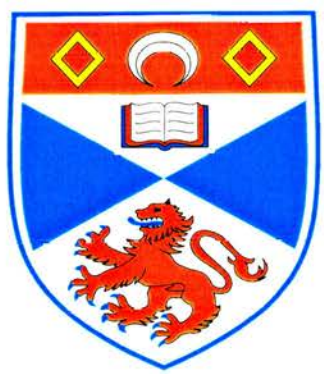
Design-based Adaptive Monitoring Strategies for Wildlife Population Assessment

Fiona Mary Underwood



Thesis submitted for the degree of

DOCTOR OF PHILOSOPHY

in the

School of Mathematics and Statistics

UNIVERSITY OF ST ANDREWS

April 2004

Declarations

1. I, *Fiona Mary Underwood*, hereby certify that this thesis, which is approximately 50,000 words in length, has been written by me, that it is the record of work carried out by me and that it has not been submitted in any previous application for a higher degree.

   date 15/4/04 signature of candidate

2. I was admitted as a research student in October 2000 and as a candidate for the degree of PhD in October 2001; the higher study for which this is a record was carried out in the University of St Andrews between 2000 and 2004.

   date 15/4/04 signature of candidate

3. I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of PhD in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree.

   date 15/4/04 signature of supervisor

4. In submitting this thesis to the University of St Andrews I understand that I am giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. I also understand that the title and abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker.

   date 15/4/04 signature of candidate

Abstract

Wildlife managers and conservationists need data about the status and trends in populations of animal or plant species to make informed decisions about their management. To obtain reliable estimates, survey design and data collection should be coordinated under a monitoring strategy. Efficient survey design is often difficult because survey design is fixed at the start of the monitoring programme when there is little information about species distribution.

A design-based sampling strategy is proposed that allows the survey design to change through time. A predictive map of expected abundance, obtained from one survey, is used to influence future survey design. A proportion of the sample is selected using simple random sampling, and the rest selected by sampling with inclusion probability proportional to predicted abundance (πps). When the total number of individuals in the population is of interest, the sample is selected independently of previous survey samples, and unconditional inclusion probabilities are estimated. To estimate the change in the population through time, part of the sample from the previous survey is maintained and the covariance between estimators of the population total from different surveys is required. Analytic estimators of covariance do not use all the sampled data. This motivated the development of a bootstrap variance estimator. An extension for estimating covariances using all of the sampled data is proposed.

Simulation indicates that the strategies give more precise estimates of the population total and change in the population total through time than standard design-based methods. Key issues are the proportion of the sample selected using πps in each survey and the accuracy of the predictive map of expected abundance. The basic strategies are developed, and tested, on populations of motile individuals, using plot sampling. The estimation of the abundance of forest elephants motivates an application to distance sampling.

Acknowledgements

My supervisor Steve Buckland gave me the space to develop my own research ideas and did not rush me at times when these ideas came slowly. My second supervisor, David Elston, was always encouraging.

Early on in my PhD David Borchers provided me with a version of the R library WiSP, developed mainly by Walter Zucchini. This helped form the basis of the simulation work in this thesis.

Danny Pfeffermann and Yves Berger of Southampton University were both generous with their time when I wished to discuss various sampling issues with them.

I was given permission to use the data from Odzala National Park by Nigel Hunter and Rene Beyers. Thanks to Rene, and to Len Thomas who originally prepared this data, and to the many fieldworkers who were involved in the data collection.

I have been lucky in my colleagues at CREEM. Camilla Dixon provided the LaTeX template for this thesis and many cakes. In times of difficulty Camilla, Sharon Hedley and Monique Mackenzie could always be depended on for friendship and support. Liz Clarke and Mike Lonergan were always willing to discuss issues of a mathematical nature and Charles Paxton has been encouraging throughout. Sam Strindberg and Rachel Atkinson reminded me of what it is like to work in conservation.

My time to carry out this thesis was funded by an Engineering and Physical Sciences Research Council studentship award.

Finally, my love and thanks go to Bob who has provided me with encouragement, confidence, love, and support on everything from stochastic processes to cups of tea and all things in between.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Problem Statement

The management of wildlife populations becomes increasingly important as human needs put greater pressure on the natural resources of the world. To be effective, wildlife managers and conservationists need information about the status and trends in populations of animal or plant species. For a particular species this information may be simple descriptive summaries of the population, for example a parameter of interest may be the population total — the number of individuals in a pre-determined area at a particular point in time — or the change in the population total through time. These parameters may be useful for setting levels of sustainable offtake, to decide whether intervention is required to stop a species decline into extinction, or to contribute to international assessments of the status of species, as is the case with the IUCN Red List of Threatened Species (Hilton-Taylor, 2000). Alternatively more detailed information about the relationships between species abundance and past management decisions or environmental changes may be required. Possible examples might be the increase in population density with altitude, or with a decline in grazing pressure.

The information required by the wildlife manager is obtained by collecting data about

the study species in a particular area, for example a reserve. Except under the unusual circumstances in which all individuals in the population are known, it is not possible to observe every individual in the population, or to fully survey the whole area. Instead, at any time a survey of the area is carried out in which only a sample of the population is observed. The survey region is partitioned into a large number of contiguous units. A sample of these units is visited and the number of individuals observed within each unit recorded. In this thesis it is assumed that all the individuals that are present within a unit are observed, i.e. the detection probability is assumed to be one. There could be many circumstances in which this would not be true, and there is a large literature and range of strategies for coping with this, see Seber (1986) and Seber (1992) for overviews of these strategies and Borchers *et al.* (2002) for the development of a framework into which these strategies fall. The counts, together with additional information about the area, if available, are used to estimate the parameters of interest.

There are two different approaches to parameter estimation. In the first approach, model-based inference, parameter estimates and estimates of their precision depend partly on the observed data but also on an assumed model of the sampled population. This could be a model of the spatial distribution, if the population total is the focus, or for trend estimation a model of how the population total changes through time. To some extent it is possible to check the assumptions underlying this model, but the choice of model is sometimes seen as arbitrary, especially in circumstances where there is some controversy, when different models may be used to present different arguments. In the second approach, design-based inference, parameter estimates and their associated estimates of precision are derived solely from the probability that each unit was included in the sample. Hence once the survey has been designed, the method of calculating estimates of precision is relatively fixed. When dealing with controversial issues this can be considered desirable as there is little scope for arriving at different results depending on the choice of model.

In a design-based framework, units to be included in the sample are selected according to a probabilistic sampling scheme. A simple and commonly used scheme is simple random

sampling in which each unit has an equal chance of being included in the final sample. The precision of estimates of the population total and of the change in the population total can be improved by implementing an unequal probability scheme in which the probability of including a particular unit in the sample (the inclusion probability) is positively correlated with the number of individuals in that unit, the observed count. Clearly at the start of the survey these counts are unknown but other auxiliary variables describing habitat, topography, and human factors for the survey region, may be available. If species density is thought to be related to these variables then they can be used as proxies for the observed counts to determine the inclusion probabilities. Stratified sampling is one such sampling scheme.

Because data collection for a survey is usually expensive, surveys are often planned to address several questions simultaneously, which implies that they are expected to estimate several parameters of interest. Some parameters can be estimated from a single survey, such as the population total at a particular time, whilst others will require data from several surveys, for example the change in the population total through time. In practice a survey may be commissioned each time an estimate of the population total is required. A set of these surveys may then be used *post hoc* to estimate the change in the population total through time. However, unless the surveys have been designed with this aim in mind, it is generally the case that there is little consistency between the surveys. For example the area covered may differ from one survey to the next, or the sampling strategies may differ. In such circumstances, there are difficulties in combining estimates from successive surveys, or in comparing estimates between surveys.

Alternatively the wildlife manager may realise that a series of surveys is required to answer both short-term and long-term questions and that the survey design and data collection need to be coordinated through time under a monitoring strategy. The monitoring strategy would consist of defining the parameters to be estimated, choosing a sampling scheme for all the surveys, which is often fixed in advance of data collection, and determining the appropriate estimators for the parameters and their precision. Although this strategy

ensures that there is consistency between the surveys, it may be inefficient in the sense that parameters are estimated with low precision. By using an appropriate unequal probability sampling scheme greater precision could be obtained. Auxiliary variables, such as habitat and topography, are often available before the monitoring strategy is implemented but the relationship between these variables and species density is often unknown. Hence an appropriate set of inclusion probabilities cannot be chosen and so the sampling scheme will remain simple and inefficient.

Through time the relationship between species density and other variables can become better understood by analysing data from successive surveys. For example predictive maps of species density or habitat suitability can be derived from the observed counts and the auxiliary variables available for the survey region. However if the sampling design is predetermined at the start of a series of surveys then a large proportion of the resources may be spent sampling areas which are known to have none of the study species. This is not only an inefficient use of resources but can also be very demoralising for the survey teams who might spend a large amount of time not seeing any individuals of the species they are surveying. For certain species, knowledge of the biology or ecology may be inadequate and it becomes important to observe as many individuals as possible during the survey so that basic data may be recorded.

Within the design-based literature, strategies such as adaptive sampling (Thompson & Seber 1996) use data collected from early on in a survey to determine which other units should be included in that particular survey. The motivation behind these strategies is to increase the number of individuals of the study species that were observed, and is particularly relevant for species that occur in clusters. However there has not been much development of sampling strategies that allow data from a previous survey to change the sampling scheme for future surveys so that parameter estimation is more efficient whilst ensuring long-term consistency for trend estimation.

1.2 Aim of the thesis

This thesis investigates the extent to which it is possible to develop an adaptive design-based monitoring strategy through time. The aim of the adaptation is to increase the efficiency of parameters to be estimated within each survey, usually the population total, and to ensure long-term consistency so that parameters requiring data from several surveys, such as change in population total through time, are also estimated as precisely as possible. A further aim is to increase the number of individuals of the species observed compared with a non-adaptive strategy. The adaptation is based on the predictive maps of species density, which use the observed counts from past surveys and auxiliary variables over the survey region. These predictions help determine the inclusion probabilities required for the sampling design. Although the population total and change in population total are estimated within a design-based framework, the predictive maps of species density can be obtained within a model-based framework.

1.3 Thesis outline

The first two chapters provide background information. Chapter 2 gives the setting for the thesis. The objectives of a generic single-species monitoring programme are defined and a notation introduced to describe both data and survey methods. Given the survey methodology we describe how the data obtained from a complete census of the area could be used to define parameters that would meet the objectives of the monitoring programme. This enables some key concepts to be described in preparation for understanding how data from only a sample of the survey region can be used to estimate these parameters. These ideas about sample survey design, how they can incorporate auxiliary information and how they are used in monitoring strategies through time are described in Chapter 3. The last section in this chapter describes the thesis problem in detail.

Chapters 4 and 5 give the basic building blocks of an adaptive, or combined sampling,

strategy. In these two chapters we explore how we can use information from an initial survey to improve survey design and hence estimation in a second survey. The general strategy of adaptation is based on sampling with inclusion probabilities proportional to species density or habitat suitability, which is estimated from the data collected in the first survey. In Chapter 4, the aim is to estimate the population total as precisely as possible. In Chapter 5 the general principle is maintained but the emphasis now is to estimate the change in the population total from one survey to another. This is a new application of the two-phase sampling strategy which is well known in standard survey sampling literature.

The methods described in the two previous chapters are applied to a simulated population in Chapter 6. A ten-year monitoring programme is implemented and various monitoring strategies are compared. This raises issues about how the beliefs about the population being surveyed will determine the monitoring strategy and how maps of species density are obtained once data from more than one survey are available.

In Chapter 7 a case study is examined. Data on elephant dung from a forest reserve in Central Africa have been collected using distance sampling. In addition there is information on human activities. One survey has been carried out and these data are used to suggest a potential second survey design. We discuss how the strategies in this thesis can be applied to line transect data.

The estimation strategy described in Chapter 5 is well known, but traditional design-based estimators do not use all the observed data to estimate the covariance between the estimators of the two population totals, required for estimating change through time. Bootstrapping was proposed as an alternative estimation strategy. The combined sampling strategies we propose incorporate unequal probability sampling designs. Current bootstrapping methods for estimating the variance of an estimator obtained from data from unequal probability samples are limited. This motivated the work in Chapter 8 where a new bootstrapping method for unequal probability sampling is developed. This is extended to suggest a potential strategy for estimating the covariance between two population totals using all the observed data.

Chapter 9 concludes this thesis by discussing the proposed strategies for use in an adaptive design-based monitoring programme and describing its limitations. Further work is also suggested to extend the methods in a number of directions.

Chapter 2

Wildlife Population Monitoring

Managers of wildlife populations require information about a species for many different reasons. Although their motivations may be different, the information required is often relatively similar and can be summarised as a set of objectives that cover a wide range of scenarios. For a particular species and environment, a monitoring programme can be designed to meet these objectives. This programme will consist of a series of surveys through time. The data collected and how they are used to meet the objectives of the programme will depend on the chosen survey method. In practice there are many survey methods each with its own assumptions about the population and with issues still to be solved. In this thesis we concentrate on the issues surrounding only one survey method, plot sampling.

In this chapter the setting for the thesis is given. Section 2.1 describes the objectives of a generic single-species monitoring programme. These objectives are related to a number of scenarios that a wildlife manager may face. A notational framework for the basic survey methodology, plot sampling, and the data available before and after the survey are described in section 2.2 and a 'generic' simulated population is introduced. Assuming each survey is a census of the survey region section 2.3 describes how the survey data can be used to meet the objectives of the monitoring programme. This introduces the concept of

a realised population and the superpopulation. We conclude this chapter, in section 2.4, by describing the type of populations and monitoring programmes that are considered in this thesis.

2.1 Monitoring of wildlife populations

A general set of objectives for a single–species monitoring programme over some pre–specified area, the survey region, are

1. Status: From one survey
   (a) To estimate the population or sub–population total
   (b) To describe the distribution of the species over the survey region
2. Trends: From a series of surveys through time:
   (a) To describe long–term trends or changes in the population, or sub–population, totals through time
   (b) To describe the change in the distribution of the species over the survey region through time.

In many cases these objectives will arise naturally from a general management strategy of a particular species, especially when little is known at the start of the management strategy. To illustrate this and how different motivations for monitoring can still lead to the same four generic objectives, the case of Gunther's Gecko (*Phelsuma guntheri*) is described. This was one of three motivating examples for the work in this thesis. Further details can be found in Burn and Underwood (2001).

Gunther's Gecko is a species of reptile found only on Round Island, an uninhabitated island that lies 22.5 km north east of Cap Malheureux on the Mauritius mainland. This island has an area of 151 ha and rises to a height of 280m. In 1957 the island was designated a nature

reserve. At that time it was overrun by rabbits and goats which had been introduced in the nineteenth century leading to severe degradation of the island's vegetation. A programme of eradication of the goats and rabbits was completed in 1986 (Merton *et al.*, 1989). The island now is the only relatively large island in the Mascarenes free of introduced mammals and of major woody weeds. It is the only known breeding ground in the Indian Ocean for the Round Island petrel (*Pterodroma arminjoniana*), and an important breeding station for the red-tailed tropic bird (*Phaethon rubricauda*), white-tailed tropic bird (*Phaethon lepturus*) and the wedge-tailed shearwater (*Puffinus padificus*). It is inhabited by eight or more species of native reptiles endemic to the Mascarene Islands, four or five of which occur only on Round Island (the reason for the uncertainty is that one species, the Burrowing Boa, may be extinct). The main aim of the Round Island management plan (Merton *et al.*, 1989) is to restore the vegetation of the island. Main activities include extending soil cover, planting seedlings of native species and the elimination of undesirable introduced plants.

There are three main reasons for monitoring Gunther's Gecko. In fact these reasons are also true for the other endemic reptile species on the island. However as each species has a separate biology and behaviour it is not practical to develop one monitoring programme for all species, so we focus on only one. Firstly Gunther's Gecko is endemic to Round Island. For a species which occurs nowhere else it is obviously important to assess its status, and estimate the total population (Obj. 1(a)). This assessment must be made repeatedly over time to provide early warning of any impending crisis in the population that may require some intervention. These repeated assessments will provide a series of population totals and so enable trends in the population through time to be observed (Obj. 2(a)). If there are signs of a consistent downward trend then further investigation will probably be required to understand the causes of these changes. In addition these assessments contribute to international classifications such as the IUCN Red List of Threatened Species[1] (*P. guntheri* is currently classified as Endangered) and the CITES[2]

[1]This categorises taxa on their relative risk of extinction. Those at high risk of global extinction are classified as Critically Endangered, Endangered or Vulnerable (Hilton-Taylor, 2000).

[2]for further explanation of the CITES appendix see section 7.1

Appendices (*P. guntheri* is currently on Appendix 2). These classifications are based on previous estimates of population size, which do not have estimates of precision attached to them and have been difficult to repeat. It is important to classify the species correctly so that if they are considered under threat, conservation action can be targeted towards them and more funding may become available to do this.

Secondly, as the restoration of Round Island continues, the habitats provided for the reptiles by the evolving patterns of vegetation are also changing. An important indicator of the success of the restoration is the quality of these habitats for the reptiles. Monitoring changes in the distribution of the species over the island and in relation to specific management practices (Obj. 1(b), 2(b)), provides a measure of the overall impact of the restoration programme. Maps of the species distribution (Obj. 1(b)) can be a helpful aid in decision making. For example, to identify high density areas that can then be avoided when new management strategies, that are not directed at *P. guntheri*, are implemented. Furthermore by studying how the distribution has changed (Obj. 2(b)) lessons can also be learnt about which habitats or management practices the species prefers.

A third reason for monitoring the population is that the idea of translocating some individuals to another island has been mooted. If successful, translocation would provide greater security for the long-term survival of the species. A necessary prerequisite for the management of translocation is that reliable population estimates (Obj. 1(a)) are available. Selection of the individuals to be translocated requires a knowledge of the species distribution (Obj. 1(b)). The gene pool of the translocated population should be as diverse as possible, so *P. guntheri* should be taken from all parts of the island. However in very low density areas the population may be vulnerable to disturbance and these areas can be identified and left alone.

Three types of study have been described here which we call Types I, II and III. A Type I study is a "snapshot" of the population at one particular instant in time. This provides data for objectives Obj. 1(a) and Obj. 1(b). A Type II study describes how the population total changes through time; the simplest scenario is to complete a number of snapshots

(Type I studies) at different times using a repeatable and consistent method. This is the core of a monitoring programme and provides data for all four objectives. A Type III study is a more detailed investigation into the population, usually initiated because the Type II study has indicated a problem. These types of study may be of a more experimental nature to understand how the species responds to particular interventions.

In the case of Round Island a monitoring strategy for Gunther's Gecko would consist of developing a Type I study that allows absolute population estimates of the species to be obtained and recorded, together with relevant information on location, habitat and climatic information using repeatable methodology. Although the biology of *P. guntheri* is well known, from studies of captive bred populations in Jersey Zoo, little is known about its behaviour in the wild. This is partly because it is a crepuscular species, and it spends a lot of its time in the depths of Latania palm trees. So that a suitable survey method for the Type I study can be proposed and tested, a greater understanding of the species ecology is required. These are detailed studies, which although have a different motivation to our standard definition could also be defined as Type III studies.

Once the basic Type I study has been identified and implemented the Type II study can start. This is a a series of Type I studies through time. Type III studies will be employed if the data from the Type II study indicates that there are problems. Ideally the Type II studies should be as informative as possible to indicate the possible direction of a Type III study. If the Type II study indicates that there is a decline in the population total, the maps of changing species distribution (Obj. 1(b), 2(b)) may indicate whether species decline is occurring over the whole of Round Island, or in particular habitats or places where there is habitat change. In the case of Round Island, habitat information must be collected at each survey because of the rapid change in vegetation.

Although the monitoring programme attempts to improve understanding about the relationship between management practices and population density, the monitoring studies that we consider are not designed to specifically compare the effectiveness of different management practices. This requires experimental design type procedures. Similarly we do

not attempt to consider how a programme of adaptive management could be monitored. Under adaptive management several management strategies are implemented at the same time on different parts of the population, usually defined by geography, with the opportunity to always adapt the management further when particular strategies are clearly not working. Further adaptations continue through time as strategies are refined, more is learnt about the species, and management objectives shift. The monitoring programme is the method of feedback for the wildlife manager, and therefore experimental design type procedures need to be employed, so that comparisons between various management strategies can be made. This can become extremely complex when there are rapid changes to management strategies, and it can be very difficult to obtain long-term trends in the population and separate out effects due to the whole combination of management decisions implemented.

The core of most monitoring programmes is the Type II study. In many cases estimates of relative abundance rather than absolute abundance are used to monitor trends in abundance as is the case with the original studies carried out on Round Island. However to assess change in abundance using relative abundance an important assumption is that the relationship between abundance and relative abundance remains constant through time. Against a background of rapid habitat change, this assumption is unlikely and so little about trends in abundance can be deduced.

An important point about the Round Island example, that is also true of many other studies, is that although Type III studies are to be implemented before the monitoring programme is initiated, these studies are mainly to help determine the appropriate type of survey design. There will still be many questions unanswered when the first survey is run - in particular how to choose the most efficient design, for the survey method chosen. The initial survey design will therefore be simple. Through time, when a number of surveys have been completed, information about how the survey could be designed more efficiently becomes available. For example the maps of species density (Obj. 1(b), 2(b)) will indicate whether most sampling effort has occurred in areas of very low density. Ideally this

information could be used to adapt survey design so that more efficient estimates of the population total are obtained, whilst still enabling comparison with early surveys so that estimates of trend can be made.

In any fieldwork, it is important to keep the observers motivated. Observers participate partly because of an interest in the species. There is little more demoralising than spending many hours in the field, under difficult conditions, only to return without having seen any individuals of the study species. Seeing more individuals will add to the general knowledge of the observers about the species. This is especially important in cases, such as Round Island, where the monitoring programme is an integral part of capacity building, the training of local staff and the principle method by which they are able to observe *P. guntheri*. So although not a specific objective of a monitoring programme, it is important to try to design a set of surveys that will enable observers to see many individuals of the study species.

The issues raised here for Round Island are true for many other populations that are to be monitored. Within this thesis we consider the set of objectives outlined above as the key reasons for implementing a single-species monitoring programme. Given these objectives and the study species, a survey method must be selected.

2.2 The surveys

A monitoring programme will consist of a series of surveys taken at regular intervals through time. For the length of the monitoring programme we will assume, in this thesis, that the area over which the surveys are taken is fixed. We call this the *survey region*.

Borchers *et al.* (2002) describe three sources of randomness in a survey. These are due to:

1. population processes such as movement, birth, death, dispersal; represented by a *state* model

2. imperfect and unequal detection of individuals; represented by an *observation* model;

3. the survey design, that is which areas of the survey region are to be sampled; represented by the *sampling process*

The objectives, defined in the previous section, imply that it is the number of individuals within a pre-defined population that is of interest, rather than the number within a survey region. However defining this population can be difficult. For *P. guntheri,* the survey region is Round Island and the population of interest is the entire population of *P. guntheri* on the island. As there is a very clear boundary to the area, the Indian Ocean, which the species cannot cross without human intervention, the population of interest is always contained within the survey region and so the objectives are clearly defined for either the population or the survey region.

In many cases the survey region will not contain all individuals of the population of interest all of the time. Most survey regions do not have clear boundaries caused either by nature — for example rivers or mountains — or by humans — for example urban areas or fences. When populations are motile, individuals may move in and out of a fixed survey region. Particular problems occur when the species is migratory and only passes through the survey region for a period of time. For example, to monitor the population of forest elephant (*Loxodonta africana cyclotis*), as described in Chapter 7, estimates of the number of elephants within various national parks or protected areas are required. Elephants move in and out of these protected areas on a daily basis, and in addition they migrate many hundreds of miles within the course of a year (Blake *et al.*, 2001). An understanding of the population processes, or state model, is important when interpreting an estimate of the population total, or change in the population total through time. Observed changes in the population total may be due to different proportions of the population being in the survey region at the time of the surveys. This is sometimes called *process error*.

To reduce the effect of process error, so that comparisons between surveys can be made, surveys need to be run at the same point in the population process, so that the same

proportion of the population is in the survey region. For example, forest elephants tend to migrate towards sources of trees in fruit so surveys could be taken at a time when the trees are in fruit each year. Whenever the timing of the surveys occurs, concomitant variables, such as the number of fruiting trees, should be recorded and further work, such as a Type III study, may be required to obtain a better understanding of the process error.

In many cases the population is expected to drift, expand or contract its range over the course of a monitoring programme; indeed this may be the very reason for its initiation. In these cases it is wise, if practical, to make the survey region large enough to include the drift or expansion. At the start of the monitoring programme the areas in which the species is known not to occur may then be sampled extensively rather than intensively. The population is then retained within the survey region for a long period of time and process error is reduced.

In each survey the population total is to be estimated; this is also known as abundance estimation. There are many different methods of estimating animal abundance; see Seber (1986, 1992) and Schwarz and Seber (1999) for an overview of many of these methods. Except under unusual conditions a census of the population cannot be undertaken so that randomness due to survey design is inherent in estimating abundance. The different methods of estimating animal abundance vary in how they deal with the sources of variability that are described by the state model and the observation model. The observation model describes the probability of detecting an individual of the species. Methods such as distance sampling (Buckland *et al.*, 2001) or mark-recapture (Otis *et al.*, 1978) are often used to estimate abundance when the probability of detection is less than one. A framework within which different strategies for dealing with imperfect detection (probability of detection less than one) and unequal detection (not all individuals have the same probability of being detected) is given by Borchers *et al.* (2002). Within this thesis the issue of varying detection probabilities is ignored, by assuming an observation model in which the probability of detection is always one; all animals within a sampled area are detected. This does not mean that the methods developed in this thesis cannot be applied to cases

when detection is less than one. In the case study, in Chapter 7, the methods are applied to a problem in which the survey method employed is distance sampling.

In addition to describing how the population changes through time the state model also describes how the species is distributed over the survey region at a particular point in time, for example at the time of a survey. Because of habitat variability and species dispersal, the species will not be evenly spread over the area. Hence sampling some areas of the survey region will give a greater estimate of the population total than others. The interaction between heterogeneity in the species density over the survey region and the survey design is the focus of this thesis.

The survey design problem is to determine which parts of the survey region are to be sampled. Let the survey region be divided into a finite set of units, also known as quadrats or plots. In survey t, $n^{(t)}$ units are selected using some scheme and the number of individuals within each of the units is recorded. As the observation model assumes perfect detection, this is recorded accurately. The size of the units is chosen for practical reasons, and is often the largest size possible that guarantees perfect detectability.

In some cases the finite set of units correspond to naturally occurring physical entities within the survey region, for example a set of lakes. A survey will consist of sampling a number of lakes and making a complete count of the study species within these lakes. With respect to the forest elephant monitoring, each national park or protected area within a region could be considered a unit, and initial sampling is to select the set of national parks to be sampled in any one survey. In this case a hierarchical sampling strategy is required, as a census of each national park cannot be taken.

The form of sampling above, in which the survey region is partitioned into a number of discrete units, is the most commonly used within standard survey sampling work. The set of units is called a list frame. An alternative approach is to consider the survey region as a continuous two-dimensional region in which a number of points will be sampled. The total number of points in the survey region is infinite. This approach is often used for sampling

soil, climatic processes, and geological resources but we do not consider it further here.

Data about the survey region will often be available from a number of different sources. This information can be categorised as:

1. Physical characteristics such as elevation, aspect and slope may be provided by digital elevation maps (??, Ops). These characteristics will remain fixed through time;

2. Habitat types based on some habitat or suitability classification. For example river habitat surveys, National Land Cover data, or the proportion of various vegetation types. These classifications may be obtained from aerial photography interpretation (Land Cover of Scotland), ground surveys, for example the National Habitat Survey of Grampian Region as used by Buckland and Elston (1993), satellite imagery (U.S. National Land Cover Data). The vegetation greenness, measured by the Normalised Difference Vegetation Index (NDVI) from remote sensing is also of use (Khaemba and Stein, 2000; ??, Ops; Buckland and Elston, 1993). In general much of this information will change very slowly through time, unless there is some form of human intervention, such as forestry logging concessions, or fire;

3. Human influences such as the distance to transport networks, roads and rivers, or human habitation (Khaemba and Stein, 2000). In Chapter 7, variables such as the degree of protection are used, for example the frequency of patrols or the distance to a park boundary. Alternatively management practices such as grazing, or removal of alien species may be recorded. For this information there will be variability both through the survey region and through time;

4. Climatic variables such as rainfall or temperature. In many cases this information is likely to be a set of summary statistics that corresponds to the survey region as a whole rather than changing over the survey region unless the area is very large. For example if surveys are annual, then useful summary statistics may be the annual rainfall, the start date of the rains etc. This information will mainly assist in the

understanding of changes in the population total through time.

Hence for a survey at time t, for $t = 1, \ldots, T$, the survey region A is divided into N quadrats, or units, indexed $i = 1, \ldots, N$. We call the set of these N units U. The location of each quadrat is defined by its midpoint. This location may be defined in terms of latitude and longitude, or using some other positional system. Let $\underline{l}_i = (l_{i1}, \ldots, l_{iL})'$ represent the vector denoting the location of unit i. Typically $L = 2$, although $L = 1$ if a linear system such as a river is being sampled or $L = 3$ if depth is also of importance, for example both the position in the ocean and the depth at that point is required. The units remain the same from one survey to another, hence unit i has the same location $\underline{l}_i$ for all surveys $t = 1, \ldots, T$. Let $y_i^{(t)}$ be the number of observed individuals in unit i at the time of survey t. Additional information about the survey region is also summarised at quadrat level. Let $x_{ij}^{(t)}$ represent the j^{th} auxiliary variable for the i^{th} unit in A at time t for $j = 1, \ldots, Q$. Hence at time t the information available about unit i before the survey is carried out is $\underline{x}_i^{(t)} = (x_{i1}^{(t)}, \ldots, x_{iQ}^{(t)}, l_{i1}, \ldots, l_{iL})'$. The issue is how can these data, the $\underline{x}_i$ for the survey region and the data $y_i^{(t)}$ for a sample $s^{(t)}$ of $n^{(t)}$ units, be used to meet the objectives of the monitoring programme.

As an example we describe a simulated population P that is used in chapter 6 to test out the monitoring strategies developed in this thesis, and as a basis for discussion in earlier chapters. This set of data contains some generic features of the type of data we consider typical of the surveys we have described. The methods used to generate these data are described in Appendix B.

The survey region is assumed to be rectangular and is partitioned into $N = 1296$ units (a grid of 36 x 36 units). There are data for four auxiliary variables, $\underline{x}_1, \ldots, \underline{x}_4$, available at the start of the monitoring programme and these remain fixed through time. These are shown in figure 2.1. Of these four auxiliary variables, three are continuous variables and one is a categorical variable, (habitat 4). Of the continuous variables, one changes relatively slowly over the survey region (variable 1) whereas others change rapidly (variables 2 and 3). Figure 2.2 shows the counts $y_i^{(1)}$ for survey 1 and figure 2.3 the counts $y_i^{(2)}, \ldots, y_i^{(10)}$

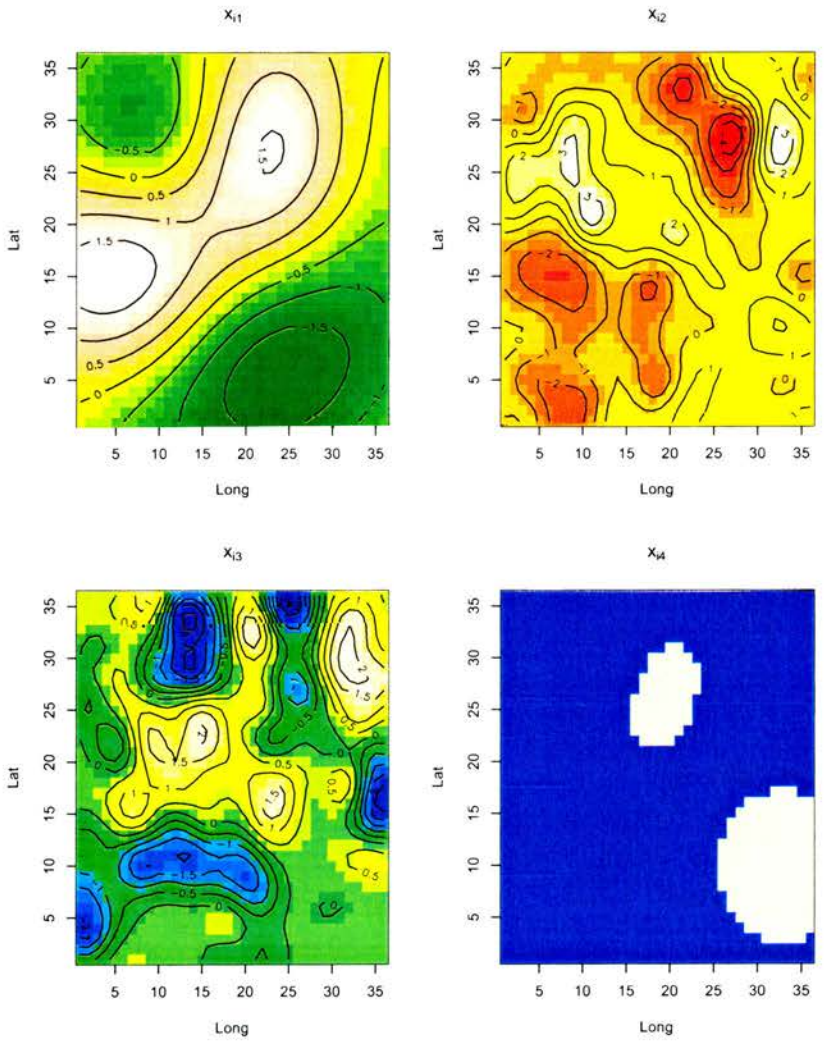


Figure 2.1: The set of auxiliary variables for population P.

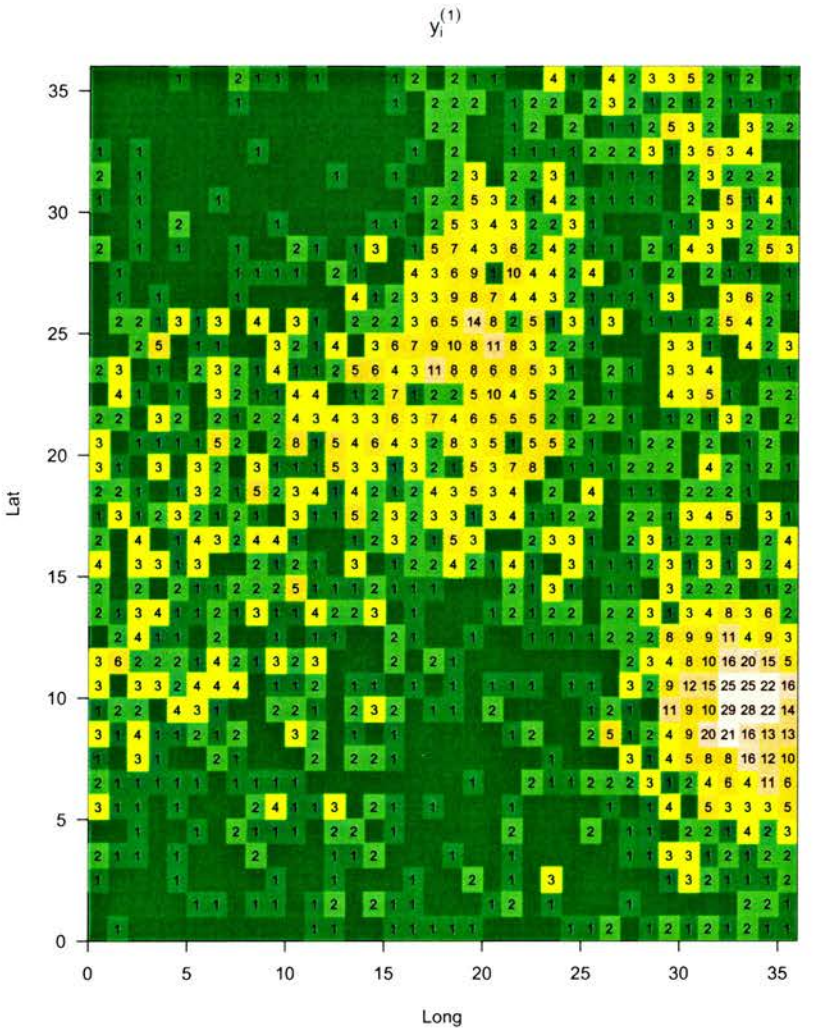


Figure 2.2: A spatial realisation of population P. Number of individuals in each unit in survey 1, $y_i^{(1)}$.

We call the data for figure 2.2 a spatial realisation of the population and the data from figure 2.3 a spatio-temporal realisation of the population.

As statisticians, it is often convenient to think of the $y_i^{(t)}$ as being generated by some unknown random process and to try to represent this process by some form of model; this is the state model of Borchers *et al.* (2002). Regression-type models such as generalised linear models (GLMs; McCullagh and Nelder (1989)) or generalised additive models (GAMs; Hastie and Tibshirani (1990)) are often used to model count data of wildlife populations. These models do not describe the processes that generate the data, for example birth-death processes of individuals, or dispersal patterns. Rather they provide an empirical description of pattern and do not allow inferences about the underlying process. In its very general form an assumed model ζ is

$$E_\zeta[Y_i^{(t)}] = \mu_i^{(t)} \qquad var_\zeta[Y_i^{(t)}] = \sigma_i^{(t)2} \qquad cov_\zeta[Y_i^{(t)}, Y_j^{(t')}] = \sigma_{i,j}^{(t,t')} \tag{2.1}$$

This type of model is often called the (quadrat) superpopulation model (QSM). For the realised population shown in figure 2.2 the correlation in the $y_i^{(t)}$ can be modelled by an autocovariance function so that $\sigma_{i,j}^{(t,t)} > 0$ or by the $\mu_i^{(t)}$, the trend, as discussed in more detail in section 2.3.1. As an operational definition, which we expand upon later, we define a population where we model $cov(Y_i^{(t)}, Y_i^{(t+1)})$ to be large as sessile and a population where $cov(Y_i^{(t)}, Y_i^{(t+1)})$ is close to zero as motile[3].

2.3 Data requirements and analysis methods under a census

Given the survey method described in the previous section, a sample, $s^{(t)} = (i_1, \ldots, i_{n^{(t)}})$, of $n^{(t)}$ units will be visited in survey t. The counts from these $n^{(t)}$ units and the auxiliary data, possibly from all N units, are used to estimate the parameters of interest. If $n^{(t)} = N$ for all t then a census of the area has been carried out so $s^{(t)} = U$ and the entire spatio-temporal realisation of quadrat data is available to calculate these parameters. In

[3]This is different to the biological definition of motile and sessile populations as, using our definition, a species with small home ranges will be defined as sessile.

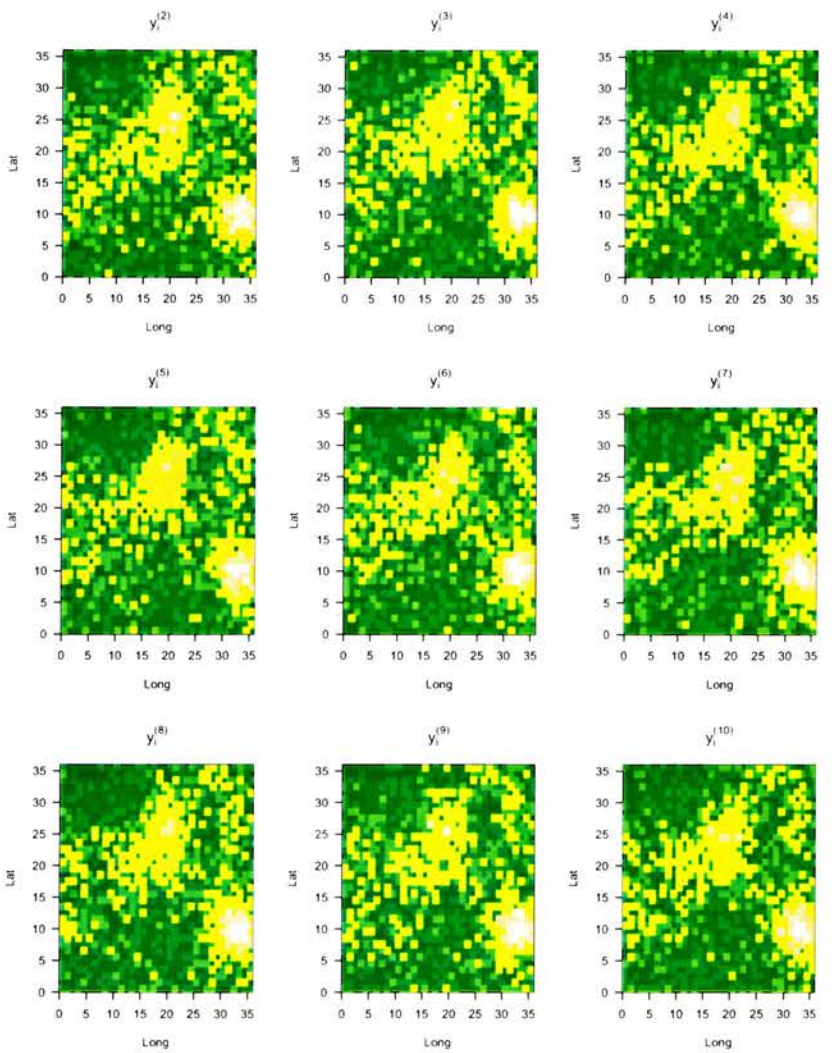


Figure 2.3: A spatio-temporal realisation for population P. The colour coding is as figure 2.2

practice $n^{(t)} < N$ and so only a portion of the spatio-temporal realisation is available for estimating these parameters. The focus of this thesis is how only a proportion of the spatio-temporal realisation can be used for parameter estimation. Current methods are described in Chapter 3. In this section, the parameters, and how they can be calculated from data obtained from a series of census surveys, are described for both a motile and a sessile population.

The following notation, which is relatively standard for survey sampling (Särndal *et al.*, 1992), is used throughout this thesis. Let $\sum_{s^{(t)}} y_i^{(t)} = \sum_{i \in s^{(t)}} y_i$ as the sum of $y_i^{(t)}$ for all units in the set $s^{(t)}$ so that if $s^{(t)} = U$ then $\sum_{U} y_i^{(t)} = \sum_{i=1}^{N} y_i^{(t)}$

2.3.1 One survey

Obj. 1(a): To estimate the population, or sub–population total
Obj. 1(b): To describe the distribution of the species over the survey region

The wildlife manager is interested in the number of individuals in the population. This is $\tau^{(t)}$ and is calculated as

$$\tau^{(t)} = \sum_{U} y_i^{(t)} \tag{2.2}$$

A sub-population may be defined by geographic area or habitat type, or alternatively by characteristics of the individuals itself, such as sex or age. Given the data collected, only sub-population totals based on characteristics of the N units, rather than characteristics of the individuals, can be calculated. For example, if $x_{ij}^{(t)} = 1$ if unit i is forest, and $x_{ij}^{(t)} = 0$ otherwise, the total number of individuals found within a forest is

$$\tau_{x_1}^{(t)} = \sum_{U} y_i^{(t)} x_{ij}^{(t)} = \sum_{s_{x.j}^{(t)}} y_i^{(t)} \text{where } s_{x.j}^{(t)} = \left\{ k : x_{kj}^{(t)} = 1 \right\} \tag{2.3}$$

The smallest sub-population contains just one unit. A map of the area showing the location of each unit and the number of individuals within each unit, for example figure

2.2, provides a description of the species distribution over the survey region. Summing over several of these units provides various sub-population totals as shown in equation 2.3.

In addition to sub-population totals, for various habitat categories, the wildlife manager may also want summary statistics about $y_i^{(t)}$ in relation to continuous auxiliary variables. For example the rate of change in $y_i^{(t)}$ with altitude x_{i1} may be of interest so that we propose a simple model

$$y_i^{(t)} = B_0^{(t)} + B_1^{(t)} x_{i1}^{(t)} + \epsilon_i^{(t)} \tag{2.4}$$

$$\Rightarrow \underline{y}_U^{(t)} = \mathbf{x}_U^{(t)} \underline{B}^{(t)} \text{where } \underline{B}^{(t)} = (B_0^{(t)}, B_1^{(t)})' \text{ and } \mathbf{x}_U^{(t)'} = \begin{pmatrix} 1 & x_{i1}^{(t)} \\ \vdots & \vdots \\ 1 & x_{iN}^{(t)} \end{pmatrix}$$

$\underline{B}^{(t)}$ can be calculated using ordinary least squares so that

$$\underline{B}^{(t)} = (\mathbf{x}_U^{(t)'} \mathbf{x}_U^{(t)})^{-1} \mathbf{x}_U^{(t)'} \underline{y}_U^{(t)} \tag{2.5}$$

$B_1^{(t)}$ describes the rate of change in $y_i^{(t)}$ with altitude for the realised population. As the survey is a census, the population parameter $B_1^{(t)}$ is calculated exactly, in the same way that the sub-population totals describe one particular aspect of the scatter in the $\underline{y}_U^{(t)}$. Not every $y_i^{(t)}$ will fall on the line $B_0^{(t)} + B_1^{(t)} x_{i1}^{(t)}$ but the general pattern observed is described. More complex models for the realised population can also be proposed and their parameters estimated.

For some populations, these descriptions of the spatial realisation of the population are of interest. This is not always the case. For example in some species, individuals are very motile, so that they can cover a large part of the survey region in a small period of time, compared to the interval between surveys. Then the spatial realisation at time $t + \epsilon$, where ϵ is a small number much less than 1, the interval between surveys, will be very different to the spatial realisation at time t even if $\tau^{(t)} = \tau^{(t+\epsilon)}$. Then a map of the realised distribution will become rapidly out-of-date and unuseful. Instead the average of several realisations for different ϵ will give an indication of high and low density areas. This is

the $\mu_i^{(t)}$ from a QSM where the QSM is assumed to hold for the random variables $Y_i^{(t)}$ and $Y_i^{(t+\epsilon)}$. As stated above the QSM is an assumed model that provides an empirical description of the pattern in the $y_i^{(t)}$. Using a GLM or a GAM the auxiliary data can be used to estimate the $\mu_i^{(t)}$. As the data are count data, a log-link function is typically used so that

$$E[Y_i^{(t)}] = \mu_i^{(t)} = exp\left(\sum_{j=0}^{Q} f_j^{(t)}(x_{ij}^{(t)})\right) \tag{2.6}$$

where the $f_j^{(t)}(x_{ij}^{(t)})$ may be a linear or smooth function of the j^{th} auxiliary variable. In its simplest form, $f_j^{(t)}(x_{ij}^{(t)}) = \beta_j x_{ij}^{(t)}$ and the β_j are estimated, unlike the description of the realised population in which parameters are known under a census. In this simplest form a GLM is most appropriate but for more complex functions, such as smoothing splines (Silverman, 1985) a GAM would allow greater flexibility in the form of the functional relationship between the auxiliary variables and $Y_i^{(t)}$. If the model assumes that the $Y_i^{(t)}$ follow a Poisson distribution, then $var[Y_i^{(t)}] = \mu_i^{(t)}$, $cov(Y_i^{(t)}, Y_j^{(t)}) = 0$ and likelihood methods can be used to fit the model. In practice the Poisson assumption often does not hold and $var[Y_i^{(t)}] > \mu_i^{(t)}$; this is called overdispersion. Depending on the beliefs of the modeller/wildlife manager about the species being monitored, the overdispersion can be dealt with using one of two strategies.

In the first strategy, it is assumed that overdispersion is caused by $\mu_i^{(t)}$ being inadequately modelled, because data for the appropriate auxiliary variables are not available. In this case a thin-plate spline of latitude and longitude is added to the model. This becomes a proxy for the unobserved auxiliary variables, that are assumed to change continuously over the survey region.

Alternatively the $\mu_i^{(t)}$ can be assumed to be adequately modelled, but the assumption about the variance is wrong. A common strategy then is to use quasi-likelihood methods (McCullagh and Nelder, 1989; Wedderburn, 1974). Rather than an explicit form for the distribution of the observations, these methods require the relationship between the variance of the observations and their mean to be specified. In a quasi-Poisson model

$var[Y_i^{(t)}] = b\mu_i^{(t)}$ and b is estimated from the data. This approach is often used when individuals are thought to associate with each other, for example offspring are found close to their parents, so that even in a relatively homogeneous environment, individuals tend to cluster.

In addition to overdispersion, a further problem may be that the assumption that the autocovariance is zero, $cov(Y_i^{(t)}, Y_j^{(t)}) = 0$, does not seem to hold once the model has been fitted. Again if the modeller or wildlife manager believes that the $\mu_i^{(t)}$ are inadequate, fitting a thin-plate regression spline of latitude and longitude may alleviate this problem. If however the $\mu_i^{(t)}$ are thought to be adequately modelled, then the alternative is to fit more complex models in which the autocovariance is non-zero. This assumption might be considered appropriate if individuals are thought to occur in clusters and the clusters are large compared to the size of the quadrats. A difficulty with this approach is that we cannot model positive autocovariance under the Poisson assumption (Besag, 1974).

In many cases a combination of these different processes may seem appropriate. When data from only one survey are available it is difficult to separate out spatial trend, change in $\mu_i^{(t)}$ over the survey region, and the effect of autocovariance. When data from several surveys are available then it becomes more possible to separate out these effects if it is assumed that spatial trend remains constant over time.

As well as beliefs about the population, the use of the $\hat{\mu}_i^{(t)}$ also determines how the QSM is specified. This is especially true when the survey is not a census of U but only a sample of U as discussed in section 3.5. In general even under a census survey, if we believe that a population is highly motile, then it is the map of the $\hat{\mu}_i^{(t)}$ rather than the $y_i^{(t)}$ that is of use to the wildlife manager as the $y_i^{(t)}$ will quickly become out-of date. As stated before, this assumes that the QSM, $\zeta^{(t)}$, holds for the random variables $Y_i^{(t)}$ and $Y_i^{(t+\epsilon)}$ where $\epsilon < 1$. In a similar manner, if sub-population totals of a motile population are to be calculated, it is debatable whether the sum of the relevant $y_i^{(t)}$ will provide the manager with useful information. To a certain extent it depends on whether the individuals are confined to a particular area defined as the sub-population, in which case the $y_i^{(t)}$ are adequate, or

whether the individuals can move freely in and out of the sub-population region, in which case the sum of the $\hat{\mu}_i^{(t)}$ may be of more use, as this represents the expected number of individuals in the sub-region. Although we could similarly argue this point for the total number of individuals in the total population so that $\sum_U \hat{\mu}_i^{(t)}$ rather than $\sum_U y_i^{(t)}$ is used, we will assume that $\tau^{(t)}$ is the focus of the survey.

2.3.2 Several surveys

Through time
Obj. 2(a): To describe long–term trends or changes in the population, or sub–population, totals
Obj. 2(b): To describe the change in the distribution of the species over the survey region

Once data from several surveys are available, estimates of change in the population or sub-population total will be of interest. As with the population total, of direct interest is the change in the realised population through time. With data from two surveys, at times t and t', an estimate of change in the population total $\delta^{(t',t)}$ is calculated as

$$\delta^{(t',t)} = \tau^{(t)} - \tau^{(t')} = \sum_U (y_i^{(t)} - y_i^{(t')}) \tag{2.7}$$

In a similar manner, a change in sub-population totals can be calculated as

$$\begin{aligned} \delta_{x_1}^{(t',t)} &= \tau_{x_1}^{(t)} - \tau_{x_1}^{(t')} \\ &= \sum_U y_i^{(t)} x_{1i}^{(t)} - \sum_U y_i^{(t')} x_{1i}^{(t')} \\ &= \sum_{s_{x_1}^{(t)}} y_i^{(t)} - \sum_{s_{x_1}^{(t')}} y_i^{(t')} \end{aligned} \tag{2.8}$$

Unless the covariate $x_{i1}^{(t)} = x_{i1}^{(t')}$ for all $i \in U$, equation 2.8 cannot be reduced in a similar way to the final form of equation 2.7 as the units in $s_{x_1}^{(t)}$ may not be the same as those in $s_{x_1}^{(t')}$. A map of the change in species distribution however is the map of the individual unit differences $y_i^{(t)} - y_i^{(t')}$.

If data from several, say T, surveys are available, a description of the trend in the population through time can be obtained. In its simplest form, this is a summary of the linear component of the trend in the population totals so that if

$$\tau^{(t)} = \eta_o + \eta_1 t$$

the linear component, η_1, can be estimated using least squares to be

$$\eta_1 = \frac{24}{T(T+1)(T-1)} \sum_{t=1}^{T} (2t - T - 1)\tau^{(t)} \tag{2.9}$$

Parameters other than the linear trend component or a description of the trend may also be required.

Even if $\delta^{(t',t)} = 0$ the distribution of the species over the survey region, the $y_i^{(t)}$, may change through time. A map of the realised population in each survey shows the change, for example figure 2.3. To summarise all changes on one map, a simple summary statistic, for example η_{i1}, can be calculated for each unit $i \in U$ where

$$y_i^{(t)} = \eta_{i0} + \eta_{i1} t$$

and η_{i1} is estimated using equation 2.9.

For a motile population, an understanding of the QSM through time may be of interest. In its simplest form, it might be assumed that $\mu_i^{(t)} = r^t \mu_i^{(0)}$, so that auxiliary variables and the relationship between the auxiliary variables and $\mu_i^{(t)}$ are assumed constant through time and the key parameter of interest is r. Note that in many monitoring programmes it is r that is of interest rather that η_1. Assuming a model of this form, with a constant relationship between auxiliary variables and $\mu_i^{(t)}$ through time, makes it easier to determine whether $cor(y_i^{(t)}, y_j^{(t)})$ can be modelled by the mean $\mu_i^{(t)}$ or by an autocovariance function $cov(Y_i^{(t)}, Y_j^{(t)})$, as for the former, the location of high density areas should remain relatively constant through time, whereas for the latter it will change. Similarly the modeller might wish that the autocovariance function $cov(Y_i^{(t)}, Y_i^{(t')})$ is high for species in which individuals are sessile, long-lived and offspring are dispersed close to their parents,

compared to species in which individuals can cover a large part of the survey region in a small period of time, or are short-lived and offspring are not necessarily found close to their parents. This corresponds to our working definition of motile and sessile species.

Alternatively, the modeller may believe that there is drift in the population through time, so that $\mu_i^{(t)}$ is not changing at the same rate over the whole survey region. This might be represented by the $x_{ij}^{(t)}$ or the $f^{(t)}$ changing through time. In these cases maps that show how $\mu_i^{(t)}$ changes through time will be of interest.

2.4 Summary

A monitoring programme consists of multiple objectives which means that different parameters need to be calculated from the survey data. These parameters will also vary depending on whether the population is assumed to be motile or sessile, also defined in this chapter. For some objectives, parameters summarise the realised population, notably the total number of individuals in the population at time t, $\tau^{(t)}$, and the change in $\tau^{(t)}$ from survey t' to survey t, $\delta^{(t',t)}$. For other objectives, those relating to the spatial distribution of the species, the parameters of interest for a motile species are based on the assumed superpopulation model, whereas for a sessile species, parameters are based on the realised population.

In this chapter we described the data required to meet all four objectives of a monitoring programme for motile and sessile species. In this thesis our focus is on the design of a monitoring programme for a single motile species. We will also assume that individuals do not naturally associate with each other. Hence we can assume a very simple QSM in which $cov(Y_i^{(t)}, Y_j^{(t')}) = 0$ and if there is evidence of overdispersion this is due to the lack of appropriate auxiliary variables which can be dealt with by adding a thin-plate spline of latitude and longitude to the model. In addition the main parameters of interest will be the population total, $\tau^{(t)}$, and the change in the population total through time, $\delta^{(t',t)}$.

We assume that the monitoring programme is a Type II study, consisting of a series of plot sampling surveys. Although plot sampling is generally not the recommended strategy for motile populations, because it can be difficult to achieve perfect detectability, the survey design issues explored in this thesis still apply when detection is imperfect. However except for the case study in Chapter 7 we assume perfect detectability so that this source of randomness is removed.

In a monitoring programme the resources to visit all units in the survey region in each survey will be vast and hence only a sample of units can be visited. The key survey design issue that we address in this thesis is how the choice of sample units, determined by the sampling process, can cope with heterogeneity in species density over the survey region, represented by the state model, so as to provide precise estimators of the various parameters of interest in the monitoring programme.

Chapter 3

Current Sampling Methods

To implement a monitoring programme using plot sampling, the statistician needs to determine the sampling scheme by which $n^{(t)}$ units are selected from U and the appropriate estimator for each parameter of interest. If the sampling fraction $f = \frac{n^{(t)}}{N} = 1$ then there is no sample selection problem and the estimators are as defined in the previous chapter. Generally the sampling fraction is small (less than 0.25) and the choice of sampling scheme, the sampling fraction and the estimator all influence the precision of the estimated parameter.

In a monitoring programme each survey consists of a design stage, a data collection stage and an estimation stage. The design stage of a survey determines the sampling scheme that is used to select a sample and uses this scheme to draw a sample $s^{(t)}$ of size $n^{(t)}$. In the data collection stage the units are visited and the $y_i^{(t)}$ observed and recorded. Generally the data collection stage occurs after the design stage once all $n^{(t)}$ units have been selected, although for adaptive sampling strategies (Thompson and Seber, 1996) and some two-phase sampling strategies this is not the case. The estimation stage refers to the period of obtaining estimates from the collected data. The appropriate estimation method depends on the design stage of the survey, in particular the scheme used to select the sample. A *sampling strategy* is defined as the combination of the choice of sample design and the

method of estimation for a particular parameter. We will define a *monitoring strategy* as the set of sampling strategies through time that are required to provide estimators for all parameters required by the monitoring programme.

The method of selecting the sample can be probabilistic or purposive. In section 3.1 the basic principles of probabilistic sampling schemes are introduced. If a probabilistic sampling scheme is selected then estimators and their precision can be estimated using a design-based or model-based strategy. Section 3.2 compares design-based and model-based estimation strategies and explains why for the two parameters, $\tau^{(t)}$ and $\delta^{(t',t)}$, a design-based strategy is employed in this thesis. For a general probabilistic sampling strategy the estimators for $\tau^{(t)}$ and its variance are described in section 3.3 and sampling strategies and estimators that improve the precision of this estimator using auxiliary variables are described. The estimate of $\delta^{(t',t)}$ can be calculated merely by taking the estimates of $\tau^{(t)}$ and $\tau^{(t')}$ and finding the difference. In some circumstances more precise estimates can be obtained by retaining some units from one survey to another. Section 3.4 describes these circumstances and some commonly used sampling strategies. These sampling strategies are generally fixed at the start of the monitoring programme. However when little is known about the species at the start of the programme, these designs may be inefficient. Adaptive sampling strategies allow the sampling strategy to change as more is learnt about the population. Section 3.6 describes the adaptive sampling strategies that are currently available to the practitioner. These strategies are few and within a design-based framework focus mainly on adaptation within one survey, rather than between surveys. Although the focus of the monitoring programme is to obtain estimates of totals and changes in total through time, which we do using design-based estimators, parameters describing the species distribution are also of interest. We briefly describe in section 3.5 how these can be estimated; in particular for a motile population, which requires a model-based approach. Section 3.7 describes the shortcomings of the current methods for designing a monitoring programme, and sets out in detail the problem that this thesis investigates.

3.1 Survey sampling; the design stage

Given a finite population of units $U = (1, \ldots, i, \ldots, N)$, a sample $s \subseteq U$ of n units must be selected[1]. A sampling scheme defines how the sample is selected for a particular survey.

Units can be selected using a purposive scheme, which may for example be based on accessibility. For example Walsh and White (1999) suggests monitoring forest elephants by using recce ("Reconnaissance") samples. Transects follow the path of 'least resistance', usually trails and paths, within the forest from some, possibly randomly located, point. An alternative method of 'purposive' sampling is one in which a set of units that are 'representative' of the survey region are selected. This representativeness is based on the opinion of an individual selecting the samples and can lead to conscious or unconscious bias which cannot be measured

Alternatively a probabilistic sample selection scheme may be employed. This form of sampling scheme is commonly used within survey sampling and the general principle is that any particular set of units, s, has a specified probability of being the selected sample. A more formal definition is that a *probability sample*, is a sample s that is realised under the four conditions described below as given by Särndal *et al.* (1992, pp 8):

1. The finite set of all possible samples $\mathcal{S} = \{s_1, \ldots, s_M\}$ that can be obtained using the sample selection scheme can be defined;
2. Every possible sample $s_k \in \mathcal{S}$ for $k = 1, \ldots, M$ has a known probability of selection $p(s_k)$;
3. Every unit in the population has a non-zero probability of being selected;
4. One sample is selected using a random mechanism under which each possible s receives exactly the probability $p(s)$.

[1] In this section we ignore the superscript t that denotes the period in which the survey was taken, as the results hold true for all surveys $t = 1, \ldots, T$

The function $p(.)$ is defined as the *sampling design* so that $p(s)$ is the probability of selecting sample s. Different sample selection schemes can lead to the same sampling design $p(.)$. For a given sampling design $p(.)$, the sample s is the outcome of a random variable S whose probability distribution is defined by the sampling design $p(.)$:

$$Pr(S = s) = p(s) \text{ for any } s \in \mathcal{S}$$

Because $p(s)$ is a probability distribution

$$p(s) \geq 0, \forall s \in \mathcal{S} \text{ and } \sum_{s \in \mathcal{S}} p(s) = 1$$

Given a particular sampling design $p(.)$, the probability that an individual unit is included in the sample, π_i, can be defined. Let the random variable I_i denote the inclusion of a particular unit i in a sample where

$$I_i = \begin{cases} 1 & \text{if } i \in s \\ 0 & \text{otherwise} \end{cases}$$

so that $E[I_i] = 0Pr(I_i = 0) + 1Pr(I_i = 1) = \pi_i$

where in conventional survey sampling notation $i \in s$ denotes that the unit i is in the sample s and, for future reference, $s \ni i$ denotes the samples that contain the unit i. The probability that unit i will be included in a sample, π_i, is obtained from the design by

$$\pi_i = Pr(i \in s) = Pr(I_i = 1) = \sum_{s \ni i} p(s)$$

In a similar manner the joint inclusion probability

$$\pi_{ij} = Pr(i \& j \in S) = Pr(I_i I_j = 1) = \sum_{s \ni i \& j} p(s)$$

and we note that $\pi_{ij} = \pi_{ji}$ and $\pi_{ii} = Pr(I_i^2 = 1) = Pr(I_i = 1) = \pi_i$

There are two different types of sampling scheme. In a with-replacement scheme the sample s may contain the same unit more than once. The effective sample size ν is the number of unique units in the sample and in a with-replacement scheme $\nu \leq n$. The

inclusion probabilities π_i represent the probability that unit i is included at least once in s. In a without-replacement sampling scheme $\nu = n$ as any unit can only occur once in s.

Under the sample selection strategy of simple random sampling without replacement, which we denote *srswor*, the sample design $p(.)$ is such that each possible sample has the same probability of being selected. Given the size of the sample n

$$p(s) = \frac{1}{\binom{N}{n}}$$

and so the inclusion probabilities are of the form

$$\pi_i = \sum_{s \ni i} p(s) = \binom{N-1}{n-1} / \binom{N}{n} = \frac{n}{N}$$

and

$$\pi_{ij} = \sum_{s \ni i \& j} p(s) = \binom{N-2}{n-2} / \binom{N}{n} = \frac{n(n-1)}{N(N-1)}$$

3.2 Design-based and Model-based Estimators

Suppose the population total $\tau = \sum_U y_i$ is to be estimated. As only a sample has been taken, only an estimate of the parameter, $\hat{\tau}$, can be obtained. For a particular estimator of τ, different 'samples' will give different estimates of a parameter; so $\hat{\tau}$ is a random variable. The variation in $\hat{\tau}$ can be attributed to one of two sources which distinguish the two approaches to estimation. Working in a design–based framework, the variation is considered to come only from the sampling design $p(.)$. That is, for the same spatial realisation, different estimates are obtained by taking a different sample of units s. In a model–based framework, randomness arises because the $\underline{y}$ are generated from a random variable $\underline{Y}$. For the same sample of units s, a different estimate of τ would be obtained under two different spatial realisations, as the $\underline{y}$ would be different.

Desirable properties of an estimator are that it is unbiased and has low variance. These properties can be summarised using the mean–square error which is the sum of the variance

and the squared bias of an estimator. An estimator, $\hat{\tau}$, of the population total, τ, is design-unbiased if the expectation over all possible sample sets generated by the design $p(.)$ is the population total, τ, for that particular set of data $\underline{y}_U$

$$E_p[\hat{\tau}] = E[\hat{\tau}|\underline{y}] = \tau \tag{3.1}$$

and so the design-based mean-square error is

$$MSE_p(\hat{\tau}) = E[(\hat{\tau} - \tau)^2|\underline{y}]. \tag{3.2}$$

For example if a sample is selected using *srswor*, a common estimator of the population total is

$$\hat{\tau}_p = \frac{N}{n}\sum_s y_i = N\bar{y}_s. \tag{3.3}$$

Using the inclusion probabilities defined in the previous section, the expectation over all possible samples is

$$\begin{aligned} E_p[\hat{\tau}] &= \frac{N}{n}E_p\left[\sum_U I_i y_i\right] \\ &= \frac{N}{n}\sum_U y_i E[I_i] \\ &= \frac{N}{n}\sum_U y_i \pi_i \\ \Rightarrow E_p[\hat{\tau}] &= \sum_U y_i = \tau \text{ so it is design–unbiased} \end{aligned}$$

In a model-based framework, the estimator $\hat{\tau}$ is a model-unbiased estimate of τ if the expectation over all realisations of an assumed model ξ is the expected total for that model so that for a specific model,

$$E_\xi[\hat{\tau}] = E_\xi[\tau|s] = \sum_U \mu_i \tag{3.4}$$

and the model–based mean–square error is

$$MSE_\xi[\hat{\tau}] = E[(\hat{\tau} - \tau)^2|s] \tag{3.5}$$

The model can be the QSM, ζ, described in section 2.2. To illustrate this approach, assume a simple QSM model ζ of the form

$$E_\zeta[Y_i] = \alpha \qquad var_\zeta(Y_i) = \sigma^2 \tag{3.6}$$

As it is the total of the realised population that is of interest, we use the prediction approach first described by Royall (1970) so that the estimator is of the form

$$\hat{\tau} = \sum_s y_i + \sum_{s_c} \hat{y}_i \tag{3.7}$$

where s_c are the units that are not sampled and the $\hat{y}_i$ are predicted from ζ.

In this case $\hat{y}_i = \hat{\alpha}$ and a best linear unbiased predictor of α is $\hat{\alpha} = \bar{y}_s$. The estimated population total $\hat{\tau}$ is therefore

$$\hat{\tau}_\zeta = n\bar{y}_s + (N-n)\bar{y}_s = N\bar{y}$$

the same as that under the design-based framework. Taking expectations over realisations gives

$$\begin{aligned} E_\zeta[\hat{\tau}|\underline{s}] &= NE_\zeta[\bar{y}_s] \\ &= \frac{N}{n}\sum_s E[Y_i] = \frac{N}{n}\sum_s \alpha \\ \Rightarrow E_\zeta[\hat{\tau}|\underline{s}] &= N\alpha = \sum_U E[Y_i] = E_\zeta[\tau] \text{ so it is model-unbiased} \end{aligned}$$

The advantages and disadvantages of a model-based vs a design-based framework have been discussed at great length in the literature, see for example Brus and de Gruijter (1997); Cassel *et al.* (1977); Hansen *et al.* (1983); Särndal (1978); Smith (1976); Thompson (2002); Thompson and Seber (1996). The most compelling argument for using a design-based rather than a model-based framework for estimation of parameters of the realised population is the potential for obtaining unbiased estimates without needing to make

any assumptions about the population. Given the sample design there is, in a design-based framework, little question about the appropriate estimators to be used. Another advantage is that the strategy requires a probabilistic sampling selection scheme so that human selection biases do not play a part in choosing the sample. The main advantages of a model-based framework are that auxiliary information can be easily included in the estimators, and that for an assumed state model, the efficiency of an estimator can be calculated for different sets of sample units so that an optimal sample can be specified; an optimal sample is one that maximises the precision of the estimator under an assumed state model. However model-based estimates will depend on the assumptions about the population. If poor assumptions are made, estimators can be biased and imprecise. For many wildlife populations, there is not enough information about a species to provide a consensus on the appropriate population model and interested groups out to support one view or another could with equally justifiable reasoning propose very different models leading to very different estimates. In particular they may lead to very different estimates of precision. Hence the design-based framework in which these population assumptions do not matter is appealing.

In this thesis we will therefore concentrate on estimating the parameters $\tau^{(t)}$, $\delta^{(t',t)}$ within a design-based framework. For maps of the spatial distribution of the species over the survey region we need to use a model-based approach, as for motile populations it is the $\mu_i^{(t)}$ rather than the $y_i^{(t)}$ of the realised population which are required.

3.3 Estimating τ

For any sampling design $p(.)$, in which π_i is the probability that unit i is included in the sample s, an unbiased estimator of τ, is the Horvitz-Thompson estimator[2](Horvitz and

[2]Most of the estimators in this section are found in many textbooks. However in this section some are derived for completeness, so that work in future chapters has a solid base.

Thompson, 1952)

$$\hat{\tau} = \sum_{i=1}^{\nu} \frac{y_i}{\pi_i} \tag{3.8}$$

where s_ν is the set of ν unique units in s. It is simple to show that this estimate is design-unbiased as

$$E[\hat{\tau}] = E\left[\sum_{s_\nu} \frac{y_i}{\pi_i}\right] = E\left[\sum_{U} \frac{y_i}{\pi_i} I_i\right] = \sum_{U} \frac{y_i}{\pi_i} E[I_i] = \sum_{U} \frac{y_i}{\pi_i} \pi_i = \sum_{U} y_i = \tau$$

The variance of this estimator is

$$\begin{aligned}
var(\hat{\tau}) =& E[(\hat{\tau} - E[\hat{\tau}])^2] = E[\hat{\tau}^2] - E[\hat{\tau}]^2 \\
=& E\left[\left(\sum_{s_\nu} \frac{y_i}{\pi_i}\right)^2\right] - \left(\sum_{U} y_i\right)^2 \\
=& \sum_{U}\sum_{U} \frac{y_i}{\pi_i}\frac{y_j}{\pi_j} E[I_{ij}] - \sum_{U}\sum_{U} y_i y_j \\
=& \sum_{U}\sum_{U} \frac{y_i y_j}{\pi_i \pi_j} \pi_{ij} - \sum_{U}\sum_{U} y_i y_j \\
\Rightarrow var(\hat{\tau}) =& \sum_{U}\sum_{U} y_i y_j \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j}\right)
\end{aligned} \tag{3.9}$$

Provided that $\pi_{ij} > 0$ for all $i, j \in U$ an unbiased estimate of this variance is

$$\widehat{var}_{HT}(\hat{\tau}) = \sum_{s_\nu}\sum_{s_\nu} \frac{y_i y_j}{\pi_{ij}} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j}\right) = \sum_{s_\nu}\sum_{s_\nu} y_i y_j \left(\frac{1}{\pi_i \pi_j} - \frac{1}{\pi_{ij}}\right) \tag{3.10}$$

This is unbiased as

$$E[\widehat{var}_{HT}(\hat{\tau})] - \sum_{U}\sum_{U} E[I_{ij}] \frac{y_i y_j}{\pi_{ij}} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j}\right) \text{ and } E[I_{ij}] - \pi_{ij} \tag{3.11}$$

If for the proposed sampling strategy $\nu = n$ for all possible samples, that is a without replacement sampling scheme with all possible samples having the same fixed sample size, so that $s_\nu = s$, the variance can be rewritten as

$$var(\hat{\tau}) = \frac{1}{2} \sum_{U}\sum_{U} (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 \tag{3.12}$$

This reformulation was proposed by Yates and Grundy (1953) and Sen (1973) and an estimator for this is

$$\widehat{var}_{SYG}(\hat{\tau}) = \frac{1}{2}\sum_s\sum_s \frac{(\pi_i\pi_j - \pi_{ij})}{\pi_{ij}}\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 \tag{3.13}$$

Using the same argument as given in equation 3.11 this is an unbiased estimator of $var(\hat{\tau})$. This estimator is more robust than that proposed by Horvitz-Thompson, equation 3.10. In particular under many sampling designs, $\Delta_{ij} = \pi_{ij} - \pi_i\pi_j \leq 0\ \forall i, j$, so that the Sen-Yates-Grundy variance estimator, $\widehat{var}_{SYG}(\hat{\tau})$, is guaranteed to be non-negative. In contrast the Horvitz-Thompson estimator can be negative even when this condition holds. Unless stated the Sen-Yates-Grundy variance estimator will be used throughout this thesis. If an *srswor* design is employed, the estimator of τ, the variance and variance estimator[3] are

$$\hat{\tau} = \frac{N}{n}\sum_s y_i = N\bar{y} \tag{3.14}$$

$$var(\hat{\tau}) = \frac{N(N-n)}{n}\frac{\sum_U(y_i - \bar{y}_U)^2}{N-1} = \frac{N(N-n)}{n}S_U^2 \tag{3.15}$$

$$\widehat{var}(\hat{\tau}) = \frac{N(N-n)}{n}\frac{\sum_s(y_i - \bar{y}_s)^2}{n-1} = \frac{N(N-n)}{n}S_s^2 \tag{3.16}$$

where for a set of units s of size n_s, so that when $s = U$, $n_s = N$

$$\bar{y}_s = \frac{1}{n_s}\sum_s y_i \text{ and } S_s^2 = \frac{1}{n_s - 1}\sum_s(y_i - \bar{y}_s)^2$$

Under *srswor* if there is a large amount of heterogeneity in the y_i, that is if S_U^2 is large, the precision of $\hat{\tau}$ will be poor as the values of $var(\hat{\tau})$ and $\widehat{var}(\hat{\tau})$, in equations 3.12 and 3.13, will be large. Inspection of the variance estimators, equations 3.10 and 3.13, shows that if $y_i \propto \pi_i$ then $var(\hat{\tau}) = \widehat{var}(\hat{\tau}) = 0$. The y_i are unknown at the design stage but if the y_i are correlated to known auxiliary variables, such as habitat $\underline{x}_{i1}$ say, a sampling strategy in which inclusion probabilities are a function of the $\underline{x}_{i1}$ would lead to a reduction in $var(\hat{\tau})$ compared to simple random sampling. Stratified random sampling and sampling with inclusion probability proportional to size are some of the common strategies that

[3] In this case $\widehat{var}_{HT}(\hat{\tau}) = \widehat{var}_{SYG}(\hat{\tau})$

are employed within survey sampling. Many sampling texts such as Cochran (1977) and Thompson (2002) describe these and other strategies in detail.

Stratified random sampling, which we denote *strs*, allows information about several auxiliary variables to be incorporated into the sampling design, by defining a number of strata. The aim is for there to be little variability in the y_i within a stratum but large between-stratum variability. Hence units with similar values of a set of auxiliary variables, that are assumed to be correlated with y_i, are in the same stratum. For example if auxiliary variable x_{i4} in figure 2.1 was thought to be related to y_i, then two strata, one for each value that x_{i4} takes, would be defined. If auxiliary variable x_{i1} and x_{i4} were considered important, additional strata for high and low values of x_{i1} within the two strata defined by x_{i4} may be defined. Alternatively geographical strata may be suggested so that the survey region is divided, for example, into four strata created by splitting the survey region into four quarters. Such stratification however rarely leads to low within stratum variance and high between stratum variance.

When sampling with inclusion probability proportional to size, which we denote πps, the sample design allocates units which are expected to have high values of y_i a high inclusion probability, π_i, and units with a low value of y_i a low π_i, the ideal is that $y_i \propto \pi_i$ so that $var(\hat{\tau}) = 0$. As the value of y_i is unknown the inclusion probabilities are instead based on the value of an auxiliary variable, z say. This we denote πpz and if $cor(y_i, z_i)$ is positive, units with large values of z_i will have a large inclusion probability and units with a small value of z_i will have a small inclusion probability. If more than one auxiliary variable is considered correlated to y_i it can be difficult to formulate an appropriate set of inclusion probabilities. The πps strategy is more commonly used for practical reasons, for example each unit is a lake and the variable of interest is the total number of fish in a region. Then large lakes have a greater probability of being selected than small lakes. When there are multiple response variables, for example there are several study species being monitored, it can be difficult to implement one πps sampling design that estimates the total number of individuals of each species precisely. The πps strategy

does have potential for single variable surveys (Stehman and Overton, 1994) such as the single-species monitoring programmes we describe.

In practice, particularly when little is known about the study species, there is uncertainty as to which variables are correlated with the y_i. Therefore choosing an efficient sampling scheme can be difficult. If many variables are thought to be important, it can be difficult to incorporate them all effectively into the sample design. For example suppose all the habitat variables in figure 2.1 are considered correlated with y_i. One strategy is to derive a new variable x_{i5} that is a function of the four auxiliary variables. Then *strs* can be implemented by defining strata to be sets of units that take a particular range of x_{i5} values or πpx_5 can be implemented by sampling with π_i proportional to x_{i5}. The difficulty here is defining an appropriate function for x_{i5}. Alternatively for each of the four auxiliary variables two strata can be defined, relating to high and low values of each variable. By taking all possible combinations of these two strata for each of four auxiliary variables the survey region is partitioned into a total of $16 = 2^4$ strata. The difficulty here is that 16 is a relatively large number of strata - each having on average 80 units, although some may be very small. To allow variance estimation, a minimum sample size of 32 units, two from each stratum, would need to be taken; large strata will then be poorly sampled unless the total sample size is large. If some of these auxiliary variables are in fact not correlated with the y_i, then this degree of stratification is unnecessary, and so that it would be better to change the stratification for future surveys. However as we describe in section 3.6, this is not always easy to do.

Instead of using the auxiliary information in the design process to reduce the variance, a simple sample design, such as *srswor*, can be implemented and the auxiliary information incorporated into the estimator. These are called model-assisted strategies and are described in detail in Särndal *et al.* (1992). A common form of estimator is the generalised regression estimator (GREG), as proposed by Cassel *et al.* (1976) such that

$$\hat{\tau}_{MA} = \sum_s \left(\frac{y_i - \hat{y}_i}{\pi_i} \right) + \sum_U \hat{y}_i \tag{3.17}$$

where the $\hat{y}_i$ are estimated using a model that describes the relationship between the realised population and the auxiliary data, for example equation 2.4. One possible approximation of the variance is to use the residuals e_i where

$$e_i = y_i - \hat{y}_i \tag{3.18}$$

so that equations 3.12 and 3.13 become

$$var_{MA}(\hat{\tau}) = \frac{1}{2}\sum_U \sum_U (\pi_i\pi_j - \pi_{ij}) \left(\frac{e_i}{\pi_i} - \frac{e_j}{\pi_j}\right)^2 \tag{3.19}$$

$$\widehat{var}_{MA}(\hat{\tau}) = \frac{1}{2}\sum_s \sum_s \left(\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\right) \left(\frac{e_i}{\pi_i} - \frac{e_j}{\pi_j}\right)^2 \tag{3.20}$$

These estimators are approximately design unbiased, so even under a poorly specified model, for example if an important auxiliary variable is not available, the estimators should remain approximately unbiased. If the variability in the $\frac{e_i}{\pi_i}$ is smaller than in the $\frac{y_i}{\pi_i}$, which we would generally expect to be the case, the precision of $\hat{\tau}$ is increased, compared to the Horvitz-Thompson estimator. By fitting weighted models (weights equal to π_i) then in general $var_{MA}(\hat{\tau}) < var(\hat{\tau})$.

The model describes the relationship between the auxiliary variables and the realised population, rather than the underlying state model, hence they are more akin to the models described in equation 2.4. For example a simple model might be of the form

$$y_i = B_0 + B_1 x_{i1} + \epsilon_i \; var(y_i) = \sigma^2 \tag{3.21}$$

As shown in equation 2.5, $\underline{B} = (B_0, B_1)$ can be calculated using ordinary least squares when data from all N units are available. As only a sample has been taken, these parameters are now estimated. If Π_s is a diagonal matrix with values π_i for $i \in s$, $\underline{B}$ is estimated as

$$\hat{\underline{B}} = (\mathbf{x}_s'\Pi_s\mathbf{x}_s)^{-1}\mathbf{x}_s'\Pi_s\underline{y}_s \tag{3.22}$$

If $\pi_i = n/N$ for all i then $\sum_s e_i/\pi_i = 0$ else for unequal inclusion probabilities $\sum_s e_i/\pi_i$ is non–zero. If x_{i1} is a categorical variable then the model-assisted strategy is a form

of post-stratification. The regression estimator methods can be extended so that general parametric regression models (Wu and Sitter, 2001) and local polynomial regression models (Breidt and Opsomer, 2000) can be fitted.

The advantage of the model-assisted method is that simple sample design strategies such as *srswor* can be implemented and the heterogeneity in the y_i is dealt with in the estimation stage so that the precision of the estimators is greater than under the Horvitz-Thompson estimator. However in many cases, using a sample design such as *srswor* can mean that fieldworkers spend a large proportion of their time visiting areas in which species abundance is very low and observing very few individuals of the study species. In addition, although design-robust, the precision of the estimates will change depending on the model assumed.

3.4 Estimating $\delta^{(t',t)}$

If the change in the realised population from survey t' to survey t is

$$\delta^{(t',t)} = \tau^{(t)} - \tau^{(t')} \tag{3.23}$$

we can estimate this change and its associated variance to be

$$\hat{\delta}^{(t',t)} = \hat{\tau}^{(t)} - \hat{\tau}^{(t')} \tag{3.24}$$

$$var(\hat{\delta}^{(t',t)}) = var(\hat{\tau}^{(t)}) + var(\hat{\tau}^{(t')}) - 2cov(\hat{\tau}^{(t)}, \hat{\tau}^{(t')}) \tag{3.25}$$

$$\widehat{var}(\hat{\delta}^{(t',t)}) = \widehat{var}(\hat{\tau}^{(t)}) + \widehat{var}(\hat{\tau}^{(t')}) - 2\widehat{cov}(\hat{\tau}^{(t)}, \hat{\tau}^{(t')}) \tag{3.26}$$

For any sampling scheme the covariance between the two estimators is

$$cov(\hat{\tau}^{(t)}, \hat{\tau}^{(t')}) = E\left[\sum_U \frac{y_i^{(t)} I_i^{(t)}}{\pi_i^{(t)}} \sum_U \frac{y_j^{(t')} I_j^{(t')}}{\pi_j^{(t')}}\right] - E\left[\sum_U \frac{y_i^{(t)} I_i^{(t)}}{\pi_i^{(t)}}\right] E\left[\sum_U \frac{y_j^{(t')} I_j^{(t')}}{\pi_j^{(t')}}\right]$$

$$= \sum_U \sum_U y_i^{(t)} y_j^{(t')} \left(\frac{\pi_{ij}^{(t,t')}}{\pi_i^{(t)} \pi_j^{(t')}} - 1\right) \tag{3.27}$$

where $\pi_{ij}^{(t,t')}$ is the probability that unit i is included in sample $s^{(t')}$ and unit j is included in sample $s^{(t)}$. If for survey t the sample, $s^{(t)}$, is selected independently of the sample $s^{(t')}$ from survey t', then $\pi_{ij}^{(t',t)} = \pi_i^{(t')}\pi_j^{(t)}$ and so the covariance is zero. If however $s^{(t)} = s^{(t')}$ so that the sample is retained from one survey to another, then $\pi_{ij}^{(t',t)} = \pi_{ij}^{(t')}$. If observations on the same unit are correlated between surveys so that $cor(y_i^{(t)}, y_i^{(t')}) = \rho^{|t-t'|}$ then

$$var(\hat{\delta}^{(t',t)}) = var(\hat{\tau}^{(t')}) + var(\hat{\tau}^{(t)}) - 2\rho^{|t-t'|}\sqrt{var(\hat{\tau}^{(t')})var(\hat{\tau}^{(t)})} \tag{3.28}$$

Hence as ρ increases so $var(\hat{\delta}^{(t',t)})$ decreases and the precision of $\delta^{(t',t)}$ will be improved if sample locations are retained from one survey to another.

If abundance estimation is also an aim of the monitoring programme, then a poor design in the initial survey can lead to poor precision of the abundance estimates in each survey when the original sample is retained for future surveys. If a new sample is taken in survey t, efficiency of both $\hat{\delta}^{(t',t)}$ and $\hat{\tau}^{(t)}$ is improved by using a more efficient sampling strategy than that used at time t'.

We can compare the relative increase in the precision of $\hat{\delta}^{(t',t)}$ from using a more efficient sampling strategy at time t to retaining the units from the sample at time t'. Suppose that $var(\hat{\tau}^{(t)}) = var(\hat{\tau}^{(t)})$ if $s^{(t')} = s^{(t)}$. If a different sampling strategy is employed in survey t assume that the two strategies are independent of each other so that $cov(\hat{\tau}^{(t)}, \hat{\tau}^{(t')}) = 0$. Using this new strategy at time t let $var(\hat{\tau}^{(t)}) = evar(\hat{\tau}^{(t')})$ where e is a constant such that if $e > 1$ the precision of $\hat{\tau}^{(t')}$ is less than the precision of $\hat{\tau}^{(t)}$. Hence using a new sampling strategy will be more efficient than retaining the sample $s^{(t')}$ if

$$\begin{aligned} & 2(1 \quad \rho^{|t-t'|}) > 1 + e \\ \Rightarrow\ & \rho^{|t-t'|} < \frac{1-e}{2} \end{aligned}$$

Even if there is only a small correlation in the $y_i^{(t)}$ between surveys, say $\rho^{|t-t'|} = 0.25$, a new sampling strategy would need to increase precision by 50%, in the example given here, to give as precise an estimate of $\delta^{(t',t)}$ as would be obtained by retaining the sample from time t' to time t.

Many monitoring programmes would consider the precision of $\hat{\tau}^{(t)}$ to be as important as the precision of $\hat{\delta}^{(t',t)}$. Often a compromise design is implemented so that both parameters can be estimated as precisely as possible. The principle of these designs is that the sample $s^{(t)}$ consists of retaining a subset, or 'panel', of $m \leq n$ units from $s^{(t')}$; this sub-sample is denoted $s_m^{(t')} = s_m^{(t)}$. A 'panel' of $n - m$ new units are selected from $s_c^{(t')}$; this sub-sample is denoted $s_{n-m}^{(t)}$. These designs were first proposed by Jessen (1942), where a farm survey in 1941 retained 450 of the 900 farms surveyed in 1940 and selected an additional 450 farms from those that were not surveyed in 1940. More complex designs for surveys of more than two years have been proposed and common names for this type of design are 'panel studies' or designs where we 'sample with partial replacement'. These designs have been relatively well known in the area of survey sampling for many years and Duncan and Kalton (1987) and Binder and Hidiroglou (1988) provide a general review of these sampling designs and appropriate estimators. However they have only recently been used in environmental and ecological sampling. Skalski (1990) suggested using these types of designs in environmental modelling and Urquhart *et al.* (1993) have implemented these methods to the large Environmental Mapping and Assessment Programme (EMAP) in the USA (Overton *et al.*, 1990). There is little evidence that these designs have been used for wildlife population assessment. One example is in the assessment of the Arroyo toad (Atkinson *et al.*, 2003).

There are many variants of these designs, but the general idea is based on sets or 'panels' of units that appear in the same surveys. Common strategies are illustrated in Box 3.1. In a serially alternating design, panels are included in every k^{th} survey (in Box 3.1 $k = 3$). In a rotating panel design r panels are included in the sample and in each survey, one panel leaves the design and one panel enters the design. In Box 3.1 we show this design where $r = 3$. Hence after $r - 1 = 2$ surveys, each panel remains in the sample for 3 surveys. More complex designs are a combination of these two strategies, such as the combined panel shown here in which a panel remains in the sample for $r = 3$ surveys, is removed for $k = 2$ surveys and then returns to the sample for another $r = 3$ surveys. Augmented survey designs also have a panel of units that occur in every survey and Urquhart and Kincaid

Box 3.1: Different rotating panel designs

		Survey							
Strategy	Panel	1	2	3	4	5	6	7	8
	1	X			X			X	
Serially	2		X			X			X
Alternating	3			X			X		
	1	X							
Rotating	2	X	X						
Panel	3	X	X	X					
	4		X	X	X				
	5			X	X	X			
	6				X	X	X		
	1	X			X	X	X		
Combined	2	X	X			X	X	X	
Panel	3	X	X	X			X	X	X
	4		X	X	X			X	X
	5			X	X	X			X
	1	X							
New= Never	2		X						
Revisit	3			X					
	4				X				
	1	X	X	X	X	X	X	X	X
Augmented	2	X							
+ New	3		X						
	4			X	...				
Sentinel	1	X			X			X	
sites	2	X	X	X	X	X	X	X	X

(1999) detail other designs. The units in panel l, s_{p_l}, can be selected using any sampling design, but generally they are selected using *srswor* from the units that do not belong to panels $1, \ldots, l-1$. As the sampling strategy is determined at the start of the monitoring programme, the selection of all the panels for T surveys can be completed before data collection for the first survey has started.

Estimation of $\tau^{(t)}$ or $\delta^{(t',t)}$ can be within a design-based or model-assisted framework. In a design-based framework, $\delta^{(t',t)}$ is generally calculated using equation 3.24. Estimating the covariance $cov(\hat{\tau}^{(t)}, \hat{\tau}^{(t')})$ is difficult because part of the sample is overlapping. In practice the covariance is usually estimated from the data of the matched sample, $s_m^{(t)}$, (Holmes and Skinner, 2000). This will generally overestimate the covariance, which may lead to negative variance estimates. We note that Berger (2003b) has recently developed a variance estimator that incorporates all of the data from $s_m^{(t)}$ and $s_{n-m}^{(t)}$.

In the case of model-assisted estimation of $\delta^{(t',t)}$, if all units are selected by *srswor*, the general principle is to assume that

$$y_i^{(t')} = \alpha_0 + \alpha_1 y_i^{(t)}$$

and to use the units in $s_m^{(t)}$ to estimate the parameters α_0 and α_1 using least squares. Hence $\hat{\tau}^{(t')}$ is of the form

$$\hat{\tau}^{(t')} = \hat{\tau}_m^{(t')} + \hat{\alpha}_1(\hat{\tau}^{(t)} - \hat{\tau}_m^{(t)})$$

$$\text{where } \hat{\tau}_m^{(t)} = \frac{N}{m}\sum_{s_m^{(t)}} y_i^{(t)} \text{ and } \hat{\tau}_m^{(t')} = \frac{N}{m}\sum_{s_m^{(t')}} y_i^{(t')}$$

Särndal *et al.* (1992) provide a more general form of estimator when units within a panel are selected using unequal probability schemes. This requires the use of estimators from two-phase sampling, also known as double sampling, to calculate the estimators. Yates (1950) and Patterson (1950) extended the basic method to monitoring programmes with more than two surveys. For many survey designs, for example the rotating panel design, there will be units at time t which were sampled in several previous surveys (in the rotating panel design some units would have occurred in survey $t-1$ and survey $t-2$), hence all this

information can be used to provide improved estimates of $\tau^{(t)}$ and $\delta^{(t',t)}$. More recently, elementary estimators and modified regression estimators such as those described by Fuller and Rao (2001) have been developed. The motivation for many of these estimators is the monthly data collected from the Canadian Labour Force Survey, (Dufour *et al.*, 1998) a rotating panel design where $k = 6$ and n=53,500 households.

A feature of ecological and environmental sampling is the spatial autocorrelation between units $cor(y_i^{(t)}, y_j^{(t)})$. Urquhart *et al.* (1993) compared the effectiveness of different panel designs on the precision of $\hat{\tau}^{(t)}$ and linear trend, η_1, using a model–assisted components of variance approach for populations with varying levels of spatial and temporal correlation between units and surveys respectively. They found that the serially alternating design, rather than the rotating panel design generally gave greater precision. Augmentation was generally only of use in the first few years of a study before panels had occurred twice in the monitoring programme.

An alternative form of sampling design through time is of sentinel sites, as used in many large-scale programmes that monitor epidemics such as AIDS (Chin and Mann, 1989) . Every r years, a cross-sectional study, similar to a Type I study, is performed to estimate the prevalence of AIDs at that particular time. Between these studies, groups of patients which are easily accessible, such as blood donors, are sampled. These groups are known as sentinel sites. The sentinel sites do not have to be the most at-risk groups in the population; in fact a range of groups with different risk factors is more informative. These are followed as in a Type II study, and used to indicate areas where risks are increasing. The last design in Box 3.1 represents this strategy. In years 1, 4 and 7, estimates of abundance are obtained (panel 1) and in intervening years only the sentinel sites (panel 2) are sampled. Indices from the sentinel sites might need to be calibrated with the information obtained from the cross-sectional estimates of prevalence. This strategy has not, to my knowledge, been applied to wildlife populations although in some circumstances it may be applicable. For example on Round Island, it may be that the resources for a large scale Type I study

can only take place every few years[2]. In these years a detailed estimate of the population could be carried out. In the intervening years some small sub-populations or sub-regions, such as gullies or areas where particular management practices have been put into place, could be monitored and may indicate if there are particular problems. These are then calibrated with the results from the detailed population estimate.

The strategies described here retain part of the sample from previous surveys so that $\delta^{(t',t)}$ is precisely estimated, if $cor(y_i^{(t)}, y_i^{(t')})$ is large and positive. In addition, by allowing part of the sample to change from one survey to another, and by using observed data from previous surveys the estimate of $\tau^{(t)}$ should increase in efficiency through time.

We started this section by describing how we might wish to change the survey design to obtain more precise estimates of $\tau^{(t)}$, based on information gained from previous surveys. However, the strategies described here do not change the design based on the observed data from previous surveys. Rather, they use the data from the previous surveys to improve the estimators. Generally panels are selected using a strategy such as *srswor* so that $s^{(1)}, \ldots, s^{(T)}$ can be selected before the data collection stage of the first survey begins.

3.5 Obtaining maps of species distributions, at one time and through time

For motile populations, parameters of a QSM such as $\hat{\mu}_i^{(t)}$ can provide a useful representation of species density and how species density changes through time, as described in section 2.3. For example Khaemba and Stein (2000) uses observed counts from nine surveys and data on 12 auxiliary variables describing habitat and distances from roads, rivers and park boundaries to model the spatial distribution of elephants in a Kenyan wildlife reserve.

[2]For example a team of experts have visited the island every seventh year since 1975, and there are usually more resources available at the time of these visits

For a motile population, it is unlikely that a map of the realised population is required. If it is however, then geostatistical methods such as kriging (Cressie, 1991, pp 151–183), can be useful in estimating the $y_i^{(t)}$ for unobserved units. Borchers *et al.* (2002) suggest that fitting a thin-plate spline of latitude and longitude, when individuals are thought to occur in clusters, provides a prediction surface for the realised distribution as it does not matter whether correlation in the $y_i^{(t)}$ is due to trend or autocorrelation. This is a slightly different approach from kriging in that with a spline fit $\hat{y}_i^{(t)}$ is obtained for all units in U, even those for which $y_i^{(t)}$ was observed.

Estimation of parameters of the QSM falls within a model-based framework. One advantage of this framework, as previously stated, is that for a specified form of the linear predictor, for example one of the form

$$log(\mu_i^{(t)}) = \sum_{j=0}^{Q} \beta_j x_{ij}^{(t)},$$

it is (theoretically) possible to choose the set of units that will maximise the precision of $\hat{\mu}_i^{(t)}$, $\hat{\beta}_j$ or some function of these estimators. These methods have their basis in response surface designs. For example under a standard regression model with one auxiliary variable, x_{i1},

$$y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$$

and using differentiation and standard least squares theory we can show that the precision of $\hat{\beta}_1$ is maximised when the sample s contains the units with the most extreme values of x_{i1}. When the form of the model is not known exactly, for example some terms in the model are smoothing splines, or when the model contains several auxiliary variables, the optimal set of units cannot be (easily) specified. A general principle might be that for any auxiliary variable x_{ij} for which there is uncertainty about its functional form in the linear predictor, the sample must provide a good range and spread of values to give flexibility in determining the functional form.

Our interest is in using $\mu_i^{(t)}$ for prediction, so we would want to select the sample s that minimises the prediction error. When the functional form is known it may be possible to

specify a simple optimality criterion. However the functional form is often unknown. A heuristic choice of sample may start by selecting units close to maximum and minimum values of the auxiliary variables within the survey region so that the predicted values are not extrapolated from the model. Similarly we would wish to ensure that areas of high variability in $\mu_i^{(t)}$ are sampled. As the relationship between $\mu_i^{(t)}$ and several auxiliary variables is often unknown, a sample that provides good coverage of the X-space could be useful.

In spatial sampling, a systematic sampling design, which we denote *sys*, in which units in $s^{(t)}$ are spaced out in a regular grid over the survey region, has been shown to be effective for estimating $\tau^{(t)}$ and individual $y_i^{(t)}$ values when the $y_i^{(t)}$ are spatially correlated (Matérn, 1986). Bellhouse and Rao (1975) shows that in a design-based framework, systematic sampling is an efficient sampling strategy for estimating $\tau^{(t)}$ when the $y_i^{(t)}$ are assumed to be generated by a QSM in which $cov(Y_i^{(t)}, Y_j^{(t)}) > 0$. By spacing the units as far apart as possible, the $y_i^{(t)}$ values will be as uncorrelated as possible. If there is spatial trend in the survey region, spreading the units out should give the greatest coverage of the range of the $y_i^{(t)}$. In the survey region, many auxiliary variables will also exhibit spatial autocorrelation. Hence a systematic sample, provides good spatial spread and hence good coverage of the X-space which may be useful for minimising the prediction error of $\hat{\mu}_i^{(t)}$. In particular if a thin-plate spline of latitude and longitude is included in the model selecting units close to the edge of the survey region would ensure that there are not large areas of extrapolation. To some extent a systematic sample ensures that these units are selected. A disadvantage of systematic sampling is that autocorrelation cannot easily be estimated, as units are not close together.

A difficulty with systematic sampling is that there is usually only one random start so that an unbiased estimate of the variance of $\hat{\tau}$ cannot be obtained. One strategy is to have more than one random start. Alternatively Wolter (1985) suggests a number of variance estimators. Many practitioners assume that the ordering of units is similar to that from *srswor* and use the *srswor* variance estimator, $\widehat{var}_{srs}(\hat{\tau})$ (Stehman and Overton, 1994).

This is a reasonable assumption as long as there is no periodicity in the survey region that is aligned with the grid. For example if the survey region is a forest all the sampled units might occur on the edge of forest tracks, where species density is consistently low. In this case $\widehat{var}_{srs}(\hat{\tau})$ will be an underestimate of the true variance. If however the grid is not aligned with periodicity in the $y_i^{(t)}$, then $\widehat{var}_{srs}(\hat{\tau})$ will tend to overestimate the true variance unless individuals are randomly distributed throughout the survey region (Strindberg, 2001). As there is often a trend in the species distribution over the survey region, the use of $\widehat{var}_{srs}(\hat{\tau})$ is not a correct estimate of $var_{sys}(\hat{\tau})$.

3.6 Adaptive strategies

In section 3.4 there is no clear framework for changing the sampling design through time as more is learnt about the distribution of the species over the survey region; instead improvements in efficiency are obtained by using model-assisted estimators. Although the wildlife manager gains from this information by obtaining more precise parameter estimates, the fieldworker perceives no direct benefit, as areas in which it is known that there will be few sightings of the study species must still be visited. Adaptive cluster sampling (Thompson and Seber, 1996) is a sampling strategy that enables fieldworkers to adapt their sample design whilst in the field so that they can observe more individuals of the study species. This falls into the first type of adaptive design we describe in which adaptation occurs within the data collection component of one survey. We also describe strategies where the sample design adapts from one survey to another, based on changing priorities or past observations.

3.6.1 Adapting the sampling strategy within a survey

At time t, a sample s_1 of n_1 units is taken and the data $\underline{y}_{s_1}$ observed[3]. The information from these data is used to select, using some adding rule, a second phase sample s_2 of

[3]In this section we do not include the superscript $^{(t)}$ as everything relates to one survey only

n_2 units. In total there are $G \geq 1$ phases and the final sample consists of n units where, depending on the strategy, n is fixed at the start of the survey or is unknown until no more units meet the criteria for sampling additional units. The data $(\underline{y}_{s_1}, \ldots, \underline{y}_{s_G})$ and relevant inclusion probabilities are used to obtain an estimate of $\hat{\tau}$.

Adaptive allocation, proposed by Francis (1984) and a Jolly and Hampton (1990), is a strategy in which the sample size is fixed in advance. We describe here the strategy of Francis (1984). A sample s_1 of size n_1 is selected based on some prior stratification of H strata and the data for these units collected. Using these data, the remaining units are allocated sequentially to strata in such a way that the reduction in within–stratum variances is maximised. Suppose that in stratum h there are N_h units of which n_{1h} are initially sampled so that $\sum_{h=1}^{H} n_{1h} = n_1$. Let

$$\widehat{var}(\hat{\tau}_h) = \frac{N_h(N_h - n_h)}{n_h} s_h^2$$

where s_h^2 is the variance of the sampled data. Then the reduction in $\widehat{var}(\hat{\tau}_h)$ by sampling an extra unit from this stratum is estimated to be

$$\hat{R}[\widehat{var}(\hat{\tau}_h)] = \frac{N_h^2}{n_h(n_h+1)} s_h^2.$$

$\hat{R}[\widehat{var}(\hat{\tau}_h)]$ is calculated for all $h = 1, \ldots, H$ strata. The $(1 + \sum_{h=1}^{H} n_{1h})^{th}$ unit is sampled from stratum h^* where

$$\hat{R}[\widehat{var}(\hat{\tau}_{h*})] = max(\hat{R}[\widehat{var}(\hat{\tau}_h)]) \text{ for } h = 1, \ldots, H.$$

$\hat{R}[\widehat{var}(\hat{\tau}_{h*})]$ is now recalculated so that

$$\hat{R}[\widehat{var}(\hat{\tau}_{h*})] = \frac{N_h^2}{(n_h+1)(n_h+2)} s_h^2.$$

The next unit is again allocated to the stratum with the maximum $\hat{R}[\widehat{var}(\hat{\tau}_h)]$ and so on. Once all the remaining units have been allocated, the second phase of data collection can begin. Francis (1984), Brown (1999) and Thompson and Seber (1996) have all shown that this strategy gives negatively biased estimates, and that the similar approach of Jolly and

Hampton (1990) is also biased. In the case of Jolly and Hampton (1990), it is possible to estimate this bias and if n_1, is a large proportion of n the bias will be small.

The adaptive designs described in Thompson and Seber (1996) are strategies with no fixed final sample size. Instead, an initial sample of n_1 units is taken using, for example, *srswor* and the data for these n_1 units collected. If $y_i \geq C$, where C is some pre-defined number, the units in the 'neighbourhood' of i are added to the sample. If for any of these units, j say, $y_j \geq C$ then the units in the neighbourhood of j are added to the sample. Units continue being added to the sample until the neighbourhood of all units where $y_i \geq C$ have been sampled. The neighbourhood is defined before data collection begins. Under plot sampling the neighbourhood of unit i is often defined as the four units that share an edge with unit i. Other neighbourhoods can be defined, the key being that they are symmetric so that if unit j is in the neighbourhood of unit i, unit i will also be in the neighbourhood of unit j. The final sample will consist of three sets of units:

1. the units where $y_i \geq C$;
2. the units for which $y_i < C$ and $i \in s_1$;
3. the units for which $y_i < C$ and $i \notin s_1$.

Figure 3.1 illustrates the adaptive sampling process where in figure 3.1(a) an initial sample of $n_1 = 15$ units are taken. Those units that meet the criterion $y_i \geq C = 3$ are shaded pink, those that do not are shaded dark green. In figure 3.1(b) the units in the neighbourhood of the pink units from s_1 are added to the sample. Those that meet the criterion are shaded pink and those that do not are shaded light green. Figure 3.1(c) shows the final sample when the neighbourhood of all units that meet the criterion have been sampled. In total 30 units have been sampled. Figure 3.1(d) illustrates the final sample that would be obtained if the criterion $C = 2$ rather than 3.

The estimator of τ works by defining networks. Each network is a set of units such that if any one of these units is included in the initial sample, s_1, all other units in the network

would also be selected. In figure 3.1(c) a network is either a set of contiguous pink units (set 1), a single green unit, one that did not meet the criterion either in the initial sample (set 2, dark green) or in later samples (set 3, light green). Inclusion probabilities can be defined for the networks in which at least one unit was included in s_1. That is the units in the first two sets described above, shaded pink or dark green. An unbiased estimate of $\hat{\tau}$ is obtained by using an adaption of the Horvitz-Thompson estimator for these networks.

The adaptive cluster sampling strategy described above was originally designed for sampling rare and clustered populations. In these circumstances the fieldworker will often spend a lot of time in the field not seeing any individuals. As the species is clustered, once some individuals have been observed, others are often seen in the surrounding area. As most fieldworkers are interested in the species, they will want to observe these additional individuals. Hence a sampling scheme that incorporates this type of searching matches the behaviour of the fieldworker, and for rare species, it allows data on more individuals in the population to be obtained.

In practice this sampling strategy can be hard to implement. The choice of C is crucial for determining whether there is a feast (a never-ending number of units for which $y_i \geq C$) or a famine (there are no units for which $y_i \geq C$). Given this it can be difficult to plan how much fieldwork there will be. Salehi and Seber (1997) suggest a two-part strategy in which the survey region is divided into primary sampling units, PSUs. In the first stage, only some, n_{ps}, of these PSUs are selected. In the second stage, standard adaptive cluster sampling is applied to each of the selected PSUs. By confining clusters to remain within a PSU, a limit on the final sample size is obtained as the maximum number of units that can be sampled is $\sum_{i=1}^{n_{ps}} n_i$ where n_i is the number of units in the i^{th} PSU. Christman and Lan (2001) suggests an adaptive strategy that has a stopping rule, so that units are added until $\sum_s I[y_i \geq C] \geq \mathcal{K}$. Pollard and Buckland (1997) and Pollard *et al.* (2002) suggest a fixed effort adaptive sampling strategy for distance sampling surveys in which greater adaptation of the original sample design occurs at the start of the survey when a large proportion of resources are still available, compared to at the end of the survey.

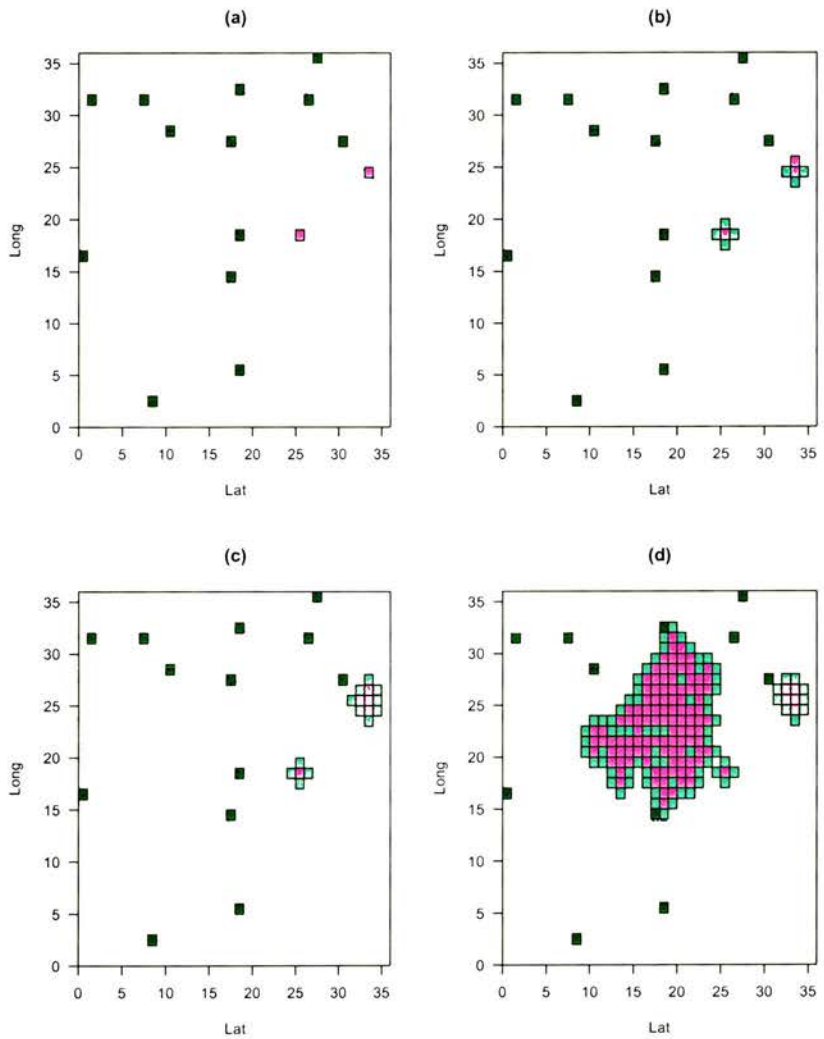


Figure 3.1: An adaptive sampling design for $\mathcal{C} = 3$. (a) An initial sample of $n_1 = 15$ units are selected. The units for which $y_i \geq 3$ are shown in pink. Those where $y_i < 3$ are shown in dark green. (b) The neighbourhood of the units for which $y_i \geq 3$ are sampled. Pink units have $y_i \geq 3$ and light green units have $y_i < 3$ (c) The final sample (d) The final sample from using an adaptive strategy with $\mathcal{C} = 2$.

The efficiency of adaptive cluster sampling compared to *srswor* or other strategies depends on the population being sampled. In general the greater the variability in the y_i within the networks the more efficient adaptive cluster sampling compared to *srswor* (Thompson, 2002). In simulation studies Christman (2000) shows that stratified sampling in which one stratum contains almost all the rare elements will do better than adaptive cluster sampling, although this is only useful when the high density areas are known; after a first survey this may be possible in the second survey. Systematic sampling also performed better than adaptive cluster sampling. One reason for the lack of efficiency is that some units, those for which $y_i < C$ and $i \notin s_1$ are not included in the estimate of abundance.

An implicit assumption seems to be that the $y_i^{(t)}$ remain constant over the whole period in which the survey is carried out. It is not clear as to the consequences of breaking this assumption, for example in highly motile species, in particular if there are large intervals between the various phases of the survey.

3.6.2 Adapting the sampling strategy between surveys

The principle idea is that the sampling strategy at time $t+1$ is adapted based on what is learnt from previous surveys at time $1, \ldots, t$. An informal process by which this happens is the pilot study in which methods are tested out in a small survey. The Type I study is then adapted based on lessons learnt from the pilot. In many cases these lessons relate to survey design as well as issues to do with data collection, but there is no formal process and in general the pilot study is not considered part of the monitoring programme. We consider here more formal methods for changing the sampling design within the course of the monitoring programme.

The main work in this area is by Overton and Stehman (1996) in which they provide a strategy for redefining the strata part way through a monitoring programme. Their motivation is that, in a long-term monitoring programme with multiple objectives, the sample can become out-of-date. This may occur if there have been habitat changes, so

that what was wetland is now agricultural land, or because the importance of various objectives of the monitoring programme may change through time. The restructuring depends on calculating $N_{h,h'}$, the number of units in the old stratum h that occur in the new stratum h'. The cross-stratum $U_{h,h'}$ is the set of these $N_{h,h'}$ units. In this cross-stratum the set of units $s_{h,h'}$ of $n_{h,h'} \subseteq N_{h,h'}$ units were in the original sample. Using principles of post-stratification, the number of units in the final sample from each of the new strata $n_{h'}$, and the number of units $n^*_{h,h'}$ that are needed from each of the cross-strata, are calculated so that $\sum_{h=1}^{H} n_{h,h'} = n_{h'}$ The new sample $s^*_{h,h'}$ from cross-stratum $U_{h,h'}$ is created by removing $n^*_{h,h'} - n_{h,h'}$ units from $s_{h,h'}$ or increased by adding an extra $n^*_{h,h'} - n_{h,h'}$ units from $U_{h,h'} - s_{h,h'}$ to $s_{h,h'}$ using an appropriate sampling scheme.

This restructuring allows for comparability between surveys that occur before and after the restructuring, but the process cannot be repeated many times within the course of a monitoring programme. Unless samples are originally selected using *srswor* within strata, the sampling to reduce or increase the relevant cross-stratum sample can be complicated. The work was motivated by the EMAP project (Overton *et al.*, 1990) which has multiple aims, based on many different indicators of environmental health. Equal probability designs, or at most simple stratification are favoured because of the difficulty of finding inclusion probabilities that are correlated with so many different indicators. They suggest that the precision of estimators should be improved using model–assisted methods appropriate to each separate indicator rather than by the design process.

Within wildlife population monitoring, Haines and Pollock (1998) describe a scheme that uses the $y_i^{(t)}$ to inform the sample design at time $y_i^{(t+1)}$. This scheme was developed for estimating the number of occupied bald eagle nests. At time $t+1$ a sample of locations at which occupied eagle nests was observed at time t are revisited, and in addition a sample of previously unvisited sites in the survey region is also visited. The list of locations where nests are already present is then updated for the following survey. This is a dual-frame strategy, the two frames being the incomplete list of locations at which nests were previously observed and an area frame which defines the survey region as a set of units.

It is assumed that for all units visited in the area frame, the number of nests is recorded accurately. Interest is in estimating the total $\tau^{(t+1)}$, rather than the change in the total through time and an adaptation of the screening estimator first developed by Hartley (1962) is used.

3.6.3 Model-based adaptive methods

If estimation is set in a model-based rather than a design-based framework it is often possible to find an optimal sampling design, where optimality is based on some criterion. For example the optimal sampling design may be one that gives the most precise estimator of a parameter or may minimise the mean-square prediction error. Zacks (1969) showed that in general the optimal sampling strategy is adaptive.

Chao and Thompson (2001) propose a two-phase adaptive sampling strategy to estimate τ under an assumed population model. A first-phase sample is selected using *sys*. Using the observed data from this first-phase sample, a second-phase sample is selected to minimise the mean-square prediction error of τ. This is not a probability sample but an optimal choice of sample locations. Data from both samples are used to estimate τ.

Within environmental monitoring there has been much work on finding model-based designs that add, remove, or change monitoring stations, units, to improve the precision of a required parameter of the future realised population, or of the stochastic process itself. These strategies learn through time using past survey data to develop models of the environmental process. Sampson and Guttorp (1992) and Zidek *et al.* (2000) use temporal information to estimate spatial covariances, that are assumed static in time. Wikle and Royle (1999) consider a dynamic design which accounts for non–separable spatial and temporal correlation, to estimate parameters of the realised population at time t, within a Bayesian context. Their strategy is equivalent to that of Huang and Cressie (1996) if the spatial and temporal correlation are separable. New locations are selected that minimise the maximum prediction variance.

The development of these type of methods within environmental monitoring rather than wildlife monitoring may be because once monitoring stations are in place many observations can be collected from that location with little increased effort. Changes to environmental monitoring networks also occur once a large amount of data is also available, whereas the adaptations in the wildlife monitoring schemes we require are after only one survey and there is a large amount of effort required to collect data even from the same location.

3.7 Discussion

The aim of this chapter was to describe current sampling strategies that can be implemented in a monitoring programme of a single motile species. The key objectives of the monitoring programme are to obtain precise and unbiased estimates of the total number of individuals in the survey region at the time of the survey, $\tau^{(t)}$, and the change in the number of individuals in the survey region through time, $\delta^{(t',t)}$. In this thesis we consider only design-based estimators of these parameters.

At the start of the monitoring programme we assume that little is known about the study species, in particular about its spatial distribution over the survey region. Hence a simple survey design, such as simple random sampling without replacement (*srswor*) may be implemented. This can give imprecise estimates of $\tau^{(t)}$ because of the variability in the samples as large areas of the survey region may not selected in any particular survey. One strategy to ensure more even coverage of the survey region is to use a systematic (*sys*) sample design, although we have noted that variance estimation is difficult and so in practice the variance of $\hat{\tau}^{(t)}$ is estimated assuming *srswor* so that there is no observable increase in precision.

Auxiliary variables may be available and if we understood how these variables are related to the spatial distribution it would be possible to use a survey design such as stratification (*strs*) or sampling with inclusion probability proportional to size (πps) as described in

section 3.3. If well specified, these sampling strategies could improve the precision of $\hat{\tau}^{(t)}$ compared with implementing *srswor*. However the relationship between the auxiliary variables and the species distribution is often unknown, or if there are several variables that are thought to be related to the species distribution it can be difficult to incorporate all of these variables into the sample design. An alternative sampling strategy is to use a simple sample design, such as *srswor*, and incorporate auxiliary information into the estimate of $\hat{\tau}$ using a model-assisted estimator. These estimators can give more precise estimates of $\hat{\tau}^{(t)}$ than a fully design-based estimate of $\hat{\tau}^{(t)}$.

To obtain an estimate of $\delta^{(t',t)}$, as well as $\tau^{(t)}$, a common sample design is that of the rotating panel design in which a sub-sample in each survey is retained from one survey to another and the rest of the sample is selected from units not previously sampled. In general these different sub-samples are selected using *srswor*. If units are correlated through time then the precision of $\delta^{(t',t)}$ will decrease compared to selecting a completely new sample in each survey using the same sample design. In addition by using model-assisted estimators, that use the observed counts from past surveys, the precision of the estimates of $\delta^{(t',t)}$ and $\hat{\tau}^{(t)}$ may be improved further. These model-assisted estimators do not however make use of auxiliary information.

So a monitoring strategy could consist of a very simple survey design such as *srswor* and estimates of $\tau^{(t)}$ could be obtained using a model-assisted estimator that incorporates auxiliary information. In addition, when an estimate of $\delta^{(t',t)}$ is also important, the rotating panel designs that could also use a model-assisted estimator based on the observed counts from past surveys could be implemented. In both these cases there is no reason to change the survey design through time; a new sample may be selected but using the same sampling scheme, for example *srswor*.

As the monitoring programme progresses more information about the spatial distribution of the species is gathered but there is no framework for integrating this information into the survey design. In fact as the sub-samples tend to be selected using *srswor*, and hence do not depend on the results of previous surveys, the samples $s^{(1)}, \ldots, s^{(T)}$ could be

selected before data collection for the first survey starts. Hence areas known to have low, or zero, abundance may still be surveyed quite intensively. For fieldworkers this can be discouraging; particularly if they are involved in the monitoring programme for more than one survey. Although (hopefully) they are informed that the monitoring programme is providing useful information, they themselves see no benefit in terms of being able to see more individuals of the study species. An additional aim of our monitoring programme is that the observers also benefit from what is learnt about the spatial distribution of the species. We will measure this by the number of individuals of the species that are observed in a survey.

The motivation for much of the development of adaptive sampling strategies, in particular those by Thompson (1990), was to enable more observations of the study species. We described adaptive sampling strategies in which adaptation occurs within one survey so that the original sample, and the additional units added to the sample all contribute to the estimate of $\tau^{(t)}$. These designs use the observed counts to determine which additional units should be added to the sample.

There has been little work on design-based adaptive sampling strategies through time. The work of Overton and Stehman (1996) provides a framework for changing a stratified sample design within a monitoring programme so that there is consistency between surveys using the old and the new strata but this is not a framework for a constantly adapting sample design through time; it should only be used sparingly within the life of a monitoring programme. The work of Haines and Pollock (1998) returns to a sample of units for which $y_i^{(t)} > C$, and also selects a new sample of previously unvisited units. However the selection of the new sample does not incorporate any information gained about the species distribution from previous surveys.

In the design-based literature there is no framework for an adaptive monitoring design through time that learns from past surveys to determine future survey design. However the adaptive sampling methods of Thompson (1990) give an indication of how such a strategy may work.

A motivating example comes from a forest inventory of *Prunus africana* on Mount Cameroon carried out in 2000. The bark of the tree is used in medicine for prostate disorders and a pharmaceutical company wished to use the bark from the trees on Mount Cameroon. The species is on CITES Appendix II so a quota needed to be set for the total amount of bark that could be taken, and hence the total number of trees from which bark could be removed. This required an estimate of the total number of trees on the mountain, τ. As the species occurs in clusters and is relatively sparse, an adaptive strip transect method was employed. This consisted of sampling all circular plots along a number of randomly chosen transects, where transects are orientated to run down the mountain. Further details can be found in Burn and Underwood (2000). Because of the terrain and as it can be difficult to keep track of which plots must be added, the survey teams were required to go into the field several times. In particular they needed to return to locations of high density where the sampled plots meet the criterion, $\mathcal{C}$ and additional plots in the neighbourhood had not as yet been sampled. As the species is sessile this was possible to do.

Within a monitoring context we can see how this strategy could be employed through time. The first survey sample $s^{(1)}$ would consist of the initial sample s_1 and an estimate of $\tau^{(1)}$ obtained. In the second survey, to estimate $\tau^{(2)}$, sample $s^{(2)}$ would consist of the neighbourhood of units for which $y_i^{(1)} \geq \mathcal{C}$. In the third survey, to estimate $\tau^{(3)}$, the sample $s^{(3)}$ would consist of neighbourhood of the units for which $y_i^{(2)} \geq \mathcal{C}$, and so on ... Although this suggests a feasible sample design, especially for a sessile species, it is not at all clear how to estimate $\tau^{(t)}$ for $t = 2, \ldots$.

We are interested in motile populations. We discussed in Chapter 2 how for a motile population an estimate of $\mu_i^{(t)}$, the QSM, may be a more useful guide to the spatial distribution of the species than $y_i^{(t)}$. Although this map of the spatial distribution is seen as useful by wildlife managers, and may be a desired output from a monitoring programme, we suggest that it is also of use to the statistician for future survey design. By definition we expect $y_i^{(t)}$ to, on average, be proportional to $\mu_i^{(t)}$ so we could use the $\mu_i^{(t)}$ in place of the $y_i^{(t)}$ in a form of scheme described above. After the first survey it is possible to

construct a QSM, ζ and we have described in section 3.5 why a systematic sample design may be useful when trying to construct ζ. The data from one survey may not be sufficient to obtain a good estimate of $\mu_i^{(t)}$. Through time as more data becomes available our model can improve as we learn more about the relationship between habitat and species abundance and we obtain a map of the spatial distribution of the species. As we have described in this chapter, there is a body of literature that investigates how a model of the superpopulation process constructed using past survey data can be used to determine the optimal sample design for a future survey.

We are working within a design-based framework. We can obtain an estimate of $\mu_i^{(t)}$ after one, or several surveys. This can be obtained for all units in the survey region. Hence unlike in adaptive sampling in which units are added only in the neighbourhood of previously sampled units we can propose a scheme in which units with high $\hat{\mu}_i^{(t)}$ anywhere in the survey region have a greater probability of being selected in the following sample. The estimate of $\mu_i^{(t)}$ is a summary of the relationship between auxiliary data and the expected number of individuals within unit i. There are design-based strategies such as *strs* and πps that can be implemented when auxiliary data that is correlated with the $y_i^{(t)}$ are available. So we can use the $\hat{\mu}_i^{(t)}$ to help determine inclusion probabilities for future surveys. Therefore in this thesis we explore how we can use the $\hat{\mu}_i^{(t)}$ to adapt future survey designs so that $\tau^{(t)}$ and $\delta^{(t',t)}$ can be estimated as precisely as possible using fully design-based estimators. In addition we wish the design to change through time as our knowledge about $\mu_i^{(t)}$ improves. We also hope that, as in the spirit of adaptive sampling, this strategy will increase the number of individuals observed.

We conclude this chapter by noting that the type of strategy we are proposing is akin to the general principle stated by Hansen *et al.* (1983) that

> ...design decisions may be guided and evaluated by models, but inferences concerning population characteristics should be made on the basis of induced randomization...

Chapter 4

Sampling strategy for efficient and robust estimation of $\tau^{(2)}$

At the start of a monitoring programme there is generally little knowledge about the distribution of the species over the survey region, even if data on auxiliary variables are available. Therefore in the first survey an appropriate sampling design is a simple scheme such as *srswor* or *sys*. Data from this survey can be used to estimate $\mu_i^{(1)}$ and so provide a predictive map of species abundance. This is a succinct summary of how the auxiliary data are related to species abundance. A potential use of this map is to assist in future survey design. This chapter describes a suite of sampling designs that make use of this map and the process that led to their development.

To explore how we can use $\hat{\mu}_i^{(1)}$ in survey design, we reduce the problem to a few essential components. Firstly we assume, in this chapter, that the only objective of the second survey is to obtain an efficient design-based estimate of $\tau^{(2)}$ (Obj, 1(a)). In addition, although we are aware that units in the survey region are spatially related, we do not consider this to be part of the survey design problem in this or the following two chapters. We will assume, for the purposes of developing the strategies in this chapter, that the survey region consists of $N = 1000$ units from which $n^{(2)} = n^{(1)} = n = 50$ units are

to be selected in each of two surveys. Appendix B describes a population A, with these characteristics that is used in this chapter. In the first survey *srswor* is used to select the sample, $s^{(1)}$. The predicted map of species abundance, $\hat{\mu}_i^{(1)}$, can be obtained using the data from this sample. In the second survey, $\hat{\mu}_i^{(1)}$ is the only auxiliary variable used to select $s^{(2)}$.

As described in section 3.3 we know that the strategies of *strs* and πps can give more precise estimates of $\tau^{(2)}$ than that obtained under *srswor*. This increase in precision occurs if the auxiliary variable that determines the stratification, or the size variable in sampling with πps, is correlated with $y_i^{(2)}$. An obvious first step is to consider how $\hat{\mu}_i^{(1)}$ can be used in a πps or *strs* design. However $\hat{\mu}_i^{(1)}$ is only an estimate of an unknown $\mu_i^{(1)}$ from some assumed QSM. We would therefore expect the efficiency of the strategies to depend on whether $\hat{\mu}_i^{(1)}$ and the assumed QSM are a good representation of how $y_i^{(2)}$ are generated.

We start in section 4.1 by considering the simpler survey design problem in which the $y_i^{(2)}$ are generated using a known QSM with mean $\mu_i^{(2)}$. The question addressed here is how the strategies of *strs* and πps be implemented using $\mu_i^{(2)}$ as the auxiliary variable. Section 4.2 investigates how robust these strategies are when we use an estimate $\hat{\mu}_i^{(2)}$ rather than $\mu_i^{(2)}$ as the auxiliary variable. By defining a simple summary statistic, b, that describes how well $\hat{\mu}_i^{(2)}$ estimates $\mu_i^{(2)}$, we show that the strategies of *strs* and πps are not that robust to model mis-specification. Therefore in section 4.3 we develop a suite of more robust combined sampling strategies. A simple simulation study in section 4.4 shows how these can be used in practice. A competing method is that of model-assisted estimation. Section 4.5 describes a comparable model-assisted strategy and compares this with the design-based combined sampling strategies.

4.1 Sampling using μ_i

Assume that for an unspecified time period the QSM, ζ, is of the form

$$E_\zeta[Y_i] = \mu_i \qquad var_\zeta[Y_i] = \sigma_i^2 \qquad cov_\zeta[Y_i, Y_j] = 0 \tag{4.1}$$

In this section we address how μ_i can be used as the size variable if sampling with πps, we denote this strategy $\pi p\mu$, or how μ_i can be used as the variable that determines the stratification process; we denote this strategy $strs\mu$.

We compare these sampling strategies using the model–averaged design variance. Given $\underline{y}_U$ we could use the design variance $var(\hat{\tau})$ to compare strategies. Under a known QSM, when it is the $\underline{\mu}_U$ rather than the $\underline{y}_U$ that are known, we require the model–averaged design variance, $E_\zeta[var(\hat{\tau})]$, see Breidt (1995) as an example. Assuming fixed n, so that the variance is as equation 3.12, the model–averaged design variance is of the form

$$E_\zeta[var(\hat{\tau})] = \frac{1}{2}\sum_U \sum_{\substack{U \\ j \neq i}} (\pi_i\pi_j - \pi_{ij}) \left(\frac{\mu_i^2 + \sigma_i^2}{\pi_i^2} + \frac{\mu_j^2 + \sigma_j^2}{\pi_j^2} - 2\frac{\mu_i\mu_j}{\pi_i\pi_j} \right) \tag{4.2}$$

Under *srswor* and the population model ζ, the model–averaged design variance is

$$\begin{aligned} E_\zeta[var(\hat{\tau})] =& \frac{N-n}{2n(N-1)} E_\zeta \left[\sum_U \sum_{\substack{U \\ j \neq i}} (Y_i - Y_j)^2 \right] \\ =& \frac{N-n}{2n(N-1)} \left[2(N-1) \sum_U E_\zeta[Y_i^2] - 2 \sum_U \sum_{\substack{U \\ j \neq i}} E_\zeta[Y_i Y_j] \right] \\ =& \frac{N-n}{n(N-1)} \left[(N-1) \sum_U (\sigma_i^2 + \mu_i^2) - \sum_U \sum_{\substack{U \\ j \neq i}} \mu_i\mu_j \right] \end{aligned} \tag{4.3}$$

If $\mu_i = \mu$ and $\sigma_i^2 = \sigma^2 \; \forall i$ then the only strategy that can be implemented is *srswor* as we have no other auxiliary data to inform inclusion probabilities and so

$$E_\zeta[var(\hat{\tau})] = \frac{N(N-n)}{n}\sigma^2$$

In general μ_i will vary over the survey region. Then strategies such as $strs\mu$ and $\pi p\mu$ can use the variability in μ_i to reduce the model–averaged design variance. As these strategies are based on μ_i rather than σ_i^2, the model–averaged design variance will have a minimum value greater than zero even when $\pi_i \propto \mu_i$, see equation 4.5. This contrasts with strategies in which $\pi_i \propto y_i$, when $var(\hat{\tau}) = 0$. The greater σ_i^2, that is the scatter of y_i values about μ_i, the greater the minimum value of the model–averaged design variance.

4.1.1 Stratified sampling, $strs\mu$

The key idea is that the area A is divided into H strata, based on a set of criteria. Stratum h contains N_h units from which n_h units are sampled independently of other strata. We will select units within each stratum using *srswor*. The model–averaged design variance is calculated as the sum of the model–averaged design variance from each stratum so that using equation 4.3

$$E_\zeta[var(\hat{\tau})] = \sum_{h=1}^{H} \frac{N_h - n_h}{n_h(N_h - 1)} \left[(N_h - 1) \sum_{U_h} (\sigma_i^2 + \mu_i^2) - \sum_{\substack{U_h \\ j \neq i}} \sum_{U_h} \mu_i \mu_j \right] \tag{4.4}$$

Hence the model–averaged design variance is low if the strata are selected so that the variability between the μ_i within a stratum is low and the variability in the μ_i between strata is high.

To implement $strs\mu$ we must make three decisions:

(i) the number of strata, H;

(ii) the definition of each stratum, U_h, i.e. how to determine which units belong to each stratum and hence the number of units within each stratum, N_h;

(iii) the number of units sampled from each stratum, n_h

These three decisions are interlinked. For example the choice of n_h depends on the units in U_h, i.e on the definition of the strata.

(i) The number of strata If μ_i takes only two values then at most two strata can be defined. If we assume that μ_i takes N unique values, the maximum number of strata is $H = n$ as we must sample from each stratum in the survey region. This strategy is used in the one-per-stratum sampling designs of Breidt (1995), of which systematic sampling is a special case. Unbiased estimates of the stratum variances cannot be obtained in these cases. Breidt (1995) specifies the variance and compares strategies using the model-averaged design variance for specific populations but gives no analytic variance estimators. Some techniques for estimating the variances under various population assumptions are described by Wolter (1985), and Särndal *et al.* (1992) describes a collapsed stratum approach in which strata are formed into pairs; this tends to overestimate the variance. In general we would like to sample at least two units within each stratum, so H can take the values $2, \ldots, n/2$ for even n. As H increases so the average value of N_h and n_h decreases. If strata are well-defined the within-stratum variance will decrease as H increases and to estimate the within stratum variance with the same precision, the sample size n_h decreases. Too great a decrease in n_h will lead to the within-stratum variance being poorly estimated; this occurs when there are too many strata. So there is a trade-off between decreasing within-sample variance, by increasing H, and having sufficient resources to estimate the within-sample variance efficiently. Cochran (1977) demonstrates this trade-off when stratifying on an auxiliary variable, x_{ij}, assumed to be linearly related to the y_i.

For simplicity we will set $H = 2$. The strata will still be quite variable and so the increase in precision of $\hat{\tau}^{(2)}$ due to stratification will be small. However we would expect the increase in the precision of $\hat{\tau}^{(2)}$ from changing $H = 1$ to $H = 2$ to be greater than the increase in the precision of $\hat{\tau}^{(2)}$ from changing $H = 2$ to $H = 3$.

(ii) Determining n_h Standard methods of determining the sample allocation are those of optimum allocation first described by Neyman (1934), proportional allocation described by Cochran (1977) and disproportional allocation described by Kalton and Anderson (1986). In these strategies it is the properties of the realised population, for example the (estimated) variance of the strata under optimum allocation, that determine the al-

location rule. Cochran (1977) compared the optimum allocation rule and proportional allocation rule and showed that the estimate of the population mean (and hence total) is more precise using the optimum allocation rule than when using the proportional allocation rule. The disproportional allocation rule was devised for populations in which a large proportion of the units in U are such that $y_i \leq C$ for some C often zero. If the two strata are such that one contains most of the units for which $y_i > C$ then the allocation rule disproportionately samples this stratum. Christman (2000) compared all three allocation rules and found that for rare and clustered populations optimum allocation with a good choice of strata tended to give the most precise estimates of τ.

In our case we wish to use characteristics of the QSM rather than the realised population to determine the sample allocation rule. If we follow the principle of optimum allocation, in which n_h is such that $var(\hat{\tau})$ is minimised, the $E_\zeta[var(\hat{\tau})]$, equation 4.4, is minimised when

$$n_h = n\frac{\sqrt{N_h A_h}}{\sum_{h'=1}^{H}\sqrt{N_{h'}A_{h'}}}$$

where

$$A_h = \sum_{U_h}(\sigma_{ih}^2 + \mu_{ih}^2) - \frac{1}{N_h - 1}\sum_{\substack{U_h \\ j\neq i}}\sum_{U_h}\mu_{ih}\mu_{jh}$$

A_h is an expression of the within-stratum variability. Strata with a large within-stratum variability and large N_h relative to other strata will have a large proportion of the total sample allocated to them. We note that this is of a similar form to the optimum allocation rule for minimising $var(\hat{\tau})$,

$$n_h = n\frac{N_h S_{U_h}}{\sum_{h'=1}^{H} N_{h'}S_{U_{h'}}}$$

in that strata with large within-stratum variability S_{U_h} and large N_h relative to other strata will have a large proportion of the final sample allocated to them. The difference in the square-root is due to the difference between the expressions of $E_\zeta[var(\hat{\tau})]$ and $var(\hat{\tau})$.

(iii) Definition of strata Cochran (1977) reviews a number of strategies for defining stratum boundaries. We use the strategy of Dalenius and Hodges (1959) also described by Cochran (1977, p. 129). This was originally developed when the y_i were assumed known, and n_h chosen using the optimum allocation rule. Cochran (1977) showed that this strategy was effective in many circumstances, for example when y_i are unknown but are linearly related to a known x_i. As the key idea is to define strata that consist of units with similar values of the variable on which stratification is based, then it would seem appropriate in this case as well. We note that the development of an optimum allocation rule specifically for minimising the model–averaged design variance, rather than using the optimum allocation rule for minimising the design variance, could be the subject of further work.

The premise for the strategy of Dalenius and Hodges (1959) is that using the optimum allocation rule the $var(\hat{\tau})$ is minimised by choosing strata so that $\sum_{h=1}^{H} \frac{N_h}{N} S_{y_h}$ is minimised where $S_{y_h}^2$ is the variance of the y-values in stratum h. The minimisation depends on the distribution of the y-values and for most assumed densities it can be difficult to calculate the exact optimum. Hence the approximation rule of Dalenius and Hodges (1959), that is also known as the cum $\sqrt{f}$ rule. If y_i was continuous then the cum $\sqrt{f}$ rule would create H strata so that each stratum accounts for $\frac{1}{H}$ of the total integral of $\sqrt{f_Y(y)}$. As y is unknown the auxiliary variable μ is used in it's place. As μ_i, and also y_i, is not continuous a histogram using the N values of μ_i is created.

Let $\mu_{[1]}$ be the minimum value of μ_i in U and $\mu_{[k]}$ the k^{th} smallest value of μ_i. The strategy defines $H-1$ values $k_1, \ldots, k_{H-1}$ that divide the μ_i into H strata. Initially we divide the interval $[\mu_{[1]}, \mu_{[N]}]$ into J, much greater than H, intervals of equal sizes. Let f_j be the number of units that fall into the j^{th} interval, this provides a frequency distribution of the N values of μ. Note that J may be so large that many intervals contain no units. To calculate the cumulative sum of $\sqrt{f_\mu}$ let $F_{j*} = \sum_{j=1}^{j*} \sqrt{f_j}$. Then if there are to be H strata in total the boundaries of the h^{th} stratum are defined by the values $[F_1 + (h-1)F_J/H, F_1 + hF_J/H)$.

Under population A, if $H = 2$ and $n = 50$, stratum 1 contains $N_1 = 558$ units where $\mu_i \leq 3.70$ and stratum 2 contains $N_2 = 442$ units. Of the 50 units in the sample, $n_1 = 20$ units are sampled from U_1 and $n_2 = 30$ units are sampled from U_2. Inclusion probabilities are illustrated in figure 4.1(a).

4.1.2 Sampling with probability proportional to size, $\pi p\mu$

The key idea of sampling with inclusion probability proportional to size (πps) is that, if we use μ_i as the size variable, the inclusion probabilities will be of the form

$$\pi_i = \frac{n\mu_i}{\sum_U \mu_i}$$

Under this scheme, the model–averaged design variance is

$$E_\zeta[var(\hat{\tau})] = \frac{\sum_U \mu_i}{2n} \sum_{\substack{U \\ j \neq i}} \sum_{U} (\pi_i \pi_j - \pi_{ij}) \left(\frac{\sigma_i^2}{\mu_i^2} + \frac{\sigma_j^2}{\mu_j^2} \right) \tag{4.5}$$

In principle $\pi p\mu$ sampling is a simple idea. In practice it is not simple to devise a fixed-size sampling scheme that has the properties of:

- being relatively simple to implement;
- giving the desired inclusion probabilities, π_i;
- enabling π_{ij} to be calculated exactly without heavy computation and is such that for all $i \neq j$ $\pi_{ij} > 0$ and $\pi_i \pi_j - \pi_{ij} > 0$.

Brewer and Hanif (1982) reviewed 50 strategies most of which work only when $n = 2$, or were computationally complex when $n > 2$. Two strategies not mentioned by Brewer and Hanif (1982) are that of Sunter (1977a,b) and Chao (1982). These are relatively simple list-sequential schemes. In this thesis, we use Sunter's strategy. This is a generalisation of a list sequential scheme for selecting units using *srswor*. Most units are selected with

$\pi p\mu$ but the $\pi p\mu$ property is relaxed for a sub-set of units. This is a reasonable approach and in fact could be considered desirable as discussed later in this section. The strategy described by Chao (1982) in which there is no relaxation of $\pi_i \propto \mu_i$ could alternatively have been used, although it is not considered in this thesis.

The general idea of Sunter's method is that units are ranked in some order. Units are selected sequentially with probability proportional to μ except for the last $k^* \geq N - n + 1$ units which are selected using *srswor* . In detail the algorithm for selecting units is:

1. Order the N elements from $k = 1, \ldots, N$ where $\mu_{[1]}$ is the first unit in the list and $\mu_{[N]}$ the last.
2. Starting with $k = 1$ and following the same procedure for the k^{th} unit:
   (a) Calculate $\pi'_{[k]} = (n - n_k)\mu_{[k]}/t_k$ where $t_k = \sum_{j=k}^{N} \mu_{[k]}$, and n_k is the number of units already selected in the previous $k - 1$ elements.
   (b) Draw $\epsilon_k \sim U[0, 1]$.
   (c) Select unit k if $\epsilon_k < \pi'_{[k]}$.
   (d) Continue until $n_k = n$ or, $k = k^* = min\{k_0, N - n + 1\}$ where k_0 is the smallest k for which $n\mu_{[k]}/t_k \geq 1$.
3. Select the remaining $n - n_{k*}$ from $\{\mu_{k*}, \ldots, \mu_N\}$ using *srswor*.

If we let $\tau_\mu = \sum_{i=1}^{N} \mu_i$, inclusion probabilities are calculated as

$$\pi_{[k]} = \begin{cases} \frac{n\mu_{[k]}}{\tau_\mu} & k = 1, \ldots, k^* - 1 \\ \\ \frac{n\bar{\mu}_{k*}}{\tau_\mu} & k = k^*, \ldots, N \qquad \text{where} \quad \bar{\mu}_{k^*} = \frac{t_{k*}}{(N - k* + 1)} \end{cases}$$

and joint inclusion probabilities as

$$\pi_{[k][l]} = \begin{cases} \frac{n(n-1)}{\tau} g_k \mu_{[k]} \mu_{[l]} & 1 \leq k < l < k^* \\ \frac{n(n-1)}{\tau} g_k \mu_{[k]} \bar{\mu}_{k^*} & 1 \leq k < k^* \leq l \leq N \\ \frac{n(n-1)}{\tau_\mu} g_{[k^*-1]} \frac{t_{k^*} - \mu_{[k^*-1]}}{t_{k^*} - \bar{\mu}_{k^*}} (\bar{\mu}_{k^*})^2 & k^* \leq k \leq l \leq N \end{cases}$$

where $g_1 = \frac{1}{t_2}$ and for $k = 2, \ldots, k^* - 1$

$$g_k = \frac{(1 - \frac{\mu_{[1]}}{t_2})(1 - \frac{\mu_{[2]}}{t_3}) \ldots (1 - \frac{\mu_{[k-1]}}{t_k})}{t_{k+1}} = g_{k-1} \left(\frac{t_k - \mu_{[k-1]}}{t_{k+1}} \right)$$

Generally we would order the units so that $\mu_{[1]} = \max(\mu_i)$ and $\mu_{[N]} = \min(\mu_i)$ This ordering tends to give a relatively large number of units with small μ_i selected with *srswor*, compared to the reverse ordering of the data in which there would be a relatively small number of units with large μ_i selected using *srswor*. As stated by Sunter it is often desirable that the smallest valued units are sampled with equal probability as the correlation between the size measure and the variable of interest can become unstable for these units. We could also think of this strategy as smoothing μ_i when μ_i is small. This may be appropriate when we use an estimate of μ_i as we might not expect small values of μ_i to be estimated very well. If the last k^* units have the same value of μ_i, then the whole population is sampled with probability proportional to size. Although Sunter notes that it might be possible to find an optimal ordering of the units so that $\pi_i \pi_j - \pi_{ij}$ is minimised, especially if there is a risk of this being large, we do not explore this further. When there are a few very large units and many small units, most of the sample is selected using *srswor*.

The inclusion probabilities for values of μ_i are shown in figure 4.1(a) in pink. Using population A, figure 4.1(a) shows the inclusion probabilities when $n = 50$ units are sampled by ordering the units from largest to smallest. Ordering the units from smallest to largest does not affect the value of the inclusion probabilities for most units in U. However the

units with the highest 50 values of μ_i would have an inclusion probability of $\pi_i = 0.14$. The different inclusion probabilities when the units are ordered in reverse are shown in pink on figure 4.1(a)

4.1.3 Comparison of strategies

Strategies are compared using the model–averaged design variance. This is a function of both μ_i and σ_i. In section 2.3 we stated that an initial model for the QSM may be very simple and of the form

$$E[Y_i] = var[Y_i] = \mu_i \qquad cov(Y_i, Y_j) = 0 \tag{4.6}$$

although in practice we would often expect $var[Y_i] > \mu_i$. Using this simple QSM, the model–averaged design variance depends only on μ_i. If $\mu_i = \mu$ for all $i \in U$ then the only strategy we can implement based on μ_i is *srswor*. The greater the variablity in the μ_i over the survey region, the greater we expect the model–averaged design variance to be under *srswor* and we would expect alternative strategies such as $strs\mu$ and $\pi p\mu$ in which $\pi_i \propto \mu_i$ to perform better.

Defining $S^2_{\mu,U} = \frac{1}{N-1}\sum_U(\mu_i - \bar{\mu}_U)^2$, we wish to compare our strategies under varying $S^2_{\mu,U}$. For a general comparison, the μ_i were treated as i.i.d. realisations from the random variable $\mathcal{M}$, drawn from a Gamma distribution with shape parameter γ_1 and scale parameter γ_2, such that

$$\mathcal{M} \sim \Gamma(\gamma_1, \gamma_2) \text{ for } i = 1, \ldots, 1000 \qquad \text{where } \gamma_1 = 2 \qquad \gamma_2 = 1, 2, \ldots, 8$$

$$\text{so } E[\mathcal{M}] = \gamma_1\gamma_2 \qquad var[\mathcal{M}] = \gamma_1\gamma_2{}^2 \text{ as } f(\mu_i) = \frac{1}{\gamma_2^{\gamma_1}\Gamma(\gamma_1)}\mu_i^{\gamma_1}\exp^{-\frac{\mu_i}{\gamma_2}}$$

$S^2_{\mu,U}$ increases linearly with $var[\mathcal{M}]$. For a given value of γ_2, 100 realisations of $\underline{\mu}_U$ were generated. For each realisation of $\underline{\mu}_U$, the model–averaged design variance was calculated under each of the sampling strategies of *srswor*, $strs\mu$ or $\pi p\mu$. The mean $E_\zeta[var(\hat{\tau})]$ over the 100 realisations is plotted against $var[\mathcal{M}]$ in figure 4.1(b).

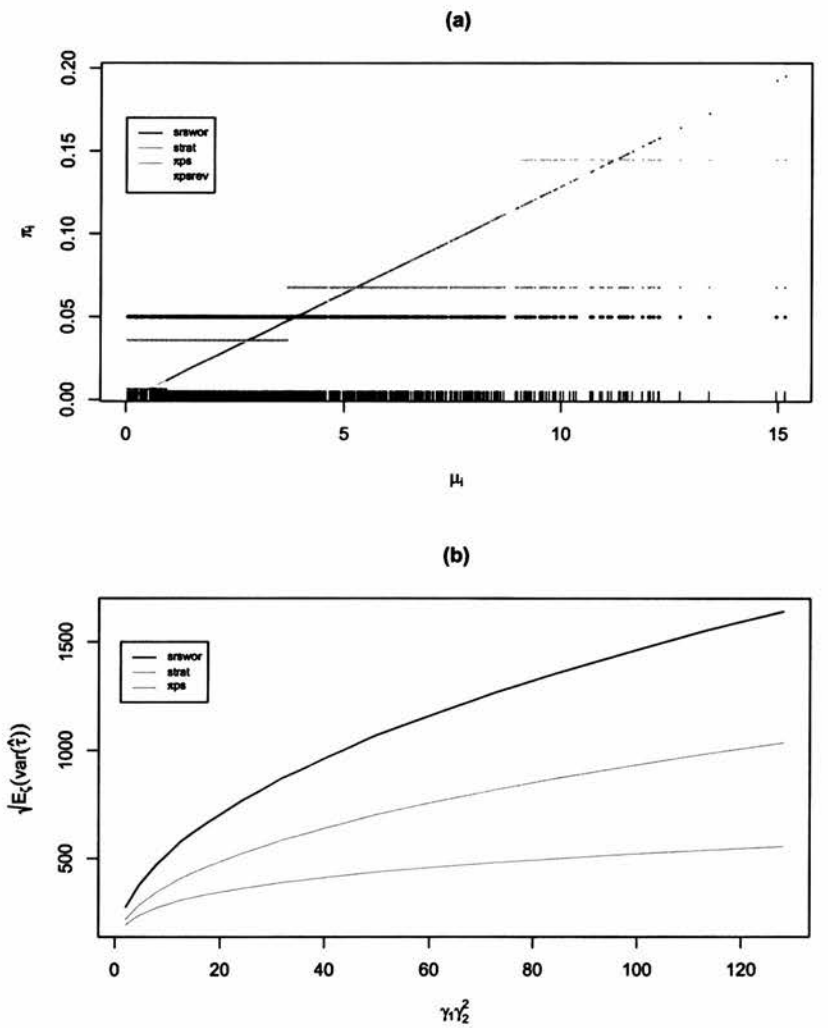


Figure 4.1: (a) Inclusion probabilities for *srswor*, *strs* μ, and $\pi p\mu$ when units are ordered largest to smallest and $\pi p\mu$ rev units ordered smallest to largest. (b) $\sqrt{E_\zeta[var(\hat{\tau})]}$ for *srswor*, *strs* μ and $\pi p\mu$ when μ_i are generated by the stochastic process $M_i \sim \Gamma(\gamma_1, \gamma_2)$

As $var[\mathcal{M}]$ increases, so the model–averaged design variance increases as expected under all strategies. The $\pi p\mu$ strategy gives the lowest model–averaged design variance because the correlation between π_i and μ_i is high. Increasing $var[\mathcal{M}]$ has little effect on the $E_\zeta[var(\hat{\tau})]$ under this strategy. Using $strs\mu$, only some of the variability is taken into account and so the model–averaged design variance is higher than $E_\zeta[var(\hat{\tau})]$ under $\pi p\mu$, but less than the $E_\zeta[var(\hat{\tau})]$ under *srswor*. As the number of strata increases we would expect the model–averaged design variance under *strs* to tend towards the model–averaged design variance for $\pi p\mu$. If μ_i is known, then sampling with $\pi p\mu_i$ is the best strategy for obtaining an efficient estimate of $\tau^{(2)}$. In addition, the decision-making process for implementing this strategy is less complicated than under $strs\mu$.

4.2 Sampling when μ_i is estimated

Assume the y_i are generated from a QSM such that

$$log(\mu_i) = \sum_{j=0}^{Q} \beta_j x_{ij}$$

An estimate of μ_i, using the same auxiliary variables, will be of the form

$$log(\hat{\mu}_i) = \sum_{j=0}^{Q} \hat{\beta}_j x_{ij}$$

If for simplicity we assume that parameter estimates are related in a linear fashion to the parameters so that

$$\hat{\beta}_j = a_j^* + b_j^* \beta_j$$

$$\text{we can write } log(\hat{\mu}_i) = \sum_{j=0}^{Q} (a_j^* + b_j^* \beta_j) x_{ij}$$

For convenience, in particular to obtain a tractable expression, we assume the relationship between parameter estimates and parameters is such that $b_j^* = b$ for $j = 0, \ldots, Q$. Then

$$\begin{aligned} log(\hat{\mu}_i) &= \sum_{j=0}^{Q} a_j^* x_{ij} + b\, log(\mu_i) \\ \Rightarrow \hat{\mu}_i &= a\mu_i^b \end{aligned} \tag{4.7}$$

This is a large oversimplification, but it lends itself to a comparison of sampling strategies under varying amounts of model mis-specification very easily. In practice b_j^* will vary for each auxiliary variable and in many cases $b_j^* = 0$. In addition the effect of b_j^* on $\hat{\mu}_i$ will depend on the relative size of the auxiliary variable. We would expect b to deviate from one when the set of auxiliary variables used to estimate $\hat{\mu}_i$ are not the set of auxiliary variables used to calculate μ_i. We call b a measure of the amount of model mis-specification. When b is close to one we define $\hat{\mu}_i$, and the model ζ, as being well specified. $\hat{\mu}_i$ and $\zeta^{(t)}$ become poorly specified as $|b-1|$ deviates from zero, and moves closer to one.

As b deviates from one, we would expect the strategies of *strs* μ^b and $\pi p\mu^b$ to give less precise estimates of τ. The parameter a is a scaling parameter. Under $\pi p\mu^b$ sampling, varying a will have no effect on the inclusion probabilities as $\pi_i = \frac{na\mu_i^b}{\sum_U a\mu_k^b} = \frac{n\mu_i^b}{\sum_U \mu_k^b}$ and so is not considered further here.

4.2.1 Comparison of strategies

For each realisation of $\underline{\mu}_U$ generated from $\mathcal{M} \sim \Gamma(2,2)$, the strategies of *strs* μ^b and $\pi p\mu^b$ were implemented for $\hat{\mu}_i = a\mu_i^b$ where $b = -2, -1.8, \ldots, 4$.

Strategies were basically implemented as described in section 4.1. In addition under *strs* μ^b the strata were defined such that $n_h \geq 5$ for either stratum, that is no more than 90% of the sampling effort could be allocated to one stratum. It was thought that a limit on N_h, the stratum sizes, would also need to be set but this was not necessary. Under $\pi p\mu^b$ if the maximum value of $\hat{\mu}_{[1]}$ was such that $n\frac{\hat{\mu}_{[1]}}{\sum_U \hat{\mu}_i} \geq 1$, the strategy of *srswor* was employed instead.

Table 4.1: N_h (n_h) for two strata under varying b using sampling strategy $strs\mu^b$ for μ generated from $\mathcal{M} \sim \Gamma(2,2)$

	b		
Stratum	-2	1	2
1	81(45)	558 (20)	690 (15)
2	919 (5)	442 (30)	310 (35)

The mean model–averaged design variance over all 100 realisations is shown for the three strategies under varying b in figure 4.2.

As the inclusion probabilities under *srswor* are not a function of μ_i^b the model–averaged design variance remains constant for all b. The lowest model–averaged design variance using *strs* μ^b occurs when $b = 1$, and so $\hat{\mu}_i = \mu_i$. Under $\pi p\mu^b$ the lowest model–averaged design variance appears to be at $b = 0.8$, although the model–averaged design variance for $b = 1$ is very similar. We would expect the model–averaged design variance to be lowest when $b = 1$ and this discrepancy between what is observed and what we expect is likely to be due to only a small number of simulations being carried out. Increasing the number of simulations should remove this apparent discrepancy. As b moves away from one, the model–averaged design variance increases. The rate of increase is more rapid under $\pi p\mu^b$ than under *strs* μ^b because the π_i under $\pi p\mu^b$ are directly related to $\hat{\mu}_i$ whereas under *strs* μ^b a small change in $\hat{\mu}_i$ will not affect the π_i.

Under *strs* μ^b table 4.1 shows how varying b affects both N_h and n_h. Let stratum 1 represent the stratum that contains the units with low values of μ_i and stratum 2 the units with higher values of μ_i. As b increases so $\hat{\mu}_i$ increases, and the proportion of the variability in the 442 units with the largest values of μ_i increases. Hence to account for the same proportion of the variability N_2 decreases, from 442 to 310. As this stratum is, according to $\hat{\mu}_i$, still very variable, n_2 will increase. Figure 4.3(a) shows that the π_i for units in stratum 2 are greater when $b = 2$ than when $b = 1$. In comparison, stratum

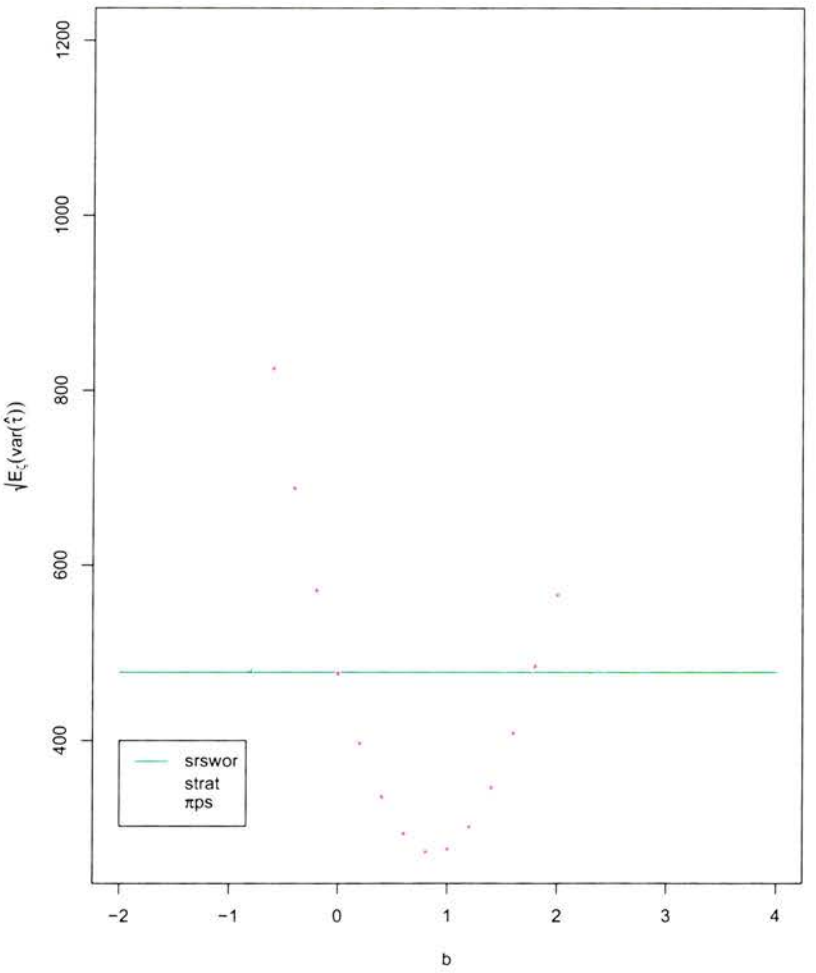


Figure 4.2: $\sqrt{E_\zeta[var(\hat{\tau})]}$ under *srswor*, *strs* μ^b and $\pi p \mu^b$ for varying b when μ are generated by the stochastic process $\mathcal{M} \sim \Gamma(2, 2)$

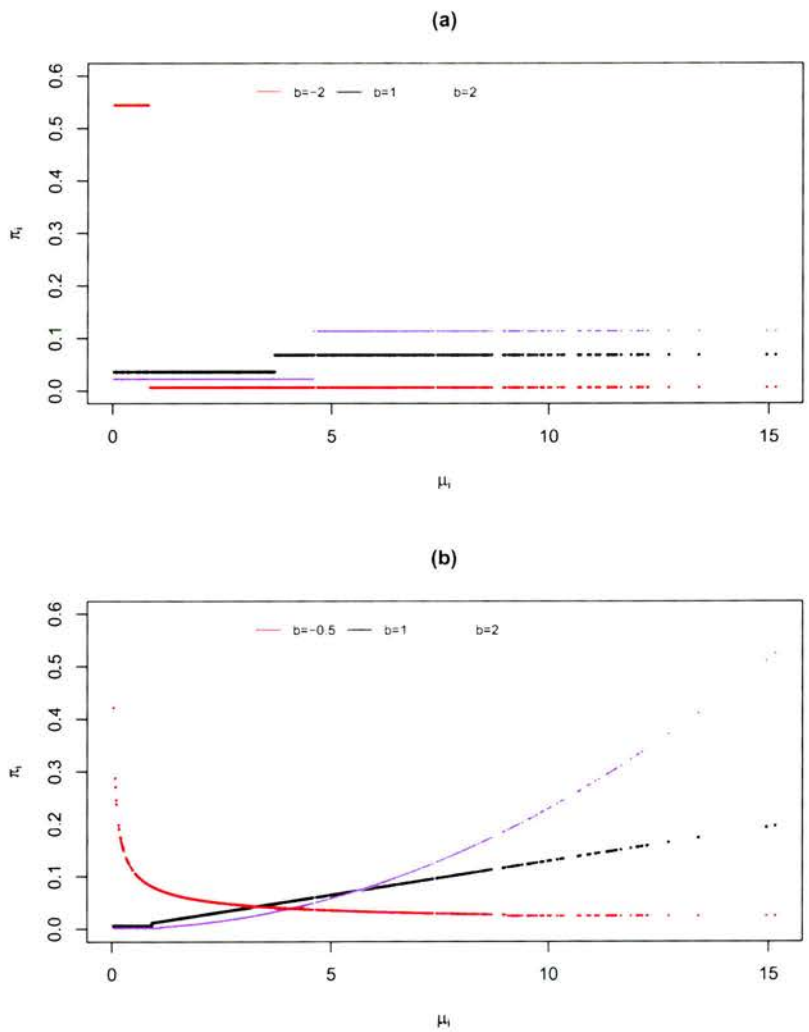


Figure 4.3: (a) π_i under *strs* μ^b for $b = -2, 1, 2$ (b) π_i under $\pi p \mu^b$ for $b = -0.5, 1, 2$ when μ are generated by the stochastic process $\mathcal{M} \sim \Gamma(2, 2)$ 4.2

1, which is large ($N_1 = 690$), has only a small proportion of the sample allocated to it ($n_1 = 15$) and so π_i for stratum 1 is smaller than when $b = 1$. The model–averaged design variance is large because this stratum has a higher proportion of the variability than $\hat{\mu}_i$ suggests and so is under-sampled.

As b decreases, then there is a "blip" when $b = 0$ as $\hat{\mu}_i = 1$ and so *srswor* is applied, i.e. there is only one stratum. If instead the sample was selected using *strs*, for some arbitrary division of units into two strata we would expect the curve to be smooth. As b decreases below zero, the model–averaged design variance becomes larger than that for *srswor* because units in stratum 1 are thought to have large values of μ. Hence a high proportion of the sample ($n_1 = 45$) is allocated to sampling a small stratum ($N_1 = 81$) which is not very variable, and a small proportion of the sample ($n_2 = 5$) is allocated to sampling a large stratum ($N_2 = 919$) which is highly variable.

Under $\pi p\mu^b$, as b increases, units with high values of μ_i have disproportionately high inclusion probabilities, see the purple line in figure 4.3(b). As b decreases, units which have low values of μ_i are given very high inclusion probabilities and units with high values of μ_i have low inclusion probabilities (the red line in figure 4.3(b))[1]. Therefore the model–averaged design variance is very high as there is a negative correlation between μ_i and π_i. For large $|b|$, there reaches a point where the value of $\frac{n\hat{\mu}_{[1]}}{\sum_U \hat{\mu}_{[i]}} \geq 1$, and in the simulation, the strategy returns to *srswor*. An alternative would have been to set the inclusion probability for $\pi_k = 1$ for all units where $\pi_{[k]} > 1$ and select the rest of the sample with $\pi p\mu^b$. We would then expect the model–averaged design variance to continue to increase as $|b|$ increases.

[1]We calculate the inclusion probabilities for $b = -0.5$ and not $b = -2$, as we did under *strs* . Some units have such high values of $\hat{\mu}^{-2}$ that their inclusion probability is one. These cases are not included.

4.2.2 Discussion

Clearly *srswor* is indifferent to model mis-specification, in that the model–averaged design variance does not change depending on the value of b. The simulations in section 4.4 and chapter 6, suggests that b might vary between 0 and 2[2]. In this range the model–averaged design variance under *strs* μ^b remained less than that under *srswor*. This was not always the case under $\pi p\mu^b$, although for $b = 1 \pm 0.5$ the strategy $\pi p\mu^b$ gave more precise estimates of τ than the strategy *strs* μ^b. Given that we would expect an improvement in the model–averaged design variance under *strs* μ^b with a greater number of strata and we would expect the effect of model mis-specification on the model–averaged design variance to be less under *strs* μ^b than under $\pi p\mu^b$, we might suppose that *strs* μ^b is a more desirable strategy than $\pi p\mu^b$.

In practice there are a couple of reasons why the *strs* μ^b strategy is not so desirable. Firstly it is more complex to implement, in that a greater number of decisions, about the number of strata, their definition and allocation of sampling effort, need to be determined. Also as b varies, the change in π_i for some units will be large, as they move from one stratum to another, whereas under $\pi p\mu^b$ the π_i change as a smooth function of b. Perhaps more importantly for a given $\hat{\mu}_i$, that is relatively continuous, there will be discontinuities in the π_i under *strs* μ^b due to stratum boundaries, which are to some extent arbitrary, compared to $\pi p\mu^b$ in which the π_i change smoothly over the survey region.

Although we have here only considered a strategy for survey 2, we wish to repeat the strategy in future surveys. If we use *strs* μ^b, a new set of strata would need to be defined for each survey. This does not give any degree of consistency between surveys, unless the strategy of Overton and Stehman (1996) as described in section 3.6 is employed. However this strategy is only recommended as an occasional refocussing of the whole monitoring programme, rather than a continuous process of adjustment.

[2]Note that in the simulations the auxiliary variables, density surface and the spatial realisation of individuals are generated on a continuous surface, but the modelling is carried out on the discretized values

Also if $\hat{\mu}_i$ is well specified, the model–averaged design variance under $\pi p\hat{\mu}^{(t)}$ is less than under *strs* μ^b. Ideally then we require a sampling design that combines the robust properties of *srswor* with the increase in precision of $\pi p\mu^b$.

4.3 A robust sampling strategy

The general principle of this strategy is that not all of the sample is selected using $\pi p\hat{\mu}$. Instead, a sample s_1 of n_1 units is selected from U using *srswor* and a sample s_2 of $n_2 = n - n_1$ units is selected from the remaining $N - n_1$ units by sampling with $\pi p\hat{\mu}$ so that $s = \bigcup\{s_1, s_2\}$. We call this a combined sampling strategy and use $comb\mu(\frac{n_2}{n})$ to denote a strategy in which a sample s_1 of $n_1 = n - n_2$ units are selected from U using *srswor*, and a sample s_2 of n_2 units are selected using $\pi p\mu$ from $U - s_1$. Using population P, described in chapter 2, figure 4.4(c) illustrates the strategy $comb\hat{\mu}^{(1)}(0.5)$ where the units coloured green are in s_1, and those coloured pink are in s_2. The variable $\hat{\mu}_i^{(1)}$ is shown in figure 4.4(a). The strategies *srswor*=$comb\hat{\mu}^{(1)}(0)$ and $\pi p\hat{\mu}^{(1)}$=$comb\hat{\mu}^{(1)}(1)$ are special cases of the strategy and are illustrated in figure 4.4(b) and figure 4.4(d) respectively.

If $\pi_{i|s_1}$ is the probability that unit i is included in s given the sample s_1 has been selected, then π_i can be calculated as

$$\begin{aligned}
\pi_i =& Pr(i \in s_1) + Pr(i \in s_2) \\
=& Pr(i \in s_1) + Pr(i \in s_2 | i \notin s_1) Pr(i \notin s_1) \\
=& \sum_{s_1 \ni i} p(s_1) + \sum_{s_1 \not\ni i} \pi_{i|s_1} p(s_1) \\
=& \sum_{s_1 \in S_1} \pi_{i|s_1} p(s_1) \qquad (4.8)
\end{aligned}$$

as $\pi_{i|s_1} = 1$ when $i \in s_1$. As s_1 is selected using *srswor*, $p(s_1) = \frac{1}{\binom{N}{n_1}}$ and so taking the third line of the expansion above

$$\pi_i = \frac{\binom{N-1}{n_1-1}}{\binom{N}{n_1}} + \frac{1}{\binom{N}{n_1}} \sum_{s_1 \not\ni i} \pi_{i|s_1} = \frac{n_1}{N} + \frac{1}{\binom{N}{n_1}} \sum_{s_1 \not\ni i} \pi_{i|s_1} \qquad (4.9)$$

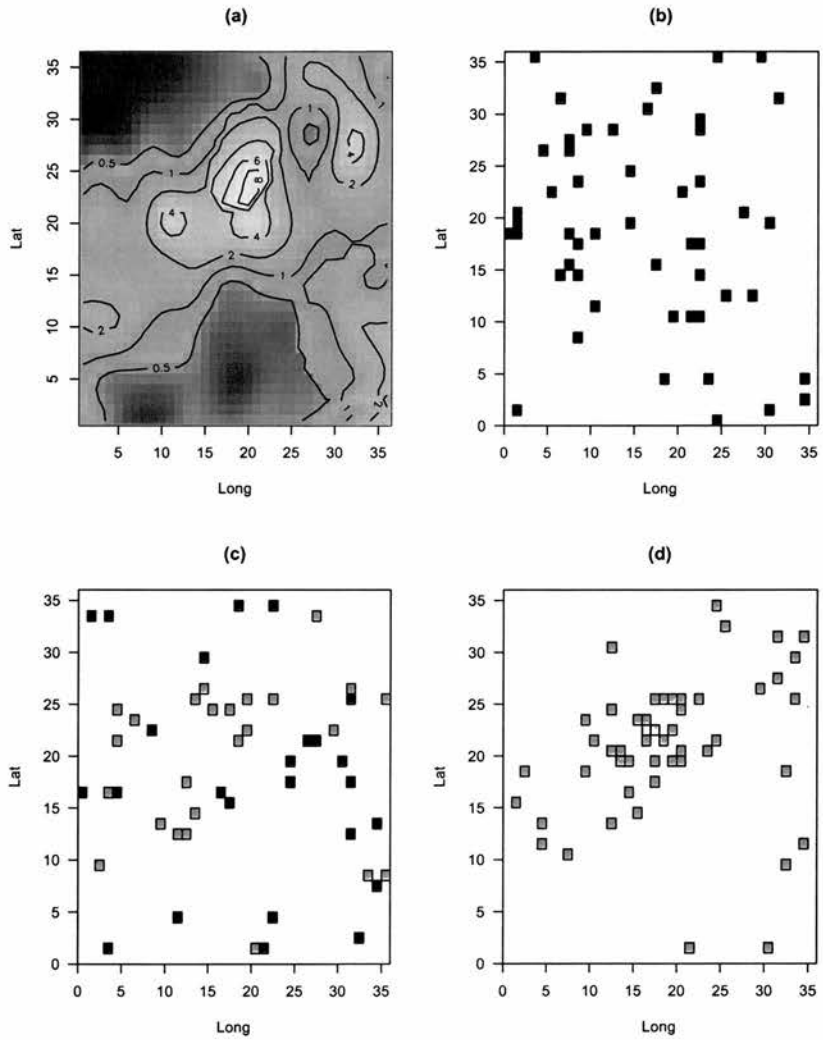


Figure 4.4: (a) $\hat{\mu}_i^{(1)}$ estimated from $\zeta^{(1)}$ constructed using data from a sample, s, of $n = 50$ units selected using *srswor* (b) Sample of $n = 50$ units selected using *srswor* (c) Sample of $n = 50$ units of which $n = 25$ selected using *srswor*, shown in green, and $n = 25$ units selected using $\pi p\hat{\mu}^{(1)}$, shown in pink (d) Sample of $n = 50$ units selected using $\pi p\hat{\mu}^{(1)}$

Second order inclusion probabilities are calculated in a similar manner so that

$$\begin{aligned}\pi_{ij} &= \sum_{s_1} \pi_{ij|s_1} p(s_1) \\ &= \sum_{\substack{s_1 \ni i \\ s_1 \ni j}} p(s_1) + \sum_{\substack{s_1 \not\ni i \\ s_1 \ni j}} \pi_{i|s_1} p(s_1) + \sum_{\substack{s_1 \ni i \\ s_1 \not\ni j}} \pi_{j|s} p(s_1) + \sum_{\substack{s_1 \not\ni i \\ s_1 \not\ni j}} \pi_{ij|s} p(s_1)\end{aligned} \tag{4.10}$$

To calculate these inclusion probabilities, all possible s_1 samples must be enumerated. For example, from equation 4.9 we see that to estimate π_i we need to know $\pi_{i|s_1}$ for all s_1 that do not include the unit i. Except when $\frac{n_1}{N}$ is very small, or very large, the number of possible samples is too large to enable a complete enumeration.

We can approximate the π_i and π_{ij} using simulation. Take a sample s_{1_r} of size n_1 using *srswor*. For this particular s_{1_r} it is possible to calculate $\pi_{i|s_{1_r}}$ for all $i \in U$ so that except for small μ_i

$$\begin{aligned}\pi_{i|s_{1_r}} &= \frac{n_1}{N} && \text{for } i \in s_{1_r} \\ \pi_{i|s_{1_r}} &= \frac{n_2 \mu_i}{\sum_{j \notin s_{1_r}} \mu_j} && \text{for } i \notin s_{1_r}\end{aligned}$$

This process is repeated for $r = 1, \ldots, R$ so that

$$\pi_i \doteq \frac{1}{R} \sum_{r=1}^{R} \pi_{i|s_{1r}}$$

as $p(s_{1_r}) = \frac{n_1}{N}$ for all r. A similar process is carried out to estimate π_{ij}.

Figure 4.5(a) shows estimated inclusion probabilities, where $R = 1000$ for the population described in A using the strategy $comb\mu(0.5)$. This is a computationally intensive strategy. We would expect units with similar μ_i to have very similar π_i but there is quite a lot of variability in these estimated inclusion probabilities. Although we have not managed to satisfactorily obtain an approximation for π_i using simulation, automated survey design routines, such as the one in DISTANCE (Thomas *et al.*, 2003) have more efficient algorithms for approximating inclusion probabilities using simulation under different survey

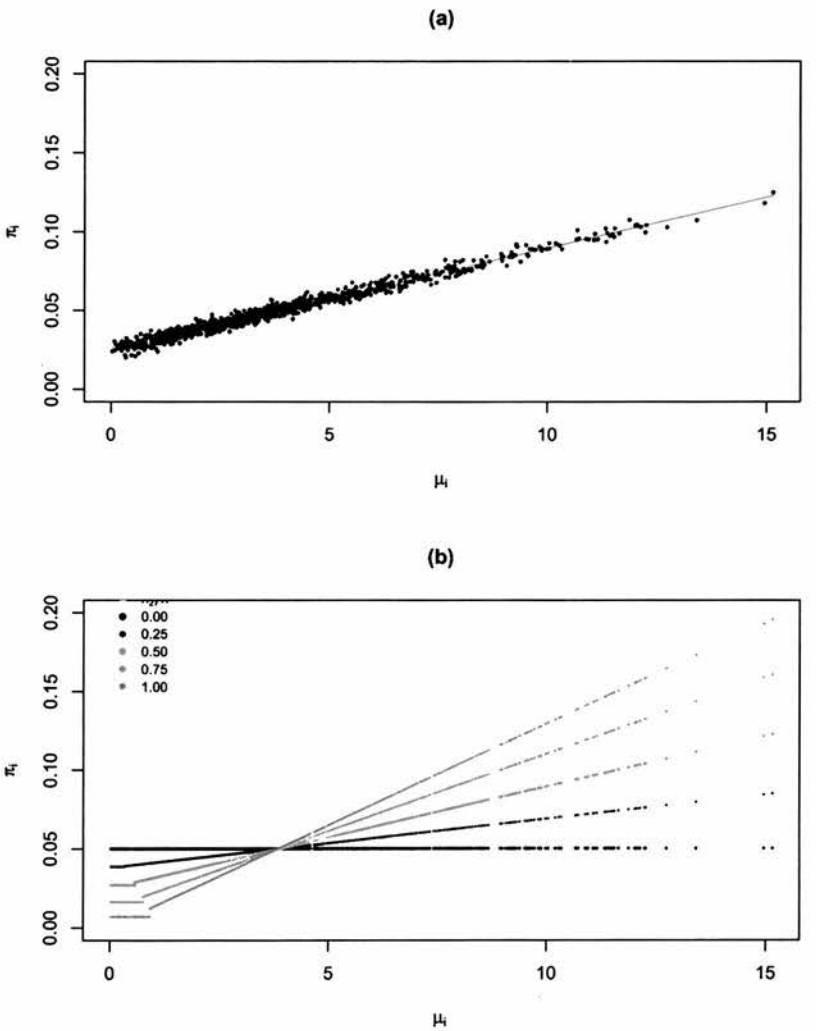


Figure 4.5: Inclusion probabilities (a) Points are results from a simulation of 1000 runs where $n_1 = n_2 = 25$. Line is the approximate inclusion probabilities using equation 4.11(b) Approximate inclusion probabilities using equation 4.11 for varying n_2 when $n = 50$

designs. If our combined sampling strategies were incorporated into such a package the unconditional inclusion probabilities could be approximated

Alternatively we can attempt to find an analytic approximation for the π_i and π_{ij}. One possibility is to calculate the minimum and maximum values of $\pi_{i|s_1}$ for unit i when $i \notin s_1$. For example let units be ordered from largest to smallest so that unit [1] is such that $\mu_{[1]} = max(\underline{\mu}_U)$. The inclusion probability $\pi_{[1]|s_1}$ is maximised when $s_1 = \{[2], \ldots, [n_1 + 1]\}$, that is s_1 contains the units with the largest values of $\underline{\mu}_U$, not including unit [1]. $\pi_{[1]|s_1}$ is minimised when $s_1 = \{[N - n_1 + 1], \ldots, [N]\}$, that is s_1 contains the units with the smallest values of $\underline{\mu}_U$. We could then take an average of these two values, $\bar{\pi}_{[i]|s_1}$, so that

$$\pi_{[1]} \doteq \frac{n_1}{N} + \frac{N - n_1}{N}\bar{\pi}_{[1]|s_1}$$

A difficulty is defining the samples that give the maximum and minimum value of $\pi_{[k]|s_1}$ for units where $[i]$ is close to k^*, the first unit under Sunter's strategy that is selected with *srswor* rather than $\pi p\mu$. Depending on the s_1 selected, unit i may be the k^{th} largest where $k < k^*$ so that $\pi_i = \frac{n_2\bar{\mu}_{k^*}}{\sum_{j \notin s_1} \mu_j}$ or may be the k'^{th} largest where $k' > k^*$ so that $\pi_i = \frac{n_2\mu_{[i]}}{\sum_{j \notin s_1} \mu_j}$.

A simpler approximation which we use was suggested by inspection of the simulated results for varying μ, n_1, n_2 and N. When n_1/N is small, an approximation for π_i can be obtained by assuming that s_2 is also taken from U rather than $U - s_1$ so that

$$\pi_i \doteq \frac{n_1}{N} + \pi_{iU} \tag{4.11}$$

$$\pi_{ij} \doteq \frac{n_1(n_1 - 1)}{N(N - 1)} + \frac{n_1}{N}(\pi_{iU} + \pi_{jU}) + \pi_{ijU} \tag{4.12}$$

where π_{iU} is the probability that unit i is included in the final sample when n_2 units are sampled from U using $\pi p\mu$.

Figure 4.5(b) shows the inclusion probabilities using the approximation 4.11 above when s is selected using $comb\mu(\frac{n_2}{n})$ with varying n_2. As $\frac{n_2}{n}$ increase, so the inclusion probabilities of units with high μ_i increase, whereas for units with low μ_i, π_i decreases. The approximation is reasonable when $\frac{n_1}{N}$ is small, and is convenient for calculating the model–averaged design variance when $\hat{\mu}_i = a\mu_i^b$.

If the unconditional inclusion probabilites π_i and π_{ij} are estimated, the standard design-based estimator of τ, the Horvitz-Thompson estimator, equation 3.8 and its variance, equation 3.13, can be used.

Given the sampling strategy described, an exact estimator of τ can be obtained by employing the principles of two-phase sampling. The motivations for these methods and details of how inclusion probabilities are calculated for general unequal probability sampling designs through time are given in Chapter 5. The basic idea is sufficient here. The key is that two separate estimators of τ are obtained, $\hat{\tau}_1, \hat{\tau}_2$, that use the data from s_1 and s_2 respectively. Similarly two variances $var(\hat{\tau}_1), var(\hat{\tau}_2)$ and the $cov(\hat{\tau}_1, \hat{\tau}_2)$ are calculated. The final estimator of τ is a weighted average of $\hat{\tau}_1$ and $\hat{\tau}_2$ where the optimal weighting depends on the relative size of the variances. Although this gives exact results, we do not use this method in this chapter because these estimators are not tractable when we wish to explore how these robust sampling strategies perform under model mis-specification. In addition the linear approximations of the inclusion probabilities, from using equations 4.11 and ?? are intuitively nicer because it allows unconditional inclusion probabilities to be calculated.

4.3.1 Comparison of strategies

Using the strategy $comb\mu^b(\frac{n_2}{n})$ for $\frac{n_2}{n} = 0.25, 0.5, 0.75$ we repeat the method of section 4.2.1 for $b = -2, -1.8, \ldots, 4$. The limiting cases when $\frac{n_2}{n} = 0$, *srswor*, and $\frac{n_2}{n} = 1$, $\pi p\mu^b$, were carried out in section 4.2.1.

Figure 4.6 illustrates the results from this simulation, by plotting the mean $E_\zeta[var(\hat{\tau})]$ over the 100 simulations against b for each strategy. When $b = 1$ the model–averaged design variance increases as $\frac{n_2}{n}$ decreases from $\frac{n_2}{n} = 1$, $\pi p\mu^b$, to $\frac{n_2}{n} = 0$, *srswor*. Under $\pi p\mu^b$ the model–averaged design variance is minimised when $b = 1$. As $\frac{n_2}{n}$ decreases the minimum model–averaged design variance increases, and the value of b at which the minimum occurs also increases. When $0 < \frac{n_2}{n} < 1$ the model–averaged design variance

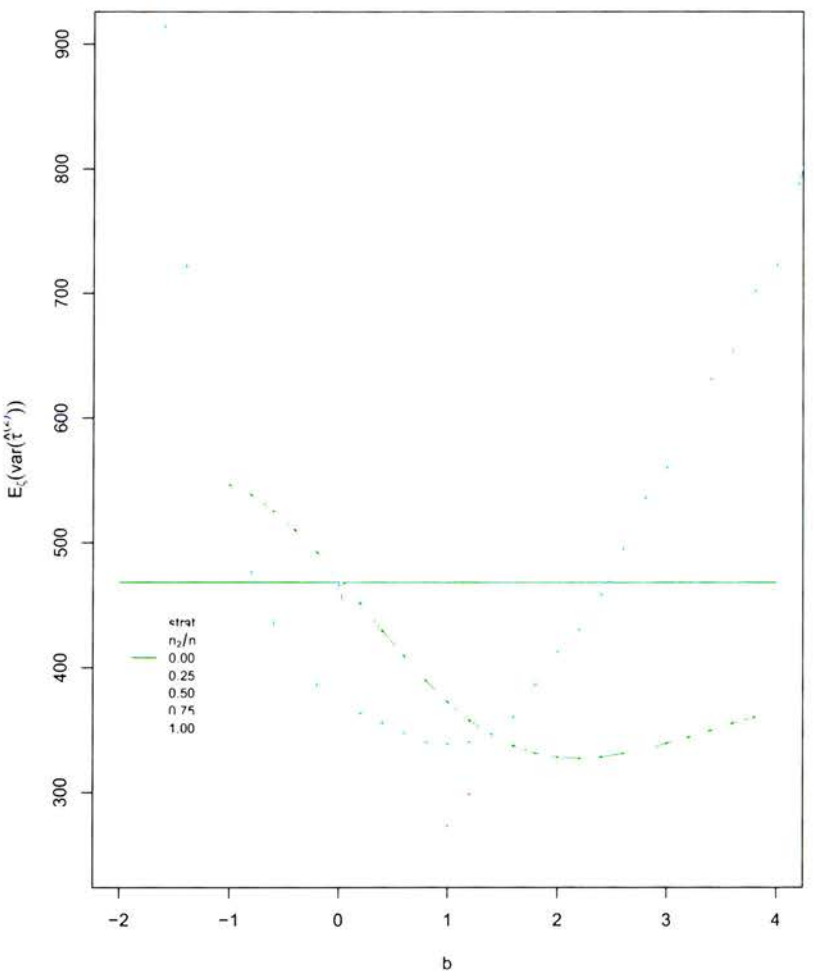


Figure 4.6: Mean $\sqrt{E_\zeta[var(\hat{\tau})]}$ for sampling strategy $comb\mu^b \frac{n_2}{n}$ with varying $\frac{n_2}{n}$ and b over 100 simulations. The μ are generated by the stochastic process $\mathcal{M} \sim \Gamma(2, 2)$

is always less under $comb\mu^b(\frac{n_2}{n})$ than the model–averaged design variance under *srswor* when $0 < b < 2$. When b is negative the $comb\mu^b(\frac{n_2}{n})$ strategies perform worse than *srswor* but perform better than $\pi p\mu^b$.

The combined sampling strategies, $comb\mu^b(\frac{n_2}{n})$ when $0 < \frac{n_2}{n} < 1$, are therefore more robust to model mis-specification than $\pi p\mu^b$ and they reduce the model–averaged design variance compared to *srswor* under the most likely conditions of $0 < b < 2$. We see in particular that the model–averaged design variance under the strategy $comb\mu^b(0.75)$ is nearly as good as that under $\pi p\mu^b$=$comb\mu^b(1)$ when $0 < b < 1$ and is considerably better when $b > 1$.

The model–averaged design variance under the combined strategies is lower than under *strs* μ^b when $b > 1$. For $b < 0$ the model–averaged design variance under *strs* μ^b is always better than under the combined strategies although in practice we would not expect $b < 0$ to occur in practice except under unusual circumstances.

4.4 Simulation

Using population A, for which there are data on three auxiliary variables, we test out the combined sampling strategies as they would be used in practice. In survey 1 a sample $s^{(1)}$ of $n = 50$ units is selected using *srswor*. Using the observed data $\underline{y}^{(1)}_{s^{(1)}}$ an estimate of $\tau^{(1)}$ is obtained using the Horvitz-Thompson estimator. In addition these data are used to estimate $\mu_i^{(1)}$. We fit a simple GLM that assumes a QSM of the form $Y_i^{(1)} \sim Po(\mu_i^{(1)})$ such that

$$log(\mu_i^{(1)}) = \sum_{j=0}^{3} \beta_j x_{ij} \tag{4.13}$$

After fitting the maximal model, a stepwise AIC procedure is used to choose the final model. This is used to calculate the predicted abundance, $\hat{\mu}_i^{(1)}$, for all $i \in U$.

In survey 2, a new sample $s^{(2)}$ of $n = 50$ units is taken using the strategy $comb\hat{\mu}_i^{(1)}(\frac{n_2}{n})$. For each $s^{(1)}$ and $\hat{\mu}_i^{(1)}$, a sample is taken for survey 2 using each of five combined sampling

strategies defined by the proportion $\frac{n_2}{n} = 0, 0.25, 0.5, 0.75, 1$ so that $n_1 = 50, 38, 25, 12, 0$ respectively and a sixth sample is taken using the *strs* $\hat{\mu}_i^{(1)}$ strategy described in section 4.1.1. For the combined sampling strategies the inclusion probabilities are calculated using equation 4.11 and 4.12. The Horvitz-Thompson estimator, equation 3.8, is used to calculate $\hat{\tau}^{(2)}$, and $var(\hat{\tau}^{(2)})$ and $\widehat{var}(\hat{\tau}^{(2)})$ are calculated using the Sen-Yates-Grundy equations 3.12 and 3.13. In addition $E_\zeta[var(\hat{\tau})]$ is calculated. The mean of $\hat{\tau}^{(2)}$, the standard error, $\sqrt{var(\hat{\tau}^{(2)})}$, an estimate of the standard error, $\sqrt{\widehat{var}(\hat{\tau}^{(2)})}$ and the model-average design variance $\sqrt{E_\zeta[var(\hat{\tau})]}$ over the 1000 simulations are given in table 4.2.

The simulations incorporate two sources of variability. The first source is from varying $s^{(2)}$ between simulations and the second source is due to the sample $s^{(1)}$ varying between simulations. Suppose instead that $s^{(1)}$ remains fixed for the 1000 simulations. From $s^{(1)}$ an estimate of $\mu_i^{(1)}$ can be obtained and these $\hat{\mu}_i^{(1)}$ would be used to select all 1000 samples of $s^{(2)}$. As $\hat{\mu}_i^{(1)}$ and so also the $\pi_i^{(2)}$ are fixed for all 1000 simulations we could calculate $\sqrt{var(\hat{\tau}^{(2)})}$, the standard error of $\hat{\tau}^{(2)}$, because in a simulation we know $y_i^{(2)}$ for all $i \in U$. From the 1000 simulations we would obtain a distribution of $\hat{\tau}^{(2)}$ values. The mean of these values should be close to $\tau^{(2)}$ if our estimator is unbiased. The standard deviation of the 1000 estimates of $\hat{\tau}^{(2)}$ is an empirical estimate of the standard error and should be close to the standard error, $\sqrt{var(\hat{\tau}^{(2)})}$. Similarly the mean of the analytic estimate of the standard error, $\sqrt{\widehat{var}(\hat{\tau}^{(2)})}$, obtained from the 1000 samples $s^{(2)}$ should be close to the standard error as we use an unbiased estimator. We can also calculate an empirical estimate of the standard deviation of $\sqrt{\widehat{var}(\hat{\tau}^{(2)})}$. This is the standard deviation of the 1000 values of $\sqrt{\widehat{var}(\hat{\tau}^{(2)})}$.

If $s^{(1)}$ also varies between simulations then $\hat{\mu}_i^{(1)}$ will vary between simulations. Then the inclusion probabilities will vary from one simulation to another and so $var(\hat{\tau}^{(2)})$ will vary between simulations. In table 4.2 the standard deviation of $\sqrt{var(\hat{\tau}^{(2)})}$ describes the variability in the $var(\hat{\tau}^{(2)})$ from varying $\hat{\mu}_i^{(1)}$. Under *srswor*, $\frac{n_2}{n} = 0$ the $\sqrt{var(\hat{\tau}^{(2)})}$ does not vary between simulations as the inclusion probabilities are not a function of $\hat{\mu}_i^{(1)}$. The $\sqrt{var(\hat{\tau}^{(2)})}$ varies most under the sample design $\pi p\hat{\mu}^{(1)}$, $\frac{n_2}{n} = 1$, as the inclusion

probabilities are a function of the $\hat{\mu}_i^{(1)}$. As $\frac{n_2}{n}$ decreases so the variability of $\sqrt{var(\hat{\tau}^{(2)})}$ reduces as the $\hat{\mu}_i^{(1)}$ only partly determine the $\pi_i^{(2)}$. A similar argument applies for the $E_\zeta[var(\hat{\tau})]$.

In table 4.2 the empirical estimate of the standard error of $\hat{\tau}^{(2)}$, the standard deviation of the 1000 $\hat{\tau}^{(2)}$ values, incorporates both sources of variability. We would expect the mean of the $\hat{\tau}^{(2)}$ to be an unbiased estimate of $\tau^{(2)}$ and the empirical estimate of the standard error of $\hat{\tau}^{(2)}$ to be a good estimate of the mean value of $\sqrt{var(\hat{\tau}^{(2)})}$. We would expect the mean of the $\sqrt{\widehat{var}(\hat{\tau}^{(2)})}$ to be an unbiased estimate of $\sqrt{var(\hat{\tau}^{(2)})}$ and the empirical estimate of the standard error of $\sqrt{\widehat{var}(\hat{\tau}^{(2)})}$ to be greater than the standard deviation of the $\sqrt{var(\hat{\tau}^{(2)})}$ over the 1000 simulations as it incorporates both sources of variability.

Under each combined sampling strategy the $\hat{\tau}^{(2)}$ appear to be normally distributed. Asymptotic 95% confidence intervals suggest that the strategies give unbiased estimates of $\tau^{(2)}$ as we would expect using the Horvitz-Thompson estimator. In a similar manner $\sqrt{\widehat{var}(\hat{\tau}^{(2)})}$ is an unbiased estimate of $\sqrt{var(\hat{\tau}^{(2)})}$. The strategies that give the highest precision are those for which 50% or 75% of the sample are selected using $\pi p\hat{\mu}_i^{(1)}$. Stratification does better than *srswor* and better than $comb\hat{\mu}^{(1)}(0.25)$.

If we compare $var(\hat{\tau}^{(2)})$ under each strategy with $var(\hat{\tau}^{(2)})$ under *srswor* we obtain a measure of efficiency. By dividing the mean variance under each strategy by the mean variance from sampling using *srswor* we obtain efficiencies of 0.86, 0.79, 0.77 and 0.78 for $\frac{n_2}{n} = 0.26, 0.50, 0.74$ and 1.00 respectively. Therefore using a combined sampling strategy in which at least 50% of units were selected using $\pi p\hat{\mu}^{(1)}$ increased the precision of the estimates by over 20%.

This increase in precision is not reflected in a noticeable decrease in c.v. For example when $\frac{n_2}{n} = 0.74$ the mean c.v. is 9.35%, compared to selecting the sample using *srswor*, the mean c.v. is 10.6%. The reason for this is that the simulated data were not very heterogeneous, the variance of the data is 2.38. We would expect a greater improvement in the c.v. when the $y_i^{(2)}$ are more variable. Therefore further work using a more heterogeneous population

Table 4.2: Mean (s.d.) for various parameters and estimators obtained from 1000 simulations using the sampling strategies $comb\hat{\mu}_i^{(1)}(\frac{n_2}{n})$ for varying $\frac{n_2}{n}$ or $strs\hat{\mu}_i^{(1)}$ where $n = 50$. $\tau^{(2)} = 2546$. The final row are for a model-assisted strategy where $s^{(2)}$ is selected using *srswor*.

$\frac{n_2}{n}$	$\hat{\tau}^{(2)}$	$\sqrt{var(\hat{\tau}^{(2)})}$	$\sqrt{\widehat{var}(\hat{\tau}^{(2)})}$	$\sqrt{E_\zeta[var(\hat{\tau})]}$
0.00	2540 (274)	270 (0)	269 (36)	267 (0)
0.25	2543 (247)	250 (9)	248 (29)	248 (9)
0.50	2537 (243)	238 (13)	237 (29)	237 (12)
0.75	2562 (226)	235 (14)	234 (29)	235 (13)
1.00	2547 (240)	242 (17)	238 (37)	242 (17)
Stratification	2549 (235)	243 (9)	241 (29)	243 (9)
Model-assisted	2553 (225)	228 (12)	207 (24)	–

could be useful.

As the data are simulated the $\mu_i^{(1)}$ are known. For each of the 1000 surveys at time $t = 1$, the value b can be calculated where $\hat{\mu}_i^{(1)} = a\mu_i^{(1)b}$. Figure 4.7 plots $var(\hat{\tau}^{(2)})$ against b for each of the combined sampling strategies, except *srswor*, and for *strs* $\hat{\mu}_i^{(1)}$. Note that b is contained within the range $(0, 2)$ and that the number of auxiliary variables included in the final model is shown in colour. Looking first at the results for the strategy $\pi p\hat{\mu}_i^{(1)}$, we see that as the number of variables in the model decreases so the value of b tends towards zero. When the correct terms are in the model, the effect of b is equivalent to that shown in figure 4.6 for $\pi p\mu$. As the number of terms decreases the curve shifts to the left and the variance increases. This pattern is repeated in the combined sampling strategies. The combined strategies are relatively robust to model mis-specification compared to $\pi p\hat{\mu}^{(1)}$ and almost all cases give more precise estimates of $\hat{\tau}^{(t)}$ than that obtained under *srswor*. The strategy of *strs* $\hat{\mu}_i^{(1)}$ is robust to model mis-specification in that $\sqrt{var(\hat{\tau}^{(2)})}$ is less than under *srswor*.

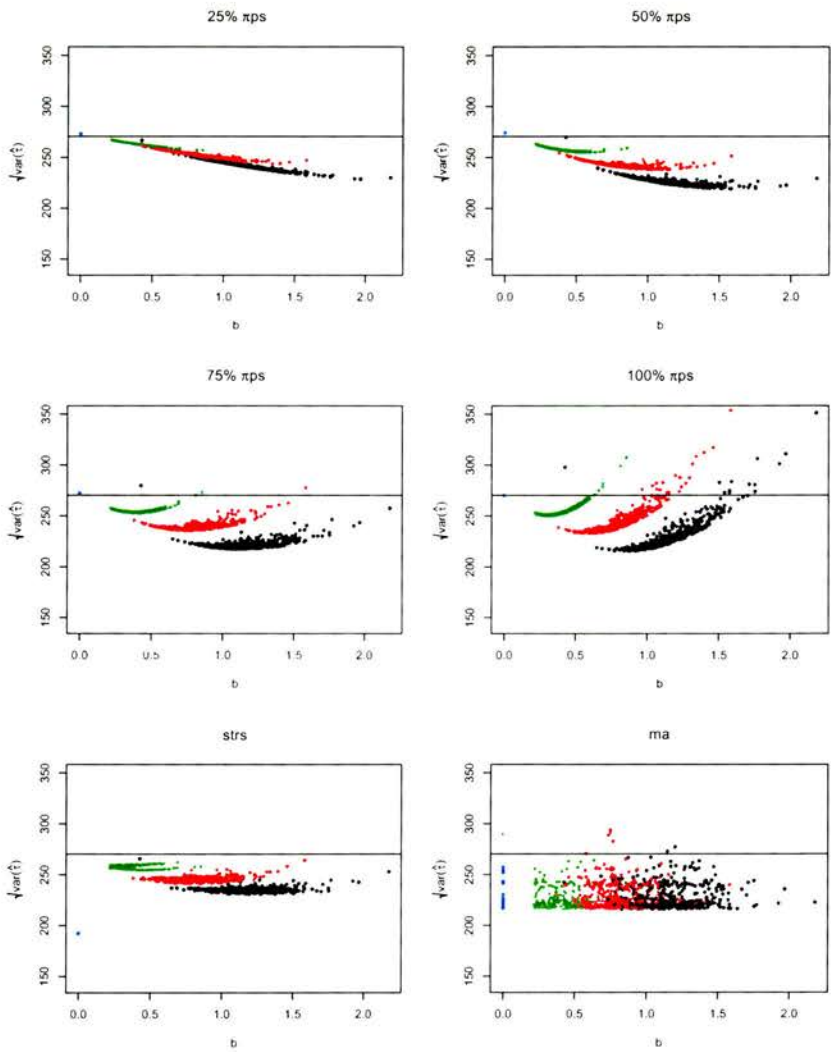


Figure 4.7: $\sqrt{var}(\hat{\tau}^{(2)})$ plotted against b for various sampling strategies for population A. b is estimated from the model ζ constructed using the data from $s^{(1)}$ selected using *srswor*. The horizontal line is $\sqrt{var(\hat{\tau}^{(2)})}$ under *srswor*. Colours represent the number of auxiliary variables in the final model where black=3, red=2, green=1, blue=0. The correct model has 3 auxiliary variables. "ma" are the results from a model-assisted strategy in which s is selected using *srswor* and $\tau^{(2)}$ is estimated using a model-assisted estimator.

4.5 Model–assisted survey sampling

A model-assisted sampling strategy is sometimes used as an alternative to a design-based strategy, such as *strs*, to increase the precision of $\hat{\tau}^{(2)}$. Hence it seems appropriate to compare our combined sampling strategies with a model-assisted strategy.

The model-assisted strategy we will employ consists of selecting $s^{(2)}$ using *srswor*, and uses a model-assisted estimator to estimate $\hat{\tau}^{(2)}$. The equations for $\hat{\tau}^{(2)}$, its variance, $var(\hat{\tau}^{(2)})$, and its estimate, $\widehat{var}(\hat{\tau}^{(2)})$, are of the same form as equations 3.17, 3.19 and 3.20 so that

$$\hat{\tau}^{(2)} = \sum_s \left(\frac{y_i^{(2)} - \hat{y}_i^{(2)}}{\pi_i} \right) + \sum_U \hat{y}_i^{(2)} \tag{4.14}$$

$$var(\hat{\tau}) = \frac{1}{2} \sum_U \sum_U (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i^{(2)} - \hat{y}_i^{(2)}}{\pi_i} - \frac{y_j^{(2)} - \hat{y_j}^{(2)}}{\pi_j} \right)^2 \tag{4.15}$$

$$\widehat{var}(\hat{\tau}) = \frac{1}{2} \sum_s \sum_s \left(\frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \right) \left(\frac{y_i^{(2)} - \hat{y}_i^{(2)}}{\pi_i} - \frac{y_j^{(2)} - \hat{y_j}^{(2)}}{\pi_j} \right)^2 \tag{4.16}$$

In general the model that describes the relationship between the auxiliary variables and the $y_i^{(2)}$, so that $\hat{y}_i^{(2)}$ can be calculated, does not need to be the QSM. As discussed in section 2.3.1, it is sometimes inappropriate to use the QSM, because the model-assisted strategy wishes to estimate $\hat{y}_i^{(2)}$, whereas using the QSM it is $\hat{\mu}_i^{(2)}$ that is estimated. Unless $cov(Y_i^{(2)}, Y_j^{(2)}) = 0$, $\hat{y}_i^{(2)} \neq \hat{\mu}_i^{(2)}$. In our case we assume a simple population model in which this covariance is zero, and so we can use the QSM to estimate $\hat{y}_i^{(2)}$. We could use the generalised nonparametric regression method of Breidt and Opsomer (2000) that is applied in Opsomer *et al.* (2003), although we use a parametric regression.

The procedure for estimating $\hat{y}_i^{(2)}$ is as follows. Using the data from $s^{(2)}$ and working within a GLM framework, see equation 4.13, the appropriate terms for the linear predictor are selected using, for example, a stepwise AIC procedure. Given the final model a design-consistent estimate of $\hat{\mu}_i^{(2)}$ is obtained by fitting a weighted version of the model where the weights correspond to the inclusion probabilities $\pi_i^{(2)}$. Using this weighted model, we estimate $\hat{\mu}_i^{(2)}$ for all $i \in U$. Note that when $s^{(2)}$ is selected using *srswor*, the parameters

of the weighted model are the same as that for the unweighted model.

The model is then calibrated by estimating the parameters a_1 and a_2 where

$$y_i^{(2)} = a_1 + a_2\hat{\mu}_i^{(2)} + \epsilon_i^{(2)} \qquad \text{for } i \in s^{(2)}$$

so that

$$\hat{y}_i^{(2)} = \hat{\mu}_i^{(2)*} = \hat{a}_1 + \hat{a}_2\hat{\mu}_i^{(2)} \qquad \text{for } i \in U$$

This calibration is similar to that proposed by Wu and Sitter (2001) and is stated by Opsomer *et al.* (2003) to remove the effect of bias and to allow the estimator to be a linear combination of the observed $y_i^{(2)}$s. The $\hat{y}_i^{(2)}$ are then used in equations 4.14, 4.15 and 4.16

Results from implementing the model-assisted strategy in our simulation study are given in the last line of table 4.2. They suggest that the model-assisted strategy gives a more precise estimate of $\hat{\tau}^{(2)}$ than under any of the other strategies described in this chapter. In addition, figure 4.7(f) shows how $\sqrt{var(\hat{\tau}^{(2)})}$ varies with b, where in this case b is estimated from survey 2 so that $\hat{\mu}_i^{(2)} = a\mu_i^{(2)b}$. We see that the strategy appears robust to model mis-specification.

This would suggest that applying the model-assisted estimator to the combined sampling strategies may also lead to an improvement in $var(\hat{\tau}^{(2)})$. In table 4.3 results from applying the model-assisted estimators to the combined sampling strategies are given. There is a large discrepancy between the standard error of $\hat{\tau}^{(2)}$, $var(\hat{\tau}^{(2)})$, and the empirical estimate of the standard error, the standard deviation of the 1000 estimates of $\tau^{(2)}$. In Särndal *et al.* (1992, pp 234–238) the estimated variance of a model-assisted estimator incorporates weighted values of the auxiliary variables, where the weights are related to inclusion probabilities. Under *srswor* these weights disappear but under $\pi p\mu$ sampling they may significantly change the estimator. These weights are not incorporated in the estimator described by Opsomer *et al.* (2003) and could be the cause of the discrepancy between the empirical estimates of the standard error and the analytic standard errors as this discrepancy is greatest when a proportion of the sample is selected using $\pi p\hat{\mu}^{(1)}$. Fur-

Table 4.3: Mean (s.d.) for various parameters and estimators obtained from 1000 simulations using the sample design $comb\hat{\mu}_i^{(1)}(\frac{n_2}{n})$ and a model-assisted strategy to estimate $\tau^{(2)} = 2546$ when $n = 50$.

$\frac{n_2}{n}$	$\hat{\tau}^{(2)}$	$\sqrt{var(\hat{\tau}^{(2)})}$	$\sqrt{\widehat{var}(\hat{\tau}^{(2)})}$
0.00	2553 (225)	228 (12)	207 (24)
0.25	2566 (274)	231 (22)	216 (33)
0.50	2582 (294)	229 (19)	215 (29)
0.75	2581 (297)	235 (24)	221 (33)
1.00	2642 (350)	241 (29)	228 (40)
Stratification	2541 (252)	226 (16)	214 (29)

ther work is required to apply these type of weights from the simple regression estimator to the model-assisted variance estimator proposed here.

Comparing the empirical estimates of the standard error of $\hat{\tau}^{(2)}$ for the different sampling designs we see that the model-assisted estimator does not increase the precision of $\hat{\tau}^{(2)}$ when a proportion of the sample is selected using $\pi p\hat{\mu}^{(1)}$. This is because the $e_i^{(2)} = y_i^{(2)} - \hat{y}_i^{(2)}$ are not correlated with π_i. When we use a design-based estimator the increase in the precision of $\hat{\tau}^{(2)}$ occurs because $y_i^{(2)}$ is correlated with $\pi_i^{(2)}$. When $s^{(2)}$ is selected using *srswor* the model-assisted estimator is more efficient that the Horvitz-Thompson design-based estimator as $\frac{e_i^{(2)}}{\pi_i^{(2)}} = e_i^{(2)}\frac{N}{n}$ are less than $y_i^{(2)}\frac{N}{n}$. This is not the case for unequal $\pi_i^{(2)}$ in the combined sampling strategies.

4.6 Discussion

This chapter has explored how we can use $\hat{\mu}_i^{(1)}$ to determine the sampling strategy in survey 2, where the aim of the survey is to provide an efficient estimate of $\tau^{(2)}$. Simple strategies such as *strs* $\hat{\mu}^{(1)}$ or $\pi p\hat{\mu}^{(1)}$ improves the efficiency of $\hat{\tau}^{(2)}$ compared to using

srswor. One difficulty is that if $\hat{\mu}_i^{(1)}$ is a poor estimate of $\mu_i^{(2)}$ then the increase in precision is reduced, and in fact $var(\hat{\tau}^{(2)})$ using either strategy may be greater than under *srswor*. This is particularly the case under $\pi p\hat{\mu}_i^{(1)}$. This was illustrated by deriving a measure b where $\hat{\mu}_i^{(1)} = \mu_i^{(2)b}$ that represents how good an estimate $\hat{\mu}_i^{(1)}$ is of $\mu_i^{(2)}$. As b moves away from one the model is less well specified.

We have proposed a combined sampling strategy, in which a proportion of the sample n_1 is selected using *srswor*, and the rest n_2 is selected using $\pi p\hat{\mu}_i^{(1)}$. This is more robust to model mis-specification than a fully $\pi p\hat{\mu}_i^{(1)}$ strategy, and the precision of $\hat{\tau}^{(2)}$ is greater than under *srswor* or *strs* $\hat{\mu}_i^{(1)}$ when $b > 0$.

Compared to *strs* $\hat{\mu}^{(1)}$ in which H, n_h and the definition of U_h need to be decided, the combined sampling strategies only require a decision on the proportion of the sample that is selected using *srswor*. We look in more detail at how the proportion $\frac{n_2}{n}$ should be chosen in Chapter 6. Conceptually the *srswor* portion of the sample, $s_1^{(2)}$, is, on average, equivalent to stratified sampling using the proportional allocation rule. The $\pi p\hat{\mu}_i^{(1)}$ portion of the sample, $s_2^{(2)}$, then acts by disproportionately sampling the high abundance stratum. Rather than the disjoint inclusion probabilities that arise from *strs* $\hat{\mu}^{(1)}$, the $comb\hat{\mu}^{(1)}(\frac{n_2}{n})$ strategy allows a smooth transition in the inclusion probabilities, with changing $\hat{\mu}_i^{(1)}$ and $\frac{n_2}{n}$.

When we applied the $strs\mu$ strategy using the optimum allocation rule, the sample sizes, n_h, were determined by the variability in both μ_i and σ_i^2. In our $\pi p\mu$ strategy we only use μ_i to determine our sampling strategy and it is of interest to see if π_i can also be a function of σ_i^2 in either the $\pi p\mu$ or the combined sampling strategies [3].

The simulation suggested that the model-assisted strategy, when s was selected using *srswor*, was more effective than our combined sampling strategy in increasing the precision of $\hat{\tau}^{(2)}$. However one of our motivations for developing the combined sampling strategies is

[3]Using regression estimators Särndal *et al.* (1992) suggest that an estimate of the anticipated variance (Isaki and Fuller, 1989) $E_\zeta E_p[(\hat{\tau}^{(2)} - \tau^{(2)})^2]$ is minimised when $\pi_i \propto \sigma_i$

that we wish to increase the number of individuals that are observed and reduce the period of time that observers spend in very low density areas. Heuristically we would expect to observe more individuals when the sample is selected using a combined sampling design than when the sample is selected using *srswor*, although this will depend on the distribution of the $\mu_i^{(1)}$. For example suppose that $\mu_i^{(t)}$ have a uniform distribution. Under *srswor* the distribution of the $\mu_i^{(1)}$ in the sample would also have a uniform distribution. When sampling with $\pi p\mu^{(1)}$ the distribution of the $\mu_i^{(1)}$ in the sample would have a negatively skewed distribution, that is more units with high rather than low values of $\mu_i^{(1)}$ would be sampled, and so we would expect $\sum_s y_i^{(2)}$ to be higher than if s was selected using *srswor*. Except when the distribution of the $\mu_i^{(1)}$ in the survey region has negative skew, we would expect to see more individuals from a sample selected using $\pi p\mu^{(1)}$ than from a sample selected using *srswor*. When samples are selected using $\pi p\hat{\mu}^{(1)}$ we would expect a similar result unless $\hat{\mu}_i^{(1)}$ is poorly specified. In many wildlife populations the distribution of the y_i often has positive skew rather than negative skew. That is there are many units with low values of y_i and only a few units with high values of y_i. So under a model-assisted strategy when $s^{(2)}$ is selected using *srswor* we would generally expect to see fewer individuals than under a combined sampling strategy. If the precision of $\hat{\tau}^{(t)}$ is all that is of interest then a model-assisted strategy in which $s^{(2)}$ is selected using *srswor* might be more appropriate than a combined sampling strategy. If there are other considerations, such as increasing the number of individuals observed, then a combined strategy may be more appropriate.

We also wish to use a combined sampling strategy for several surveys. We would not expect $\mu_i^{(2)}$ to be very different from $\mu_i^{(1)}$, unless there has been an extreme change in the survey region. Even then the QSM may be of a similar form, but $x_{ij}^{(2)} \neq x_{ij}^{(1)}$. We would expect to be able to use data from both surveys to estimate $\hat{\mu}_i^{(2)}$ and so through time our estimate $\hat{\mu}_i^{(t)}$ should improve and therefore we expect the precision of $\hat{\tau}^{(t)}$ to increase. We look at modelling through time in more detail in Chapter 6. In comparison, although model-assisted strategies are popular for sample surveys through time, we are not aware of appropriate strategies that use survey data from time $t-1$ and auxiliary data to improve the estimate at time t. The modified regression estimator of Fuller and Rao (2001) is a

first step but does allow for the use of the right QSM. So currently the model-assisted strategy does not learn through time about the relationship between auxiliary variables and $y_i^{(t)}$ to improve the estimation of $\tau^{(t)}$

The combined sampling strategies we have developed are a useful suite of sampling strategies that incorporate $\hat{\mu}_i^{(t)}$, a summary of the relationship between the $y_i^{(t)}$ and auxiliary variables, into the survey design. This increases the precision of the estimate of $\tau^{(t)}$ compared to using an *srswor* design. As only part of the sample is selected using $\pi p\hat{\mu}^{(t)}$ the effect of a poor estimate of $\hat{\mu}_i^{(t)}$ on the precision of $\hat{\tau}^{(t)}$ is small compared to using a fully $\pi p\hat{\mu}^{(t)}$ design. As the combined sampling strategies target areas of the survey region that are expected to have high values of $y_i^{(t)}$ it will, in cases in which the distribution of the $y_i^{(t)}$ is not right skew, lead to the observers seeing more individuals than a sampling strategy in which $s^{(t)}$ is selected using *srswor*.

Currently our strategy has been developed to estimate $\tau^{(2)}$ as precisely as possible. We need to extend this strategy so that $\delta^{(1,2)} = \tau^{(2)} - \tau^{(1)}$ can be precisely estimated. We also investigate how the combined strategies can be implemented for $t > 2$ surveys. This requires consideration of the choice of $\frac{n_2}{n}$ and the estimation of $\hat{\mu}_i^{(t-1)}$.

Chapter 5

Sampling for trend estimation

In chapter 4 a suite of combined sampling strategies that used the predicted abundance from the first survey, $\hat{\mu}_i^{(1)}$, to assist in the design of survey 2 was described. By selecting part of the sample using *srswor* and part of the sample with $\pi p\hat{\mu}^{(1)}$, the precision of $\hat{\tau}^{(2)}$ was improved, compared with selecting the whole sample using *srswor*. In a monitoring programme an estimate of change in abundance from one survey to another, $\hat{\delta}^{(t',t)}$, or an estimate of trend, $\hat{\eta}_1$, is also of interest (Obj. 2(a)). In section 3.4 we saw that a sampling strategy in which part of the sample from survey t is maintained in survey t' will give a more precise estimate of $\delta^{(t',t)}$ than a strategy in which $s^{(t)}$ is selected independently of $s^{(t')}$ when $cor(y_i^{(t)}, y_i^{(t')}) > 0$.

The aim of this chapter is to investigate how the combined sampling strategy outlined in chapter 4 can be modified, so that part of the sample is retained from one survey to another, to improve estimation of $\delta^{(1,2)}$, whilst also selecting part of the sample using $\pi p\hat{\mu}^{(1)}$ for more precise estimation of $\hat{\tau}^{(2)}$. Section 5.1 describes such a sampling design. Sections 5.2 and 5.3 describe how design-based estimates of $\tau^{(2)}$ and $\delta^{(1,2)}$ respectively, can be obtained. Section 5.4 compares these new sampling strategies with strategies in which no part of the sample is maintained, the strategies described in Chapter 4, and with standard two-phase sampling designs through time. Interest is in the precision of

$\hat{\tau}^{(2)}$ and $\delta^{(1,2)}$ and how robust the strategies are to model mis-specification. As we expect the precision of $\delta^{(t',t)}$ to increase as $cor(y_i^{(t)}, y_i^{(t')})$ increases we use both population A, used in the previous chapter, where $cor(y_i^{(t)}, y_i^{(t')}) = 0.28$ and population B, described in Appendix B, where $cor(y_i^{(t)}, y_i^{(t')}) = 0.71$

5.1 The sample scheme

In the first survey, the sample $s^{(1)}$ is obtained by taking a simple random sample of size n. When the only aim is to improve the precision of $\hat{\tau}^{(2)}$, survey 2 consisted of selecting a sub-sample $s_1^{(2)}$ of $n_1^{(2)} \leq n$ units using *srswor* and a sub-sample $s_2^{(2)}$ of $n_2^{(2)} = n - n_1^{(2)}$ units using $\pi p\hat{\mu}^{(1)}$. When we also wish to obtain efficient estimates of the change in abundance, $\delta^{(1,2)} = \hat{\tau}^{(2)} - \hat{\tau}^{(1)}$, then one of these sub-samples needs to be selected from $s^{(1)}$ and the other from $s_c^{(1)} = U - s^{(1)}$. The key issue is whether we select the units from $s^{(1)}$ using *srswor* and from $s_c^{(1)}$ using $\pi p\hat{\mu}^{(1)}$ or alternatively the units from $s^{(1)}$ using $\pi p\hat{\mu}^{(1)}$ and the units from $s_c^{(1)}$ using *srswor*.

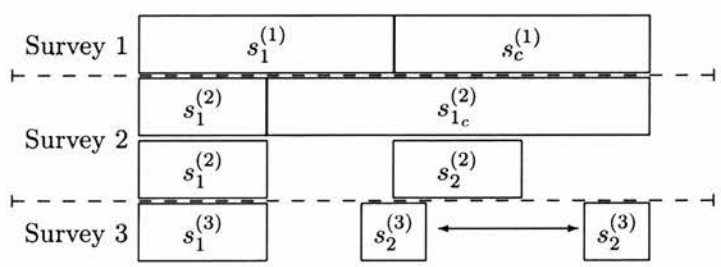
Assume that the sample $s_1^{(2)}$ is being selected to estimate $\sum_{s^{(1)}} y_i^{(2)}$ and the sample $s_2^{(2)}$ to estimate $\sum_{s_c^{(1)}} y_i^{(2)}$ and suppose that $n_1 = n_2 = \frac{n}{2}$. Usually $N - n > n$ and so $s_c^{(1)}$ is larger than $s^{(1)}$ and hence the sampling fraction $f_1 = \frac{n_1}{n}$ will be greater than $f_2 = \frac{n_2}{N-n}$. If both $s_1^{(2)}$ and $s_2^{(2)}$ are selected using *srswor* then $\sum_{s^{(1)}} y_i^{(2)}$ will be estimated more precisely than $\sum_{s_c^{(1)}} y_i^{(2)}$. It would therefore be desirable to use a sample design that gives more precise estimation of $\sum_{s_c^{(1)}} y_i^{(2)}$. So $s_1^{(2)}$ is selected using *srswor* from $s_1^{(2)}$ and $s_2^{(2)}$ is selected using $\pi p\hat{\mu}^{(t-1)}$ from $s_c^{(1)}$.

For future surveys we want a design that enables $\delta^{(t',t)}$ to be estimated as precisely as possible for any t and t'. Hence for $t > 2$ we propose that the sample $s_1^{(t)} = s_1^{(t-1)} = s_1^{(2)}$ and that $n_2^{(t)} = n_2$ units are selected from $s_{1_c}^{(t-1)} = s_{1_c}^{(t)} = U - s_1^{(t)}$ using $\pi p\hat{\mu}^{(t-1)}$. These schemes are summarised in box 5.1

We denote this new sampling scheme to select a sample $s^{(t)}$ of size n as $comb\hat{\mu}^{(t')}(\frac{n_2}{n})(s_1^{(t')}, s_{1_c}^{(t')})$.

Box 5.1: A two-phase combined sampling design through time

			Sampling Strategy	
Survey	Sample	Size	Method	From
$t = 1$	$s^{(1)}$	n	*srswor*	U
$t = 2$	$s_1^{(2)}$	$n_1^{(2)}$	*srswor*	$s^{(1)}$
	$s_2^{(2)}$	$n_2^{(2)} = n - n_1^{(2)}$	$\pi p \hat{\mu}^{(1)}$	$s_c^{(1)}$
$t > 2$	$s_1^{(t)}$	$n_1^{(t)}$	Retain	$s_1^{(t-1)} = s_1^{(2)}$
	$s_2^{(t)}$	$n_2^{(t)} = n - n_1^{(t)}$	$\pi p \hat{\mu}^{(t-1)}$	$s_{1_c}^{(t)} = s_{1_c}^{(t-1)} = s_{1_c}^{(2)}$



That is $s_1^{(t)}$ of $n - n_2$ units is selected from $s_1^{(t')}$ and $s_2^{(t)}$ of n_2 units is selected using $\pi p\hat{\mu}^{(t')}$from $s_{1_c}^{(t')}$. The strategies of the previous chapter in which none of the sample is retained from one survey to another, except by change, can therefore be denoted $comb\hat{\mu}^{(t')}(\frac{n_2}{n})(U, s_{1_c}^{(t)})$.

5.2 Estimating $\tau^{(t)}$

Using the standard Horvitz-Thompson estimator

$$\hat{\tau}^{(t)} = \sum_{s^{(t)}} \frac{y_i^{(t)}}{\pi_i^{(t)}}$$

we need to find the unconditional inclusion probability $\pi_i^{(t)}$. This is of the form

$$\begin{aligned}\pi_i^{(t)} =& Pr(i \in s^{(t)}) = Pr(i \in s_1^{(t)}) + Pr(i \in s_2^{(t)}) \\ =& Pr(i \in s_1^{(t)} | i \in s_1^{(t-1)}) Pr(i \in s_1^{(t-1)}) + Pr(i \in s_2^{(t)} | i \in s_{1_c}^{(t-1)}) Pr(i \in s_{1_c}^{(t-1)})\end{aligned}$$

At times we will denote $s^{(1)}$ as $s_1^{(1)}$ and $s_c^{(1)}$ as $s_{1_c}^{(1)}$ so that these equations can be applied for all t. When $t = 2$

$$\pi_i^{(2)} = \frac{n_1}{n}\frac{n}{N} + \sum_{s_c^{(1)} \ni i} \pi_{i_2|s^{(1)}}^{(2)} p(s^{(1)}) \tag{5.1}$$

where $\pi_{i_2|s^{(1)}}^{(2)}$ is the probability that unit i is included in $s_2^{(2)}$ given that the sample $s^{(1)}$ has been selected. Note that $\pi_{i_2|s^{(1)}}^{(2)} = \pi_{i_2|s_c^{(1)}}^{(2)}$ because this formulation only specifies that the sample $s^{(1)}$ has been selected and not whether unit $i \in s^{(1)}$ or $i \in s_c^{(1)}$. This formulation is very similar to that in the previous chapter where in equation 4.9

$$\pi_i^{(2)} = \frac{n_1}{n}\frac{n}{N} + \sum_{s_c^{(2)} \ni i} \pi_{i_2|s_1^{(2)}}^{(2)} p(s_1^{(2)}) \tag{5.2}$$

and we could find an approximation for $\pi_i^{(2)}$ because $\pi_{i_2|s_1^{(2)}}^{(2)}$ is a function of $\hat{\mu}_i^{(1)}$ only, which does not vary with $s_1^{(2)}$. For our new sampling scheme in which part of the sample

is retained from $s^{(1)}$ we require $\pi^{(2)}_{i_2|s_1^{(1)}}$ which is also a function of $\hat{\mu}_i^{(1)}$. But $\hat{\mu}_i^{(1)}$ is estimated using the data in $s^{(1)}$ and a different $s^{(1)}$ will give a different estimate of $\mu_i^{(1)}$ and so $\pi^{(2)}_{i_2|s^{(1)}}$ will vary with $s^{(1)}$. We do not know how $\hat{\mu}_i^{(1)}$, and so also how $\pi^{(2)}_{i_2|s^{(1)}}$, varies with $s^{(1)}$ so we cannot calculate the unconditional inclusion probabilities and we require a different estimation method.

The sampling schemes described in this chapter and the previous chapter are examples of two-phase sampling schemes. The general principle of these types of scheme is that two samples $s_1^{(t)}$ and $s_2^{(t)}$ are taken. The second-phase sample $s_2^{(t)}$ is in some sense conditional on the first-phase sample $s_1^{(t)}$. Two-phase estimators enable unbiased estimation of parameters such as $\tau^{(t)}$ when the probability that a unit is included in $s_2^{(t)}$ is conditional on the data collected in the first-phase sample $s_1^{(t)}$. As only one first-phase sample can be taken, unconditional inclusion probabilities, calculated over all possible first-phase samples, cannot be obtained. For example suppose that an auxiliary variable x_i, that is to be used as the variable for stratification, is not available at the start of the survey but is relatively quick to measure, but limited resources mean that all N units cannot be surveyed. Instead a first-phase sample, s_x, is taken of $n_x > n$ units in which only the variable x_i is observed. The second-phase sample s_y is selected from s_x using the stratified sampling design *strsx*. Until s_x has been taken, the probability that unit $i \in s_y$ is unknown. This probability will vary as s_x varies and so we cannot calculate at the start of the survey the probability that unit i will be included in s_y. Note that if instead s_y was selected from s_x using *srswor*, and the auxiliary data x_i are only used in the estimator, for example a regression estimator, then the two-phase estimator is not required because the probability that unit i is included in s_y does not depend on the x_i.

In our sampling design, both $s_1^{(2)}$ and $s_2^{(2)}$ are conditional on the sample $s^{(1)}$. In addition $\pi^{(2)}_{i_2|s^{(1)}}$ is a function of $\hat{\mu}_i^{(1)}$ which is determined by $s^{(1)}$ and it is this that leads us to use the two-phase estimator. If both $s_1^{(2)}$ and $s_2^{(2)}$ were selected using *srswor* from $s^{(1)}$ and

$s_c^{(1)}$ respectively, then the two-phase estimators are not required as

$$\begin{aligned}\pi_i^{(2)} =& Pr(i \in s_1^{(2)} | i \in s_1^{(1)}) Pr(i \in s_1^{(1)}) + Pr(i \in s_2^{(2)} | i \in s_2^{(1)}) Pr(i \in s_2^{(1)}) \\ \Rightarrow \pi_i^{(2)} =& \frac{n_1}{n}\frac{n}{N} + \frac{n_2}{N-n}\frac{N-n}{N} = \frac{n}{N}\end{aligned}$$

and given the unconditional $\pi_i^{(2)}$, we can estimate $\tau^{(2)}$ using the Horvitz-Thompson estimator.

Two-phase sampling is described in detail in Särndal *et al.* (1992). They provide design-based estimators for standard two-phase unequal probability sampling designs implemented in one survey and regression estimators for two-phase unequal probability sampling designs through time, where $s^{(2)}$ is conditional on $s^{(1)}$. Here we describe design-based estimators that use the principles of two-phase estimation to estimate $\tau^{(t)}$ and $\delta^{(t',t)}$.

The estimate $\hat{\tau}^{(t)}$ is obtained by taking a weighted sum of two estimates of $\tau^{(t)}$ so that

$$\hat{\tau}^{(t)} = \omega\hat{\tau}_1^{(t)} + (1-\omega)\hat{\tau}_2^{(t)} \qquad 0 \leq \omega \leq 1 \tag{5.3}$$

where $\hat{\tau}_k^{(t)}$ uses the data from $s_k^{(t)}$. The variance of $\hat{\tau}^{(t)}$ is

$$var(\hat{\tau}^{(t)}) = \omega^2 var(\hat{\tau}_1^{(t)}) + (1-\omega)^2 var(\hat{\tau}_2^{(t)}) + 2\omega(1-\omega)cov(\hat{\tau}_1^{(t)}, \hat{\tau}_2^{(t)}) \tag{5.4}$$

By differentiation we can show that the variance $var(\hat{\tau}^{(t)})$ can be minimised, given $var(\hat{\tau}_k^{(t)})$, $var(\hat{\tau}^{(t)})$ and $cov(\hat{\tau}_1^{(t)}, \hat{\tau}_2^{(t)})$, by setting ω to be

$$\omega = \frac{var(\hat{\tau}_2^{(t)}) - cov(\hat{\tau}_1^{(t)}, \hat{\tau}_2^{(t)})}{var(\hat{\tau}_1^{(t)}) + var(\hat{\tau}_2^{(t)}) - 2cov(\hat{\tau}_1^{(t)}, \hat{\tau}_2^{(t)})} \tag{5.5}$$

As $s_1^{(t)}$ is selected using *srswor* we can obtain the estimate $\hat{\tau}_1^{(t)}$ very simply, as the unconditional inclusion probability $\pi_{i_1}^{(t)}$, the probability that $i \in s_1^{(t)}$, is:

$$\begin{aligned}\pi_{i_1}^{(t)} &= \sum_{s_1^{(t)} \ni i} p(s_1^{(t)}) \\ &= \sum_{s_1^{(t)} \ni i} \sum_{s_1^{(t-1)} \ni i} p(s_1^{(t)} | s_1^{(t-1)}) p(s_1^{(t-1)}) \\ &= \sum_{s_1^{(t-1)} \ni i} \pi_{i_1 | s_1^{(t-1)}}^{(t)} p(s_l^{(t-1)})\end{aligned}$$

When $t = 2$, then $s_1^{(t-1)} = s^{(1)}$ and so $\pi^{(2)}_{i_1|s^{(1)}} = \frac{n_1}{n}$, whereas when $t > 2$, the inclusion probability $\pi^{(t)}_{i_1|s_1^{(t-1)}} = 1$ as units in $s_1^{(t)}$ are retained from one survey to another, so

$$\pi^{(2)}_{i_1} = \frac{n_1}{N} = \pi^{(t)}_{i_1} \text{ for } t \geq 2 \tag{5.6}$$

Hence we can use the Horvitz-Thompson estimator to obtain $\hat{\tau}_1^{(t)}$, and its variance and corresponding estimate are given by the Sen-Yates-Grundy formulation, equations 3.12 and 3.13

To estimate $\hat{\tau}_2^{(t)}$ using the Horvitz-Thompson estimator, we would require the unconditional inclusion probability

$$\pi^{(t)}_{i2} = \sum_{s_{1_c}^{(t-1)} \ni i} p(s_1^{(t-1)}) \pi^{(t)}_{i_2|s_1^{(t-1)}} \tag{5.7}$$

As previously stated, we cannot calculate $\pi^{(t)}_{i_2|s_1^{(t-1)}}$, because $\hat{\mu}_i^{(1)}$ depends on $s^{(t-1)}$. Instead we use the two-phase estimator. The design-based two-phase sampling estimator of $\tau_2^{(t)}$, its variance and estimator and the covariance $cov(\hat{\tau}_1^{(t)}, \hat{\tau}_2^{(t)})$ are stated in box 5.2. For notational simplicity, $t' = t - 1$ and except where specified the superscript is (t) . For the case when $t' = 1$ we let $s_1^{(1)} = s^{(1)}$. We let $E_{s_k^{(t')}}$ represent the expectation over all possible samples $s_k^{(t')}$ which may be shortened to E_k for $E_{s_k^{(t)}}$.

We illustrate the principle behind these estimators by demonstrating how $\hat{\tau}_2^{(t)}$, equation 5.8, is an unbiased estimator of $\tau^{(t)}$.

As $s_1^{(t')}$ is a probability sample, $s_{1_c}^{(t')} = U - s^{(t')}$ is also a probability sample, and if we knew y_i for all $i \in s_{1_c}^{(t')}$, then an unbiased estimate of τ is

$$\tau = \sum_{s_{1_c}^{(t')}} \frac{y_i}{\pi^{(t')}_{i_{1_c}}}$$

as

$$E_{s_{1_c}^{(t')}} \left[\sum_{s_{1_c}^{(t')}} \frac{y_i}{\pi^{(t')}_{i_{1_c}}} \right] = \sum_{U} \frac{y_i}{\pi^{(t')}_{i_{1_c}}} E_{s_{1_c}^{(t')}} \left[I^{(t')}_{i_{1_c}} \right] = \tau \text{ where } I^{(t')}_{i_{1_c}} = \begin{cases} 1 & \text{if } i \in s_{1_c}^{(t')} \\ 0 & \text{otherwise} \end{cases}$$

Box 5.2: Two-phase estimators for $\tau_2^{(t)}$, $var(\hat{\tau}_2^{(t)})$ and $cov(\hat{\tau}_2^{(t)}, \hat{\tau}_2^{(t)})$ derived for the sample design $comb\hat{\mu}^{(t-1)}(\frac{n_2}{n})(s_1^{(t-1)}, s_{1_c}^{(t-1)})$. For notational simplicity $t-1$ is denoted t' and the suffix (t) is omitted.

$$\hat{\tau}_2 = \sum_{s_2^{(t)}} \frac{y_i}{\pi_i^\dagger} \tag{5.8}$$

$$var(\hat{\tau}_2) = \sum_U \sum_U \Delta_{(ij)1_c}^{(t')} \frac{y_i y_j}{\pi_{i_{1_c}}^{(t')} \pi_{j_{1_c}}^{(t')}} + E_{s_{1_c}^{(t')}} \left[\sum_{s_{1_c}^{(t')}} \sum_{s_{1_c}^{(t')}} \Delta_{(ij)_{2|s_1^{(t')}}} \frac{y_i}{\pi_i^\dagger} \frac{y_j}{\pi_j^\dagger} \right] \tag{5.9}$$

$$\widehat{var}(\hat{\tau}_2) = \sum_{s_2} \sum_{s_2} \frac{\Delta_{(ij)1_c}^{(t')}}{\pi_{ij}^\dagger} \frac{y_i y_j}{\pi_{i_{1_c}}^{(t')} \pi_{j_{1_c}}^{(t')}} + \sum_{s_2} \sum_{s_2} \frac{\Delta_{(ij)_{2|s_1^{(t')}}}}{\pi_{(ij)2|s_1^{(t')}}} \frac{y_i}{\pi_i^\dagger} \frac{y_j}{\pi_j^\dagger} \tag{5.10}$$

$$cov(\hat{\tau}_1, \hat{\tau}_2) = -\sum_U \sum_U \Delta_{(ij)_1}^{(t')} \frac{y_i}{\pi_{i_1}^{(t')}} \frac{y_j}{\pi_{j_{1_c}}^{(t')}} \tag{5.11}$$

$$\widehat{cov}(\hat{\tau}_1, \hat{\tau}_2) = \sum_{s_1} \sum_{s_1} \frac{\Delta_{(ij)_1}^{(t')}}{\pi_{(ij)_{1|s_1^{(t')}}} \pi_{(ij)_1}^{(t')}} \frac{y_i}{\pi_{i_1}^{(t')}} \frac{y_j}{\pi_{j_{1_c}}^{(t')}} \tag{5.12}$$

where

$$\pi_i^\dagger = Pr(i \in s_2 | s_1^{(t')}) Pr(i \in s_{1_c}^{(t')}) = \pi_{i_2|s_1^{(t')}} \pi_{i_{1_c}}^{(t')} \tag{5.13}$$

$$\pi_{ij}^\dagger = Pr(i\&j \in s_2 | s_1^{(t')}) Pr(i\&j \in s_{1_c}^{(t')}) = \pi_{(ij)_2|s_1^{(t')}} \pi_{(ij)_{1_c}}^{(t')} \tag{5.14}$$

$$\Delta_{ij}^{(t)\dagger} = \pi_{ij}^{(t)\dagger} - \pi_i^{(t)\dagger} \pi_j^{(t)\dagger} \tag{5.15}$$

$$\Delta_{(ij)_k}^{(t')} = \pi_{(ij)_k}^{(t')} - \pi_{i_k}^{(t')} \pi_{j_k}^{(t')} \tag{5.16}$$

However we only know y_i for a sample s_2 of n_2 units from $s_{1_c}^{(t')}$. As s_2 is also a probability sample, an unbiased estimate of $\sum_{s_{1_c}^{(t')}} y_i$ is

$$\sum_{s_{1_c}^{(t')}} y_i = \sum_{s_2} \frac{y_i}{\pi_{i_{2|s_1^{(t')}}}}$$

as

$$E_{2|s_{1_c}^{(t')}}\left[\sum_{s_2} \frac{y_i}{\pi_{i_{2|s_1^{(t')}}}}\right] = E_{2|s_{1_c}^{(t')}}\left[\sum_{s_{1_c}^{(t')}} \frac{y_i}{\pi_{i_{2|s_1^{(t')}}}} I_{i_{2|s_1^{(t')}}}\right] = \sum_{s_1^{(t')}} y_i \text{ where } I_{i_k}^{(t')} = \begin{cases} 1 & \text{if } i \in s_k^{(t')} \\ 0 & \text{otherwise} \end{cases}$$

By combining these two results so that we take the expectation over both $s_{2|s_{1_c}^{(t')}}$ and $s_{1_c}^{(t')}$, we find that

$$\begin{aligned} E[\hat{\tau}_2] &= E_{s_{1_c}^{(t')}}\left[E_{2|s_{1_c}^{(t')}}\left[\sum_{i \in s_2} \frac{y_i}{\pi_{i_{2|s_{1_c}^{(t')}}} \pi_{i_{1_c}}^{(t')}}\right]\right] \\ &= E_{s_{1_c}^{(t')}}\left[\sum_{s_{1_c}} \frac{y_i}{\pi_{i_{2|s_{1_c}^{(t')}}} \pi_{i_{1_c}}^{(t')}} E_{2|s_{1_c}^{(t')}}[I_{i_{2|s_1^{(t')}}}^{(t)}]\right] \\ &= E_{s_{1_c}^{(t')}}\left[\sum_{s_{1_c}} \frac{y_i}{\pi_{i_{1_c}}^{(t')}}\right] = \tau \end{aligned}$$

For notational simplicity the equations 5.13 - 5.15 summarise the double inclusion probabilities.

Using the same principle of conditioning for the variance,

$$
\begin{aligned}
var(\hat{\tau}_2) =& var_{s_{1c}^{(t')}}\left(E_{2|s_{1c}^{(t')}}[\hat{\tau}_2]\right) + E_{s_{1c}^{(t')}}\left[var_{2|s_{1c}^{(t')}}(\hat{\tau}_2)\right] \\
=& var_{s_{1c}^{(t')}}\left[\sum_{s_{1c}^{(t')}} \frac{y_i}{\pi_{i_{1c}}^{(t')}}\right] + E_{s_{1c}^{(t')}}\left[E_{2|s_{1c}^{(t')}}\left[\sum_{s_2}\sum_{s_2} \frac{y_i y_j}{\pi_i^{\dagger}\pi_j^{\dagger}}\right] - E_{2|s_{1c}^{(t')}}\left[\sum_{s_2} \frac{y_i}{\pi_i^{\dagger}}\right] E_{2|s_{1c}^{(t')}}\left[\sum_{s_2} \frac{y_j}{\pi_j^{\dagger}}\right]\right] \\
=& \sum_U \sum_U \left(\pi_{(ij)_{1c}}^{(t')} - \pi_{i_{1c}}^{(t')}\pi_{j_{1c}}^{(t')}\right) \frac{y_i y_j}{\pi_{i_{1c}}^{(t')}\pi_{j_{1c}}^{(t')}} \\
&+ E_{s_{1c}^{(t')}}\left[E_{2|s_{1c}^{(t')}}\left[\sum_{s_{1c}^{(t')}}\sum_{s_{1c}^{(t')}} \frac{y_i y_j}{\pi_i^{\dagger}\pi_j^{\dagger}}\left(I_{(ij)_{2|s_{1c}^{(t')}}} - I_{i_{2|s_{1c}^{(t')}}} I_{j_{2|s_{1c}^{(t')}}}\right)\right]\right] \\
=& \sum_U \sum_U \Delta_{(ij)_{1c}}^{(t')} \frac{y_i y_j}{\pi_{i_{1c}}^{(t')}\pi_{j_{1c}}^{(t')}} + E_{s_{1c}^{(t')}}\left[\sum_{s_{1c}^{(t')}}\sum_{s_{1c}^{(t')}} \Delta_{(ij)_{2|s_{1c}^{(t')}}} \frac{y_i}{\pi_i^{\dagger}} \frac{y_j}{\pi_j^{\dagger}}\right]
\end{aligned}
$$

Note that this cannot be calculated explicitly except when the inclusion probabilities $\pi_{i_2}^{(t)}$ are not dependent on $s_1^{(t)}$ or $s^{(t')}$, for example using the sample design *srswor*.

We can show that equation 5.10 is an unbiased estimate of the variance, equation 5.9, by taking expectations of the estimated variance

$$
\begin{aligned}
E\left[\widehat{var}(\hat{\tau}_2)\right] =& E_{s_{1c}^{(t')}}\left[E_{2|s_{1c}^{(t')}}\left[\sum_{s_2}\sum_{s_2} \frac{\Delta_{(ij)_{1c}}^{(t')}}{\pi_{(ij)_{2|s_{1c}^{(t')}}}\pi_{(ij)_{1c}}^{(t')}} \frac{y_i y_j}{\pi_{i_{1c}}^{(t')}\pi_{j_{1c}}^{(t')}}\right]\right] \\
&+ E_{s_{1c}^{(t')}}\left[E_{2|s_{1c}^{(t')}}\left[\sum_{s_2}\sum_{s_2} \frac{\Delta_{(ij)_{2|s_{1c}^{(t')}}}}{\pi_{(ij)_{2|s_{1c}^{(t')}}}} \frac{y_i}{\pi_i^{\dagger}} \frac{y_j}{\pi_j^{\dagger}}\right]\right] \\
=& E_{s_{1c}^{(t')}}\left[\sum_{s_{1c}^{(t')}}\sum_{s_{1c}^{(t')}} \frac{\Delta_{(ij)_{1c}}^{(t')}}{\pi_{(ij)_{1c}}^{(t')}} \frac{y_i y_j}{\pi_{i_{1c}}^{(t')}\pi_{j_{1c}}^{(t')}}\right] + E_{s_{1c}^{(t')}}\left[\sum_{s_{1c}}\sum_{s_{1c}} \Delta_{(ij)_{2|s_{1c}^{(t')}}} \frac{y_i}{\pi_i^{\dagger}} \frac{y_j}{\pi_j^{\dagger}}\right] \\
=& \sum_U \sum_U \Delta_{(ij)_{1c}}^{(t')} \frac{y_i y_j}{\pi_{i_{1c}}^{(t')}\pi_{j_{1c}}^{(t')}} + E_{s_{1c}^{(t')}}\left[\sum_{s_{1c}^{(t')}}\sum_{s_{1c}^{(t')}} \Delta_{(ij)_{2|s_{1c}^{(t')}}} \frac{y_i}{\pi_i^{\dagger}} \frac{y_j}{\pi_j^{\dagger}}\right]
\end{aligned}
$$

Using similar principles for covariances, we show that

$$\begin{aligned}
cov(\hat{\tau}_1, \hat{\tau}_2) =& E_{s_1^{(t')}}\left[E_{s|s_1^{(t')}}\left[\hat{\tau}_1\hat{\tau}_2\right]\right] - E_{s_1^{(t')}}\left[E_{s|s_1^{(t')}}\left[\hat{\tau}_1\right]\right] E_{s_1^{(t')}}\left[E_{s|s_1^{(t')}}\left[\hat{\tau}_2\right]\right] \\
=& E_{s_1^{(t')}}\left[E_{s|s_1^{(t')}}\left[\sum_{s_1}\sum_{s_2}\frac{y_i}{\pi_{i_1|s_1^{(t')}}\pi_{i_1}^{(t')}}\frac{y_j}{\pi_j^{\dagger}}\right]\right] \\
& - E_{s_1^{(t')}}\left[\sum_{s_1}\frac{y_i}{\pi_{i_1|s_1^{(t')}}\pi_{i_1}^{(t')}}\right] E_{s_1^{(t')}}\left[E_{s|s_1^{(t')}}\left[\sum_{s_2}\frac{y_j}{\pi_j^{\dagger}}\right]\right] \\
=& E_{s_1^{(t')}}\left[\sum_{s_1^{(t')}}\sum_{s_{1_c}^{(t')}}\frac{y_i}{\pi_{i_1}^{(t')}}\frac{y_j}{\pi_{j_{1_c}}^{(t')}}\right] - E_{s_1^{(t')}}\left[\sum_{s_{1_c}^{(t')}}\frac{y_i}{\pi_{i_1}^{(t')}}\right] E_{s_{1_c}^{(t')}}\left[\frac{y_j}{\pi_{j_{1_c}}^{(t')}}\right] \\
=& \sum_U\sum_U \pi_{i_1 j_{1_c}}^{(t')}\frac{y_i}{\pi_{i_1}^{(t')}}\frac{y_j}{\pi_{j_{1_c}}^{(t')}} - \sum_U\sum_U y_i y_j
\end{aligned}$$

where $\pi_{i1j1_c}^{(t')} = \pi_{i1}^{(t')} - \pi_{(ij)_1}^{(t')}$ is the probability that $i \in s^{(t')}$ and $j \in s_{1_c}^{(t')}$ so that

$$cov(\hat{\tau}_1, \hat{\tau}_2) = -\sum_U\sum_U\left(\pi_{(ij)_1}^{(t')} - \pi_{i_1}^{(t')}\pi_{j_1}^{(t')}\right)\frac{y_i}{\pi_{i_1}^{(t')}}\frac{y_j}{\pi_{j_{1_c}}^{(t')}}$$

Again by taking expectations we see that equation 5.12 is an unbiased estimate of the covariance, equation 5.11

$$\begin{aligned}
E[\widehat{cov}(\hat{\tau}_1, \hat{\tau}_2)] =& -E_{s_1^{(t')}}\left[E_{s_1|s_1^{(t')}}\left[\sum_{s_1}\sum_{s_1}\frac{\Delta_{(ij)_1}^{(t')}}{\pi_{(ij)_{1|s_{1_c}^{(t')}}}\pi_{(ij)_1}^{(t')}}\frac{y_i}{\pi_{i_1}^{(t')}}\frac{y_j}{\pi_{j_{1_c}}^{(t')}}\right]\right] \\
=& E_{s_1^{(t')}}\left[\sum_{s_1^{(t')}}\sum_{s_1^{(t')}}\frac{\Delta_{(ij)_1}^{(t')}}{\pi_{(ij)_1}^{(t')}}\frac{y_i}{\pi_{i_1}^{(t')}}\frac{y_j}{\pi_{j_{1_c}}^{(t')}}\right] \\
=& -\sum_U\sum_U \Delta_{(ij)_1}^{(t')}\frac{y_i}{\pi_{i_1}^{(t')}}\frac{y_j}{\pi_{j_{1_c}}^{(t')}}
\end{aligned}$$

The estimator $cov(\hat{\tau}_1, \hat{\tau}_2)$ only uses data in s_1. This is reasonable as s_1 is a sample from U and so we use these data as representative of the population to estimate the covariance. However given that the data from s_2 are not used, it does not make the best use of the data available. For example when n_1 is small and n_2 is large, then we would not expect the estimator to be very precise. As only the data in s_1 are used, we can easily express the covariance and its estimator in terms of n, n_1 and N. For example when $t' = 1$ and $t = 2$, then $s_1^{(1)} = s^{(1)}$ and so

$$cov(\hat{\tau}_1^{(2)}, \hat{\tau}_2^{(2)}) = -\left(\sum_U y_i^{(2)2} - \frac{1}{N-1} \sum_U \sum_{\substack{U \\ i \neq j}} y_i^{(2)} y_j^{(2)} \right) \tag{5.17}$$

$$\widehat{cov}(\hat{\tau}_1^{(2)}, \hat{\tau}_2^{(2)}) = -\frac{N}{n_1}\left(\sum_{s_1^{(2)}} y_i^{(2)2} - \frac{1}{n_1 - 1} \sum_{s_1^{(2)}} \sum_{\substack{s_1^{(2)} \\ i \neq j}} y_i^{(1)} y_j^{(2)} \right) \tag{5.18}$$

The term $-cov(\hat{\tau}_1^{(2)}, \hat{\tau}_2^{(2)})$ is the expression for the variance of the data in U. If all N units had the same value so that $y_i^{(2)} = y^{(2)}$, then the covariance would be zero as the totals $\hat{\tau}_1^{(2)}$ and $\hat{\tau}_2^{(2)}$ would remain the same for all possible s_1 and s_2.

5.3 Estimating $\delta^{(1,2)}$

The change in total abundance between surveys and its variance are, as described in section 3.4,

$$\hat{\delta}^{(t',t)} = \hat{\tau}^{(t)} - \hat{\tau}^{(t')} \tag{5.19}$$

$$var(\hat{\delta}^{(t',t)}) = var(\hat{\tau}^{(t)}) + var(\hat{\tau}^{(t')}) - 2cov(\hat{\tau}^{(t)}, \hat{\tau}^{(t')}) \tag{5.20}$$

When $\hat{\tau}^{(t)}$ is estimated using the Horvitz-Thompson estimator, the covariance and its estimator are:

$$cov(\hat{\tau}^{(t')}, \hat{\tau}^{(t)}) = \sum_U \sum_U \frac{y_i^{(t')}}{\pi_i^{(t')}} \frac{y_j^{(t)}}{\pi_j^{(t)}} \left(\pi_{ij}^{(t',t)} - \pi_i^{(t')} \pi_j^{(t)} \right) \tag{5.21}$$

$$\widehat{cov}(\hat{\tau}^{(t')}, \hat{\tau}^{(t)}) = \sum_{s^{(t)}} \sum_{s^{(t')}} \frac{y_i^{(t')}}{\pi_i^{(t')}} \frac{y_j^{(t)}}{\pi_j^{(t)}} \left(\frac{\pi_{ij}^{(t',t)} - \pi_i^{(t')} \pi_j^{(t)}}{\pi_{ij}^{(t',t)}} \right) \tag{5.22}$$

When the full sample is retained from one survey to another so that $s = s^{(t')} = s^{(t)}$ and s is selected using *srswor*,

$$cov(\hat{\tau}^{(t')}, \hat{\tau}^{(t)}) = \frac{N-n}{n} \left[\sum_U y_i^{(t')} y_i^{(t)} - \frac{1}{N-1} \sum_U \sum_{\substack{U \\ i \neq j}} y_i^{(t')} y_j^{(t)} \right] \tag{5.23}$$

$$\widehat{cov}(\hat{\tau}^{(t')}, \hat{\tau}^{(t)}) = \frac{N(N-n)}{n^2} \left[\sum_s y_i^{(t')} y_i^{(t)} - \frac{1}{n-1} \sum_s \sum_{\substack{s \\ i \neq j}} y_i^{(t')} y_j^{(t)} \right] \tag{5.24}$$

In general only part of the sample is retained from one survey to another so the covariance is of the form:

$$\begin{aligned} cov(\hat{\tau}^{(t')}, \hat{\tau}^{(t)}) = & \omega^{(t')} \omega^{(t)} cov(\hat{\tau}_1^{(t')}, \hat{\tau}_1^{(t)}) + (1 - \omega^{(t')}) \omega^{(t)} cov(\hat{\tau}_2^{(t')}, \hat{\tau}_1^{(t)}) \\ & + \omega^{(t')} (1 - \omega^{(t)}) cov(\hat{\tau}_1^{(t')}, \hat{\tau}_2^{(t)}) + (1 - \omega^{(t')})(1 - \omega^{(t)}) cov(\hat{\tau}_2^{(t')}, \hat{\tau}_2^{(t)}) \end{aligned} \tag{5.25}$$

In this section we consider in detail the case in which $t' = 1$ and $t = 2$ so that

$$cov(\hat{\tau}^{(1)}, \hat{\tau}^{(2)}) = \omega^{(2)} cov(\hat{\tau}^{(1)}, \hat{\tau}_1^{(2)}) + (1 - \omega^{(2)}) cov(\hat{\tau}^{(1)}, \hat{\tau}_2^{(2)}) \tag{5.26}$$

Both $\tau^{(1)}$ and the estimate of $\tau^{(2)}$ using the data from $s_1^{(t)}$ only are estimated using the Horvitz-Thompson estimator and so the covariance and its estimator are of the form given in equations 5.21 and 5.22. The inclusion probabilities $\pi_i^{(1)}$, $\pi_{j_1}^{(2)}$ and $\pi_{ij_1}^{(1,2)}$, the probability that unit i is included in $s^{(1)}$ and unit j is included in $s_1^{(2)}$, are

$$\pi_i^{(1)} = \frac{n}{N} \qquad \pi_{j_1}^{(2)} = \frac{n_1}{n}$$

$$\pi_{ij_1}^{(1,2)} = Pr(j \in s_1^{(2)} | i\&j \in s^{(1)}) Pr(i\&j \in s^{(1)}) = \begin{cases} \frac{n_1}{n}\frac{n(n-1)}{N(N-1)} = \frac{n_1(n-1)}{N(N-1)} & i \neq j \\ \frac{n_1}{n}\frac{n}{N} = \frac{n_1}{N} & i = j \end{cases}$$

and so the covariance and its estimator are

$$cov(\hat{\tau}^{(1)}, \hat{\tau}_1^{(2)}) = \frac{N-n}{n}\left[\sum_U y_i^{(1)} y_i^{(2)} - \frac{1}{N-1}\sum_U \sum_{\substack{U \\ i \neq j}} y_i^{(1)} y_j^{(2)}\right] \tag{5.27}$$

$$\widehat{cov}(\hat{\tau}^{(1)}, \hat{\tau}_1^{(2)}) = \frac{(N-n)N}{nn_1}\left[\sum_{s_1^{(2)}} y_i^{(1)} y_i^{(2)} - \frac{1}{n-1}\sum_{s^{(1)}} \sum_{\substack{s_1^{(2)} \\ i \neq j}} y_i^{(1)} y_j^{(2)}\right] \tag{5.28}$$

When the whole sample is retained from one survey to another so that $s_1^{(2)} = s^{(2)}$ and $n_1 = n$, then these equations are equivalent to equations 5.23 and 5.24.

To obtain $cov(\hat{\tau}^{(1)}, \hat{\tau}^{(2)})$, we also require the term $cov(\hat{\tau}^{(1)}, \hat{\tau}_2^{(2)})$

$$
\begin{aligned}
cov(\hat{\tau}^{(1)}, \hat{\tau}_2^{(2)}) =& E_{s^{(1)}}\left[E_{s_2^{(2)}|s_1^{(t')}}\left[\sum_{s^{(1)}}\sum_{s_2^{(2)}} \frac{y_i^{(1)}}{\pi_i^{(1)}} \frac{y_j^{(2)}}{\pi^{(2)}_{j|s^{(1)}}\pi^{(1)}_{j_c}}\right]\right] \\
& - E_{s^{(1)}}\left[E_{s_2^{(2)}|s_1^{(t')}}\left[\sum_{s_1^{(t')}} \frac{y_i^{(1)}}{\pi_i^{(1)}}\right]\right] E_{s^{(1)}}\left[E_{s_2^{(2)}|s_1^{(t')}}\left[\sum_{s_2^{(2)}} \frac{y_j^{(2)}}{\pi^{(2)}_{j|s^{(1)}}\pi^{(1)}_{j_c}}\right]\right] \\
=& E_{s^{(1)}}\left[\sum_{s^{(1)}}\sum_{s_c^{(1)}} \frac{y_i^{(1)}}{\pi_i^{(1)}} \frac{y_j^{(2)}}{\pi^{(1)}_{j_c}}\right] - E_{s^{(1)}}\left[\sum_{s_1^{(t')}} \frac{y_i^{(1)}}{\pi_i^{(1)}}\right] E_{s^{(1)}}\left[\sum_{s_c^{(1)}} \frac{y_j^{(2)}}{\pi^{(1)}_{j_c}}\right] \\
=& \sum_U \sum_U \pi^{(1)}_{ij_c} \frac{y_i^{(1)}}{\pi_i^{(1)}} \frac{y_j^{(2)}}{\pi^{(1)}_{j_c}} - \sum_U y_i^{(1)} \sum_U y_j^{(2)} \\
=& \sum_U \sum_U \left(\pi_{ij}^{(1)} - \pi_i^{(1)}\pi_i^{(1)}\right) \frac{y_i^{(1)}}{\pi_i^{(1)}} \frac{y_j^{(2)}}{\pi^{(1)}_{j_c}} \qquad (5.29)
\end{aligned}
$$

which is the same form as equation 5.11. An unbiased estimator is obtained by changing the summation to be for the units in $s_1^{(2)}$ only and dividing by the joint inclusion probabilities $\pi^{(2)}_{(ij)_{1|s_1^{(1)}}}\pi_{ij}^{(1)}$. This is the same procedure as was used to obtain an unbiased estimator of equation 5.11, so that

$$
\widehat{cov}(\hat{\tau}^{(1)}, \hat{\tau}_2^{(2)}) = -\sum_{s_1^{(2)}}\sum_{s_1^{(2)}} \frac{(\pi_{(ij)}^{(1)} - \pi_i^{(1)}\pi_j^{(1)})}{\pi^{(2)}_{(ij)_{1|s^{(1)}}}\pi_{ij}^{(1)}} \frac{y_i^{(1)}}{\pi_i^{(1)}} \frac{y_i^{(2)}}{\pi^{(1)}_{j_c}} \qquad (5.30)
$$

and hence

$$
cov(\hat{\tau}^{(1)}, \hat{\tau}_2^{(2)}) = -\left(\sum_U y_i^{(1)} y_i^{(2)} - \frac{1}{N-1}\sum_U \sum_{\substack{U \\ i\neq j}} y_i^{(1)} y_j^{(2)}\right) \qquad (5.31)
$$

$$
\widehat{cov}(\hat{\tau}^{(1)}, \hat{\tau}_2^{(2)}) = -\frac{N}{n_1}\left(\sum_{s_1^{(2)}} y_i^{(1)} y_i^{(2)} - \frac{1}{n_1 - 1}\sum_{s_1^{(2)}} \sum_{\substack{s_1^{(2)} \\ i\neq j}} y_i^{(1)} y_j^{(2)}\right) \qquad (5.32)
$$

These are the same form as equations 5.17 and 5.18 except that $y_i^{(2)}$ is replaced with $y_i^{(1)}$. This covariance, $cov(\hat{\tau}^{(1)}, \hat{\tau}_2^{(2)})$, is equivalent to the negative sample covariance of

$y_i^{(1)}$ and $y_i^{(2)}$. We note as we did for the estimator $\widehat{cov}(\hat{\tau}_1^{(2)}, \hat{\tau}_2^{(2)})$ that the estimator for $\widehat{cov}(\hat{\tau}^{(1)}, \hat{\tau}_2^{(2)})$ does not use any data from $s_2^{(2)}$ so that we would expect the covariance to be poorly estimated when $s_1^{(2)}$ is small.

We have given the explicit form of the covariance for the case when $t' = 1$ and $t = 2$. For other values of t' and t, the covariance is more difficult to derive, particularly when neither t' nor t is equal to 1. As $s_1^{(t)} = s_1^{(2)}$ for $t = 2, \ldots,$ the covariance $cov(\hat{\tau}^{(1)}, \hat{\tau}_1^{(t)})$ and its estimator are as given in equation 5.27 and 5.28. Calculation of the other covariances is more complex. In fact rather than estimate all four covariance terms, given in equation 5.25, we use a similar strategy to that of Holmes and Skinner (2000) in that only the units from the matched sample are used to estimate the covariance. That is we approximate the covariance so that

$$cov(\hat{\tau}^{(t')}, \hat{\tau}^{(t)}) \approx \omega^{(t')}\omega^{(t)} cov(\hat{\tau}_1^{(t')}, \hat{\tau}_1^{(t)}) \tag{5.33}$$

This is reasonable because most of the covariance is due to the matched units and we note that further work to obtain the full set of covariance terms is required.

5.3.1 Estimating $\delta^{(1,2)}$ when $s_1^{(t)}$ and $s_2^{(t)}$ are selected using *srswor*.

In general the standard rotating panel designs used for estimation of $\tau^{(t)}$ and $\delta^{(t',t)}$ select all sub-samples using *srswor*. For two surveys, these designs are equivalent to our sampling design when $s_2^{(2)}$ is selected using *srswor* rather than $\pi p \hat{\mu}^{(t)}$. For these designs, it is possible to calculate unconditional inclusion probabilities and so $\tau^{(t)}$ can be estimated using the Horvitz-Thompson estimator and for the difference $\hat{\delta}^{(t',t)}$, the covariance $cov(\hat{\tau}^{(t')}, \hat{\tau}^{(t)})$ can be calculated using equation 5.21. This requires the joint inclusion probability

$$\begin{aligned}\pi_{ij}^{(t',t)} =& Pr(j \in s^{(t)} | i \& j \in s^{(t')}) Pr(i \& j \in s^{(t')}) \\ &+ Pr(j \in s^{(t)} | i \in s^{(t')}, j \in s_c^{(t')}) Pr(i \& j \in s_c^{(t')})\end{aligned}$$

When $t' = 1$ and $t = 2$

$$\pi_{ij}^{(1,2)} = \begin{cases} \frac{n^2 - n_1}{N(N-1)} & i \neq j \\ \frac{n_1}{N} & i = j \end{cases}$$

so that the covariance and its estimator are

$$cov(\hat{\tau}^{(1)}, \hat{\tau}^{(2)}) = \frac{n_1 N - n^2}{n^2} \left[\sum_U y_i^{(1)} y_i^{(2)} - \frac{1}{N-1} \sum_U \sum_{\substack{U \\ i \neq j}} y_i^{(1)} y_j^{(2)} \right] \tag{5.34}$$

$$\widehat{cov}(\hat{\tau}^{(1)}, \hat{\tau}^{(2)}) = \frac{N(n_1 N - n^2)}{n_1 n^2} \left[\sum_{s_1} y_i^{(1)} y_i^{(2)} - \frac{n_1}{n^2 - n_1} \sum_{s_{1c}^{(2)}} \sum_{\substack{s_2^{(2)} \\ i \neq j}} y_i^{(1)} y_j^{(2)} \right] \tag{5.35}$$

Again when $s^{(t')} = s^{(t)}$, these are of the form of equations 5.23 and 5.24.

5.4 Simulation

The following sampling design is implemented on population A.

1. Select a sample, $s^{(1)}$, of $n = 50$ units using *srswor*.
2. Predict $\hat{\mu}_i^{(1)}$
3. Select $s_1^{(2)}$ of n_1 units from $s^{(1)}$ using *srswor* where $n_1 = 50, 38, 25, 12, 0$
4. Select $s_2^{(2)}$ of $n_2 = n - n_1$ units from $s_c^{(1)}$ using $\pi p \hat{\mu}^{(1)}$.

We expect a strategy in which part of the sample is retained from one survey to another to be more effective when $cor(y_i^{(1)}, y_i^{(2)})$ is high. For population A, $cor(y_i^{(1)}, y_i^{(2)}) = 0.28$, so we also implement the sampling design on population B described in Appendix B.3.1 for which $cor(y_i^{(1)}, y_i^{(2)}) = 0.71$. The same $s^{(1)}$ and hence $\hat{\mu}_i^{(1)}$ as were obtained in the simulation in section 4.4 are used for both populations as it is the the $y_i^{(2)}$ that vary between the two populations.

Table 5.1: Mean (s.d.) for various parameters and estimators obtained from 1000 simulations using the sampling strategy $comb\hat{\mu}^{(1)}(\frac{n_2}{n})(s^{(1)}, s_c^{(1)})$ where a sample $s_1^{(2)}$ of n_1 units is selected from $s^{(1)}$ using *srswor* and a sample $s_2^{(2)}$ of n_2 units from $s_c^{(1)}$ using $\pi p\hat{\mu}^{(1)}$. $\tau^{(2)} = 2546$ and $\widehat{cov}_s$ is the empirical estimate of the covariance $cov(\hat{\tau}_1^{(2)}, \hat{\tau}_2^{(2)}) = -3804$ using the 1000 estimates of $\hat{\tau}^{(1)}$ and $\hat{\tau}^{(2)}$. The estimate $\hat{\tau}^{(2)}$ is calculated to be $\omega\hat{\tau}_1^{(2)} + (1-\omega)\hat{\tau}_2^{(2)}$.

$\frac{n_2}{n}$	$\hat{\tau}^{(2)}$	$\sqrt{\widehat{var}(\hat{\tau}^{(2)})}$	$\hat{\tau}_1^{(2)}$	$\hat{\tau}_2^{(2)}$	$\widehat{cov}$	$\widehat{cov}_s$	ω
0.00	2557 (264)	267 (35)	2557(264)	–	–	–	1.00 (0.00)
0.26	2551 (249)	249 (35)	2561 (308)	2653 (530)	-3816 (1168)	-2383	0.68 (0.12)
0.50	2597 (242)	242 (34)	2556 (387)	2686 (355)	-3804 (1505)	-3967	0.45 (0.12)
0.74	2626 (239)	239 (34)	2595 (561)	2683 (288)	-3906 (2224)	-5711	0.23 (0.11)
1.00	2547 (240)	238 (37)	–	2547 (240)	–	–	0.00 (0.00)

5.4.1 Estimation of $\tau^{(2)}$

Equation 5.3 is used to estimate $\tau^{(2)}$ where $\hat{\tau}_1^{(2)}$ is estimated using the Horvitz-Thompson estimator, and equation 5.8 to estimate $\hat{\tau}_2^{(2)}$. Table 5.1 summarises these results for population A. The estimates $\hat{\tau}_1^{(2)}$ are unbiased and become more variable as $\frac{n_2}{n}$ increases, as they are calculated using a smaller sample. Similarly the estimates $\hat{\tau}_2^{(2)}$ are unbiased and become less variable as $\frac{n_2}{n}$, increases as they are calculated using a larger sample. We only give the empirical estimates of the standard error of $\hat{\tau}_1^{(2)}$ and $\hat{\tau}_2^{(2)}$ in table 5.1 but the analytic values are similar: 372 (56), 459 (87) and 659 (184) for $\sqrt{\widehat{var}(\hat{\tau}_1^{(2)})}$; 612 (204), 435 (112) and 349 (69) for $\sqrt{\widehat{var}(\hat{\tau}_2^{(2)})}$. When $\frac{n_2}{n} = 0$, all of $s^{(1)}$ is retained in survey 2. When $\frac{n_2}{n} = 1$, none of $s^{(1)}$ is retained and $s^{(2)}$ is selected using $\pi p\hat{\mu}^{(1)}$.

The covariance $cov(\hat{\tau}_1^{(2)}, \hat{\tau}_2^{(2)})$ does not vary with $\frac{n_2}{n}$ as it is a function of N, see equation 5.17. As $\frac{n_2}{n}$ increases the variability in the estimate of $\widehat{cov}(\hat{\tau}_1^{(2)}, \hat{\tau}_2^{(2)})$ increases, because a smaller sample is used for its estimation. This variability may not be too important as the estimated correlation between $\hat{\tau}_1^{(2)}$ and $\hat{\tau}_2^{(2)}$ is approximately 0.03 for all three values

of $\frac{n_2}{n}$. Over these 1000 simulations, the empirical estimate of the covariance, estimated using the 1000 values of $\hat{\tau}_1^{(2)}$ $\hat{\tau}_2^{(2)}$, is different from the analytic estimate of covariance. Further inspection of these values indicates that the covariance of the simulated totals is extremely variable. This is due to both the variability in the $\hat{\mu}_i^{(1)}$, which are used to select the $s_2^{(2)}$, and the variability in the $s_2^{(2)}$, which occurs even when $\hat{\mu}_i^{(1)}$ remains fixed across simulations. A much greater number of simulations would be needed before the empirical estimate of the covariance more closely matched the true covariance because of the high variability between samples.

The value of ω is determined using the estimates $\widehat{var}(\hat{\tau}_1^{(2)})$, $\widehat{var}(\hat{\tau}_2^{(2)})$ and $\widehat{cov}(\hat{\tau}_1^{(2)}, \hat{\tau}_2^{(2)})$ in equation 5.5. As $\frac{n_2}{n}$ increases, ω decreases as greater weight is given to $\hat{\tau}_2^{(2)}$. If we select $s_2^{(2)}$ using *srswor*, i.e. if $\hat{\mu}^{(1)}$ is constant, the average value of ω is close to $\frac{n_1}{n}$. From a set of 1000 such simulations, in which $s_2^{(2)}$ is selected using *srswor*, we obtained mean values for ω of 0.72, 0.50 and 0.27 for $\frac{n_1}{n}$ taking the values 0.74,0.50 and 0.26 respectively. We would expect this as when both $s_1^{(2)}$ and $s_2^{(2)}$ are selected using *srswor* the relative precision of the two estimators depends only on $\frac{n_1}{n}$. When $s_2^{(2)}$ is selected using $\pi p\hat{\mu}^{(1)}$, the average value of ω is less than $\frac{n_1}{n}$, values of 0.68, 0.45 and 0.2 respectively, because the precision of $\hat{\tau}_2^{(2)}$ is greater than would be obtained if $s_2^{(2)}$ was selected using *srswor*.

The estimates of $\hat{\tau}^{(2)}$ are unbiased. It appears from the results of the simulation that under the combined sampling strategy $\hat{\tau}^{(2)}$ increases with $\frac{n_2}{n}$. This is a feature of this particular set of samples and does not seem to be the case in general. Again by increasing the number of simulations we would expect this effect to disappear.

In the previous chapter we required a sampling strategy that was robust to model misspecification. This was investigated by looking at the model-averaged design variance when $\hat{\mu}_i = \mu_i^b$ for varying b. Using the two-phase sampling estimators, it is not possible to calculate the model-averaged design variance as the variance, equation 5.4, cannot be expressed explicitly. Instead we calculate $\widehat{var}(\hat{\tau}^{(2)})$ for a sample $s^{(2)}$ where $s_2^{(2)}$ is selected using $\pi p\hat{\mu}^{(1)}$ using the same $\hat{\mu}_i$ as section 4.3.1. For a fixed $s^{(1)}$, we repeat the process of selecting $s^{(2)}$ and estimating the variance 1000 times. Figure 5.1 gives the mean value

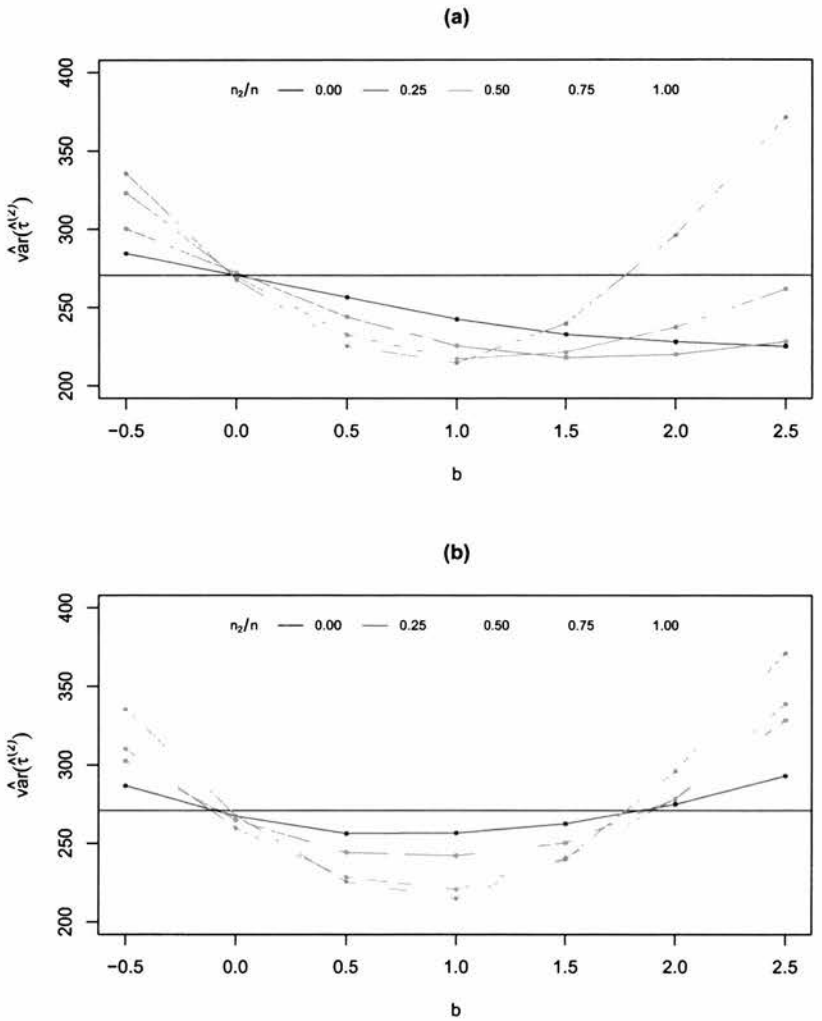


Figure 5.1: Effect of model mis-specification on $\widehat{var}(\hat{\tau}^{(2)})$ for varying b and n_2 using strategy: (a) $comb\mu^b(\frac{n_2}{n})(U, s_1^{(2)})$ where $s_1^{(2)}$ is selected from U using *srswor* and $s_2^{(2)}$ is selected from $s_{1_c}^{(2)}$ using $\pi p\hat{\mu}^b$; (b) $comb\mu^b(\frac{n_2}{n})(s^{(1)}, s_c^{(1)})$ where $s_1^{(2)}$ is selected from $s^{(1)}$ using *srswor* and $s_2^{(2)}$ is selected from $s_c^{(1)}$ using $\pi p\hat{\mu}^b$

of $\hat{\tau}^{(2)}$ for $b = -0.5, 0, \ldots, 2.5$ and $\frac{n_2}{n} = 0, 0.25, 0.5, 0.75, 1$. For comparison figure 5.1(a) shows the results from using the sampling design of the previous chapter in which $s_1^{(2)}$ is selected from U as described in detail in section 4.3. Under either scheme, the minimum value of $\widehat{var}(\hat{\tau}^{(2)})$ is obtained using the fully $\pi p\hat{\mu}^{(1)}$ sample design when $b = 1$. When $s_1^{(2)}$ is selected from $s^{(1)}$, the minimum value of $\widehat{var}(\hat{\tau}^{(2)})$ occurs when $b = 1$ for all n_2 and $\widehat{var}(\hat{\tau}^{(2)})$ increases as a greater proportion of the sample is selected using *srswor*, i.e. as n_2 decreases. As b moves away from one, $\widehat{var}(\hat{\tau}^{(2)})$ increases most rapidly when $\frac{n_2}{n} = 1$, the fully $\pi p\hat{\mu}^{(1)}$design, and the rate of increase decreases as n_2 decreases. For all values of n_2 the effect of changing b is roughly symmetric about $b = 1$. The sampling design *srswor* is better than the combined sampling design when b is less than zero and greater than approximately two.

The effect of model mis-specification on the precision of $\hat{\tau}^{(2)}$ is different under the two sampling strategies. When $s_1^{(2)}$ is selected from U,

$$\hat{\tau}^{(2)} = N\tau_{U\hat{\mu}} \sum_{s^{(2)}} \frac{y_i^{(2)}}{n_1\tau_{U\hat{\mu}} + Nn_2\hat{\mu}_i^{(2)}}$$

and the effect of model mis-specification is reduced for all units in $s^{(2)}$. When $s_1^{(2)}$ is selected from $s^{(1)}$ rather than from U and $\omega = \frac{n_1}{n}$,

$$\hat{\tau}^{(2)} = \frac{N}{n} \sum_{s_1^{(2)}} y_i^{(2)} + \frac{N\tau_{s_c^{(1)}\hat{\mu}}}{n(N-n)} \sum_{s_2^{(2)}} \frac{y_i^{(2)}}{\hat{\mu}_i^{(2)}}$$

and $\hat{\tau}_2^{(2)}$, estimated using the units in $s_2^{(2)}$, behaves as a fully $\pi p\hat{\mu}^{(1)}$estimate of $\hat{\tau}^{(2)}$, so that the complete effect of model mis-specification is applied to these n_2 units. The effect of this is that the most precise estimates of $\hat{\tau}^{(2)}$ are obtained by implementing a fully $\pi p\hat{\mu}^{(1)}$ design, rather than a design in which part of the sample is selected using *srswor*. By retaining units from one survey to another the precision of our estimate of $\tau^{(2)}$ is less than if none of the sample was retained from the previous survey when $b > 1$.

Table 5.2: Mean (s.d) of $\widehat{cov}(\hat{\tau}^{(1)}, \hat{\tau}_k^{(2)})$ for $k = 1, 2$ for populations A and B over 1000 simulations using strategy $comb\hat{\mu}^{(1)}(\frac{n_2}{n})(s^{(1)}, s_c^{(1)})$ where $s_1^{(2)}$ is selected using *srswor* from $s^{(1)}$ and $s_2^{(2)}$ is selected using $\pi p\hat{\mu}^{(1)}$. $\widehat{cov}(\hat{\tau}^{(1)}, \hat{\tau}_k^{(2)})$ is estimated using equation 5.28 for $k = 1$ and equation 5.32 for $k = 2$ and $\widehat{cov}_s$ is the empirical estimate of the covariance $cov(\hat{\tau}^{(1)}, \hat{\tau}_k^{(2)})$

	$cov(\hat{\tau}^{(1)}, \hat{\tau}_1^{(2)})$					
	A			B		
	$cor(y_i^{(1)}, y_i^{(2)}) = 0.28$			$cor(y_i^{(1)}, y_i^{(2)}) = 0.71$		
$\frac{n_2}{n}$	$\widehat{cov}$		$\widehat{cov}_s$	$\widehat{cov}$		$\widehat{cov}_s$
1.00	14,269	(9,298)	12,845	44,519	(12,824)	41,419
0.26	14,349	(12,846)	14,360	45,114	(17,307)	42,822
0.50	14,888	(18,135)	12,818	45,419	(24,507)	43,088
0.74	15,522	(30,627)	20,285	45,910	(40,366)	49,667
cov	14,734			44,417		
	$cov(\hat{\tau}^{(1)}, \hat{\tau}_2^{(2)})$					
0.26	-1,079	(1,366)	-2079	-2,921	(1,799)	-4,389
0.50	-1,027	(937)	2162	-2,875	(1,236)	-1,841
0.74	-1,081	(752)	786	-2,905	(991)	-2,257
cov	-775			-2,338		

5.4.2 Estimation of $cov(\hat{\tau}^{(1)}, \hat{\tau}^{(2)})$

Table 5.2 shows the mean estimated covariances $\widehat{cov}(\hat{\tau}^{(1)}, \hat{\tau}_1^{(2)})$ and $\widehat{cov}(\hat{\tau}^{(1)}, \hat{\tau}_2^{(2)})$ using equations 5.28 and 5.32 respectively for populations A and B. As $cor(y_i^{(1)}, y_i^{(2)})$ is greater for population B than population A, the covariances for population B are greater than for population A. The covariances that are estimated are functions of n and N and so remain fixed for all n_1. It is the weight $w < \frac{n_1}{n}$ that determines the final covariance, see table 5.3. As n_1 decreases, so the precision of the estimates decreases because the data in $s_2^{(2)}$ are not used. When $\frac{n_2}{n}$ is large, the empirical estimate of the covariances do not match that

Table 5.3: $cov(\hat{\tau}^{(1)}, \hat{\tau}^{(2)})$ and mean (s.d) of $\widehat{cov}(\hat{\tau}^{(1)}, \hat{\tau}^{(2)})$ over 1000 simulations for populations A and B using strategy $comb\hat{\mu}^{(1)}(\frac{n_2}{n})(s^{(1)}, s_c^{(1)})$ where $s_1^{(2)}$ is selected using *srswor* from $s^{(1)}$ and $s_2^{(2)}$ is selected using $\pi p\hat{\mu}^{(1)}$. $\widehat{cov}$ is calculated using equation 5.26 and $\widehat{cov}_s$ is the empirical estimate of the covariance.

	A $cor(y_i^{(1)}, y_i^{(2)}) = 0.28$			B $cor(y_i^{(1)}, y_i^{(2)}) = 0.71$		
$\frac{n_2}{n}$	cov	$\widehat{cov}$	$\widehat{cov}_s$	cov	$\widehat{cov}$	$\widehat{cov}_s$
1.00	14,734	14,269 (9,298)	12,845	44,417	44,519 (12,834)	41,419
0.26	9,771	9,074 (8,393)	8,453	29,923	29,464 (11,226)	29,458
0.50	6,204	5,492 (7,573)	7,549	19,637	18,330 (9,813)	18,898
0.74	2,792	1,841 (6,194)	6,361	9,351	6,617 (7,220)	13,987

closely with the true covariance. However the empirical estimates of covariance can vary greatly and this is due to the variability in $s_2^{(2)}$. More importantly the analytic estimates of covarianceare not very precise when $\frac{n_2}{n}$ is high. This is because they are estimated using the data in $s_1^{(2)}$ only.

5.4.3 Estimation of $\delta^{(1,2)}$

We wish to compare the results of this new sampling design, $comb\hat{\mu}^{(1)}(\frac{n_2}{n})(s^{(1)}, s_c^{(1)})$, in which $s_1^{(2)}$ is selected from $s^{(1)}$ and $s_2^{(2)}$ is selected using $\pi p\hat{\mu}^{(1)}$ with the combined sampling design $comb\hat{\mu}^{(1)}(\frac{n_2}{n})(U, s_{1_c}^{(2)})$ in which $s_1^{(2)}$ is selected from U and $s_2^{(2)}$ is selected using $\pi p\hat{\mu}^{(1)}$, and with the strategy in which both $s_1^{(2)}$ and $s_2^{(2)}$ are selected using *srswor* from $s^{(1)}$ and $s_c^{(1)}$ respectively, $comb1(\frac{n_2}{n})(s^{(1)}, s_c^{(1)})$. Table 5.4 gives the mean value of $\hat{\delta}^{(1,2)}$ over 1,000 simulations for each of these three strategies for both populations A and B. These estimates are unbiased.

The means (s.d) of the estimated variances, $\widehat{var}(\hat{\delta}^{(1,2)})$, are given in table 5.5 for populations A and B. The most precise estimates of $\delta^{(1,2)}$ are obtained when $s_1^{(2)}$ is selected from

Table 5.4: Mean (s.d) of $\delta^{(1,2)} = 996$ for populations A and B over 1000 simulations where $s^{(1)}$ is selected using *srswor* and $s^{(2)}$ is selected using (a) rotating panel design so that $s_1^{(2)}$ and $s_2^{(2)}$ are selected using *srswor* from $s^{(1)}$ and $s_c^{(1)}$ respectively (b) $s_1^{(2)}$ is selected from $s^{(1)}$ using *srswor* and $s_2^{(2)}$ is selected from $s_c^{(1)}$ using $\pi p\hat{\mu}^{(1)}$ (c) $s_1^{(2)}$ is selected from U using *srswor* and $s_2^{(2)}$ is selected from $s_{1_c}^{(2)}$ using $\pi p\hat{\mu}^{(1)}$.

	A			B		
	$s_1^{(2)} \subseteq s^{(1)}$		$s_1^{(2)} \subseteq U$	$s_1^{(2)} \subseteq s^{(1)}$		$s_1^{(2)} \subseteq U$
$\frac{n_2}{n}$	(a) *srswor*	(b) $\pi p\hat{\mu}^{(1)}$	(c) $\pi p\hat{\mu}^{(1)}$	(a) *srswor*	(b) $\pi p\hat{\mu}^{(1)}$	(c) $\pi p\hat{\mu}^{(1)}$
0.00	996 (279)	996 (279)	979 (329)	1080 (228)	1080 (228)	1076 (375)
0.26	991 (289)	991 (290)	982 (311)	1077 (274)	1055 (296)	1080 (351)
0.50	992 (308)	1036 (294)	976 (307)	1083 (315)	1098 (325)	1060 (345)
0.74	1005 (307)	1065 (303)	1001 (290)	1079 (344)	1116 (346)	1079 (361)
1.00	979 (329)	987 (303)	987 (303)	1076 (375)	1084 (355)	1084 (355)

$s^{(1)}$ and $s_2^{(2)}$ is selected using $\pi p\hat{\mu}^{(1)}$. When $s_1^{(2)} \subseteq s^{(1)}$, the precision of $\hat{\delta}^{(1,2)}$ decreases as n_2 increases because the covariance $cov(\hat{\tau}^{(1)}, \hat{\tau}^{(2)})$ decreases as a smaller proportion of $s^{(1)}$ is retained in survey 2. In comparison, when none of the sample is retained from $s^{(1)}$, the precision of $\hat{\delta}^{(1,2)}$ increases with increasing n_2. Under population A, where $cor(y_i^{(1)}, y_i^{(2)})$ is low, the increase in the precision of $\hat{\delta}^{(1,2)}$ from retaining part of $s^{(1)}$ but selecting $s_2^{(2)}$ using *srswor*, is less than the increase in precision from not retaining any of $s^{(1)}$ but selecting $s_2^{(2)}$ using $\pi p\hat{\mu}^{(1)}$. This is because $cov(\hat{\tau}^{(1)}, \hat{\tau}^{(2)})$ is small and the precision of $\hat{\tau}^{(2)}$ remains constant with increasing $\frac{n_2}{n}$ when $s_2^{(2)}$ is selected using *srswor*, but decreases when $s_2^{(2)}$ is selected using $\pi p\hat{\mu}^{(1)}$. Under population B, when $cor(y_i^{(1)}, y_i^{(2)}) = 0.71$, the increase in the precision of $\hat{\delta}^{(1,2)}$ from retaining part of $s^{(1)}$ under a fully *srswor* design is greater than the increase in precision from selecting $s_2^{(2)}$ using $\pi p\hat{\mu}^{(1)}$, as $cov(\hat{\tau}^{(1)}, \hat{\tau}^{(2)})$ is large. For this population, the standard rotating panel design gives a more precise estimate of $\hat{\delta}^{(1,2)}$ than a sampling design in which $s^{(2)}$ does not retain any units from $s^{(1)}$. The sampling design developed in this chapter combines both the benefits of retaining the sample and

Table 5.5: Mean (s.d) of $\sqrt{\widehat{var}(\hat{\delta}^{(1,2)})}$ from 1000 simulations where $s^{(1)}$ is selected using *srswor* and $s^{(2)}$ is selected using (a) rotating panel design so that $s_1^{(2)}$ and $s_2^{(2)}$ are selected using *srswor* from $s^{(1)}$ and $s_c^{(1)}$ respectively (b) $s_1^{(2)}$ is selected using *srswor* from $s^{(1)}$ and $s_2^{(2)}$ is selected from $s_c^{(1)}$ using $\pi p\hat{\mu}^{(1)}$ (c) $s_1^{(2)}$ is selected from U using *srswor* and $s_2^{(2)}$ is selected from $s_{1_c}^{(2)}$ using $\pi p\hat{\mu}^{(1)}$.

	A $cor(y_i^{(1)}, y_i^{(2)}) = 0.28$			B $cor(y_i^{(1)}, y_i^{(2)}) = 0.71)$		
	$s_1^{(2)} \subseteq s^{(1)}$		$s_1^{(2)} \subseteq U$	$s_1^{(2)} \subseteq s^{(1)}$		$s_1^{(2)} \subseteq U$
$\frac{n_2}{n}$	*srswor*	$\pi p\hat{\mu}^{(1)}$	$\pi p\hat{\mu}^{(1)}$	*srswor*	$\pi p\hat{\mu}^{(1)}$	$\pi p\hat{\mu}^{(1)}$
0.00	283 (33)	283 (33)	331 (33)	230 (31)	230 (31)	375 (40)
0.26	293 (40)	284 (35)	314 (26)	270 (44)	265 (34)	362 (35)
0.50	306 (40)	291 (34)	305 (25)	312 (48)	303 (37)	354 (33)
0.74	321 (38)	301 (34)	303 (25)	351 (44)	335 (39)	354 (37)
1.00	331 (33)	306 (32)	306 (32)	375 (40)	358 (45)	358 (45)

the benefits of selecting $s_2^{(2)}$ using $\pi p\hat{\mu}^{(1)}$ and in these simulations does better than either of the two competing strategies, even though it is less robust to model mis-specification than the other two strategies.

5.5 Discussion

In this chapter, we have extended the sampling design of Chapter 4, so that part of the sample in survey t is selected from survey $t-1$ using *srswor* and the rest is selected using $\pi p\hat{\mu}^{(t-1)}$ from $s_{1_c}^{(t-1)}$. This strategy gives more precise estimates of $\delta^{(1,2)}$ than our previous sample design in which none of the sample is retained from the previous survey. In addition it gives more precise estimates of both $\tau^{(2)}$ and $\delta^{(1,2)}$ than a design-based estimate of a standard rotating panel design, selected using *srswor*. As $\frac{n_2}{n}$ increases, the precision of $\hat{\delta}^{(1,2)}$ decreases and the precision of $\hat{\tau}^{(2)}$ increases.

The two-phase sampling strategy in which units are retained from the previous survey behaves differently with respect to model mis-specification than a strategy in which none of the sample is retained. When $0 < b < 2$ and units are retained from the previous survey a fully $\pi p\hat{\mu}^{(1)}$ strategy gives the most precise estimates of $\tau^{(t)}$ and the precision of $\hat{\tau}^{(2)}$ decreases as $\frac{n_2}{n}$ decreases. When units are not retained and b is greater than one a fully $\pi p\hat{\mu}^{(1)}$ strategy gives less precise estimates of $\tau^{(2)}$ than a combined sampling strategy in which $0 < \frac{n_2}{n} < 1$.

The sampling scheme from Chapter 4 in which $s_1^{(2)}$ is selected from U is also a two-phase sampling scheme. Rather than estimating $\tau^{(2)}$ by approximating the unconditional inclusion probabilities, $\pi_i^{(2)}$, an alternative estimator would be to use a form of two-phase estimator described here. We would, because of the effect of model mis-specification, expect the precision of the estimates to be poorer using these estimators than using the approximation.

In this chapter it was necessary to use a two-phase sampling estimator to estimate $\tau^{(2)}$ and $\delta^{(1,2)}$ because the probability that unit $i \in s_2^{(2)}$ is conditionally dependent on $\hat{\mu}_i^{(1)}$, $s^{(1)}$, $\pi_i^{(1)}$ and $\frac{n_2}{n}$. Given $\hat{\mu}_i^{(1)}$, the probability that $i \in s_2^{(2)}$ also depends on the units in the sample $s_c^{(1)}$. This has led to a relatively complex estimator of $var(\hat{\tau}^{(2)})$ as it requires the estimation of a covariance term. If the variability in $\hat{\mu}_i^{(1)}$ is small the approximation $\pi_i^{(t)} \doteq \frac{n_1}{N} + \pi_{i_U}^{(t)}$ as used in the previous chapter could be appropriate. However as illustrated in figure 4.7 the model mis-specification varied from $b = 0$ to $b = 2$ over our 1000 samples of $s^{(1)}$. This suggests that $\hat{\mu}_i^{(1)}$ varies considerably, and so in this case we require the two-phase estimators suggested in this chapter.

If the monitoring programme continues for many surveys, we might under some conditions assume that $\hat{\mu}_i^{(t)}$ will not vary with the choice of $s^{(t-1)}$. For example if we assume that $\mu_i^{(t)} = r^t \mu_i^{(0)}$ we may use the data from all surveys $1, \ldots, t$ to estimate $\hat{\mu}_i^{(t)}$. Each survey will only contribute a small amount to the estimate of $\hat{\mu}_i^{(t)}$. When t is large, we might assume that the same $\hat{\mu}_i^{(t)}$ could be obtained given any particular set of samples $s^{(1)}, \ldots, s^{(t)}$. If this is the case, the approximation for the unconditional inclusion probability $\pi_i^{(t)}$ can

be used and $\hat{\tau}^{(t)}$ estimated using the Horvitz-Thompson estimator. Estimation of the covariance $cov(\hat{\tau}^{(t')}, \hat{\tau}^{(t)})$ would be more difficult as it uses equation 5.28 and the inclusion probability $\pi_{ij}^{(t,t')}$ must be specified. A common solution is that of Holmes and Skinner (2000) in which the covariance is estimated using the $y_i^{(1)}$ and $y_i^{(2)}$ for $i \in s_1^{(t)} \bigcap s_1^{(t-1)}$. The recent development of Berger (2003a) which uses all of the data in $s^{(1)}$ and $s^{(2)}$ may be an alternative solution.

None of the estimators of covariance use the data in $s_2^{(t)}$. As a greater proportion of the sample is allocated to $s_2^{(t)}$, the precision of the covariance will decrease. The c.v. of the estimated covariance $cov(\hat{\tau}^{(1)}, \hat{\tau}^{(2)})$ may seem low (3%-6%) but it is ten times the c.v. of the estimated variance when $\frac{n_2}{n} = 0.74$. The covariance will also decrease in size as $\frac{n_2}{n}$ increases but is still large enough to influence the precision of $\hat{\delta}^{(1,2)}$. For example under population B when $\frac{n_2}{n} = 0.74$ the correlation $cor(\hat{\tau}^{(1)}, \hat{\tau}^{(2)})$ is approximately 0.16. The mean estimate of the correlation, using $\widehat{cov}(\hat{\tau}^{(1)}, \hat{\tau}^{(2)})$, is 0.11 so that the precision of $\hat{\delta}^{(t',t)}$ will be underestimated. It is not clear whether the use of $s_2^{(t)}$ to estimate the covariance will increase the precision of the covariance because of the high variability in the $s_2^{(t)}$, as shown by the empirical estimates of covariance when $\frac{n_2}{n}$ is high. In Chapter 8 we propose a bootstrap estimator that incorporates all of the sample data in the estimation of the covariance.

We have compared our new sampling strategies with the standard rotating panel designs when $\hat{\tau}^{(2)}$ and $\hat{\delta}^{(1,2)}$ are design-based estimators. In practice design-based estimation of $\delta^{(1,2)}$ or $\tau^{(2)}$ under a rotating panel design will use a model-assisted estimator rather than the design-based estimators proposed here. As demonstrated in Chapter 4 we might expect model-assisted estimators to be more precise than fully design-based estimators when both samples are selected using *srswor*. In addition they may do better than our combined sampling strategies in which a design-based estimator is employed. The standard model-assisted estimator assumes a regression model of the form

$$y_i^{(2)} = \alpha_0 + \alpha_1 y_i^{(1)}$$

The populations that we use are of the form

$$Y_i^{(2)} \sim Po(r^2 \mu_i^{(0)})$$

and so we would wish our model to be

$$y_i^{(2)} = r y_i^{(1)}$$

These models only work across surveys and do not use auxiliary information within a survey to improve estimation, neither do they take the form of the superpopulation model into account.

In the previous chapter, the model-assisted estimators we used (Breidt and Opsomer, 2000) did take the form of the superpopulation model into account. In addition they used auxiliary variables to improve the estimation. These strategies only worked efficiently when data from one survey were used. It would be of interest to try to incorporate this auxiliary information in the model-assisted estimators through time. The modified regression estimators of Fuller and Rao (2001) may be a first attempt at this although they do not use the correct model formulation as we require $log(\hat{\mu}_i^{(2)}) = \sum_{j=1}^{Q} f_j(x_{ij}^{(t)})$ as is possible using the methods of Breidt and Opsomer (2000).

Chapter 6

Long-term monitoring strategies

Two combined sampling designs have been described that use a map of predicted abundance to determine the sample in the following survey. So far we have demonstrated how, given an estimate of $\mu_i^{(2)}$, these designs can be used to select $s^{(2)}$. In addition we have shown how varying the proportion selected using $\pi p\hat{\mu}^{(t)}$ affects the precision of $\tau^{(2)}$ and $\delta^{(1,2)}$ when these are estimated using appropriate design-based estimators. The aim of this thesis was to develop sampling strategies that could be used within a long-term monitoring programme. In this chapter we look at how the combined sampling strategies we have developed can be implemented within a monitoring programme. We demonstrate the methods using population P, first described in chapter 2 and illustrated in figures 2.1,2.2 and 2.3. Further details are given in Appendix B, section B.3.2. We assume that there are the resources to take a sample of $n = 81$ units in each survey. Unlike the populations A and B, on which the sampling designs were originally tested, the units in this population have a spatial location. As described in chapter 3, a systematic sampling design is often used when sampling over space. In section 6.1, we show how our two sampling designs can be implemented when $s^{(1)}$ and $s_1^{(t)}$ are selected using *sys* rather than *srswor*. In particular we demonstrate how we might retain part of the sample from $s^{(1)}$ to $s_1^{(2)}$ so that $s_1^{(2)}$ is also a systematic sample. In the following section 6.2 we describe how our combined

sampling designs can be implemented within a long-term monitoring programme. The three key issues that must be addressed for selecting our sample $s^{(t)}$ are: how we estimate $\mu_i^{(t)}$, section 6.3; whether part of the sample is to be retained from one survey to another, section 6.4; and the proportion of the sample that is selected using $\pi p\hat{\mu}^{(t)}$ in survey t, section 6.5. Section 6.6 implements a number of monitoring strategies on our population over a period of $T = 10$ surveys. Section 6.7 discusses the implications of the results from the simulation for the design of monitoring programmes.

6.1 Use of a systematic sampling design in place of *srswor*.

When an area is to be surveyed, systematic sampling, which we denote *sys*, is commonly used as it provides good coverage of the survey region compared with *srswor*. This is useful when we wish to estimate $\hat{\mu}_i^{(t)}$ for $i \in U$. As stated in section 3.5, and shown by Matérn (1986), *sys* will provide efficient estimates of $\hat{\tau}^{(t)}$ when there is spatial autocorrelation. Care must be taken to ensure that the systematic sample is not aligned with any periodic variation in $y_i^{(t)}$ over the survey region. Figure 6.1(a) illustrates an *sys* design of 9x9 units. Section 3.5 discussed issues of variance estimation when sampling using *sys*. Here there are only 16 possible samples but we assume that for variance estimation the samples are selected using *srswor*. The first row in table 6.1 compares the results from 128 simulations when $s^{(1)}$ is selected using *srswor*, or *sys*. Each of the 16 possible *sys* designs was replicated 8 times. After the first survey the selected sample will vary between the 8 replicates with the same $s^{(1)}$. We see that $\hat{\tau}^{(2)}$ is unbiased for both sampling designs, and that the standard deviation is greater under *srswor* than under *sys*. Hence we tend to overestimate the variance by assuming *srswor* when our sample has been selected using *sys*.

When we retain part of the sample from $s^{(1)}$ to $s^{(2)}$, we could select the units from the *sys* sample, $s^{(1)}$, at random. If we wish to retain some form of systematic coverage in $s^{(2)}$, then only certain units in the sample can be retained. Figures 6.1(b-d) illustrate, in green, a set of systematic sampling designs in which 75%, 50% and 25% of units have

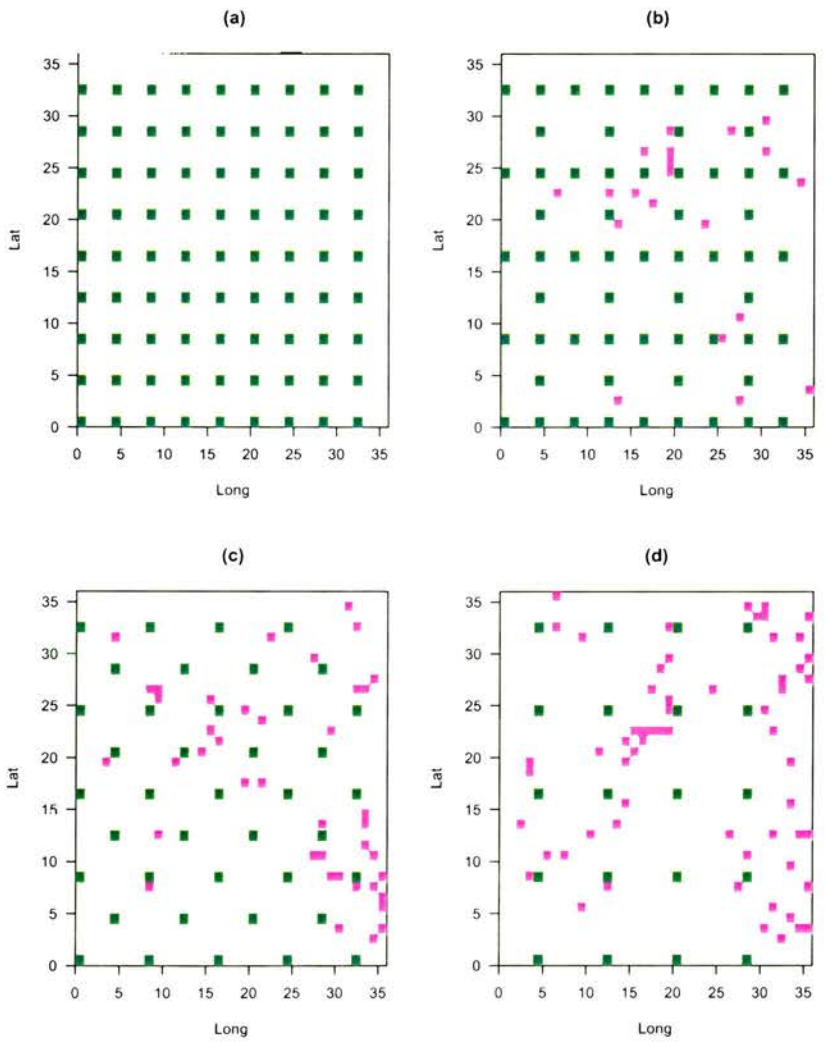


Figure 6.1: Sample designs for $s^{(2)}$ where $s_1^{(2)}$ of size n_1 is selected from $s^{(1)}$, a systematic sample, and n_1 varies in size. The sub-sample $s_2^{(2)}$ of size $n_2 = n - n_1$ is selected using $\pi p\hat{\mu}^{(1)}$ from $s_c^{(1)}$ where $\hat{\mu}^{(1)}$ is predicted from $s^{(1)}$ (a) $n_1 = 81$, $n_2 = 0$ (b) $n_1 = 61$, $n_2 = 40$ (c) $n_1 = 40$, $n_2 = 41$ (d) $n_1 = 0$, $n_2 = 81$

Table 6.1: Summary of results for survey 2 of population P. $s^{(1)}$ is selected using *sys* or *srswor* and $s_1^{(2)}$ is selected using *sys* or *srswor* respectively from U and $s_2^{(2)}$ is selected using $\pi p\hat{\mu}^{(1)}$. Results are mean (s.d) of 128 simulations.

	sys		*srswor*	
$\frac{n_2^{(t)}}{n}$	$\hat{\tau}^{(2)}$	$\sqrt{\widehat{var}(\hat{\tau}^{(2)})}$	$\hat{\tau}^{(2)}$	$\sqrt{\widehat{var}(\hat{\tau}^{(2)})}$
0.00	2588 (202)	431 (75)	2572 (458)	403 (127)
0.25	2607 (255)	322 (45)	2507 (295)	297 (51)
0.50	2543 (248)	259 (33)	2547 (259)	262 (30)
0.75	2567 (237)	236 (27)	2535 (260)	232 (27)
1.00	2618 (275)	254 (51)	2589 (284)	251 (70)

been retained from $s^{(1)}$. The remaining units in the sample, shown in pink, have been selected using $\pi p\hat{\mu}_i^{(1)}$. The second, third and fourth rows of table 6.1 show the estimates of $\hat{\tau}^{(1)}$ and $\widehat{var}(\hat{\tau}^{(1)})$ when only a proportion of $s^{(1)}$ is selected using *sys* or *srswor* and the remaining proportion selected using $\pi p\hat{\mu}^{(1)}$. These again indicate that if we estimate $var(\hat{\tau}^{(2)})$ assuming that $s_1^{(2)}$ is selected using *srswor* we obtain an overestimate of $var(\hat{\tau}^{(2)})$; the empirical estimate of the standard error is less than than the analytic estimate. The difference between the analytic and empirical estimates reduces as the proportion of the sample selected using *sys* decreases.

6.2 Long-term monitoring strategies

In chapter 2 we stated that a common set of objectives for a wildlife monitoring programme were:

1. Status: From one survey
   (a) To estimate the population or sub–population total
   (b) To describe the distribution of the species over the area of interest

2. Trends: From a series of surveys through time:

   (a) To describe long–term trends or changes in the population, or sub–population, totals through time

   (b) To describe the change in the distribution of the species over the area of interest through time.

We have assumed in this thesis that the principle aims of the monitoring programme are to estimate $\tau^{(t)}$ and $\delta^{(t',t)}$, Obj. 1(a) and 1(b), as efficiently as possible within a design-based framework. We have also assumed that the population is motile and does not occur in clusters. The choice of monitoring strategy we implement will depend on other beliefs we have about the population, in particular how we think it changes through time.

Assuming for the moment that $s^{(1)}$ and $s_1^{(t)}$ are selected using *srswor* rather than *sys*, then the basic monitoring strategy will be of the form:

1. Survey 1:

   (a) Take a sample $s^{(1)}$ using *srswor* of n units from U

   (b) Develop a model $\zeta^{(1)}$ that uses the data in $s^{(1)}$

   (c) Predict $\hat{\mu}_i^{(2)}$ using the model $\zeta^{(1)}$

2. Survey 2:

   (a) Take a sample $s^{(2)}$, of which $n_1^{(2)}$ units are selected using *srswor*, from U or $s^{(1)}$, and $n_2^{(2)}$ are selected using $\pi p\hat{\mu}^{(1)}$, from $s_1^{(2)}$ or $s_c^{(1)}$

   (b) Develop a model $\zeta^{(2)}$ using the data from $s^{(2)}$, and possibly the data from $s^{(1)}$.

   (c) Predict $\hat{\mu}_i^{(3)}$

3. Survey $t > 2$:

   (a) Take a sample $s^{(t)}$, of which $n_1^{(t)}$ units are selected, using *srswor*, from U or $s_1^{(t-1)}$ and $n_2^{(t)}$ units are selected using $\pi p\hat{\mu}^{(t-1)}$ from $s_{1_c}^{(t)}$ or $s_{1_c}^{(t-1)}$

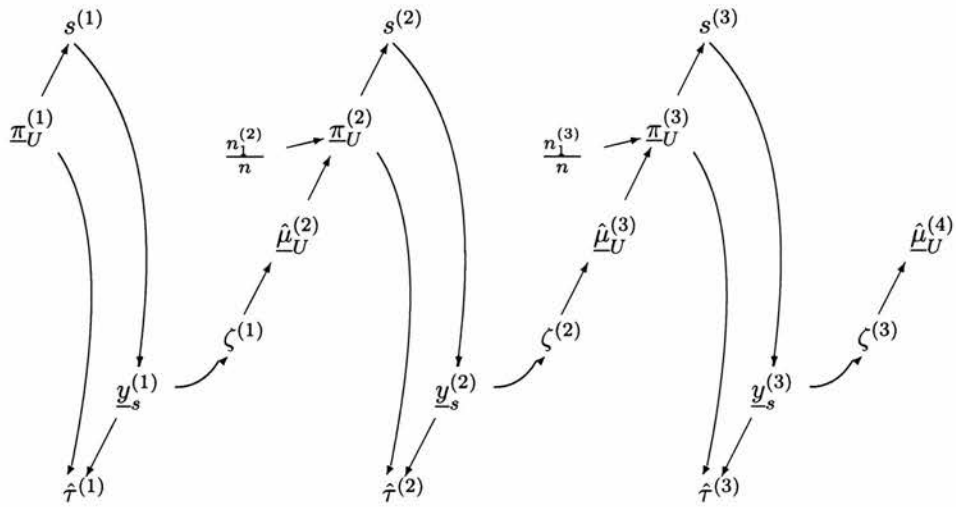


Figure 6.2: Representation of a monitoring strategy through time

(b) Develop model $\zeta^{(t)}$ that uses the data in $s^{(t)}$, and possibly the data from $s^{(t-1)}, \ldots, s^{(1)}$

(c) Predict $\hat{\mu}_i^{(t+1)}$

We illustrate the simplest scenario in Figure 6.2. The sample $s^{(1)}$ is selected from U. The sample selected depends on the inclusion probabilities $\pi_i^{(1)}$ for $i \in U$ which we denote $\underline{\pi}_U^{(1)}$. These inclusion probabilities and the observed data $y_i^{(1)}$ for $i \in s^{(1)}$, which we denote $\underline{y}_s^{(1)}$, are used to estimate $\hat{\tau}^{(1)}$. The sample data $\underline{y}_s^{(1)}$ are used to construct the model $\zeta^{(1)}$ from which an estimate of $\mu_i^{(2)}$ for $i \in U$, denoted $\underline{\hat{\mu}}_U^{(2)}$, can be obtained. Note that to construct the model $\zeta^{(1)}$, the auxiliary variables $\underline{x}_{s1}^{(1)}, \ldots, \underline{x}_{sQ}^{(1)}$ are required, and to estimate $\underline{\hat{\mu}}_U^{(2)}$, the auxiliary variables $\underline{x}_{U1}^{(1)}, \ldots, \underline{x}_{UQ}^{(1)}$ are needed. For simplicity, these are not included in the diagram. The probability of including unit i in $s^{(2)}$, $\pi_i^{(2)}$, depends on $\frac{n_2^{(2)}}{n}$, the proportion of the sample allocated to $s_2^{(2)}$, and $\hat{\mu}_i^{(2)}$. Given the inclusion probabilities, a sample $s^{(2)}$ is selected and the observed data $\underline{y}_s^{(2)}$ and the $\underline{\pi}_U^{(2)}$ are used to estimate $\hat{\tau}^{(2)}$. To construct

the model $\zeta^{(2)}$, only $\underline{y}_s^{(2)}$, the data from $s^{(2)}$, are used. These enable $\underline{\hat{\mu}}_U^{(3)}$ to be calculated which contribute to the inclusion probabilities $\underline{\pi}_U^{(3)}$ and so on ...

We can easily expand this model so that $\zeta^{(t)}$ is a function of the data obtained in surveys $t = 1, \ldots, t-1$ as well as from survey t.

A more complex representation is required when part of $s^{(t)}$ is selected from $s^{(t-1)}$. In this case $\pi_{i_1}^{(t)}$ is a function of $\frac{n_1}{n}$ and $\pi_{i_1}^{(t-1)}$, and $\pi_{i_2}^{(t)}$ depends on $\hat{\mu}_i^{(t-1)}$, $\pi_{i_1}^{(t-1)}$, $\hat{\mu}_i^{(t-1)}$, $s_{1_c}^{(t-1)}$ and $\frac{n_2}{n}$.

For a fixed n, and given that $s_1^{(t)}$ is selected using *srswor* or *sys*, the choices that need to be made about the sample design at time t are:

1. Which data are used to obtain the model $\zeta^{(t)}$, and what form does the model take?
2. Do we select $s_1^{(t)}$ from U or from $s_1^{(t')}$ for some $t' < t$?
3. What are the relative sizes of $s_1^{(t)}$ and $s_2^{(t)}$?

These decisions can be made at the start of the monitoring programme and applied to all surveys, or alternatively different decisions may be made for each survey as the relative importance of the objectives change. The decisions made depend partly on the belief about how the spatial distribution of species abundance changes over the survey region through time.

6.3 Issue 1: Estimating $\mu_i^{(t)}$

The aim of the modelling process is to draw inferences about the superpopulation model $\zeta^{(t)}$ so that an estimate of $\mu_i^{(t+1)}$ can be obtained for $i \in U$. The main use of $\hat{\mu}_i^{(t+1)}$, in our work, is to determine the inclusion probabilities for survey $t+1$, although interest in $\hat{\mu}_i^{(t)}$ to present a map of the species distribution, Obj. 1(b), may also be important. The estimate $\hat{\mu}_i^{(t+1)}$ is a good estimate of $\mu_i^{(t+1)}$, where $\hat{\mu}_i^{(t+1)} = a\mu_i^{(t+1)b}$, when b is close to one. An estimate $\tau^{(t+1)}$ will be more precise, using our combined sampling designs,

when b is close to one. Both the construction of the model $\zeta^{(t)}$ and the calculation of the predictions $\hat{\mu}_i^{(t+1)}$ are considered here.

Our general population model $\zeta^{(t)}$ is such that $E[Y_i^{(t)}] = \mu_i^{(t)}$ where

$$log(\mu_i^{(t)}) = \sum_{j=0}^{Q} f_j^{(t)}(x_{ij}^{(t)}) + g(t)$$

where $f_j^{(t)}$ is a linear or smooth function of the j^{th} auxiliary variable and $g(t)$ is a function, linear or smooth, of time. In section 2.3 we outlined some simple models that we might assume to describe $log(\mu_i^{(t)})$. These are given below, and the lack of superscript for $f^{(t)}$ or $x_{ij}^{(t)}$ indicates that these components remain fixed through time:

1. Constant population;
   $log(\mu_i^{(t)}) = \sum_{j=0}^{Q} f_j(x_{ij})$
2. Increasing or decreasing population at the same rate over the whole survey region;
   $log(\mu_i^{(t)}) = \sum_{j=0}^{Q} f_j(x_{ij}) + g(t)$
3. Varying change in the population over the survey region due to habitat change;
   $log(\mu_i^{(t)}) = \sum_{j=0}^{Q} f_j(x_{ij}^{(t)})$
4. Varying change in the population over the survey region due to changing relationship between abundance and habitat;
   $log(\mu_i^{(t)}) = \sum_{j=0}^{Q} f_j^{(t)}(x_{ij})$

These models are descriptions of how the population density changes over the survey region through time. They do not describe population processes, such as birth and death, rather they describe the variability in the species density in response to habitat. Models 1–3 assume that the functional relationship between species density and habitat remains constant through time. The change in species density over time may not be constant over the survey region because of habitat change in some areas, model 3. This does not necessarily mean that relative species density over the survey region remains constant through time. If there is habitat change over the survey region, the relative species density

will change. Because the functional relationship remains constant through time, we can use the data from surveys $1, \ldots, t$ to construct the model $\zeta^{(t)}$, and our knowledge about the relationship between species density and habitat increases through time, as long as there is data available about the changing habitat. Hence we would expect model mis-specification to decrease and our estimate of $\tau^{(t)}$ to become more precise through time.

Under model 4, the relationship between habitat and species density changes with time. This form of model is motivated by the effects of climate change on species as different species will respond at different rates to climate change depending on their characteristics. For example suppose that a motile species moves north as the climate becomes warmer. The ecosystem on which the species lives changes at a slower rate, and so the species moves north at a faster rate than the habitat and chooses sub-optimal habitat. Hence the functional relationship between habitat and species density changes through time. In a simple model, this may be expressed as

$$log(\mu_i^{(t)}) = \sum_{j=0}^{Q} \beta_j^{(t)}(x_{ij})$$

where $\beta_j^{(t)}$ is itself a stochastic process which must be modelled. For example, $\beta_j^{(t)}$ may depend on $\beta_j^{(t-1)}$. Unlike models 1–3 in which data from all surveys can contribute to the construction of $\zeta^{(t)}$, data from past surveys will contribute less to the construction of $\zeta^{(t)}$. In the extreme case when the change in the relationship between habitat and abundance is rapid, only the current survey t may be useful for constructing the model $\zeta^{(t)}$. As we learn less from previous surveys, we do not necessarily expect model mis-specification to reduce through time.

Unless $\frac{n_2^{(t)}}{n} = 0$ part of $s^{(t)}$ will be selected with unequal inclusion probabilities. In the literature, there is much discussion about whether estimation of the parameters of the superpopulation model, $\zeta^{(t)}$, should incorporate information about the survey design. A simple strategy for incorporating the sample design in the estimation process is to find weighted estimates of the parameters, where the weights are the inclusion probabilities; see section 4.5 when $\tau^{(2)}$ was estimated using a model-assisted approach. If $s^{(t)}$ was selected

using *strs*, then a variable identifying the stratum of each unit would be included as a candidate variable when constructing $\zeta^{(t)}$ although it may be removed if not significant. We do not consider this issue here but note that it may require further investigation.

To calculate the inclusion probabilities $\pi_i^{(t+1)}$, we use the model $\zeta^{(t)}$ to obtain an estimate of $\mu_i^{(t+1)}$. Consider first the case when we assume that the relationship between habitat and abundance remains constant from one survey to another. If our most current data on habitat are from time t, or if we believe that habitat has remained unchanged from time t to time $t+1$, then $\hat{\mu}_i^{(t)}$ and $\hat{\mu}_i^{(t+1)}$ will both be adequate estimates of $\mu_i^{(t+1)}$ as they only vary by a proportion, $\hat{\mu}_i^{(t+1)} = r\hat{\mu}_i^{(t)}$; the constant disappears when sampling $\pi p\hat{\mu}^{(t)}$ as $\pi_i^{(t)} = n\frac{\hat{\mu}_i^{(t)}}{\sum_U \hat{\mu}_i^{(t)}} = n\frac{\hat{\mu}_i^{(t+1)}}{\sum_U \hat{\mu}_i^{(t+1)}}$. If habitat has changed from survey t to survey $t+1$, then we can estimate $\hat{\mu}_i^{(t+1)}$ if data on the habitat at time $t+1$ are available. Else we can only estimate $\hat{\mu}_i^{(t)}$ and accept that there will be some errors. If we assume model 4, in which the relationship between habitat and abundance changes through time, then prediction of $\mu_i^{(t+1)}$ will require the functions $f_j^{(t+1)}$ to be specified. If this is not possible, then the best estimate of $\mu_i^{(t+1)}$ may be $\hat{\mu}_i^{(t)}$.

6.4 Issue 2: Is $s_1^{(t)}$ selected from U or $s_1^{(t')}$?

We retain units from survey t' to survey t because we wish to obtain a precise estimate of $\delta^{(t',t)}$ and we believe that the units $y_i^{(t')}$ and $y_i^{(t)}$ are correlated. If estimation of trend is not of great importance, then the extra complexity, both in the effort to relocate samples and in the estimation process, may indicate that this is not appropriate.

The retention of units can only be considered if it is possible to relocate units from one survey to another. As Global Positioning Systems (GPS) technology has become cheaper and more precise, point relocation becomes more practical, except in very dense habitats where GPS does not work well. In some cases it may be possible to leave a permanent marker which, once found, clearly indicates the location of the sampling unit. The precision with which the sampling unit must be relocated depends on the relative scale of the

sampling unit and the habitat characteristics that lead to high species density. For example the Bojers skink *Cryptoblephaus boutonii*, found on Round Island, is generally found in shade, such as under rocks or in and by shearwater burrows. Sampling units will tend to contain several such places and it is therefore important when returning in a future survey that the sample unit has exactly the same features, if they are still present, as the previous survey, so these units must be very carefully relocated, to the nearest centimetre say. When the size of the sampling unit is small compared to the habitat characteristics that determine species density, relocation does not need to be as accurate. For example if monitoring forest elephants, see Chapter 7, the accuracy of the relocation can be within metres rather than centimetres. Although it is clearly important to locate all sampling units as accurately as possible, whether units are retained from the previous survey or not, the precision of locating new samples (as distinct from revisiting sample locations) is not so important as long as there is no systematic bias in their location, because the issue of correlation between surveys does not arise.

In this thesis we have assumed that the cost, in terms of time and money, required to sample a unit that is retained from one survey to another is the same as the cost of selecting a new sampling unit. If the costs are different, because it costs time and effort to set up a permanent marker or because new transects have to be cut if an entirely new sample is taken, then this may become an additional consideration as to whether units are retained from one survey to another. If it costs more to retain units rather than select new units, the question arises of whether the increase in precision is considered worth the additional cost.

The reduction in the variance of $\hat{\delta}^{(t',t)}$ that occurs when units are retained from one survey to another, occurs because the units are assumed to retain the same characteristics from one survey to another so that $y_i^{(t')}$ and $y_i^{(t)}$ are correlated. If there is drift in the distribution of the species over the survey region, because of habitat or climate change as described by models 3 or 4 above, then the correlation $cor(y_i^{(t')}, y_i^{(t)})$ may not be high and so the advantage of sample retention is reduced. We note that retention of the units,

particularly when selected using *sys* so that there is coverage of the survey region, will enable descriptions of how the species is changing over the survey region, Obj. 2(b).

6.5 Issue 3: How do we determine the proportion $\frac{n_2^{(t)}}{n}$?

The proportion $\frac{n_2^{(t)}}{n}$ depends on the relative importance of the objectives, in particular the precision of $\hat{\tau}^{(t)}$ and $\delta^{(t',t)}$, our beliefs about how the population is changing through time, and how well $\hat{\mu}_i^{(t)}$ estimates $\mu_i^{(t)}$.

If trend estimation is of highest importance and it is possible to retain units from one survey to another, then decreasing $\frac{n_2^{(t)}}{n}$ will increase the precision of $\delta^{(t',t)}$, especially when $cor(y_i^{(t')}, y_i^{(t)})$ is high. If the relative species density changes through time, model 3 or model 4, then the correlation will be lower and some of the advantage of retaining units will be reduced. If $\frac{n_2^{(t)}}{n}$ remains low and the original units are selected using *sys*, the change in the distribution over the survey region can be observed.

Suppose the estimate of $\tau^{(t)}$ is also of interest. When the model $\zeta^{(t)}$ is well specified so that $\hat{\mu}_i^{(t)} = a\mu_i^{(t)b}$ and b is close to one, more precise estimates of $\tau^{(t)}$ can be obtained by setting $\frac{n_2^{(t)}}{n}$ to be close to one. Figure 6.3 is a repeat of figure 5.1 that demonstrates how the precision of $\hat{\tau}^{(t)}$ changes with b. If the model $\zeta^{(t)}$ is poorly specified, b is not close to one, then setting $\frac{n_2^{(t)}}{n}$ close to zero may give more precise estimates of $\tau^{(t)}$. This is slightly complicated however when units are not retained from one survey to another and if $0 < b < 1$. The most precise estimates are then obtained with $\frac{n_2^{(t)}}{n} = 1$. In practice the value b cannot be determined and so we cannot determine the optimal $\frac{n_2^{(t)}}{n}$. We might expect our model $\zeta^{(t)}$ to improve through time, especially if the model is of type 1–3, in which case a strategy where $\frac{n_2^{(t)}}{n}$ increases through time would be appropriate.

The choice of $\frac{n_2^{(t)}}{n}$ can also contribute to the precision of $\hat{\tau}^{(t+1)}$. Suppose that the model $\zeta^{(t)}$ is constructed using only the data from survey t. We discussed in section 3.5 how we expect our model $\zeta^{(t)}$ to be better specified if the data used to construct $\zeta^{(t)}$ provide good

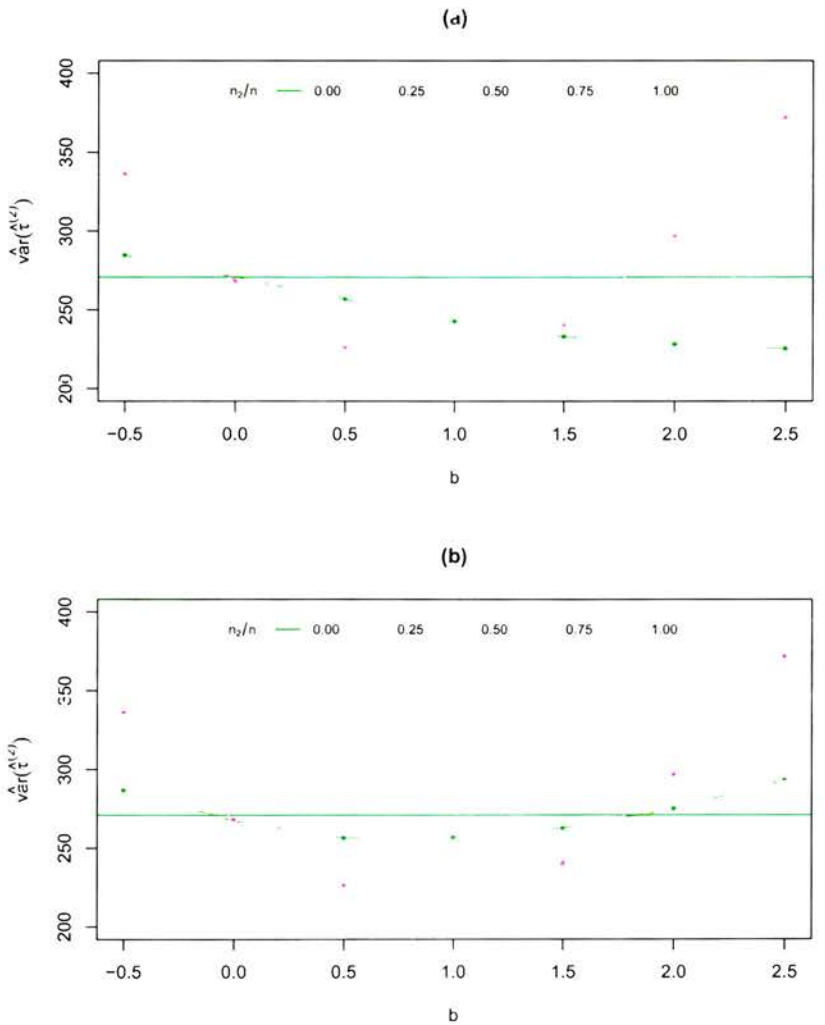


Figure 6.3: Effect of model mis-specification on $\sqrt{\widehat{var}(\hat{\tau}^{(2)})}$ for varying b and n_2 using strategy: (a) $comb\mu^b(\frac{n_2}{n})(U, s_1^{(2)})$ where $s_1^{(2)}$ is selected from U using *srswor* and $s_2^{(2)}$ is selected from $s_{1_c}^{(2)}$ using $\pi p\hat{\mu}^b$; (b) $comb\mu^b(\frac{n_2}{n})(s^{(1)}, s_c^{(1)})$ where $s_1^{(2)}$ is selected from $s^{(1)}$ using *srswor* and $s_2^{(2)}$ is selected from $s_c^{(1)}$ using $\pi p\hat{\mu}^b$

coverage of the X-space. If $\frac{n_2^{(t)}}{n}$ is close to or equal to one then most of $s^{(t)}$ will be selected using $\pi p\hat{\mu}^{(t)}$. Hence most of the units selected will have large values of $\hat{\mu}_i^{(t)}$. As $\hat{\mu}_i^{(t)}$ is a function of the auxiliary variables, then in the extreme case when only large values of $\hat{\mu}_i^{(t)}$ have been selected, the auxiliary variables for the units in the sample $s^{(t)}$ will take only a small range of values. Hence the model $\zeta^{(t)}$ will be constructed using data from only a small part of the X-space, which is likely to lead to a poor model. In addition for many $i \in U$, the prediction $\hat{\mu}_i^{(t+1)}$ will be an extrapolation from $\zeta^{(t)}$ and so we do not expect the model to be well specified; b will not be close to one. So setting $\frac{n_2^{(t)}}{n}$ high can lead to poor model specification of $\hat{\mu}_i^{(t+1)}$ and hence the precision of $\hat{\tau}^{(t+1)}$ will be poor (unless $\frac{n_2^{(t+1)}}{n}$ is decreased to close to zero). If $\frac{n_2^{(t)}}{n}$ is low, then we would expect greater coverage of the X-space, especially if $s_1^{(t)}$ is selected using *sys*, and we would expect model mis-specification to be low, with b close to one. Hence in scenarios where we assume that there is rapid change in the relationship between habitat and species density, there is a trade-off between high precision at time t obtained by setting $\frac{n_2^{(t)}}{n}$ close to one, and high precision at time $t+1$, obtained by setting $\frac{n_2^{(t)}}{n}$ to be close to zero.

A good model is more likely to be obtained when we have good coverage of the X-space. If $\zeta^{(t)}$ is constructed using data from several surveys, that is we assume a constant relationship between habitat and abundance, then by taking a combined sample in which $\frac{n_2^{(t)}}{n} < 1$, we would expect our model to improve through time as coverage of the X-space will increase. With an improved model, we could increase $\frac{n_2^{(t)}}{n}$ through time to lead to greater improvements in the precision of $\tau^{(t)}$. Hence when the relationship between habitat and species density remains constant, there is not the same trade-off between choosing whether the precision of $\hat{\tau}^{(t)}$ or $\hat{\tau}^{(t+1)}$ should dominate the selection of $\frac{n_2^{(t)}}{n}$.

When both $\tau^{(t)}$ and $\delta^{(t',t)}$ are of importance, there is some conflict. Suppose that we assume a constant relationship between habitat and abundance so that $\zeta^{(t)}$ can use data from all past surveys. If we wish $\delta^{(t',t)}$ to be precise, we would retain units from one survey to another. If $\tau^{(t)}$ is to be precise, we would want to take a new sample in each survey so that we obtain more information about the relationship between habitat and abundance;

$\hat{\mu}_i^{(t)}$ is better specified and so $\hat{\tau}^{(t)}$ is more precisely estimated. The combined sampling design will enable both retention of old units and selection of new units. One difficulty is that new units are most likely to be in areas of high abundance, as $s_2^{(t)}$ is selected with $\pi p \hat{\mu}_i^{(t)}$ and so there will be greater updating and information about units with high $\hat{\mu}_i^{(t)}$ than for the units selected using *sys* that give good coverage of the survey region.

6.6 Simulation

Using population P, in which $\mathcal{T}$ the expected number of individuals in the population at time t, has a Poisson distribution with mean Λ, three different monitoring programmes were considered for a period of $T = 10$ surveys. Sample 1, $s^{(1)}$, was selected using *sys* and for each of the 16 possible *sys* samples, 8 simulations were run, so that in total 128 simulations were run for each possible monitoring strategy.

The differences between the three monitoring strategies correspond to the number of surveys used to construct the model $\zeta^{(t)}$ and whether $s_1^{(t)}$ was selected from U or $s^{(1)}$ so that the three strategies are:

Str. 1. Select $s_1^{(t)}$ from U and $s_2^{(t)}$ from $s_{1_c}^{(t)}$ and construct $\zeta^{(t)}$ using the data from $s^{(t)}$ only;

Str. 2. Select $s_1^{(t)}$ from U and $s_2^{(t)}$ from $s_{1_c}^{(t)}$ and construct $\zeta^{(t)}$ using data from $s^{(1)}, \ldots, s^{(t)}$;

Str. 3. Select $s_1^{(t)}$ from $s^{(1)}$ and $s_2^{(t)}$ from $s_{1_c}^{(t-1)}$ and construct $\zeta^{(t)}$ using data from $s^{(1)}, \ldots, s^{(t)}$.

In each case $s_2^{(t)}$ was selected using $\pi p \hat{\mu}_i^{(t-1)}$ where $\hat{\mu}_i^{(t-1)}$ is predicted from model $\zeta^{(t-1)}$. This model was selected by initially fitting a GLM with all the auxiliary variables. Although we know that $\mu_i^{(t)}$ does not change through time we included time as a candidate variable in the construction of models under strategies 2 and 3. When $t \leq 4$ the variable was included as a factor and when $t > 4$ it was included as a continuous variable. The final model was selected using the automated step-wise procedure in R, where model selection is based on minimum AIC.

For each of these monitoring strategies, the proportion $\frac{n_2^{(t)}}{n}$ was fixed from survey 2 onwards. The monitoring strategy was repeated for $\frac{n_2^{(t)}}{n} = 0, 0.25, 0.5, 0.75$ and 1 using the survey designs shown in figure 6.1. The estimates of $\tau^{(t)}$ for $t = 1, \ldots, 10$ and $\delta^{(1,10)}$ were calculated using the appropriate estimators, as described in Chapters 4 and 5. For strategy 3 when $t > 2$ the covariance $cov(\hat{\tau}^{(1)}, \hat{\tau}^{(t)})$ is estimated using equation 5.33.

The first monitoring strategy represents the extreme case when there is believed to be rapid change in the relationship between habitat and species density so that $\zeta^{(t)}$ is constructed using the data from $s^{(t)}$ only. Figure 6.4 gives a summary of $\hat{\tau}^{(t)}$ over the 128 simulations for all 10 surveys for each of the five values of $\frac{n_2^{(t)}}{n}$. Figure 6.4(b) is a summary of the estimated variance, $\widehat{var}(\hat{\tau}^{(t)})$. These results indicate that when $\frac{n_2^{(t)}}{n} = 1$, that is $s^{(t)}$ is selected using $\pi p \hat{\mu}_i^{(t-1)}$ the variability of the estimates increases through time. This is because of model mis-specification. Figure 6.5 shows the mean values of model mis-specification for each survey under varying $\frac{n_2^{(t)}}{n}$. When $\frac{n_2^{(t)}}{n}$ is high, only a small part of the X-space can be used to construct the model $\zeta^{(t)}$ and hence it is poorly specified. The reason for the high variability in the b is that, once the model starts being misspecified, the wrong part of the survey region is sampled in the following survey, and hence the mis-specification increases. Note that as soon as $\frac{n_2^{(t)}}{n}$ is not equal to one, model mis-specification reduces because there is at least some coverage of the X-space. The effect of model mis-specification on the estimates is considerably less because the combined sampling designs are relatively robust to model mis-specification. When $\frac{n_2^{(t)}}{n} = 1$ the estimates are biased because the distribution of the $\hat{\tau}^{(t)}$ has positive skew. The median and interquartile ranges for these estimates are also given.

The estimates of $var(\hat{\tau}^{(t)})$ are lowest when $\frac{n_2^{(t)}}{n} = 0.5$ or $\frac{n_2^{(t)}}{n} = 0.75$. Because $s_1^{(t)}$ is selected using *sys*, the analytic estimate of $var(\hat{\tau}^{(t)})$ when $\frac{n_2^{(t)}}{n} = 0$ overestimates $var(\hat{\tau}^{(t)})$, as $\widehat{var}(\hat{\tau}^{(t)})$ has been estimated assuming $s^{(t)}$ was selected using *srswor*. In some cases, the empirical estimate of the standard error when $\frac{n_2^{(t)}}{n} = 0$, shown by the size of the bars on the estimate of $\hat{\tau}^{(t)}$, is smaller than the empirical estimate for our combined sampling strategies when $\frac{n_2^{(t)}}{n} > 0$. This suggests that the strategy in which $\frac{n_2^{(t)}}{n} = 0$ gives a more

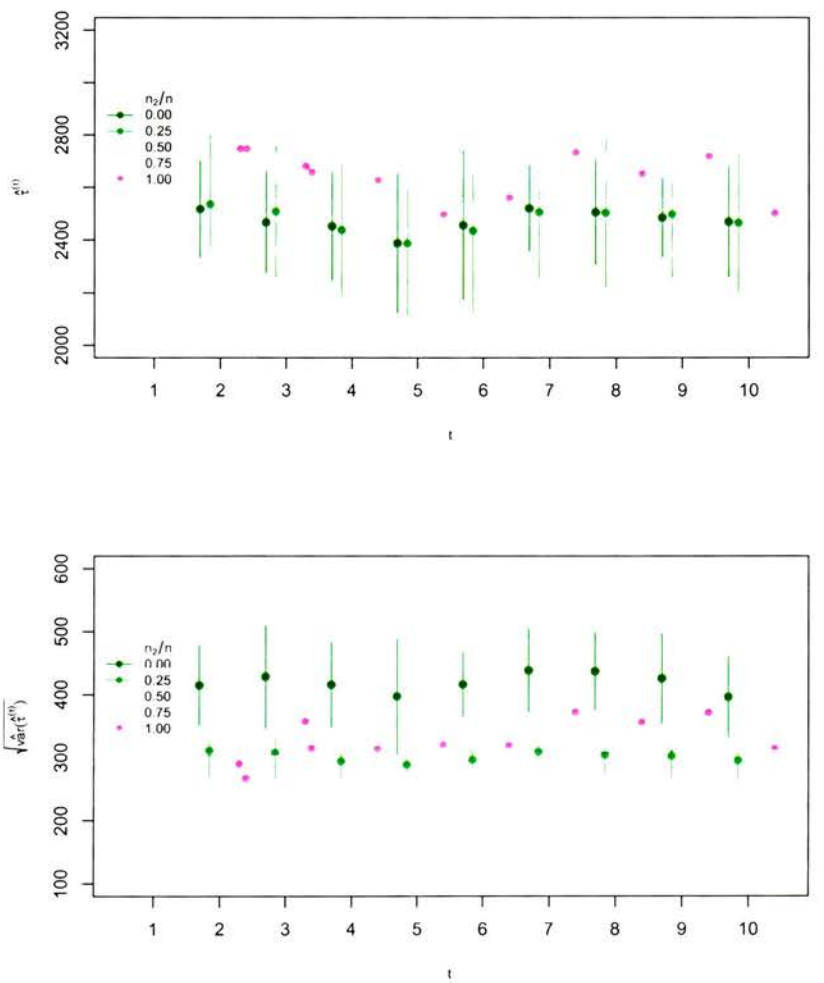


Figure 6.4: Mean • (± s.d –)of (a) $\hat{\tau}^{(t)}$ (b) $\sqrt{\widehat{var}(\hat{\tau}^{(t)})}$ for surveys $t = 2, \ldots, 10$ over 128 simulations using monitoring strategy 1. $s_1^{(t)}$ is selected from U using *srswor* and $s_2^{(t)}$ is selected from $s_{1_c}^{(t)}$ using $\pi p \hat{\mu}_i^{(t-1)}$ where $\hat{\mu}_i^{(t-1)}$ is estimated using the data from $s^{(t-1)}$ only. When $\frac{n_2}{n} = 1.00$ the results for $t > 3$ have positive skew. Dashed lines represent lower and upper quartile range and • is the median

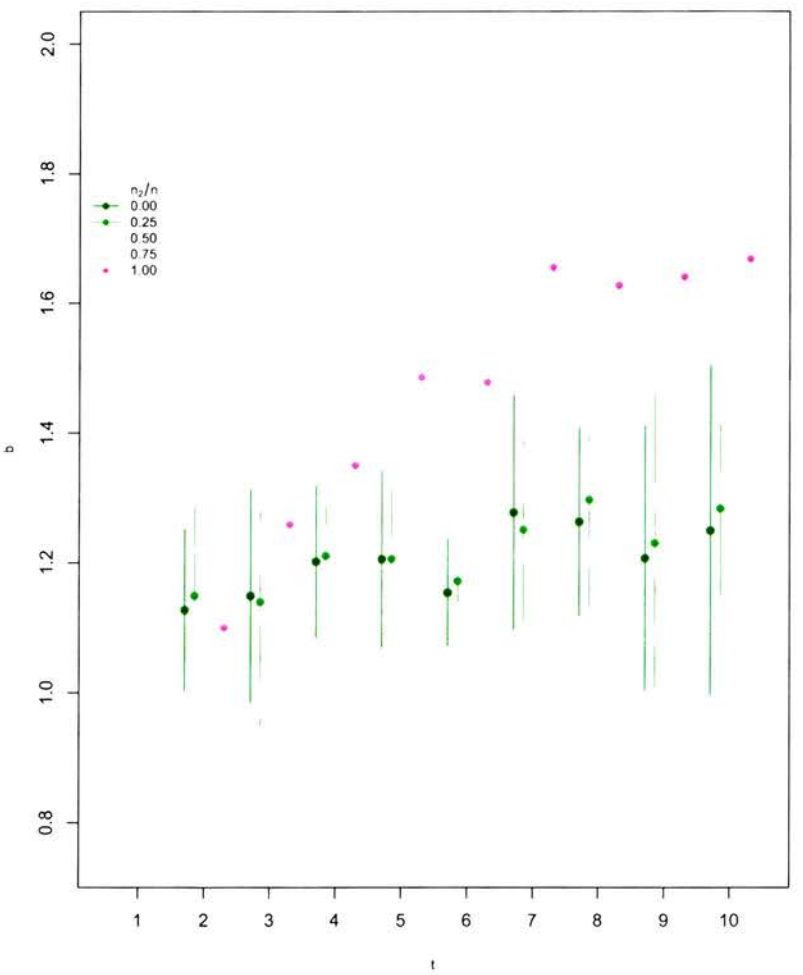


Figure 6.5: Mean • (± s.d –)of b using where $\hat{\mu}_i^{(t)} = a\mu_i^{(b)}$ over 128 simulations using monitoring strategy 1. $s_1^{(t)}$ is selected from $s^{(t-1)}$ and $s_2^{(t)}$ is selected from $s_{1_c}^{(t-1)}$ using $\pi p\hat{\mu}^{(t-1)}$ where $\hat{\mu}^{(t-1)}$ is estimated using the data from $s^{(t-1)}$ only.

precise estimate of $\hat{\tau}^{(t)}$ than the combined sampling strategies in which $\frac{n_2^{(t)}}{n} > 0$. Caution must therefore be exercised if the *srswor* variances are used as these overestimate the true variances and hence indicate that a strategy other than *sys* is optimal when this may not be the case in practice. However this also suggests that the correct estimate of the variance must be obtained if the sample is selected using *sys* for this to be observed in practice.

When we assume a constant relationship between habitat and abundance through time, we can employ the second monitoring strategy in which $\zeta^{(t)}$ is constructed using all past survey data. Figure 6.6 summarises the estimates of $\tau^{(t)}$ and $var(\hat{\tau}^{(t)})$ for $t = 1, \ldots, T$ and the different values of $\frac{n_2^{(t)}}{n}$. Because we can learn from one survey to another, we see that even the strategy in which $\frac{n_2^{(t)}}{n} = 1$ it is possible to obtain precise estimates of $\hat{\tau}^{(t)}$, although the strategy in which $\frac{n_2^{(t)}}{n} = 0.75$ gives the most precise estimates of $\tau^{(t)}$ as it is most robust to model mis-specification.

We might expect an increase in the precision of $\hat{\tau}^{(t)}$ through time if the model is initially poorly specified, possibly because the population is sparse and so there is only a small amount of information about the relationship between habitat and density. Then if $\hat{\mu}_i^{(t)}$ improves through time, we would expect to see this reflected in the precision of $\hat{\tau}^{(t)}$. However we note that there does not seem to be an obvious decrease in $\widehat{var}(\hat{\tau}^{(t)})$ over time for any of the combined strategies. This is because on average $\hat{\mu}_i^{(2)}$ is already a reasonably good estimate of $\mu_i^{(2)}$ and so there is little scope for improvement.

As the model seems reasonably well specified even after the first survey, the results for the monitoring strategy in which only the data in $s^{(t)}$ are used to construct the model ζ may give a false indication of how well the combined sampling designs perform when $\frac{n_2^{(t)}}{n}$ is high, but less than one. If data are sparse, then a sampling design in which $\frac{n_2^{(t)}}{n}$ is relatively low may be a more cautious and successful strategy.

Figure 6.7 summarises the estimates of $\tau^{(t)}$ and $\sqrt{var(\hat{\tau}^{(t)})}$ for $t = 1, \ldots, T$ and the different values of $\frac{n_2^{(t)}}{n}$ using the third monitoring strategy, where $s_1^{(t)}$ is selected from

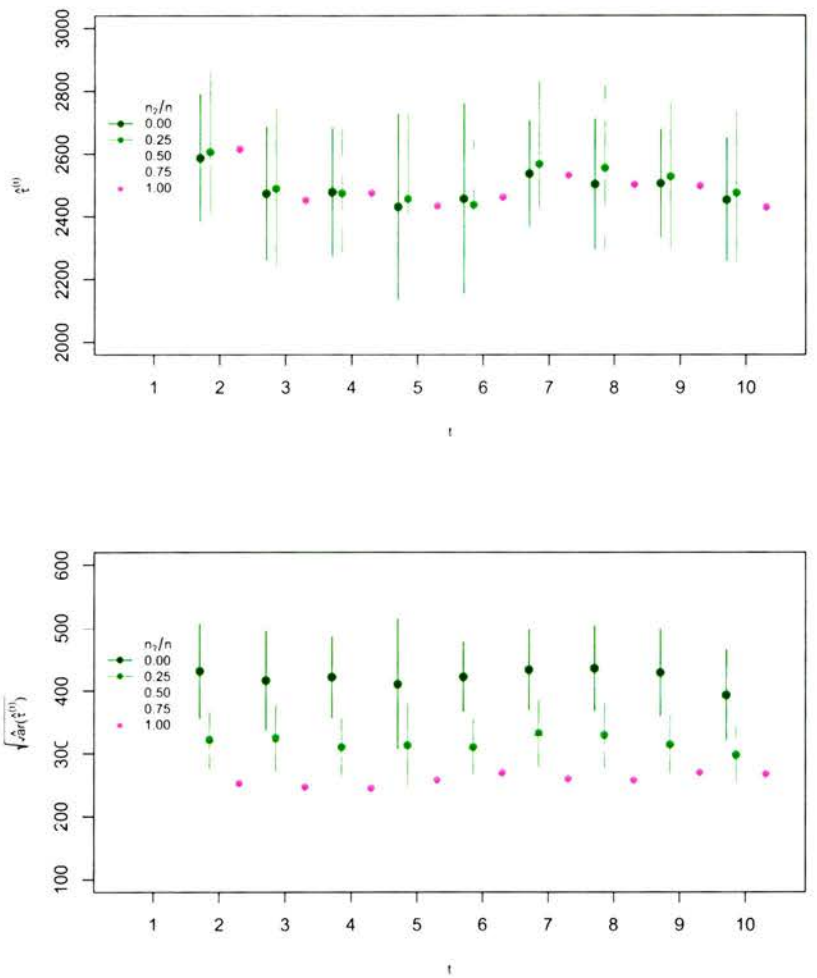


Figure 6.6: Mean • ($\pm$ s.d –)of (a) $\hat{\tau}^{(t)}$ (b) $\sqrt{\widehat{var}(\hat{\tau}^{(t)})}$ for surveys $t = 2, \ldots, 10$ over 128 simulations using monitoring strategy 2. $s_1^{(t)}$ is selected from U using *srswor* and $s_2^{(t)}$ is selected from $s_{1c}^{(t)}$ using $\pi p \hat{\mu}_i^{(t-1)}$ where $\hat{\mu}_i^{(t-1)}$ is estimated using the data from $s^{(1)}, \ldots, s^{(t-1)}$.

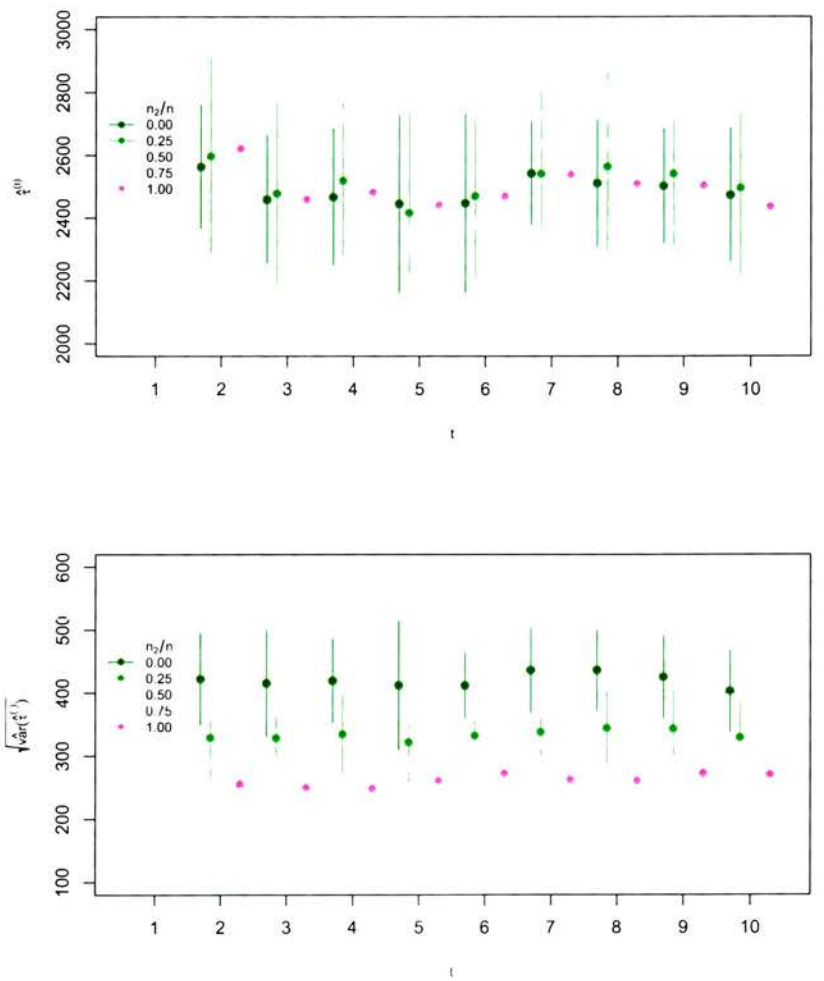


Figure 6.7: Mean • (± s.d –)of (a) $\hat{\tau}^{(t)}$ (b) $\sqrt{\widehat{var}(\hat{\tau}^{(t)})}$ for surveys $t = 2, \ldots, 10$ over 128 simulations using monitoring strategy 3. $s_1^{(t)}$ is selected from $s_1^{(t-1)}$ using *srswor* and $s_2^{(t)}$ is selected from $s_{1_c}^{(t-1)}$ using $\pi p \hat{\mu}_i^{(t-1)}$ where $\hat{\mu}_i^{(t-1)}$ is estimated using the data from $s^{(1)}, \ldots, s^{(t-1)}$.

$s_1^{(t-1)}$. Here we see a similar pattern to that of figure 6.6, when no units were retained from $s^{(1)}$, as in both cases the model improves through time. As expected, from the behaviour of the estimator under model mis-specification, the precision of $\tau^{(t)}$ when $\frac{n_2^{(t)}}{n} < 1$ is less than when $\frac{n_2^{(t)}}{n} = 1$.

Figure 6.8(a) gives the estimates of $\delta^{(1,10)}$ for these two monitoring strategies and figure 6.8(b) the estimate of $var(\hat{\delta}^{(1,10)})$. We see that the precision of $\delta^{(t',t)}$ is much greater when a large proportion of the sample is retained from one survey to another. As the proportion decreases so the difference in the precision of $\hat{\delta}^{(t',t)}$, between strategies in which $s^{(t-1)}$ is retained from $s^{(1)}$ and when it is not retained, decreases. Note that the precision of $\delta^{(t',t)}$ depends on the correlation $cor(y_i^{(t')}, y_i^{(t)})$ and the precision of $\tau^{(t)}$ depends on the correlation $cor(\hat{\mu}_i^{(t)}, y_i^{(t)})$.

Finally in figure 6.9 we summarise the average number of individuals observed when $s_1^{(t)}$ is selected from U using sys and $s_2^{(t)}$ is selected from $s_{1_c}^{(t)}$ using $\pi p \hat{\mu}_i^{(t-1)}$ and $\hat{\mu}_i^{(t-1)}$ is estimated using the data from all previous surveys. Although not directly stated as an aim of the monitoring programme, an additional motivation for these methods was to increase the number of observations of individuals in the population. We see that as $\frac{n_2^{(t)}}{n}$ increases so the number of individuals observed increases. Under the strategy in which $\frac{n_2^{(t)}}{n} = 1$ twice as many individuals are observed as when $\frac{n_2^{(t)}}{n} = 0$. This is as we expect, as explained in section 4.6 as the distribution of the $y_i^{(t)}$ is positively skewed.

6.7 Discussion

The aim of this chapter was to illustrate how the sampling designs we have proposed can be implemented in a monitoring programme for a motile species. The three key issues are which data are used to construct the model $\zeta^{(t)}$, whether units are retained from one survey to another and the proportion $\frac{n_2^{(t)}}{n}$.

When only $\tau^{(t)}$ is of interest, and so $s_1^{(t)}$ is selected from U, we see that the application of

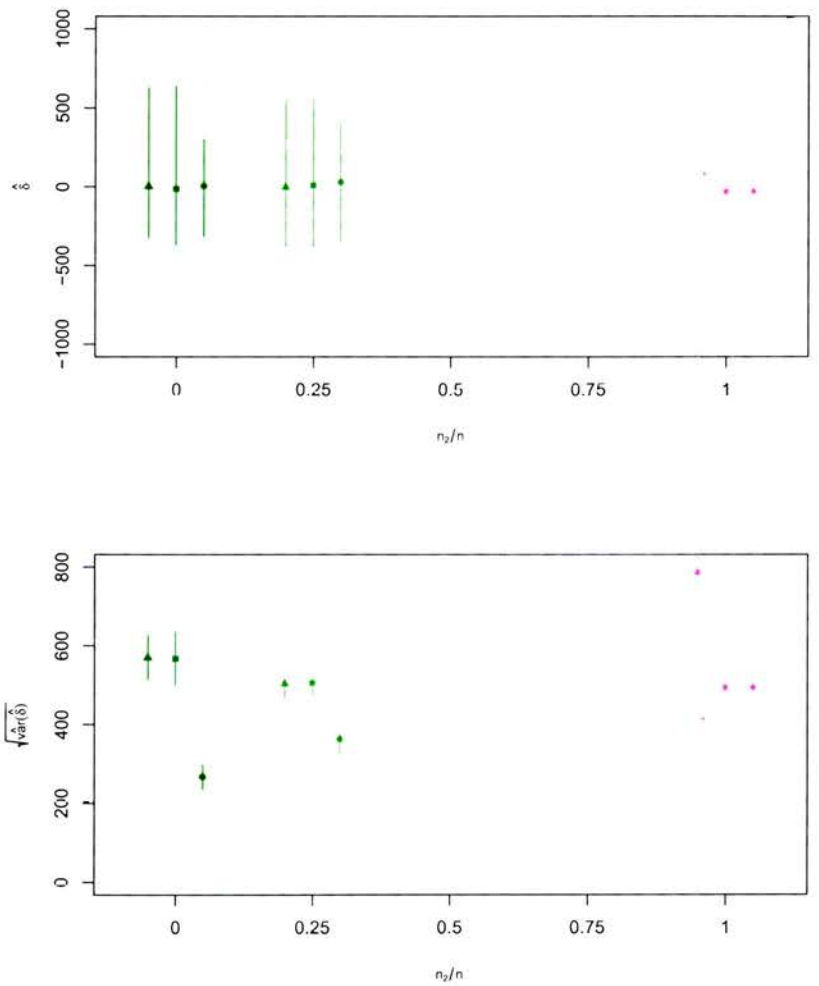


Figure 6.8: Mean ($\pm$ s.d) of (a) $\delta^{(1,10)}$ (b) $\sqrt{\widehat{var}(\hat{\delta}^{(1,10)})}$ over 128 simulations.

▲ Monitoring strategy 1, $s_1^{(t)}$ is selected from U using *srswor* and $s_2^{(t)}$ is selected from $s_{1_c}^{(t)}$ using $\pi p\hat{\mu}_i^{(t)}$ where $\hat{\mu}_i^{(t)}$ is estimated using the data from $s^{(t-1)}$ and $\triangle$ shows the median (quartile).

■ Monitoring strategy 2, $s_1^{(t)}$ is selected from U using *srswor* and $s_2^{(t)}$ is selected from $s_{1_c}^{(t)}$ using $\pi p\hat{\mu}_i^{(t)}$ where $\hat{\mu}_i^{(t)}$ is estimated using the data from $s^{(1)}, \ldots, s^{(t-1)}$.

• Monitoring strategy 3, $s_1^{(t)}$ is selected from $s_1^{(t-1)}$ using *srswor* and $s_2^{(t)}$ is selected from $s_{1_c}^{(t-1)}$ using $\pi p\hat{\mu}^{(t)}$ where $\hat{\mu}_i^{(t)}$ is estimated using data from $s^{(t-1)}, \ldots, s^{(1)}$.

the combined sampling strategy gives as precise results as sampling with $\pi p\hat{\mu}^{(t)}$ when we can learn through time about the species distribution. When we can only learn from the last survey the combined sampling strategy does considerably better than sampling with $\pi p\hat{\mu}^{(t)}$, due to the robustness of the strategy to model mis-specification. In the population we have used here, the relationship between habitat and abundance remains constant through time and data from just one survey can give a good estimate $\hat{\mu}_i^{(t)}$, that is one in which model mis-specification is reasonably low. Then a combined sampling strategy in which $\frac{n_2^{(2)}}{n}$ is high can be used in survey 2 to obtain precise estimates of $\tau^{(2)}$. When the population is sparse we would not expect data from survey 1 to give a good estimate of $\mu_i^{(1)}$, so that in survey 2 a more cautious approach would be required to avoid giving imprecise estimates of $\tau^{(2)}$. Similarly when data from only the last survey can be used to construct $\zeta^{(t)}$, we would expect to use the more conservative designs of $\frac{n_2^{(t)}}{n} = 0.25$ than when data from all surveys can be used to construct $\zeta^{(t)}$. Therefore the less we can learn about the species density over the survey region from previous surveys, the less improvement we expect in the precision of $\hat{\tau}^{(t)}$. This implies that when the data from several surveys can be used to construct $\zeta^{(t)}$, we might increase $\frac{n_2^{(t)}}{n}$ through time from zero to one as our knowledge of the species density improves so that the precision of $\hat{\tau}^{(t)}$ can increase in precision.

When $\tau^{(t)}$ and $\delta^{(1,10)}$ are both of interest then the decision about $\frac{n_2^{(t)}}{n}$ depends on which of these two parameters is of most importance. As $\frac{n_2^{(t)}}{n}$ increases so the precision of $\hat{\tau}^{(t)}$ increases and the precision of $\delta^{(1,10)}$ decreases, when the $y_i^{(1)}$ and $y_i^{(10)}$ are correlated. In particular we see that when $\frac{n_2^{(t)}}{n} \geq 0.5$ the precision of $\delta^{(t',t)}$ is not much different to if $\frac{n_2^{(t)}}{n} = 1$.

Given the varying objectives of a monitoring programme, we would suggest that it is possible to change the proportion $\frac{n_2^{(t)}}{n}$ and whether units are retained throughout the monitoring programme to meet the most pressing objectives. In particular the purpose of individual surveys can be determined with the whole monitoring programme in mind. For example some surveys could be dedicated to obtaining an improved estimate of $\hat{\mu}_i^{(t)}$,

so that an entirely new sample is selected using *sys* or *srswor*, or indeed an adaptation of our design in which $s_1^{(t)}$ is selected using *sys* and $s_2^{(t)}$ is selected using *srswor*. In some other surveys, the objective may be to obtain precise estimates of $\hat{\tau}^{(t)}$ so that a combined sampling design with a high $\frac{n_2^{(t)}}{n}$ may be selected. Other surveys may retain all of the units from the first survey so that a good estimate of $\delta^{(1,t)}$ can be obtained. This type of strategy may be particularly desirable when the population is quite sparse. In this case, data from several surveys may be required to obtain a reasonable estimate of $\hat{\mu}_i^{(t)}$. Similarly if the population is declining, as described in model 2, it would be important to obtain as much information as possible about the spatial distribution of the species at the start of the monitoring programme as over time

Further work is required to explore how different monitoring strategies could operate for populations with different behaviours and for different objectives. We suggest that our combined sampling designs provide a framework from which these more complicated monitoring strategies can be developed.

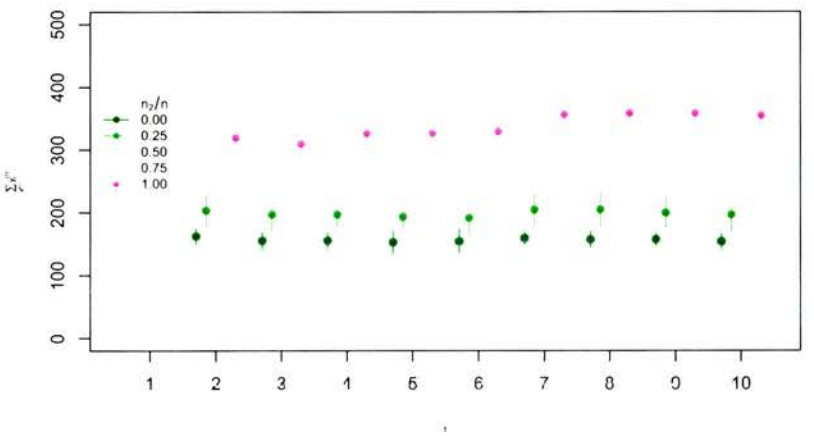


Figure 6.9: Mean (± s.d –) of $\sum_s y_i^{(t)}$ for varying $\frac{n_2}{n}$ using monitoring strategy 2. $s_1^{(t)}$ is selected from U using *srswor* and $s_2^{(t)}$ is selected from $s_{1_c}^{(t)}$ using $\pi p \hat{\mu}_i^{(t-1)}$ where $\hat{\mu}_i^{(t-1)}$ is estimated using the data from $s^{(1)}, \ldots, s^{(t-1)}$

Chapter 7

A case study: Monitoring forest elephants in Central Africa

In this thesis we have applied our sampling strategies to simulated data. It is of interest to see how the methods can be applied to a set of real data. Ideally we would like census data from several surveys for a motile species. Alternatively the methods could be used within a monitoring programme and the difficulties and results presented here. Neither of these were possible, so instead we propose a sample design for a second survey, based on data observed from a first survey. The case study selected was a motivating example for the thesis.

Although the strategies developed in this thesis were based on plot sampling, in which the probability of detection is one, there is no reason why the methods cannot be applied to problems in which there is imperfect detectability; for example for surveys that use distance sampling. This survey method is particularly appropriate for motile species.

The example given here is part of an international monitoring programme to monitor the

illegal killing of elephants (MIKE). The background to this monitoring programme and its objectives are given in section 7.1. The scale of the monitoring programme means that much of the monitoring is carried out at a number of selected sites. One component of the monitoring programme is to estimate the numbers of elephants at each site, so that the change in elephant numbers through time can be estimated. The case study we present here is based on a pilot study to estimate the number of forest elephants, *Loxodonta africana cyclotis*, in Odzala, a national park in Central Africa. As forest elephants are elusive, population estimation is based on an estimate of dung density, obtained using distance sampling. In section 7.2, we describe how the standard distance sampling estimator, in which it is assumed that all units are selected using *srswor*, can be extended to incorporate our combined sampling designs in which units are selected using unequal probability sampling. In section 7.3, the pilot study and its results are described and a map of predicted abundance obtained. Using this map, we suggest a possible sample scheme for the second survey in section 7.4.

7.1 Background

The Convention on International Trade in Endangered Species of wild fauna and flora (CITES) is an international agreement between governments to regulate the international trade in wildlife and wildlife products of endangered species. It works by allowing trade of these species only under permit. The species protected by CITES are listed on one of three Appendices, depending on the level of protection required. Species on Appendix I are threatened with extinction and trade in specimens is only permitted in exceptional circumstances. Species on Appendix II are not necessarily threatened with extinction, but trade is controlled to ensure that harvesting is sustainable. Species on Appendix III are those in which a country already controls trade in the species, but requires cooperation from other countries. Some species may occur on more than one Appendix depending on their status in different countries.

Information about the species is obtained from many sources. Although data are collected to determine the listing of a particular species, the impact of decisions made by CITES (at the biennial Conference of the Parties (CoP)) on the suspension or commencement of trade of particular species has not, as yet, been assessed in a systematic or detailed manner.

In 1990 all elephants (African, *Loxodonta africana,* and Asian, *Elephus maximus*) were listed on Appendix I. At the 10^{th} meeting of the Conference of the Parties to CITES (CoP10), resolution 10.10 was passed that allowed the commencement of trade in ivory from Botswana, Zimbabwe and Namibia under tightly regulated controls. *L. africana* is listed on Appendix II in these countries. This was, and is, a controversial decision and so a condition of this resolution was that a monitoring system would be put in place across Africa and Asia to facilitate decision-making regarding the the protected status of all elephants.

At CoP11, South Africa was also added to the list of countries in which tightly controlled legal trade was permitted. In addition, the monitoring system known as MIKE (Monitoring the Illegal Killing of Elephants) was given a broader range of objectives as quoted in the revised version of Resolution 10.10:

(i) measuring and recording levels and trends, and changes in levels and trends of illegal hunting and trade in ivory in elephant range States, and in trade entrepots;

(ii) assessing whether and to what extent observed trends are related to changes in the listing of elephant populations in the CITES appendices and/or the resumption of legal international trade in ivory;

(iii) establishing an information base to support the making of decisions on appropriate management, protection and enforcement needs; and

(iv) building capacity in range States.

The decision was taken that these objectives should be met by a site-based system. A

set of 60 sites was originally selected from the four regions of Africa (Eastern, Central, Western and Southern) and from Asia (IUCN/SSC *et al.*, 1999). Many of these sites are protected national parks, although in many sites, part of the area outside the national park is also to be monitored.

Many different sources of data are required to meet the objectives of the monitoring programme. The key data requirements are; elephant population numbers; mortality rates, both natural and due to illegal killing; measures of protection and law enforcement effort; socio-economic and socio-political data such as civil strife and community involvement in conservation. The data analysis strategy (Burn and Underwood, 2003) proposes a strategy for analysis and integration of these different data sources to meet the MIKE objectives.

One component of the MIKE programme is a series of population surveys at each site; these are Type II studies using our definition in Chapter 2. The objectives of these surveys are to obtain estimates of population totals, $\hat{\tau}^{(t)}$, and trend, η_1, through time. The case study described in this chapter is based on a population survey at one site, Odzala (Congo) in Central Africa, a forested site where the forest elephant is found. The first survey was part of a pilot study (a Type I study) carried out by the Wildlife Conservation Society on behalf of CITES to test and evaluate protocols for population surveys of forest elephants. A description of the pilot study, results and conclusions are given in Beyers *et al.* (2001)

7.2 Estimating elephant density

Forest elephants are difficult to observe and the standard method of estimating forest elephant abundance is to count elephant dung-piles so that the total number of elephants in the population, τ is estimated as

$$\hat{\tau} = A\frac{\hat{D}}{\hat{\theta}\hat{\kappa}} \tag{7.1}$$

$$\widehat{var}(\hat{\tau}) = A^2\hat{D}^2\left[\frac{\widehat{var}(\hat{D})}{\hat{D}^2} + \frac{\widehat{var}(\hat{\theta})}{\hat{\theta}^2} + \frac{\widehat{var}(\hat{\kappa})}{\hat{\kappa}^2}\right] \tag{7.2}$$

where $\hat{D}$ is the estimated density of dung-piles over the survey region, of area A, $\hat{\theta}$ is the estimated mean time a dung-pile takes to decay, and $\hat{\kappa}$ is the estimated daily production of dung-piles by one animal. The parameters θ and κ are to be estimated from separate studies. The current proposed method for estimating $\hat{\theta}$ is that of Laing *et al.* (2003). An estimate $\kappa = 17$ (Wing and Buss, 1970) is commonly used.

Distance sampling is used to obtain the estimate of D, as counting dung-piles in thick forest can be difficult. Distance sampling requires transects to be cut on a compass bearing. The observer walks along the transect and records any dung-piles observed. For the k^{th} dung-pile, the perpendicular distance from the centre of the dung-pile to the transect line, u_k, is recorded. The aim is to detect all dung-piles that are close to and on the line and it is expected that fewer dung-piles will be detected as the perpendicular distance from the line increases.

The survey region is divided into N strips each of width $2w$ and length l. Using some probability design, a sample s of n transects is selected so that π_i is the probability that the i^{th} strip is included in the sample. The line transect runs down the centre of the strip. We define y_i to be the number of dung-piles observed in the i^{th} transect.

Under the simplest scenario we assume that:

1. Each transect is selected using *srswor*;
2. The probability of detecting a dung-pile depends only on the distance from the transect line;
3. All dung-piles on the line are detected and distances recorded accurately.

Using the standard distance sampling estimator, the density of dung-piles in the survey region will be estimated as

$$\hat{D} = \frac{\sum_s y_i}{2wnl\hat{P}_a} \tag{7.3}$$

where $2wnl$ is the total area surveyed and P_a is the proportion of dung-piles that is detected within the survey strips. P_a is estimated using the observed distances u_k for $k = 1, \ldots \sum_s y_i$ dung-piles. If $g(u)$ is the probability that a dung-pile is detected at distance u from the transect line, then

$$P_a = \frac{\int_0^w g(u)du}{w}$$

so that

$$\hat{D} = \frac{\sum_s y_i}{2nl \int_0^w \hat{g}(u)du} \tag{7.4}$$

To estimate $g(u)$, we define $f(u)$, the probability density function of the u_k, as

$$f(u) = \frac{g(u)}{\int_0^w g(u)du} \tag{7.5}$$

As perfect detection is assumed on the transect line, $g(0) = 1$ and

$$f(0) = \frac{1}{\int_0^w g(u)du} \tag{7.6}$$

so P_a is estimated by finding the p.d.f of the u_k so that

$$\hat{P}_a = \frac{1}{w\hat{f}(0)} \tag{7.7}$$

$$\text{and } \hat{D} = \frac{\hat{f}(0)\sum_s y_i}{2nl} \tag{7.8}$$

Further details of how to estimate $f(0)$ are given in Buckland *et al.* (2001). An estimate of the variance of $\hat{D}$ is

$$\widehat{var}(\hat{D}) = \hat{D}^2 . \left\{ \frac{\widehat{var}\left[\sum_s y_i\right]}{\left(\sum_s y_i\right)^2} + \frac{\widehat{var}\left[\hat{f}(0)\right]}{\left(\hat{f}(0)\right)^2} \right\} \tag{7.9}$$

Now suppose that transects are selected using an unequal probability design $p(.)$ where π_i is the probability that the i^{th} unit is included in the sample. If we assumed perfect detection, we would estimate the density of dung-piles in the survey region to be the estimated total number of dung-piles in the survey region, $\sum_s \frac{y_i}{\pi_i}$ divided by the area of the survey region, $A = 2wNl$ so that

$$\hat{D} = \frac{1}{2wNl} \sum_s \frac{y_i}{\pi_i} = \frac{\hat{\tau}}{2wNl} \tag{7.10}$$

Note that when transects are selected using *srswor* this estimator is equivalent to equation 7.3 as we have assumed perfect detectability so that $\hat{P}_a = 1$. An estimate of the variance is

$$\widehat{var}(\hat{D}) = \hat{D}^2 . \left\{ \frac{\widehat{var}(\hat{\tau})}{\hat{\tau}^2} \right\} \tag{7.11}$$

where $\widehat{var}(\hat{\tau})$ is an estimate of $var(\hat{\tau})$ such as the Sen-Yates-Grundy variance estimator. This represents the variability in the encounter rate, adjusted for the probability of selecting the transect lines. With imperfect detection, when we assume that the probability of detection is a function of distance from the line only,

$$\hat{D} = \frac{\hat{f}(0)}{2Nl} \sum_s \frac{y_i}{\pi_i} = \frac{\hat{f}(0)}{2Nl} \hat{\tau} \tag{7.12}$$

Note that when $\pi_i = \frac{n}{N}$, this gives equation 7.8. The variance is estimated as

$$\begin{aligned} \widehat{var}(\hat{D}) =& \hat{D}^2 . \left\{ \frac{\widehat{var}\left[\sum_s \frac{y_i}{\pi_i}\right]}{\left(\sum_s \frac{y_i}{\pi_i}\right)^2} + \frac{\widehat{var}\left[\hat{f}(0)\right]}{\left(\hat{f}(0)\right)^2} \right\} \\ =& \hat{D}^2 . \left\{ \frac{\widehat{var}(\hat{\tau})}{\hat{\tau}^2} + \frac{\widehat{var}\left[\hat{f}(0)\right]}{\left(\hat{f}(0)\right)^2} \right\} \end{aligned} \tag{7.13}$$

These estimators only hold when we assume that the probability of detecting an individual dung-pile depends on the distance from the transect line and not on any other variables. Hence if we select our sample using a combined sampling design, $\widehat{var}(\hat{\tau}^{(t)})$ would be the relevant variance estimator reflecting whether part of the sample has been retained from

a previous survey or not. From the work in this thesis, we would expect the precision of $\hat{D}$ to increase under a combined sampling strategy compared to selecting $s^{(2)}$ using *srswor*. In addition, as demonstrated in Chapter 6 and discussed in section 4.6, we would expect the number of observed dung-piles to increase as $\frac{n_2^{(t)}}{n}$ increases as the distribution of the observed y_i has positive skew. The probability density function tends to be better estimated the greater the number of observations and hence we also expect the precision of $\hat{f}(0)$ to increase using a combined sampling design compared to selecting $s^{(2)}$ using *srswor*.

7.3 The first survey

Using distance sampling methods, the cutting and walking of line transects within a forest is time consuming. Walsh and White (1999) proposed an alternative for forests "*Recce* sampling". The "Recces" (short for reconnaissance walks) allow the observer to deviate from a strict compass bearing and to take the line of least resistance through the forest. This often means that the observer follows paths or animal trails. All dung-piles observed are recorded, but the distance from the dung-pile to the recce does not need to be measured. Walsh *et al.* (2001) tested these methods and found that a combination of the two strategies worked at a limited spatial scale.

Before implementing a full-scale monitoring programme at all the Central African forest sites in the MIKE system, the sampling strategy needed to be tested. A combined strategy using both recces and line transects was implemented. This was carried out at three sites, Odzala being one of them. The conclusion, given in Beyers *et al.* (2001), was that the recces did not increase the precision of $\hat{D}$ at Odzala, because adjacent recce and line transect counts were highly correlated. Here we concentrate only on the line transect data from the first survey, and propose a second survey that is based on the predicted abundance estimated from the line transect data only.

The survey region is an area of 7,750 km^2, part of which falls outside the Odzala National

Park boundary. Inside the national park, patrols are carried out regularly in an attempt to prevent poaching. On a small scale, elephant density has been shown to vary by vegetation type (Barnes *et al.*, 1991). On a larger scale, studies in many parts of Central Africa (Barnes *et al.*, 1995; Fay, 1991; Fay and Agnagna, 1991) suggest that the density of forest elephants increases with distance from human disturbance, such as transport networks and villages. Hence Barnes (1993) suggests that stratification should be based on human disturbance such as intensity of poaching, distance from villages and transport networks.

From GIS base maps, produced as part of the WCS Central Africa pilot study (Beyers and Hart, 2001), data on several variables relating to human disturbance were made available. For the survey region, these could be summarised as a set of distances from:

x_{i1} park boundary, where the distance from the park boundary for areas outside the national park are negative;

x_{i2} nearest village;

x_{i3} nearest non-conservation village. Conservation villages engage in anti-poaching activities; such as education programmes. This variable identifies villages that may be a source of poaching;

x_{i4} nearest park guard camp;

x_{i5} regular park guard patrol route;

x_{i6} road.

Past work is not detailed enough to suggest a model of the relationship between these auxiliary variables and elephant dung density, or if all of these auxiliary variables are important. Variables x_{i2} and x_{i3} are highly correlated, as are variables x_{i4} and x_{i5}, so only four of these variables were considered for stratification. These are illustrated in figure 7.1. Four variables were considered too many for stratification and so a proxy based

on accessibility was used to divide the survey region into three strata, as indicated by the three blocks in figure 7.1. One of these areas, the most southerly block, is outside the park, and of the two inside, the most northerly block is farther from park boundaries, non-conservation villages and roads.

For sample selection however, it appears that an *sys* sample was selected over the whole survey region and that stratification was introduced post-hoc for estimation purposes only. Figure 7.2 shows the proposed sample units. Note that the number of sample units in each stratum is very close to what we would expect under *strs* using proportional allocation. Each sample unit was 5 km in length from which five 200m line transect were taken at 1km intervals. The rest of the 5 km sample unit was sampled using recce transects as an aim of the pilot study was to investigate how the two could be used together. The results of the five 200m line transects are treated as one 1km transect in the analysis.

In the field, it was not possible to visit all of these transects. In addition some transects were not correctly located in the field. The results from the 44 transects taken are given in figure 7.2. It is clear that outside the park, the density of elephant dung is low, whereas inside the park, density is much higher. An estimate of the density of dung-piles over the survey region was calculated to be 23.8 km^{-2} (c.v.=13.9%) and there was no evidence of P_a varying across strata. Most (86%) of the variation in $\hat{D}$ was due to variability in the y_i, the number of dung-piles observed in each stratum (Beyers *et al.*, 2001).

A model of dung-density, ζ was obtained using the observed counts from each transect, the auxiliary variables shown in figure 7.1 and the location of each transect. In Beyers and Hart (2001, Annex 7) model selection was carried out by initially fitting a generalised additive model (GAM) with these 4 covariates and latitude and longitude. The final model was selected using the automated stepwise procedure in S-Plus; this uses the lowest AIC as the selection criterion, and the model included the distance from patrol and a smooth of the distance from the road. Here we select a model with the minimum generalised cross-validation score, using the mgcv function from the mgcv library (Wood, 2001) in R. The same variables, distance from road and distance from patrol routes, were retained in

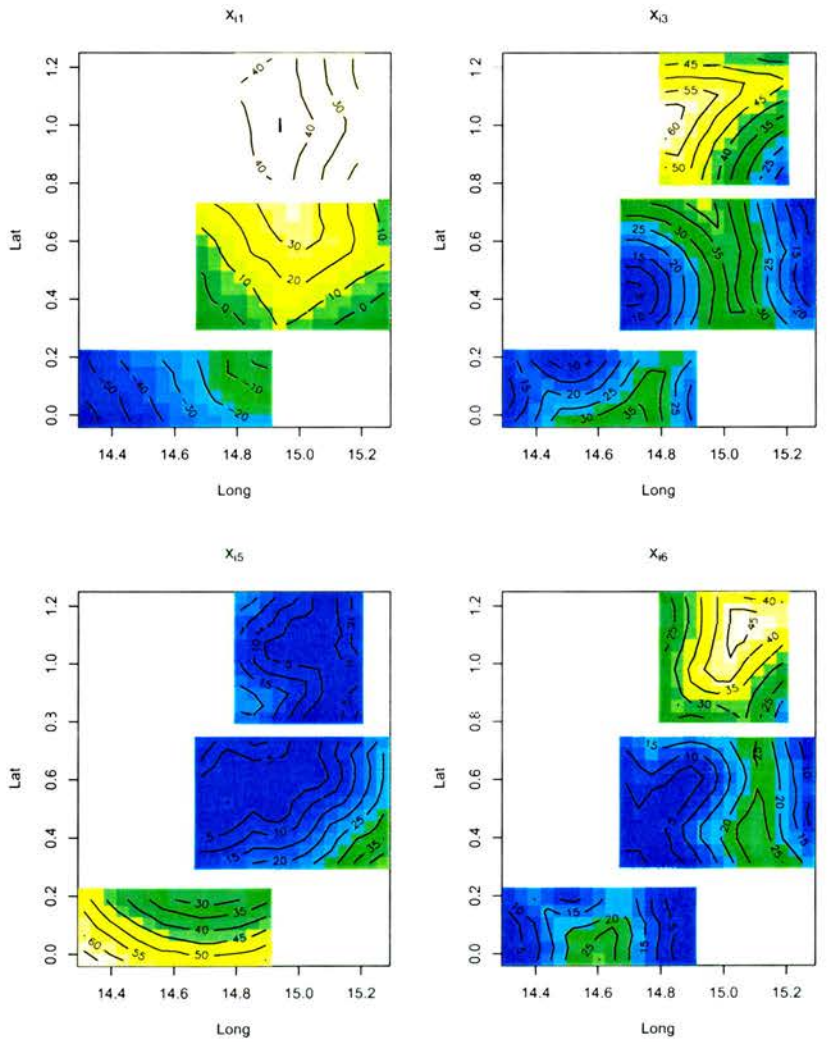


Figure 7.1: Auxiliary variables x_{i1}, distance from park boundary; x_{i3} distance from a non-conservation village ; x_{i5} distance from patrol route; x_{i6} distance from road.

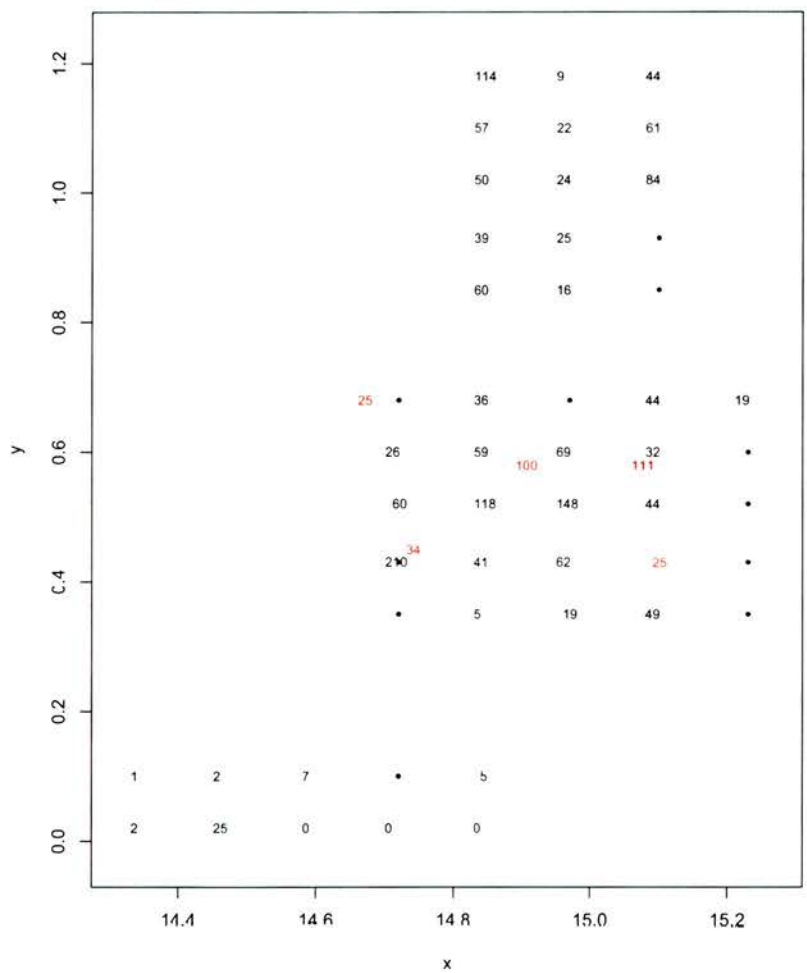


Figure 7.2: Number of dung-piles detected in each sampling unit. Sampling units that were omitted are shown as a dot, and units that were wrongly sampled shown in red.

the model and the distance from the nearest conservation village was also retained in the model.

To obtain a map of predicted dung density, $\hat{\mu}_i^{(1)}$, the values of the auxiliary variables have been extracted from the GIS for cells of size 21.25 km^2. This is shown in figure 7.3. In Beyers and Hart (2001, Annex 7), the aim of producing a map of predicted dung density was to demonstrate that model-based estimates of density can be obtained. Our emphasis is different. We require the map of predicted dung density to help determine the inclusion probabilities for the following survey. We note that the presentation of a map of species density, Obj. 1(b), is of use to site managers, especially when a key component of the data analysis strategy is that any analysis of data collected from a site should be returned to the site manager.

7.4 Design of a second survey

We assume that the total number of transects that can be taken is the same as in the first survey, so that a potential 50 transects will be sampled. The set of units from which we sample will be restricted to the 365 cells of 21.25 km^2 for which auxiliary data are available, although it would be possible to use a finer grid, particularly if auxiliary data were made available at a smaller scale.

The aim of a second survey is to obtain an estimate of elephant density, and in the long-term an estimate of change in density over the survey region. Work at another park, Ituri, suggests that it is possible to return to the same locations (Beyers *et al.*, 2001) although whether this is possible at Odzala is not clear. In addition the correlation between $y_i^{(1)}$ and $y_i^{(2)}$ is currently not known and so the benefit of retaining units from one survey to another, which will take more effort is unknown.

The sample $s_1^{(2)}$ will be selected using *sys*. This should ensure reasonable coverage of the survey region so that a model of density in the following year can be obtained, and also to

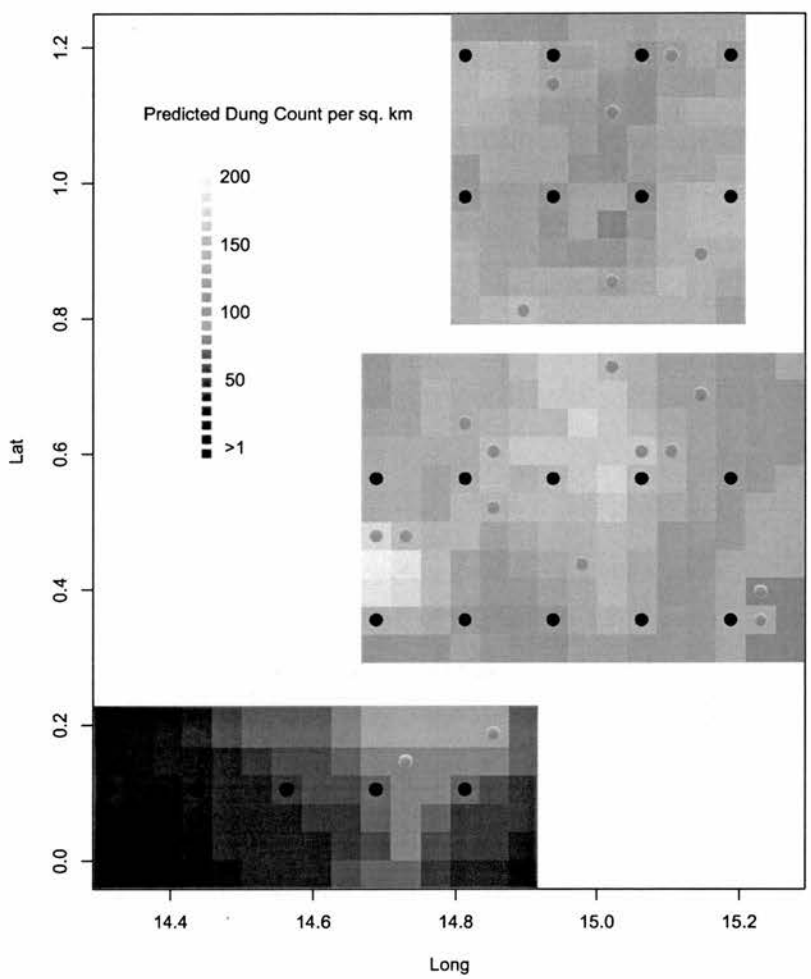


Figure 7.3: Predicted density, $\hat{\mu}_i^{(1)}$ and proposed sample design for survey 2. Units selected using *sys* are shown in green and those selected using $\pi p \hat{\mu}_i^{(1)}$ in pink.

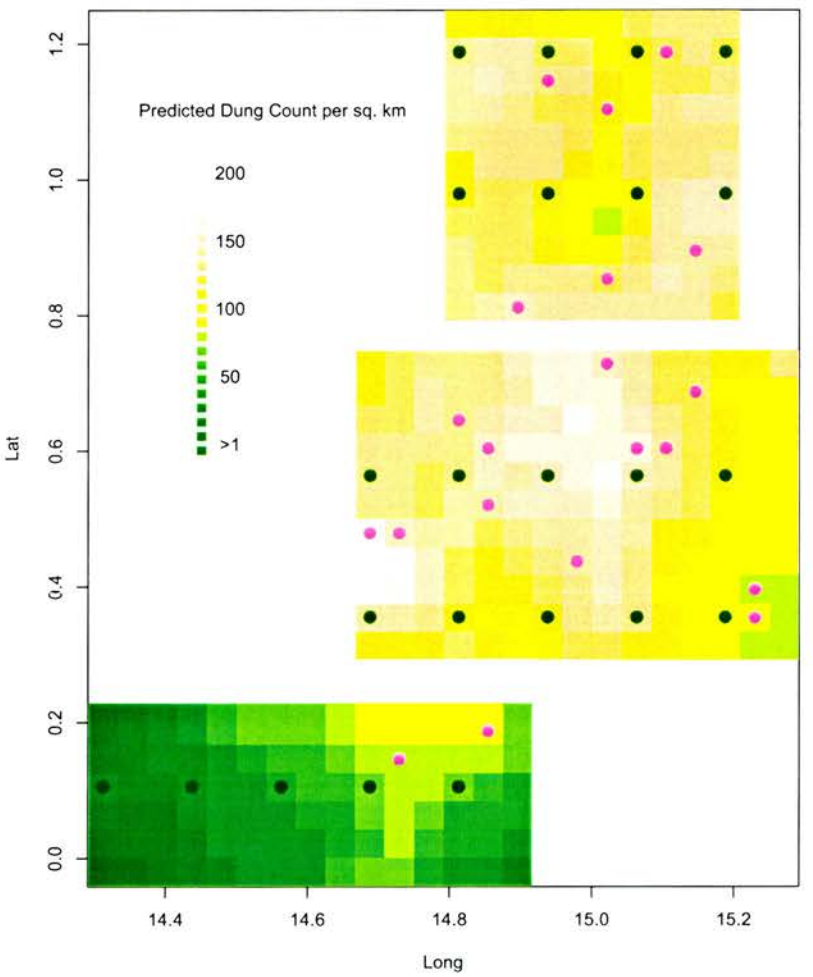


Figure 7.3: Predicted density, $\hat{\mu}_i^{(1)}$ and proposed sample design for survey 2. Units selected using *sys* are shown in green and those selected using $\pi p \hat{\mu}_i^{(1)}$ in pink.

the model and the distance from the nearest conservation village was also retained in the model.

To obtain a map of predicted dung density, $\hat{\mu}_i^{(1)}$, the values of the auxiliary variables have been extracted from the GIS for cells of size 21.25 km^2. This is shown in figure 7.3. In Beyers and Hart (2001, Annex 7), the aim of producing a map of predicted dung density was to demonstrate that model-based estimates of density can be obtained. Our emphasis is different. We require the map of predicted dung density to help determine the inclusion probabilities for the following survey. We note that the presentation of a map of species density, Obj. 1(b), is of use to site managers, especially when a key component of the data analysis strategy is that any analysis of data collected from a site should be returned to the site manager.

7.4 Design of a second survey

We assume that the total number of transects that can be taken is the same as in the first survey, so that a potential 50 transects will be sampled. The set of units from which we sample will be restricted to the 365 cells of 21.25 km^2 for which auxiliary data are available, although it would be possible to use a finer grid, particularly if auxiliary data were made available at a smaller scale.

The aim of a second survey is to obtain an estimate of elephant density, and in the long-term an estimate of change in density over the survey region. Work at another park, Ituri, suggests that it is possible to return to the same locations (Beyers *et al.*, 2001) although whether this is possible at Odzala is not clear. In addition the correlation between $y_i^{(1)}$ and $y_i^{(2)}$ is currently not known and so the benefit of retaining units from one survey to another, which will take more effort is unknown.

The sample $s_1^{(2)}$ will be selected using *sys*. This should ensure reasonable coverage of the survey region so that a model of density in the following year can be obtained, and also to

ensure consistency with $s^{(1)}$. Being relatively cautious we will consider a survey design in which $\frac{n_2^{(2)}}{n} = 0.5$ so that half the sampling effort is allocated to providing good coverage of the survey region. When the survey region is not entirely regular the number of units in the systematic sample cannot be strictly identified before the sample is taken. We select the systematic sample by considering the rectangle that contains all of the survey region; this is approximately twice the area of the survey region. We select a systematic sample of 48 units (8 x 6) over the whole rectangle and retain only those units that occur in the survey region. The remaining units are selected using $\pi p\hat{\mu}^{(1)}$. As we only have the current habitat data, then we cannot do better than estimate $\mu_i^{(2)}$ using $\hat{\mu}_i^{(1)}$.

Figure 7.3 shows a sample selected using this sample design. The green dots represent the units selected using *sys*, of which there are 28, and the pink dots the units selected using $\pi p\hat{\mu}^{(1)}$. As under survey 1, the green dots are approximately allocated according to the size of the strata. There are most pink dots in the second stratum, which generally has the highest predicted density. This illustrates how we could consider our combined sampling designs as a form of stratified sampling in which the number of units in each stratum is first selected using proportional allocation, and additional units are then added to strata in which abundance is greatest.

7.5 Discussion

In this chapter we have described a global monitoring programme, part of which is to estimate trends in elephant population numbers through time. As the state of elephant populations is extremely controversial, with different factions clearly wanting very different outcomes a design-based estimate of elephant numbers and changes in elephant numbers provides a useful and non-controversial basis for discussion.

Forest elephant abundance is estimated by counting elephant dung-piles using distance sampling surveys. We have shown in this chapter how the combined sampling designs can be used to select transect lines and how this is included in the estimation of D. This can

only be applied when the probability of detecting a dung-pile depends only on its distance from the transect line. We have stated that we would expect the precision of $\hat{D}$ to increase when transects are selected under a combined sampling design when $\frac{n_2}{n} > 0$ compared to an *srswor* design, $\frac{n_2}{n} = 0$. This is firstly because the variability in the encounter rate decreases, but secondly because we expect $\sum_s y_i$ to increase.

Pollard *et al.* (2002) developed a fixed-effort adaptive distance sampling method in which the path of the line transect changes when more than C individuals are detected in a transect. In simulations using this strategy they demonstrated that more individuals are observed than under a standard line transect survey and so the precision of $\hat{f}(0)$ increases. However the variability in the encounter rate did not decrease because of the extra effort in sampling; this is equivalent to the edge units in adaptive cluster sampling. As the greatest proportion of the variability in $\hat{D}$ depends on encounter rate the strategy did not show much, if any increase in the precision of $\hat{D}$. We note that the motivation for this strategy is different to our combined sampling strategy. Under the adaptive distance sampling strategy the aim is to detect clusters of individuals that are associated, whereas the combined sampling strategy aims to sample areas of high density that occur due to spatial trend.

To predict $\hat{\mu}_i^{(1)}$, the model $\zeta^{(1)}$ was constructed using the observed counts for each transect, the y_i, and did not take the probability of detection or the distance from the transect line into account. Further work is required if the probability of detection varies because of differences between observers, field conditions, the habitat or individuals of the species. Using one of the two strategies described by Hedley *et al.* (1999), that incorporate imperfect detection into the modelling process, the model $\zeta^{(1)}$ can be constructed so that $\hat{\mu}_i^{(1)}$ can be obtained and $s^{(2)}$ selected, but it is not clear how $\hat{D}^{(2)}$ would be estimated as

$$\hat{D}^{(2)} = \frac{1}{2Nl} \sum_s \hat{f}_i(0) \frac{y_i}{\pi_i}$$

as $f(0)$ varies between individuals and so it is not clear how the variance would be estimated.

Chapter 8

Covariance estimation using the bootstrap

The two-phase sampling strategy described in the previous chapter provided a method of estimating abundance $\tau^{(t)}$ and change in abundance $\delta^{(t',t)} = \tau^{(t)}\tau^{(t')}$. To calculate the precision of these estimators several covariances must be estimated. In Chapter 5, we saw that the analytic estimators of covariance only used the data in $s_1^{(2)}$ and none of the data in $s_2^{(2)}$. In addition, the estimation of some covariances can become quite complex. In this chapter, we investigate whether an alternative method can be employed to estimate these covariances.

There are many different strategies used to estimate the variance $var(\hat{\tau}^{(t)})$. Wolter (1985) and Särndal *et al.* (1992) provide reviews of several strategies including the use of random groups; balanced half-samples; generalised variance functions; Taylor series methods; the jackknife; and the bootstrap. These strategies are generally applied to complex surveys or non-linear estimators of $\hat{\tau}^{(t)}$ in which analytic variance estimators cannot be used. This is either because an expression for the variance cannot be found or because an appropriate unbiased estimator of the variance cannot be expressed analytically. All of the strategies tend to have some small amount of bias but are flexible so that they can accommodate

most features of a complex survey or estimator. It is difficult to select one strategy over any other for all survey designs and estimators. Although strategies for variance estimation are numerous, it is not always easy to apply the same strategies to the estimation of covariance.

We require a strategy for covariance estimation that uses all of the data in the sample and is relatively simple to implement. Bootstrapping is one such method that is often applied to complex problems, and in this chapter we explore a new approach to bootstrapping for survey sampling with particular reference to two-phase unequal probability sampling described in Chapter 5. Bootstrapping has been applied to many areas of statistics but there have been only a few applications to finite population sampling. For our sampling design, we need to build on bootstrapping methods for variance estimation from samples selected using unequal probability sampling, for example the πps sampling design, before attempting to estimate covariances. We start this chapter, section 8.1, by describing current bootstrapping methods in survey sampling and state why they are not appropriate for more general πps sampling designs. Section 8.2 describes a simple strategy that can be applied to unequal probability sampling. There are variations on this strategy and we calculate the bias of these variations when s is selected using *srswor*; it is not possible to estimate bias analytically for samples selected using πps sampling. In section 8.3 we consider how the basic bootstrapping strategy can be applied to two-phase sampling designs to estimate the covariance $cov(\hat{\tau}_1^{(t)}, \hat{\tau}_2^{(t)})$. We extend the methods further in section 8.4 to indicate how the covariance $cov(\hat{\tau}^{(t')}, \hat{\tau}^{(t)})$ could be estimated.

8.1 Current methods of bootstrapping

Although we are interested in estimating covariances, we start by considering how we might obtain $\widehat{var}_{BS}(\hat{\tau}_s)$, a bootstrap estimate of the variance, for the simple case in which τ_s is an estimate of τ from a sample s of size n. We let $\widehat{var}_a(\hat{\tau}_s)$ be an analytic estimate of $var(\hat{\tau}_s)$.

Davison and Hinkley (1997, pps. 92–100) review the general principles of bootstrapping in survey sampling. If we knew U, we could estimate $\widehat{var}_a(\hat{\tau}_s)$ by taking a sample s of size n from U and calculate $\hat{\tau}_s$. If we repeat this a large number, say B, of times so that $\underline{\hat{\tau}_s} = (\hat{\tau}_{s1}, \ldots, \hat{\tau}_{s2}, \ldots, \hat{\tau}_{sB})^2$ then $var(\hat{\tau}_s)$ can be estimated by finding the sample variance of the $\underline{\hat{\tau}_s}$ so that

$$\widehat{var}(\hat{\tau}_s) = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\tau}_{sb} - \frac{1}{B} \sum_{b=1}^{B} \hat{\tau}_{sb} \right)$$

In practice only s rather than U is known and so we use s to construct a pseudo-population, U^*, instead. The general bootstrapping procedure is:

1. Construct a pseudo-population U_b^* assumed to mimic the real, but unknown population U.
2. Take a bootstrap resample, s_b^*, from U_b^* using some strategy and calculate $\hat{\tau}_b^*$.
3. Repeat either steps 1 and 2, or just step 2, B times, to give $\underline{\hat{\tau}}^* = (\hat{\tau}_1^*, \ldots, \hat{\tau}_B^*)$
4. Use the observed distribution $\underline{\hat{\tau}}^*$ as an estimate of the sampling distribution of $\hat{\tau}$ and estimate the parameter of interest accordingly. In this case

$$\widehat{var}_{BS}(\hat{\tau}) = var(\underline{\hat{\tau}}^*) = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\tau}_b^* - \frac{1}{B} \sum_{b=1}^{B} \hat{\tau}_b^* \right)^2$$

Different bootstrap estimators vary as follows: the method of constructing the pseudo-population, U_b^*, step BS(1); the sampling design used to select the samples s_b^*, step BS(2); and whether each sample s_b^* is selected from the same pseudo-population, which we will denote U^*, or a different pseudo-population, which we will denote U_b^*, step BS(3).

In the literature, there are several proposed strategies for estimating $\widehat{var}_a(\hat{\tau}_s)$ when s has been selected using *srswor*. In all cases, the resamples, s_b^*, are selected from the pseudo-population using *srswor*. Most of these strategies use the principle of the population bootstrap (Davison and Hinkley, 1997), generally referred to as the Bootstrap WithOut

replacement (BWO) in survey sampling and described by Gross (1980). Let $N = cn + d$ where $0 \leq d < n$ and $c > 0$ is an integer. When $d = 0$ one pseudo-population, U^*, is created that consists of $c = N/n$ repeats of each unit $i \in s$. From this one pseudo-population, B samples, $s_1^*, \ldots, s_B^*$ are drawn, each sample being selected using *srswor*. This strategy gives a biased estimate of variance so that $\widehat{var}_{pb}(\hat{\tau}_s)$ the estimated variance using the population bootstrap is

$$\widehat{var}_{pb}(\hat{\tau}_s) = \frac{N(n-1)}{(N-1)n} \widehat{var}_a(\hat{\tau}_s) \tag{8.1}$$

When $d > 0$, equivalent to N/n being non-integer, then a number of adaptations to the strategy have been suggested. Several of these are described in Holmberg (1998) and we summarise the main strategies here. Davison and Hinkley (1997) and Booth *et al.* (1994) suggest that cn units are created as described above, so that each unit in s is replicated n times, and then d units are selected from s using *srswor*. Chao and Lo (1984, 1994) propose a similar method except that the d units are selected using simple random sampling with replacement. In either case each bootstrap resample, s_b^*, is selected from a new pseudo-population. The pseudo-populations vary only in the last $d < n$ units. Bickel and Freedman (1984) and Sitter (1992) generate two pseudo–populations U_1^* and U_2^*. The pseudo-population U_1^* consists of c_1 replicates of the units in s and U_2^* consists of c_2 replicates of the units in s. Bickel and Freedman (1984) set $c_1 = c$ and $c_2 = c + 1$, whereas Sitter (1992) sets $c_1 < c* < c_2$ where c_1 and c_2 are consecutive integers and $c* = \frac{1}{n^2}(Nn - N - n)$. Given these two potential pseudo-populations, U_1^* and U_2^*, only one is selected for use as U^*, by selecting U_1^* with probability p and U_2^* with probability $1 - p$. Bickel and Freedman (1984) use one pseudo-population, U^*, for all bootstrap samples whereas in Sitter (1992) each bootstrap resample, s_b^*, is selected from a new pseudo-population, U_b^*. If p is chosen correctly the resulting estimator, $\widehat{var}_{BS}(\hat{\tau}_s)$, is unbiased.

An alternative approach is the superpopulation bootstrap, as described by Davison and Hinkley (1997). Compared to the population bootstrap in which U^* consists of $c = N/n$ replicates of each unit in the sample, the superpopulation bootstrap generates a pseudo-

population by sampling with replacement from s. Each bootstrap resample, s_b^*, is selected from a new pseudo-population so on average each unit in s will occur an equal number of times in s. The mean variance is stated, without proof, by Davison and Hinkley (1997) to be

$$\widehat{var}_{BS}(\hat{\tau}) = \frac{(n-1)}{n} \widehat{var}_a(\hat{\tau}_s) \tag{8.2}$$

In section 8.2.1 we show that equation 8.2 only holds when all B bootstrap resamples are selected from the same pseudo-population, U^*. Unlike the BWO there are no restrictions, or adjustments to be made to the superpopulation method if N/n is not an integer, hence it is a more flexible strategy.

When s is selected using a πpz sampling design, for some auxiliary variable z_i, the bootstrap methods proposed in the literature are extensions of the population bootstrap. In general these are applied to very specific πpz sampling designs. Kuk (1989) proposes a strategy for systematic πpz sampling, and Rao and Wu (1984, 1988), Rao *et al.* (1992) and Sitter (1992, 1992b) propose strategies for the Rao-Hartley unequal probability sampling design. Holmberg (1998) suggests a general extension of the population bootstrap, and demonstrates its applicability to two πpz sampling designs one of which is that of Sunter (1977a) which we have employed in this thesis. Holmberg's bootstrapping strategy creates U^* by replicating the i^{th} unit in s $w_i = 1/\pi_i$ times. This replication occurs for both the y_i and the z_i so that the pseudo-population has units with counts of $\underline{y}^*$ and auxiliary data $\underline{z}^*$. The pseudo-population, U^*, will be of size $\hat{N} = \sum_s w_i$. This is not equal to N except under very simple sampling designs such as *srswor*. The resamples s_b^* are selected by sampling with πpz^*. Särndal *et al.* (1992) proposes the same scheme for generating U^* but the resamples, s_b^*, are selected by sampling with replacement from U^* with selection probability $p_{i*} = \pi_i^*/n$ where π_i^* are the inclusion probabilities by sampling with πpz^*. They show that using this strategy the resulting variance estimator has a bias of $\frac{n-1}{n}$.

One difficulty with the bootstrap strategies proposed for unequal probability sampling is that they have only been developed for cases where w_i is an integer. This is generally not the case in practice. Also the pseudo-population will not be the same size as the original

population, hence it is not always clear how employing the original sampling strategy to obtain the s_b^* will give an appropriate estimate of variance. If however an alternative strategy is suggested for obtaining s_b^*, such as that of Särndal *et al.* (1992), then there is no simple method for extending the bootstrapping strategy to more complex sampling schemes such as the two-phase sampling strategy described in Chapter 5.

8.2 A bootstrap strategy for unequal probability sampling

Suppose the sample s is selected using πpz and we wish to estimate $var(\hat{\tau})$. We propose a strategy for generating the pseudo-population that is similar in spirit to that of Holmberg (1998), in that the inclusion probabilities determine how U^* is generated. We use the superpopulation approach rather than the population bootstrap approach as we sample with replacement from s.

The principle is that a pseudo-population, U^*, is generated by sampling with replacement from s where the probability that unit i is selected, p_i, is proportional to the reciprocal of its inclusion probability, π_i. Hence if $w_i = 1/\pi_i$ then

$$p_i = \frac{w_i}{\sum_{k \in s} w_k} = \frac{1/\pi_i}{\sum_{k \in s} 1/\pi_k}. \tag{8.3}$$

We denote this strategy $ppswr(w_i, N, s)$, that is we select N units from s by sampling with replacement with probability proportional to w_i.

Although not strictly necessary for the estimation of $var(\hat{\tau})$, we separate the selection of the units in U^* from the values of y_i or z_i that these units take. This is important when we propose bootstrap strategies for the estimation of $cov(\hat{\tau}^{(t')}, \hat{\tau}^{(t)})$ in which the same pseudo-population will take different $\underline{y}^*$ values depending on the survey of interest. We let $U^* = (i_1^*, \ldots, i_N^*)$ represent the set of labels identifying which units in s occur in U^*, selected using $ppswr(w_i, N, s)$. The y_i^* values of these N units are specified so that

$$\underline{y}^* = (y_{i_1}^*, \ldots, y_{i_N}^*) \text{ where } y_{i_j}^* = y_k \text{ if } i_j = k \text{ for } i_j \in U^*, k \in s$$

we define this as $pspv(\underline{y}, s, U^*)$ so that

$$\underline{y}^* = pspv(\underline{y}, s, U^*)$$

Similarly the z_i that were used to select the original sample must also be allocated to the units in U^* so that

$$\underline{z}^* = (z^*_{i_1}, \ldots, z^*_{i_N}) = pspv(\underline{z}, s, U^*)$$

In other words if the k^{th} unit in s has been selected as the i^{*th} unit in the pseudo-population the y_k and z_k values are also attributed to the i^{*th} unit in the pseudo-population. Once the pseudo-population U^* has been created the bootstrap resample s^* is selected using πpz^*.

The generation of U^* is based on the same logic as the Horvitz-Thompson estimator $\hat{\tau}_{HT} = \sum_s \frac{y_i}{\pi_i}$ in that a unit $i \in s$ represents $w_i = \frac{1}{\pi_i}$ such units in the population U (Overton and Stehman, 1995). For example if a unit in the sample has an inclusion probability of $\pi_i = 0.25$, we take this to mean that there are $4 = \frac{1}{0.25}$ such units in U.

The aim of creating the pseudo-population is to replicate the real population U as well as possible. In the bootstrap strategy we have proposed above, all N units of the pseudo-population are generated by sampling from s. Therefore some of the units in s may not occur in U especially if the sampling fraction is high, for example $n = 0.5N$. In this case, half the units in the population are known so U^* would be a better replica of U if n units in the pseudo-population are the units from the sample s and the sample complement of $N - n$ units are generated using the bootstrap strategy defined above. Then if a unit in the sample has an inclusion probability of $\pi_i = 0.25$, we would expect there to be 4 such units in $s_c = U - s$. As one of these units occurs in s, then we would expect there to be another $3 = \frac{1}{0.25} - 1 = \frac{1-\pi_i}{\pi_i} = w_i - 1$ units in U. Hence U^* is created using a strategy we define as $ppswr_c(w_i, N, s)$, in which U^* consists of the n units in s and a pseudo-sample-complement, s^*_c, of $N - n$ units where s^*_c is selected using the strategy $ppswr(w_i - 1, N - n, s)$. The full bootstrapping procedures are given in box 8.1

Box 8.1: Bootstrapping strategy to estimate $var(\hat{\tau})$ when s is selected using πpz

1. Generate a pseudo-population U_b^* of size N so that U_b^* is selected using either

   (a) $ppswr(w_i, N, s)$

   (b) $ppswr_c(w_i, N, s)$

2. Generate the response values $\underline{y}_b^*$ and the auxiliary data $\underline{z}_b^*$ using

$$\underline{y}_b^* = pspv(\underline{y}, s, U_b^*)$$
$$\underline{z}_b^* = pspv(\underline{z}, s, U_b^*)$$

3. Select a sample s_b^* of size n from U_b^* using the sampling design πpz. Calculate

$$\hat{\tau}_b^* = \sum_{s_b^*} \frac{y_{ib}^*}{\pi_{ib}^*}$$

4. Repeat steps 1 to 3, or just step 3, B times to give

$$\hat{\tau}^* = (\hat{\tau}_1^*, \ldots, \hat{\tau}_B^*)$$

5. Estimate $var(\hat{\tau})$ to be

$$\widehat{var}_{BS}(\hat{\tau}) = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\tau}_b^* - \frac{1}{B} \sum_{b=1}^{B} \hat{\tau}_b^* \right)^2$$

$ppswr(w_i, N, s)$	Select N units from s, sample with replacement, probability $\frac{w_i}{\sum_{k \in s} w_k}$
$ppswr_c(w_i, N, s)$	Retain sample s and select $N - n$ units using $ppswr(w_i - 1, N - n, s)$
$pspv(\underline{y}, s, U^*)$	$y_{i_j}^* = y_k$ if $i_j = k$ for $i_j \in U^*, k \in s$

8.2.1 Which bootstrap strategy?

In Box 8.1 there is a choice in how both step 1 and step 4 are implemented. This gives four possible bootstrapping strategies.

$U_{[1]}$: Generate one pseudo-population for all bootstrap resamples.
Step 1(a) and repeat step 3 B times

$U_{[B]}$: Generate a new pseudo-population for each bootstrap resample.
Step 1(a) and repeat steps 1(a), 2 and 3 B times

$s_{[c1]}$: Generate one pseudo-sample-complement for all bootstrap resamples.
Step 1(b) and repeat step 3 B times

$s_{[cB]}$: Generate a new pseudo-sample-complement for each bootstrap resample.
Step 1(b) and repeat steps 1(b), 2 and 3 B times

Independently of our work, Sverchkov and Pfeffermann (2003) proposed the strategy $U_{[1]}$. The probability proportional to size extension to the superpopulation bootstrap of Davison and Hinkley (1997) is $U_{[B]}$. Before any simulations had taken place, we favoured the strategy $s_{[cB]}$ as this generated only the units which we did not know about, s_c, and accounted for the variability in the generation of U_b^* in the estimation of $\widehat{var}_B S(\hat{\tau})$. We can compare the four strategies using simulation. In addition, the strategies $U_{[1]}$ and $U_{[B]}$ can be compared analytically if the original sample was selected using *srswor*. For this purpose we denote the estimated analytic variance of $\hat{\tau}_s$ as

$$\widehat{var}_a(\hat{\tau}_s) = \frac{N(N-n)}{n(n-1)} \sum_s (y_i - \bar{y}_s)^2 \tag{8.4}$$

where

$$\bar{y}_s = \frac{1}{n} \sum_s y_i = \frac{1}{N} \hat{\tau}_s \tag{8.5}$$

Under the strategy $U_{[1]}$

$$\widehat{var}_{U_{[1]}}(\hat{\tau}_s) = var(\hat{\underline{t}}^*|U^*)$$

so the expected variance is the expectation over all possible U^*s so that

$$E[\widehat{var}(\hat{\tau}_s)_{U_{[1]}}] = E_{U^*}[var(\hat{\tau}_b^*|U^*)] \tag{8.6}$$

We consider the simple scenario in which the original sample, $s = (y_{i_1}, \ldots, y_{i_n})$, is selected using *srswor*, hence the pseudo-population, $U^* = (y_1^*, \ldots, y_N^*)$, is selected using simple random sampling with replacement and the $\hat{\tau}_b^*$ are obtained by selecting the resamples, $s_b^* = (y_{i_1}^*, \ldots, y_{i_n}^*)$, with *srswor*. Hence

$$var(\hat{\underline{t}}^*|U^*) = \frac{N(N-n)}{n(N-1)} \sum_{U^*} (y_i^* - \frac{1}{N} \sum_{U^*} y_i^*)^2 \tag{8.7}$$

and we use the results of with-replacement sampling to find an expression for the expected variance. Let

$$t = \sum_s y_i = n\bar{y}_s \tag{8.8}$$

This is estimated by taking a sample $U^* = (y_1^*, \ldots, y_N^*)$ of N units from $s = (y_1, \ldots, y_n)$[1]. Using the Hansen-Hurwitz estimator (Hansen and Hurwitz, 1943), we can estimate t to be

$$\hat{t} = \frac{n}{N} \sum_{U^*} y_i^*$$

This has a variance of

$$var(\hat{t}) = \frac{1}{nN} \sum_s (ny_i - t)^2$$

Using equation 8.8 then

$$var(\hat{t}) = \frac{n}{N} \sum_s (y_i - \bar{y}_s)^2$$

[1]This is a reversal of our usual sampling problem as here our "sample" U_b^* is greater than our "population" s, The purpose here though is to regenerate U from s.

and from equation 8.4

$$var(\hat{t}) = \frac{n^2(n-1)}{N^2(N-n)} \widehat{var}_a(\hat{\tau}_s) \tag{8.9}$$

The Hansen-Hurvitz variance estimator is of the form

$$\widehat{var}(\hat{t}) = \frac{1}{N(N-1)} \sum_{U_b^*} (ny_i^* - \hat{t})^2 = \frac{n^2}{N(N-1)} \sum_{U_b^*} (y_i^* - \frac{1}{N} \sum_{U^*} y_i^*)^2$$

and subsituting this into equation 8.7 gives

$$var(\hat{\tau}_b^* | U^*) = \left[\frac{N^2(N-n)}{n^3} \widehat{var}(\hat{t}) \right]$$

As $\widehat{var}(\hat{t})$ is an unbiased estimate of $var(\hat{t})$ then

$$E_{U_b^*}[var(\hat{\tau}_b^* | U^*)] = \frac{N^2(N-n)}{n^3} var(\hat{t})$$

and so from equations 8.6 and 8.9

$$E[\widehat{var}(\hat{\tau}_s)_{U_{[1]}}] = \frac{n-1}{n} \widehat{var}_a(\hat{\tau}_s) \tag{8.10}$$

Under the bootstrap strategy $U_{[B]}$, any estimate of $\widehat{var}(\hat{\tau}_s)_{U_{[B]}}$ is already taken over many U_b^*. Hence we need to take expectations over the U_b^* so that

$$E[\widehat{var}_{U_{[B]}(\hat{\tau}_s)}] = E_{U_b^*}[var(\hat{\tau}_b^* | U_b^*)] + var_{U_b^*}(E[\hat{\tau}_b^* | U_b^*]) \tag{8.11}$$

In addition to $E_{U_b^*}[var(\hat{\tau}_b^* | U_b^*)]$ given in equation 8.10 we also require $var_{U_b^*}(E[\hat{\tau}_b^* | U_b^*])$. Now $\hat{\tau}_b^*$ is an unbiased estimte of τ_b^* where

$$\tau_b^* = \sum_{U_b^*} y_i^* = \frac{N}{n} \hat{t} \tag{8.12}$$

$$\Rightarrow var(\tau_b^*) = \frac{N^2}{n^2} var(\hat{t}) \tag{8.13}$$

and so

$$var_{U_b^*}(E[\hat{\tau}_b^* | U_b^*]) = var_{U_b^*}(\tau_b^*) = \frac{N^2}{n^2} var(\hat{t}) \tag{8.14}$$

Using equation 8.9

$$var_{U_b^*}(E[\hat{\tau}_b^*|U_b^*]) = \frac{n-1}{N-n}\widehat{var}_a(\hat{\tau}_s) \tag{8.15}$$

Combining this with equation 8.10 we obtain

$$E[\widehat{var}(\hat{\tau}_s)_{U[B]}] = \frac{(n-1)N}{n(N-n)}\widehat{var}_a(\hat{\tau}_s) \tag{8.16}$$

Davison and Hinkley (1997) give a bias of the mean value under the strategy $U_{[B]}$, equation 8.2. Our working would suggest that this is in fact the bias under the strategy $U_{[1]}$. The work by Sverchkov and Pfeffermann (2003) does not give an estimate of any bias for their proposed strategy $U_{[1]}$.

Although we can calculate the expected variance under the bootstrap when s has been selected using *srswor*, we cannot do this when s has been selected using πps. Neither can we calculate analytic results for the bootstrapping strategies $s_{c[1]}$ and $s_{c[B]}$. Instead we can only obtain results using simulation. Using the $y_i^{(2)}$ values from population A, table 8.1 summarises results from applying all four bootstrapping strategies when s has been selected using *srswor* or $\pi p\hat{\mu}^{(1)}$. For each sampling design, a sample of size $n = 50$ was taken and $\widehat{var}_a(\hat{\tau}_s)$ estimated using each of the bootstrapping strategies described above. This procedure was repeated 1000 times. The means of $\widehat{var}(\hat{\tau}_s)$ and their standard deviations are given in the table. In addition to a sample of size $n = 50$, the samples of size $n = 250, 500, 750, 950$ were also taken using the sample design of *srswor*. One disadvantage of using Sunter's sampling strategy (Sunter, 1977a) for πps sampling is that it performs poorly when the sampling fraction is large, as most of the sample is then selected using *srswor*. Under the bootstrapping strategies, almost all of the sample was selected using *srswor* and estimates were very poor. We have not included these results here, but further exploration of the bootstrapping strategy using Chao's sampling strategy (Chao, 1982) would be useful.

As n increases the bias under $U_{[1]}$ tends to zero. When the sampling fraction is large, although this is rare in practice, the bias for strategy $U_{[B]}$ is large as $\frac{N}{N-n}$ is large. The

Table 8.1: Comparison of different bootstrapping strategies to estimate $\sqrt{var(\hat{\tau}_s)}$ for population A. The sample s is selected using $\pi p\mu^{(1)}$, when $n = 50$, or *srswor*, for varying n. Results are the mean (s.d) of $\sqrt{\widehat{var}_{BS}(\hat{\tau}_s)}$ over 1000 samples of s. Each bootstrap estimate of variance is obtained using $B = 1000$ resamples. $\widehat{var}_a(\hat{\tau}_s)$ is the analytic estimate of $var(\hat{\tau}_s)$. To demonstrate bias, the values of $\sqrt{\frac{(n-1)}{n} var_a(\hat{\tau}_s)}$ and $\sqrt{\frac{n(N-n)}{(n-1)N} var(\hat{\tau}_s)}$ are given.

	$\pi p\mu^{(1)}$	*srswor*				
$\sqrt{\widehat{var}}$ n	50	50	250	500	750	950
$\sqrt{var}$	176	198	79	45	26	10
$\sqrt{\frac{n-1}{n}\widehat{var}_a}$	173 (25)	194 (27)	79 (5)	45 (2)	26 (1)	10 (0.2)
$U_{[1]}$	172 (24)	194 (27)	79 (5)	45 (2)	26 (1)	10 (0.3)
$s_{c[1]}$	173 (25)	194 (27)	79 (5)	45 (2)	26 (1)	10 (0.2)
$s_{c[B]}$	179 (24)	199 (27)	88 (5)	56 (2)	35 (1)	15 (0.3)
$U_{[B]}$	178 (23)	200 (27)	91 (5)	64 (2)	52 (1)	47 (1.2)
$\sqrt{\frac{(n-1)N}{n(N-n)} var}$	179	201	91	64	52	47

strategy $s_{c[1]}$ performs similarly to $U_{[1]}$. The strategy $s_{c[B]}$ has less bias than strategy $U_{[B]}$ because part of U_b^* is the same for all $U_1^*, \ldots, U_B^*$. This is especially clear under strategies in which n is large as the population varies very little from one survey to another. By taking 1000 new samples, we are able to compare the variability in the bootstrapping method with the variability in the analytic method. We see that the variability in the bootstrap estimator is similar to that of the analytic estimator. The table also shows that the bootstrap estimators when s is selected using $\pi p\hat{\mu}^{(1)}$ is as effective as when s is selected using *srswor*.

These results would suggest that generation of only one pseudo-population, $U_{[1]}$, or pseudo-sample-complement, $s_{c[1]}$, are the most appropriate strategies for estimating the variance $\widehat{var}(\hat{\tau})$. If the bias is known, as it is for $U_{[1]}$ under *srswor*, this can be accounted for in the estimator.

The advantage of these bootstrapping strategies over other unequal probability sampling strategies is that they can, in principle, be implemented whatever sample design is used rather than requiring a different bootstrapping strategy for a particular sample design. In addition there is no restriction on w_i being integer. We note that there are difficulties with the implementation of Sunter's sample design because part of the sample is selected using *srswor*.

8.3 Bootstrap estimation of $cov(\hat{\tau}_1, \hat{\tau}_2)$

Consider a simple two-phase sampling scheme in which a sample s_1 of n_1 units is selected using *srswor* and a sample s_2 of n_2 units is selected from s_{1_c} using πpz, for some z. This is equivalent to our combined sampling scheme that we proposed in chapter 4 in which the sample is selected independently of previous samples. Although we estimated $\hat{\tau}$ by calculating unconditional inclusion probabilities, we now consider the use of a two-phase sampling estimator. We consider this partly as a precursor to developing a bootstrapping strategy for estimation of $cov(\hat{\tau}^{(1)}, \hat{\tau}^{(2)})$ in which $\tau^{(2)}$ is estimated using a two-phase es-

timator. We expand on this in section 8.4. We estimate $\hat{\tau}_1$ using the Horvitz-Thompson estimator and $\hat{\tau}_2$ using the two-phase estimator so that

$$\hat{\tau}_2 = \sum_{s_2} \frac{y_i}{\pi_{2|s_1} \pi_{1_c}}$$

An analytic estimate of the covariance $cov(\hat{\tau}_1, \hat{\tau}_2)$ only uses the data in s_1.

Based on the bootstrapping method for unequal probability sampling proposed in the previous section, we describe a bootstrapping strategy that can estimate $cov(\hat{\tau}_1, \hat{\tau}_2)$ that uses the data from both s_1 and s_2. To generate a pseudo-population, U^*, we construct two sub-pseudo-populations using the data from s_1 and s_2 respectively and the strategy described in the previous section. The sizes of the sub-pseudo-populations are based on the relative sizes of the sub-samples so that U_1^* is of size $N\frac{n_1}{n}$ and sub-pseudo-population U_2^* is of size $N\frac{n_2}{n}$ units. These two sub-pseudo-populations are combined to give $U^* = \bigcup\{U_1^*, U_2^*\}$. The $\underline{y}^*$ values are easily obtained using $pspv(\underline{y}, s_k, U_k^*)$ for each sub-pseudo-population, $k = 1, 2$. As s_2 is selected using πpz, where $z_i = \hat{\mu}_i^{(1)}$, we also require $\underline{z}^*$ for all units in U^*. In this case the $\hat{\mu}_i^{(1)}$ are known for all units in U so the $\underline{z}^*$ are constructed in the same manner as $\underline{y}^*$. Bootstrap resamples are selected using the original sampling scheme so that a sample s_1^* of n_1 units is selected using *srswor* from U^* and a sample s_2^* of n_2 units is selected from $s_{1_c}^* = U^* - s_1^*$ using πpz^*. As it is the covariance $cov(\hat{\tau}_1, \hat{\tau}_2)$ that is of interest, which we estimate using the covariance of the $\underline{\hat{\tau}_1^*}$ and $\underline{\hat{\tau}_2^*}$, we must retain the same U^* for all bootstrap resamples. If each resample is selected from a different U_b^*, then the variability between the U_b^* would be too large to detect any pattern of covariance. The full bootstrapping procedure is outlined in box 8.2.

Table 8.2 shows the results of applying this bootstrap method. Using population B for a given $\hat{\mu}_i^{(1)}$, estimated using data from a sample $s^{(1)}$ selected using *srswor*, a sample $s_1^{(2)}$ of n_1 units was selected from U using *srswor* and a sample $s_2^{(2)}$ of n_2 units was selected from $s_{1_c}^{(2)}$ using $\pi p \hat{\mu}_i^{(1)}$. (This is the same strategy as implemented in chapter 4, although in that case, unconditional inclusion probabilities $\pi_i^{(2)}$ were approximated so that $\hat{\tau}^{(2)}$ was estimated using the Horvitz-Thompson estimator.) It is assumed that $\hat{\tau}^{(2)}$ will be

Box 8.2: A bootstrap strategy to estimate $cov(\hat{\tau}_1, \hat{\tau}_2)$ which s is selected using a two-phase sampling scheme.

1. Generate the bootstrap population $U^* = (U_1^*, U_2^*)$ where

   U_1^* is constructed using $ppswr(\frac{1}{pii[i1]} = \frac{N}{n_1}, \frac{Nn_1}{n_1+n_2}, s_1)$

   U_2^* is constructed using $ppswr(\frac{1}{\pi_{i2|1_c}}, \frac{Nn_2}{n_1+n_2}, s_2)$

2. Generate $\underline{y}^* = (\underline{y}_{1b}^*, \underline{y}_2^*)$ where

   $$\underline{y}_k^* = pspv(\underline{y}, s_k, U_k^*) \text{ for } k = 1, 2$$

3. Generate $\underline{z}^* = (\underline{z}_1^*, \underline{z}_2^*)$ where

   $$\underline{z}_k^* = pspv(\underline{z}, s_k, U_k^*) \text{ for } k = 1, 2$$

4. Take a sample, s_{1b}^*, of size n_1 using *srswor* from U_b^* and calculate

   $$\hat{\tau}_{1b}^* = \frac{N}{n_1} \sum_{i \in s_{1b}^*} y_{ib}^*$$

5. Take a sample, s_{2b}^* of size n_2 from $s_{1_cb}^* = U_b^* - s_{1b}^*$ using πpz_b^*, and calculate

   $$\hat{\tau}_{2b}^* = \frac{N}{N - n_1} \sum_{i \in s_{2b}^*} \frac{y_{ib}^*}{\pi_{i2|1_cb}^*}$$

6. Repeat steps 4 to 5, $b = 1, \ldots, B$ times to get

   $$\underline{\hat{\tau}}_1^* = (\hat{\tau}_{11}^*, \ldots, \hat{\tau}_{1B}^*)$$

   and $\underline{\hat{\tau}}_2^* = (\hat{\tau}_{21}^*, \ldots, \hat{\tau}_{2B}^*)$

7. Calculate the covariance between the two estimators as

   $$\widehat{cov}(\hat{\tau}_1, \hat{\tau}_2) = \frac{1}{B} \sum_{i=1}^{B} (\hat{\tau}_{1i}^* - \frac{1}{B} \sum_{j=1}^{B} \hat{\tau}_{1j}^*)(\hat{\tau}_{2i}^* - \frac{1}{B} \sum_{j=1}^{B} \hat{\tau}_{2j}^*)$$

Table 8.2: Bootstrap results for estimation of $cov(\hat{\tau}_1, \hat{\tau}_2)$ for population B using a two-phase sampling strategy. Results are mean (s.d) over 1000 simulations. $\widehat{cov}_{BS}$ is estimated from $B = 1000$ resamples. $\widehat{cov}_a$ is the analytic estimate of covariance using equation 8.17. cov_s is the empirical estimate of covariance from the 1000 estimates of τ using the data from s_1 and 1000 estimate using the data from s_{1_c}

$\frac{n_2}{n}$	$\widehat{cov}_{BS}$	$\widehat{cov}_a$	cov_s
0.26	-5,762 (10,479)	-5,605 (1,715)	-5,151
0.50	-5,948 (9,427)	-5,603 (2,143)	-5,487
0.74	-5,710 (11,076)	-5,688 (3,224)	-5,114

estimated using the two-phase estimator so that

$$\hat{\tau}_1^{(2)} = \sum_{s_1^{(2)}} \frac{y_i^{(2)}}{\pi_{i_1}^{(2)}}$$

$$\hat{\tau}_2^{(2)} = \sum_{s_2^{(2)}} \frac{y_i^{(2)}}{\pi_{i|s_{1_c}^{(2)}}^{(2)} \pi_{i_{1_c}}^{(2)}}$$

where $\pi_{i_1}^{(2)} = \frac{n_1}{N}$. Using the data in $s_1^{(2)}$ and $s_2^{(2)}$, the covariance $cov(\hat{\tau}_1^{(2)}, \hat{\tau}_2^{(2)})$ is estimated using the bootstrap estimator described in box 8.2 where $B = 1000$, to give $cov_{BS}(\hat{\tau}_1^{(2)}, \hat{\tau}_2^{(2)})$. In addition an analytic estimate of the covariance is calculated using

$$\widehat{cov}_a(\hat{\tau}_1^{(2)}, \hat{\tau}_2^{(2)}) = -\frac{N}{n_1}\left(\sum_{s_1^{(2)}} y_i^{(2)2} - \frac{1}{n_1 - 1}\sum_{s_1^{(2)}}\sum_{\substack{s_1^{(2)} \\ j \neq i}} y_i^{(2)} y_j^{(2)}\right) \tag{8.17}$$

The results presented are the mean (and standard deviation) from 1000 such samples.

Using the bootstrap estimator, we see that the results are much more variable than the analytic results. We note that the variability in the bootstrap estimators is greater when $\frac{n_2}{n}$ is not equal to 0.5. This may be because in these cases one of the sub-pseudo-populations is obtained from a sub-sample s_1 or s_2 of only 12 units. The analytic estimator is less

variable than the bootstrap estimator. The empirical estimates of covariance are not based on the covariance of τ_1 and τ_2 because of the high variability in τ_2. Instead of estimating τ_2 using the data in s_2 we estimate τ_2' using the data from s_{1_c} which is collected using *srswor*. This covariance appears to be an underestimate of the true covariance.

8.4 A bootstrap strategy for covariance estimation for a two-phase sampling scheme through time

Here we consider the sample design described in Chapter 5 in which $s_1^{(2)}$ is selected from $s^{(1)}$ using *srswor*, and $s_2^{(2)}$ is selected from $s_c^{(1)}$ using $\pi p\hat{\mu}^{(1)}$. Our aim is to propose a bootstrapping strategy that would enable the covariance $cov(\hat{\tau}^{(1)}, \hat{\tau}^{(2)})$ to be estimated using all of the data from $s^{(1)}$ and $s^{(2)}$. In the bootstrapping strategies proposed above, an essential property was that the bootstrap resamples are selected from the pseudo-population using the same sampling design as used to select the original samples. Also the estimates of the pseudo-totals are calculated using the original estimators. Hence to estimate $cov(\hat{\tau}^{(1)}, \hat{\tau}^{(2)})$ we will require a pseudo-population U^* that has two sets of pseudo-values.

$$\underline{y}^{(1)*} = \left\{y_1^{(1)*}, \ldots, y_N^{(1)*}\right\} \text{ with auxiliary variables } \mathbf{x}^{(1)*} = \left\{\underline{x}_1^{(1)*}, \ldots, \underline{x}_N^{(1)*}\right\}$$
$$\text{and } \underline{y}^{(2)*} = \left\{y_1^{(2)*}, \ldots, y_N^{(2)*}\right\} \text{ with auxiliary variables } \mathbf{x}^{(2)*} = \left\{\underline{x}_1^{(2)*}, \ldots, \underline{x}_N^{(2)*}\right\}$$

The main focus of this section is how the pseudo-population and pseudo-values are created but first we describe how the bootstrapping will proceed once they have been created. The bootstrapping procedure will be:

1. Take a sample $s_b^{(1)*}$ selected using *srswor* from the pseudo-population U^* and estimate $\hat{\tau}_b^{(1)*}$;
2. Construct a model $\zeta^{(1)*}$ and estimate $\hat{\mu}_i^{(1)*}$ for $i \in U^*$;
3. Take a sample $s_{1,b}^{(2)*}$ selected from $s_b^{(1)*}$ using *srswor* and estimate $\hat{\tau}_{1,b}^{(2)*}$;

4. Take a sample $s_{2,b}^{(2)*}$ selected from $s_{b_c}^{(1)*} = U^* - s_b^{(1)*}$ using $\pi p \hat{\mu}^{(1)*}$ and estimate $\hat{\tau}_{2,b}^{(2)*}$;

5. Estimate $\hat{\tau}^{(2)*}$;

6. Repeat the above steps $b = 1, \ldots, B$ times and estimate

$$\begin{aligned}\widehat{cov}_{BS}(\hat{\tau}^{(1)}, \hat{\tau}^{(2)}) =& cov(\underline{\hat{t}}^{(1)*}, \underline{\hat{t}}^{(2)*}) \\ =& \left(\sum_{b=1}^{B} \tau_b^{(1)*} - \frac{1}{B}\sum_{b=1}^{B} \tau_b^{(1)*}\right)\left(\sum_{b'=1}^{B} \tau_{b'}^{(2)*} - \frac{1}{B}\sum_{b'=1}^{B} \tau_{b'}^{(2)*}\right)\end{aligned}$$

Step 2 is important because in the original sampling strategy we could not obtain the unconditional inclusion probability that $i \in s^{(2)}$ as this requires $\hat{\mu}_i^{(1)}$ for all possible $s^{(1)}$, and it is only known for the specific $s^{(1)}$ that is selected. Because of this we had to estimate $\hat{\tau}^{(2)}$ using the two-phase estimator in which we require the inclusion probability that $i \in s_2^{(2)}$ given the sample $s^{(1)}$. As $s^{(1)}$ varies so does $\hat{\mu}_i^{(1)}$ and hence this variability needs to be incorporated into the bootstrap strategy.

We are estimating the covariance as part of $s^{(2)}$ is selected from $s^{(1)}$ and we believe the $y_i^{(1)}$ and $y_i^{(2)}$ to be correlated. Therefore the pseudo-values $y_i^{(1)*}$ and $y_i^{(2)*}$ must also exhibit the same correlation pattern. The pseudo-values, and the pseudo-population, are generated using the sample data. The data available are from the samples

(a) $s_1^{(1)} = s_1^{(2)}$, the set of matched units which we will denote s_1. Unlike previous chapters in which $s^{(1)} = s_1^{(1)}$ we separate $s^{(1)}$ into its matched and unmatched components

(b) $s_2^{(1)}$ selected using *srswor*

(c) $s_2^{(2)}$ selected using $\pi p \hat{\mu}^{(t-1)}$.

We have information about both $y_i^{(1)}$ and $y_i^{(2)}$ for the units in the matched sample s_1. We could therefore create the pseudo-population and pseudo-values using only the data from this matched sample. Our aim however is to use all the data in $s^{(1)}$ and $s^{(2)}$ to create the pseudo-populations.

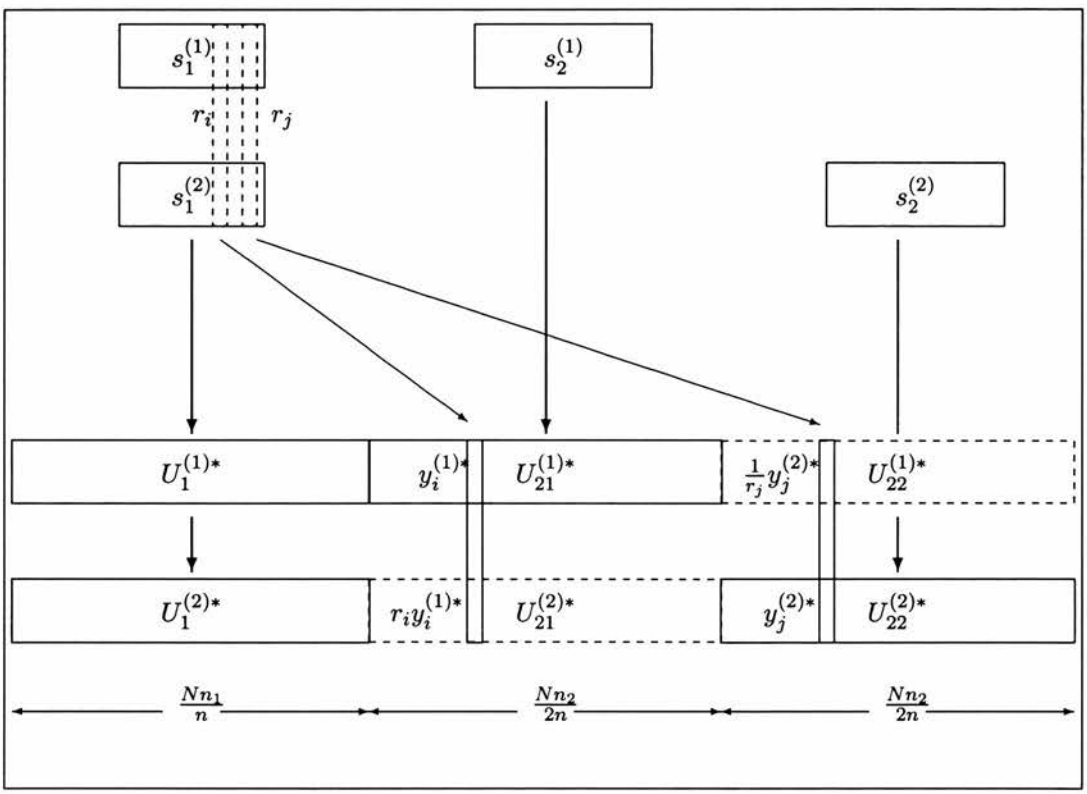


Figure 8.1: A bootstrap strategy for sampling through time. The samples, $s_1^{(t)}$ and $s_2^{(t)}$ for $t = 1, 2$, are used to create a pseudo-population with two sets of data $U^{(1)*}$ and $U^{(2)*}$. Further explanation is provided in the text.

To do this we generate three sub-pseudo-populations $\{U_1^*, U_{21}^*, U_{22}^*\}$ using the units from s_1, $s_2^{(1)}$ and $s_2^{(2)}$ respectively. As both $s^{(1)}$ and $s^{(2)}$ are of equal size we want both samples to contribute equally to the generation of U^*. Hence $\frac{N}{2}$ units are generated using the data from $s^{(1)}$. Using the same process as the previous section we want to generate $\frac{n_1}{n}\frac{N}{2}$ units from $s_1^{(1)}$ and $\frac{n_2}{n}\frac{N}{2}$ units from $s_2^{(1)}$. We apply the same allocation rule for $s^{(2)}$ as $n_1^{(1)} = n_1^{(2)}$ so that in total $\frac{n_1}{nN}$ units are generated from $s_1 = s_1^{(1)} = s_1^{(2)}$. The pseudo-population is therefore made-up of:

(i) U_1^* consisting of $\frac{Nn_1}{n}$ units from s_1 $\quad ppswr(\frac{N}{n_1}, \frac{Nn_1}{n}, s_1)$

(ii) U_{21}^* consisting of $\frac{Nn_2}{2n}$ units from $s_c^{(1)}$ $\quad ppswr(\frac{n}{n_1}, \frac{Nn_1}{n}, s_c^{(1)})$

(iii) U_{22}^* consisting of $\frac{Nn_2}{2n}$ units from $s_2^{(t)}$ $\quad ppswr(\frac{1}{\pi^{(2)}_{2|s_c^{(1)}}}, \frac{Nn_2}{n}, s_c^{(1)})$

The generation of the pseudo-population and of the pseudo-values is shown in figure 8.1. The original samples are illustrated at the top of the figure and the vertical arrows point to the corresponding pseudo-populations at the bottom of the figure. For example: $\boxed{s_1^{(2)}} \downarrow \boxed{U_{21}^{(1)*}}$

As the sub-pseudo-population U_1^* is generated from $s_1 = s_1^{(1)} = s_1^{(2)}$ both $y_i^{(1)}$ and $y_i^{(2)}$ are known for all $i \in s_1$. Hence we can obtain the pseudo-values $y_i^{(1)*}$, using the process $pspv(\underline{y}^{(1)}, s_1, U_1^*)$, and the pseudo-values $y_i^{(2)*}$, using the process $pspv(\underline{y}^{(2)}, s_1, U_1^*)$.

The sub-pseudo-population U_{21}^* is generated from $s_2^{(1)}$. For $i \in s_2^{(1)}$ we only know $y_i^{(1)}$ and not $y_i^{(2)}$ and so although we can obtain the pseudo-values $y_i^{(1)*}$, using the process $pspv(\underline{y}^{(1)}, s_2^{(1)}, U_{21}^*)$, we cannot obtain the pseudo-values $y_i^{(2)*}$ this way. Instead we need to find a strategy to "create" the pseudo-population. In a similar manner, for the third sub-pseudo-population, U_{22}^*, we can obtain the pseudo-values $y_i^{(2)*}$ using the process $pspv(\underline{y}^{(2)}, s_2^{(2)}, U_{22}^*)$ and we require a strategy to "create" the pseudo-values $y_i^{(1)*}$. In figure 8.1 the pseudo-values that are "created" are represented as $\boxed{\boxed{U_{21}^{(2)*}}}$ and $\boxed{\boxed{U_{22}^{(1)*}}}$ whereas the pseudo-values that are generated from the sample values are represented as

$\boxed{U_{21}^{(1)*}}$ and $\boxed{U_{22}^{(2)*}}$.

For $i \in U_{21}^{*}$ the unknown pseudo-values $y_i^{(2)*}$ must be correlated with the $y_i^{(1)*}$. Although we do not know the value of $y_i^{(2)}$ for the sub-sample $s_2^{(1)}$, from which U_{21}^{*} is created, we do know the relationship between $y_i^{(1)}$ and $y_i^{(2)}$ for the matched sample. For the types of population we are studying we are particularly interested in the ratio

$$r_i = \frac{y_i^{(2)}}{y_i^{(1)}} \text{ for } i \in s_1$$

For the j^{th} unit in U_{21}^{*} we know the pseudo-value $y_j^{(1)*}$ but not the pseudo-value $y_j^{(2)*}$. However if we had a ratio r_j we could use this to create the pseudo-value $y_i^{(2)*}$ where

$$y_i^{(2)*} = r_j y_i^{(1)*}$$

We can obtain a ratio r_j by selecting a unit j from s_1 using *srswor* and calculating the ratio using $y_j^{(1)}$ and $y_j^{(2)}$. This strategy of firstly selecting a ratio at random from s_1 and then using this to calculate a new pseudo-value can be repeated for all units in U_{21}^{*}. In a similar manner we can create the pseudo-values $y_j^{(1)*}$ in U_{22}^{*} where

$$y_j^{(1)*} = \frac{1}{r_i} y_j^{(2)*}$$

This is shown in figure 8.1 by the dashed boxes in $s^{(1)}$ and $s^{(2)}$ giving the r_i and r_j that are used to generate pseudo-values. One issue that arises is how r_i should be calculated if $y_i^{(t')} = 0$. Rather than seeking a replacement unit i from s_1 the average ratio $\bar{r}_{s_1}$ is used where

$$\bar{r}_{s_1} = \frac{1}{n_1} \sum_{s_1} \frac{y_i^{(2)}}{y_i^{(1)}}$$

Each unit in U has a vector of auxiliary data $\underline{x}_i^{(t)}$ associated with it for $t = 1, 2$. These data are attributed to the units in U^* so that

$$\underline{x}_i^{(t)*} - pspv(\underline{x}_i^{(t)}, U, U^*) \text{ for } t = 1, 2$$

The strategy described above can be extended if $cov(\hat{\tau}^{(t')}, \hat{\tau}^{(t)})$ must be estimated for many different surveys $t' = 1, \ldots, T-1,\ t = 2, \ldots, T$. With T surveys in total, each unmatched sample $s_2^{(t)}$ contributes $\frac{Nn_2}{Tn}$ units to U^*. Box 8.3 summarises the strategy to generate U_b^* described above in the general case when there are T surveys.

We have assumed that the ratio r_i is the appropriate method for describing the relationship between the $y_i^{(t')}$ and $y_i^{(t)}$. This is appropriate for the populations A and B where $log(\mu_i^{(t)}) = r^t \mu_i^{(0)}$. However for other populations the difference, $y_i^{(t)} - y_i^{(t')}$, or some other measure may be deemed more appropriate. A further assumption is that the ratio r_i does not depend on the auxiliary variable x_i. That is the rate of change from one survey to another is constant over the whole population and does not change more rapidly in some areas than others. This will not always be a valid assumption particularly if T is large.

Simulation work is needed to explore how effective this strategy is for covariance estimation. One potential difficulty is that the estimation of $\hat{\mu}^{(t*)}$ may be poor, because the $y_i^{(t)*}$ are constructed independently of the $x_i^{(t)*}$. This may lead to a high variability in the $s_2^{(t)}$ so that the estimated covariance is more variable than the analytic strategy in which the variability in $s_2^{(t)}$ is not included in the estimates of covariance.

8.5 Discussion

The motivation for this chapter was to estimate covariances such as $cov(\hat{\tau}_1^{(t)}, \hat{\tau}_2^{(t)})$ or $cov(\hat{\tau}^{(t')}, \hat{\tau}^{(t)})$ using all of the data from surveys $s^{(t')}$ and $s^{(t)}$, rather than using none of the data in $s_2^{(t)}$.

We choose the method of bootstrapping as it is relatively simple to implement and can easily be applied to different forms of sampling designs. To be able to develop a bootstrapping strategy for covariance estimation, we require a strategy for estimating $var(\hat{\tau}^{(t)})$ when $s^{(t)}$ is selected using πpz. Current strategies cannot be applied when $\frac{1}{\pi_i}$ is non-integer, and in addition they cannot be easily adapted to more complex sampling schemes.

Box 8.3: A bootstrapping strategy to generate U^* to estimate of $cov(\hat{\tau}^{(t')}, \hat{\tau}^{(t)})$ for $t' = 1, \ldots, T-1$ and $t = 2, \ldots, T$

1. Generate a pseudo-population $U^* = \bigcup(U_1^*, U_{21}^*, \ldots, U_{2t}^*, \ldots, U_{2T}^*)$ where

$$U_1^* \text{ using } ppswr(\pi_{i1} = \frac{n_1}{N}, N_1^* = \frac{Nn_1}{n_1 + n_2}, s_1)$$

$$U_{2t}^* \text{ using } ppswr(\pi_{i2|1^c}^{(t)}, N_{2t}^* = \frac{Nn_2}{T(n_1 + n_2)}, s_2^{(t)}) \text{ for } t = 1, \ldots, T$$

2. Obtain the auxiliary data

$$uux_{U^*}^{(t)*} = pspv(\underline{x}^{(t)}, U, U^*) \text{ for } t = 1, \ldots = T$$

3. Obtain reference pseudo–populations, U_{rt}^*, and ratio data, $\underline{r}_t^{*(t')}$, for $t = 1, \ldots, T$, and $t' \neq t$

$$U_{rt}^* \text{ using } ppswr(\frac{n_1}{N}, N_{2t}^*, s_1)$$
$$\underline{r}_t^{*(t')} = pspv(\underline{y}^{(t')}/\underline{y}^{(t)}, s_1^{(t',t)}, U_{rt}^*)$$

If $y_i^{(t)} = 0$ then $r_{it}^{*(t')}$ is replaced by the average ratio

4. Obtain $\underline{y}^{*(t)}$ for $t = 1, \ldots, T$

$$\underline{y}_1^{*(t)} = pspv(\underline{y}^{(t)}, s_1^{(t)}, U_1^*)$$
$$\underline{y}_{2t}^{*(t)} = pspv(\underline{y}^{(t)}, s_2^{(t)}, U_{2t}^*)$$
$$\underline{y}_{2t'}^{*(t)} = \underline{r}_t^{*(t')} pspv(\underline{y}^{(t')}, s_2^{(t')}, U_{2t'}^*) \text{ for } t' \neq t$$

We have proposed a set of four bootstrap strategies that can be used when s is selected using πpz. All four strategies estimate $var_a(\hat{\tau}_s)$ with a small amount of bias when n/N is small. The strategy of Sverchkov and Pfeffermann (2003) is one of the proposed methods and has a low bias. We show that an improvement to their strategy is to construct a pseudo-population in which n of the units are the sample s. The reduction in bias is small when n is small. These strategies are an improvement on previous strategies in that they can be applied when $\frac{1}{\pi_i}$ is non-integer and can be adapted for more complex sampling schemes.

We have shown how this bootstrapping strategy can be used to estimate the covariance $cov(\hat{\tau}_1^{(t)}, \hat{\tau}_2^{(t)})$ when $s_1^{(t)}$ is selected independently of $s^{(t-1)}$. However the estimated covariance is far more variable than that using the analytic estimator, although we also note that we require further work to fully understand these covariances. However we would suggest that this method of generating the pseudo-population may be appropriate for more complex estimators when an analytic expression cannot be obtained.

When $s_1^{(t)}$ is selected from $s^{(t-1)}$ and $s_2^{(t)}$ is selected so that $\pi^{(t)}_{i|s_{1c}^{(t-1)}}$ is a function of $\hat{\mu}_i^{(t-1)}$ we saw that a more complex bootstrapping strategy is required to estimate $cov(\hat{\tau}_1^{(t)}, \hat{\tau}_2^{(t)})$ or $cov(\hat{\tau}^{(t')}, \hat{\tau}^{(t)})$. This requires the generation of y_i^* values, and hence some form of assumption about how the population changes through time. The strategy has not as yet been tested but we have noted potential problems with it and it is not clear whether the benefit from using all of the sample data in the bootstrapping strategy will lead to more precise estimates of covariance than analytic strategies, because of the assumptions that need to be made about the population.

The original motivation for developing these bootstrap strategies was to estimate the covariances in our two-phase combined sampling strategy. Although there are no results here for the covariance estimation using the bootstrap when sampling through time, which assumes that $\hat{\mu}_i^{(2)}$ varies with $s^{(1)}$, the results for covariance estimation assuming a fixed $\hat{\mu}_i^{(2)}$ would suggest that the bootstrap estimate of covariance is much more variable than the analytic estimate of covariance. Hence although testing of the bootstrap estimate of

covariance through time will be of interest we would in practice expect to use the analytic estimate of covariance.

Chapter 9

Discussion and Further Work

9.1 Discussion

The aim of this thesis is to develop an adaptive design-based sampling strategy that can be used in a single-species monitoring programme. The primary objectives of the programme are to estimate $\tau^{(t)}$, the total number of individuals in the area of interest at the time of survey, t for $t = 1, \ldots$, and to estimate $\delta^{(t',t)}$ the change in the total number of individuals in the survey region from time t' to time t. In addition we wish to observe as many individuals of the study species as possible.

We started from the premise that although auxiliary data may be available at the start of the monitoring programme we know little about the species distribution over the survey region. However as monitoring progresses we expect to learn more about the spatial distribution of the species over the survey region and in this thesis we have investigated how this knowledge can be incorporated into the survey design. In general we have restricted this investigation to plot sampling of motile species that do not occur in clusters. Therefore, a useful summary of the spatial distribution of the species is the expected number of individuals in a sampling unit, $\mu_i^{(t)}$. Using data from the first survey we can construct a model $\zeta^{(1)}$ that describes the relationship between auxiliary variables x_{ij} and $\mu_i^{(1)}$. Using

this model we can estimate $\mu_i^{(1)}$ for all units in the survey region.

In Chapter 4 we demonstrated how $\mu_i^{(2)}$ could be used as the variable for stratification (*strs* $\mu^{(2)}$) or for sampling with inclusion probability proportional to $\hat{\mu}_i^{(2)}$ ($\pi p\mu^{(2)}$) to estimate $\tau^{(2)}$. This was most precisely estimating using $\pi p\mu^{(2)}$. In practice $\mu_i^{(2)}$ can only be estimated so we derived a measure b to represent how good an estimate $\hat{\mu}_i^{(2)}$ is of $\mu_i^{(2)}$ where $\hat{\mu}_i^{(2)} = \mu_i^{(2)b}$. When $b = 1$ the model is well specified. As b moves away from one the model is less well specified and the precision of $\hat{\tau}^{(2)}$ decreases. The rate of change is greater under $\pi p\mu^b$ than under *strs* μ^b. Empirical evidence suggests that b commonly takes values between zero and two and when $|b-1|$ is greater than approximately 0.5 the precision of $\hat{\tau}^{(t)}$ is greater under *strs* μ^b than under $\pi p\mu^b$.

The combined sampling design we have proposed consists of selecting a sub-sample $s_1^{(t)}$, of size $n_1^{(t)}$, using an equal probability design, such as *srswor* or *sys*, and selecting a second sub-sample $s_2^{(t)}$, of size $n_2^{(t)} = n - n_1^{(t)}$ units, using $\pi p\hat{\mu}^{(t)}$. When $\frac{n_2^{(t)}}{n} = 0$ the sample design is *srswor*, or *sys*, and when $\frac{n_2^{(t)}}{n} = 1$ the sample design is $\pi p\hat{\mu}^{(t)}$. The first sub-sample $s_1^{(t)}$ is selected either from U or from $s_1^{(t-1)}$ and the second sub-sample $s_2^{(t)}$ is selected from $s_{1_c}^{(t)}$ or $s_{1_c}^{(t-1)}$ respectively. The latter approach is taken when an estimate of $\delta^{(t',t)}$ is of interest.

An estimate of $\tau^{(t)}$ is relatively easy to obtain when $s_1^{(t)}$ is selected from U. The unconditional inclusion probabilities are approximated and the Horvitz-Thompson estimator used to estimate $\tau^{(t)}$. When $b = 1$ the precision of $\hat{\tau}^{(t)}$ decreases as $\frac{n_2^{(t)}}{n}$ decreases. When $b > 1$ the combined sampling strategy where $0 < \frac{n_2^{(t)}}{n} < 1$ gives more precise estimates of $\tau^{(t)}$ than under $\pi p\hat{\mu}^{(t)}$ sampling strategy. So the combined sampling strategy is more robust to model mis-specification.

When $\delta^{(t',t)}$ and $\tau^{(t)}$ are both of interest the combined sampling strategy in which $s_1^{(t)}$ is selected from $s_1^{(t-1)}$ is employed. The estimate of $\tau^{(t)}$ requires the use of a two-phase sampling estimator because the probability that unit i is included in $s_2^{(t)}$ depends on the sample $s_1^{(t-1)}$, and different $s_1^{(t-1)}$ will give different estimates of $\hat{\mu}_i^{(t)}$. Only the estimate

of $\hat{\mu}_i^{(t)}$ from the sample $s^{(t-1)}$ is known and so unconditional inclusion probabilities that average over all possible $s^{(t-1)}$ samples cannot be obtained. The two-phase estimator provides a strategy for coping with this difficulty. Using this strategy the precision of $\hat{\tau}^{(t)}$ increases as $\frac{n_2^{(t)}}{n}$ increases for $0 < b < 2$. In contrast the estimate of $\delta^{(t',t)}$ becomes more precise as $\frac{n_2^{(t)}}{n}$ decreases as a greater proportion of the sample is retained from one survey to another. Hence the relative importance of the two parameters will partly determine the selection of the appropriate value of $\frac{n_2^{(t)}}{n}$ and will also depend on the correlation between $y_i^{(t)}$ and $y_i^{(t')}$. The greater this correlation, the greater the potential increase in precision in $\delta^{(t',t)}$ by setting $\frac{n_2^{(t)}}{n}$ to be low. In comparison the precision of $\tau^{(t)}$ will increase as the correlation between $y_i^{(t)}$ and $\hat{\mu}_i^{(t)}$ increases and so to estimate $\tau^{(t)}$ precisely $\frac{n_2^{(t)}}{n}$ should be high.

In a monitoring programme, the choice of sampling strategy for survey t depends on the beliefs about how the spatial distribution of the population changes through time as well as the relative importance of the parameters to be estimated and whether units can be retained from one survey to another. If we believe that the change in the spatial distribution of the population through time can be modelled, so that data from past surveys can be used in the construction of the model $\zeta^{(t)}$, then we would expect $\hat{\mu}_i^{(t)}$ to become more accurate, and hence better specified (that is b close to one), through time. In these circumstances, a combined sampling strategy can increase the precision of $\hat{\tau}^{(t)}$ through time. In particular as more is learnt about the spatial distribution, the proportion $\frac{n_2^{(t)}}{n}$ can increase, further increasing the precision of $\tau^{(t)}$. If we cannot learn through time, because we believe the model $\zeta^{(t)}$ changes rapidly through time, only the data from the most recent survey can be used to estimate $\zeta^{(t)}$. Then the increase in the precision of $\tau^{(t)}$ is small through time as $\frac{n_2^{(t)}}{n}$ must be kept low. If $\frac{n_2^{(t)}}{n}$ is high, the sample will be concentrated in a small part of the survey region and model mis-specification is likely to increase through time and hence the precision of $\hat{\tau}^{(t)}$ will decrease.

In practice, a monitoring strategy that is commonly used to obtain precise estimates of $\tau^{(t)}$ and $\delta^{(t',t)}$, is to use a form of rotating panel design and to estimate $\tau^{(t)}$ and $\delta^{(t',t)}$ using

model-assisted estimators. In the rotating panel design, some units are retained from one survey to another and new units are added into the sample in each survey. Usually units are removed and selected using simple sampling design such as *srswor* so that the samples for many surveys could be selected at the start of the monitoring programme. Data from past surveys are used to improve the estimates of $\tau^{(t)}$ and $\delta^{(t',t)}$ rather than to inform survey design and sample selection.

These strategies are often employed when the objectives of the monitoring programme require the estimation of many parameters for a number of different variables, for example the EMAP project (Overton *et al.*, 1990). In this case it is important that the survey design is relatively simple, as different parameters have very different spatial distributions over the survey region. The choice of inclusion probabilities that would lead to precise estimates of one parameter of interest could lead to very imprecise estimates of another parameter of interest. When samples are selected using *srswor* the precision of the estimates can be increased by using the appropriate model-assisted estimator. Generally these estimators use data from previous surveys and auxiliary variables are not incorporated into the estimation process although recent work by Fuller and Rao (2001) and Singh (1996) does enable auxiliary data to be incorporated when the regression model is simple.

When $s^{(t)}$ has been selected using *srswor*, a model-assisted estimate of $\tau^{(t)}$ will generally be more precise than a design-based estimate. The reason is that $var(\hat{\tau}^{(t)})$ is proportional to $e_i^2 = (y_i - \hat{y}_i)^2$ in the model-assisted case and with y_i^2 when $\tau^{(t)}$ is estimated using a design-based estimator, and under a reasonable model $e_i^2 < y_i^2$. When $s^{(t)}$ is selected using an unequal probability design, a model-assisted estimator will often not increase the precision of $\tau^{(t)}$ compared to a design-based estimator because $var(\hat{\tau}^{(t)})$ will be low when e_i or y_i respectively, is correlated with π_i. Although the $y_i^{(t)}$ are correlated with π_i, so that $var(\hat{\tau}^{(t)})$ is low, e_i is not necessarily correlated with π_i and so the model-assisted estimate of $\tau^{(t)}$ is less precise than the design-based estimate.

Therefore, under a combined sampling design it is generally not possible to increase the precision of the estimate of $\hat{\tau}^{(t)}$ once the data collection stage has been completed by using

a model-assisted estimator. This was demonstrated in Chapter 4. For the estimation of $\tau^{(t)}$ and $\delta^{(t',t)}$, our combined sampling strategies are relatively robust to model mis-specification, so using a design-based estimator under our sampling scheme will give precise estimates of $\tau^{(t)}$ and $\delta^{(t',t)}$. If other parameters are of interest, which are based on variables other than $y_i^{(t)}$ then design-based estimates of these parameters may be inefficient and may not necessarily be improved by using a model-assisted estimator. Hence our combined sampling designs are good for the initially defined parameters, but may be less successful at estimating other parameters which post-hoc are deemed to be of interest. This compares to the model-assisted strategy in which estimates of other parameters can also be estimated precisely using a model-assisted estimator.

A disadvantage of the model-assisted strategies for estimating $\tau^{(t)}$ or $\delta^{(t',t)}$ is that a simple regression model is often not sufficient for describing the relationship between the observed counts, the $y_i^{(t)}$, and the auxiliary variables, $x_{ij}^{(t)}$. In particular, a suitable model $\zeta^{(t)}$ could be of the form $E[Y_i^{(t)}] = \mu_i^{(t)} = log(\sum_{j=0}^{Q} f_j^{(t)}(x_{ij}^{(t)}))$ where $f_j^{(t)}$ is a non-linear or semi-parametric function. Recent methods of Wu and Sitter (2001) have developed model-assisted strategies for the case when $g(\mu_i^{(t)}) = \sum_{j=0}^{Q} f_j^{(t)}(x_{ij}^{(t)})$ and Breidt and Opsomer (2000) can incorporate semi-parameteric functions into the estimators. In both cases these have only been developed for estimating $\tau^{(t)}$ using data from survey t and have not been extended, or incorporated into the model-assisted estimators through time of Fuller and Rao (2001) and Singh (1996). Hence current model-assisted methods can either learn through time by estimating $\tau^{(t)}$ or $\delta^{(t',t)}$ using past $y_i^{(t)}$, or alternatively can use the $y_i^{(t)}$ and auxiliary data from survey t to estimate $\tau^{(t)}$ but cannot use data from previous surveys to improve the estimate. In comparison, the combined sampling strategies we propose can use auxiliary data and the observed counts from past surveys to estimate $\hat{\mu}_i^{(t)}$ and so increase the precision of $\tau^{(t)}$

An additional objective of our monitoring programmes is to increase the number of individuals that are observed as more is learnt about the spatial distribution of the species over the survey region. Under the combined sampling strategies we expect $\sum_{s^{(t)}} y_i^{(t)}$ to in-

crease as $\frac{n_2^{(t)}}{n}$ increases if the distribution of the $y_i^{(t)}$ is not negative skew and $\hat{\mu}_i^{(t)}$ is well specified as future samples target areas that are expected to have high values of $y_i^{(t)}$. This is not the case under a model-assisted strategy in which units are selected using *srswor*.

In Chapter 3 we stated that it was important that our sampling strategy enabled fieldworkers to see as many individuals of the study species as possible, so that they are not discouraged by constantly returning to areas that have low, or zero, abundance. Assuming a population generated by a superpopulation model that remains constant through time, we would expect $\sum_{s^{(t)}} y_i^{(t)}$ to increase if $\hat{\mu}_i^{(t)}$ is well-specified when $s^{(t)}$ is selected using a combined sampling design in which $\frac{n_2^{(t)}}{n} > 0$. We would not expect $\hat{\mu}_i^{(t)}$ to be well specified if $\sum_{s^{(1)}} y_i^{(1)}$ is very low. In this case our premise that our strategy is successful in preventing fieldworkers from returning to areas of low abundance may not hold. Either fieldworkers see few individuals and continue to do so for several surveys until $\hat{\mu}_i^{(t)}$ is better specified, or fieldworkers start by seeing many individuals so that the original problem of sampling low abundance areas does not exist. In this case $\hat{\mu}_i^{(1)}$ will be well-specified and more individuals are observed in future surveys.

Increasing $\sum_{s^{(t)}} y_i^{(t)}$ through time can be useful even if there are not areas of the survey region which are sparsely populated at time $t = 1$. If the population is decreasing through time, particularly if the population is decreasing but the relationship between auxiliary variables and species density remains constant, then fieldworkers may in future surveys spend large amounts of time surveying areas that are sparsely populated unless a combined sampling strategy is implemented. For any population, whether it is increasing or decreasing, other data may be recorded about the individuals observed and so parameters related to characteristics of individuals of the species can be estimated more precisely by observing more individuals, particularly if these estimates are obtained within a model-based framework.

We have shown how the combined sampling strategies can be applied to distance sampling as well as plot sampling to estimate the density and hence total number of individuals in the survey region. This has been limited to assuming that the probability of detecting

an individual depends only on the distance of the individual from a transect line. There is a clear advantage here to increasing $\sum_{s^{(t)}} y_i^{(t)}$ as this can increase the precision of $\hat{f}(0)$, the probability of detecting an individual on the transect line. Under combined sampling we expect the variance of both $\hat{f}(0)$ and of the encounter rate to decrease compared to selecting the transects using *srswor*.

Our sampling strategy is adaptive as it changes the survey design through time based on previously observed data. In Chapter 3 we described how adaptive cluster sampling could be implemented through time for a sessile species, although we do not know how to derive an estimator for $\tau^{(t)}$ using this sample design. In the first survey, $s^{(1)}$ is selected using *srswor* and the data used to estimate $\tau^{(1)}$. In the second survey sample, $s^{(2)}$ consists of units in the neighbourhood of units in $s^{(1)}$ where $y_i^{(1)} > \mathcal{C}$ and in the third survey sample, $s^{(3)}$ consists of units in the neighbourhood of units in $s^{(2)}$ where $y_i^{(2)} > \mathcal{C}$ and so on .... Compare this to our combined sampling strategy when the model $\zeta^{(1)}$ is constructed using the data from $s^{(1)}$ and the only auxiliary variables available are latitude and longitude. One strategy for modelling $\mu_i^{(t)}$ is to fit a thin-plate spline of latitude and longitude. The predicted values of $\hat{\mu}_i^{(t)}$ would only be high for units close to the units $i \in s^{(1)}$ where $y_i^{(1)}$ is high. If $s^{(2)}$ is selected using $\pi p \hat{\mu}^{(1)}$ then most units in $s^{(2)}$ will occur close to units in $s^{(1)}$. So the sample selected would closely resemble the adaptive cluster sampling design through time in which $s^{(2)}$ is obtained by sampling the neighbourhood of the units for which $y_i^{(1)} > \mathcal{C}$. In a combined sampling strategy, where $\frac{n_2^{(t)}}{n} < 1$ then part of the sample would also be selected using *srswor*.

A sample obtained using the combined sampling design also resembles samples obtained using the model-based adaptive sampling strategy of Chao and Thompson (2001). In both cases a sample $s_1^{(t)}$ is selected using *srswor* or *sys* and a second sample is $s_2^{(t)}$ is selected where most units occur in areas of high abundance. Chao and Thompson (2001) seek a selection of units $s_2^{(t)}$ that minimise the mean-square prediction error by given a model constructed using the observed data in $s_1^{(t)}$. Both the data from $s_1^{(t)}$ and $s_2^{(t)}$ are used to estimate $\hat{\tau}^{(t)}$. In comparison our model is constructed from the data obtained in survey

$s^{(t-1)}$ and we do not seek an optimal sample.

The basic components of our strategy indicates a potential sampling design that falls between the fully design-based adaptive sampling strategy and the optimal adaptive sampling strategy for estimating $\tau^{(t)}$. A sub-sample $s_1^{(t)}$ could be selected using *srswor* and the data from this sub-sample used to construct a model $\zeta^{(1)}$ and therefore estimate $\hat{\mu}_i^{(1)}$. A second sub-sample $s_2^{(t)}$ would then be selected using $\pi p \hat{\mu}^{(1)}$. The data from both $s_1^{(t)}$ and $s_2^{(t)}$ would be used to estimate $\tau^{(1)}$. This estimate may be biased as are the strategies of Francis (1984) and Jolly and Hampton (1990) in which units in $s_2^{(t)}$ are allocated to strata based on the $y_i^{(1)}$ observed for $i \in s_1^{(t)}$. This proposed strategy demonstrates how our combined sampling strategy is different to standard adaptive cluster sampling in that $\hat{\mu}_i^{(t)}$ can be obtained for all units in U rather than using the observed counts, $y_i^{(t)}$ for units in $s^{(t)}$ and is different to model-based optimal sampling designs as the model helps guide design decisions but $s^{(t)}$ is still a probability sample, rather than choosing a sample that minimises some criterion.

Our form of combined sampling strategy has a different philosophy to other sampling strategies. The simplest sampling strategies are design-based strategies in which $s^{(t)}$ is selected using a probability sampling scheme, that may take account of auxiliary variables, and $\tau^{(t)}$ is estimated using a design-based estimator such as the Horvitz-Thompson estimator. This has no formal framework for learning through time. An alternative estimator for a sample selected using a simple probability sampling scheme is to use a model-assisted estimator. This may incorporate auxiliary information and the estimator can 'learn' through time as past survey data can be used to model how the $y_i^{(t)}$ change through time. The sampling units may change through time, but sample selection is carried out independently of the observed data. In the adaptive framework, an initial design-based sample is selected and the observed $y_i^{(t)}$ used: to select more units in that sample the adaptive cluster sampling strategies of Thompson and Seber (1996) or of Francis (1984) and Jolly and Hampton (1990) to estimate $\tau^{(t)}$; or to select the sample at time $t+1$ to estimate $\tau^{(t+1)}$, the adaptive strategy of Haines and Pollock (1998) in which some some new units

are selected and some units in $s^{(t)}$ for which $y_i^{(t)} > C$ are returned to. Alternatively within the adaptive framework an initial probability sample is selected that is used to estimate $\mu_i^{(t)}$. From this an optimal sample is selected to obtain a model-based estimate of $\tau^{(t)}$, i.e. the strategy of Chao and Thompson (2001), or is used to adapt environmental monitoring networks such as the work by Wikle and Royle (1999).

In comparison we initially take a probability sample $s^{(1)}$ and obtain a design-based estimate of $\hat{\tau}^{(1)}$. At all future times t we use a model $\zeta^{(t-1)}$ from which we estimate $\mu_i^{(t)}$ which are used to determine the inclusion probabilities $\pi_i^{(t)}$ so that $s^{(t)}$ is a probability sample from which we obtain a design-based estimate of $\tau^{(t+1)}$. The model informs or guides survey design and enables it to adapt through time, but all inference is design-based.

9.2 Further Work

There are three sections of further work. The first relates to parts of the thesis where, for completeness, greater investigation of the issues is required. This work is based on the combined sampling strategies developed for monitoring a motile population where individuals behave independently. The second section looks at the application of combined sampling strategies to different types of population. In the third, extensions or changes to the basic combined sampling strategy are suggested.

9.2.1 Greater investigation of the issues in this thesis

A heuristic argument in Chapter 4 stated that unless the distribution of the $y_i^{(t)}$ were negatively skewed, and if $\hat{\mu}_i^{(t)}$ was well-specified we would expect $\sum_{s^{(t)}} y_i^{(t)}$ to be greater under a combined sampling strategy in which $\frac{n_2^{(t)}}{n} > 0$ than under *srswor*. A more formal argument, or a simulation exercise is needed to justify this statement.

Population A was not very heterogeneous and so the change in the c.v. of $\tau^{(2)}$ from using

a combined sampling strategy with $\frac{n_2}{n} = 0$ to a combined sampling strategy in which $\frac{n_2}{n} = 0.74$. Further testing of the strategies on a much more heterogeneous population might show a greater change in the c.v.

In several cases, such as the estimation of covariances in Chapter 5 and the calculation of the model–averaged design variance in Chapter 4 the number of simulations that were carried out was too small so that the some results were not quite as expected because of Monte Carlo error. In future work a greater number of simulations must be carried out.

When estimation of $\delta^{(t',t)}$ is of importance, the combined sampling strategy selected $s_1^{(t)}$ from $s_1^{(t-1)}$. Estimation of $\tau^{(t)}$ and $\delta^{(t',t)}$ require the covariances $cov(\hat{\tau}_1^{(t)}, \hat{\tau}_2^{(t)})$ and $cov(\hat{\tau}^{(t')}, \hat{\tau}^{(t)})$ to be estimated, as shown in Chapter 5. To estimate $cov(\hat{\tau}^{(t')}, \hat{\tau}^{(t)})$, when $t', t \geq 2$, four separate covariance terms are required. Only one of these covariance terms has currently been determined, the other three still need to be specified.

Analytic estimates of covariance only use the data from $s_1^{(t)}$ and $s_1^{(t')}$. In Chapter 8 we proposed a bootstrapping strategy for estimating these covariances using all the data in $s^{(t)}$ and $s^{(t')}$. This strategy is yet to be implemented and tested. In addition a greater understanding of the variability in the empirical estimates of covariance is required.

In Chapter 7 estimators of population density were derived when distance sampling, rather than plot sampling is employed. We stated that we would expect to see an increase in the precision of the density when transects were selected using a combined sampling strategy rather than *srswor*, and that the precision of $\hat{f}(0)$ would increase because $\sum_{s^{(t)}} y_i^{(t)}$ increases. This needs to be tested.

The combined sampling strategies have been developed for motile populations where individuals behave independently. In Chapter 6 we tested the combined sampling strategies on population P in which the superpopulation model from which the $y_i^{(t)}$ were generated remained constant through time. Further simulation work is required to investigate appropriate monitoring strategies when the superpopulation model changes through time, using the model formulations (2–4) described in section 6.3. By selecting $s^{(1)}$ of 81 units

using *sys*, $\hat{\mu}_i^{(1)}$ was a good estimate of $\mu_i^{(2)}$ so that there was little improvement in $\hat{\mu}_i^{(1)}$ after the first survey. An investigation of how various monitoring strategies perform when $\hat{\mu}_i^{(1)}$ is not a good estimate of $\mu_i^{(2)}$ is needed. This could be carried out using population P where n is smaller than 81, or by omitting one of the auxiliary variables from the model construction.

The development of suitable monitoring strategies for different populations requires two distinct investigations. In the first investigation, monitoring strategies can be compared on populations with known behaviour, this is similar to the work in Chapter 6. In addition an exploration of more complex monitoring strategies in which $\frac{n_2^{(t)}}{n}$ changes through time would be useful, to understand how each survey can contribute to the long-term and short-term aims of the monitoring programme can be met so that the objective of each survey may differ. These types of monitoring programme were described in section 6.7.

In the second investigation, we need to mimic what happens in practice; that is the behaviour of the population is not known and so the choice of the appropriate monitoring strategy is difficult. For example the degree of model mis-specification will be unknown and so the choice of the proportion $\frac{n_2^{(t)}}{n}$ can be difficult. This may be especially so when the relationship between auxiliary variables and $\mu_i^{(t)}$ is semi-parametric and non-linear. The development of some simple diagnostics could aid the decision-making process. For example, a potential strategy to determine $\frac{n_2^{(3)}}{n}$ that keeps model mis-specification low, when the model includes non-parametric functions of auxiliary variables would be useful as it may be that $\mu_i^{(t)}$ varies rapidly over particular regions of the X-space. Suppose we set $\frac{n_2^{(2)}}{n} = 0.5$. Then we can construct a model $\zeta^{(2)}$ using the data from $s_1^{(2)}$ only, the data from $s_2^{(2)}$ only or for some proportion p of data from each sub-sample. In each case we can then calculate the sum of the prediction errors, PEV, for all the data in $s^{(2)}$. The proportion, $\frac{n_2^{(3)}}{n}$, is selected based on the proportion p that minimises the PEV. An additional measure would be required to indicate the expected precision of $\hat{\tau}^{(3)}$.

We have stated that model-assisted strategies are an alternative strategy to use in a monitoring programme. In this thesis we have only compared the model-assisted and combined

sampling strategies when estimating $\tau^{(2)}$. A more detailed comparison is required. Although simple regression estimators could be used in the model-assisted estimator a more appropriate strategy would be to develop model-assisted estimators that learn from past surveys and incorporate auxiliary information. These could build on the estimators of Fuller and Rao (2001) which take auxiliary information and learn through time and the estimators of Opsomer *et al.* (2003) which take into account the type of model in which $log(\mu_i^{(t)}) = \sum_{j=1}^{Q} f_j(x_{ij}^{(t)})$.

9.2.2 Application of combined sampling strategies to different populations

We have limited our work to motile populations that do not occur in clusters. In Chapter 2 we described the considerations for constructing a quadrat superpopulation model ζ for different types of populations, in particular for more clustered or sessile populations. In these cases it is important to separate out trend, which is caused by auxiliary variables, and spatial autocovariance, caused by individuals that do not act independently of each other.

To monitor these types of populations additional issues may need to be considered if we wish to use the combined sampling strategy described here. For motile populations that occur in clusters the key problem will be the modelling of $\hat{\mu}_i^{(t)}$ in that until data from several surveys are available it will be difficult to separate the effect of autocorrelation and the effect of unknown auxiliary variables, particularly if smoothing splines of latitude and longitude are included in the model. Hence $\hat{\mu}_i^{(t)}$ may be poorly specified and so we might wish to use a combined sampling strategy in which $\frac{n_2^{(t)}}{n}$ is low.

When the population is sessile it becomes even more difficult to separate the effects of spatial correlation and spatial trend and model mis-specification may be large. Consideration of the extreme case in which $y_i^{(1)} = y_i^{(2)} \forall i \in U$ may indicate possible sources of bias. For this extreme case it would be useful to explore two different strategies. The

combined sampling strategy in which $s^{(1)}$ is used to construct $\zeta^{(1)}$ and predict $\hat{\mu}_i^{(1)}$ and $s^{(2)}$ is then selected using $\pi p\hat{\mu}_i^{(1)}$, $\frac{n_2^{(2)}}{n} = 1$ and the outline of the adaptive strategy given in the discussion. In this second strategy the data from $s_1^{(2)}$ are used to construct the model $\zeta^{(t)}$ and estimate $\hat{\mu}_i^{(t)}$ and $s_2^{(2)}$ is selected using $\pi p\hat{\mu}_i^{(2)}$. The difference between the two strategies is in the data that are used to estimate $\tau^{(2)}$. This exploration may help in the development of a more general framework that links standard adaptive sampling with these combined sampling strategies. We note that an important difference is that the combined sampling strategies are designed to exploit spatial trend, that is variability in $\mu_i^{(t)}$ due to variability in auxiliary variables. On the other hand the adaptive strategies exploit spatial auto-correlation, whether that is spatial trend, or due to auto-covariance from individuals associating with each other.

9.2.3 Extensions to the combined sampling strategy.

The method of $\pi p\hat{\mu}^{(t)}$ sampling is based on the sample selection scheme of Sunter (1977a) in which some units in U have equal inclusion probabilities, equivalent to being selected using *srswor*, even if the $\hat{\mu}_i^{(t)}$ are unequal. This strategy was chosen because it was simple to implement and because units with small $\hat{\mu}_i^{(t)}$, which are likely to be poorly estimated and are not of great interest, are given the same values of π_i. Under the sample selection scheme of Chao (1982), π_i is proportional to $\hat{\mu}_i^{(t)}$ for all $i \in U$. If the combined sampling strategies implemented this sampling selection scheme we would expect the approximation of the unconditional inclusion probabilities to be simpler. In addition it was difficult to apply the bootstrapping strategy for unequal probability sampling, except when the sampling fraction was extremely small, as the proportion selected using *srswor* tended to be very large. This would not be the case under Chao's sample selection scheme.

Two different strategies for incorporating auxiliary information into the design process were considered; that of stratification and of sampling with inclusion probability proportional to size. Ranked set sampling (Patil *et al.*, 1994) is an alternative strategy for incorporating auxiliary information into the design process. This may be a useful strategy

for selecting $s_2^{(t)}$ when the model $\zeta^{(t)}$ is poorly specified as it only requires an ordering of the units in U.

In the one-per-stratum sampling strategies of Breidt (1995) the survey region is partitioned into n strata of contiguous units and one unit from each stratum is selected. Under systematic sampling it is the unit in the same position in each stratum that is selected. More complex strategies also exist in which the position of the unit selected in stratum k is a function of the position of the units selected from adjacent strata, based on Markov chains. In our sampling design the selection of $s_2^{(t)}$ is from the whole survey region so that it is units with the highest values of $\hat{\mu}_i^{(t)}$ that have the greatest probability of being selected. A possible strategy for ensuring better spatial coverage, which also gives units with relatively higher $\hat{\mu}_i^{(t)}$ a greater probability of being selected, might be based on the one-per-stratum strategy described above. Some strata could be part of a systematic design and in other strata a unit is selected with $\pi p \hat{\mu}^{(t)}$.

Finally the construction of the model $\zeta^{(t)}$ has been within a frequentist framework. As we are interested in improving our estimate of $\mu_i^{(t)}$ through time, as more data becomes available, a Bayesian framework might be a more natural strategy for model construction. The posterior distribution of the parameters in $\zeta^{(t)}$ from survey t become the priors for these parameters in survey $t+1$. A Bayesian approach has been taken in the construction of monitoring networks using model-based adaptive sampling strategies such as that of Wikle *et al.* (2001).

9.3 Conclusions

The design-based combined monitoring strategies developed in this thesis provide a flexible framework in which a monitoring programme can adapt its survey design through time. The strategies have been applied to motile species where individuals behave independently of each other. These populations are surveyed using plot sampling, although we have also indicated how the strategies can be applied when the species is surveyed using distance

sampling. Auxiliary information about the survey region is available and information about the spatial distribution of the species will increase over time as survey data becomes available. As information about the spatial distribution of the species increases through time the precision of $\hat{\tau}^{(t)}$ and $\hat{\delta}^{(t',t)}$ increases. The more that can be learnt from past surveys the greater the increase in precision. If the system changes rapidly then less can be learnt from past surveys and so a more cautious monitoring strategy must be employed and so the less the increase in precision. For most populations the number of individuals that are observed in a survey is expected to increase using a combined sampling strategy compared to a sample design in which units are selected using *srswor*. Further work is required to understand and develop monitoring strategies that use these combined sampling strategies under different population models and to enable the selection of the appropriate strategy in practice.

An additional outcome of this thesis was the development of a bootstrapping strategy for the estimation of the variance of $\hat{\tau}^{(t)}$ when samples are selected using an unequal probability sampling scheme that is more flexible than previous bootstrapping strategies as the restriction that $\frac{1}{\pi_i}$ is integer is not required. Further work is needed to test out proposed extensions of the strategy for the estimation of the covariance $cov(\hat{\tau}^{(t')}, \hat{\tau}^{(t)})$ when samples are selected using a combined sampling strategy in which some units are retained from one survey to another.

The general philosophy of the combined monitoring strategies is to use past survey data to construct a model that describes the expected number of individuals in a sampling unit. The predicted number of individuals can be found for all units in the survey region and these predicted values are used in the construction of inclusion probabilities for future survey design. This use of a model to inform future survey design, whilst retaining all inference about the parameters, $\tau^{(t)}$ and $\delta^{(t',t)}$ to a design-based framework is different to existing sampling strategies. This type of philosophy provides a framework for developing different forms of sampling strategy for different types of population and can be seen as a form of adaptive sampling. Within a monitoring context, the flexibility of the sampling

scheme should enable each survey in a monitoring programme to be tailored to meet both the short-term objectives of that survey and the long-term objectives of the monitoring programme.

Bibliography

Atkinson, A. J., B. S. Yang, R. N. Fisher, E. Ervin, T. J. Case, N. Scott, and H. B. Shaffer (2003). MCB Camp Pendleton Arroyo Toad monitoring protocol. Western Ecological Research Center, U.S Geological Survey.

Barnes, R. F. W. (1993). Indirect methods for counting elephants in forests. *Pachyderm 16*, 24–30.

Barnes, R. F. W., K. L. Barnes, M. P. T. Alers, and A. Blom (1991). Man determines the distribution of elephants in the rain forests of northeastern Gabon. *African Journal of Ecology 29*, 54–63.

Barnes, R. F. W., A. Blom, M. P. T. Alers, and K. L. Barnes (1995). An estimate of the numbers of forest elephants in Gabon. *Journal of Tropical Ecology 11*, 27–37.

Bellhouse, D. R. and J. N. K. Rao (1975). Systematic sampling in the presence of a trend. *Biometrika 62*, 694–697.

Berger, Y. G. (2003a). A simple variance estimator for unequal probability sampling without replacement. Methodology Working Paper M03/09, Southampton Statistical Sciences Research Institute.

Berger, Y. G. (2003b). Variance estimation for measures of change in probability sampling. Methodology Working Paper M03/10, Southampton Statistical Sciences Research Institute.

Besag, J. E. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, B 36*, 192–236.

Beyers, R. and J. Hart (2001). Central African pilot project. Technical Report no 4. Site base maps, data bases and MIKE site reports. CITES/MIKE, Nairobi.

Beyers, R., L. Thomas, J. Hart, and S. Buckland (2001). Central African pilot project. Technical Report no. 2. Recomendations for ground based survey methods for elephants in the Central African forest region. CITES/MIKE, Nairobi.

Bickel, P. and D. Freedman (1984). Asymptotic normality and the bootstrap in stratified sampling. *Annals of Statistics 12*, 470–482.

Binder, D. A. and M. A. Hidiroglou (1988). Sampling in time. In *Handbook of Statistics*, Volume 6, pp. 187–211. Elsevier Science Publishers, North Holland.

Blake, S., I. Douglas-Hamilton, and W. B. Karesh (2001). GPS telemetry of forest elephants in Central Africa: results of a preliminary study. *African Journal of Ecology 39*(2), 178–186.

Booth, J. G., T. W. Butler, and P. Hall (1994). Bootstrap methods for finite populations. *Journal of the American Statistical Association 89*, 1282–1289.

Borchers, D. L., S. T. Buckland, and W. Zucchini (2002). *Estimating animal abundance: Closed populations*. Springer-Verlag, London.

Breidt, F. J. (1995). Markov chain designs for one–per–stratum sampling. *Survey Methodology 21*, 63–70.

Breidt, F. J. and J. D. Opsomer (2000). Local polynomial regression estimators in survey sampling. *Annals of Statistics 28*, 1026–1053.

Brewer, K. R. W. and M. Hanif (1982). *Sampling with Unequal Probabilities*, Volume 15 of *Lecture notes in Statistics*. Springer–Verlag, New York.

Brown, J. A. (1999). A comparison of two adaptive sampling designs. *Australian and New Zealand Journal of Statistics 41*, 395–403.

Brus, D. J. and J. J. de Gruijter (1997). Random sampling or geostatistical modelling? Choosing between design–based and model–based sampling strfategies for soil. *Geoderma 80*, 1–44.

Buckland, S., D. Anderson, K. Burnham, J. Laake, D. L. Borchers, and L. Thomas (2001). *Introduction to Distance Sampling: estimating abundance of biological populations.* Oxford University Press, Oxford.

Buckland, S. T. and D. A. Elston (1993). Empirical models for the spatial distribution of wildlife. *Journal of Applied Ecology 30*, 478–495.

Burn, R. W. and F. M. Underwood (2000). Biometric aspects of sampling for the Cameroon Inventory of Prunus africana. Department for International Development.

Burn, R. W. and F. M. Underwood (2001). Monitoring Round Island reptile populations: A preliminary report. Report for Darwin Initiative. Mauritian Wildlife Foundation and National Parks and Conservation Service, Mauritius.

Burn, R. W. and F. M. Underwood (2003). MIKE data analysis strategy. CITES/MIKE, Nairobi.

Cassel, C. M., C. E. Särndal, and J. H. Wretman (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika 63*, 615–620.

Cassel, C. M., C. E. Särndal, and J. H. Wretman (1977). *Foundations of inference in survey sampling.* Probability and mathematical statistics. Wiley, New York.

Chao, C.-T. and S. K. Thompson (2001). Optimal adaptive selection of sampling sites. *Environmetrics 12*, 517–538.

Chao, M. T. (1982). A general purpose unequal probability sampling plan. *Biometrika 69*, 653–656.

Chao, M.-T. and S.-H. Lo (1984). A bootstrap method for finite population. *Sankya Series A 47*, 399–405.

Chin, J. and J. M. Mann (1989). Global surveillance and forecasting of AIDS. *Bulletin of the World Health Organization 67*, 1–7.

Christman, M. C. (2000). A review of quadrat–based sampling of rare, geographically clustered populations. *Journal of Agricultural, Biological, and Environmental Statistics 5*, 168–201.

Christman, M. C. and F. Lan (2001). Inverse adaptive cluster sampling. *Biometrics 57*, 1096–1105.

Cochran, W. G. (1977). *Sampling Techniques* (3 ed.). Wiley, New York.

Cox, D. R. and V. Isham (1980). *Point Processes.* Chapman and Hall, London.

Cressie, N. A. C. (1991). *Statistics for Spatial Data.* Wiley, New York.

Dalenius, T. and J. L. J. Hodges (1959). Minimum variance stratification. *Journal of the American Statistical Association 54*, 88–101.

Davison, A. C. and D. V. Hinkley (1997). *Bootstrap methods and their application.* Cambridge University Press, Cambridge.

Diggle, P. J. (1983). *Statistical Analysis of Spatial Point Patterns.* Academic Press, London.

Dufour, J., J. Gambino, B. Kennedy, J. Lindeyer, and M. P. Singh (1998). Methodology of the Canadian Labour Force Survey. Technical Report 71-526-XPB, Statistics Canada.

Duncan, G. J. and G. Kalton (1987). Issues of design and analysis of surveys. *International Statistical Review 55*, 97–117.

Fay, J. M. (1991). An elephant *Loxodonta africana* survey using dung counts in the forests of the Central African Republic. *Journal of Tropical Ecology 7*, 25–36.

Fay, J. M. and M. Agnagna (1991). A population survy of forest elephants *Loxodonta africana cyclotis* in Northern Congo. *African Journal of Ecology 29*, 177–187.

Francis, R. I. C. C. (1984). An adaptive strategy for stratified random trawl surveys. *New Zealand Journal of Marine and Freshwater Research 18*, 59–71.

Fuller, W. A. and J. N. K. Rao (2001). A regression composite estimator with application to the Canadian Labour Force. *Survey Methodology 27*, 45–51.

Gross, S. (1980). Median estimation in sample surveys. *Proceedings of the Section on Survey Research Methods American Statistical Association*, 181–184.

Haines, D. E. and K. H. Pollock (1998). Estimating the number of active and successful bald eagle nests: an application of the dual frame method. *Environmental and Ecological Statistics 5*, 245–256.

Hansen, M. H., W. G. Madow, and B. J. Tepping (1983). An evaluation of model–dependent and probability–sampling inferences in sample surveys. *Journal of the American Statistical Association 78*, 776–793.

Hansen, M. M. and W. N. Hurwitz (1943). On the theory of sampling from finite poulations. *Annals of Mathematical Statistics 14*, 333–362.

Hartley, H. O. (1962). Multiple frame surveys. *Proceedings of the Social Statistics Section, American Statistical Association*, 203–206.

Hastie, T. J. and R. J. Tibshirani (1990). *Generalised Additive Models.* Chapman and Hall, London.

Hedley, S. L., S. T. Buckland, and D. L. Borchers (1999). Spatial modelling from line transect data. *Journal of Cetacean Resource Management 1*, 255–264.

Hilton-Taylor, C. (2000). *IUCN Red List of Threatened Species.* IUCN.

Holmberg, A. (1998). A bootstrap approach to probability proportional–to–size sampling. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 378–383.

Holmes, D. J. and C. J. Skinner (2000). Variance estimation for labour force survey estimates of level and change. *UK Government Statistical Service Methodology Series No. 21.*

Horvitz, D. and D. J. Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association 47*, 663–685.

Huang, H. C. and N. Cressie (1996). Spatio-temporal prediction of snow water equivalent using the Kalman filter. *Computational Statistics and Data Analysis 22*, 159–175.

Isaki, C. T. and W. A. Fuller (1989). Survey design under the regression superpopulation models. *Journal of the American Statistical Association 77*(377), 89–96.

IUCN/SSC, African, and Asian Elephant Specialist Groups (1999). Proposal for establishing a long term system for monitoring the illegal killing of elephants MIKE. CITES/MIKE, Nairobi.

Jessen, R. J. (1942). Statistical investigation of a farm survey for obtaining farm facts. *Iowa Agricultural Station Research Bulletin 304*, 54–59.

Jolly, G. M. and I. Hampton (1990). A stratified random transect design for acoustic surveys of fish stocks. *Canadian Journal of Fisheries and Aquatic Science 47*, 1282–1291.

Kalton, G. and D. W. Anderson (1986). Sampling rare populations. *Journal of the Royal Statistical Society, Series A 149*, 65–82.

Khaemba, W. M. and A. Stein (2000). Use of GIS for a spatial and temporal analysis of Kenyan wildlife with generalised linear modelling. *International Journal of Geographical Information Science 14*, 833–853.

Kuk, A. Y. C. (1989). Double bootstrap estimation of variance under systematic sampling with probability proportional to size. *Journal of Statistical Computation and Simulation 31*, 73–82.

Laing, S. E., S. T. Buckland, R. W. Burn, D. L. . Lambie, and A. Amphlett (2003). Dung and nest surveys: estimating decay rates. *Journal of Applied Ecology 40*, 1102–1111.

Matérn, B. (1986). *Spatial Variation* (2nd ed.). Springer-Verlag,.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models* (2nd ed.). Chapman and Hall, London.

Merton, D. V., I. A. E. Atkinson, W. Strahmm, C. J. Jones, R. Empson, Y. Mungroo, M. Dulloo, and R. Lewis (1989). A management plan for the restoration of Round Island Mauritius. Jersey Wildlife Preservation Trust and The Ministry of Agriculture, Fisheries and Natural Resources, Mauritius.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society 97*, 558–606.

Neyman, J. and E. L. Scott (1958). Statistical approach to problems of cosmology (with discussion). *Journal of the Royal Statistical Society, B 20*, 1–43.

Opsomer, J. D., F. J. Breidt, G. G. Moisen, and J. Y. Kim (2003). Model-assisted estimation of forest resources with generalized additive models. Preprint series 03-05, Department of Statistics, Iowa State University.

Otis, D., K. P. Burnham, G. C. White, and D. Anderson (1978). Statistical inference from capture data on closed animal populations. *Wildlife Monographs 62*, 1–135.

Overton, W. S. and S. V. Stehman (1995). The Horvitz-Thompson theorem as a unifying perspective for probability sampling: With examples from natural resource sampling. *The American Statistician 49*, 261–268.

Overton, W. S. and S. V. Stehman (1996). Desirable design characteristics for long–term monitoring of ecological variables. *Environmental and Ecological Statistics 3*, 349–361.

Overton, W. S., D. White, and D. L. Stevens Jr. (1990). Design report for EMAP. Environmental Monitoring and Assessment Program. EPA/600/3-91/053, U.S. Environmental Protection Agency.

Patil, G. and C. Rao (1994). *Handbook of Statistics*, Volume 12. Elsevier Science Publishers, North Holland.

Patil, G. P., A. K. Sinha, and C. Taillie (1994). Ranked set sampling. See Patil and Rao (1994), pp. 167–200.

Patterson, H. D. (1950). Sampling on successive occasions with partial replacements of units. *Journal of the Royal Statistical Society, B 12*, 241–255.

Pollard, J. H. and S. T. Buckland (1997). A strategy for adaptive sampling in shipboard line transect surveys. *Rep. Int. Whal. Commn. 47*, 921–931.

Pollard, J. H., D. Palka, and S. T. Buckland (2002). Adaptive line transect sampling. *Biometrics 58*, 862–870.

Rao, J. N. K. and C. F. J. Wu (1984). Bootstrap inference for sample surveys. In *ASA proceedings of the section of survey research methods*, pp. 106–112. American Statistical Association.

Rao, J. N. K., C. F. J. Wu, and K. Yue (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology 18*, 209–217.

Royall, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika 57*, 377–387.

Salehi, M. and G. A. F. Seber (1997). Two-stage adaptive sampling. *Biometrics 53*, 959–970.

Sampson, P. D. and P. Guttorp (1992). Nonparametrics estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association 87*, 108–119.

Särndal, C. (1978). Design–based and model–based inference in survey sampling. *Scandinavian Journal of Statistics 5*, 27–52.

Särndal, C.-E., B. Swennsson, and J. Wretman (1992). *Model Assisted Survey Sampling.* Springer–Verlag, New York.

Schwarz, C. J. and G. A. F. Seber (1999). Estimating animal abundance: review III. *Statistical Science 14*, 427–456.

Seber, G. A. F. (1986). A review of estimating animal abundance. *Biometrics 42*, 267–292.

Seber, G. A. F. (1992). A review of estimating animal abundance II. *International Statistical Review 60*, 129–166.

Sen, A. R. (1973). Theory and applications of sampling on repeated occasions with several auxiliary variables. *Biometrics*.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion). *Journal of the Royal Statistical Society, B 47*, 1–52.

Singh, A. C. (1996). Combining information in survey sampling by modified regression. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 120–129.

Sitter, R. (1992). Comparing three bootstrap methods for survey data. *Canadian Journal of Statistics 20*, 135–154.

Skalski, J. R. (1990). A design for long–term status and trends monitoring. *Journal of Environmental Management 30*, 139–144.

Smith, T. M. F. (1976). The foundations of survey sampling. *Journal of the Royal Statistical Society, Series A 139*, 183–204.

Stehman, S. V. and W. S. Overton (1994). Environmental sampling and monitoring. See Patil and Rao (1994), pp. 263–294.

Strindberg, S. (2001). *Optimized automated survey design in wildlife population assessment.* Ph. D. thesis, School of Mathematics and Statistics, University of St. Andrews.

Sunter, A. (1977a). List sequential sampling with equal or unequal probabilities without replacement. *Applied Statistics 26*, 261–268.

Sunter, A. (1977b). Response burden, sample rotation, and classification renewal in economic surveys. *International Statistical Review 45*, 209–222.

Sverchkov, M. and D. Pfeffermann (2003). Prediction of finite population totals based on the sample distribution. Methodology Working Paper M03/06, Southampton Statistical Sciences Research Institute.

Thomas, L., J. L. Laake, S. Strindberg, F. F. C. Marques, S. T. Buckland, D. L. Borchers, D. R. Anderson, K. P. Burnham, S. L. Hedley, J. H. Pollard, and J. R. B. Bishop (2003). Distance 4.1. release 1. Research Unit for Wildlife Population Assessment, University of St. Andrews, UK. http://www.ruwp.st-and.ac.uk/distance/.

Thompson, S. K. (1990). Adaptive cluster sampling. *Journal of the American Statistical Association 85*, 1050–1059.

Thompson, S. K. (2002). *Sampling* (3rd ed.). Wiley, New York.

Thompson, S. K. and G. A. F. Seber (1996). *Adaptive Sampling.* Wiley, New York.

Urquhart, N. S. and T. M. Kincaid (1999). Designs for detecting trend from repeated surveys of ecological resources. *Journal of the American Statistical Association 4*, 404–414.

Urquhart, N. S., W. S. Overton, and D. S. Birkes (1993). Comparing sampling designs for monitoring ecological status and trends: Impact of temporal patterns. In *Statistics for the Environment*, pp. 7–85. Wiley, New York.

Walsh, P. D. and L. J. T. White (1999). What will it take to monitor forest elephant populations? *Conservation Biology 13*(5), 1194.

Walsh, P. D., L. J. T. White, C. Mbina, D. Idiata, Y. Mihindou, F. Maisels, and M. Thibault (2001). Estimates of forest elephant abundance: projecting the relationship between precision and effort. *Journal of Applied Ecology 38*(1), 217–227.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the gauss-newton method. *Biometrika 61*, 439–447.

Wikle, C. K., R. F. Milliff, D. Nychka, and L. M. Berliner (2001). Spatiotemporal hierarchical Bayesian modelling: tropical ocean surface winds. *Journal of the American Statistical Association 96*, 382–394.

Wikle, C. K. and J. A. Royle (1999). Space-time dynamic design of environmental monitoring networks. *Journal of Agricultural, Biological, and Environmental Statistics 4*, 489–507.

Wing, L. D. and I. O. Buss (1970). Elephants and forests. *Wildlife Monographs 19*, 1–92.

Wolter, K. M. (1985). *Introduction to Variance Estimation.* Springer Series in Statistics. Springer–Verlag, New York.

Wood, S. N. (2001). mgcv:GAMs and Generalized Ridge Regression for R. *R News 1*(2), 20–25.

Wu, C. and R. R. Sitter (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association 96*, 185–193.

Yates, F. (1950). *Sampling methods for censuses and surveys.* Charles Griffin and Co., London.

Yates, F. and P. M. Grundy (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society, B 15*, 235–261.

Zacks, S. (1969). Bayes sequentiual design of fixed size samples from finite populations. *Journal of the American Statistical Association 64*, 1342–1969.

Zidek, J. V., W. M. Sun, and N. D. Le (2000). Designing and integrating composite networks for monitoring multivariate Gaussian pollution fields. *Journal of the Royal Statistical Society, C 49*, 63–79.

Zucchini, W., M. Erdelmeier, and D. Borchers (2002). WiSP. Institut fur Statistik und Okonometrie, Geror-August-Universitat Gottingen, Platz der Gottinger Seiben 5, Gottingen, Germany.

Appendix A

Notation

Much of this notation is based on that of Särndal *et al.* (1992).

Summation

$\sum_A c_i = \sum_{i \in A} c_i$	Sum of c_i for all elements of i in the set A, which we write $i \in A$
$\sum_A \sum_A c_{ij} = \sum_{i \in A} \sum_{j \in A} c_{ij}$	Sum of c_{ij} where $i \in A$ and $j \in A$
$\sum_A \sum_{\substack{A \\ i \neq j}} c_{ij} = \sum_{i \in A} \sum_{j \in A} c_{ij}$	Sum of c_{ij} where $i \in A$ and $j \in A$ but $i \neq j$

The survey region

A	The survey region
N	Number of units in A
i	Label of units in A
$U = \{i : i = 1, \ldots, N\}$	The set of units in A indexed $i = 1, \ldots, N$
t	Index of time - each survey increases the increment by 1
$y_i^{(t)}$	Number of individuals in unit i at time t
$\underline{x}_i^{(t)}$	Set of covariates for unit i at time t
$x_{ij}^{(t)}$	Value of covariate j for unit i at time t

$\tau^{(t)} = \sum_U y_i^{(t)}$	The total number of individuals in the survey region at time t
$\delta^{(t',t)} = \tau^{(t)} - \tau^{(t')}$	Change in the total number of individuals from time t' to t

A sample	*We omit the superscript* $^{(t)}$
n	Number of sampled units
$s = \{i_1, \ldots, i_n\}$	The set of sampled units
$s_c = U - s$	The sample complement
s_k	A subset of s
$s_{k_c} = U - s_k$	The complement of s_k
n_k	Number of units in sub-sample s_k
$p(.)$	A sample design
$p(s)$	Probability of selecting the sample s
$\mathcal{S}$	Set of all possible samples obtained using $p(s)$
$I_{is} = \begin{cases} I_{is} = 0 & i \notin s \\ I_{is} = 1 & i \in s \end{cases}$	Random variable indicating whether unit $i \in s$
$\pi_i = \sum_{s \in \mathcal{S}} I_{is} = \sum_{s \ni i} p(s)$	Probability that unit i is included in sample s
$\pi_{ij} = \sum_{s \ni i,j} p(s)$	Probability that units i and j are included in sample s
$\pi_{i\|s_1^{(t')}}$	Probability that unit $i \in s$ given sample $s_1^{(t')}$ has been selected
E_p	Expectation over all samples that can be selected using $p(s)$
E_{s_k}	Expectation over all possible samples s_k
$\bar{y}_s = \frac{1}{n} \sum_s y_i$	Sample mean
$S_s^2 = \frac{1}{n-1} \sum_s (y_i - \bar{y}_s)^2$	Sample variance

Sample Designs

srswor	simple random sampling without replacement
sys	systematic sampling
strs	stratified random sampling

strs z	*strs* where stratification is based on the variable z_i
πps	sampling with inclusion probability proportional to z_i
$combz(\frac{n_2}{n})$	Combined sampling strategy:
$= combz(\frac{n_2}{n})(U, s_1^{(t)})$	n_1 units sampled from U using *srswor*
	and n_2 units from $s_{1_c}^{(t)}$ using πpz.
$combz(\frac{n_2}{n})(s_1^{(t-1)}, s_{1_c}^{(t-1)})$	Combined sampling strategy through time:
	$n_1 = n - n_2$ units from $s_1^{(t-1)}$ using *srswor*
	and n_2 units from $s_{1_c}^{(t-1)}$ using πpz.

QSM *Quadrat Superpopulation Model*

ζ	The QSM
$Y_i^{(t)}$	Random variable of number of individuals in unit i at time t
$\mu_i = E_\zeta[Y_i^{(t)}]$	Expected number of individuals in unit i at time t
$\sigma_i^{(t)2} = var_\zeta[Y_i^{(t)}]$	Variance of $Y_i^{(t)}$
$\sigma_{i,j}^{(t,t')} = cov_\zeta[Y_i^{(t)}, Y_j^{(t')}]$	Covariance between $Y_i^{(t)}$ and $Y_j^{(t')}$

Distance sampling

D	Density of individuals in A
n	Number of line transects sampled
ω	Half width of line transect
l	Length of line transect
P_a	Proportion of individuals observed in the n line transects
u_k	Distance of k^{th} individual from transect line
y_i	Number of individuals observed in the i^{th} line transect
$g(u)$	Probability that an individual is observed at distance u from the line
$f(u)$	Probability density function of the u_k
$var(\tau)$	Variability in the encounter rate
	(adjusted for probability of selecting transect lines)

Bootstrapping

U_b^*	Pseudo-population
s_b^*	Pseudo-sample
y_i^*	Pseudo-values of y_i
$\tau_b^* = \sum_{U_b^*} y_i^*$	Total for pseudo-population U_b^*
z_i^*	Pseudo-values of z_i
$ppswr(w_i = \frac{1}{\pi_i}, N, s)$	Select N units from s by sampling with replacement with selection probability $\frac{w_i}{\sum_s w_k}$
$ppswr_c(w_i = \frac{1}{\pi_i}, N, s)$	Retain the sample s and select $N - n$ units using $ppswr(w_i - 1, N - n, s)$
$pspv(\underline{y}, s, U^*)$	$y_{i_j}^* = y_k$ if $i_j = k$ for $i_j \in U^*, k \in s$

Appendix B

Population Simulation

In this thesis, three simulated populations have been used to test out the combined sampling strategies. This appendix describes how the populations were generated.

In Chapter 2 a quadrat superpopulation model (QSM) is introduced that gives an empirical description of the pattern of the $y_i^{(t)}$ over the survey region and through time. This does not describe population processes such as birth and death rates, or dispersal. Instead it describes how the $y_i^{(t)}$ change through time. Populations A and B were created by defining a QSM and generating the $y_i^{(t)}$ for $t = 1, 2$ using this QSM.

The $y_i^{(t)}$ are a count of the number of individuals in unit i at time t. For population P a model that generates the location of each individual in the survey region is needed and spatial point processes are used for this purpose. Section B.1 describes some basic spatial point processes. For all three populations the same basic population model is assumed and this is described in section B.2. The details of each population and how it was simulated are given in section B.3.

All computing work in this thesis was carried out in the statistical computing programme

R, available from

`http://www.r-project.org`

A small set of functions, taken from the WiSP library (for Wildlife survey Simulation Package) (Zucchini *et al.*, 2002) were used in the generation of populations described in section B.3. In addition components of the functions for producing plots of sampled units, for example figure 6.1 and for taking a systematic sampling were also used. All other functions for sample selection and estimation were written by the author of this thesis.

B.1 Spatial Point Processes

A spatial point process is a stochastic mechanism for generating a countable set of events in the plane. Cox and Isham (1980) give an introduction to the theory of spatial point processes and Diggle (1983) describes various processes that have been used to model spatial point patterns in biology.

The simplest model is the homogeneous planar Poisson process from which more complex models can develop. There are two key postulates for this process. For a homogeneous survey region, A, of size $|A|$ let the number of individuals in the population at time t be $\tau^{(t)}$. Then

1. The random variable $\mathcal{T}^{(t)}$, the number of individuals in the population at time t has a Poisson distribution with mean $\Lambda^{(t)}$ so that

$$\mathcal{T}^{(t)} \sim Po(\Lambda^{(t)}) \tag{B.1}$$

2. Given $\mathcal{T}^{(t)} = \tau^{(t)}$ the $\tau^{(t)}$ individuals form an independent random sample from the uniform distribution on A.

Hence individuals behave independently of each other and the density at the point $\mathcal{L}$ is $\lambda^{(t)} = \Lambda^{(t)}/|A| \forall \mathcal{L} \in A$

This is an extremely simple model There are two ways in which complexity is introduced into the spatial point pattern. Firstly, individuals are not independent of each other. Secondly the environment is not homogeneous so that the species distribution varies over the survey region.

Assuming that the environment is homogeneous over A, the Poisson cluster process introduced by Neyman and Scott (1958) provides a framework for modelling aggregated spatial patterns. In its simplest form,

1. the number of cluster parents form a Poisson process with intensity $\mathcal{T}^{(t)}/\Gamma^{(t)}$.
2. Each cluster parent produces a random number of offspring $\gamma^{(t)}$ realised independently and identically for each cluster parent according to some probability distribution $\Gamma_{\gamma}^{(t)}, \gamma = 0, 1, 2, \ldots$.
3. The positions of the offspring relative to the cluster parent are independently and identically distribution according to a bivariate probability distribution function $h(.)$

Unless explicitly stated the convention is that the only individuals in the population are the offspring. Populations exhibiting these characteristics are not used in this thesis.

Suppose now, that the environment is heterogeneous over A. The heterogeneity is summarised by a suitability index for a particular species. At a particular point, $\mathcal{L}$, in A, the suitability $\lambda_{\mathcal{L}}^{(t)}$ is a function of Q^* variables such as habitat, terrain and climate that influence the distribution of the species of interest. In the same survey region a different suitability index would be derived for different species. In general suitability will be a relatively smooth function over A, although changes in habitat, due to natural or man–made boundaries, may lead to discontinuities.

As individuals are assumed to behave independently of each other, the $\tau^{(t)}$ individuals form an independent random sample from the distribution on A with a probability density function proportional to $\lambda_{\mathcal{L}}^{(t)}$. This is the inhomogeneous Poisson process.

Returning to the homogeneous planar Poisson process, if $\lambda_{\mathcal{L}}^{(t)}$ represents the population density at a single point, the expected number within quadrat i, $\mu_i^{(t)}$ with an area A_i is

$$\mu_i^{(t)} = \int_{A_i} \lambda^{(t)} d\mathcal{L} = A_i \lambda^{(t)} = \frac{A_i}{A} \Lambda^{(t)} \tag{B.2}$$

Assuming $A_i = A_0$ for all i, then

$$\mu_i^{(t)} = \mu_{(0)}^{(t)} \tag{B.3}$$

and a QSM ζ is of the form

$$Y_i^{(t)} \sim Po(\mu^{(t)}) \tag{B.4}$$

Under the inhomogeneous Poisson process if we assume that $\lambda_{\mathcal{L}} = \lambda_i$ for all $\mathcal{L}$ in the unit $A_i = A_0$ then

$$\mu_i^{(t)} = \int_{A_i} \lambda^{(t)} d\mathcal{L} = \lambda_i^{(t)2} \tag{B.5}$$

B.2 Description of the population

In this thesis, we consider populations of highly motile animals within a heterogeneous environment. By highly motile we mean that all individuals can easily cover A in a period much shorter than the interval between surveys. Although we would generally not recommend plot sampling for surveying highly motile animals we use it to develop our sampling designs.

We assume that individuals are exchangeable and therefore model the dynamics of the whole population, rather than modelling individual life histories. The effect of environment would anyway be small as individuals are highly motile, and so all are affected in the same way. Although age, sex and other individual characteristics may influence survival this is ignored for simplicity. The number of individuals in the population at time t, $\tau^{(t)}$ is assumed to be a random variable generated from a Poisson distribution with a mean $\Lambda^{(t)}$

that changes by a constant proportion r from one period to another. This proportion r is an intrinsic feature of the population rather than the effect of changing environment.

In fact although the environment is heterogeneous over A we assume that it remains constant through time, or at least is the same at each survey. The environment is summarised with respect to the species of interest using a suitability index. At a particular point, $\mathcal{L}$, in A, suitability is a function of variables such as habitat, terrain and climate that influence the distribution of the species of interest. In the same environment a different suitability index would be derived for different species. In general suitability will be a relatively smooth function over A although changes in habitat, due to natural or man–made boundaries may lead to discontinuities. In this thesis we assume the relationship between log(suitability) and habitat is linear.

Given suitability it is assumed that the location of an individual is conditionally independent of the locations of all other individuals in A. This is unrealistic, as most species move around in groups or have territories. However it serves as a baseline scenario, and we could alternatively think of each "individual" as representing a group of animals. The aim is then to estimate the number of groups in A.

As individuals are highly motile it is also assumed that the location of an individual at time t is independent of the location of an individual at time $t-1$. From a modelling perspective only the number of individuals at time t, $\tau^{(t)}$, is required and their locations are generated using an inhomogeneous Poisson process with intensity proportional to suitability.

A formal model Ξ of the population and the environment is described below

1. Population characteristics through time
   - $\mathcal{T}^{(t)}$ has a Poisson distribution with mean $\Lambda^{(t)}$ for the area A
   - The mean, $\Lambda^{(t)}$ changes by a proportion r, the intrinsic growth rate so that $\Lambda^{(t)} = r\Lambda^{(t-1)}$
2. Suitability
   - At location $\mathcal{L}$ suitability at time t is $\lambda_{\mathcal{L}}^{(t)}$ so that $\int_A \lambda_{\mathcal{L}}^{(t)} d\mathcal{L} = \Lambda^{(t)}$
   - The change in suitability through time is $\lambda_{\mathcal{L}}^{(t)} = r\lambda_{\mathcal{L}}^{(t-1)} = r^t\lambda_{\mathcal{L}}^{(0)}$
   - Suitability can be defined as a function of Q^* covariates $x_1, \ldots, x_{Q^*}$ so that $\lambda_{\mathcal{L}}^{(0)} = r^{(0)}\lambda^{(0)}(x_{\mathcal{L}1}, \ldots, x_{\mathcal{L}Q^*})$
3. Relationship between suitability and spatial distribution of individuals
   - Given $\mathcal{T}^{(t)} = \tau^{(t)}$ the $\tau^{(t)}$ individuals form an independent random sample from the distribution on A with pdf proportional to $\lambda_{\mathcal{L}}^{(t)}$

The QSM is of the form

$$Y_i^{(t)} \sim Po(r^t\mu_i^{(0)}) \tag{B.6}$$

$$\Rightarrow E[Y_i^{(t)}] = r^t\mu_i^{(0)} \qquad var[Y_i^{(t)}] = r^t\mu_i^{(0)} \qquad cov[Y_i^{(t)}, Y_j^{(t'0)}] = 0 \tag{B.7}$$

$$log(\mu_i^{(t)}) = \sum_{j=0}^{Q} \beta_j x_{ij} \tag{B.8}$$

B.3 Population simulation

B.3.1 Simulation of populations A and B

Populations A and B are used in Chapters 4 and 5 to test out the basic combined sampling strategies, where the focus is how $\hat{\mu}_i^{(1)}$ can be used in the design of survey 2. These populations are also used in Chapter 8 where the focus is variance and covariance estimation for the combined sampling strategy. Although units in a survey region will be spatially related, this is not part of the survey design or estimation problem in these three

chapters. The essential components that are required from the populations used for these investigations are a set of N units where for each unit Q auxiliary variables are available. Then $\mu_i^{(t)}$ is a function of these Q auxiliary variables as described in equation B.8 so that $y_i^{(t)}$ is generated from a QSM of the form B.6. For populations A and B, $N = 1000$ and $Q = 3$ and we require $y_i^{(t)}$ for $t = 1, 2$.

The auxiliary variables were generated from different distributions:

- x_{i1} from a uniform distribution, Unif$[0.5, 10]$;
- x_{i2} from a normal distribution, $N(\mu = 3, \sigma = 1)$;
- x_{i3} from a normal distribution, $N(\mu = 20, \sigma = 5)$.

The auxiliary variables are illustrated in figure B.1 and the correlation between auxiliary variables in figure B.2.

Suitability, $\mu_i^{(1)}$ the expected number of individuals in unit i at time 1, is calculated as

$$log(\mu_i^{(1)}) = 0.1 + 0.1x_{i1} + 0.25x_{i2} - 0.05x_{i3} \tag{B.9}$$

For population A we use a population model in which $\Lambda^{(2)} = 1.5\Lambda^{(1)}$ so that

$$log(\mu_i^{(1)}) = 0.6 + 0.1x_{i1} + 0.25x_{i2} - 0.05x_{i3} \tag{B.10}$$

The correlation between $y_i^{(1)}$ and $y_i^{(2)}$ is 0.28. As $\delta^{(1,2)}$ is also to be estimated and we expect the precision of $\delta^{(1,2)}$ to increase when $y_i^{(1)}$ and $y_i^{(2)}$ are highly correlated we generate population B where $y_i^{(1)}$ is as described for population A and $y_i^{(2)}$ is generated so that

$$log(\mu_i^{(2)}) = -0.43 + log(\mu_i^{(1)}) + log(y_i^{(1)}) \tag{B.11}$$

The correlation between $y_i^{(1)}$ and $y_i^{(2)}$ for population B is 0.71. We note that this does not follow our usual assumption that $cov(Y_i^{(t)}, Y_i^{(t')}) = 0$ Figure B.2 shows the relationship between the auxiliary variables and $\mu_i^{(1)}$ and figure B.3 shows the distribution of the $y_i^{(1)}$ and $y_i^{(2)}$, the relationship between $y_i^{(t)}$ for $t = 1, 2$ and $\mu_i^{(1)}$ and between the $y_i^{(1)}$ and $y_i^{(2)}$ for populations A and B.

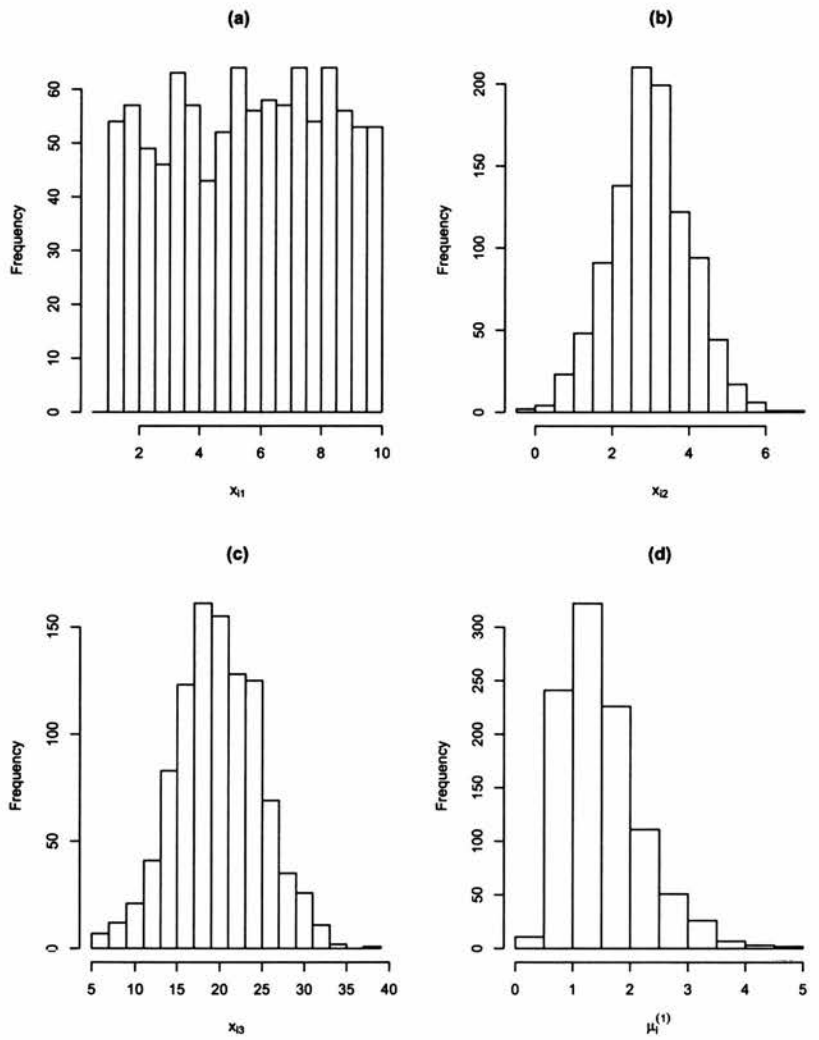


Figure B.1: Auxiliary variables (a) x_{i1}, (b) x_{i2} (c) x_{i3} and (d) $\mu_i^{(1)}$ for populations A and B.

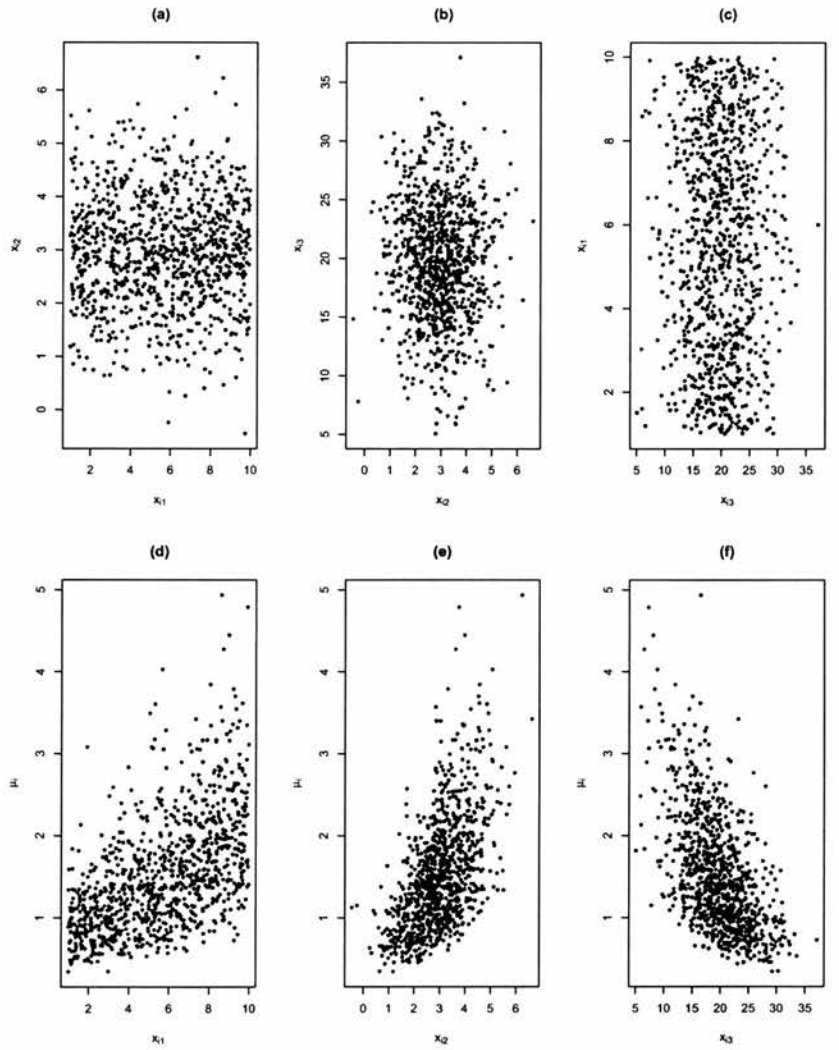


Figure B.2: Relationship between auxiliary variables (a) x_{i1} and x_{i2}, (b) x_{i2} and x_{i3} (c) x_{i3} and x_{i1} and between $\mu_i^{(1)}$ and (d) x_{i1} (e) x_{i2} (f) x_{i3} for populations A and B.

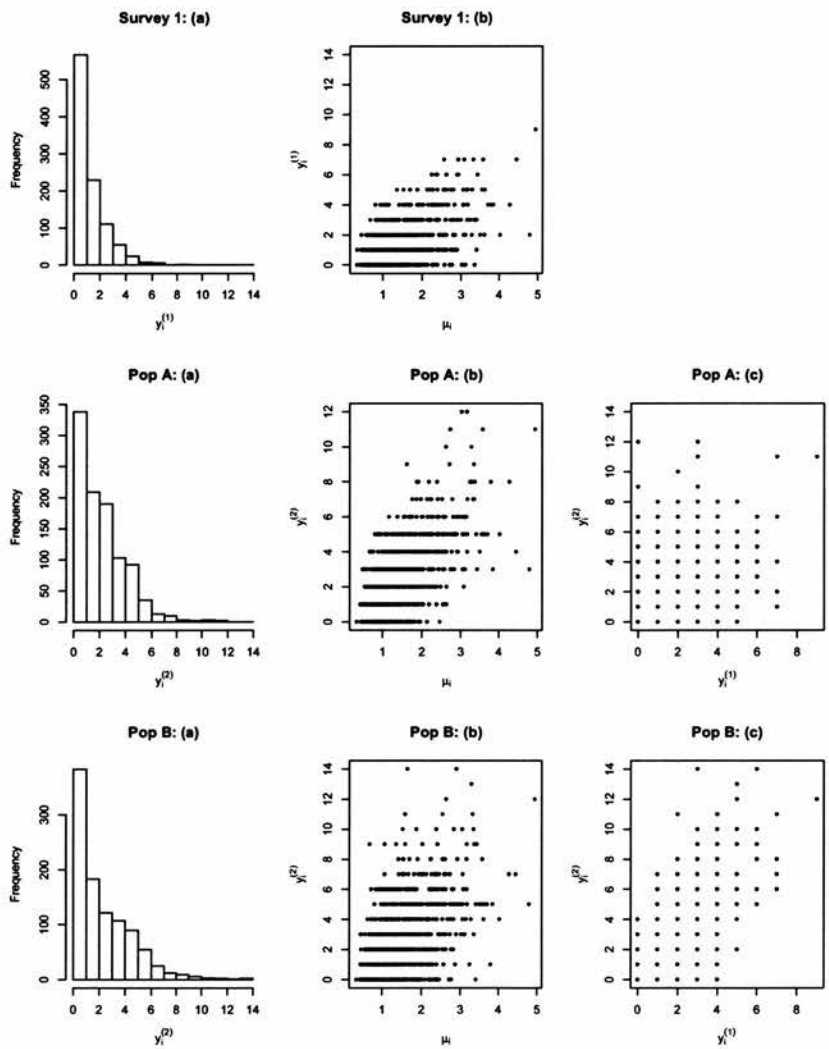


Figure B.3: (a) Distribution of $y_i^{(t)}$ (b) Relationship between $\mu_i^{(1)}$ and $y_i^{(t)}$ (c) Relationship between $y_i^{(1)}$ and $y_i^{(2)}$, for populations A and B and for $t = 1, 2$

B.3.2 Simulation of population P

This population is used in Chapter 6 to explore how the combined sampling strategies can be used within a monitoring programme. Here it is important to represent a generic population and so the spatial relationship between the units in the survey region is important. In this case the distribution of the individuals is generated.

The survey region is 36 x 36 in size which is summarised into $N = 1296$ units each of size one. To generate individuals, auxiliary data must be defined as a surface over the survey region before being summarised into quadrat information. Data on $Q = 4$ auxiliary variables are created. The basic functions for setting up the survey region, generating basic density functions, the λ_L, the spatial point patterns and the summarising of the individual locations into quadrat counts use functions from the WiSP package.

The auxiliary variable $x_{\mathcal{L}_i j}$ is the sum of K_j variables $z_{\mathcal{L}_i k}$ where $\underline{z}_k$ has a maximum value with distribution $N(h_k, \sigma_{h_k})$ at a randomly located point in the survey region $\mathcal{L}_{0k} = (x_{0k}, y_{0k})$. The value of $\underline{z}_k$ at a point $\mathcal{L}_i = (x_i, y_i)$ is a function $d(\sqrt{((x_{0k}-x_i)^2+(y_{0k}-y_i)^2)})$ where d is the Normal density function with mean 0 and variance 0.0625. For each of the four auxiliary variables the number of hotspots and their parameters are:

- x_{i1} $K_1 = 50$ hotspots $N(h_k, \sigma_{h_k}) = N(0, 8)$;
- x_{i2a} $K_2 = 100$ hotspots $N(h_k, \sigma_{h_k}) = N(0, 2)$;
- x_{i2b} $K_2 = 25$ hotspots $N(h_k, \sigma_{h_k}) = N(0, 4)$;
- x_{i3} $K_3 = 100$ hotspots $N(h_k, \sigma_{h_k}) = N(0, 2)$;
- x_{i4} $K_4 = 5$ hotspots $N(h_k, \sigma_{h_k}) = N(0, 4)$.

Auxiliary variables x_{i2a} and x_{i2b} were added together to give auxiliary variable x_{i2}. Auxiliary variable x_{i4} is converted into a categorical variable by taking all values of $x_{i4} > 0.85$ as category 1 and units $x_{i4} \leq 0.85$ as category 2.

Suitability $\lambda_{\mathcal{L}}$ uses auxiliary variables x_{i1}, x_{i2} and x_{i4} so that

$$log(\lambda_{\mathcal{L}}) = 1.25x_{i1} + 0.5x_{i2} + 1.25x_{i4} \tag{B.12}$$

Auxiliary variable x_{i3} is not included in the definition of $\lambda_{\mathcal{L}}$ so as to mimic the common scenario in which not all the auxiliary variables that are available to the statistician are relevant for model construction. The population of individuals at time t are generated using an inhomogeneous Poisson process with intensity $\lambda_{\mathcal{L}}$.

This model is retained for all surveys $t = 1, \ldots, 10$. Data is summarised at quadrat level so that x_{ij} is the value of each auxiliary variable at the midpoint of each quadrat, and $y_i^{(t)}$ is the number of individuals within each quadrat in survey t. The spatial point pattern for $t = 1$ is shown in figure B.4

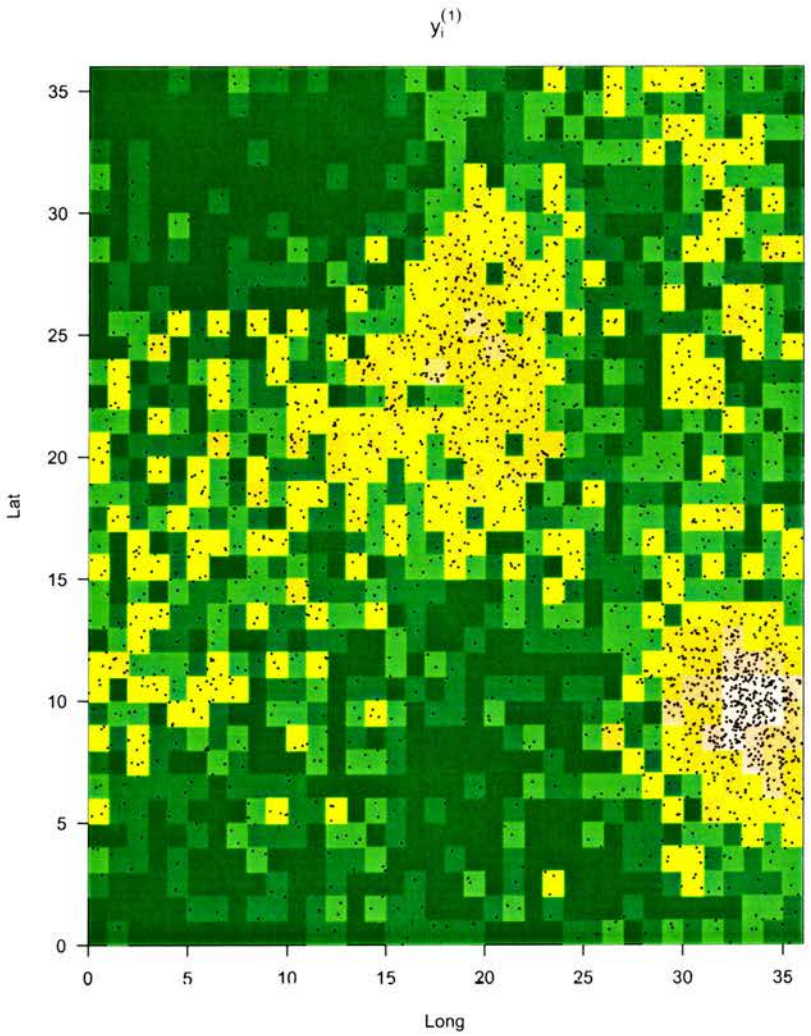


Figure B.4: The spatial point pattern for population P for $t = 1$ and summarised counts $y_i^{(1)}$