or covariance function deserves some further attention. A variogram like (16.1) may be difficult to infer from the available space–time data, when sampling has been dense in time and sparse

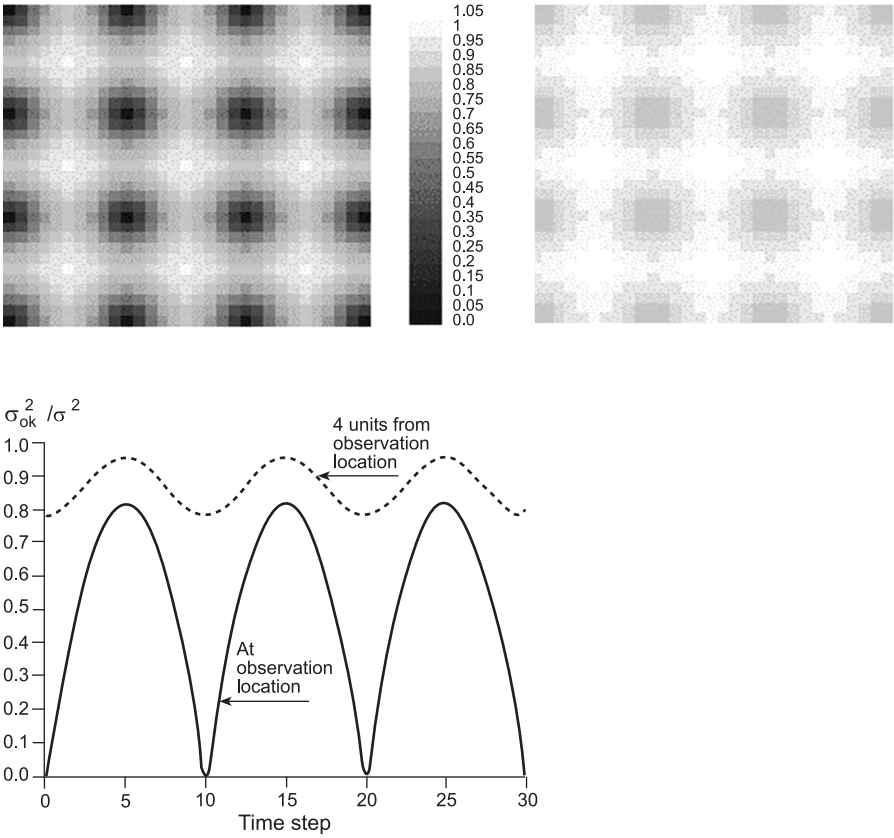


Fig. 16.1. Spatial variation of the space–time ordinary kriging variance-ratio $\sigma^2_{\mathrm{ok}}/\sigma^2$ at a sampling time (upper left) and between sampling times (upper right), and time series of the space–time ordinary kriging variance-ratio $\sigma^2_{\mathrm{ok}}/\sigma^2$ at a sampling location and between sampling locations; $\Delta s/a_{\mathrm{s}} = \Delta t/a_{\mathrm{t}} = 1$

in space (e.g., time series at a few locations). For this reason, a so-called separable covariance model is often adopted. This means that $Z(\mathbf{s}, t)$ is thought to consist of either the sum or the product of a purely spatial and a purely temporal Stochastic Function, which are assumed to be mutually independent. This yield the following two covariance models:

$$C(\mathbf{h}, \tau) = C_{\mathrm{s}}(\mathbf{h}) + C_{\mathrm{t}}(\tau) \,, \tag{16.6}$$

$$C(\mathbf{h}, \tau) = C_{\mathrm{s}}(\mathbf{h}) \cdot C_{\mathrm{t}}(\tau) \,. \tag{16.7}$$

The advantage of such a model is that the data can be averaged over time first to obtain the spatial covariance function, and vice versa. However, the additive model (16.6) in particular, is problematic. First, as explained by

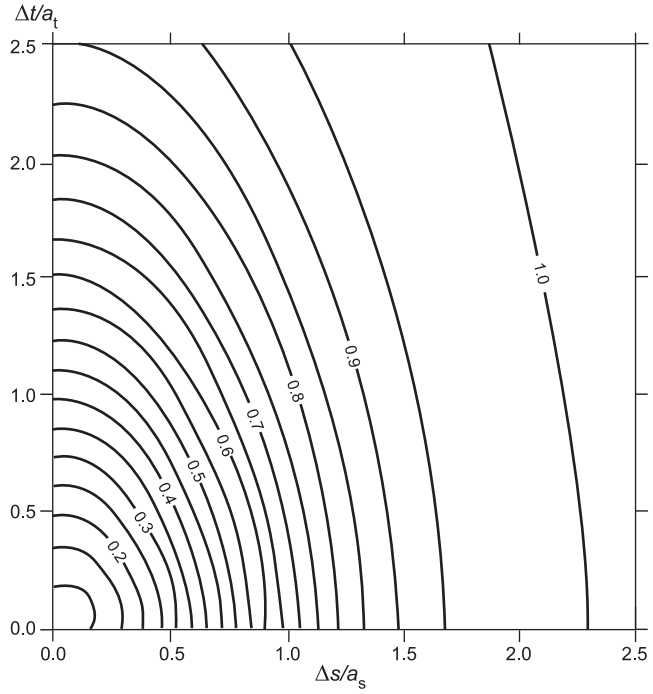


Fig. 16.2. Average variance ratio $\overline{\sigma^2_{\mathrm{ok}}}/\sigma^2$ for sampling on a space–time grid and space–time ordinary kriging, as a function of the ratios $\Delta s/a_{\mathrm{s}}$ and $\Delta t/a_{\mathrm{t}}$

Rouhani and Myers (1990) and Heuvelink et al. (1997), this model leads to singularity problems in the case of Grid Sampling in space–time. Second, Heuvelink et al. (1997) also point out that this model leads to unrealistic spatio-temporal variation, involving time series at different locations varying completely parallel in time; a spatial image that does not change its form with time but only changes its mean level. Myers (2002) also works with combinations of (16.6) and (16.7), leading to more realistic results. Rodríguez-Iturbe and Meíja (1974) use the following form of (16.7):

$$C(\mathbf{h}, \tau) = \sigma^2 \cdot \rho_{\mathrm{s}}(\mathbf{h}) \cdot \rho_{\mathrm{t}}(\tau) \; . \tag{16.8}$$

This is a convenient model, but care should be taken when such a model is used with Grid Sampling, since both $\rho_{\mathrm{s}}(\mathbf{h})$ and $\rho_{\mathrm{t}}(\tau)$ are represented by single covariance models (instead of, e.g., a linear combination of two or more models, see Appendix B). If observations are made at fixed time steps, simple space–time kriging is equivalent to co-kriging, using every time slice as a covariable (Bogaert, 1996). If all locations have been observed at all time slices (a case in co-kriging called 'isotopy'), and $\rho_{\mathrm{s}}(\mathbf{h})$ and $\rho_{\mathrm{t}}(\tau)$ are represented by

single covariance models (called 'intrinsic correlation' in co-kriging), then the entire prediction problem is 'autokrigeable' (Wackernagel, 1994). This means that co-kriging, and therefore also space–time kriging, yield exactly the same results as spatial kriging of every time slice separately, and no use is made of the temporal correlation to improve predictions.

Similar to sampling in space, there may be prior sample data from locations that do not match with the grid. In that case one might prefer a spatial coverage pattern for the sampling locations. Another possible reason for using this type of pattern is an irregular shape of the area. Again one must decide on the sampling density in space and sampling frequency. The diagrams in Appendix D can also be used to approximate these sample characteristics.

16.2.3 Space–Time Kriging Current Maps

The previous section treats the design of space–time samples for the whole monitoring period, simultaneously for all sampling times. This section deals with the situation where a spatial sample is designed for the next sampling time only, in other words *spatial* samples are designed sequentially.

Static-synchronous patterns such as space–time grids are a good option only when the spatial autocorrelation strongly dominates the temporal autocorrelation. In this case the sampling problem can be treated as one of sampling in space, see Sect. 8.3.

In the reverse case, there are two simple solutions. The first solution is an interpenetrating space–time grid. The required spacing of the grids at each sampling time might be approximated by calculating the average or maximum space–time kriging variance for a range of grid spacings. This average or maximum is calculated over the space domain only. The second solution is to design a spatial infill sample with k-means, using the locations of all sampling times that are temporally autocorrelated as prior locations (Sect. 8.3.3).

If neither the temporal nor the spatial autocorrelation is dominant, then the pattern of the locations might be optimized with simulated annealing, using the previous sampling events as prior data. The quality measure to be minimized is again the average or maximum space–time kriging variance over the space domain.

If temporal autocorrelation is weak, and consequently it does not pay to use the data of the second-last and earlier sample times to predict the current values, then co-kriging current maps can be a good alternative to space–time kriging. We refer to Sect. 15.3.3 for optimization of sample patterns with co-kriging. For the point co-kriging equations we refer to Goovaerts (1997).

16.2.4 Kalman Filtering the Space–Time Map

Kalman filtering is a method for optimal prediction of dynamic (temporally varying) linear stochastic systems. If a dynamic model can be cast into a linear state–space formulation and the error sources of the model can be identified,

Kalman filtering can be used to predict the model's output variables using observations. As will be shown in this section, the Kalman filter can also be used for prediction in space and time, especially if the spatio-temporal variation of the target variable is described by a partial differential equation. Merging observations and model predictions from partial-differential equations is referred to as 'data-assimilation', a term first coined in the meteorological and oceanographical literature. The Kalman filter is one particular form of data-assimilation as is further explained in Sect. 16.2.6. Here, the space–time Kalman filter is explained using an example from groundwater hydrology.

First a rather extensive description of the Kalman filter is given. After that we describe how the Kalman filter can be used in space–time sampling problems. Two sampling goals are treated: 1) optimizing a static synchronous design (space-time grid) for space-time mapping; 2) optimizing a synchronous dynamic design for predicting current values.

The example described here is partly based on the work of van Geer et al. (1991) and Zhou et al. (1991). The target quantity is the hydraulic head h (water table elevation with respect to some reference level) at a location $\mathbf{s} \in \mathcal{S} \subset \mathbb{R}^2$ and a time $t \in \mathcal{T} \subset \mathbb{R}$. The spatio-temporal variation is modelled with a Stochastic Function $H(\mathbf{s}, t)$ as follows.

Groundwater head $h(\mathbf{s}, t)$ for two-dimensional flow in a (semi-)confined aquifer is described with the following partial-differential equation:

$$S(\mathbf{s})\frac{\partial h}{\partial t} = \nabla.\,[\mathbf{T}(\mathbf{s})\nabla h(\mathbf{s}, t)] + Q(\mathbf{s}, t)\ , \tag{16.9}$$

where S is the storage coefficient [T], $\mathbf{T}$ a matrix (tensor) with transmissivities [$\mathrm{L^2T^{-1}}$] whose elements T_{ij} relate the flow in direction i to a head gradient in direction j, and Q [$\mathrm{LT^{-1}}$] is a source/sink term representing, for instance, groundwater withdrawal and groundwater recharge. As can be seen, S and $\mathbf{T}$ may vary in space and Q may vary in both space and time. In practice, a partial-differential equation like (16.9) is solved numerically using finite difference or finite element algorithms. In both cases, $h(\mathbf{s}, t)$ is solved at a finite number of locations and time steps discretizing $\mathcal{S} \times \mathcal{T} : h(\mathbf{s}_i, t_k), \quad i = 1, \ldots, N, \quad k = 1, \ldots, K$. After numerical approximation of (16.9) it is possible to write $h(\mathbf{s}_i, t_k), \quad i = 1, \ldots, N$ in a matrix–vector (state–space) formulation:

$$\mathbf{h}_k = \mathbf{A}_k\,\mathbf{h}_{k-1} + \mathbf{B}_k\,\mathbf{u}_k\ , \tag{16.10}$$

where $\mathbf{h}_k = (h(\mathbf{s}_1, t_k), \ldots, h(\mathbf{s}_N, t_k))'$ are the groundwater heads at time step k at the nodes discretizing the universe, $\mathbf{u}_k$ is a vector containing all source terms and boundary conditions and $\mathbf{A}_k$ and $\mathbf{B}_k$ are matrices whose elements depend on the space and time discretization, the storage coefficients and the transmissivities (note that if the time and space discretization are fixed, $\mathbf{A}_k = \mathbf{A}$ and $\mathbf{B}_k = \mathbf{B}$ are constant too).

The Stochastic Function is now defined for the discretized universe only by adding a noise vector to (16.10):

$$\mathbf{h}_k = \mathbf{A}_k\,\mathbf{h}_{k-1} + \mathbf{B}_k\,\mathbf{u}_k + \mathbf{w}_k\ , \tag{16.11}$$

where $\mathbf{w}_k$ is a vector of zero-mean discrete white noise (called system noise) assumed to be white in time and coloured in space: $E[\mathbf{w}_k] = \mathbf{0}$, $E[\mathbf{w}_k\mathbf{w}'_\ell] = \mathbf{0}$ if $k \neq \ell$ and $E[\mathbf{w}_k\mathbf{w}'_k] = \mathbf{Q}_k$, $\mathbf{Q}_k$ being the covariance matrix. It is usually assumed that $\mathbf{w}_k$ is stationary in time (although not necessary and probably not likely; see Senegas et al. (2001)) so that $\mathbf{Q}_k = \mathbf{Q}$). Note that by adding $\mathbf{w}_k$ the vector $\mathbf{h}_k = (H(\mathbf{s}_1, t_k), \ldots, H(\mathbf{s}_N, t_k))'$ becomes stochastic.

At certain time steps, observations are made. If observations $Y(\mathbf{s}_j, t_k), \quad j = 1, \ldots, n_{sk}$ are made at time step k at n_{sk} locations, they can be put in the observations vector $\mathbf{y}_k = (Y(\mathbf{s}_1, t_k), \ldots, Y(\mathbf{s}_{n_{sk}}, t_k))'$. If it is assumed that the observations are linear combinations of the state variables plus some measurement noise, the following measurement equation can be formulated:

$$\mathbf{y}_k = \mathbf{C}_k\,\mathbf{h}_k + \mathbf{v}_k\ , \tag{16.12}$$

where $\mathbf{C}_k$ is the measurement matrix and $\mathbf{v}_k$ a zero mean vector of measurement errors, which may be mutually dependent but are all white in time: $E[\mathbf{v}_k] = \mathbf{0}$, $E[\mathbf{v}_k\mathbf{v}'_\ell] = \mathbf{0}$ if $k \neq \ell$, $E[\mathbf{v}_k\mathbf{v}'_k] = \mathbf{R}_k$, $\mathbf{R}_k$ being the covariance matrix of the measurement errors. The measurement errors are often assumed to be mutually independent, so that $\mathbf{R}_k$ is a diagonal matrix. The measurement errors and the system noise are also mutually independent: $E[\mathbf{w}_k\mathbf{v}'_\ell] = \mathbf{0} \ \ \forall k, \ell$. Note that the observations can be of any type, as long as they are linearly related to the state. If observations are actual measurements at locations of hydraulic head (assuming that they are made at grid nodes only), then $\mathbf{C}_k$ is simply a $n_{sk} \times N$ matrix with for each row a 1 at the location number where an observation is made at time k and zeros everywhere else. However, an observation could also be some (weighted) average value of the state, for instance obtained with geophysical measurements, in which case the row of the corresponding observation in $\mathbf{C}_k$ contains weights for the state nodes that have contributed to the observation and zeros everywhere else. It can thus be seen that $\mathbf{C}_k$ actually represents the sampling locations which defines for each time step where observations are made. If no observations are made at a particular time step, then $\mathbf{C}_k = \mathbf{0}$, the null-matrix.

Equations (16.11) and (16.12) form the basis of our prediction problem, i.e., optimal mapping of h for the space–time coordinates $(\mathbf{s}_i, t_k)$, $i = 1, \ldots, N$, $k = 1, \ldots, K$. To solve this problem, predictions of the groundwater model somehow have to be merged with the observations. Kalman (1960) derived a recursive estimation procedure to achieve this. This procedure, referred to as the Kalman filter, gives best linear unbiased predictions (in least squares sense) of $\mathbf{h}_k$ using all observations up to and including time step k. The Kalman filter is described next. We first define the following variables:
$\tilde{\mathbf{h}}_{k|k}$ is the measurement update, i.e., the unbiased least squares prediction of $\mathbf{h}_k$ using all observations up to and including time step k;
$\tilde{\mathbf{h}}_{k|k-1}$is the time update, i.e., the unbiased least squares prediction of $\mathbf{h}_k$ using all observation up to and including time step $k-1$;

$\mathbf{P}_k$ is the covariance matrix of the error in the measurement update at time k: $E[(\tilde{\mathbf{h}}_{k|k} - \mathbf{h}_k)(\tilde{\mathbf{h}}_{k|k} - \mathbf{h}_k)']$;
$\mathbf{M}_k$ is the covariance matrix of the error in the time update at time k: $E[(\tilde{\mathbf{h}}_{k|k-1} - \mathbf{h}_k)(\tilde{\mathbf{h}}_{k|k-1} - \mathbf{h}_k)']$.

Assuming that the initial conditions $\mathbf{h}_0$ and $\mathbf{P}_0$ are given, the Kalman filter algorithm is given by:

$$\tilde{\mathbf{h}}_{k|k-1} = \mathbf{A}_k \, \tilde{\mathbf{h}}_{k-1|k-1} + \mathbf{B}_k \, \mathbf{u}_k \ , \tag{16.13}$$

$$\mathbf{M}_k = \mathbf{A}_k \, \mathbf{P}_{k-1} \, \mathbf{A}_k' + \mathbf{Q}_k \ , \tag{16.14}$$

$$\mathbf{K}_k = \mathbf{M}_k \, \mathbf{C}_k' \, \left(\mathbf{C}_k \, \mathbf{M}_k \, \mathbf{C}_k' + \mathbf{R}_k\right)^{-1} \ , \tag{16.15}$$

$$\tilde{\mathbf{h}}_{k|k} = \tilde{\mathbf{h}}_{k|k-1} + \mathbf{K}_k \, \left(\mathbf{y}_k - \mathbf{C}_k \, \tilde{\mathbf{h}}_{k|k-1}\right) \ , \tag{16.16}$$

$$\mathbf{P}_k = \left(\mathbf{I} - \mathbf{K}_k \, \mathbf{C}_k\right) \mathbf{M}_k \ , \tag{16.17}$$

where $\mathbf{I}$ is an identity matrix and $\mathbf{K}_k$ a weighting matrix called the 'Kalman gain'.

Equations (16.13) and (16.16) show that the Kalman filter is in fact a predictor–corrector method, where a prediction is made by the mechanistic model (16.13) first. Next, measurements are used to update this prediction using a weighted average (16.16). The magnitude of the update depends on the ratio of the prediction error and the measurement error through (16.15). If the measurement error (through $\mathbf{R}_k$) is very large or the error in the model prediction (through $\mathbf{M}_k$) is small, then $\mathbf{K}_k \to \mathbf{0}$ and no updating will take place. Conversely, if the measurement error is small compared to the model prediction error, updating will be considerable. The updating is performed in a physically consistent way because it depends (through $\mathbf{M}_k$) not only on the correlation structure of noise $\mathbf{Q}_k$ but also on the system matrix $\mathbf{A}_k$, which represents the system dynamics based on physical laws, such as Darcy's law and the conservation of mass in our groundwater example. Such an update could be viewed as creating additional sources/sinks at the nodes where $\mathbf{w}_k$ is defined, which are then propagated through the system by the deterministic model (16.13). For this reason, noise is often put only on the boundary of the model area, thus representing unknown boundary conditions. The temporal persistence of the update depends on the magnitude (i.e., the trace) of the system matrix $\mathbf{A}_k$, while the spatial influence of an observation depends both on the correlation length of $\mathbf{Q}_k$ and the relative magnitude of the off-diagonal elements of $\mathbf{A}_k$.

If both $\mathbf{w}_k$ and $\mathbf{v}_k$ are multivariate Gaussian and stationary in space, the Kalman filter not only gives Best Linear Unbiased Predictions, but also the best unbiased prediction, in which case the time update is the conditional mean $\tilde{\mathbf{h}}_{k|k} = E[\mathbf{h}_k|\mathbf{y}_j, \ldots, \mathbf{y}_k]$ and its covariance the conditional covariance: $\mathbf{P}_k = C[\mathbf{h}_k|\mathbf{y}_j, \ldots, \mathbf{y}_k]$. In this case, $\tilde{\mathbf{h}}_{k|k}$ is also a maximum likelihood prediction.

The Kalman filter was developed in the context of control theory, usually dealing with small state vectors. However, in data assimilation, distributed dynamic models are used, often consisting of thousands of nodes. The Kalman filter then commonly leads to both storage problems (large covariance matrices $\mathbf{M}_k$ and $\mathbf{P}_k$) and computational problems (CPU time) mainly when evaluating the covariance of the time update (16.14). Several adaptations and approximations to the original formulation of the Kalman filter have been developed to solve such problems (e.g., Morf et al., 1974; Heemink and Kloosterhuis, 1990; Evensen and van Leeuwen, 1995; Verlaan and Heemink, 1997; Bierkens et al., 2001; Heemink et al., 2001). A comprehensive overview of methods is given by Verlaan (1998).

The Kalman filter is not only suitable for prediction in a data-assimilation framework, but can also be used together with purely statistical descriptions of $Z(\mathbf{s}, t)$. For instance, Huang and Cressie (1996) and Bierkens et al. (2001) formulated autoregressive statistical space–time models in a state–space framework to apply the Kalman filter for space–time prediction. Wikle and Cressie (1999) used a dimension reduction technique to reduce an autoregressive space–time model to a small number of spatially correlated time series which are then predicted using a Kalman filter. The advantage of these approaches is that they solve some of the dimension problems mentioned above and can therefore be applied to large systems.

The Kalman Filter and Space–Time Sampling

To return to the sampling problem: if the goal is to map the hydraulic head in space and time and an optimal sampling design is sought to achieve this, the numerical groundwater model and the Kalman filter can be used to direct the sampling. For space-time mapping, a static-synchronous design seems advantageous and we will provide a framework for optimizing such a design using the Kalman filter.

If one were to define the quality measure as the sum of all squared prediction errors $\sum_k \sum_i [\tilde{h}(\mathbf{s}_i, t_k) - h(\mathbf{s}_i, t_k)]^2$ then it follows that this quality measure is given by the sum of the traces of $\mathbf{P}_k$:

$$J = \sum_{k=1}^{K} \mathrm{tr}(\mathbf{P}_k) \; . \tag{16.18}$$

For practical reasons, the covariance matrices of system noise and measurement error are usually assumed to be constant in time: $\mathbf{Q}_k = \mathbf{Q}$ and $\mathbf{R}_k = \mathbf{R}$, while the system matrix is usually constant because constant grid sizes and time steps are used: $\mathbf{A}_k = \mathbf{A}$. Given some initial condition $\mathbf{P}_0$, covariance matrices $\mathbf{Q}$ and $\mathbf{R}$ and system matrix $\mathbf{A}$, this quality measure can thus be fully evaluated for some sample $\mathbf{C}_k, \quad k = 1, \ldots, K$ with (16.14), (16.15) and (16.17):

$$\mathbf{M}_k = \mathbf{A}\, \mathbf{P}_{k-1}\, \mathbf{A}' + \mathbf{Q} \; , \tag{16.19}$$

$$\mathbf{K}_k = \mathbf{M}_k \, \mathbf{C}_k' \, (\mathbf{C}_k \, \mathbf{M}_k \, \mathbf{C}_k' + \mathbf{R})^{-1} \ , \tag{16.20}$$

$$\mathbf{P}_k = (\mathbf{I} - \mathbf{K}_k \, \mathbf{C}_k) \, \mathbf{M}_k \ . \tag{16.21}$$

As can be seen, these equations do not depend on the actual predictions, but only on the system and noise characteristics and the sample $\mathbf{C}_k$, $k = 1, \ldots, K$. So, once a calibrated groundwater model is available and the noise statistics have been identified, it is possible to perform monitoring-network optimization, not only by leaving out existing sampling locations and sampling times but also by adding additional sampling locations and sampling times (e.g., Zhou et al., 1991).

Of course, before any sampling optimization can be implemented, it is necessary to identify $\mathbf{A}$, $\mathbf{Q}$ and $\mathbf{R}$. The system matrix $\mathbf{A}$ is built from the physical properties of the subsurface ($S(\mathbf{s})$ and $\mathbf{T}(\mathbf{s})$) and the discretization of the numerical model. Usually, $S(\mathbf{s})$ and $\mathbf{T}(\mathbf{s})$ are determined by making geophysical observations (pumping tests, slug tests) at a number of locations and then interpolated in space by means of splines, kriging, inverse distance or geological stratification. The values can be further adapted by means of calibration of the groundwater model to head measurements from an existing network. The measurement error covariance matrix is often assumed to be diagonal, with the diagonal showing the variances of measurement errors as derived from experience. The greatest problem is the identification of the matrix of system noise $\mathbf{Q}$. As in the case of the system matrix, some regularization is needed. This can for instance be achieved by a continuous covariance function whose parameters are then estimated through a trial and error procedure from the innovations of the Kalman filter $\mathbf{y}_k - \mathbf{C}_k \, \tilde{\mathbf{h}}_{k|k-1}$ when applied to an existing network (cf. van Geer et al., 1991). More advanced methods are possible. For instance, te Stroet (1995) applied inverse estimation theory and adaptive filtering techniques to estimate regularized versions of $\mathbf{A}$ and $\mathbf{Q}$ concurrently.

To illustrate the evolution of the variance of prediction error when using a Kalman filter, we present a similar set-up of a space–time universe as described for the space–time kriging: discretized with 10×10 nodes in space and 10 time steps. To avoid border effects, the universe was placed in the centre of a larger spatio-temporal universe of 30×30 grid nodes in space and 30 time steps. If we assume that the spatial universe is a homogeneous and isotropic porous medium (conductivity and storage coefficient are constant in space and conductivity is constant with direction) with groundwater flow described by (16.9), then, using a finite difference approximation of (16.9), the inverse of the system matrix is equal to the following band matrix:

$$\mathbf{A}^{-1} = \begin{bmatrix} (4/\alpha)+1 & -1/\alpha & 0 & \cdot & 0 & -1/\alpha & 0 & \cdot & \cdot & 0 \\ -1/\alpha & (4/\alpha)+1 & -1/\alpha & & & & -1/\alpha & & & \cdot \\ 0 & -1/\alpha & \cdot & \cdot & & & & \cdot & & \cdot \\ \cdot & & \cdot & \cdot & \cdot & & & & \cdot & 0 \\ 0 & & & \cdot & \cdot & \cdot & & & & -1/\alpha \\ -1/\alpha & & & & \cdot & \cdot & \cdot & & & 0 \\ 0 & -1/\alpha & & & & \cdot & \cdot & \cdot & & \cdot \\ \cdot & & \cdot & & & & \cdot & \cdot & -1/\alpha & 0 \\ \cdot & & & \cdot & & & & -1/\alpha & (4/\alpha)+1 & -1/\alpha \\ 0 & 0 & \cdot & 0 & -1/\alpha & 0 & \cdot & 0 & -1/\alpha & (4/\alpha)+1 \end{bmatrix} \tag{16.22}$$

where α is a dimensionless parameter that describes the characteristic response time:

$$\alpha = \frac{S}{T} \cdot \frac{(\delta s)^2}{\delta t} \,, \tag{16.23}$$

S and T being the storage coefficient and transmissivity of the domain (assumed constant), and δs and δt being the finite difference discretizations in space and time, respectively. Here we have neglected to explicitly account for boundary conditions, so that the model boundaries are in fact no flow. Figure 16.3 shows the results of the evolving prediction-error variance for this system with $\alpha = 1$. The system noise matrix has been modelled with a spherical covariance model with range $a_s = 10\delta s$ and variance $\sigma_Q^2 = 1$, and the matrix of measurement noise with a diagonal with variance $\sigma_R^2 = 0.1$ for every sampling location. Figure 16.3 shows the prediction-error variance for the large domain at two time steps, at a sampling time (upper left) and between sampling times (upper right), as well as the temporal variation at two locations, one location close to a sampling location and one further away. The influence of the boundaries can be clearly seen in the upper right figure. The time series below shows the typical saw-tooth shape of the evolution of the prediction-error variance, first growing to a maximum value and then reduced to a value smaller than the observation error after updating at a sampling time.

The characteristic response time α determines how quick the groundwater system reacts to a change in forcing, such as a change in boundary conditions, precipitation or a change in pumping rate of a groundwater withdrawal. In case α is small, a disturbance travels fast and a change in for instance precipitation rate will quickly lead to a new equilibrium hydraulic head distribution. In case α is large, the system takes a lot of time to reach a new equilibrium after a change. For example, confined aquifers that are covered on top with an impermeable layer have very small storage coefficients and will reach a new head distribution within days or even hours after groundwater pumping is increased. On the other hand, very thick unconfined aquifers with large storage coefficients and transmissivities can still experience years later the influence of a succession of wet or dry years.

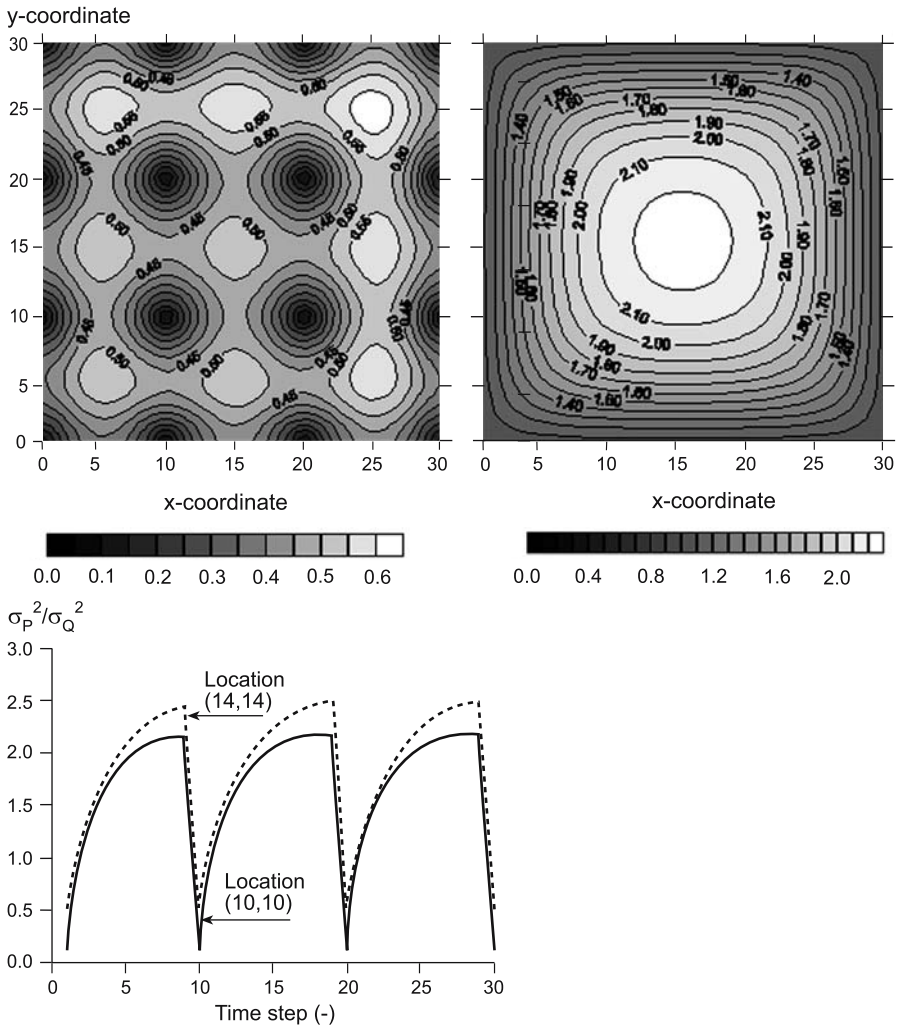


Fig. 16.3. Spatial variation of prediction-error variance-ratio for a linear space–time Kalman filter applied to two-dimensional groundwater flow in a confined aquifer ($\alpha = 1$, $a_s = 10\,\delta s$, $\sigma_Q^2 = 1$ and $\sigma_R^2 = 0.1$) at a sampling time (upper left) and between sampling times (upper right), and time series of prediction-error variance-ratio close to sampling location and distant from sampling locations (lower left)

Figure 16.4 shows that the characteristic response time also influences the accumulation of model errors. The figure shows how the variance of the total prediction error $\sigma_{\mathbf{M}}^2$ develops with time in case at each time a model error is made with a fixed variance $\sigma_{\mathbf{Q}}^2$ and no updating with observations is performed. It can be seen that slow reacting systems tend to accumulate

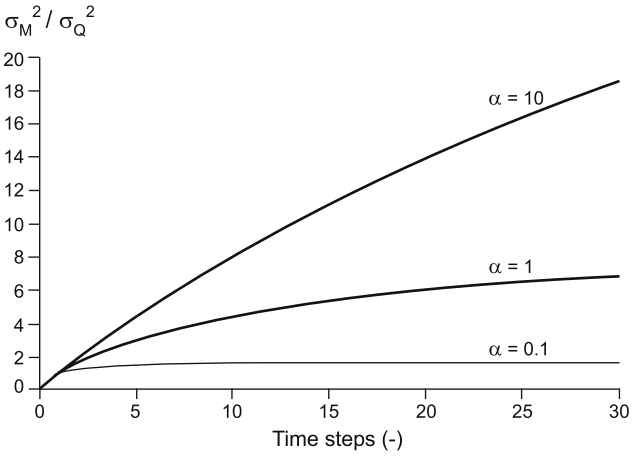


Fig. 16.4. Accumulation of prediction errors in a stochastic dynamic system described by (16.11) with varying response times α and $\mathbf{Q}$ modelled using a spherical covariance function with $a_s = 10\,\delta s$ and variance $\sigma_Q^2 = 1$

errors much more than fast reacting ones; the have a larger 'memory' for past errors, which are thus propagated more strongly in time. It can also be seen that the maximum prediction error is reached much faster for a fast reacting system than for slow reacting one. Clearly, in case of a slow reacting system one achieves an enormous variance reduction with just a few sampling rounds, while in case of fast reacting systems a higher sampling frequency is needed to obtain any improvement at all. The characteristic response time thus determines strongly how a given number of measurements should be optimally distributed in space and time.

The sampling density in space and time not only depends on α, but also on the spatial correlation of the error process in $\mathbf{w}_k$. This can be seen in Fig. 16.5, which shows isolines of the average variance ratio $\overline{\sigma_{\mathrm{kf}}^2}/\sigma_{\mathbf{Q}}^2$ of the central universe of our example for $a_{\mathrm{s}}/\delta s = 10$ and $\alpha = 0.1$ as a function of the sampling density and sampling frequency. Here

$$\overline{\sigma_{\mathrm{kf}}^2} = \frac{1}{NK} \sum_{k=1}^{K} \mathrm{tr}(\mathbf{P}_k)\ , \tag{16.24}$$

where a_s and σ_Q^2 are the parameters used to model covariance matrix $\mathbf{Q}$ with a spherical covariance function. Note that some small corrections have been applied when calculating σ_{kf}^2 for the central universe in order to account for the effect of the boundary conditions (maximum correction up to 5%).

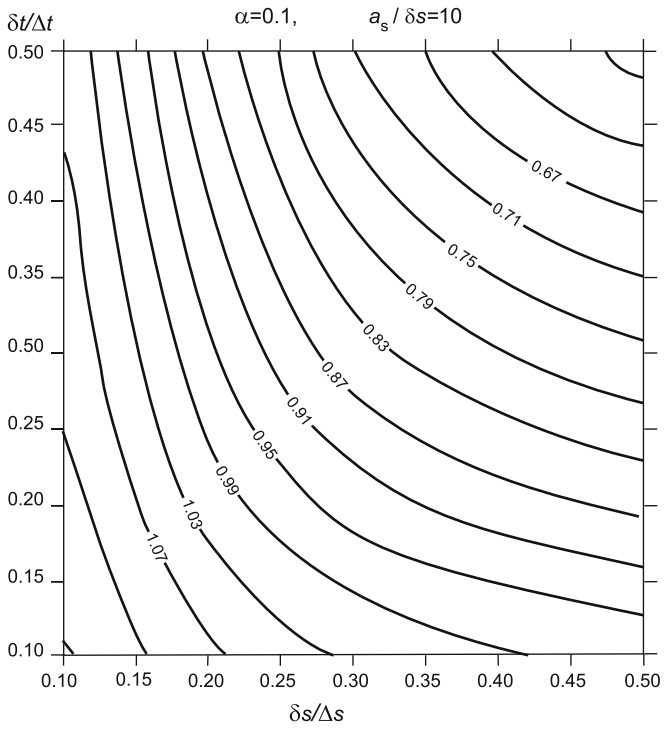


Fig. 16.5. Lines of equal normalized mean prediction-error variance $\overline{\sigma^2_{\mathrm{kf}}}/\sigma^2_{\mathbf{Q}}$ for sampling on a space–time grid and prediction of groundwater heads with a Kalman filter. The covariance model is spherical with parameters a_s and σ^2_Q. Here $a_{\mathrm{s}}/\delta s = 10$ and the system property $\alpha = 0.1$. Figures with other combinations of $a_{\mathrm{s}}/\delta s$ and α are presented in Appendix D.

Figure 16.5 can be used as follows. If the hydraulic parameters S, T are given, as well as discretization $\delta s, \delta t$ (and therefore α), a combination of Δs and Δt can be chosen, with $\Delta s \in \{\delta s, 2\delta s, 3\delta s, \ldots\}$ and $\Delta t \in \{\delta t, 2\delta t, 3\delta t, \ldots\}$, in order to achieve a given required mean variance ratio $\overline{\sigma^2_{\mathrm{kf}}}/\sigma^2_{\mathbf{Q}}$. Figure 16.5 is valid for $a_{\mathrm{s}}/\delta s = 10$ and $\alpha = 0.1$. Appendix D provides a more complete set of figures for various combinations of $a_{\mathrm{s}}/\delta s$ and α. These figures provide a tool that can be used to design samples for mapping hydraulic heads with a Kalman filter. Also, we envisage an even broader use of these diagrams. Many dynamic systems in the earth sciences can be described by an equation that is similar to (16.11), with a system matrix having the structure of (16.22). Space–time samples for such systems could thus be designed using the diagrams in Appendix D.

16.2.5 Kalman Filtering Current Maps

This section deals with designing a spatial sample for mapping the values at the next sampling time, using the sample data at the previous sampling times as prior data. So, contrary to the previous section where a space–time sample is designed for the whole monitoring period before the start of the first sampling round, in this section *spatial* samples are designed sequentially. The sample for the second sampling time is designed after the data of the first sampling time have been collected, etc.

Wikle and Royle (1999) have shown that for Kalman filtering current maps synchronous patterns are generally more efficient than static-synchronous patterns. Synchronous patterns are particularly advantageous if the characteristic response time of the system (i.e., the diagonal of the system matrix $\mathbf{A}_k$) is large (long memory) and the spatial correlation length of the system noise $\mathbf{w}_k$ (off-diagonals of $\mathbf{Q}_k$) is not too large. Of course, such a pattern requires that the costs and time involved in moving sampling locations are not a problem: e.g., sensors for measuring air quality may be easily moved, while the time required to install piezometer tubes measuring groundwater head prohibits moving them around.

Wikle and Royle (1999) discuss also an intermediate case where there are a limited number of sampling locations (e.g., piezometer tubes) and an even smaller number of measurement devices (e.g., groundwater quality sensors) that can be moved freely between the sampling locations at each time step. Such a synchronous pattern is especially suitable if both the costs of installing sampling locations and the observation costs contribute significantly to the total monitoring costs.

16.2.6 Models and Predictors in Space–Time

The many methods of modelling spatio-temporal variation reported in the literature differ greatly with respect to the modelling of $Z(\mathbf{s}, t)$ and the predictor used. Here, we present a taxonomy of space–time methods, restricting ourselves to prediction at points. The proposed classification is partly based on an extensive survey of geostatistical models given by Kyriakidis and Journel (1999) and partly on papers by McLauglin (1995) and Houser et al. (1998). The suitability of a method for use in sampling design depends on whether a quality measure can be constructed based on the space–time coordinates of projected sampling events only, i.e., without the sample data. Methods that meet this requirement are marked by asterisks. All other methods require a stochastic simulation approach to be used for sampling. The suitability of a space–time prediction method for sampling also depends on the complexity of the quality measure, i.e., the way $Z(\mathbf{s}, t)$ is modelled and the form of the predictor. Methods that are complex are marked with the symbol ^{c}. Note that the latter is only our personal opinion, not a fact. Thus, methods marked with an asterisk only are, in our opinion, most easily adapted for use in sampling.

We stress again that in model-based sampling, the prediction method and the sampling design are intertwined: the sampling design is only 'optimal' if, after sampling, the inference is performed with the prediction method associated with the quality measure.

Statistical Models

Statistical models are entirely based on the data. Statistical models may be purely empirical, as in kriging or regression. However, models that are built from preconceived ideas about the spatio-temporal variation (e.g., data-based mechanistic modelling as described in Young and Beven (1994)), but do not explicitly include physical laws are also classified as statistical models. If the deterministic part of the model is a partial-differential equation and model prediction is performed in a physically consistent manner (conserving mass, momentum or energy) it is classified as a data-assimilation method. Statistical models of space–time can be classified into the following categories (Kyriakidis and Journel, 1999; Bierkens, 2001).

*1. Multiple Time-Series Models**

Here, time series are modelled at the sampling locations, using either temporal geostatistical models or Box–Jenkins type models (Box and Jenkins, 1976). Prediction is restricted to the sampling locations (where the time series have been observed) and is applied at non-observed times. In Box–Jenkins modelling, prediction is performed by using the multivariate time-series model in forecasting mode (Bras et al., 1983) or Kalman filtering (Priestley, 1981). In a geostatistical framework, co-kriging is used to predict time series at unobserved times (Rouhani and Wackernagel, 1990).

2. Regionalized Time-Series Models

Multiple time-series modelling restricts prediction to the sampling locations. By contrast, adaptation of the concept of multiple time series, called 'regionalized time-series models', enables prediction at unobserved locations. To this end, stochastic time-series models are defined at a grid. First, parameters of the time-series model are inferred at the sampling locations. Next, these parameters are interpolated onto the grid. Using the interpolated parameters, the time-series model is used to predict time series at every grid node. Differences between methods concentrate on the type of time series model used, the type of method used to interpolate the parameters and the prediction method. Examples of methods used are:

- (Huang and Cressie, 1996)*. These authors assume an autoregressive process of order p (AR(p)) at each location. Autoregressive parameters are assumed to be the same for all locations and estimated from spatially averaged temporal covariances obtained from all sampling locations. The

noise process driving the AR(p) process is white in time and spatially coloured. Predictions at unobserved locations for a given time step t_k are made with a space–time Kalman filter using all available observations up to and including time step t_k.

- Knotters and Bierkens (2001). The time-series model is a simple transfer-function noise model (called ARX model). Parameters are interpolated with ordinary kriging and kriging with an external drift. Predictions are made by running the deterministic part of the ARX model.
- Bierkens et al. (2001)*; Knotters and Bierkens (2002)*. This approach builds on the approach by Knotters and Bierkens (2001). The regionalized time series at the grid nodes are embedded in a space–time Kalman filter for space–time prediction, where the noise process driving the ARX model is white in time, spatially coloured and heteroscedastic. Apart from spatially varying time-series parameters and spatially varying noise variance, the approach by Bierkens et al. (2001) and Knotters and Bierkens (2001) is equivalent to the method described by Wikle and Royle (1999). Through the use of the multiple time-series concept, these approaches are particularly suitable for predicting onto large grids, despite the use of the Kalman filter.
- Angulo et al. (1998)*,c. These authors present a combined geostatistical and discrete time-series approach. The time-series models consist of a temporal trend and an ARMA model. The trend parameters and the ARMA parameters are interpolated using inverse distance weighting. Prediction is done in two steps: (1) spatial interpolation of noise time series obtained at the sampling locations; (2) use of the interpolated noise at the unobserved locations in the (interpolated) ARMA model for predicting residuals, which are then added to the (interpolated) trend. A similar approach was used by Meiring et al. (1998)*,c, adding the option to model a spatially and temporally nonstationary noise process.
- Kyriakidis and Journel (2001)c use a conditional simulation approach, in which trend parameters are determined at locations with time series. Full spatial coverage of trend parameters is achieved through geostatistical conditional simulation (after transformation of the trend parameters to independent variates using principal component analysis). The residual process is modelled as a collection of spatially correlated time series, whose covariance parameters are again interpolated with conditional stochastic simulation. The resulting spatio-temporal noise is modelled by simulating at each location Gaussian deviates that are white in time but spatially coloured and conditioned on normal scores of residuals.

3. Multiple Spatial Random Fields

The universe is only considered at a finite number of time sections: $\mathbf{s} \in \mathcal{S} \subset \mathbb{R}^k$ and $t \in \{t_1, t_2, \ldots, t_N,\}$ and the stochastic model consists of a collection of multiple mutually dependent spatial random fields. Prediction is performed by:

- co-kriging multiple time sections*. Examples can be found in Bogaert (1996) and Bogaert and Christakos (1997b);
- (co-)kriging differences between two time sections*. See, for instance, Sølna and Switzer (1996) and Papritz and Flühler (1994).

4. Space–Time Random Fields

Here, $z(\mathbf{s}, t)$ is modelled with a complete space–time stochastic function $Z(\mathbf{s}, t)$, $\mathbf{s} \in \mathcal{S} \subset \mathbb{R}^k$, $t \in \mathcal{T} \subset \mathbb{R}$. Two ways of obtaining predictions are commonly used in the literature.

- Space–time kriging*. Time is treated as an additional coordinate and kriging is performed using data from different locations and different time steps. In certain cases, nonstationarity in time is accounted for by universal kriging. Examples are given in papers by Rouhani and Hall (1989), Bogaert (1996), Bogaert and Christakos (1997a), Heuvelink et al. (1997) and Snepvangers et al. (2003). As described in the previous section, the disadvantage of these methods is that particular features of time (its unidirectionality, causality) are not taken into account and that nonstationarity in time can lead to complicated models and inference problems. Also, notice again that under certain circumstances space–time kriging and co-kriging are equivalent when considering a finite number of fixed time sections (Bogaert, 1996).
- Space–time Kalman filtering*. Here, a space–time model is formulated as a spatially continuous and temporally autoregressive process, where time is considered at discrete time steps only. The advantage is that the unidirectional nature of time is embedded in the model formulation[1]. An example of this approach is provided in the paper by Wikle and Cressie (1999). A special feature of their method is that it can be applied to large problems by using dimension reduction techniques (see also the previous section).

Data Assimilation

Readers not familiar with the Kalman filter are advised to read the example in Sect. 16.2.4 first. McLauglin (1995) provides the following description of data assimilation:

"*The objective of data assimilation is to provide physically consistent estimates of spatially distributed environmental variables. These estimates are derived from scattered observations taken at a limited number of times and locations, supplemented by any additional 'prior information' that may guide the*

[1] In this sense, the stochastic model used is close to the multiple random field approach, albeit with a specific unidirectional Markov-type covariance structure between time steps.

estimation process. The terms 'spatially distributed' and 'physically consistent' are important here. Spatially distributed estimates are defined everywhere, not just at sampling locations, and they can usually be viewed as continuous functions of time and location. Physically consistent estimates are compatible with physical laws such as conservation of mass, momentum, and energy. This is reflected in both their spatial and temporal (or dynamic) structure."

Thus, the distinctive feature of data-assimilation methods is that space–time predictions are made in a physically consistent way, that is, making sure that the predictions obey certain physical laws. In data-assimilation methods, the space–time Stochastic Function $Z(\mathbf{s}, t)$ is generally described with a stochastic partial-differential equation. When discretized in space (on a grid size N) and time (K time steps), this stochastic partial differential equation can be represented by the following discrete stochastic system in state space:

$$\mathbf{z}_k = \varphi_k(\mathbf{z}_{k-1}, \mathbf{a}_k, \mathbf{u}_k) + \mathbf{w}_k \ , \tag{16.25}$$

where $\mathbf{z}_k = (Z(\mathbf{s}_1, t_k), \ldots, Z(\mathbf{s}_N, t_k))'$, $k = 1, \ldots, K$, is the state vector, $\mathbf{u}_k$ is a vector of (possibly random) inputs such as boundary conditions sinks and sources, $\mathbf{a}_k$ is a vector of (possibly random) model parameters, $\mathbf{w}_k$ is a vector with system noise representing errors due to inadequacy of the model description and $\varphi_k(\cdot)$ is a (generally non-linear) vector function relating $\mathbf{z}_k$ to $\mathbf{z}_{k-1}$, $\mathbf{a}_k$, $\mathbf{u}_k$ and $\mathbf{w}_k$. Note that if φ_k is linear and $\mathbf{a}_k$ is deterministic, (16.25) reduces to (16.11).

Apart from an equation describing the physics behind $Z(\mathbf{s}, t)$, there are also observations $y(\mathbf{s}_j, t_k)$, $j = 1, \ldots, n_k$ that tell us something about $z(\mathbf{s}, t)$. If it is possible to find some relation between observations and the state, a general equation describing this relation looks like:

$$\mathbf{y}_k = \mathbf{h}_k(\mathbf{z}_k) + \mathbf{v}_k \ , \tag{16.26}$$

where $\mathbf{y}_k = (Y(\mathbf{s}_1, t_k), \ldots, Y(\mathbf{s}_{n_k}, t_k))'$ is a vector of observations at time k, $\mathbf{v}_k$ is a vector with observation errors and $\mathbf{h}_k(\cdot)$ is a (generally non-linear) vector function relating the observations to the state. If the observations are the actual state variables or weighted averages of these variables, then $\mathbf{h}_k$ is linear and (16.26) reduces to (16.12). However, if our observations are, e.g., radar reflectivity measurements which are related to soil moisture (our state variable), then $\mathbf{h}_k(\cdot)$ may represent a complex non-linear reflectivity model (e.g., Hoeben and Troch, 2000). Equations (16.25) and (16.26) form the basis of all data assimilation methods aiming to the optimal merger of predictions from physical models with data in order to arrive at physically consistent space–time predictions. An overview of data-assimilation methods is presented below.

1. Heuristic (Non-Optimal) Methods

There are various heuristic methods that can be used to assimilate data into physical–mechanistic models. An overview of such methods applied to the

modelling of soil moisture content is given by Houser et al. (1998). These methods are called sub-optimal because the updating of model predictions with data is not done in order to minimize some error-related quality measure. Instead, updating is done in a heuristic manner. Therefore, these methods are not very suitable for sampling design. Examples of such methods are:

- 'Direct insertion', where observations are directly inserted in the state vector, and no measurement error is taken into account. It is assumed that a positive influence of the inserted correction will propagate to the other state elements through the system dynamics.
- 'Statistical correction' is very much like direct insertion, except that observation errors can be taken into account by inserting a weighted average of the observation and the model prediction.
- 'Newtonian nudging' (Stauffer and Seaman, 1990) is a technique in which the physical–mechanistic model is augmented with an additional source term that is adjusted while advancing the model in time in order to 'nudge' the model predictions towards the observations. Observation errors can be taken into account.

2. Optimal and Approximately Optimal Methods

In optimal data-assimilation methods, the observations and the state are merged in such a way that some error-related quality measure is minimized. For instance, a Kalman filter minimizes the mean squared difference between the actual (but unknown) state and the model predictions. Many of the optimal methods provide quality measures that are a function of the coordinates only and are therefore suitable for spatio-temporal sampling design. Optimal methods can be divided into filters and smoothers.

2a. Filtering

In filtering, observations are used that are made at times up to and including the time step at which the prediction is made: $\mathbf{y}_1, \ldots, \mathbf{y}_k$ to predict $\mathbf{z}_k$. Filters are very suitable if the aim is to forecast in time while observations are taken along the way. Optimal filters thus extrapolate in time. However, as filter algorithms are easier to implement than smoothing algorithms, they are often used for prediction in space and time when all observations have already been made. Distinction can be made between linear and non-linear systems.

- Linear systems*. If the stochastic system and the measurement equation are both linear, the linear Kalman filter can be used (van Geer et al., 1991; Zhou et al., 1991). In case a system or the measurement equation is slightly non-linear, linearized versions can be formulated in order to apply the Kalman filter (e.g., Walker and Houser, 2001).
- Non-linear systems[c]. If the stochastic system or the measurement equation are non-linear and linearization is not an option (i.e., gives poor results), an extended Kalman filter can be used. The extended Kalman filter uses

the non-linear state and measurement equations, but it uses a linearized version of these equations to calculate the covariance matrix of the time update and the Kalman gain. Linearization is done around the optimal prediction of the previous time step. Because the linearization requires actual predictions, the extended Kalman filter is not very suitable for sampling, except for thinning out. For examples of the extended Kalman filter, we refer to Evensen (1992) and Hoeben and Troch (2000). Linearization can lead to system instability (prediction blow-up), especially in the case of few or imprecise observations. The ensemble Kalman filter (Evensen, 1994) is a solution to this problem. Instead of the 'optimal trajectory', an ensemble of trajectories (realizations of $\mathbf{z}_k$ by simulating error vectors $\mathbf{w}_k$ and possibly $\mathbf{a}_k$ and $\mathbf{u}_k$ in (16.25)) are propagated with the model. The covariance of the time update is obtained by estimating it directly from the realizations. The covariance matrix is used to calculate the Kalman gain, which in turn is used to update each realization with the observations. The updated realizations are each propagated with the model (16.25) and so on. The ensemble Kalman filter is increasingly used in data assimilation. A recent addition to non-linear filtering is the particle filter (or sequential importance resampling filter; van Leeuwen (2003)). The particle filter is much simpler to implement than the ensemble Kalman filter and has the added value that it also works for variables that have an extremely non-Gaussian (e.g., multimodal) distribution. It is however less suitable if state vectors are large. Unfortunately, non-linear filtering methods are not very suitable for sampling design (except for thinning out), as they also require actual observations.

2b. Smoothing

If all observations have already been made, then it makes sense to predict $\mathbf{z}_k$ using all observations made before and after time step k: $\mathbf{y}_1, \ldots, \mathbf{y}_k, \ldots, \mathbf{y}_K$. There are two ways of doing this: recursively or off-line.

- Recursively: the Kalman smoother^{*,c}. The algorithm that can provide optimal predictions of $\mathbf{z}_k$ using all observations made before and after time step k in a recursive manner is called the 'Kalman smoother'. It basically consists of running the Kalman filter forward and backward in time, and optimally averaging the predictions obtained with the forward and the backward filter for each time step. We refer to Schweppe (1973) for the algorithm and to Zhang et al. (1999) for a recent application. Like the Kalman filter, the Kalman smoother requires linear state equations and measurement equations.
- Off-line: variational data assimilation^{*,c}. Suppose that the uncertainty in the predictions is caused by an error in the parameters $\mathbf{a} - \tilde{\mathbf{a}}$, with $\tilde{\mathbf{a}}$ some prior estimate of the parameters with covariance matrix $\mathbf{C}_\mathbf{a}$, initial conditions $\mathbf{z}_0 - \tilde{\mathbf{z}}_0$ with covariance matrix $\mathbf{P}_0$, measurement errors $\mathbf{v}_k$ with covariance matrix $\mathbf{R}_k$, and a model error (system noise) $\mathbf{w}_k$ with covariance

matrix $\mathbf{Q}_k$. The goal is then to minimize a quality measure of the following form:

$$
\begin{aligned}
J = & \sum_{k=1}^{K} [\mathbf{y}_k - \mathbf{h}_k(\mathbf{z}_k)]' \, \mathbf{R}_k^{-1} \, [\mathbf{y}_k - \mathbf{h}_k(\mathbf{z}_k)] + \sum_{k=1}^{K} \mathbf{w}_k' \, \mathbf{Q}_k^{-1} \, \mathbf{w}_k \\
& + [\mathbf{z}_0 - \tilde{\mathbf{z}}_0]' \, \mathbf{P}_0^{-1} \, [\mathbf{z}_0 - \tilde{\mathbf{z}}_0] + [\mathbf{a}_0 - \tilde{\mathbf{a}}_0]' \, \mathbf{C}_{\mathbf{a}}^{-1} \, [\mathbf{a}_0 - \tilde{\mathbf{a}}_0] \\
& + 2 \sum_{k=1}^{K} \lambda_k \, [\mathbf{z}_k - \varphi_k(\mathbf{z}_{k-1}, \mathbf{a}_k, \mathbf{u}_k) - \mathbf{w}_k]' \cdot \mathbf{1}
\end{aligned}
\tag{16.27}
$$

where the first term represents measurement error, the second the model error, the third term the error in initial conditions and the fourth the error in our initial estimate of the parameters. Errors in boundary conditions and source terms $\mathbf{u}_k$ could be added in a similar way. In this formulation, the goal is to estimate both the model parameters and initial conditions and to predict the state optimally. Thus, (16.27) is also suitable for inverse estimation of parameters (Valstar, 2001). The fifth term (with $\mathbf{1}$ a vector of ones) is added to make sure that the optimal predictions obey the physical laws as described by the model. It is a constraint very much like that added in the quality measure that is minimized with ordinary kriging (see Appendix B), i.e., to ensure that the predictor is unbiased. The vector λ_k is a vector of Lagrange multipliers. However, the fifth term is a weak constraint: the prediction does not have to be exactly equal to a solution of $\mathbf{z}_k = \varphi_k(\mathbf{z}_{k-1}, \mathbf{a}_k, \mathbf{u}_k)$ because of the added model error $\mathbf{w}_k$. To minimize (16.27), so called 'variational methods' are used to derive explicit equations for its minimum: the so called Euler–Lagrange equations. These equations are solved using all available measurements simultaneously, so that variational data assimilation is an off-line method, not a recursive one. To solve the Euler–Lagrange equations , we also need the vector λ_k for every time step. For this we need to derive a so called 'adjoint model' which is similar to our model $\varphi_k(\cdot)$ but using as its inputs the differences $\mathbf{y}_k - \mathbf{h}_k(\tilde{\mathbf{z}}_k)$ between model predictions and observations. The Euler–Lagrange equations must be solved in an iterative manner, requiring several runs of the mechanistic model $\varphi_k(\cdot)$ as well as the adjoint model. We refer to Reichle et al. (2001) for a description of the method as well as an illustrative example. Standard references for descriptions of this method are Daley (1991), Bennet (1992) and McLauglin (1995).

From this description it is clear that variational data assimilation is not easy to apply. In particular, the derivation and implementation of an adjoint model from a numerical model code of some distributed mechanistic model requires much effort and time. However, the method is very popular for modelling large-dimensional systems, such as those used in numerical weather prediction[2], because it does not require propagation and storage of large covariance matrices. The method can be used for both linear and non-linear systems (in

[2] In numerical weather prediction, variational data assimilation is used to re-initialize the state (e.g., pressure, temperature, air moisture content) using all

the case of a linear system, and using only the first, second and fifth terms, variational data assimilation is mathematically equivalent to fixed interval Kalman smoothing). For linear systems, it is possible to also derive an expression for the covariance matrix of the prediction error, which is a function of the space–time sample positions only. This makes variational data assimilation suitable for space–time sampling design. It can also be used for non-linear systems by using an approximation of the covariance matrix of the prediction error obtained by linearization around some nominal value of $\mathbf{h}_k$.

We end this description of optimal data-assimilation methods by noting that there are recursive algorithms (both filtering and smoothing) which are approximately optimal but very efficient in terms of computer storage as well as CPU demand (see Verlaan (1998) for an overview). These algorithms can be used in large-dimensional systems, but they do not require the tedious derivation and implementation of an adjoint state model. Examples of such approximate methods can be found in Verlaan and Heemink (1997)[*,c], Zhang et al. (1999)[*,c] and Heemink (2000)[*,c].

---

the observations available. A similar equation as (16.27) is used, using only the first, third and fifth terms without the model error (hard constraint). The goal is to obtain better forecasts by improving initial conditions: in weather prediction, initial conditions are by far the largest source of uncertainty due to chaotic behaviour

Appendices

A

Optimization Algorithms for Model-Based Sampling

The aim of optimization under model-based sampling is to find the sample size n and the sampling events $\{(\mathbf{s}_1, t_1), (\mathbf{s}_2, t_2), \ldots, (\mathbf{s}_n, t_n)\}$ (assuming sampling in space and time at point support) such that some quality measure is optimized (minimized or maximized). In some situations the sample size is given, usually because of a fixed budget. In that case the problem is restricted to finding the sampling events that optimize the quality measure. In other situations the problem may be to find a sample such that the quality measure is smaller than a certain value (e.g. in testing problems) or the sample size may be part of the quality measure itself, e.g. if the quality measure contains a cost function. In these situations both the sample size n and the sampling events have to be optimized. The following description of optimization methods only focuses on the first case: the sample size n is known and the sampling events have to be found such that the quality measure is optimized. The sample size can be included in the optimization quite easily by running the optimization algorithms repeatedly with increasing sample size until the optimized quality measure becomes smaller (larger) than the target value.

The algorithms described below all assume that the sampling events can be chosen freely within the universe. Optimization of sampling grids is not treated here. We refer to McBratney et al. (1981) for optimizing spatial grids, and to Zhou et al. (1991) and Sect. 11.3 for optimizing the sampling interval in time. Also, in the following description we assume that we are dealing with a quality measure that has to be minimized. Of course, the same algorithms can be used to maximize quality measures, just by rewriting the minimization problem into a maximization problem.

When looking for algorithms to optimize a sample of size n, three types of algorithms come to mind. These are described in the next sections.

A.1 Standard Methods for Minimizing Functions

If the sample size is small, standard methods can be used to minimize the quality measure subject to the sampling events $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n\}$. Gradient-based methods, such as Conjugate Gradient methods, are usually fastest. This, however, requires the partial derivatives of the quality measure with respect to the coordinates. If it is not possible to analytically calculate these derivatives, a numerical approximation using finite differences may be better used. Alternatively methods can be used that do not require gradients, such as the Downhill Simplex method. Optimization algorithms and programs are for instance given in Press et al. (1988).

A.2 Simulated Annealing

Simulated annealing is a combinatorial optimization algorithm in which a sequence of combinations is generated by changing the previous combination slightly and randomly. Each time a combination is generated, the quality measure is evaluated and compared with the quality measure value of the previous combination. The new combination is always accepted if the quality measure value is decreased by the change. A distinctive property of the annealing algorithm is that it accepts also some changes that increase the quality measure value. This is to avoid being trapped at local minima. The probability of accepting a worse combination is given by

$$P = e^{\frac{-\Delta J}{T}}, \tag{A.1}$$

where ΔJ is the increase of the quality measure (objective function) and T is a control parameter which, by analogy with the original application of simulating the cooling of a metal into a minimum energy crystalline structure, is known as the system temperature. This temperature is lowered stepwise during the optimization. Equation (A.1) shows that, given the temperature, the larger the increase of the quality measure value, the smaller the acceptance probability. The temperature remains constant during a fixed number of transitions (a Markov chain), and is then lowered, reducing the probability of accepting changes that increase J (given ΔJ). The simplest and most commonly used cooling scheme is the exponential cooling scheme:

$$T_{k+1} = \alpha T_k, \tag{A.2}$$

where α is a constant, close to and smaller than 1. Besides this parameter, there are three more parameters to be specified by the user:

- the initial temperature T_0;
- the final temperature or stopping quality measure;
- the length of the Markov chains.

A suitable initial temperature results in an average probability of accepting an increase of the quality measure value of 0.8 (Kirpatrick, 1984). This initial temperature can be calculated by performing an initial search in which all increases are accepted. From the result the average of objective-function increases $\overline{\Delta J^+}$ is calculated. The initial temperature is then given by

$$T_0 = \frac{\overline{\Delta J^+}}{\ln(\chi_0)}, \tag{A.3}$$

where χ_0 is the average increase acceptance probability.

The final temperature is not directly specified, but indirectly via the number of Markov chains or the number of combinations to be evaluated. Alternatively, a stopping criterion may be specified for lack of progress. A commonly used definition for lack of progress is no improvement of the quality measure within a Markov chain or an average increase acceptance probability falling below a given small value χ_f. The length of the Markov chains, i.e., the number of combinations generated at one temperature, can be chosen as fixed, i.e., independent of the Markov chain. Alternatively, one may specify a minimum number of acceptances for each Markov chain. In this case the length of the Markov chains increases with the chain number, because the average increase acceptance probability decreases.

Sacks and Schiller (1988) showed how the simulated annealing algorithm can be used for optimizing sampling patterns in situations where observations are spatially dependent. Van Groenigen and Stein (1998) modified the simulated annealing algorithm by changing the solution generator, i.e., the mechanism to generate a new solution by a random perturbation of one of the variables of the previous solution. This is done by moving one randomly selected sampling location over vector $\mathbf{h}$, with a random direction, and a random length with probability distribution Uniform(0,$h_{\max}$). Initially, $h_{\max}$ is half the length of the target area, and decreases with the chain-number.

A.3 Genetic Algorithms

Genetic Algorithms (Gershenfeld, 1999) are a third way of optimizing large samples. They start with defining not one initial sample, but a large ensemble (say 100), called 'population'. For each member of the population the quality measure is calculated. The value of the quality measure is called the fitness. The smaller the value the fitter the population member. Next three things may happen:

1. *Reproduction.* A new generation of the same number of members is created. Depending on the fitness a population member may have a larger or smaller chance of reproduction, and therefore many duplicates or few or even no duplicates. The way the probability of reproduction is linked to the fitness determines the selectivity of the scheme. This has the same

function as the cooling down scheme in annealing. If the selectivity is high, than only a few population members get all the offspring and the danger of ending up in a local minimum is large. If selectivity is low all members procreate and evolution (to the best solution) is slow.

2. *Crossover.* This is not always included in genetic algorithms. Here members of the population can share parameter sets. After being randomly chosen, usually according to a probability related to their fitness, two parents give random subsets of there respective parameters to their offspring: in our case the children share for instance $n/2$ randomly chosen sampling locations from one parent and $n/2$ from the other.
3. *Mutation.* In this step, parameters are randomly changed. So, in this case, for some randomly chosen members the positions of one or more points are (randomly or purposively) changed.

These steps are repeated until no significant changes in average fitness occur between generations. Genetic algorithms thus yield a population of samples. From this one can choose the fittest one, but it may also be the case that several samples are equally fit and therefore eligible. The versatility of genetic algorithms is also a disadvantage. The user must make many choices: the relation between fitness and probability of reproduction (selectivity), if and how crossover is implemented and how mutations are performed, i.e., which sampling locations and how they are shifted. In this way there is an almost infinite number of variants of the genetic algorithm that can be thought of. It is not clear from the start which variant would be most successful in finding the true but unknown global minimum. See Cieniawski et al. (1995) for an application of genetic algorithms in monitoring.

The shifts of coordinates of sampling locations that are proposed in the simulated annealing and the genetic algorithms can be performed in a random manner or in a purposive manner. Random shifts are of course most easily implemented, but generally lead to slow convergence, especially when getting close to a minimum. To speed up the convergence, shifts can be made purposively, using standard optimization methods such as the Conjugate Gradient or the Downhill Simplex method.

B

Kriging

Kriging or geostatistical interpolation aims at predicting the unknown value of Z at a non-observed location $\mathbf{s}_0$ using the values z_i observed at surrounding locations. To do this we use a Stochastic Function (SF) as a model of spatial variation so that the actual but unknown value $z(\mathbf{s}_0)$ and the values at the surrounding locations $z(\mathbf{s}_i)$, $i = 1, \ldots, n$ are spatially dependent random variables (RVs) $Z(\mathbf{s}_0)$ and $Z(\mathbf{s}_i)$, $i = 1, \ldots, n$. There exist a large number of different kriging methods. For an introduction to stochastic functions and kriging we refer to Isaaks and Srivastava (1989). A complete overview of all types of kriging can be found in Goovaerts (1997). Deutsch and Journel (1998) and Pebesma (2004) provide kriging software.

B.1 Simple Kriging

B.1.1 Theory

The most elementary of kriging methods is called simple kriging and is treated here. Simple kriging is based on a SF that is second-order stationary . A second-order stationary SF has the following properties

$$E[Z(\mathbf{s})] = \mu = \text{ constant} \tag{B.1}$$

$$V[Z(\mathbf{s})] = E[Z(\mathbf{s}) - \mu]^2 = \sigma^2 = \text{ constant (and finite)} \tag{B.2}$$

$$C[Z(\mathbf{s}_1), Z(\mathbf{s}_2)] = E[\{Z(\mathbf{s}_1) - \mu\}\{Z(\mathbf{s}_2) - \mu\}] = C(\mathbf{s}_2 - \mathbf{s}_1) = C(\mathbf{h}_{12}) \tag{B.3}$$

These properties show that in case of second-order stationarity, the mean and variance are constant, the variance has a finite value and the covariance is not dependent on the absolute locations $\mathbf{s}_1$ and $\mathbf{s}_2$, but only on the separation vector $\mathbf{h} = \mathbf{s}_2 - \mathbf{s}_1$. The function $C(\mathbf{h})$ is called the covariance function. Furthermore we have $C(0) = \sigma^2$. In Sect. 9.4 it is shown how covariance functions can be estimated from data.

Simple kriging is the appropriate kriging method if the SF is second-order stationary and the mean of the SF $E[Z(\mathbf{s})] = \mu$ is known without error. With simple kriging a predictor $\widetilde{Z}(\mathbf{s}_0)$ is sought that

1. is a linear function of the surrounding data,
2. is unbiased: $E[\widetilde{Z}(\mathbf{s}_0) - Z(\mathbf{s}_0)] = 0$,
3. and has the smallest possible error, i.e., the variance of $\widetilde{Z}(\mathbf{s}_0) - Z(\mathbf{s}_0)$ is minimal.

A linear and unbiased predictor is obtained when considering the following weighted average of deviations from the mean:

$$\widetilde{Z}(\mathbf{s}_0) = \mu + \sum_{i=1}^{n} \lambda_i \left[Z(\mathbf{s}_i) - \mu \right] , \tag{B.4}$$

with $Z(\mathbf{s}_i)$ the values of $Z(\mathbf{s})$ at the surrounding sampling locations. Usually, not all sampling locations are included in the predictor, but only a limited number within a given search neighbourhood. Predictor (B.4) is unbiased by definition:

$$\begin{aligned} E[\widetilde{Z}(\mathbf{s}_0) - Z(\mathbf{s}_0)] &= \mu + \sum_{i=1}^{n} \lambda_i \, E[Z(\mathbf{s}_i) - \mu] - E[Z(\mathbf{s}_0)] \\ &= \mu + \sum_{i=1}^{n} \lambda_i \left[\mu - \mu \right] - \mu = 0 \, . \end{aligned} \tag{B.5}$$

The weights λ_i should be chosen such that the prediction-error variance is minimal. The variance of the prediction error (simple kriging variance) can be written as

$$\begin{aligned} V[\widetilde{Z}(\mathbf{s}_0) - Z(\mathbf{s}_0)] = E[\widetilde{Z}(\mathbf{s}_0) - Z(\mathbf{s}_0)]^2 = E\left[\sum_{i=1}^{n} \lambda_i \left[Z(\mathbf{s}_i) - \mu \right] - \left[Z(\mathbf{s}_0) - \mu \right] \right]^2 = \\ \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \, \lambda_j \, E\{ [Z(\mathbf{s}_i) - \mu][Z(\mathbf{s}_j) - \mu] \} - 2 \sum_{i=1}^{n} \lambda_i \, E[\{ Z(\mathbf{s}_i) - \mu \}\{ Z(\mathbf{s}_0) - \mu \}] \\ + E[Z(\mathbf{s}_0) - \mu]^2 \, . \end{aligned} \tag{B.6}$$

Using the definition of the covariance of a second-order stationary SF $E[\{Z(\mathbf{s}_i) - \mu\}\{Z(\mathbf{s}_j) - \mu\}] = C(\mathbf{h}_{ij})$ and $C(0) = \sigma^2$, we obtain for the variance of the prediction error:

$$V[\widetilde{Z}(\mathbf{s}_0) - Z(\mathbf{s}_0)] = \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \, \lambda_j \, C(\mathbf{h}_{ij}) - 2 \sum_{i=1}^{n} \lambda_i \, C(\mathbf{h}_{i0}) + \sigma^2 \, . \tag{B.7}$$

To obtain the minimum of (B.7) we have to equate all its partial derivatives with respect to the λ_i to zero:

$$\frac{\partial}{\partial \lambda_i} V[\widetilde{Z}(\mathbf{s}_0) - Z(\mathbf{s}_0)] = 2 \sum_{j=1}^{n} \lambda_j \, C(\mathbf{h}_{ij}) - 2C(\mathbf{h}_{i0}) = 0 \qquad i = 1, \ldots, n \; . \tag{B.8}$$

This results in the following system of n equations, referred to as the 'simple kriging system':

$$\sum_{j=1}^{n} \lambda_j \, C(\mathbf{h}_{ij}) = C(\mathbf{h}_{i0}) \qquad\qquad i = 1, \ldots, n \; . \tag{B.9}$$

The n unknown values λ_i can be uniquely solved from these n equations if all the $\mathbf{s}_i$ are different. The predictor (B.4) with the λ_i found from solving (B.9) is the one with the minimum prediction-error variance. This variance can be calculated using (B.7). However, it can be shown (e.g. de Marsily, 1986, p. 290) that the simple kriging variance can be written in a simpler form as:

$$V[\widetilde{Z}(\mathbf{s}_0) - Z(\mathbf{s}_0)] = \sigma^2 - \sum_{i=1}^{n} \lambda_i \, C(\mathbf{h}_{i0}) \; . \tag{B.10}$$

The error variance shows very nicely how kriging takes advantage of the spatial dependence of $Z(\mathbf{s}_i)$. If only the marginal probability distribution had been estimated from the data and the spatial coordinates had not been taken into account, the best prediction for every non-observed location would have been the mean μ. Consequently, the variance of the prediction error would have been equal to σ^2. As the values of the product $\lambda_i \, C(\mathbf{s}_i - \mathbf{s}_0)$ are positive, it can be seen from (B.10) that the prediction error variance of the kriging predictor is always smaller than the variance of the SF.

To obtain a positive error variance using (B.10) the function $C(\mathbf{h})$ must be positive definite. This means that for all possible $\mathbf{s}_1, \ldots, \mathbf{s}_n \in \mathbb{R}^k$ (k equals 1, 2 or 3) and for all $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ the following inequality must hold:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \, \lambda_j \, C(\mathbf{h}_{ij}) \geq 0 \; . \tag{B.11}$$

It is difficult to ensure that this is the case for any chosen function. Therefore, we cannot just estimate a covariance function directly from the data for a limited number of separation distances and then obtain a continuous function by linear or spline interpolation between the sampling locations. If such a covariance function were used in (B.9) and (B.10), then this would not necessarily lead to a positive estimate of the prediction error variance. In fact, there are only a limited number of functions for which it has been proven that inequality (B.11) will always hold. So the practical solution used in kriging is to take one of these 'permissible' functions and fit it through the points of the experimental covariance function. Next, the values of the fitted function are used to build the kriging system (B.9) and to estimate the kriging variance using (B.10). Table B.1 gives a number of covariance functions that can be

Table B.1. Permissible covariance models for simple kriging. Here h denotes the length of the lag vector $\mathbf{h}$.

(a) spherical model	$C(h) = \begin{cases} c\left[1 - \frac{3}{2}\left(\frac{h}{a}\right) + \frac{1}{2}\left(\frac{h}{a}\right)^3\right] & \text{if } h \le a \\ 0 & \text{if } h > a \end{cases}$
(b) exponential model	$C(h) = c\exp(-h/a)$
(c) Gaussian model	$C(h) = c\exp(-[h/a]^2)$
(d) nugget model	$C(h) = \begin{cases} c & \text{if } h = 0 \\ 0 & \text{if } h > 0 \end{cases}$

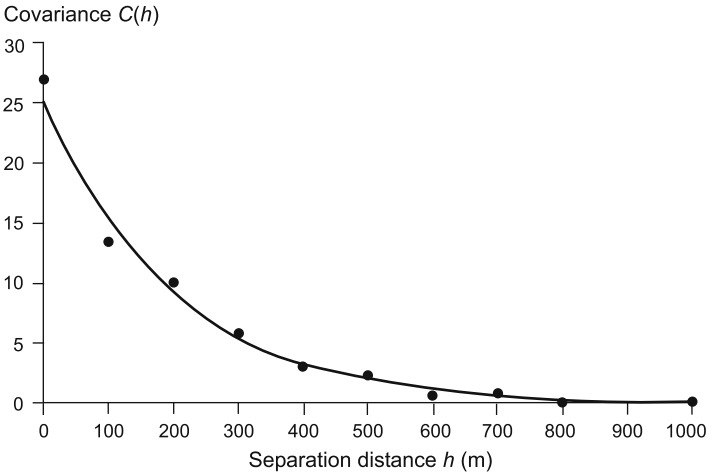


Fig. B.1. Example of an exponential covariance model fitted to estimated covariances

used for simple kriging (i.e., using a second-order stationary SF). Figure B.1 shows an example of an exponential model that is fitted to estimated covariances. Of course, in case of second-order stationarity the parameter c should be equal to the variance of the SF: $c = \sigma^2$.

Any linear combination of a permissible covariance model is a permissible covariance model itself. Often a combination of a nugget model and another model is observed:

$$C(h) = \begin{cases} c_0 + c_1 & \text{if } h = 0 \\ c_1 \exp(-h/a) & \text{if } h > 0 \end{cases} \tag{B.12}$$

where $c_0 + c_1 = \sigma^2$. In this case c_0 is often used to model the part of the variance that is attributable to observation errors and spatial variation that

occurs at distances smaller than the minimal distance between sampling locations.

B.1.2 Practice

The practical application of simple kriging would involve the mapping of some variable observed at a limited number of locations. In practice, the kriging routine would consist of the following steps:

1. *Estimate the mean and the covariance function from the data*
2. *Fit a permissible covariance model to the experimental covariance function*

If kriging is used for making maps, the locations where the predictions are made are usually located on a grid. So, when in the following steps we refer to a prediction location $\mathbf{s}_0$ we refer to a node of this grid. Thus, the following steps are repeated for every grid node:

3. *Solve the simple kriging system*
   Using (B.9) and the covariance function $C(\mathbf{h})$, the λ_i are obtained for location $\mathbf{s}_0$.
4. *Predict the value* $Z(\mathbf{s}_0)$
   With the λ_i, the observed values $z(\mathbf{s}_i)$ and the estimated value of μ in (B.4) the unknown value of $Z(\mathbf{s}_0)$ is predicted
5. *Calculate the variance of the prediction error at* $\mathbf{s}_0$
   Using $\lambda_i(\mathbf{s}_0)$, $C(\mathbf{h})$ and σ^2 the variance of the prediction error is calculated with (B.10).

The result is a map of predicted properties on a grid and a map of associated error variances.

B.2 Ordinary kriging

B.2.1 Theory

Ordinary kriging can be used if

1. $Z(\mathbf{s})$ is a second-order stationary SF but the mean of $Z(\mathbf{s})$ is unknown, or
2. $Z(\mathbf{s})$ is an intrinsic SF.

An intrinsic SF has the following properties (Journel and Huijbregts, 1978):

$$E[Z(\mathbf{s}_2) - Z(\mathbf{s}_1)] = 0 \tag{B.13}$$

$$E[Z(\mathbf{s}_2) - Z(\mathbf{s}_1)]^2 = 2\gamma(\mathbf{h}_{21}) = 2\gamma(\mathbf{h}) \tag{B.14}$$

The mean difference between the RVs at any two locations is zero (i.e., constant mean) and the variance of this difference is a function that only depends

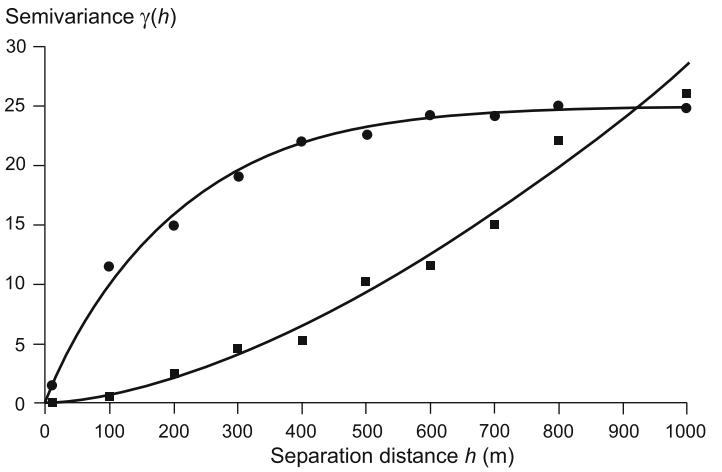


Fig. B.2. Example of permissible variogram models fitted to estimated semivariances; one variogram represents a second-order stationary SF and reaches a sill, the other represents a SF that is intrinsic but not second-order stationary

on the separation vector $\mathbf{h}$. The function $\gamma(\mathbf{h}) = \frac{1}{2}E[Z(\mathbf{s}) - Z(\mathbf{s}+\mathbf{h})]^2$ is called the semivariogram or variogram. If a SF is second-order stationary, it is also intrinsic, but not vice versa. If an intrinsic SF is also second-order stationary than the following relation exists between the variogram and the covariance function:

$$\gamma(\mathbf{h}) = \sigma^2 - C(\mathbf{h}) : \forall\ \mathbf{h}\ . \tag{B.15}$$

The difference between intrinsicity and second-order stationarity is further explained by Fig. B.2. The figure shows two variograms. The first variogram is the mirror image of the covariance function in Fig. B.1 (according (B.15)) and represents the variogram of a second-order stationary SF. Here, the variogram reaches a constant value called the *sill* which is equal to $\sigma^2 = 25$. The separation distance for which this sill is reached is called the *range* (here 1000 m). The range is the maximum distance over which the RVs at two locations are correlated. The second variogram does not reach a sill and belongs to an intrinsic SF that is not second-order stationary.

The ordinary kriging predictor is a weighted average of the surrounding observations:

$$\widetilde{Z}(\mathbf{s}_0) = \sum_{i=1}^{n} \lambda_i\, Z(\mathbf{s}_i)\ , \tag{B.16}$$

with $Z(\mathbf{s}_i)$ the values of $Z(\mathbf{s})$ at the observation locations (usually within a limited search neighbourhood). As with the simple kriging predictor we want (B.16) to be unbiased:

$$E[\widetilde{Z}(\mathbf{s}_0) - Z(\mathbf{s}_0)] = E\left[\sum_{i=1}^{n} \lambda_i \, Z(\mathbf{s}_i) - Z(\mathbf{s}_0)\right] = \sum_{i=1}^{n} \lambda_i \, E[Z(\mathbf{s}_i)] - E[Z(\mathbf{s}_0)] = 0 \tag{B.17}$$

As the unknown mean is constant, i.e., $E[Z(\mathbf{s}_i)] = E[Z(\mathbf{s}_0)] \, \forall \, \mathbf{s}_i, \mathbf{s}_0$, we find the following 'unbiasedness constraint' for the λ_i:

$$\sum_{i=1}^{n} \lambda_i = 1 \; . \tag{B.18}$$

Apart from being unbiased we also want to have a predictor with a minimum variance of the prediction error. The variance of the error for predictor (B.16) (ordinary kriging variance) can be written in terms of the semivariance as (see de Marsily, 1986, for a complete derivation):

$$V[\widetilde{Z}(\mathbf{s}_0) - Z(\mathbf{s}_0)] = E[\widetilde{Z}(\mathbf{s}_0) - Z(\mathbf{s}_0)]^2 = -\sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \, \lambda_j \, \gamma(\mathbf{h}_{ij}) + 2 \sum_{i=1}^{n} \lambda_i \, \gamma(\mathbf{h}_{i0}) \; . \tag{B.19}$$

We want to minimize the error variance subject to the constraint (B.18). In other words, we want to find the set of values λ_i, $i = 1, \ldots, n$ for which (B.19) is minimum without violating constraint (B.18). To find these, a mathematical trick is used. First the expression of the error variance is extended as follows:

$$\begin{aligned} E[\widetilde{Z}(\mathbf{s}_0) - Z(\mathbf{s}_0)]^2 = &-\sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \, \lambda_j \, \gamma(\mathbf{s}_i - \mathbf{s}_j) + 2 \sum_{i=1}^{n} \lambda_i \, \gamma(\mathbf{s}_i - \mathbf{s}_0) \\ &- 2\nu \left\{ \sum_{i=1}^{n} \lambda_i - 1 \right\} \; . \end{aligned} \tag{B.20}$$

If the predictor is unbiased, nothing has happened to the error variance as the added term is zero. The dummy variable ν is called the *Lagrange multiplier*. It can be shown that if we find the set of λ_i, $i = 1, \ldots, n$ and the value of ν for which (B.20) has its minimum value, we have also have the set of λ_i, $i = 1, \ldots, n$ for which the error variance of the ordinary kriging predictor is minimal, while at the same time $\sum \lambda_i = 1$. As with simple kriging, the minimum value is found by partial differentiation of (B.19) with respect to λ_i, $i = 1, \ldots, n$ and ν and equating the partial derivatives to zero. This results in the following system of $(n + 1)$ linear equations:

$$\begin{aligned} &\sum_{j=1}^{n} \lambda_j \, \gamma(\mathbf{h}_{ij}) + \nu = \gamma(\mathbf{h}_{i0}) \qquad\qquad i = 1, \ldots, n \\ &\sum_{i=1}^{n} \lambda_i = 1 \end{aligned} \tag{B.21}$$

Using the Lagrange multiplier, the value for the (minimum) ordinary kriging variance can be conveniently written as

Table B.2. Permissible variogram models for ordinary kriging; here h denotes the length of the lag vector $\mathbf{h}$.

(a) spherical model	$\gamma(h) = \begin{cases} c\left[\frac{3}{2}\left(\frac{h}{a}\right) + \frac{1}{2}\left(\frac{h}{a}\right)^3\right] & \text{if } h \le a \\ c & \text{if } h > a \end{cases}$	
(b) exponential model	$\gamma(h) = c\,[1 - \exp(-h/a)]$	
(c) Gaussian model	$\gamma(h) = c\,\{1 - \exp(-[h/a]^2)\}$	
(d) nugget model	$\gamma(h) = \begin{cases} 0 & \text{if } h = 0 \\ 1 & \text{if } h > 0 \end{cases}$	
(e) power model	$\gamma(h) = c\,h^{\omega}$	$0 < \omega < 2$

$$V[\widetilde{Z}(\mathbf{s}_0) - Z(\mathbf{s}_0)] = \sum_{i=1}^{n} \lambda_i\,\gamma(\mathbf{h}_{i0}) + \nu\ . \tag{B.22}$$

A unique solution of the system (B.21) and a positive kriging variance is only ensured if the variogram function is 'conditionally non-negative definite'. This means that for all possible $\mathbf{s}_1, \ldots, \mathbf{s}_n \in \mathbb{R}^k$ (k equals 1, 2 or 3) and for all $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ such that $\sum_i \lambda_i = 1$, the following inequality must hold:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i\,\lambda_j\,\gamma(\mathbf{h}_{ij}) \ge 0\ . \tag{B.23}$$

This is ensured if one of the permissible variogram models (Table B.2) is fitted to the experimental variogram data.

Models (a) to (d) are also permissible in case the SF is second-order stationary. The power model, which does not reach a sill, can be used in case of an intrinsic SF but not in case of a second-order stationary SF.

The unknown mean μ and the Lagrange multiplier ν require some further explanation. If all the data are used to obtain predictions at every location, at all locations the same unknown mean μ is implicitly estimated by the ordinary kriging predictor. The Lagrange multiplier represents the additional uncertainty that is added to the kriging prediction by the fact that the mean is unknown and must be estimated. Therefore, if the SF is second-order stationary, the variance of the prediction error for ordinary kriging is larger than that for simple kriging, the difference being the Lagrange multiplier. This can be deduced from substituting in (B.22) $\gamma(\mathbf{h}) = \sigma^2 - C(\mathbf{h})$ and taking into account that $\sum \lambda_i = 1$. This means that, whenever the mean is not exactly known and has to be estimated from the data it is better to use ordinary kriging, so that the added uncertainty about the mean is taken into account.

Even in simple kriging one rarely uses all data to obtain kriging predictions. Usually only a limited number of data points close to the prediction

location are used. This is to avoid that the kriging systems becomes too large and the mapping too slow. The most common way of selecting data is to centre an area or volume at the prediction location $\mathbf{s}_0$. Usually the radius is taken about the size of the variogram range. A limited number of data points that fall within the search area are retained for the kriging prediction. This means that the number of data points becomes a function of the prediction location: $n = n(\mathbf{s}_0)$. Also, if ordinary kriging is used, a local mean is implicitly estimated that changes with $\mathbf{s}_0$. So we have $\mu = \mu(\mathbf{s}_0)$ and $\nu = \nu(\mathbf{s}_0)$[1]. This shows that, apart from correcting for the uncertainty in the mean and being able to cope with a weaker form of stationarity, ordinary kriging has a third advantage when compared to simple kriging: even though the intrinsic hypothesis assumes that the mean is constant, using ordinary kriging with a search neighbourhood enables one to correct for local deviations in the mean. This makes the ordinary kriging predictor more robust to trends in the data than the simple kriging predictor.

B.2.2 Practice

In practice ordinary kriging consists of the following steps:

1. *Estimate the variogram*
2. *Fit a permissible variogram model*

For every node on the grid repeat:

3. *Solve the kriging equations*
   Using the fitted variogram model $\gamma(\mathbf{h})$ in the $n+1$ linear equations (B.21) yields, after solving them, the kriging weights $\lambda_i, i = 1, \ldots, n$ and the Lagrange multiplier ν.
4. *Predict the value* $Z(\mathbf{s}_0)$
   With the λ_i, the observed values $z(\mathbf{s}_i)$ (usually within the search neighbourhood), and (B.16) the unknown value of $Z(\mathbf{s}_0)$ is predicted.
5. *Calculate the variance of the prediction error*
   Using λ_i, $\gamma(\mathbf{h})$ and ν the variance of the prediction error is calculated with (B.22).

B.3 Block-Kriging

Up to now we have been concerned with predicting the target variable at the same support as the observations, usually point support. However, in many

[1] For briefness of notation we will use n and ν in the kriging equations, instead of $n(\mathbf{s}_0)$ and $\nu(\mathbf{s}_0)$. The reader should be aware that in most cases, both the number of observations and the Lagrange multipliers depend on the prediction location $\mathbf{s}_0$, except for those rare occasions that a global search neighbourhood is used.

cases one may be interested in the mean value of the target variable for some area or volume much larger than the support of the observations. For instance, one may be interested in the average porosity of a model block that is used in a numerical groundwater model, or the average precipitation of a catchment. These average quantities can be predicted using block kriging. The term 'block-kriging' is used as opposed to 'point-kriging' or 'punctual kriging' where variables are predicted at the same support as the observations. Any form of kriging has a point form and a block form. So, there is simple point-kriging and simple block-kriging and ordinary point-kriging and ordinary block-kriging etc. Usually, the term 'point' is omitted and the term 'block' is added only if the block-kriging form is used.

Consider the problem of predicting the mean $\overline{Z}$ of the target variable z that varies with spatial coordinate $\mathbf{s}$ for some domain $\mathcal{D}$ with size $|\mathcal{D}|$ (length, area or volume):

$$\overline{Z} = \frac{1}{|\mathcal{D}|} \int_{\mathbf{s}\in\mathcal{D}} Z(\mathbf{s})\,\mathrm{d}\mathbf{s}\ . \tag{B.24}$$

In case $\mathcal{D}$ is a block in three dimensions with upper and lower boundaries $s_{1\mathrm{l}}$, $s_{1\mathrm{u}}$, $s_{2\mathrm{l}}$, $s_{2\mathrm{u}}$, $s_{3\mathrm{l}}$, $s_{3\mathrm{u}}$ the spatial integral (B.24) stands for

$$\frac{1}{|\mathcal{D}|} \int_{\mathbf{s}\in\mathcal{D}} Z(\mathbf{s})\,\mathrm{d}\mathbf{s} = \frac{1}{|s_{1\mathrm{u}} - s_{1\mathrm{l}}||s_{2\mathrm{u}} - s_{2\mathrm{l}}||s_{3\mathrm{u}} - s_{3\mathrm{l}}|} \int_{s_{3\mathrm{l}}}^{s_{3\mathrm{u}}} \int_{s_{2\mathrm{l}}}^{s_{2\mathrm{u}}} \int_{s_{1\mathrm{l}}}^{s_{1\mathrm{u}}} Z(s_1, s_2, s_3)\,\mathrm{d}s_1\,\mathrm{d}s_2\,\mathrm{d}s_3 \tag{B.25}$$

Of course, the domain $\mathcal{D}$ can be of any form, in which case a more complicated spatial integral is used.

Similar to point-kriging, the unknown value of $\overline{Z}$ can be predicted as linear combination of the observations by assuming that the predictant and the observations are partial realizations of a SF. So, the ordinary block kriging predictor becomes:

$$\widetilde{\overline{Z}} = \sum_{i=1}^{n} \lambda_i\, Z(\mathbf{s}_i)\ , \tag{B.26}$$

where the block-kriging weights λ_i are determined such that $\widetilde{\overline{Z}}$ is unbiased and the prediction-error variance $V[\widetilde{\overline{Z}} - \overline{Z}]$ is minimal. This is achieved by solving the λ_i from the ordinary block-kriging system:

$$\begin{aligned} &\sum_{j=1}^{n} \lambda_j\, \gamma(\mathbf{h}_{ij}) + \nu = \gamma(\mathbf{s}_i, \mathcal{D}) \qquad\qquad i = 1, \ldots, n \\ &\sum_{i=1}^{n} \lambda_i = 1 \end{aligned} \quad . \tag{B.27}$$

It can be seen that the ordinary block-kriging system looks almost the same as the ordinary kriging system, except for the term on the right hand side

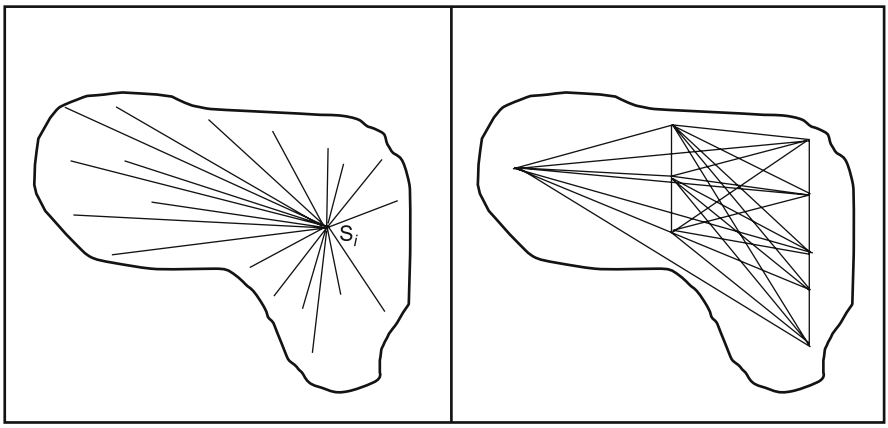


Fig. B.3. Schematic examples of calculating the point–block semivariance (left) and the within-block semivariance (right)

which is the average semivariance between a location $\mathbf{s}_i$ and all the locations inside the domain $\mathcal{D}$:

$$\gamma(\mathbf{s}_i, \mathcal{D}) = \frac{1}{|\mathcal{D}|} \int_{\mathbf{s} \in \mathcal{D}} \gamma(\mathbf{s}_i - \mathbf{s}) \, \mathrm{d}\mathbf{s} \; . \tag{B.28}$$

When building the block-kriging system, the integral in (B.28) is usually not solved. Instead, it is approximated by first discretizing the area of interest in a limited number of grid nodes. Second, the semivariances are calculated between the sampling location and the N nodes $\mathbf{s}_j$ discretizing $\mathcal{D}$ (see Fig. B.3, left figure). Third, the average semivariance is approximated by averaging these semivariances as:

$$\gamma(\mathbf{s}_i, \mathcal{D}) \simeq \frac{1}{N} \sum_{j=1}^{N} \gamma(\mathbf{h}_{ij}) \; . \tag{B.29}$$

The variance of the prediction error (block-kriging variance) of is given by

$$V[\widetilde{\overline{Z}} - \overline{Z}] = E[\widetilde{\overline{Z}} - \overline{Z}]^2 = \sum_{i=1}^{n} \lambda_i \, \gamma(\mathbf{s}_i, \mathcal{D}) + \nu - \gamma(\mathcal{D}, \mathcal{D}) \; , \tag{B.30}$$

where $\gamma(\mathcal{D}, \mathcal{D})$ is the average semivariance within the area of interest, i.e., the average semivariance between all locations with $\mathcal{D}$:

$$\gamma(\mathcal{D}, \mathcal{D}) = \frac{1}{|\mathcal{D}|^2} \int_{\mathbf{s}_2 \in \mathcal{D}} \int_{\mathbf{s}_1 \in \mathcal{D}} \gamma(\mathbf{h}_{12}) \, \mathrm{d}\mathbf{s}_1 \, \mathrm{d}\mathbf{s}_2 \; , \tag{B.31}$$

which in practice is approximated by N grid nodes $\mathbf{s}_i$ discretizing $\mathcal{D}$ as (see also Fig. B.3, right figure)

$$\gamma(\mathcal{D},\mathcal{D}) \simeq \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\gamma(\mathbf{h}_{ij})\ . \tag{B.32}$$

Here we have given the equations for ordinary block-kriging. The simple block-kriging equations can be deduced in a similar manner from the simple kriging equations (B.9) by replacing the covariance on the right hand side by the point–block covariance $C(\mathbf{s}_i,\mathcal{D})$. The prediction error variance is given by (B.10) with σ^2 replaced by the within-block variance $C(\mathcal{D},\mathcal{D})$ (the average covariance of grid nodes within $\mathcal{D}$) and $C(\mathbf{s}_i-\mathbf{s}_0)$ by $C(\mathbf{s}_i,\mathcal{D})$. The point–block covariance and the within block covariance are defined as in (B.28) and (B.31), with $\gamma(\mathbf{s}_1-\mathbf{s}_2)$ replaced by $C(\mathbf{s}_1-\mathbf{s}_2)$.

B.4 Block Indicator Kriging the Spatial Cumulative Distribution Function

Kriging can also be used to predict the Spatial Cumulative Distribution Function (SCDF) of a target variable for an area or volume from observations taken at a limited number of sampling locations. The SCDF gives for a given value of the target variable (called 'threshold'), the fraction of the universe (area or volume) with a value smaller than or equal to this threshold. To predict the SCDF the geostatistical technique called 'block indicator kriging' can be used. Again it is assumed that the unknown spatial variation of target variable z is a realization of a Stochastic Function $Z(\mathbf{s})$. In the following we are interested in predicting the SCDF of some universe $\mathcal{U}$. For a given threshold z_k the fraction $F_{\mathcal{U}}(z_k)$ of $\mathcal{U}$ with $Z(\mathbf{s}) \leq z_k$ can be calculated with the following spatial integral over $\mathcal{U}$:

$$F_{\mathcal{U}}(z_k) = \int\limits_{\mathbf{s}\in\mathcal{U}} I(\mathbf{s};z_k)\,\mathrm{d}\mathbf{s}\ , \tag{B.33}$$

where $I(\mathbf{s};z_k)$ is the indicator transform for threshold z_k:

$$I(\mathbf{s};z_k) = \begin{cases} 1 & \text{if } Z(\mathbf{s}) \leq z_k \\ 0 & \text{if } Z(\mathbf{s}) > z_k \end{cases} \tag{B.34}$$

Based on (B.33) the fraction $F_{\mathcal{U}}(z_k)$ can be estimated from n observations $Z(\mathbf{s}_i)$ with the following ordinary block indicator kriging predictor:

$$\hat{F}_{\mathcal{U}}(z_k) = \sum_{i=1}^{n}\lambda_i\, I(\mathbf{s}_i;z_k)\ , \tag{B.35}$$

where λ_i are the block indicator kriging weights.

To calculate kriging weights the indicator variogram for threshold z_k is required, which is defined as:

$$\gamma_I(\mathbf{s}_i, \mathbf{s}_j; z_k) = \frac{1}{2} E[I(\mathbf{s}_i; z_k) - I(\mathbf{s}_j; z_k)]^2 \ . \tag{B.36}$$

As was the case with the regular variogram, if $I(\mathbf{s}; z_k)$ is considered to be an intrinsic Stochastic Function, the indicator variogram only depends on the difference $\mathbf{s}_i - \mathbf{s}_j$ and we have:

$$\gamma_I(\mathbf{s}_j, \mathbf{s}_i; z_k) = \gamma_I(\mathbf{h}_{ij}; z_k) = \gamma_I(\mathbf{h}_{ji}; z_k) \ . \tag{B.37}$$

Besides the point–point indicator semivariance also the point–block indicator semivariance and the within-block indicator semivariance are required, defined respectively as:

$$\gamma_I(\mathbf{s}_i, \mathcal{U}; z_k) = \frac{1}{|\mathcal{U}|} \int\limits_{\mathbf{s} \in \mathcal{U}} \gamma(\mathbf{s}_i - \mathbf{s}; z_k) \, \mathrm{d}\mathbf{s} \tag{B.38}$$

$$\gamma_I(\mathcal{U}, \mathcal{U}; z_k) = \frac{1}{|\mathcal{U}|^2} \int\limits_{\mathbf{s}_2 \in \mathcal{U}} \int\limits_{\mathbf{s}_1 \in \mathcal{U}} \gamma(\mathbf{h}_{12}; z_k) \, \mathrm{d}\mathbf{s}_1 \, \mathrm{d}\mathbf{s}_2 \ . \tag{B.39}$$

The kriging weights are obtained by solving a set of linear equations that, apart from the semivariances being replaced by indicator semivariances, is similar to (B.21):

$$\begin{aligned} &\sum_{j=1}^{n} \lambda_j \, \gamma_I(\mathbf{h}_{ij}; z_k) + \nu_I = \gamma_I(\mathbf{s}_i, \mathcal{U}; z_k) \qquad\qquad i = 1, \ldots, n \\ &\sum_{i=1}^{n} \lambda_i = 1 \end{aligned} \tag{B.40}$$

where the Lagrange multiplier ν_I is added to account for the additional unbiasedness condition $\sum \lambda_i = 1$.

Finally, the variance of the prediction error is given by:

$$V[\hat{F}_{\mathcal{U}}(z_k) - F_{\mathcal{U}}(z_k)] = \sum_{i=1}^{n} \lambda_i \, \gamma_I(\mathbf{s}_i, \mathcal{U}; z_k) + \nu_I - \gamma_I(\mathcal{U}, \mathcal{U}; z_k) \ . \tag{B.41}$$

The SCDF can be estimated by dividing the range of variation of $Z(\mathbf{s})$ into $K+1$ classes based on K thresholds z_k, $k = 1, \ldots, K$ (where $z_1 < z_2 < \cdots < z_k < z_{k+1} < \cdots < z_K$). By repeating for each threshold the block indicator kriging with the indicator variograms belonging to that threshold, a series of fractions below threshold $\hat{F}_{\mathcal{U}}(z_k)$, $k = 1, \ldots, K$ is obtained. From these a discrete approximation of the SCDF is constructed.

We end this appendix with some words on indicator variograms. Indicator variograms must obey certain rules. If the Stochastic Function $I(\mathbf{s}; z_k)$ is stationary, the indicator variogram will reach a sill which, theoretically, is equal to $E[I(\mathbf{s}; z_k)]\{1 - E[I(\mathbf{s}; z_k)]\}$, i.e., the variance of a binomially distributed variable. Of course, as $E[I(\mathbf{s}; z_k)]$ is not known beforehand, we cannot fit a

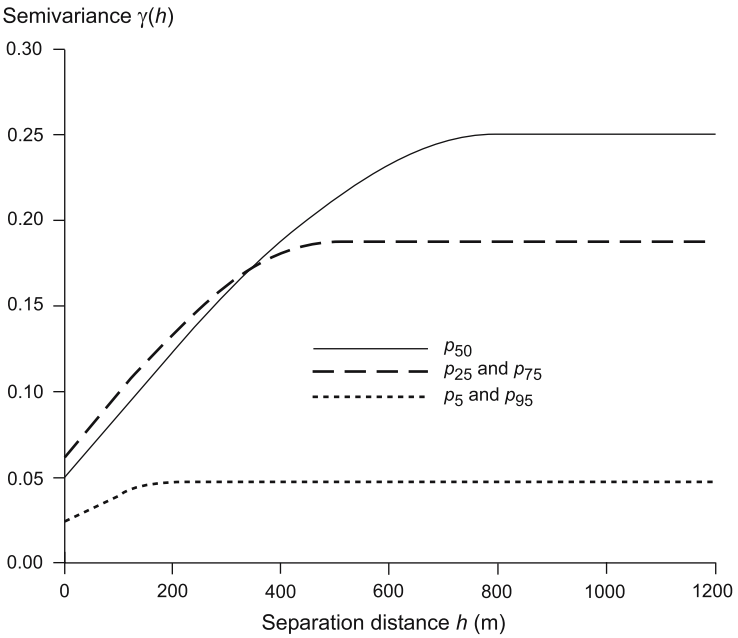


Fig. B.4. Examples of indicator variograms of a multivariate Gaussian Stochastic Function

permissible function with exactly that sill. If $Z(\mathbf{s})$ is stationary and multivariate Gaussian the indicator variograms will behave as shown in Fig. B.4. For a threshold at the median value of the SCDF (p_{50} in Fig. B.4.) the sill is largest (sill ≈ 0.25), the relative nugget (ratio of nugget to sill) smallest and the range is largest.

When moving towards the tail of the distribution (low or high values) (p_{25}, p_{75} and p_5, p_{95}) the sill and range decrease and the relative nugget increases. This shows that the values close to the median exhibit more spatial contiguity than large or small values. For non-Gaussian $Z(\mathbf{s})$ this does not have to be the case. For instance, in case of heterogeneous hydraulic conductivity in certain geological settings there may be a background of less contiguous smaller and median values in combination with large values that are very well spatially connected, forming preferential flow paths.

C

Processes and Time-Series Modelling

C.1 Introduction to Time-Series Modelling

The state of many phenomena in nature changes with time. This dynamic 'behaviour' can be described by time-series models, which can be used to estimate target parameters. These may include expected values at certain times such as the start of the growing season, or probabilities that critical levels are exceeded at certain times or during certain periods. These target parameters are estimated with the purpose of obtaining characteristics of the development of a certain universe in time. Such characteristics can, for instance, be extrapolated to future situations. Inherently, the universe is assumed to develop in time, following a process about which information is obtained by an observed time series. Because of restricted knowledge, there is no certainty about the 'true' process, if any, along which a universe develops in time. Therefore, the 'assumed' process is referred to as a *stochastic* process. Typically, this process is described by a *model*. One general class of models is that of time-series models as described by Box and Jenkins (1976) and Hipel and McLeod (1994). It is emphasized here that in time-series modelling, the observed time series itself is regarded as the realization of a process. This process, which depends on the sampling interval, should not be confused with the underlying physical processes which cause the variation in the observed time series, as well as observation error. It is stressed that the assumption of a stochastic process does not imply that nature is stochastic. Basically, stochastic processes are data-based, and they are imaginary, enabling statistical inference.

C.2 Stationary Processes

A process is said to be stationary if its statistical properties do not change with time. It is important to note that stationarity is not found in nature, whether in geological, evolutionary or any other processes. Stationarity can only be assumed, given the length of the period and the length of time intervals.

Strong or *strict stationarity* means that all statistical properties are time-independent, so they do not change after time shifts. It is often sufficient to assume *weak stationarity of order k*, which means that the statistical moments up to order k only depend on differences in time and not on time as such. *Second-order stationarity* means that the stochastic process can be described by the mean, the variance and the autocorrelation function. This is also called *covariance stationarity*.

We now consider a discrete-time second-order stationary stochastic process. Suppose that we have an equidistant time series of n observations, $z_1, z_2, z_3, \ldots, z_n$. The process cannot be exactly described, so $\{z_t\}$ is considered to be a realization of a stochastic process $\{Z_t\}$. The mean is defined as the expected value of Z_t

$$\mu = E[Z_t] \,, \tag{C.1}$$

which can be estimated from an observed time series by the simple estimator

$$\hat{\mu} = \bar{z} = \frac{1}{n} \sum_{t=1}^{n} z_t \,. \tag{C.2}$$

The variance of the stochastic process $\{Z_t\}$ is defined as the expected value of the squared deviations from the mean:

$$\sigma_Z^2 = E[(Z_t - \mu)^2] \,. \tag{C.3}$$

σ_Z^2 can be estimated by

$$\widehat{\sigma_Z^2} = \frac{1}{n-1} \sum_{t=1}^{n} (z_t - \bar{z})^2 \,. \tag{C.4}$$

The autocovariance for lag k is defined by

$$\begin{aligned} C_k &= E[(Z_t - \mu)(Z_{t+k} - \mu)] \,, \\ C_0 &= \sigma_Z^2 \,. \end{aligned} \tag{C.5}$$

For lag 0 the autocovariance equals the variance. The autocorrelation function (ACF) for lag k is a scaled form of the autocovariance:

$$\rho_k = \frac{C_k}{C_0} \,. \tag{C.6}$$

The sample autocovariance function for lag k can be calculated from a time series by

$$\widehat{C}_k = \frac{1}{n_k} \sum_{t=1}^{n-k} (z_t - \bar{z})(z_{t+k} - \bar{z}) \,, \tag{C.7}$$

where n_k is the number of summed terms, with a maximum of $n - k$; terms for which a value of z_t or z_{t+k} is missing are excluded. The sample ACF is estimated by

$$\hat{\rho}_k = \left(1 - \frac{k}{n}\right) \frac{\widehat{C}_k}{\widehat{C}_0} \,. \tag{C.8}$$

C.2.1 Autoregressive (AR) processes

For many environmental processes it is likely that the state at a particular time is correlated with the state at previous times. These processes are referred to as autoregressive (AR) processes. An autoregressive process of order 1, an AR(1) process or Markov process, is given by

$$Z_t - \mu = \phi_1(Z_{t-1} - \mu) + \epsilon_t \ , \tag{C.9}$$

where μ is the mean level, ϕ_1 is the AR parameter, and ϵ_t is the error term with zero mean and variance σ_ϵ^2. ϵ_t is assumed to be identically and independently distributed (IID), so

$$E[\epsilon_t \epsilon_{t-k}] = \begin{cases} \sigma_\epsilon^2 & \text{if } k = 0 \\ 0 & \text{if } k \neq 0 \end{cases} \tag{C.10}$$

for all t.

Using the backward shift operator B, (C.9) can be written as

$$Z_t - \mu = \phi_1(BZ_t - \mu) + \epsilon_t \ , \tag{C.11}$$

where $B^k Z_t = Z_{t-k}$. (C.11) can also be written as

$$\phi(B)(Z_t - \mu) = \epsilon_t \ , \tag{C.12}$$

with $\phi(B) = 1 - \phi_1 B$.

An autoregressive process of order p, an AR(p) process, is given by

$$Z_t - \mu = \phi_1(Z_{t-1} - \mu) + \phi_2(Z_{t-2} - \mu) + \cdots + \phi_p(Z_{t-p} - \mu) + \epsilon_t \ , \tag{C.13}$$

or using the backward shift operator:

$$\phi(B)(Z_t - \mu) = \epsilon_t \ , \tag{C.14}$$

where $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$ is the autoregressive operator of order p.

To obey the assumption of stationarity, the values of the AR parameters are restricted. For an AR(1) process, this restriction is $|\phi_1| < 1$.

Important tools to identify an AR(p) process from an observed time series are the ACF and the partial autocorrelation function (PACF). The theoretical ACF for an AR(p) process is derived as follows. First, the terms of the AR(p) process in (C.13) are multiplied by $(Z_{t-k} - \mu)$:

$$\begin{aligned}(Z_{t-k} - \mu)(Z_t - \mu) = \phi_1(Z_{t-k} - \mu)(Z_{t-1} - \mu) + \phi_2(Z_{t-k} - \mu)(Z_{t-2} - \mu) \\ + \cdots + \phi_p(Z_{t-k} - \mu)(Z_{t-p} - \mu) + (Z_{t-k} - \mu)\epsilon_t \ .\end{aligned} \tag{C.15}$$

By taking expectations of the terms in (C.15) we obtain

$$C_k = \phi_1 C_{k-1} + \phi_2 C_{k-2} + \cdots + \phi_p C_{k-p} \ , \tag{C.16}$$

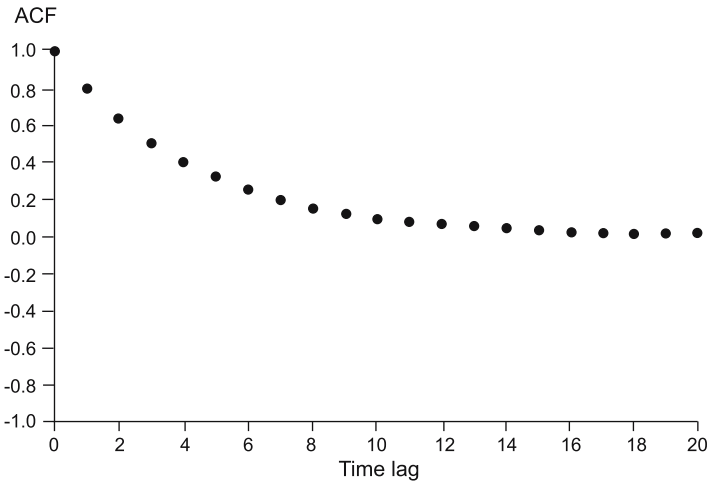


Fig. C.1. Theoretical ACF for an AR(1) process with $\phi_1 = 0.8$

with $k > 0$. $E[(Z_{t-k} - \mu)\epsilon_t]$ equals zero for $k > 0$, because Z_{t-k} only depends on the error process up to and including $t - k$ and is uncorrelated with a_t. The theoretical ACF is obtained by dividing (C.16) by $C(0)$:

$$\rho_k = \phi_1 \rho_{k-1} + \phi_2 \rho_{k-2} + \cdots + \phi_p \rho_{k-p} \ , \tag{C.17}$$

with $k > 0$. Figures C.1 and C.2 give the theoretical ACFs for AR(1) processes with $\phi_1 = 0.8$ and $\phi_1 = -0.8$, respectively (see (C.9)).

For a derivation of the PACF we refer to Hipel and McLeod (1994). Because of its definition, the PACF must be equal to zero after lag p for an AR(p) process. The sample ACF and PACF of a deseasonalized series of water table depths are given in Fig. C.3, which indicate an AR(1) process.

C.2.2 Moving average (MA) processes

In moving average processes the state at a certain time depends on a random shock at that time and a random shock which occurred at one or more previous times. A first-order moving average process, MA(1), is given by

$$Z_t - \mu = \epsilon_t - \theta_1 \epsilon_{t-1} \ . \tag{C.18}$$

Here ϵ_t and ϵ_{t-1} are random shocks which form part of a white noise process with zero mean and finite and constant variance. Using the backward shift operator, (C.18) can be written as

$$Z_t - \mu = \theta(B)\epsilon_t \ , \tag{C.19}$$

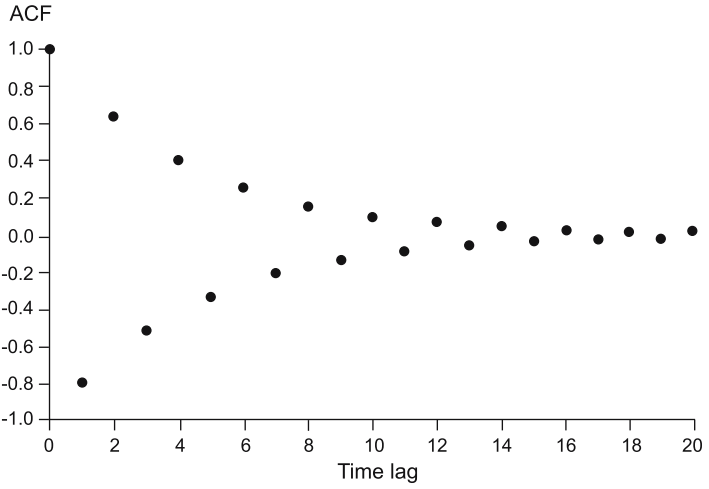


Fig. C.2. Theoretical ACF for an AR(1) process with $\phi_1 = -0.8$

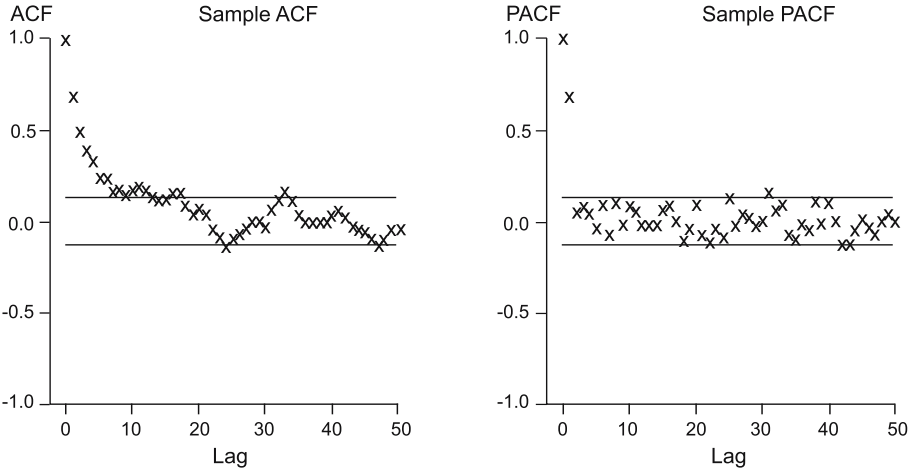


Fig. C.3. Sample ACF and PACF for a deseasonalized time series of water table depths.

where $\theta(B) = 1-\theta_1 B$ is the MA operator of order one. The process is invertible if $|\theta_1| < 1$. The process is stationary for all values of θ_1 since ϵ_t is stationary. The theoretical ACF for a MA(1) process with $\theta_1 = 0.8$ is given in Fig. C.4.

An MA(q) process is given by

$$Z_t - \mu = \epsilon_t - \theta_1\epsilon_{t-1} - \theta_2\epsilon_{t-2} - \cdots - \theta_q\epsilon_{t-q} \,, \tag{C.20}$$

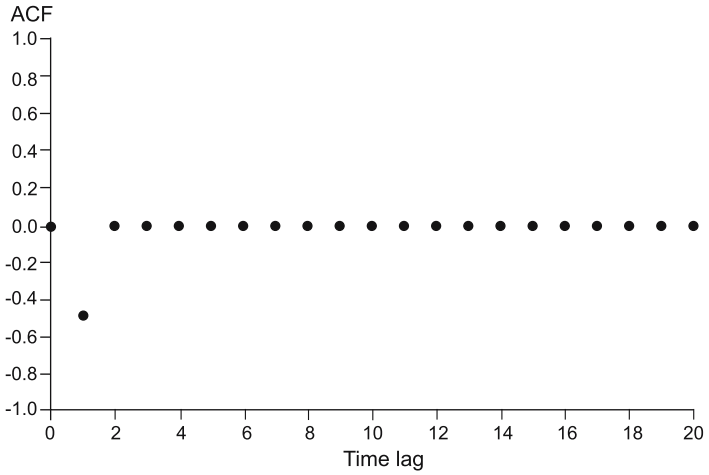


Fig. C.4. Theoretical ACF for a MA(1) process with $\theta_1 = 0.8$

or

$$Z_t - \mu = \theta(B)\epsilon_t \ , \tag{C.21}$$

where $\theta(B)$ is the MA operator of order q. The process is stationary for all values of the MA parameters.

C.2.3 Autoregressive moving average (ARMA) process

A time series may contain properties of an autoregressive process as well as a moving average process. An autoregressive moving average ARMA(1,1) process is given by

$$Z_t - \mu = \phi_1(Z_{t-1} - \mu) + \epsilon_t - \theta_1\epsilon_{t-1} \ . \tag{C.22}$$

The ARMA(p, q) process is given by

$$\phi(B)(Z_t - \mu) = \theta(B)\epsilon_t \ , \tag{C.23}$$

where $\phi(B)$ and $\theta(B)$ are the AR(p) and the MA(q) operator, respectively.

C.3 Nonstationary Processes

Calculating differences allows a trend to be removed from a series:

$$\nabla Z_t = (Z_t - \mu) - (Z_{t-1} - \mu) \tag{C.24}$$

$$\nabla^2 Z_t = \nabla Z_t - \nabla Z_{t-1} \tag{C.25}$$

and so on, until a series of differences is obtained with a constant mean in time.

C.3.1 Autoregressive Integrated Moving Average (ARIMA) processes

Basically, an ARIMA model is an ARMA model for stationary differences:

$$(\nabla^d Z_t - \mu) = \frac{\theta(B)}{\phi(B)}\epsilon_t \ . \tag{C.26}$$

C.3.2 Seasonal nonstationary processes

A form of nonstationarity often encountered in environmental processes is seasonality. Besides seasonal variation of the mean, the variance itself may vary seasonally. For example, shallow water tables in the wet season may vary more than deep water tables in the dry season, due to reduced storage capacity of the unsaturated zone in the wet season. If the variance varies in time, i.e., there is heteroscedasticity, the variance should be made constant by an appropriate deseasonalization procedure or by a Box–Cox transformation of the time series (Hipel and McLeod, 1994).

C.3.3 Seasonal integrated autoregressive moving average (SARIMA) processes

In the case of a seasonal autoregressive moving average process, differences are calculated for the so-called seasonal distance, with the aim of removing a seasonal trend. For example, the seasonal distance for monthly values is twelve. The general notation of a SARIMA(p,d,q)×(P,D,Q) model is

$$(\nabla^d \nabla_s^D Z_t - \mu) = \frac{\phi(B)\Phi(B^s)}{\theta(B)\Theta(B^s)}\epsilon_t \ . \tag{C.27}$$

C.4 Transfer Function–Noise Processes

A class of time-series models which describe the linear dynamic relationship between one or more input series and an output series is that of the transfer function model with added noise (TFN) developed by Box and Jenkins (1976). For applications to environmental series, we refer to Hipel and McLeod (1994). The general TFN model is given schematically in Fig. C.5.

If one input series $\{X_t\}$ is considered, the TFN model is defined as

$$Z_t = Z_t^* + N_t \ , \tag{C.28}$$

where

$$Z_t^* = \sum_{i=1}^{r} \delta_i Z_{t-i}^* + \omega_0 X_{t-b} - \sum_{j=1}^{s} \omega_j X_{t-j-b} \tag{C.29}$$

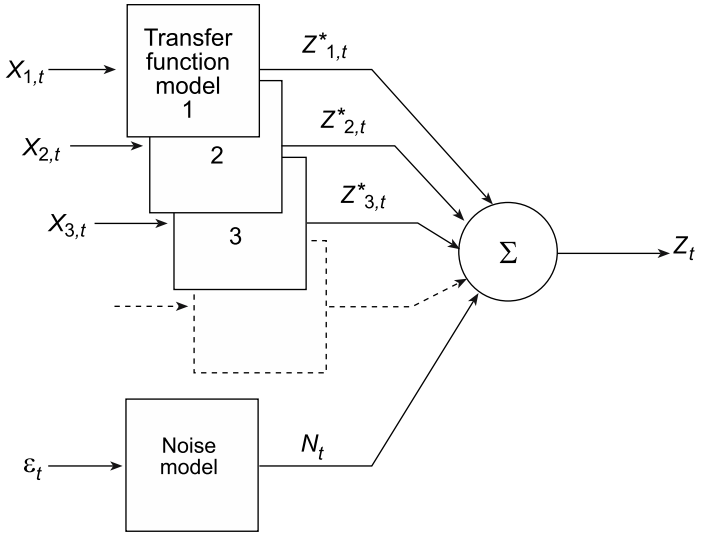


Fig. C.5. Schematic representation of a transfer function model with added noise. $X_{1,t}, X_{2,t}, X_{3,t}, \ldots$ are input variables. $Z^*_{1,t}, Z^*_{2,t}, Z^*_{3,t}, \ldots$ are transfer components. ϵ_t is an error term which forms a series of independent and identically distributed disturbances, with finite and constant variance σ^2_ϵ. N_t is the noise component. Z_t is the output variable

is the transfer component, and

$$N_t - \mu = \sum_{i=1}^{p} \phi_i (N_{t-i} - \mu) + \epsilon_t - \sum_{j=1}^{q} \theta_j \epsilon_{t-j} \ , \tag{C.30}$$

is the noise component. The *subscript b* is a pure delay, which is the number of time steps after which a reaction to an input change is observed in the output. The extension to more input series is straightforward.

The transfer component in (C.28) can be written as

$$\begin{aligned} Z^*_t &= \nu_0 X_t + \nu_1 X_{t-1} + \nu_2 X_{t-2} + \cdots \\ &= \nu(B) X_t \ . \end{aligned} \tag{C.31}$$

The weights $\nu_0, \nu_1, \nu_2, \ldots$ form the impulse–response function $\nu(B)$:

$$\nu(B) = \frac{\omega(B)}{\delta(B)} = \frac{\omega_0 - \omega_1 B - \omega_2 B^2 - \cdots - \omega_s B^s}{1 - \delta_1 B - \delta_2 B^2 - \cdots - \delta_r B^r} \ . \tag{C.32}$$

The theoretical impulse–response function reflects the same autoregressive and moving average characteristics as the theoretical autocorrelation function, given in Sect. C.2.

Box and Jenkins (1976) present a procedure for identifying the order of TFN models. This procedure is summarized by the following steps:

1. An appropriate univariate time-series model is fitted to the input series $\{x_t\}$. The resulting white noise sequence of residuals is called the prewhitened input series $\{\alpha_t\}$.
2. The output series $\{z_t\}$ is filtered by the univariate time-series model for the input series obtained in the previous step. This results in a series $\{\beta_t\}$.
3. The residual cross-correlation function $\hat{\rho}_{\alpha\beta}(k)$ (residual CCF) is calculated for the $\{\alpha_t\}$ and $\{\beta_t\}$ series:

$$\hat{\rho}_{\alpha\beta}(k) = \left(1 - \frac{k}{n}\right) \frac{\widehat{C}_{\alpha\beta}(k)}{\sqrt{\widehat{C}_{\alpha}(0)\widehat{C}_{\beta}(0)}} , \tag{C.33}$$

   where

$$\widehat{C}_{\alpha\beta}(k) = \frac{1}{n_k} \sum_{t=1}^{n-k} \alpha_t \beta_{t+k} \tag{C.34}$$

   for positive lags, and n_k is the number of summed terms. Terms with missing values are excluded. $\widehat{C}_{\alpha}(0)$ and $\widehat{C}_{\beta}(0)$ are the sample variances of the α series and the β series, respectively, calculated by (C.4).
4. Based on the residual CCF given by (C.33), the parameters required in the transfer function $\nu(B)$ in (C.32) are identified. Box and Jenkins (1976) show that the theoretical CCF between α_t and β_t is directly proportional to $\nu(B)$.
5. Next, a noise model is identified for the series

$$\hat{n}_t = (z_t - \bar{z}) - \hat{\nu}(B)(x_t - \bar{x}) , \tag{C.35}$$

   by using the sample ACF and sample PACF for $\hat{n}_t$.

D

Diagrams on Sample Sizes for Sampling in Space–Time

In this appendix diagrams are presented that can be used for optimizing spatio-temporal sampling. Figures D.1 to D.4 can be used for optimizing sampling in space and time in order to estimate the spatio-temporal mean using ordinary kriging. The underlying spatio-temporal Stochastic Function is assumed to be second-order stationary with a spherical semivariogram. Given the (surface) area $|\mathcal{S}|$, period length $|\mathcal{T}|$, variance σ^2 of the underlying Stochastic Function, spatial range a_{s}, and temporal range a_{t}, the figure shows isolines of the ratio of prediction-error variance and the variance of the Stochastic Function, $\sigma^2_{\mathrm{obk}}/\sigma^2$. For more explanation we refer to Sect. 15.3.4.

Figures D.5 to D.8 can be used for optimizing sampling in space and time when predicting hydraulic head using a numerical groundwater flow model and the Kalman filter. These figures can be used as follows. If the hydraulic parameters S (storage coefficient) and T (transmissivity) are given as well as the discretization in space (δs) and time (δt), a combination of sampling distances in space (Δs) and time (Δt) can be chosen to achieve a given ratio between mean prediction-error variance and the variance of the model error, $\sigma^2_{\mathrm{kf}}/\sigma^2_{\mathbf{Q}}$. See Sect. 16.2.4 for more explanation.

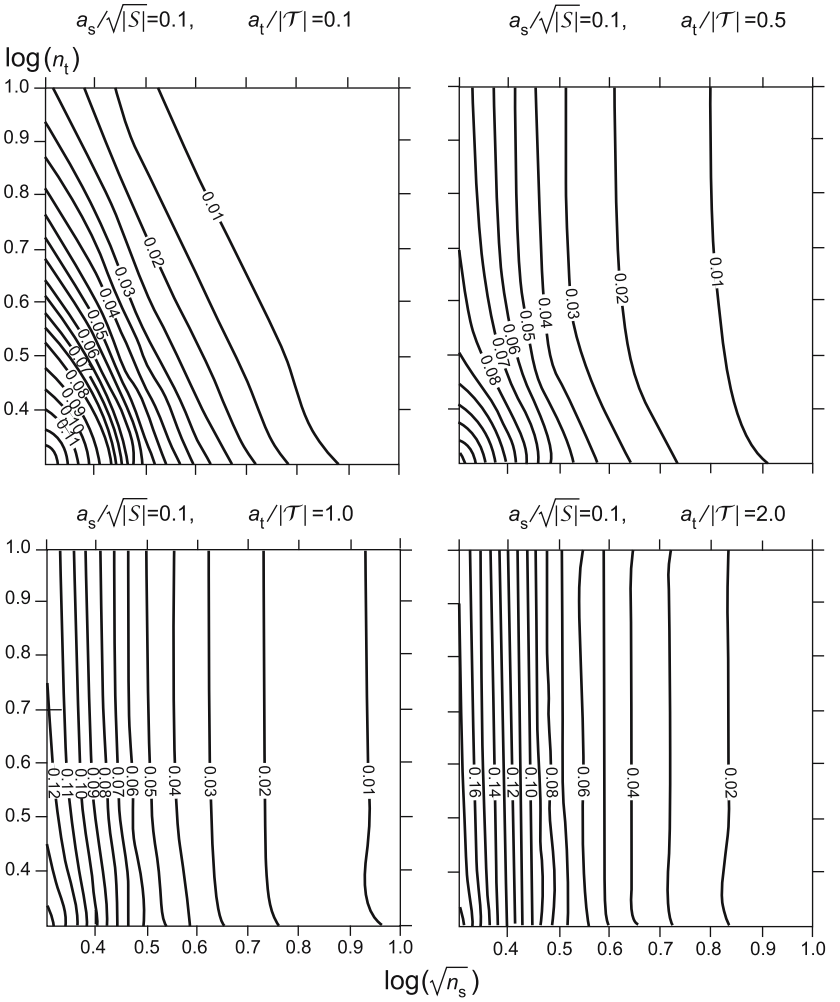


Fig. D.1. Sampling on a space–time grid (square grid pattern in space) for predicting the spatio-temporal mean by space–time kriging. The figure shows the variance ratio $\sigma^2_{\mathrm{obk}}/\sigma^2$ for $a_s/\sqrt{|\mathcal{S}|} = 0.1$ and $a_t/|\mathcal{T}| = 0.1, 0.5, 1.0, 2.0$ as a function of the number of sampling locations n_s and sampling times n_t.

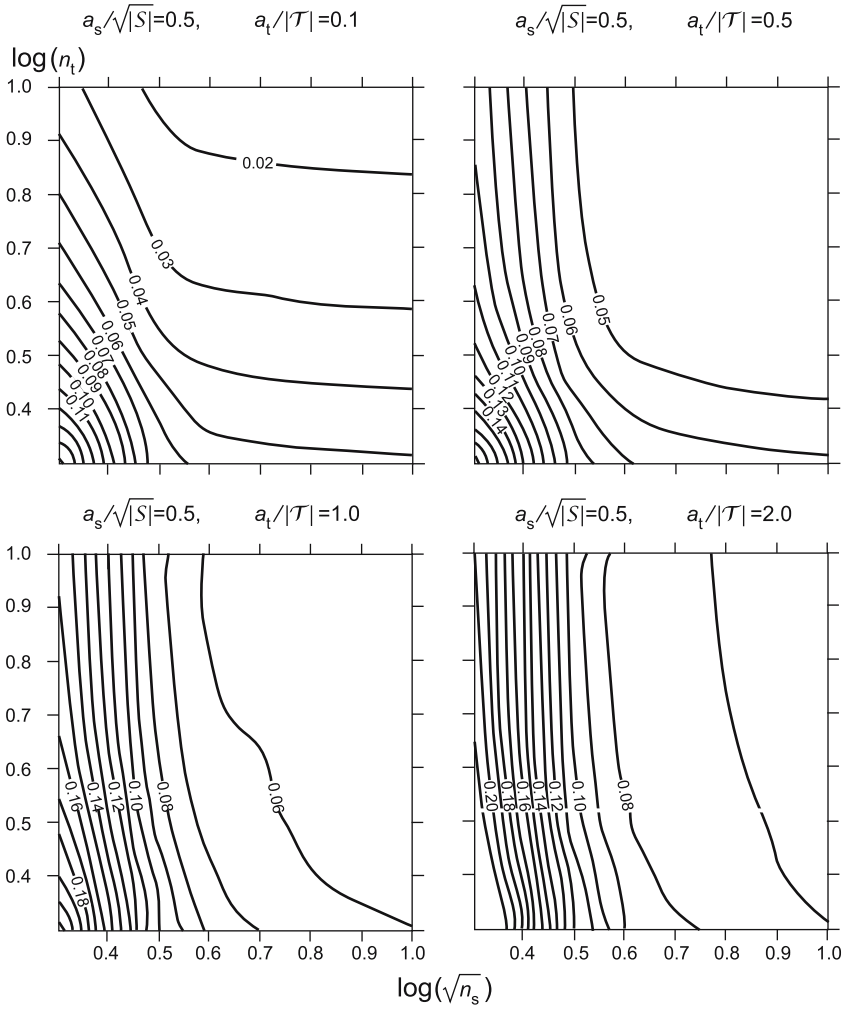


Fig. D.2. Sampling on a space–time grid (square grid pattern in space) for predicting the spatio-temporal mean by space–time kriging. The figure shows the variance ratio $\sigma^2_{\text{obk}}/\sigma^2$ for $a_s/\sqrt{|\mathcal{S}|} = 0.5$ and $a_t/|\mathcal{T}| = 0.1, 0.5, 1.0, 2.0$ as a function of the number of sampling locations n_s and sampling times n_t.

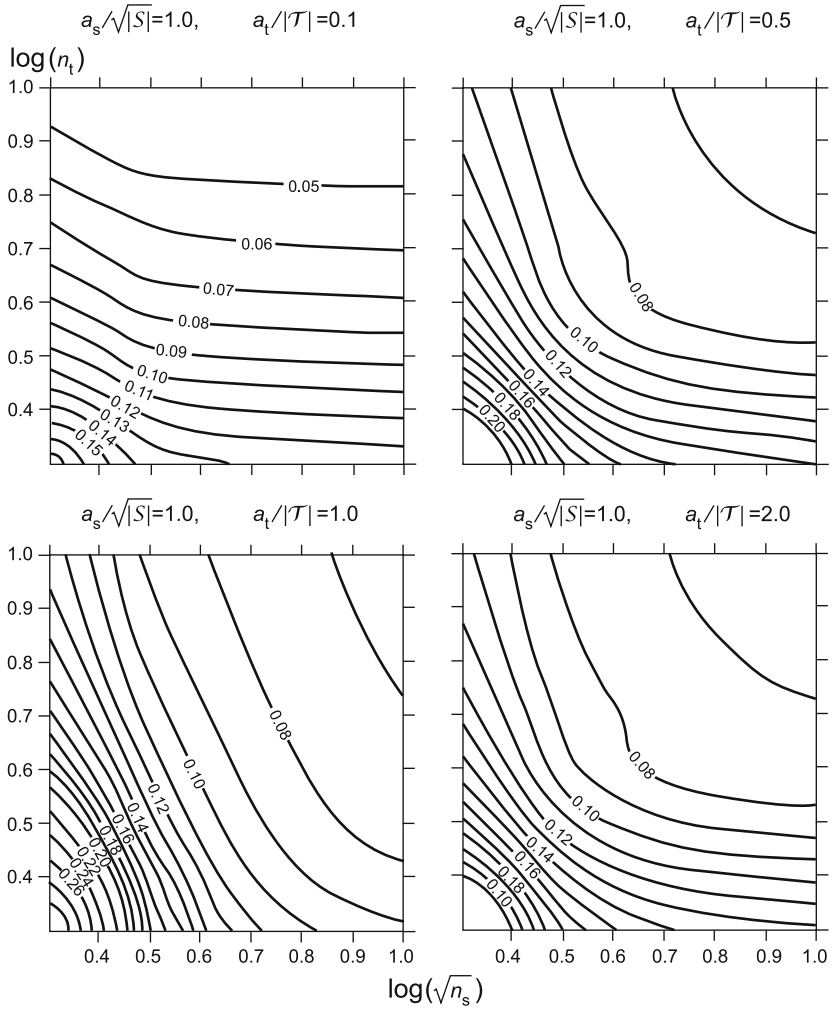


Fig. D.3. Sampling on a space–time grid (square grid pattern in space) for predicting the spatio-temporal mean by space–time kriging. The figure shows the variance ratio $\sigma^2_{\mathrm{obk}}/\sigma^2$ for $a_s/\sqrt{|\mathcal{S}|} = 1$ and $a_t/|\mathcal{T}| = 0.1, 0.5, 1.0, 2.0$ as a function of the number of sampling locations n_s and sampling times n_t.

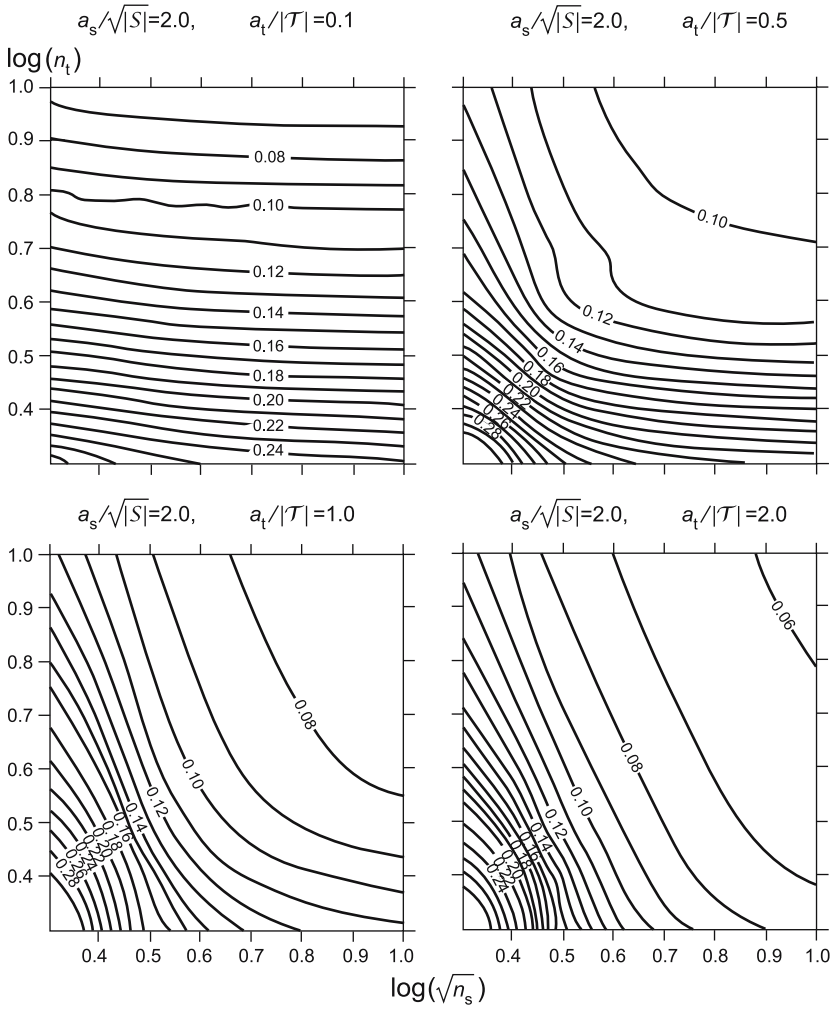


Fig. D.4. Sampling on a space–time grid (square grid pattern in space) for predicting the spatio-temporal mean by space–time kriging. The figure shows the variance ratio $\sigma^2_{\mathrm{obk}}/\sigma^2$ for $a_s/\sqrt{|\mathcal{S}|} = 2$ and $a_t/|\mathcal{T}| = 0.1, 0.5, 1.0, 2.0$ as a function of the number of sampling locations n_s and sampling times n_t.

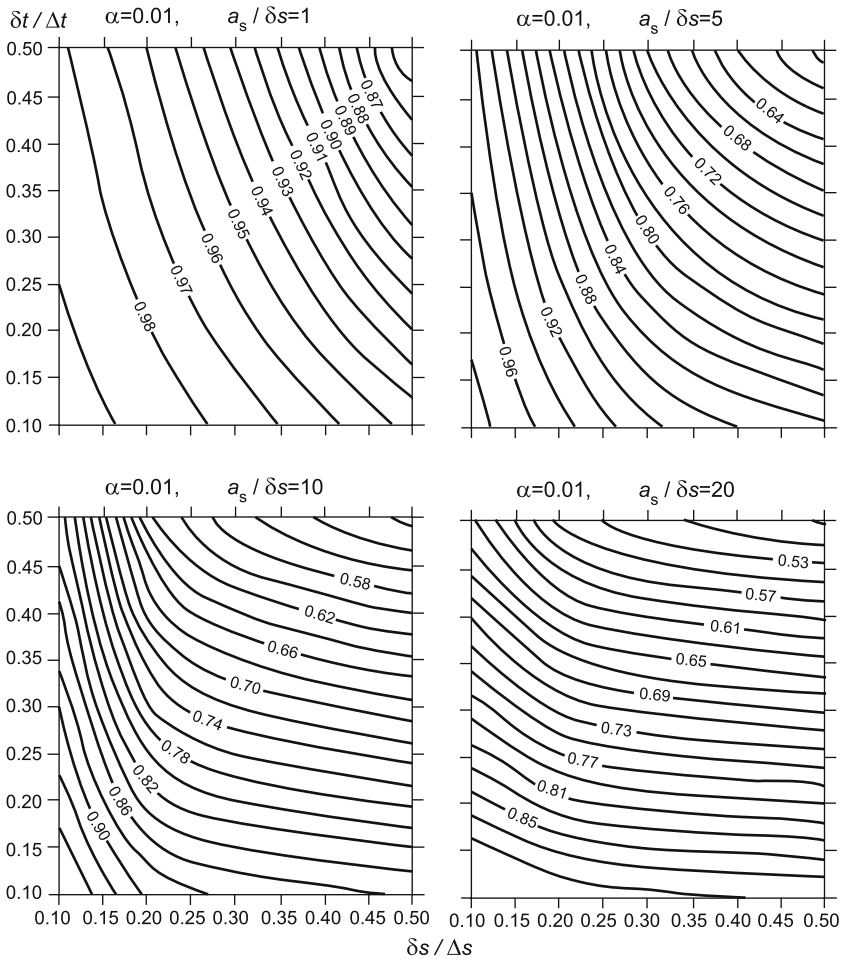


Fig. D.5. Lines of equal normalised prediction-error variance $\overline{\sigma^2_{\mathrm{kf}}}/\sigma^2_{\mathbf{Q}}$ for space–time prediction of groundwater heads with a Kalman filter using Grid Sampling. The covariance model is spherical with parameters a_s and $\sigma^2_{\mathbf{Q}}$. Here $a_s/\delta s = 1$, 5, 10 and 20 and the system property $\alpha = 0.01$

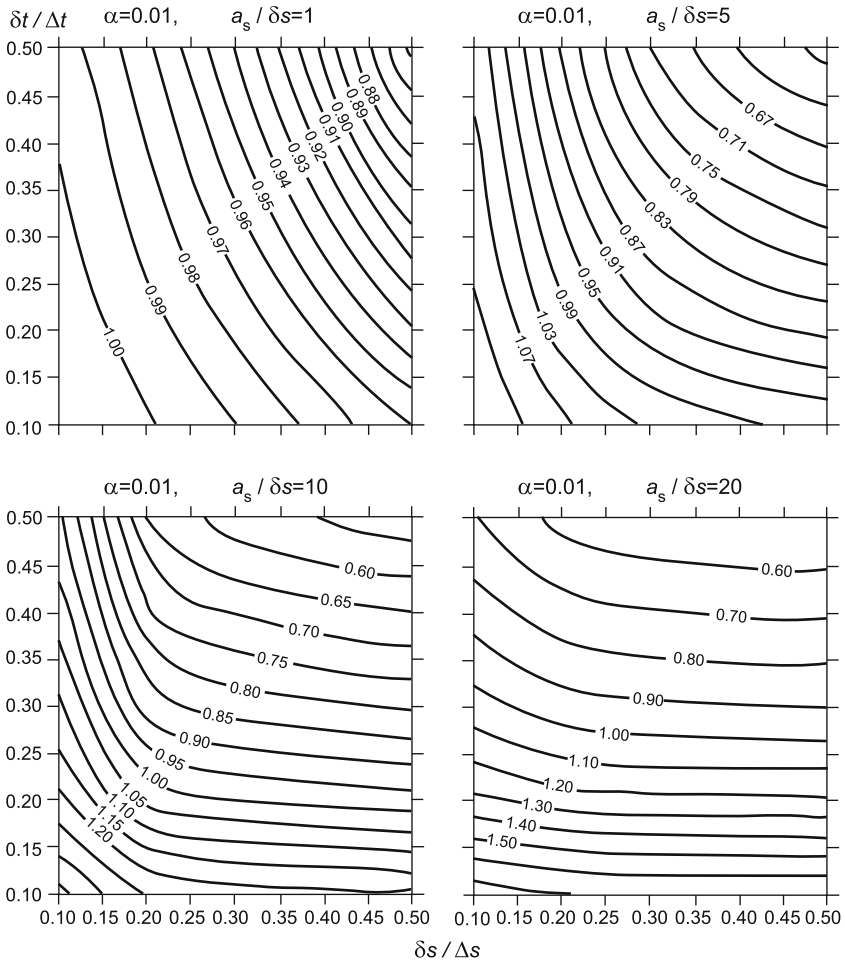


Fig. D.6. Lines of equal normalised prediction-error variance $\overline{\sigma^2_{\mathrm{kf}}}/\sigma^2_{\mathbf{Q}}$ for space–time prediction of groundwater heads with a Kalman filter using Grid Sampling. The covariance model is spherical with parameters a_s and $\sigma^2_{\mathbf{Q}}$. Here $a_s/\delta s = 1$, 5, 10 and 20 and the system property $\alpha = 0.1$

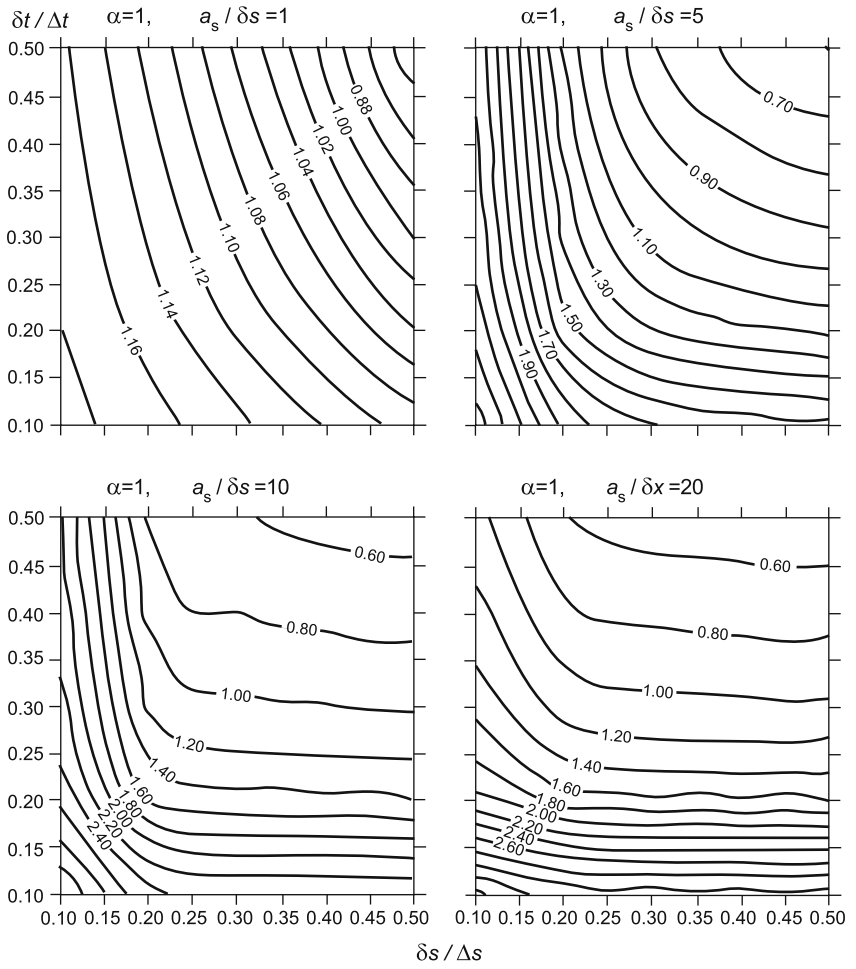


Fig. D.7. Lines of equal normalised prediction-error variance $\overline{\sigma^2_{\mathrm{kf}}}/\sigma^2_{\mathbf{Q}}$ for space–time prediction of groundwater heads with a Kalman filter using Grid Sampling. The covariance model is spherical with parameters a_s and $\sigma^2_{\mathbf{Q}}$. Here $a_s/\delta s = 1$, 5, 10 and 20 and the system property $\alpha = 1$

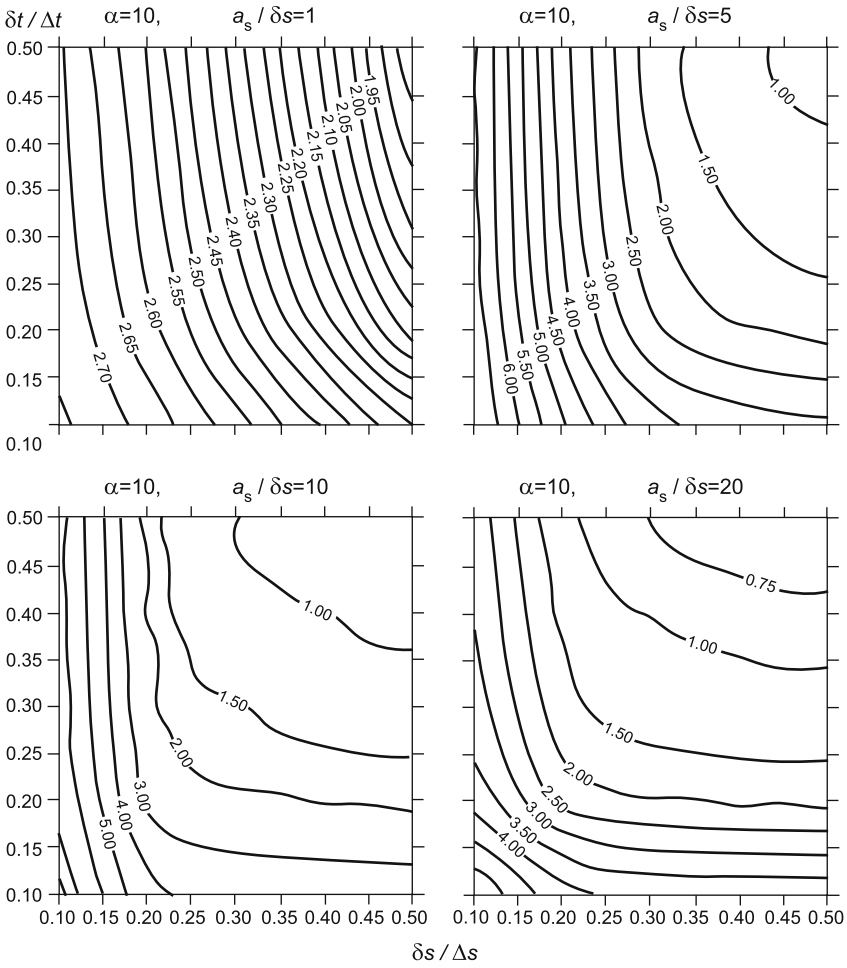


Fig. D.8. Lines of equal normalised prediction-error variance $\overline{\sigma^2_{\mathrm{kf}}}/\sigma^2_{\mathbf{Q}}$ for space–time prediction of groundwater heads with a Kalman filter using Grid Sampling. The covariance model is spherical with parameters a_s and $\sigma^2_{\mathbf{Q}}$. Here $a_{\mathrm{s}}/\delta s = 1$, 5, 10 and 20 and the system property $\alpha = 10$

References

Angulo, J. M., González-Manteiga, W., Febrero-Bande, M., and Alonso, F. J. (1998). Semi-parametric statistical approaches for space-time prediction. *Environmental and Ecological Statistics*, 5:297–316.

Aspie, D. and Barnes, R. J. (1990). Infill-sampling design and the cost of classification errors. *Mathematical Geology*, 22:915–932.

Baker, A. S., Kuo, S., and Chae, Y. M. (1981). Comparison of arithmetic soil pH values with the pH values of composite samples. *Soil Science Society of America Journal*, 45:828–830.

Bayley, G. V. and Hammersley, J. M. (1946). The effective number of independent observations in an autocorrelated time series. *Journal of the Royal Statistical Society*, 8:184–197.

Bear, J. (1972). *Dynamics of Fluids in Porous Materials*. Elsevier Science, New York.

Benedetti, R. and Palma, D. (1995). Optimal sampling designs for dependent spatial units. *Environmetrics*, 6:101–114.

Bennet, A. F. (1992). *Inverse Methods in Physical Oceanography*. Cambridge University Press, New York.

Bie, S. W. and Ulph, A. (1972). The economic value of soil survey information. *Journal of Agricultural Economics*, 23:285–297.

Bierkens, M. F. P. (2001). Spatio-temporal modelling of the soil water balance using a stochastic model and soil profile descriptions. *Geoderma*, 103:27–50.

Bierkens, M. F. P. (2002). Optimising a monitoring network for groundwater pollution using stochastic simulation and a cost model. In Hassanizadeh, S. M., Schotting, R. J., Gray, W. G., and Pinder, G. F., editors, *Computational Methods in Water Resources, Proceedings of the XIVth International Conference on Computational Methods in Water Resources, Delft, The Netherlands, June 23-28, 2002*, volume 47 of *Developments in Water Science*, pages 1443–1450. Elsevier, Amsterdam.

Bierkens, M. F. P., Knotters, M., and Hoogland, T. (2001). Space-time modelling of water table depth using a regionalized time series model and the Kalman filter. *Water Resources Research*, 37:1277–1290.

Bierkens, M. F. P., Knotters, M., and van Geer, F. C. (1999). Calibration of transfer function-noise models to sparsely or irregularly observed time series. *Water Resources Research*, 35:1741–1750.

Binder, D. A. and Hidiroglou, M. A. (1988). Sampling in time. In Krishnaiah, P. R. and Rao, C. R., editors, *Handbook of Statistics*, volume 6, pages 187–211. North-Holland, Amsterdam.

Boer, E. P. J., Hendrix, E. M. T., and Rasch, D. A. M. K. (2001). Optimization of monitoring networks for the estimation of the semivariance function. In Atkinson, A. C., Hackl, P., and Müller, W. G., editors, *mODa 6 - Advances in Model-Oriented Design and Analysis, Proceedings of the 6th international workshop on model-oriented design and analysis held in Puchberg, June 25-29, 2001*, pages 21–28. Physica Verlag, Heidelberg.

Bogaert, P. and Christakos, G. (1997a). Spatio-temporal analysis and processing of thermometric data over Belgium. *Journal of Geophysical Research*, 102:25831–25846.

Bogaert, P. and Christakos, G. (1997b). Stochastic analysis of spatiotemporal solute concent measurements using a regression model. *Stochastic Hydrology and Hydraulics*, 11:267–295.

Bogaert, P. and Russo, D. (1999). Optimal spatial sampling design for the estimation of the variogram based on a least squares approach. *Water Resources Research*, 35:1275–1289.

Boswell, M. T., Gore, S. D., Lovison, G., and Patil, G. P. (1996). Annotated bibliography of composite sampling, Part A. *Environmental and Ecological Statistics*, 3:1–49.

Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis, Forecasting and Control*. Holden-Day, San Francisco. Revised edition.

Bras, R. L., Buchanan, R., and Curry, K. C. (1983). Real time adaptive closed-loop control of reservoirs with the High Aswan Dam as a case study. *Water Resources Research*, 19:33–52.

Breidt, F. J. (1995). Markov chain designs for one-per-stratum spatial sampling. *Survey Methodology*, 21:63–70.

Brown, G. H. and Fisher, N. I. (1972). Subsampling a mixture of sampled material. *Technometrics*, 14:663–668.

Brus, D. J. (1994). Improving design-based estimation of spatial means by soil map stratification. A case study of phosphate saturation. *Geoderma*, 62:233–246.

Brus, D. J. (2000). Using regression models in design-based estimation of spatial means of soil properties. *European Journal of Soil Science*, 51:159–172.

Brus, D. J. and de Gruijter, J. J. (1994). Estimation of nonergodic variograms and their sampling variance by design-based sampling strategies. *Mathematical Geology*, 26:437–454.

Brus, D. J. and de Gruijter, J. J. (1997). Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with Discussion). *Geoderma*, 80:1–59.

Brus, D. J. and de Gruijter, J. J. (2003). A method to combine non-probability sample data with probability sample data in estimating spatial means of environmental variables. *Environmental Monitoring and Assessment*, 83:303–317.

Brus, D. J., de Gruijter, J. J., and van Groenigen, J. W. (2003). Designing spatial coverage samples by the k-means clustering algorithm. In *Proceedings of 8th International FZK/TNO conference on contaminated soil (ConSoil 2003)*, pages 504–509, Gent, Belgium.

Brus, D. J., Jansen, M. J. W., and de Gruijter, J. J. (2002). Optimizing two- and three-stage designs for spatial inventories of natural resources by simulated annealing. *Environmental and Ecological Statistics*, 9:71–88.

Brus, D. J., Spätjens, L. E. E. M., and de Gruijter, J. J. (1999). A sampling scheme for estimating the mean extractable phosphorus concentration of fields for environmental regulation. *Geoderma*, 89:129–148.

Brus, D. J. and te Riele, W. J. M. (2001). Design-based regression estimators for spatial means of soil properties: the use of two-phase sampling when the means of the auxiliary variables are unknown. *Geoderma*, 104:257–279.

Bueso, M. C., Angulo, J. M., and Alonso, F. J. (1998). A state-space model approach to optimum spatial sampling design based on entropy. *Environmental and Ecological Statistics*, 5:29–44.

Cameron, D. R., Nyborg, M., Toogood, J. A., and Laverty, D. H. (1971). Accuracy of field sampling for soil tests. *Canadian Journal of Soil Science*, 51:165–175.

Carson, J. H. (2001). Analysis of composite sampling data using the principle of maximum entropy. *Environmental and Ecological Statistics*, 8:201–211.

Carter, R. E. and Lowe, L. E. (1986). Lateral variability of forest floor properties under second-growth Douglas fir stands and the usefulness of composite sampling techniques. *Canadian Journal of Forest Research*, 16:1128–1132.

Chao, C.-T. and Thompson, S. K. (2001). Optimal adaptive selection of sampling sites. *Environmetrics*, 12:517–538.

Chauduri, A. (1994). Small domain statistics: a review. *Statistica Neerlandica*, 48:215–236.

Chen, Z., Bai, Z., and Sinha, B. K. (2004). *Ranked Set Sampling. Theory and Applications*. Springer, New York.

Christakos, G. and Killam, B. R. (1993). Sampling design for classifying contaminant level using annealing search algorithms. *Water Resources Research*, 29:4063–4076.

Cieniawski, S. E., Ehaert, J. W., and Ranjithan, S. (1995). Using genetic algorithms to solve a multiobjective groundwater monitoring problem. *Water Resources Research*, 31:399–409.

Cochran, W. G. (1977). *Sampling Techniques*. Wiley, New York.

Correll, R. L. (2001). The use of composite sampling in contaminated sites - a case study. *Environmental and Ecological Statistics*, 8:185–200.

Corsten, L. C. A. and Stein, A. (1994). Nested sampling for estimating spatial semivariograms compared to other designs. *Applied Stochastic Models and Data Analysis*, 10:103–122.

Courtin, P., Feller, M. C., and Klinka, K. (1983). Lateral variability in some properties of disturbed forest soils in southwestern British Columbia. *Canadian Journal of Soil Science*, 63:529–539.

Cressie, N. A. C. (1985). Fitting variogram models by weighted least squares. *Mathematical Geology*, 17:563–586.

Cressie, N. A. C. (1993). *Statistics for Spatial Data*. Wiley, New York. Revised edition.

Daley, R. (1991). *Atmospheric Data Analysis*. Cambridge University Press, New York.

De Cesare, L., Myers, D. E., and Posa, D. (2001). Product–sum covariance for space–time modeling: an environmental application. *Environmetrics*, 12:11–23.

de Gruijter, J. J. and Marsman, B. A. (1984). Transect sampling for reliable information on mapping units. In Bouma, J., editor, *Proceedings of the SSSA-ISSS Workshop on Spatial Variability, Las Vegas, U.S.A., Nov. 30 - Dec. 1, 1984*. Pudoc, Wageningen.

de Gruijter, J. J. and ter Braak, C. J. F. (1990). Model-free estimation from spatial samples: a reappraisal of classical sampling theory. *Mathematical Geology*, 22:407–415.

De Iaco, S., Myers, D. E., and Posa, D. (2003). The linear coregionalization model and the product–sum space–time variogram. *Mathematical Geology*, 35:25–38.

de Marsily, G. (1986). *Quantitative Hydrogeology; Groundwater Hydrology for Engineers*. Academic Press, Orlando.

Deutsch, C. V. and Journel, A. G. (1998). *GSLIB, Geostatistical Software Library and User's Guide*. Oxford University Press, New York, second edition.

Dixon, W. and Chiswell, B. (1996). Review of aquatic monitoring program design. *Water Research*, 30:1935–1948.

Domburg, P., de Gruijter, J. J., and Brus, D. J. (1994). A structured approach to designing soil survey schemes with prediction of sampling error from variograms. *Geoderma*, 62:151–164.

Domburg, P., de Gruijter, J. J., and van Beek, P. (1997). Designing efficient soil survey schemes with a knowledge-based system using dynamic programming. *Geoderma*, 75:183–201.

Draper, N. R. and Smith, H. (1981). *Applied Regression Analysis*. Wiley, New York.

Duncan, G. J. and Kalton, G. (1987). Issues of design and analysis of surveys across time. *International Statistical Review*, 55:97–117.

Elder, R. S., Thompson, W. O., and Myers, R. H. (1980). Properties of composite sampling procedures. *Technometrics*, 22:179–186.

Englund, E. J. and Heravi, N. (1994). Phased sampling for soil remediation. *Environmental and Ecological Statistics*, 1:247–263.

EPA (2000). Guidance for the Data Quality Objectives Process. Technical Document EPA QA/G-4, United States Environmental Protection Agency, Washington, DC 20460.

Evensen, G. (1992). Using the extended Kalman filter with a multilayer quasi-geostrophic ocean model. *Journal of Geophysical Research*, 97:17905–17924.

Evensen, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research*, 99:10143–10162.

Evensen, G. and van Leeuwen, P. J. (1995). Advanced data assimilation based on ensemble statistics. In *Second International Symposium on Assimilation of Observations in Meteorology and Ocenaography*, pages 153–158, Tokyo. W.M.O.

Fuller, W. A. (1999). Environmental surveys over time. *Journal of Agricultural, Biological, and Environmental Statistics*, 4:331–345.

Gershenfeld, N. (1999). *The Nature of Mathematical Modeling.* Cambridge University Press, Cambridge.

Gilbert, R. O. (1987). *Statistical Methods for Environment Pollution Monitoring.* Van Nostrand Reinhold, New York.

Gilbert, R. O., Wilson, J. E., O'Brien, R. F., Carlson, D. K., Bates, D. J., Pulsipher, B. A., and McKinstry, C. A. (2002). Version 2.0 Visual Sample Plan (VSP) models and code verification. Report PNNL-13991, Pacific Northwest National Laboratory, Richland Washington.

Gneiting, T. (2002). Non-separable, stationary covariance functions for space–time data. *Journal of the American Statistical Association*, 97:590–600.

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

Gore, S. D. and Patil, G. P. (1994). Identifying extremely large values using composite sample data. *Environmental and Ecological Statistics*, 1:227–245.

Gore, S. D., Patil, G. P., and Taillie, C. (1996). Identification of the largest individual sample value using composite sample data and certain modifications of the sweep-out method. *Environmental and Ecological Statistics*, 3:219–234.

Gore, S. D., Patil, G. P., and Taillie, C. (2001). Identifying the largest individual sample value from a two-way composite sampling design. *Environmental and Ecological Statistics*, 8:151–162.

Gy, P. M. (1979). *Sampling of Particulate Materials.* Elsevier, Amsterdam.

Hansen, M. H., Madow, W. G., and Tepping, B. J. (1983). An evaluation of model-dependent and probability sampling inferences in sample-surveys. *Journal of the American Statistical Association*, 78:805–807.

Hartfield, M. I. and Gunst, R. F. (2003). Identification of model components for a class of continuous spatiotemporal models. *Journal of Agricultural, Biological, and Environmental Statistics*, 8:105–121.

Hartigan, J. A. (1975). *Clustering Algorithms.* Wiley, New York.

Heemink, A. W. (2000). Modeling and prediction of environmental data in space and time using Kalman filtering. In Heuvelink, G. B. M. and Lemmens, M. J. P. M., editors, *Accuracy 2000, Proceedings of the 4th Internatiuonal Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Amsterdam, July 2000*, pages 283–291. Delft University Press, Delft.

Heemink, A. W. and Kloosterhuis, H. (1990). Data-assimilation for non-linear tidal models. *International Journal for Numerical Methods in Fluids*, 11:1097–1112.

Heemink, A. W., Verlaan, M., and Segers, A. J. (2001). Variance reduced ensemble Kalman filtering. *Monthly Weather Review*, 129:1718–1728.

Heuvelink, G. B. M., Musters, P., and Pebesma, E. J. (1997). Spatio-temporal kriging of soil water content. In Baafi, E. and Schofield, N., editors, *Geostatistics Wollongong '96*, volume 1, pages 1020–1030. Kluwer Academic Publishers, Dordrecht.

Hidiroglou, M. A. and Särndal, C. E. (1985). An empirical study of some regression estimators for small domains. *Survey Methodology*, 11:65–77.

Hipel, K. W., Lennox, W. C., Unny, T. E., and McLeod, A. I. (1975). Intervention analysis in water resources. *Water Resources Research*, 11:855–861.

Hipel, K. W. and McLeod, A. I. (1994). *Time Series Modelling of Water Resources and Environmental Systems*. Elsevier, New York.

Hirsch, R. M. and Slack, J. R. (1984). A nonparametric trend test for seasonal data with serial dependence. *Water Resources Research*, 20:727–732.

Hirsch, R. M., Slack, J. R., and Smith, R. A. (1982). Techniques of trend analysis for monthly water quality data. *Water Resources Research*, 18:107–121.

Hoeben, R. and Troch, P. A. (2000). Assimilation of microwave observation data for soil moisture profile estimation. *Water Resources Research*, 36:2805–2819.

Houser, P. R., Shuttleworth, W. J., Famiglietti, J. S., Gupta, H. V., Syed, K. H., and Goodrich, D. C. (1998). Integration of soil moisture remote sensing and hydrologic modeling using data assimilation. *Water Resources Research*, 34:3405–3420.

Hsiao, C. K., Juang, K. W., and Lee, D. Y. (2000). Estimating the second-stage sample size and the most probable number of hot spots from a first-stage sample of heavy-metal contaminated soil. *Geoderma*, 95:73–88.

Huang, H.-C. and Cressie, N. (1996). Spatio-temporal prediction of snow water equivalent using the Kalman filter. *Computational Statistics & Data Analysis*, 22:159–175.

Huerta, G., Sansó, B., and Stroud, J. R. (2004). A spatiotemporal model for Mexico City ozone levels. *Applied Statistics*, 53:231–248.

Isaaks, E. H. and Srivastava, R. H. (1988). Spatial continuity measures for probabilistic and deterministic geostatistics. *Mathematical Geology*, 20:313–341.

Isaaks, E. H. and Srivastava, R. H. (1989). *An Introduction to Applied Geostatistics*. Oxford University Press, New York.

Journel, A. G. and Huijbregts, C. J. (1978). *Mining Geostatistics*. Academic Press, New York.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering (ASME)*, 82D:35–45.

Kirpatrick, S. (1984). Optimization by simulated annealing: quantitative studies. *Journal of Statistical Physics*, 34:975–986.

Knotters, M. and Bierkens, M. F. P. (2001). Predicting water table depths in space and time using a regionalised time series model. *Geoderma*, 103:51–77.

Knotters, M. and Bierkens, M. F. P. (2002). Accuracy of spatio-temporal RARX model predictions of water table depth. *Stochastic Environmental Research and Risk Assessment*, 16:112–126.

Knotters, M., Brus, D. J., and Voshaar, J. H. O. (1995). A comparison of kriging, co-kriging and kriging combined with regression for spatial interpolation of horizon depth with censored observations. *Geoderma*, 67:227–246.

Koop, J. C. (1990). Systematic sampling of two-dimensional surfaces and related problems. *Communications in Statistics. Part A, Theory and Methods*, 19:1701–1750.

Krishnaiah, P. R. and Rao, C. R., editors (1988). *Sampling*, volume 6 of *Handbook of Statistics*. North-Holland, Amsterdam.

Kyriakidis, P. C. and Journel, A. G. (1999). Geostatistical space-time models: a review. *Mathematical Geology*, 31:651–684.

Kyriakidis, P. C. and Journel, A. G. (2001). Stochastic modeling of atmospheric pollution: a spatial time-series framework: 1, methodology; 2, application to monitoring monthly sulfate deposition over Europe. *Atmospheric Environment*, 35:2331–2348.

Lark, R. M. (2000). Estimating variograms of soil properties by the method-of-moments and maximum likelihood. *European Journal of Soil Science*, 51:717–728.

Lark, R. M. (2002). Optimized spatial sampling of soil for estimation of the variogram by maximum likelihood. *Geoderma*, 105:49–80.

Lettenmaier, D. P. (1976). Detection of trends in water quality data with dependent observations. *Water Resources Research*, 12:1037–1046.

Lettenmaier, D. P. (1978). Design considerations for ambient stream quality monitoring. *Water Resources Bulletin*, 14:884–902.

Lettenmaier, D. P., Hipel, K. W., and McLeod, A. I. (1978). Assessment of environmental impacts. Part two: data collection. *Environmental Management*, 2:537–554.

Loaiciga, H. A., Charbeneau, R. J., Everett, L. G., Fogg, G. E., Hobbs, B. F., and Rouhani, S. (1992). Review of ground-water quality monitoring network design. *Journal of Hydraulic Engineering*, 118:11–37.

Lophaven, S., Carstensen, J., and Rootzén, H. (2004). Space-time modeling of environmental monitoring data. *Environmental and Ecological Statistics*, 11:237–256.

Matalas, N. C. and Langbein, W. B. (1962). Information content of the mean. *Journal of Geophysical Research*, 67:3441–3448.

Matérn, B. (1986). *Spatial Variation*, volume 36 of *Lecture Notes in Statistics*. Springer-Verlag, New York, second edition.

Bogaert, P. (1996). Comparison of kriging techniques in a space-time context. *Mathematical Geology*, 28:73–86.

Duncan, A. J. (1962). Bulk sampling. Problems and lines of attack. *Technometrics*, 4:319–343.

McBratney, A. B. and Webster, R. (1981). The design of optimal sampling schemes for local estimation and mapping of regionalized variables: II Program and examples. *Computers & Geosciences*, 7:335–365.

McBratney, A. B., Webster, R., and Burgess, T. M. (1981). The design of optimal sampling schemes for local estimation and mapping of regionalized variables: I Theory and method. *Computers & Geosciences*, 7:331–334.

McLauglin, D. B. (1995). Recent advances in hydrologic data assimilation. *Reviews in Geophysics*, 33:977–984.

Meiring, W., Guttorp, P., and Sampson, P. D. (1998). Space-time estimation of grid-cell hourly ozone levels for assessment of a deterministic model. *Environmental and Ecological Statistics*, 5:197–222.

Meyer, P. D. and Bril, E. D. (1988). A method for locating wells in a groundwater monitoring network under conditions of uncertainty. *Water Resources Research*, 23:1277–1282.

Miesch, A. T. (1975). Variograms and variance components in geochemistry and ore evaluation. *Geological Society of America, Memoir*, 142:333–340.

Mode, N. A., Conquest, L. L., and Marker, D. A. (2002). Incorporating prior knowledge in environmental sampling: ranked set sampling and other double sampling procedures. *Environmetrics*, 13:513–521.

Morf, M., Lévy, B., and Kailath, T. (1974). Some new algorithms for recursive estimation in constant, linear, discrete-time systems. *IEEE Transactions on Automatic Control*, 19:315–323.

Mukhopadhyay, N., Bendel, R. B., Nikolaidis, N. P., and Chattopadhyay, S. (1992). Efficient sequential sampling strategies for environmental monitoring. *Water Resources Research*, 28:2245–2256.

Müller, W. G. (2001). *Collecting Spatial Data: Optimum Design of Experiments for Random Fields*. Physica-Verlag, Heidelberg, second edition.

Müller, W. G. and Zimmerman, D. L. (1999). Optimal designs for variogram estimation. *Environmetrics*, 10:23–27.

Myers, D. E. (2002). Space-time correlation models and contaminant plumes. *Environmetrics*, 13:535–553.

Nychka, D., Yang, Q., and Royle, J. A. (1997). Constructing spatial designs for monitoring air pollution using regression subset selection. In Barnett, V. and Turkman, K. F., editors, *Pollution Assessment and Control*, volume 3 of *Statistics for the Environment*. Wiley, Chichester.

Oliver, M. A. (1984). *Soil variation in the Wyre Forest: its elucidation and measurement*. PhD thesis, University of Birmingham.

Overton, W. S. and Stehman, S. V. (1996). Desirable design characteristics for long-term monitoring of ecological variables. *Environmental and Ecological Statistics*, 3:349–361.

Papritz, A. and Flühler, H. (1994). Temporal change of spatially autocorrelated soil properties: optimal estimation by cokriging. *Geoderma*, 62:29–43.

Papritz, A. and Webster, R. (1995). Estimating temporal change in soil monitoring: I. statistical theory. *European Journal of Soil Science*, 46:1–12.

Pardo-Igúzquiza, E. and Dowd, P. A. (1998). Maximum likelihood inference of spatial covariance parameters of soil properties. *Soil Science*, 163:212–219.

Patil, G. P. and Rao, C. R., editors (1994). *Environmental statistics*, volume 12 of *Handbook of Statistics*. North-Holland, Amsterdam.

Patil, G. P., Sinha, A. K., and Taillie, C. (1993). Relative precision of ranked set sampling: a comparison with the regression estimator. *Environmetrics*, 4:399–412.

Patil, G. P., Sinha, A. K., and Taillie, C. (1994). Ranked set sampling. In Patil, G. P. and Rao, C. R., editors, *Environmental Statistics*, volume 12 of *Handbook of Statistics*, pages 167–200. North-Holland, Amsterdam.

Pebesma, E. J. (2004). Multivariable geostatistics in S: the gstat package. *Computers & Geosciences*, 30:683–691.

Pettit, A. N. and McBratney, A. B. (1993). Sampling designs for estimating spatial variance components. *Applied Statistics*, 42:185–209.

Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T. (1988). *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press, New York.

Priestley, M. B. (1981). *Spectral Analysis and Time Series*. Academic Press, London.

Quenouille, M. H. (1949). Problems in plane sampling. *Annals of Mathematical Statistics*, 20:355–375.

Reed, J. F. and Rigney, J. A. (1947). Soil sampling from fields of uniform and nonuniform appearance and soil type. *Journal of the American Society of Agronomists*, 39:26–40.

Reichle, R. H., Entekhabi, D., and McLauglin, D. B. (2001). Downscaling of radio brightness measurements for soil moisture estimation: A four dimensional variational data assimilation approach. *Water Resources Research*, 37:2353–2364.

Rodríguez-Iturbe, I. and Meíja, J. M. (1974). The design of rainfall networks in space and time. *Water Resources Research*, 10:713–728.

Rohde, C. A. (1976). Composite sampling. *Biometrics*, 32:273–282.

Rouhani, S. and Hall, T. J. (1989). Space-time kriging of groundwater data. In Armstrong, M., editor, *Geostatistics Avignon '88*, pages 639–650. Kluwer Academic Publishers, Dordrecht.

Rouhani, S. and Myers, D. E. (1990). Problems in space-time kriging of geohydrological data. *Mathematical Geology*, 22:611–623.

Rouhani, S. and Wackernagel, H. (1990). Multivariate geostatistical approach to space-time data analysis. *Water Resources Research*, 26:585–591.

Royle, J. A. and Nychka, D. (1998). An algorithm for the construction of spatial coverage designs with implementation in SPLUS. *Computers & Geosciences*, 24:479–488.

Ruark, G. A., Mader, D. L., and Tattar, T. A. (1982). A composite sampling technique to assess urban soils under roadside trees. *Journal of Arboriculture*, 8:96–99.

Sacks, J. and Schiller, S. (1988). Spatial designs. In Gupta, S. and Berger, J., editors, *Statistical Decision Theory and Related Topics IV*, volume 2. Springer Verlag, New York.

Sanders, T. G. and Adrian, D. D. (1978). Sampling frequency for river quality monitoring. *Water Resources Research*, 14:569–576.

Särndal, C. E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer Verlag, New York.

Schreuder, H. T., Gregoire, T. G., and Wood, G. B. (1993). *Sampling Methods for Multiresource Forest Inventory*. Wiley, New York.

Schreuder, H. T., Li, H. G., and Scott, C. T. (1987). Jackknife and bootstrap estimation for sampling with partial replacement. *Forest Science*, 33:676–689.

Schweppe, F. C. (1973). *Uncertain Dynamic Systems*. Prentice-Hall, Englewood Cliffs.

Searle, S. R. (1997). *Linear Models*. Wiley, New York.

Senegas, J., Wackernagel, H., Rosenthal, W., and Wolf, T. (2001). Error covariance modeling in sequential data assimilation. *Stochastic Environmental Research and Risk Assessment*, 15:65–86.

Snepvangers, J. J. J. C., Heuvelink, G. B. M., and Huisman, J. A. (2003). Soil water content interpolation using spatio-temporal kriging with external drift. *Geoderma*, 112:253–271.

Sølna, K. and Switzer, P. (1996). Time trend estimation for a geographic region. *Journal of the American Statistical Association*, 19:577–589.

Splitstone, D. E. (2001). Sample support and related scale issues in composite sampling. *Environmental and Ecological Statistics*, 8:137–149.

Stauffer, D. R. and Seaman, N. L. (1990). Use of four-dimensional data-assimilation in a limited area mesoscale model: I, experiments with synoptic-scale data. *Monthly Weather Review*, 118:1250–1277.

Stevens, D. L. (1997). Variable density grid-based sampling designs for continuous spatial populations. *Environmetrics*, 8:167–195.

Stevens, D. L. and Olsen, A. R. (2003a). Spatially restricted surveys over time for aquatic resources. *Journal of Agricultural, Biological, and Environmental Statistics*, 4:415–428.

Stevens, D. L. and Olsen, A. R. (2003b). Variance estimation for spatially balanced samples of environmental resources. *Environmetrics*, 14:593–610.

Stewart-Oaten, A. and Bence, J. R. (2001). Temporal and spatial variation in environmental impact assessment. *Ecological Monographs*, 71:305–339.

Sunter, A. B. (1977). List sequential sampling with equal or unequal probabilities without replacement. *Applied Statistics*, 26:261–268.

te Stroet, C. B. M. (1995). *Calibration of Stochastic Groundwater Flow Models; Estimation of System Noise Statistics and Model Parameters*. PhD thesis, Delft University of Technology.

Thompson, S. K. (2002). *Sampling*. Wiley, New York, second edition.

Thompson, S. K. and Seber, G. A. F. (1996). *Adaptive Sampling*. Wiley, New York.

Tucker, P., Ferguson, C. C., Abbachi, A., and Nathanai, P. (1996). A decision support system for designing sampling strategies for potentially contaminated sites. In Kraak, M. J. and Molenaar, M., editors, *Spatial Data handling. Proceedings of the Seventh International Symposium on Spatial Data Handling, Delft, The Netherlands, August 1996*, volume 1, pages 1–12. International Geographical Union Commission on Geographic Information Systems, Association for Geographical Information.

Underwood, A. J. (1994). On beyond BACI: sampling designs that might reliably detect environmental disturbances. *Ecological Applications*, 4:3–15.

Urquhart, N. S. and Kincaid, T. M. (1999). Designs for detecting trend from repeated surveys of ecological resources. *Journal of Agricultural, Biological and Environmental Statistics*, 4:404–414.

Valstar, J. (2001). *Inverse problem theory of coupled groundwater flow and mass transport*. PhD thesis, Delft University of Technology.

van Belle, G. and Hughes, J. P. (1984). Nonparametric tests for trend in water quality. *Water Resources Research*, 20:127–136.

van Geer, F. C., te Stroet, C. B. M., and Zhou, Y. (1991). Using Kalman filtering to improve and quantify the uncertainty of numerical groundwater simulations, 1: the role of system noise and its calibration. *Water Resources Research*, 27:1987–1994.

van Groenigen, J. W. and Stein, A. (1998). Constrained optimization of spatial sampling using continuous simulated annealing. *Journal of Environmental Quality*, 27:1078–1086.

van Leeuwen, P. J. (2003). A variance-minimizing filter for large-scale applications. *Monthly Weather Review*, 131:2071–2084.

van Groenigen, J. W., Siderius, W., and Stein, A. (1999). Constrained optimisation of soil sampling for minimisation of the kriging variance. *Geoderma*, 87:239–259.

Verlaan, M. (1998). *Efficient Kalman filtering algorithms for hydrodynamic models*. PhD thesis, Delft University of Technology.

Verlaan, M. and Heemink, A. W. (1997). Tidal flow forecasting using reduced-rank square-root filters. *Stochastic Hydrology and Hydraulics*, 11:349–368.

Vos, J. W. E. (1964). Sampling in space and time. *Review of the International Statistical Institute*, 32:226–241.

Wackernagel, H. (1994). Cokriging versus kriging in regionalized multivariate data analysis. *Geoderma*, 62:83–92.

Walker, J. P. and Houser, P. R. (2001). A methodology for initializing soil moisture in a global climate model: assimilation of near-surface soil moisture observations. *Journal of Geophysical Research*, 106:11761–11774.

Warrick, A. W. and Myers, D. E. (1987). Optimization of sampling locations for variogram calculations. *Water Resources Research*, 23:496–500.

Watson, A. G. and Barnes, R. J. (1995). Infill sampling criteria to locate extremes. *Mathematical Geology*, 27:589–608.

Webster, R. and Burgess, T. M. (1984). Sampling and bulking strategies for estimating soil properties in small regions. *Journal of Soil Science*, 35:127–140.

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

Webster, R. and Oliver, M. A. (1992). Sample adequately to estimate variograms of soil properties. *Journal of Soil Science*, 43:177–192.

Whitehead, J. (1997). *The Design and Analysis of Sequential Clinical Trials, Revised 2nd ed.* Wiley, Chichester.

Wikle, C. K. and Cressie, N. (1999). A dimension-reduced approach to space-time Kalman filtering. *Biometrika*, 86:815–829.

Wikle, C. K. and Royle, J. A. (1999). Space–time dynamic design of environmental monitoring networks. *Journal of Agricultural, Biological, and Environmental Statistics*, 4:489–507.

Williams, L. R., Leggett, R. W., Espegren, M. L., and Little, C. A. (1989). Optimization of sampling for the determination of mean radium-226 concentration in surface soil. *Environmental Monitoring and Assessment*, 12:83–96.

Yates, F. (1981). *Sampling Methods for Censuses and Surveys*. Griffin, London.

Yfantis, E. A., Flatman, G. T., and Behar, J. V. (1987). Efficiency of kriging estimation for square, triangular and hexagonal grids. *Mathematical Geology*, 19:183–205.

Young, P. and Beven, K. (1994). Data-based mechanistic modelling and the rainfall-flow non-linearity. *Environmetrics*, 5:335–363.

Yue, S., Pilon, P., and Cavadias, G. (2002). Power of the Mann-Kendall and Spearman's rho tests for detecting monotonic trends in hydrological series. *Journal of Hydrology*, 259:254–271.

Zhang, X. F., Heemink, A. W., Janssen, L. H. J. M., Janssen, P. H. M., and Sauter, F. J. (1999). A computationally efficient Kalman smoother for the evaluation of the ch4 budget in europe. *Applied Mathematical Modelling*, 23:109–129.

Zhou, Y. (1996). Sampling frequency for monitoring the actual state of groundwater systems. *Journal of Hydrology*, 180:301–318.

Zhou, Y., te Stroet, C. B. M., and van Geer, F. C. (1991). Using Kalman filtering to improve and quantify the uncertainty of numerical groundwater simulations, 2, application to monitoring network design. *Water Resources Research*, 27:1995–2006.

Zimmerman, D. L. and Homer, K. E. (1991). A network design criterion for estimating selected attributes of the semivariogram. *Environmetrics*, 2:425–441.

Index