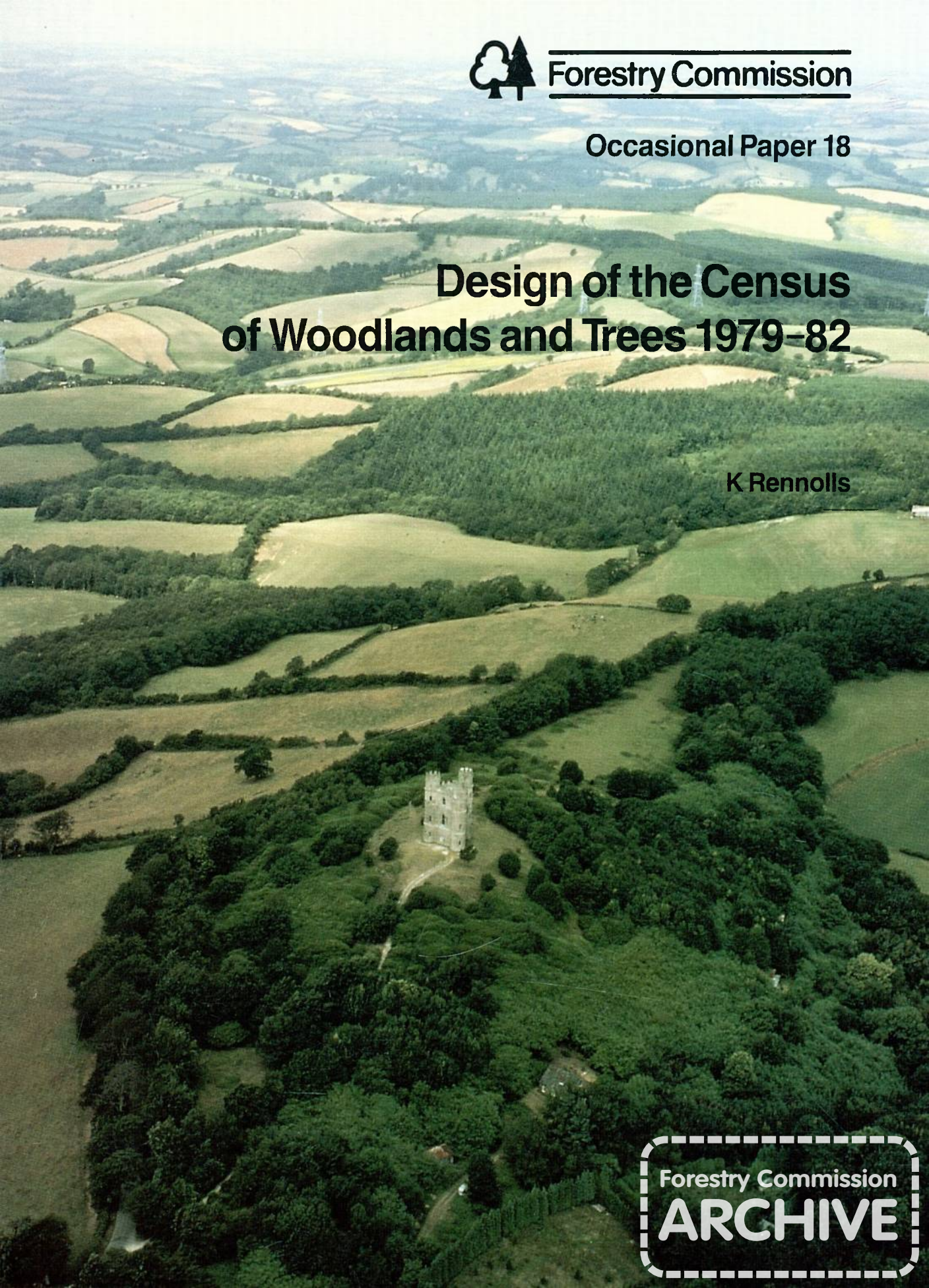


Forestry Commission

Occasional Paper 18

Design of the Census of Woodlands and Trees 1979–82

K Rennolls

Forestry Commission ARCHIVE

Forestry Commission Occasional Paper 18

Design of the Census of Woodlands and Trees 1979–82

K. Rennolls
former Statistician,
Forestry Commission
now Principal Lecturer in Statistics,
Thames Polytechnic

FORESTRY COMMISSION, EDINBURGH

© *Crown copyright 1989*
First published 1989

ISBN 0 85538 220 1

ODC 524.6: (410): U519.24

Keywords: Survey design, Forestry

Enquiries relating to this publication
should be addressed to the
Technical Publications Officer,
Forestry Commission,
Forest Research Station,
Alice Holt Lodge, Wrecclesham,
Farnham, Surrey GU10 4LH

FRONT COVER: Aerial photograph of woodlands and trees in the landscape around Lawrence Castle, Devon. *(31279)*

Contents

CHAPTER 1

Main features of the 1979–82 survey design

Introduction

The Forestry Commission carried out censuses of Woodlands and Trees in 1924, 1938, 1947–49, 1965–67 and most recently between 1979 and 1982. Only two of these data collection and collation exercises came close to being a census in the sense of being a complete enumeration. The first was in 1924 when questionnaires were submitted by owners owning blocks of 2 hectares or more and the second was the 1947 census in which all woodland blocks of 5 acres and over shown on 6 inch to 1 mile Ordnance Survey maps were visited and assessed. The remainder of the censuses have in fact taken the form of sample surveys, the sampling aspect becoming most pronounced in the last two surveys. The purpose of this report is to describe the details of the sample survey aspects of the 1979–82 Census of Woodlands and Trees.

This chapter reviews general survey methodology, most elements of which occur in any survey in any field. Figure 1 illustrates the logical structure of conducting a sample survey and this is used as a basis for the discussion of the methodological aspects of conducting the woodland and tree survey. The general description given here has been applied to both the Woodland and Non-Woodland survey, preliminary details of which are given in the next chapter, and mathematical details in Chapters 5 and 6 respectively. This report does not discuss aspects of technique and mensuration which are covered elsewhere, Locke (1987).

Stage 1: Survey Design

Survey design is the first stage of any survey and is logically prior to the operational planning of the survey and its implementation. We may subdivide survey design into two parts. The first is concerned with the objectives and structure of the survey. The structure of a survey is determined by the choice of sampling units and the way that they are grouped together, possibly in strata. The whole set of possible sample units is termed the 'sampling frame', and it is important that the sampling frame covers the population to be surveyed without any of the sample units overlapping. The second part of survey design is concerned with the determination of the sample size required to meet objectives and the optimal distribution of this sample over the sampling frame of the chosen design.

Part 1: Survey Objectives and Overall Structure

The initial considerations of the first part (Part 1 in Figure 1) of the survey design must be dominated by the requirements of the ultimate user of the survey results. It is necessary to know for what purpose the results are needed. In a national survey of a multipurpose nature the number of potential questions is very high and their range very wide. Also the precision required, and hence the necessary sample sizes, will vary from one question to another. Hence a survey of a given size may answer some questions satisfactorily but not others. Resources are invariably limited and a complete enumeration is not possible; it is therefore necessary to assign priorities to the questions which may be asked of the survey results.

For those questions of highest priority we must decide how accurate the survey results need to be in order to provide satisfactory answers and specified precisions must be set for the most important results to be estimated. The survey must be designed so that it does not obtain un-needed precision by collecting too much information. It is thus necessary to state target precisions for estimates, usually in terms of standard errors as a percentage of the estimates. These target precisions represent, for statistical purposes, the objectives of the survey, and are indicated by 'A' in Figure 1. Those adopted for the present survey have been given in detail by Locke (1987) and are briefly reviewed in the next chapter

It is worth pointing out briefly the interpretation to be given to standard errors and some of the possible consequences of stating objectives in terms of percentage standard error. For example, suppose we are estimating a population value of about 1000 (in suitable units) and our objective is to obtain a standard error of 10%, that is ± 100 (in suitable units). This means, approximately, that we wish to be 67% and 95% sure that the population value lies in the ranges (1000 ± 100) and (1000 ± 200) respectively. However, since the survey is based on random samples we cannot be sure that our target

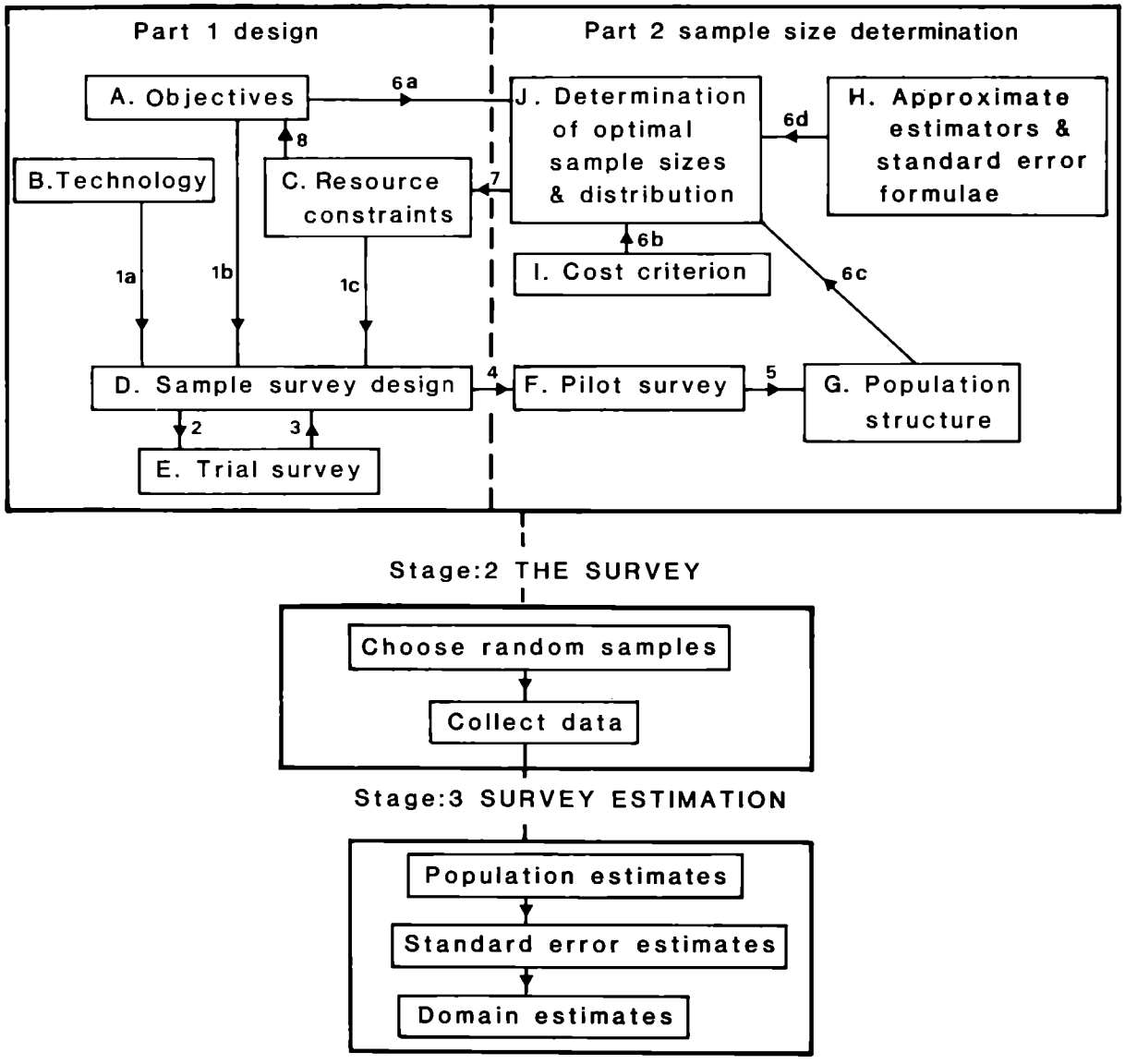


Figure 1. Survey methodology.

precision will be achieved. It is possible that the result of the survey will be (900±120) (13.4% s.e.) or possibly (1100±80) (7.3% s.e.). The target precision represents the precision we hope to get and would, under suitable conditions, be approximately equal to the average precision obtained if the survey were to be done repeatedly. It follows that if we wish to be sure, say 95% sure, that the precision actually obtained will be 10% or less then we must set our target precision well below 10%.

These considerations were taken into account in the definition of the objectives of this survey and it was to be expected that some of the precisions obtained, and given later in this report, and in Locke (1987), would be greater than the corresponding target precisions.

The next two items of Figure 1 are termed 'B. Technology' and 'C. Resource Constraints'. Consideration of the part that 'technology' should play in the survey should include, for example, whether a good map of the population exists; whether satellite or aerial photography is available and other aspects which enhance the technical efficiency of the survey. Technological considerations will affect the entire structure which the survey will have. On the other hand resource constraints will affect the overall size and shape of the survey but will have little direct effect on the basic structure of the survey. Examples of such resource constraints are limitations in time, staff and money, and factors such as the necessity of having to obtain approval from land owners before a ground assessment can be made. These aspects of the survey design, together with the survey objectives, determine the range of sample survey designs that are practically feasible.

The sample survey design includes the specification of convenient sample units, the definition of suitable sampling frames and the manner in which sample units are selected from the sampling frames. There are no hard and fast rules in doing this. Previous experience provides some guidance and before the final adoption of a design for this survey, studies were made in order to determine the best sizes of sample units and to assess the usefulness of certain new techniques. Also within Part 1 (E in Figure 1) of the survey design a preliminary trial survey was done in W. Sussex to evaluate some tentative sample survey designs. The design used was also subsequently modified as a result of experience gained in the first few counties which were fully surveyed.

Part 2: Sample Size Determination

Once the sample survey design is decided the size of the sample to be taken has to be determined. Since the sample survey is more complex than 'simple random', a set of sample sizes was required and in this context 'sample size' means this whole set of sample sizes.

From the statistical viewpoint sample size determination is the most difficult aspect of survey design. It is necessary to take into account simultaneously the chosen sample survey design, the approximate population structure to be surveyed, the objectives of the survey, the limiting resource constraints, as well as to have mathematical expressions for the population estimators and an expression for the 'cost' of the survey. This combination is achieved in practice by following the path, in Figure 1, from D through links 4, 5 and 6 ((6a), (6b), (6c) and (6d) simultaneously) to obtain a required sample size which will approximately satisfy the design objectives. The required sample size (obtained at J) has then to be compared with the resource constraints (of C). If the required size of the sample is too large to satisfy these constraints, then a decision has to be made between allocating more resources to the survey or modifying the objectives and aiming for less precise estimates. If the precision targets are modified the sequence 6. 7. 8 has to be iterated until a satisfactory compromise between targets and resources is obtained.

The first item in Part 2 of Stage 1, (Sample Size Determination) is a pilot survey, (F in Figure 1). This is necessary since it is not possible to determine the sample size required to reach given precision targets without already knowing, approximately, the average sample unit values, and their variabilities. The pilot survey gives us a preliminary idea of population structure. (G in Figure 1). This process might seem rather circular, but the dilemma may be resolved by a sequential approach in which the pilot survey is done in such a way that the pilot survey data may be combined with the final survey data.

The next task within the sample size determination process requires that mathematical expressions for the eventual estimators be obtained together with formulae for their standard errors, H in Figure 1. These expressions are in terms of the unknown sample sizes to be taken in different parts of the survey design, and for a complex design, with complex estimators, as in this survey, the formulae can be very complicated. It was therefore sometimes necessary to make approximations to the sample estimators so that they could be handled conveniently.

By making use of the approximate population structure (G) in conjunction with the proposed estimation formulae, (H), it is possible to determine the precision that will be obtained from a given sample size. By comparing this with the target precisions it is possible to decide if more samples are required. The sample size may then be varied until we have a sample size which will achieve the precision targets. As mentioned above, the problem is complicated by the fact that a set of sample sizes are required. For example, a sample size will have to be determined for each of the strata of the survey, and if the survey is multistage then sample sizes at each stage will have to be determined. The balance between these sample numbers has to be chosen in order to obtain the target precisions at minimum cost. Hence a cost criterion (I in Figure 1) is required which may be used in order to determine the least costly of the possible distributions of sample sizes. For the very simplest designs it is possible to achieve this task of optimal sample size determination allocation by mathematical means. However, the complexity of this survey, and of its estimators, meant that this task had to be done using computer optimisation techniques.

Stage 2: The Survey

Once the Survey design has been completed, the second stage begins with the random selection of particular sample units. This may be done using a pseudorandom number generator on a computer. The main work of the survey, the data collection, may then take place; details for this survey are given in Locke (1987).

Stage 3: Survey Estimation

The final stage of the survey involves the estimation of population values together with standard errors. The estimation methods in this survey were, roughly speaking, either by 'expansion' or 'regression'. Technical details of these estimation methods may be found in Chapters 5 and 6 of this report.

CHAPTER 2

The population to be sampled

The population of interest can be naturally divided into two sub-populations distinguished as 'Woodland' and 'Non-woodland'. It was therefore necessary to consider whether to use a single survey to cover both categories, or whether greater efficiency might be obtained by having separate sub-surveys for each.

The objectives for 'Woodland' and 'Non-woodland' sub-populations differ. For Woodlands, defined as those groups of trees with a ground cover of not less than 0.25 ha, the precision objective was defined in terms of those woodlands (termed 'Other' Woodlands) which were neither owned by the Forestry Commission, nor managed under the Dedication or Approved Woodland Schemes. For the two last named categories the Forestry Commission had essentially complete data available from existing records and therefore these needed no sampling. The guiding precision targets for 'Other' woodland were for a 5% s.e. on total area and 15% on the total area of the main forest type.

On the other hand the non-woodland target precisions for non-woodland features would certainly have required different sampling intensities within a single survey. Also, the amount of prior information available on Woodlands, from Ordnance Survey (OS) maps was considerable more than that on Non-woodlands. In order to make full use of this information it was felt to be desirable to make at least a formal distinction between the Woodland and Non-woodland surveys. The price paid for the increased efficiency of adopting a two survey structure is the increased travel cost that results. To minimise this cost requires careful logistical planning to ensure that neighbouring sample units of the two surveys are assessed at the same visit.

Though there are essentially two sample surveys targeted on the woodland and non-woodland populations it will be seen later that the OS map does not provide a sampling frame able to completely cover the woodland population. So, even though we refer to the woodland and non-woodland surveys as distinct, they are both necessary contributors to the final estimation of the woodland population.

CHAPTER 3

The survey design for woodlands

Choice of Sampling Strategy

In considering how to define a sample unit and choose a sampling design, the possibility of following precisely the same procedures as were used in the 1965 census had certain advantages:

1. the simplicity of a simple random design would mean that sample survey design and final estimation were straightforward;
2. there was less chance of misapplication in the field of a simple random design than a more complex design;
3. by choosing a sample which overlapped with the 1965 sample it would be possible to obtain relatively precise estimates of the change that had taken place; and
4. some assurance would seemingly be provided that the surveys were more directly comparable than if the survey design were changed.

However, there were corresponding disadvantages, which led, eventually, to the choice of a more complex design than simple random. These were:

1. The simple random sample would be very inefficient. It does not allow the full use of information, known prior to the survey, when selecting samples. In principle the soil maps and OS maps both provide a good basis for stratification and hence improvement in the precision of estimates. Such auxiliary information could be used at the estimation stage of a simple random sample but this would in itself involve a considerable deviation from the methods used in 1965. In view of the requirements for more precise estimates than had been previously obtained and the fact that these estimates had to be based upon a data set collected by a smaller work force, a change in survey design was inevitable. The greater processing power of the Research Division's computer meant that the more complicated calculations required for a complex design could be quickly and accurately performed.
2. Though an overlapping sample would have provided relatively accurate estimates of change, the estimate of the current woodland stock would have been relatively imprecise, as indicated above. It was decided to aim for the most accurate estimate of current woodland stock and to obtain the estimates of change between two surveys, by differencing their separate results.
3. It is of course as valid to compare results from two independent unbiased surveys which have different sampling designs as when the survey designs are identical. The main points at which comparability needs to be ensured result from the change in the minimum size of a woodland (from 0.4 to 0.25 ha) and the change of the regions over which the survey was to be conducted (from marketing regions to counties). Neither of these aspects affects the statistical comparability of surveys having different designs.

The precision objectives mentioned in Chapter 2 (5% and 15% standard errors for total area and main forest type) in general required different sampling intensities, usually with the higher intensity being required to attain the former precision target. After preliminary investigation it became fairly clear that most of the data required for the estimation of the total area would be obtained most economically from OS maps and recent aerial photographs, while data necessary for estimation of types, species, etc., would largely have to be based upon ground survey data. The woodland survey was therefore designed in two phases which are illustrated in Figure 2. The first phase, involving the use of OS maps and aerial survey, had as its specific objective the estimation of the total woodland area (to within 5% standard error) and the second phase (a ground survey) was designed to obtain a 15% standard error on estimates of the area of the most widely represented forest type.

First Phase Sample

Sample units and sampling frame.

The use of a fixed area sampling unit was not well suited to making efficient use of the information on woods which was available from recent OS maps. Randomly distributed fixed area sample units would not focus sufficiently on those regions of the county in which it was known that woods were clustered.

A natural, and easy to use, sampling unit is a woodland block, as indicated on the 1:50 000 OS map, and this was adopted. The 'sampling frame' for the woodland survey is based on the set of woods indicated on the OS map. An immediate objection to the sampling frame which consisted of such blocks was that it did not truly cover the whole of the woodland population. There were some woodlands which did exist but had not been indicated on the map. Therefore the map only represented a partial sampling frame which is termed here 'the Woodland partial sampling frame'. Those woods not covered by this frame (termed 'extra' woods) were covered by the sampling frame for the Non-woodland survey.

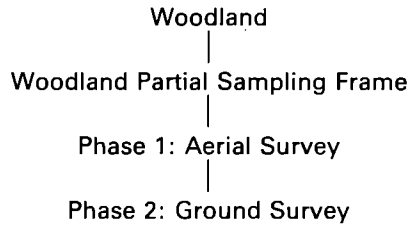


Figure 2. The levels of information in the Woodland survey.

If a block had a map area of less than 0.25 ha then it was very likely to have an actual area of less than 0.25 ha. Such small map blocks were excluded from the Woodland partial sampling frame. Woodlands of an actual area greater than 0.25 ha but with a map area of less than 0.25 ha were also included as 'extra' woods, and were sampled using the non-woodland sampling frame to be described in the next chapter. The woodland partial sampling frame was also expected to contain spurious elements, since some elements shown as woodland on the map were in fact not so, according to the definition of woodland adopted. This aspect was satisfactorily taken into account at the final estimation stage.

Sample design

A number of possible sampling strategies were considered for the sampling to estimate the woodland total area. Each of these makes some use of the map areas available from the 1:50 000 OS map. The following discussion details some of the reasons they were not chosen, and the way in which these influenced the final choice of design. Cochran (1963) gives full details of the methods mentioned.

1. Selection of units for actual-area measurement with probability proportional to the unit's map area. Such a design, though efficient, presents theoretical problems in optimal sample size determination because the requisite theory is not available. The theoretical difficulties of estimation and sample size determination for the subsampled second phase would be even more severe. Relating first and second phase woodland survey estimates would also have been very awkward technically.
2. Double sampling, in which a large simple random sample of units is selected from the partial sampling frame and the map areas of these units related to the actual areas of those units in a smaller ground measured subsample. Even though the initial sample might be large, a considerable amount of information available on the map would be discarded.
3. The woodland blocks in the partial sampling frame might be stratified by the value of the block's map area so that a stratified random sample could be selected for area determination. Such an approach has the advantage of simplicity and represents a 'safe' strategy within which robust unbiased 'expansion' estimation is always feasible. However, such an approach would require that the map areas of all units in the sampling frame would have to be measured, and yet, unless the stratification were fairly fine, this information obtained from maps would not be fully used.
4. Once a suitable sample of units has been selected the full map information might be utilised by using a regression estimation approach.

Two major, and related, factors which influenced the final choice of the design were the availability of astrafoil copies of the woodland plates at 1:50 000 scale for the whole country and the fact that these maps could, with current computer technology, be represented in computer-digitised form. It was eventually decided that the maps would be digitised, partly in order to save manpower, and partly with a view to improving the efficiency of the subsequent analysis. Within this framework it was consequently decided that option (iii), a stratified sample, would be taken, retaining the option of using a regression estimator (iv) at the final estimation stage.

Early trials devoted much effort to evaluating the possible use of aerial photographs to determine the areas of forest types within woodland blocks. With photography, taken in good conditions and of suitable scale, a fairly reliable distinction can be made between coniferous and broadleaved crops. However, much of the available photo cover was at small scale. A trial was run to test the degree of discrimination between forest types by using an objective classification of photos based on the tone, texture and pattern. It was found that the discriminating power between some types was rather low. For example the chances of distinguishing between broadleaved high forest and scrub was as low as 50%. Though such an imprecise discriminatory technique could have been efficiently included in a valid sample design, it was decided that only boundaries and area should be checked from aerial photos and forest type and species proportion determined by ground assessment.

The calculation to determine the sample size (and optimal distribution) necessary within this design, to achieve an expected precision of 5% on total area estimate, made use of the digitised area data and assumed that a regression type estimator was to be used. Further details about sample size determination are given in Chapter 5 (see also Rennolls, 1981).

Second Phase Sample

The sample unit for ground survey was chosen to be the same as for aerial survey, that is, a woodland block, and the ground sample taken was a random subsample from the first stage sample. Hence, the ground survey was also a stratified random sample. The size and distribution of this subsample between the strata were determined by a modified Neyman optimal allocation procedure so as to ensure that the expansion estimate of the main forest type would have an expected standard error of 15%. For details of the sample size determination method and the estimator actually used, once the data had been collected, see Chapter 5.

CHAPTER 4

The survey design for non-woodlands

Introduction

The objective of the non-woodland survey was to obtain quantitative information on the distribution of isolated trees and small woods which were not represented on the OS maps. There were very many features of isolated trees and small woods which had to be measured in order to answer the many questions posed relating to the ecology and landscape of the British countryside. Precision targets had been set on the number of measurable isolated trees in predefined groups of counties (20% s.e.), in counties (25% s.e.), and of a predefined species, usually the most widely represented in a county, (30% s.e.).

A further objective of the non-woodland survey was to complement the 'woodland partial sampling frame' by allowing woodland blocks which were either digitised from the OS map as less than 0.25 ha or which were not represented there to be sampled. This ensured that the woodland population was completely covered by two complementary sampling frames.

Before considering the particular circumstances and constraints which resulted in the design used in this survey, it is worth mentioning briefly the survey designs used in the 1951 and 1965–67 surveys of non-woodland trees. It would have been simplest to have adopted the same survey design as used previously. There would have been a limited statistical design and analysis requirement, and comparisons between the different survey results would have been simpler.

In 1951 the assessment unit had been a strip of land one mile long and two chains (44 yards) wide. The position of an assessment unit was taken at a fixed central position of a six inch (1:10 560) map which itself had been selected systematically from the six inch maps covering the country. Three such systematic selections of maps were made, the starting point of each sequence being chosen randomly. The data from all three selections of maps were combined for population estimation, and standard errors were obtained from the variation between the estimates from the three separate selections of maps.

The same assessment unit was used in the 1965–67 survey as in 1951 and a third of those selected corresponded to one of the systematic sets obtained in 1951. One theoretical advantage of the later design, involving a partial replacement of the 1951 sample units, was that the comparison of measurements on the same assessment unit on different dates potentially gave very precise estimates of change. However, it was found to be difficult to ensure that exactly the same area was being assessed on the two dates. Also since the time interval was fairly large, changes in assessments had been considerable and variable. Hence little correlation between assessments of the same unit was found and comparisons between the two survey estimates did not make use of the overlapping nature of the samples. The samples were treated as independent random samples.

The main potential advantage of overlapping samples was therefore not realised and this suggested that there was little to be gained by constraining the 1979–1982 survey to overlap the 1965–67 survey. The 1979–1982 survey was treated as a sample survey independent of previous surveys and hence the choice of sampling unit and sample design had to be made in the light of current circumstances, resources and constraints.

Preliminary Considerations

Previous non-woodland surveys had been conducted largely by the method of ground visit and assessment and necessarily demanded a large input of manpower. In view of the strict limit of this resource it was considered almost inevitable from the start that the survey of non-woodlands would make substantial use of aerial photography to collect the required data and could be used both for the measurement and counting aspects of the survey of non-woodlands. However, it would have been even more difficult to identify species in the non-woodland survey than in the woodland one. Also the measurements from photographs necessarily involved errors in interpretation and hence some ground truth calibration of aerial-photo measurements and counts was considered essential. Hence, it was decided that species determination was to be done entirely by ground visit and these ground visits were also to be a means of validating extensive aerial photo measurements.

A further factor which encouraged the use of aerial photographs in the non-woodland survey was that there already existed a substantial body of commercially available photographs both in black and white and colour. It was expected that such material could be easily and cheaply acquired and even though it might have been not entirely current, the measurements of such features as the length of hedgerows and counts of large individual trees would be closely related to actual values, and hence form a sound basis for estimation.

The Design

The general structure of the design

The survey was required to make use of a sample of aerial photographs and to calibrate the resulting data by ground truth assessments on subsamples of photographic units. There were therefore two sets of data which both measured the same variable on the same part of the population, though the more extensively used measure was the less accurate. Thus the non-woodland survey had a two-phase structure, as did the woodland survey, in which the use of aerial photographs was the first phase and the ground assessment the second stage.

The target population of the non-woodland survey was set as for the woodland survey, at the level of a county. However, within a county, which is largely an administrative unit, there are considerable variations in the type of non-woodland environment. It was therefore felt that substantial gains in efficiency would be obtained by stratifying the population in a suitable manner.

It was clear that different regions had very different levels and variabilities of individual trees and hedgerow cover. It was decided that the best method of defining potentially useful sampling strata was to use the already available soil maps as a starting point. Besides affecting such features as individual trees and hedgerows it was also felt that they would have a strong effect upon the presence or absence of various species and hence increase precision of some species estimates from the second phase ground data. Some soil groups were in fact combined together in order to obtain sampling groups well suited to the efficient achievement of the non-woodland survey objectives. The decision on which soil groups to combine in order to obtain sampling strata was made as a result of a pilot survey conducted at the level of county group, a concept described below. The distributions of numbers of individual trees observed in pilot primary units were compared and those found to be not significantly different were combined to produce the sampling strata of the survey.

Though the prime target population was chosen to be at county level it was decided that a higher level population should also feature in the objectives of the survey. This higher level was termed a 'county-group' and consisted of a contiguous set of counties which might have reasonably have been expected to be similar. Part of the reason for the introduction of this regional population level was that when initial pilot data were analysed it was found that the degree of sampling necessary to obtain the initially defined precision objectives would have been excessive. The county-group was a means of obaining the same numerical accuracy but on a more extended regional population. The population structure therefore consisted of county groups subdivided into counties, with each of these counties divided into sampling strata.

The two-phase sample design

It was expected that the two-phase design would enable the efficient use of data from the two assessment levels (aerial photographs and ground assessments). They would be combined to estimate accurately such features as the number of individual trees and length of hedgerow features. However, there were other population values, such as the number of trees of a given species, size and health class, for which estimation was required but for which the aerial photographs provided no useful information for estimation purposes. For such population values it was necessary to rely entirely upon the ground assessed data which should itself constitute a well defined and valid probability sampling structure.

The large expense involved in travelling between ground sampling units suggested the usefulness of a clustered sample design in which the secondary units are sufficiently small to be easily assessable. Figure 3 illustrates the general nature of the two-stage sampling structure within one sampling stratum.

First, the stratum was divided up into a number of primary units, for convenience numbered in Figure 3 by 1, 2, 3 . . . etc. in an obvious way (from left to right, top to bottom). Each of these primary units in fact consists of a number of subunits, termed secondary units (in this example, four): a primary unit is a *cluster* of secondary units. A clustered random sample of secondary units was selected by following a two-stage procedure. Suppose five primary units are selected at random from the available 37 primaries and they are 11, 15, 21, 24 and 29 as indicated in Figure 3. These primaries would be assessed using aerial photographic interpretation to give first phase information. From these five primaries an appropriate subset (three in this example) of primary units would then be randomly selected, each of which will have a secondary stage of subsampling. Each of the three primary units selected for secondary subsampling has four secondary units. In this illustration of Figure 3 we have randomly selected two of these secondaries from each of the three primary units to obtain our clustered sampling of secondary units. These are shown in black in Figure 3.

The design can therefore be seen to be rather complex since it involves stratification, a two-stage sampling structure for secondaries and a two-phase data structure. There are several sample size numbers given in this illustrative example which have to be determined in such a way that the estimates from the collected data will match the precision objectives of the survey.

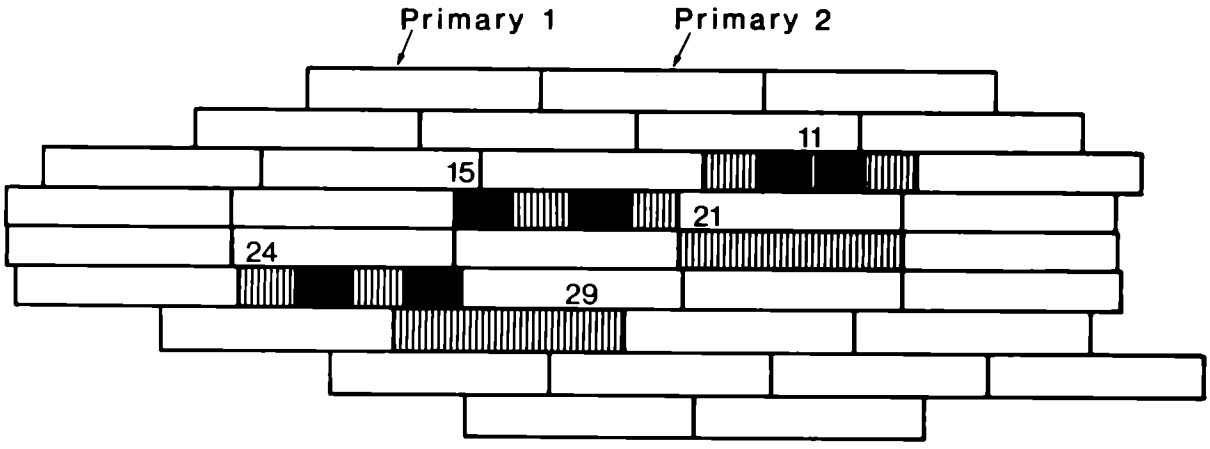


Figure 3. The sample structure of the non-woodland survey within one sampling stratum.

Several of the aspects of the numerical structure of the design were determined by practical limitations and judgements concerning what would be a reliable and robust procedure. Some of these choices evolved as the survey progressed throughout the country to take into account different circumstances such as availability of photographic cover, resource limits and improving knowledge and assessment efficiency.

In a preliminary study of survey methodology conducted in W. Sussex in 1976/77 the primary unit was chosen to be a kilometre square and the initial intention had been to assess the whole of a sample of such units on the ground, thus collapsing the clustered sample structure of the design. A study was conducted in order to determine the best number of secondaries to be selected when ten 100 m × 1 km secondary strips made up the primary unit. The intraclass correlation was so low that maximal clustering seemed best, resulting in the elimination of the need for secondary subsampling.

However, in subsequent counties the 1 km square was not found to be a suitable primary unit for obtaining first phase aerial data. From the point of view of flying, a strip of secondary units was most suitable. It was regarded as prudent to get as wide a dispersal of ground assessed secondary units as possible and hence it was decided that only two secondaries would be selected per primary. This was the minimum figure to allow a valid calculation of estimates and standard errors using the ground data as a two-stage stand-alone survey.

Practical convenience, both for photography and ground assessment, led to some early counties having a primary strip of six secondary squares each 500 × 500 m (25 ha). Economic limitations later led to a change to the use of a primary strip consisting of twelve squares, each 250 × 250 m (6.25 ha), with some loss in precision. In Scotland the lack of recent aerial photography and the more difficult terrain meant the main criterion for the choice of primary unit was the minimisation of walking time between ground assessed secondary units. Hence a primary strip comprised eight 250 × 250 m secondaries in a 1 km × 500 m format. Experience in the earlier counties surveyed confirmed the impression of low intra-class correlations. Even so, the secondary sampling intensity was maintained at two per primary, in order to obtain a wide ranging distribution of ground-truth data.

In Chapter 5 where a mathematical treatment is given, all of these possible sample structures are included within one formulation.

Sample Size Determination

The methods of sample size determination used for the non-woodland survey are approximate since the final and most complete estimator from the complex design is not amenable to a rigorous treatment. Three different sets of estimating equations of increasing complexity were used when calculating the final estimates and their standard errors.

The simplest, termed Mode 1, made an estimate for a county using the simple expansion estimator on the data from the stratified data obtained from the aerial primary units. Mode 2 estimation made use solely of the data collected from ground visited secondary units using the expansion estimator for a stratified two-stage design. Mode 3 was a regression type estimator which made simultaneous use of the complete two-phase structured data.

Sample size determination assumed Mode 1 estimators or an approximation of the Mode 2 estimators. The approximation was in fact to ignore the clustered structure of the secondary units and to assume the use of the standard expansion estimator for a simple random sample of the same number of secondary units.

The whole process of sample size determination can therefore be regarded as a number of optimum allocation calculations subject to a constraint on the minimum numbers of samples in the strata.

The first constraint imposed on sample sizes was that there should be a minimum of four clusters per sampling stratum. This constraint was set above the minimum of two units per stratum in order to obtain fairly reliable information on the variability within the sampling strata. A further reason for adopting this minimum was that it was envisaged that in certain cases Mode 3 estimation using the close regressive relationship between the variates recorded at the two phases might be used predictively to make estimates for a sampling stratum. It was thought that eight points would be a minimum with which to establish any confidence in such a relationship.

Pilot data were used to obtain approximate values for the mean and variance of the number of trees of different species in each of the sampling strata. These data also gave guiding values for the mean and variance of the total number of isolated trees in a primary unit in each sampling stratum. These were used to determine the required number of primary samples and their optimum distribution between strata in order to obtain an expected s.e. of 25%. A Mode 1 estimator for the total number of trees was assumed and the optimum allocation was subjected to the constraint of a minimum four primaries per stratum. Details of the constrained optimal allocation algorithm are given in Appendix 5A.

Once the sample sizes required in a county had been so determined these were taken as constraining minimal sample sizes in a repeated optimal allocation to ensure an expected s.e. target of 20% on the total number of isolated trees in a county group. In some county groups it was found necessary to increase the county sample sizes beyond that required to satisfy the county-level precision targets, hence increasing the expected county level precisions.

The final stage of the sampling size determination process for non-woodlands was to calculate the number of units that needed to be visited on the ground. To do this, pilot data on the mean and variance of the numbers of trees of the three main species in secondary units were used. The strata sample sizes already obtained from the two previous stages of constrained optimal allocation were used as constraints in the calculation of the number of secondaries required in order to attain standard errors of 10%, 20% and 30% on the total numbers of trees of each species. A simple random selection of secondaries was assumed.

The resulting sets of sample size distributions were then used, together with knowledge of resource and time limitations, to decide the eventual sample size. The actual selection of the sample units depended on the repeated use of a pseudo-random number generator.

The quantity of pilot data on individual tree species was often very limited and hence there was some doubt on embarking on the field work as to what the final attained precisions would be. The sample size was therefore sometimes modified in a sequential manner. After a certain number of units had been assessed, the data were used in conjunction with the pilot data to give firmer input values for the sample size determination programs. In this way a grossly imprecise estimate for the number of isolated trees of a particularly important species could be avoided.

CHAPTER 5

Mathematical details of sample size determination for the woodland survey

This Chapter sets out the mathematics underlying the structure of the survey design described in Chapter 3. Details are given of the sample size determination methods for both phases of the woodland survey.

The Woodlands Design

To specify the Woodland Survey design we need to define four aspects. First, the population of elements to be surveyed must be defined. Second, the sample unit(s) must be defined, and thirdly it is necessary that the collection of sampling units (i.e. the sampling frame) covers the population without duplication. Fourthly, any structure to be imposed upon the sampling frame, before the random selection of the sample, must be specified.

Briefly, our population is the set of woodlands in a county, we choose the OS represented blocks of sufficient size to be our partial sampling frame, and we stratify this frame according to the measured map area of the woodland. More formally:

1. The population to be surveyed is the set of woodland blocks, a block being defined as having an area greater than, or equal to, 0.25 ha. Let y_i denote the actual area of the i^{th} woodland block and Y represent the sum of these areas in the whole population. The estimation of Y is one of the main objectives to be achieved by the first phase of the woodland survey.
2. This population is completely covered by two non-overlapping partial sampling frames, the first of which is derived from the green regions on the 1:50 000 OS map and the other is that used in the 'non-woodland' survey (see Chapter 6 for details). It is important to remember that the final estimate of Y, i.e. $\hat{Y}$, is obtained by adding together the estimates from the two partial sampling frames. Since the samples in these two frames are independent of each other the variance of $\hat{Y}$ is obtained by adding the variances of the estimates obtained from the separate frames.

The OS based partial sampling frame consists of those regions of green which have a map area of 0.25 ha or more: these constitute the sampling units. Suppose there are N such sample units in the partial frame and x_i denotes the map area of the i^{th} sample unit. This 'auxiliary' variable is used to stratify the sampling frame into H strata, so that the i^{th} sampling unit is in the h^{th} stratum if $b_{h-1} \leq x_i < b_h$, where $h \in \{1, \ldots, H\}$ and the b_h are the limiting sizes which define the strata, with $b_0 = 0.25$ and $b_H = \infty$. Denote the number of units in the h^{th} stratum by N_h. The stratum limits, b_h, were chosen using a number of criteria, though they were not formally determined in an optimal manner.

The first consideration was that, because woodland block map areas were highly correlated with actual areas, the estimation of total area would be primarily based upon a regression method of estimation. Such estimates are most precise if the sampling proportions are highest at the ends of the x-variable distribution, so to some extent the boundaries were chosen to ensure that this would happen (by imposing minimal sample sizes per stratum). Also, a number of intermediate strata were defined to allow enough intermediate data to be collected to ensure that the regression model used in estimation was of an appropriate form.

Furthermore, the stratified sampling frame was to be the basis upon which forest-type estimates were to be made, using 'expansion type' estimators. If the strata boundaries are predefined then the optimal allocation of samples depends on the number of units in each stratum, on the variability of the units within strata, and the cost of sampling the units in the various strata. This process was partially reversed in a rather intuitive manner, the boundaries being determined so that a roughly equal variance of the forest-type area per block was obtained in each stratum.

Trials were made of the possibility of a two-way stratification of the sampling frame, first by block area and second by soil type. This seemed to result in no significant increase in precision of the estimates, so this approach was not pursued. Since soil type of sample blocks was recorded, it is still possible to construct estimators of the variates within soil types.

Sample Size Determination

The operational sequence of conducting the woodland survey started with measuring the map areas of woodland blocks, followed by the selection of blocks which have their actual areas measured (usually from aerial photographs), followed in

turn by the selection of a sample of woodland blocks which were ground visited to determine the proportions of the blocks in the different forest-type classes. However, for sample size determination purposes, this sequence is followed in the opposite direction. First, we determine the number of blocks needing to be ground visited, in order to achieve major forest-type precision targets. These sample sizes are then treated as minima in the determination of the number of extra blocks which have to be aerially assessed, to determine their total area.

Ground sample size determination

The main problem at this stage was the almost total absence of reliable pilot data on means and variances of forest-type areas on which to base calculations. Three approaches were adopted in order to give a suitable range of backgrounds from which the eventual choice of sample distribution might be made. The first two approaches, based upon what are termed 'proportional' and 'binomial' forest types, assumed no pilot data and depended for their use upon the 'experience and judgement' of the forest survey managers, together with a knowledge of the means and variances of the map areas, whilst the third approach, the 'empirical', depended upon data from pilot surveys.

Note that the means and variances may be calculated from:

$$\overline{X}_h = \frac{1}{N_h} \sum_{j=1}^{N_h} x_{hj} \tag{5.1}$$

$$S_h^2 = \frac{1}{(N_h - 1)} \sum_{j=1}^{N_h} (x_{hj} - \overline{X}_h)^2 \tag{5.2}$$

where x_{hj} is the map area of the j^{th} unit in the h^{th} stratum. The notation used here and subsequently is the usual for sample survey theory as may be found in Cochran (1963): $\overline{X}_h$ is the population mean of the x-values in the h^{th} stratum.

i. Model 1: Proportional forest types

Let z_{hj} denote the area of the j^{th} block in the h^{th} stratum which is of a particular forest type. The type chosen is arbitrary, but for the Phase II sample size determination process the type of major importance was chosen.

Suppose that a proportion p_{1h} of each block in the h^{th} stratum is of the forest type of major interest, choice of p_{1h} relies on the experience and judgement of the forest surveyors. Then

$$z_{hj} = p_{1h} x_{hj}; \quad j = 1 \ldots N_h, \quad h = 1 \ldots H \tag{5.3}$$

for all h and j. It is assumed that x_{hj} is sufficiently close to y_{hj} for sample size determination purposes. It follows that the mean value of z in the h^{th} stratum is given by

$$\overline{Z1}_h = E(z_{hj}) = p_{1h} E(x_{hj}) = p_{1h} \overline{X}_h \tag{5.4}$$

The '1' in $\overline{Z1}_h$ and $S1_h$ indicates that we are working with Model '1' – the proportional model.

The population variance of z in the h^{th} stratum, denoted by $S1_h^2$ is given by

$$S1_h = p_{1h}^2 S_h^2 \tag{5.5}$$

However, equations (5.4) and (5.5) are not by themselves sufficient for our purposes since the model given in (5.3) assumes that all units in the frame really are woodlands. This is not in general the case since some blocks in our partial frame might have been misrepresented on the OS map and have $y<0.25$ even though $x \geq 0.25$. The problem of an indeterminate population and sample size is avoided in the following way.

Suppose that the probability of a unit randomly chosen from the h^{th} stratum having $y_{hj} \geq 0.25$ is p_{4h}. If the block's map area is x_{hj}, then

$$\begin{aligned} z_{hj} &= p_{1h}.x_{hj} \quad \text{with probability } p_{4h} \\ &= 0 \qquad\quad \text{with probability } (1-p_{4h}) \end{aligned} \tag{5.6}$$

Hence,

$$\overline{Z1}_h = E(z_{hj}) = p_{4h}.p_{1h}.E(x_{hj}) + (1 - p_{4h}).0$$

that is,

$$\overline{Z1}_h = p_{4h}.p_{1h}.X_h \tag{5.7}$$

Similarly,

$$\begin{aligned} S1_h^2 &= E(z_{hj}^2) - E^2(z_{hj}) \\ &= p_{4h}.p_{1h}^2.E(x_{hj}^2) - p_{4h}^2.p_{1h}^2.E^2(x_{hj}) \end{aligned}$$

that is

$$S1_h^2 = p_{4h}.p_{1h}^2 [S_h^2 + \overline{X}_h^2 (1 - p_{4h})] \tag{5.8}$$

Expressions (5.7) and (5.8) give the mean and variances assuming the 'proportional model'.

ii. Model 2: Binomial forest types

The assumption of proportionality is likely to underestimate the true variability of the areas of the forest type. The binomial model represents the opposite extreme where p_{1h} is taken to be a probability value and where we assume that

$$\left.\begin{aligned} z_{hj} &= x_{hj} \text{ with probability } p_{4h}.p_{1h} \\ &= 0 \quad \text{with probability } (1 - p_{4h}.p_{1h}) \end{aligned}\right\} \tag{5.9}$$

and p_{4h} is as before the probability of a unit being a real wood.

Using the same methods as for the proportional model, we obtain

$$\overline{Z2}_h = p_{4h}.p_{1h}.\overline{X}_h \tag{5.10}$$

the same as for the proportional model, and

$$S2_h^2 = p_{4h}.p_{1h}[S_h^2 + \overline{X}_h^2 (1 - p_{4h}.p_{1h})] \tag{5.11}$$

The use of $S2_h^2$ will produce a larger sample size than will the use of $S1_h^2$. In most cases, in the absence of real pilot data $S2_h^2$ was adopted as the basis of sample size determination.

iii. Empirical approach

A third approach to getting pilot values for the variance of the z–variate is to use those figures which have been obtained in similar counties in which the survey has already been completed. In addition to the use of these approaches, a sequential approach was used in many counties. The data obtained from the first few sampled units were used as pilot data for the calculation of optimal sample sizes. This procedure was executed in such a way that the samples taken at any stage constituted a valid random sample. For convenience we denote the empirically determined mean and variance of the z–variate in the h^{th} stratum by $\overline{X3}_h$ and $S3_h^2$.

Sampling optimisation

Using ZI_h and SI_h^2 for, $I = 1,2,3$ and a range of target precisions the optimum sample size and distribution were calculated using the analytic method given in Appendix A. These optimal sample distributions achieved the target precision subject to imposed constraints on minimal sample sizes per stratum and did so at minimal cost. The costs of assessing units in different size strata had to be specified. This was not always easy to do. For though a large block will take longer to assess than a smaller block, and hence cost more, it will also generally yield more information on a range of forest types. In the absence of any clear rationale for differing costs of assessment per block they were taken to be equal.

Finally, a further constraint was imposed upon the total sample size: namely that it should be no less than 30. This arbitrary minimal figure was chosen in view of the rather weak pilot data available for this part of the survey. This constraint was achieved by starting from the optimum sample allocation resulting from the analytical method and, if necessary, repeatedly using the numerical method (of Appendix A) until the total sample size reached the required minimum size.

Aerial photo sample size determination

The methodology was very similar to that used in the previous section. The digitised map areas provide a preliminary knowledge of the area distribution of woodland blocks. The target precision could then be used to find an approximate variance target for this part of the survey.

Below, we set out an approximation to the variance of the estimator resulting from a combined use of a 'regression' estimator and a 'Binomial-model-expansion' estimator. This variance formula is a complex function of the sample sizes and an analytic determination of the optimal sample distribution is not possible. The sample sizes determined for ground visit are regarded as minimum sample sizes and sample sizes are increased by one sample at a time until the target precision is obtained. The stratum in which this extra sample is chosen is determined by maximising the decrease in estimated variance per unit cost at each step. The equivalence of this numerical approach and an exact mathematical solution for the case of stratified random sampling provides some justification for the validity of this approach.

There is generally a very close relationship between map area of a block and its actual area (see Figure 4a). However, this relationship is less strong for the stratum of woodland sample units having smallest size, i.e. <2.00 ha, (Figure 4b). Hence stratum 1 is treated in the same way as was described in the previous section and a regression predictor is used to estimate for those blocks in the larger size classes.

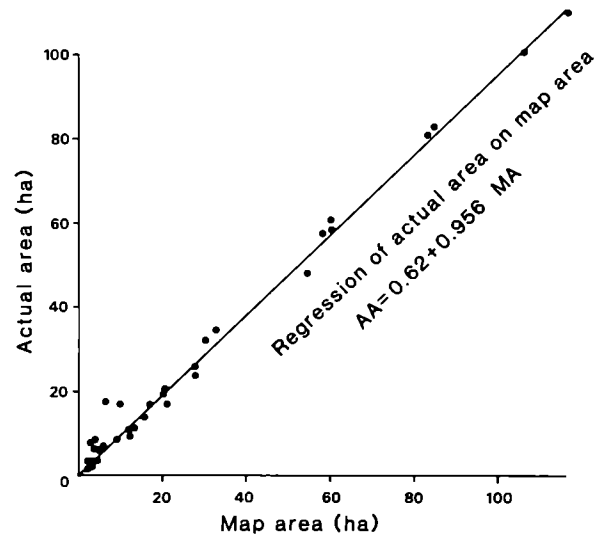


Figure 4a. Scatter diagram of actual area against map area for Avon.

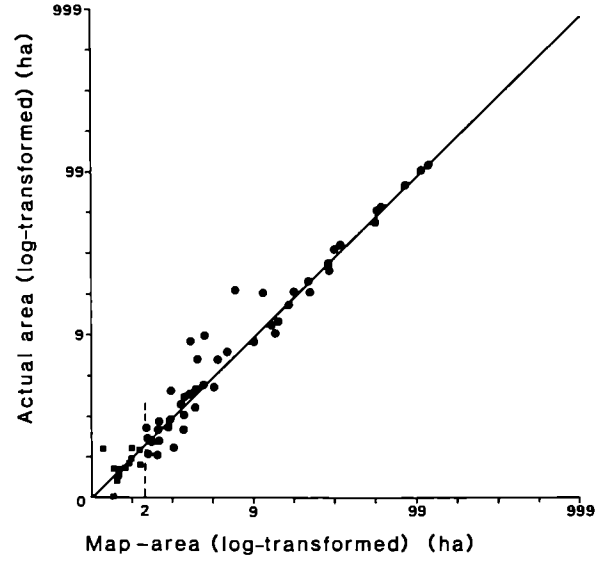


Figure 4b. Log_{10} transformed woodland data for Avon.

The justification for the exact form of this estimator is given in Chapter 6 and Appendices 5B and 5C. Using the notation of Section 5.2.1 we have approximations,

$$\overline{Y}_1 \simeq p_{41}.\overline{X}_1 \tag{5.12}$$

$$\text{and } \text{var}(\overline{Y}_1) \simeq p_{41}[S_1^2 + X_1^2(1-p_{41})]/n_1 \tag{5.13}$$

for stratum 1. Note that these are obtained from (5.10) and (5.11) by setting $p_{11}=1$.

For strata 2, . . ., H (denoted 1^+) we denote the mean block area $\overline{Y}_{1+}$,

$$\overline{Y}_{1+} = \frac{1}{N'} \sum_{h=2}^{H} \sum_{i=1}^{N_h} X_{hi} \tag{5.14}$$

and using Appendix 5C we have.

$$\text{var}\,(\hat{\overline{Y}}_{1+}) \simeq \frac{\sigma^2(1-f)}{n'}\,[1+(1-f)G] \tag{5.15}$$

where

$$G = \frac{(\overline{X}_R - \overline{X}_f)^2}{\sum_{h=2}^{H} (w_h S_h^2 + w_h(\overline{X}_h - \overline{X}_R)^2 - S_h^2/n')}$$

$$n' = \sum_{h=2}^{H} n_h \qquad N' = \sum_{h=2}^{H} N_h,$$

$$f = \frac{n'}{N'},$$

$$\overline{X}_R = \sum_{h=2}^{H} w_h \overline{X}_h,$$

$$w_h = \frac{n_h}{n'},$$

and

$$\overline{X}_f = [(\sum_{h=2}^{H} N_h \overline{X}_h) - n'\overline{X}_R]/(N'-n')$$

If the coefficient of variation of our precision target, expressed as a proportion, is p_3, then the stopping condition for our numerical optimisation is given by,

$$N_1^2\,\text{var}\,(\hat{\overline{Y}}_1) + (N')^2\,\text{var}\,(\hat{\overline{Y}}_{1+}) \leq [p_3(N_1\hat{\overline{Y}}_1 + N'\hat{\overline{Y}}_{1+})]^2 \tag{5.16}$$

Appendix 5A

Constrained Optimal Allocation of a Stratified Sample

Using the standard notation, with Y as a general population variate, and y as the corresponding sample value, the estimator of the population mean is given by

$$\hat{\bar{Y}}_{st} = \sum_h W_h \bar{y}_h \quad ; \quad W_h = N_h/N \tag{5A1}$$

and its variance is estimated by

$$\hat{\text{var}}\,(\hat{\bar{Y}}_{st}) = \sum_h \frac{W_h S_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right) \tag{5A2}$$

The aim is to ensure that there is a minimal sample size m_h in the h^{th} stratum and to make an estimate of the $\hat{\bar{Y}}_{st}$ which has a standard error equal to $p_2\,\hat{\bar{Y}}_{st}$. Furthermore we wish to do this at minimum cost. Suppose that the 'cost' of taking a sample in the h^{th} stratum is c_h. Thus the target precision is to be attained subject to the constraints $n_h \geq m_h$ whilst minimising the total cost,

$$C = \sum_h n_h c_h \tag{5A3}$$

Two alternative methods can be used to determine the constrained optimal sample sizes. The first method is numerical whilst the second method is analytical. Both are presented, since we had occasion to use both techniques.

Numerical method

If $\text{var}(\hat{\bar{Y}}_{st})$ calculated from (5A2) with $n_h = m_h$ is less than $(p_2.\hat{\bar{Y}}_{st})^2$ then we need take no more than the minimal sample sizes. Otherwise, we take an additional sample unit in that stratum such that the decrease in variance per unit cost is maximal.

From (5A2) this quantity is given by

$$\triangle V_h = \frac{W_h^2 S_h^2}{n_h (n_h + 1) c_h} \tag{5A4}$$

in the h^{th} stratum. This is repeated until the required precision is obtained.

Analytical method

This method was derived by A. Abakuks (1979, personal communication).

The precision target may be re-expressed by using (5A1) and (5A2) to give

$$U = \sum_h \frac{W_h S_h^2}{n_h} = (p_2 \sum W_h \bar{y}_h)^2 + \frac{1}{N}\sum W_h S_h^2 \tag{5A5}$$

Calculate the values,

$$\left.\begin{aligned} g_h &= \frac{W_h S_h}{m_h \sqrt{c_h}} \quad ; \quad h = 1, \ldots, H \\ \text{and set } g_0 &\equiv 0 \end{aligned}\right\} \tag{5A6}$$

Then order the g–values to give

$$0 \equiv g_0 \leq g_{(1)} \leq g_{(2)} \leq \ldots \leq g_{(H)} \tag{5A7}$$

For the strata in the same order as assigned in (5A7), define

$$\phi_r = \frac{U - \sum_{(h)=(1)}^{(r)} \frac{W_{(h)}^2 S_{(h)}^2}{m_{(h)}}}{\sum_{(h)=r+1}^{(H)} W_{(h)} S_{(h)} \sqrt{c_{(h)}}} \; ; r = 1, \ldots, H \qquad \textbf{(5A8(i))}$$

and

$$\phi_0 = \frac{U}{\sum_{(h)=1}^{(H)} W_{(h)} S_{(h)} \sqrt{c_{(h)}}} \qquad \textbf{(5A8(ii))}$$

Note that (1) represents the original stratum number of that stratum having minimum non zero g–value.

Finally, find

$$r^\dagger = \max [0 \leq r \leq (H-1) ; g_{(r)} \leq \phi_r] \qquad \textbf{(5A9)}$$

and set

$$\phi = \phi_{r\dagger}$$

The constrained optimum sample sizes are given by

$$\begin{aligned} n_{(h)} &= m_{(h)} \text{ for } 1 \leq (h) \leq r^\dagger \\ &= \frac{W_{(h)} S_{(h)}}{\phi \sqrt{c_{(h)}}} \qquad \text{for } (r^\dagger + 1) \leq (h) \leq H \end{aligned} \qquad \textbf{(5A10)}$$

If $n_h \geq N_h$ for some value of h, then we set $n_h = N_h$ and re-apply the method to the remaining strata noting that U must be decreased by $\Sigma N_h S_h^2$ where the summation is over those strata with an enumerative sample.

The analytical and numerical methods have given identical sample sizes on all comparative trials of the two methods. There is, however, no formal proof of their equivalence available even though it seems likely that this is the case.

A Pascal program implementing this method is available on request[1] as is a more extensive theoretical treatment of this method. Further work on constrained optimal allocation may be found in Hughes and Rao (1979).

[1]. This programme was written by Mr G. J. Hall.

Appendix 5B

Estimation by Prediction Using the General Linear Model

Linear model: estimation and prediction

Before considering specific regression-type estimators a short review of relevant linear-model theory is presented (Searle, 1971). For a particular element, the i^{th} say, we adopt the model,

$$E(y_i)=\boldsymbol{Z}'_i\,\beta;\quad i=1,\ldots,n \tag{5B1}$$

The regressor vector $\boldsymbol{Z}'_i$ would be $(1, x_i)$, (x_i), (1) for the linear regression, linear regression through the origin, and the minimal model. We may re-write and extend (5B1) as follows,

$$\boldsymbol{y}=\boldsymbol{X}\beta+\epsilon \tag{5B2}$$

where

$$E(\epsilon)=\boldsymbol{0},\quad \mathrm{E}(\epsilon\epsilon')=\boldsymbol{V},\quad \boldsymbol{X}=(\boldsymbol{Z}'_1,\ldots,\boldsymbol{Z}'_n)' \tag{5B3}$$

The GLS estimate of β is given by

$$\hat{\beta}=(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y} \tag{5B4}$$

with

$$\mathrm{var}(\hat{\beta})=(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1} \tag{5B5}$$

Suppose there is a single further observation $\boldsymbol{Z}_f$. Then the estimated *expected value* of y corresponding to $\boldsymbol{Z}_f$, that is y_f, is

$$E(y_f)=\boldsymbol{Z}'_f\hat{\beta} \tag{5B6}$$

Also, the *predictor* of y_f is given by,

$$\tilde{y}_f=\boldsymbol{Z}'_f\hat{\beta} \tag{5B7}$$

Though these estimates have the same form, they have differing variances,

$$\mathrm{var}[\widehat{E(y_f)}]=\boldsymbol{Z}'_f(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{Z}_f \tag{5B8}$$

$$\mathrm{var}(\tilde{y}_f)=\boldsymbol{Z}'_f(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{Z}_f+v_f \tag{5B9}$$

where $v_f=\mathrm{var}(y_f|\boldsymbol{Z}_f)$. If we have n_f independent further observations $\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_n$ having mean $\overline{\boldsymbol{Z}}_f$ then the estimated expected value of y corresponding to $\overline{\boldsymbol{Z}}_f$ and the predicted value of y corresponding to $\overline{\boldsymbol{Z}}_f$ (i.e. $\tilde{\bar{y}}_f$) are obtained from (5B6) and (5B7) by replacing $\boldsymbol{Z}_f$ by $\overline{\boldsymbol{Z}}_f$. Also it can be seen that the variance formulae become

$$\mathrm{var}[\widehat{E(\bar{y}_f)}]=\boldsymbol{Z}_f'(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\overline{\boldsymbol{Z}}_f \tag{5B10}$$

and

$$\mathrm{var}(\tilde{\bar{y}}_f)=\overline{\boldsymbol{Z}}'_f(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\overline{\boldsymbol{Z}}_f+\left[\sum_{i=1}^{n_f} v_{fi}/n_f^2\right] \tag{5B11}$$

These formulae suppose that the variance–covariance matrix of the data, $\boldsymbol{V}$, is known and that v_f is known as a function of $\boldsymbol{Z}_f$. If we take,

$$\boldsymbol{V}=\boldsymbol{I}\sigma^2 \tag{5B12}$$

than we may estimate σ^2 by,

$$\hat{\sigma}^2=(\boldsymbol{y}-\hat{\boldsymbol{y}})'(\boldsymbol{y}-\hat{\boldsymbol{y}})/(n-r)\quad ;\quad r=rk(\boldsymbol{X}) \tag{5B13}$$

where

$$\hat{\boldsymbol{y}}=\boldsymbol{X}\hat{\beta}$$

The regression estimator (i.e. $\boldsymbol{Z}'=(1,x)$)

The classical approach (Sukhatme, 1954; Cochran, 1963; Kish, 1965) uses the sample data to obtain $\hat{\beta}$ and $\hat{\sigma}^2$ from (5B4) and (5B13) (assuming (5B12). It then proceeds, rather curiously, to say; we have the mean population value of $\boldsymbol{Z}$, i.e. $\overline{\boldsymbol{Z}}=(1,\overline{\boldsymbol{X}})$; we choose as our estimator of the mean population value of y, i.e. $\overline{Y}$, from (5B6),

$$\hat{\overline{Y}}_c^r = E(\overline{\boldsymbol{y}}) = \overline{\boldsymbol{Z}}'\hat{\beta} \quad (r, \text{regression; } c, \text{classical})$$

with, from (5B8),

$$\text{vâr}(\hat{\overline{Y}}_c^r) = \boldsymbol{Z}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{Z}.\hat{\sigma}^2$$

i.e.

$$\text{vâr}(\hat{\overline{Y}}_c^r) = \frac{\hat{\sigma}^2(1-f)}{n}\left[1 + \frac{n(\bar{x}-\overline{X})^2}{\sum_1^n (x_i-\bar{x})^2}\right] \qquad \textbf{(5B14)}$$

where the introduction of $(1-f)$ to take account of sampling with non-replacement from a finite population is *ad hoc*. Sukhatme (1954, p.201), points out that (5B14) is conditional upon the sample actually drawn. He therefore averages over all simple random samples of size n, to obtain the approximate unconditional result (p.203),

$$\text{vâr}(\hat{\overline{Y}}_c^r) \simeq \frac{\sigma^2(1-f)}{n}\left\{1+(1-f)\left[1+\frac{3}{n}-\frac{6}{N}+\frac{\beta_2}{N}+2\beta_1\left(\frac{1}{n}-\frac{1}{N}\right)\right]\right\} \qquad \textbf{(5B15)}$$

where $\beta_1=\mu_3^2/\mu_2^3$ and $\beta_2=\mu_4/\mu_2^2$.

Unfortunately the approximations are valid only for a simple random sample, though Kish and Frankel (1974) conjecture that such results may be used with little loss for proportionately stratified populations. Since we intend to choose a sample from a stratified frame which is optimal with respect to a regression type estimator the conjecture is not applicable in our case.

The regression predictor

The essence of the prediction approach lies in that the inference is conditional on the data sampled, and that the prediction equations of the last section should only be used on those elements of the population which are not in the sample. The conditional approach to inference is justified by regarding the sample data as ancillary (Kalton, 1976; Holt, Smith and Winter, 1980; Cox and Hinkley, 1974). The estimation of $\hat{\beta}$ and $\hat{\sigma}^2$ is as in the previous section, but only the (N–n) units not sampled have their mean value predicted (5B7) to yield (Royall, 1970; Smith, 1976),

$$\hat{\overline{Y}}_p^r = \frac{1}{N}\left(\sum_1^n y_i + (N-n)\hat{\overline{Y}}_f\right) \qquad ,(p, \text{prediction}) \qquad \textbf{(5B16)}$$

where

$$\hat{\overline{Y}}_f = \overline{\boldsymbol{Z}}_f'\hat{\beta} = (1, \overline{X}_f)\hat{\beta} \quad ; \quad X_f = \left(X - \sum_1^n x\right) / (N-n) \qquad \textbf{(5B17)}$$

and from (5B16)

$$\text{vâr}(\hat{\overline{Y}}_p^r) = (1-f)^2 \text{ var}(\hat{\overline{Y}}_f) \qquad \textbf{(5B18)}$$

var $(\hat{\overline{Y}}_f)$ may be evaluated from (5B11) and yields

$$\hat{\text{var}}\,(\hat{\overline{Y}}_f) = \frac{\hat{\sigma}^2}{n}\left(1 + \frac{n}{N-n} + \frac{n(\bar{x}-\overline{X}_f)^2}{\sum(x_i-\bar{x})^2}\right)$$

i.e.

$$\hat{\text{var}}\,(\hat{\overline{Y}}_f) = \frac{\hat{\sigma}^2}{n(1-f)}\left\{1 + (1-f)\left[\frac{n(\bar{x}-\overline{X}_f)^2}{\sum(x_i-\bar{x})^2}\right]\right\} \qquad \textbf{(5B19)}$$

Hence combining (22) and (23) we get,

$$\text{var}\,(\hat{\overline{Y}}_p{}^r) = \frac{\hat{\sigma}^2(1-f)}{n}\left[1 + (1-f)\frac{n(\bar{x}-\overline{X}_f)^2}{\sum(x_i-\bar{x})^2}\right] \qquad \textbf{(5B20)}$$

The variance formulae (5B14) and (5B20) are similar; however, the prediction estimate has smaller variance and does not require *ad hoc* arguments to introduce the finite population correction factor. $\hat{\overline{Y}}_p{}^r$ is therefore preferable to $\hat{\overline{Y}}_c{}^r$. However, since we are still at the design stage var$(\hat{\overline{Y}}_p{}^r)$ must be averaged over all possible samples. An approximate result is given in Appendix 5C.

There is a difference in interpretation of the variance formulae (5B14) and (5B20) of the classical and predictive regression estimators. In the classical theory the variance arises from the randomisation distribution given by the repeated selection of a random sample from a finite population. In the predictive approach to estimation the variance arises from the conceptual variability of the realised population, and hence the sample about an essentially law-like relationship. This latter approach is often referred to as the superpopulation approach to sample survey estimation. Further discussion of these alternative approaches will be found in Smith (1976), Royall (1976), Cassel *et al.* (1977) and Rennolls (1981). Whichever interpretation is used we may rest assured that the numerical results of each approach will be very close, as demonstrated by the similarity of (5B14) and (5B20). When there is no prior knowledge of the population structure to guide the formulation of an appropriate model then the classical expansion estimators for a simple random sample are formally identical to those of the prediction approach. This is demonstrated in the next section.

The expansion predictor

We would like to see the form of the regression predictor when the model is minimal, i.e. on the sample data (assumed simple random for the moment).

$$\boldsymbol{y}=\boldsymbol{I}\mu+\boldsymbol{\epsilon};\quad E(\boldsymbol{\epsilon})=\boldsymbol{0};\quad E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}')=\boldsymbol{I}\sigma^2 \qquad \textbf{(5B21)}$$

It seems likely that the prediction approach can yield no better results than the classical expansion estimate. We have, $\hat{\beta}=(\mu)$, $\boldsymbol{X}=\boldsymbol{I}$ and from (5B4), (5B17) and (5B11)

$$\left.\begin{aligned} &\hat{\mu}=\bar{y};\quad \hat{\overline{Y}}_f=\bar{y};\quad \text{var}\,(\hat{\overline{Y}}_f) = \frac{\hat{\sigma}^2}{n(1-f)} \\ \text{Hence, using (5B16), (5B18) and (5B13)}\quad &\hat{\overline{Y}}_p{}^r = \bar{y};\quad \text{var}\,(\hat{\overline{Y}}_p{}^r) = \frac{\hat{\sigma}^2}{n}(1-f);\quad \hat{\sigma}^2 = S^2 \end{aligned}\right\} \qquad \textbf{(5B22)}$$

and var$(\hat{\overline{Y}}_p{}^r)$ does not change when the expectation is taken over all possible samples. So the general prediction method, when applied to the minimal model (5B21) yields the classical finite populaton results! These results generalise immediately to the stratified design giving identity between the classical expansion estimate and the predictive estimate for the minimal–main effect model.

Appendix 5C

Sample Size Determination Approximations for the Regression Predictor

For the regression predictor we have, from Appendix 5B,

$$\text{var}\,(\hat{\bar{Y}}_{\rho}{}^{r}) = \frac{\hat{\sigma}^2\,(1-f)}{n'}\left[1 + (1-f)\frac{(\bar{X}_R-\bar{X}_f)^2}{\frac{1}{n'}\sum(x_i-\bar{X}_R)^2}\right] \tag{5C1}$$

However, at the sample size determination stage we do not know $\bar{X}_R$ and the mean and variance of the sample x–values for the final survey. We find approximate expressions for these two quantities in terms of $\bar{X}_h$ and $S_h{}^2$, obtained from the digitised map areas, and the unknown sample sizes. Clearly

$$\bar{X}_R = \sum_{h=2}^{H} w_h\bar{X}_h \quad \text{where } w_h = n_h/n' \text{ and } n' = \sum_{h=2}^{H} n_h \tag{5C2}$$

Also,

$$\begin{aligned}\frac{1}{n'}\sum_i (x_i - \bar{X}_R)^2 &= \frac{1}{n'}\sum_{h=2}^{H}\sum_{j=1}^{n_h}[(x_{hj}-\bar{x}_h) + (\bar{x}_h - \bar{X}_R)]^2 \\ &\simeq \frac{1}{n'}\sum_h [(n_h - 1)\,s_h{}^2 + n_h\,(\bar{x}_h - \bar{X}_R)^2] \\ &\simeq \sum_h\left[\left(w_h - \frac{1}{n'}\right)s_h{}^2 + w_h\,(\bar{X}_h - \bar{X}_R)^2\right]\end{aligned} \tag{5C3}$$

both of which, when substituted in (5C1) give an estimate of variance as a function of map calculated values and the unknown sample sizes.

CHAPTER 6

The final estimators for the woodland survey

The sample size determination process described in Chapter 5 proceeded from the ground survey (formally termed the second phase in Chapter 3) to the aerial survey ('first phase') by a sequential constrained optimisation process. In this chapter, concerned with the final estimators which were used on the survey data, we revert to the original sequence and discuss in order first phase and then second phase estimation.

First Phase Estimation of Total Area of Woodlands (and standard error).

The primary objective is to estimate the total area of woodland in a county and to obtain information on the distribution of that area across size classes. Estimators of precision are also provided.

The woodland partial sampling frame, that is those woodland blocks having map area greater than or equal to 0.25 ha, leads to the set $\{x_i\}_{i=1\ldots N}$ termed {Diglist} where x_i is the digitised map area of the i^{th} block from amongst the N units in the frame. The frame is stratified according to the x–value into classes C_h ($h=1 \ldots H$) defined by

$$\text{unit } 'i' \text{ is in } C_h \text{ if } b_{h-1} \leqslant x_i \leqslant b_h$$

There are therefore H strata defined by $\{b_h\}_{h=0,H}$ where we set $b_0=0.25$ ha and $b_H=\infty$. The values $\{b_h\}$ were chosen at the sample size determination stage described in Chapter 3 and the final estimators must make use of these same stratification boundaries. The number of blocks in C_h is denoted by N_h and the number of samples taken from this stratum by n_h.

The sampled blocks have their actual areas determined and we denote the set of actual areas by $\{y_i\}_{i=1,n}$ where it is assumed that the ordering of x_i is such that x_i and y_i ($i \leqslant n=\sum_h n_h$) refer to the same block. Denote that subset of the frame within which units have an *actual* area between b_{h-1} and b_h by C_h'. The notation used is illustrated in Figure 5 in which the n sample units are indicated as points.

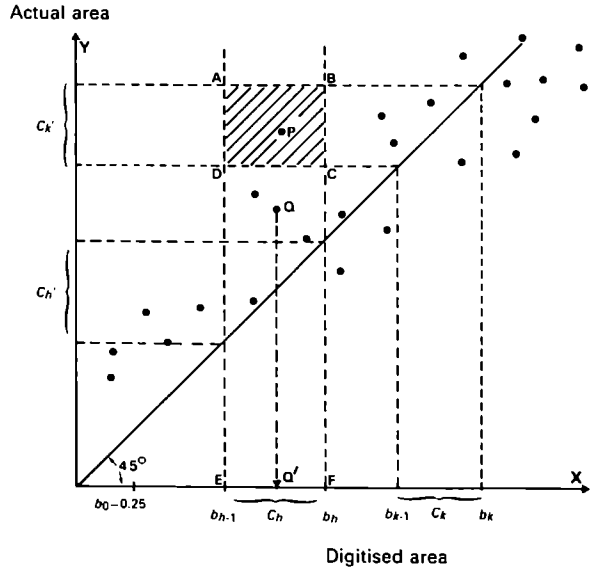


Figure 5. Illustration of notation for woodland area estimation.

The estimation of the number of woods and their total area was carried out using two different sets of estimators. The first set is based upon the classical set of expansion estimators for a stratified survey whilst the second set is based upon the use of the regression-prediction methodology. The 'expansion' approach will give generally unbiased estimators of the number and total area of woodlands in either the classes $\{C_h\}$ or $\{C_h'\}$. Estimation for the actual size classes $\{C_h'\}$ may be significantly different from that for the $\{C_h'\}$, depending upon the proportions of blocks in $\{C_h\}$ which are in $\{C_k'\}$ ($k \neq h$).

However, the regression-predictor will in general produce more precise area estimators, since it extracts more structure from the data, although the approach might involve some small and unquantified bias, depending on the adequacy of the regression model fitted. Should the data indicate that a simple regression model is inappropriate, due to an excessive number of outliers for example, then the expansion estimators provide a safe and robust alternative.

The expansion estimators

Let the number of samples with 'x' in C_h and 'y' in C_k' be denoted by m_{kh}. In Figure 5 m_{kh} is indicated by the number of sample points falling into the rectangle ABCD. Then the number of samples in C_h which really are woods is given by

$$m_h = \sum_{k=1}^{H} m_{kh} \tag{6.1}$$

First we give estimators of the numbers of blocks in the classes $\{C_h\}$. The estimated proportion of blocks in C_h which really are woods is

$$\hat{p}_h = \frac{m_h}{n_h} \tag{6.2}$$

with an estimated variance

$$\text{var}\,(\hat{p}_h) = \frac{\hat{p}_h\,(1-\hat{p}_h)}{n_h}$$

Hence the estimated number of blocks in C_h which really are woods is

$$\hat{N}_{.h} = N_h \hat{p}_h \quad \text{with} \quad \text{var}\,(\hat{N}_{.h}) = N_h^2\,\text{var}\,(\hat{p}_h) \tag{6.3}$$

yielding an expected total number of actual woods

$$\hat{N}\,.. = \sum_h \hat{N}_{.h} \quad \text{with} \quad \text{var}\,(N\,..) = \sum_h \text{var}\,(N_{.h}) \tag{6.4}$$

Secondly, we estimate the number of blocks in the actual size classes C_k'. The estimated proportion of blocks in C_h which are in C_k' is given by

$$\hat{p}_{kh} = \frac{m_{kh}}{n_h} \quad \text{with} \quad \text{var}\,(\hat{p}_{kh}) = \frac{\hat{p}_{kh}\,(1-\hat{p}_{kh})}{n_h} \tag{6.5}$$

Hence the estimated number of blocks (from the partial sampling frame) which are in C_k' is given by

$$\hat{N}_{k.} = \sum_{h=1}^{H} N_h\,\hat{p}_{kh}$$

$$\text{with} \quad \text{var}\,(\hat{N}_{k.}) \simeq \sum_{h=1}^{H} N_h\,\text{var}\,(\hat{p}_{kh}) \tag{6.6}$$

which will lead to the same total area estimator but with a different (larger) variance. The variance formula (6.6) is approximate because the correlations between the multinomial parameters $\{\hat{p}_{ij}\}$ are ignored.

We now estimate the area of woodlands in terms of the classes C_h. The mean actual area of blocks in C_h is estimated by

$$\hat{\bar{Y}}_h = \frac{1}{n_h} \sum_{C_h} y \tag{6.7}$$

so that the estimated total area in C_h, $\hat{Y}_j$, is $N_h\,\hat{\bar{Y}}_h$ with

$$\text{var}\,(\hat{Y}_h) = N_j \left[\frac{1}{n_h\,(n_h - 1)} \sum_{C_h} (y - \hat{\bar{Y}})^2 \right] \left(1 - \frac{n_h}{N_h} \right) \tag{6.8}$$

This leads to total area estimate

$$\hat{Y} = \sum_{h=1}^{H} \hat{Y}_h \text{ with var } (\hat{Y}) = \sum_{h=1}^{H} \text{var}(\hat{Y}_h) \tag{6.9}$$

On the other hand the mean actual block area of those blocks in C_k' which are also in C_h is given by

$$\overset{\triangle}{\overline{Y}}_{ij} = \frac{1}{n_h} \sum_{C_h \otimes C_k'} y \tag{6.10}$$

Hence $\hat{Y}_{kh}$ is given by $N_h \overset{\triangle}{\overline{Y}}_{kh}$ with

$$\text{var } (\hat{Y}_{kh}) = N_h^2 \left[\frac{1}{n_h (n_h - 1)} \sum_{C_h \otimes C_k'} y \right] \left(1 - \frac{n_h}{N_h} \right) \tag{6.11}$$

The estimated total area of blocks with actual area in C_k' is therefore given by

$$\hat{Y}_{k\cdot} = \sum_{h=1}^{H} Y_{kh} \text{ with var } (\hat{Y}_{k\cdot}) = \sum_{h=1}^{H} \text{var}(\hat{Y}_{kh}) \tag{6.12}$$

During the woodland survey all of these estimates were provided routinely, the precisions being expressed as a percentage coefficient of variation in order to aid comparisions with target precisions.

The regression predictor of total area

It can be seen from Figure 5 that the sample points illustrated do not fall upon a straight line at 45° through the origin. It was generally found, in practice, that the data could be represented well by a straight line but that for this line the slope and intercept, though close to 45° and 0 respectively, were significantly different. This difference reflects a wide range of causes. Blocks on the map are subject to a representation error which increases the apparent areas of small blocks in order to allow them to feature on the map. Also the blocks will possibly have been increased or decreased in size between the times of the map making and the survey.

The approach adopted is basically as follows. Ignoring the sample blocks which actually turn out not to be woodlands, the sample data for those remaining blocks are used to obtain a linear regression of y on x. This relationship is then used to predict the estimated actual area of all of the non-sampled blocks in the frame, taking into account the estimated proportions of woodlands in each size class of the frame which we expect not to turn out to be existing woodlands (i.e. actual area ≤ 0.25 ha). The results are then combined with the observed actual areas of the sample to give final area estimators. Variance formulae follow from the theory given in Appendix 5B. Technical details are given below and in the appendices.

The regression relation actually fitted was

$$\begin{aligned} y_i = a + bx_i + \epsilon_i \quad \text{where } & E(\epsilon_i) = 0 \\ & \text{var } (\epsilon_i) = \sigma^2 x_i^{\delta} \\ & \text{and cov } (\epsilon_i, \epsilon_j) = 0, (i \neq j) \end{aligned} \tag{6.13}$$

where an inhomogeneous model has been adopted to reflect observed data structure. See Figure 4 for an illustration of real data. It is assumed that ϵ_i is normally distributed and values for parameters a, b, σ^2 and δ are selected which maximise the likelihood of the observed data having occurred. The likelihood function is given in Appendix 6A. A simplifying feature of this method is that even though there are four parameters to the regression model, parameters a, b and σ^2 may be evaluated explicitly from the data and a given value of δ. In the section on aerial photo sample size determination in Chapter 5, the form of regression estimator used for sample size determination assumed $\delta = 0$, i.e. an homogeneous error structure. The value $\delta = 0$ (with $a_o = 0.0$, $b_o = 1$, $\sigma_o^2 = 80$) was used as an initial value in an iterative NAG optimisation procedure with respect to δ in order to find the maximum likelihood estimates of a, b, σ^2 and δ.

The variance formula associated with using such a non-homogeneous variance model for prediction purposes may be found in Appendix 6B. The homogeneous variance formulae used at the sample size determination stage, i.e. (5.15) and (5B20), are no longer valid and the derivations follow from the general result (5B11).

The total area estimate may be obtained in a number of ways; by using the fitted regression model on all x–values (for which y is not known) separately, or on the strata means or on the overall mean x–value. These methods all give the same estimated value since the regression is linear but may yield estimates with different standard errors. We have chosen to use the regression on the mean x–value of all further woods. However, to estimate this mean x–value in the frame it is necessary to take into account the estimated proportions of blocks in each stratum which are not actually woodlands. The value m of Appendix 6A has in practice to be estimated and we will therefore get some additional terms to the final variance estimators. These are given in Appendix 6C.

Some consideration was given as to how the regression might be used to obtain improved estimates of the total area in actual size classes. This required the calculation of the x–value which would yield a given y–value, the classical 'calibration' problem. Estimated y–values corresponding to x–values within 'backcalculated' strata could then have been obtained. However, for such an approach to be reasonably accurate conditions on the distribution of x–values and the x–y scatter would need to be satisfied. Since these could not be guaranteed in all counties it was decided to rely primarily upon the expansion estimation method to obtain definitive results on the distribution of the total area between the different size classes.

Second Phase Estimation of the Total Area of Woodland of a Particular Type

Let z_{hkt} be the area of type t occurring in the k^{th} block of stratum h, this stratum being defined in terms of the digitised area of the block. Then the mean area of type t occurring in stratum h is

$$\bar{z}_{ht} = \frac{1}{n_h} \sum z_{hkt}$$

where n_h is the number of samples in the h^{th} stratum which really are woods and which satisfy the necessary condition

$$y_{hk} = \sum_{t} z_{hkt}$$

Then the estimated total area of type 't' is, by expansion, given by

$$\hat{Z}_t = \sum_{h=1}^{H} N_h \bar{z}_{ht}$$

where we have decided to ignore the effect of non-woodlands being present in the frame. N_h could be replaced by $\hat{N}_h$ as given in previous sections with little impact upon the resulting estimates and their precisions.

If we sum over all types we obtain a total area estimate

$$Z_{\cdot} = \Sigma \hat{Z}_t$$

which will not be as precise as the total area estimate obtained by regression, $\hat{Y}_{\text{TOT}}$ say, but can be used to estimate the proportion of the total area which is of type t.

$$\hat{R}_t = \frac{\hat{Z}_t}{\hat{Z}_{\cdot}}$$

This is the combined ratio estimate from a stratified sample. Cochran (1963), pp.169–170, gives the approximate variance of this estimate which we have approximated further by assuming that

$$Z_{ht} = R_t \, Y_{ht}$$

for each stratum. The resulting variance estimator that has been used for $\hat{R}_t$ is

$$\text{var}\,(\hat{R}_t) \simeq \frac{1}{\hat{Y}_{\text{TOT}}^2} \sum \left\{ \frac{N_h^2\,(1-n_h/N_h)}{n_h\,(n_h-1)} \left[\sum_k (z_{hkt} - \hat{R}_t\, y_{hk})^2 \right] \right\}$$

where we have again ignored the fact that N_h should in fact be estimated.

Our final estimate of the total area of type t is therefore

$$\hat{Z}_t = \hat{R}_t \, \hat{Y}_{\text{TOT}}$$

with

$$\text{var}\,(\hat{Z}_t) = \hat{R}_t^2 \,\text{var}\,(\hat{Y}_{\text{TOT}}) + (Y_{\text{TOT}})^2 \,\text{var}\,(\hat{R}_t)$$

the values of $\hat{Y}_{\text{TOT}}$ and var $(\hat{Y}_{\text{TOT}})$ being available from first phase estimation and the correlation between the estimators of total area at the first phase and $\hat{R}_t$ at the second stage having been taken to equal zero.

Though the above formulae were used for final estimation of woodland types they were also used at an early stage of the survey in a monitoring role. At intermediate stages of the second phase survey, the data were fed into the program corresponding to the equations of this section in order to enable an early comparison between the target precisions used for sample size determination and the precision actually obtained. This procedure allowed a valid sequential increase in sample size in order to obtain satisfactorily precise results. This was of particular importance in the survey in view of the sparseness of pilot data on woodland types and the consequent heavy dependence on hypothesised models of variability.

Appendix 6A

Maximum Likelihood Estimation of the Regression Model

The data set consists of those woodland blocks in the sample with actual area not less than 0.25 ha in area. Suppose there are 'm' such blocks. The regression model is

$$y_i = a + bx_i + \epsilon_i \quad \text{where } \epsilon_i \sim N(o, \sigma^2 x_i^{\delta}) \text{ and cov } (\epsilon_i, \epsilon_j) = 0, (i \neq j) \tag{6A1}$$

The log-likelihood of the data is

$$\mathcal{L} = m.\ln\sqrt{2\pi} - \sum_{i=1}^{m} \left\{ \ln\sigma^2 + \delta.\ln x_i + \frac{[y_i - (a + bx_i)]^2}{\sigma^2 x_i^{\delta}} \right\} \tag{6A2}$$

Maximisation of $\mathcal{L}$ with respect to a and b leads to the estimators

$$\hat{b} = \frac{\varphi_1 \theta_0 - \varphi_2 \theta_1}{\theta_2 \theta_0 - \theta_1^2} \equiv \frac{\varphi_1 \theta_0 - \varphi_2 \theta_1}{\phi} \tag{6A3}$$

and

$$\hat{a} = \frac{\varphi_2 - \hat{b}\theta_1}{\theta_0} \tag{6A4}$$

where

$$\varphi_1 = \sum_{i=1}^{m} \frac{y_i}{x_i^{\delta-1}},$$

$$\varphi_2 = \sum_{i=1}^{m} \frac{y_i}{x_i^{\delta}},$$

and

$$\theta_l = \sum_{i=1}^{m} \frac{1}{x_i^{\delta-l}}$$

Both of these are in terms of δ. We may similarly estimate σ^2 by

$$\hat{\sigma}^2 = \frac{1}{(m-3)} \sum_{i=1}^{m} \frac{[y_i - (\hat{a} + \hat{b}x_i)]^2}{x_i^{\delta}} \tag{6A5}$$

also in terms of the data and δ. If $\delta=0$ then these estimators reduce to the standard homogeneous regression estimators.

Substitution of (6A3) and (6A5) into (6A2) gives the likelihood of the data as a function of δ alone. This was maximised by using a numerical method starting from $\delta_0=0$ (NAG, 1982).

Appendix 6B

The Variance of the Prediction Estimator Using a Regression Model with Inhomogeneous Variance

The general result of (5B11) gives the prediction estimator's variance as

$$\text{var}(\hat{\bar{y}}_f) = \left[\bar{\boldsymbol{x}}_f' \, (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1} \bar{\boldsymbol{x}}_f + \frac{\sum_{i=1}^{m} v_{ii}}{m^2} \right] \tag{6B1}$$

where $\hat{\bar{y}}_f$ is the predicted mean y–value corresponding to the mean $\bar{\boldsymbol{x}}_f$ of m data values further to those upon which the regression was based. Other terms are defined as follows

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ . & \\ . & \\ . & \\ 1 & x_n \end{pmatrix}, \quad \boldsymbol{V} = \hat{\sigma}^2 \begin{pmatrix} x_1^{\delta} & & & \\ & x_2^{\delta} & & \\ & & . & \boldsymbol{0} \\ & & & . \\ & & & . \\ & & \boldsymbol{0} & x_n^{\delta} \end{pmatrix} \tag{6B2}$$

$\boldsymbol{x}'_{fi} = (1, x_{fi})$ where x_{fi} is the map area of the i^{th} further block and $\bar{\boldsymbol{x}}_f' = (1, x_f)$. $\hat{\sigma}^2$ is the estimated variance constant estimated from the n sample points as described in Appendix 6A. Substitution of (6B2) into (6B1) and some manipulation yields,

$$\text{var}(\hat{\bar{y}}_f) = \hat{\sigma}^2 \left[\frac{\theta_2 - 2\bar{x}_f \theta_1 + \bar{x}_f^2 \theta_0}{\phi} + \frac{\sum_{i=1}^{m} x_i^{\delta}}{m^2} \right] \tag{6B3}$$

where θ and ϕ are defined in Appendix 6A.

On substitution of $\delta=0$ into (6B3) we obtain the homogenous variance predictor formula (5B20), providing a check on the validity of (6B3).

Appendix 6C

Modifications to the Variance Predictor to Take Account of Presence of Non-woodlands in the Frame

If there are $(N_f)_h$ 'further' blocks in the h^{th} stratum then we estimate the number of these which are actually woodlands by

$$\hat{N}_f = \sum_h \hat{p}_j (N_f)_h \tag{6C1}$$

Note that 'further' here means that we are referring to blocks in the frame which were not sampled. We also assume that the total x–value associated with the h^{th} stratum is $p_h(X_f)_h$ so that the estimated 'further' area is given by

$$\hat{X}_f = \sum_h p_h (X_f)_h \tag{6C2}$$

We accept that this is not completely valid when $p_j \neq 1$, since area 'x' is the main variable affecting 'p'. However, it will be an adequate approximation since p is usually equal to unity for $j > 1$.

If we adopt

$$\hat{\bar{x}}_f = \frac{\hat{X}_f}{\hat{N}_f} \tag{6C3}$$

as our estimate of the mean future digitised area of actual woods then the total area estimate is given by

$$\hat{Y}_f = (\hat{N}_f, \hat{X}_f)\,\hat{\beta} \tag{6C4}$$

where $\hat{\beta}$ is the estimated parameter vector.

Hence we approximate the variance of $\hat{Y}_f$ by

$$\text{var}(\hat{Y}_f) = \text{var}\,[(\hat{N}_f, \hat{X}_f)\,\hat{\beta}] + \left\{ \sum_h p_j \sum_i^{(N_f)_h} (v_{fi})_h \right\} \tag{6C5}$$

The second term corresponds to the second term of 6B1, but our first term expands to,

$$\text{var}\,[(N_f, X_f)\hat{\beta}] = (\hat{N}_f, X_f)\,(\text{var}\,\hat{\beta}) \begin{pmatrix} \hat{N}_f \\ \hat{X}_f \end{pmatrix} + \hat{\beta}' \begin{pmatrix} \text{var}\,\hat{N}_f, & \text{cov}\,(\hat{N}_f, \hat{X}_f) \\ \text{cov}\,(\hat{N}_f, \hat{X}_f), & \text{var}(\hat{X}_f) \end{pmatrix} \hat{\beta} \tag{6C6}$$

since $\hat{\beta}$ is independent of estimates of $\hat{N}_f$ and $\hat{X}_f$.

Our final result therefore becomes,

$$\text{var}\,(\hat{Y}_f) = \hat{\sigma}^2 \hat{N}_f^2 \left[\frac{\theta_2 - 2x_f\,\theta_1 + \bar{x}_f^2\,\theta_0}{\phi} + \frac{\sum_j p_j \left(\sum_i^{(N_f)_j} x_{ij} \hat{\delta} \right)}{\hat{N}_f^2} \right] + \hat{a}^2 C_{11} + \hat{b}^2 C_{22} + 2\hat{a}\hat{b} C_{12} \tag{6C7}$$

where

$$\begin{aligned} C_{11} &= \text{var}\,(\hat{N}_f) &&= \sum_j (N_f)_j^2\, p_j\,(1-p_j)/n_j \\ C_{22} &= \text{var}\,(\hat{X}_f) &&= \sum_j (X_f)_j^2\, p_j\,(1-p_j)/n_j \\ C_{12} &= \text{cov}\,(\hat{N}_f, \hat{X}_f) &&= \sum_j (N_f)_j (X_f)_j p_j (1-p_j)/n_j, \end{aligned} \tag{6C8}$$

where $\hat{\sigma}^2$, θ_i, ϕ, $\hat{\delta}$, $\hat{a}$ and $\hat{b}$ are as previously defined.

The estimated value of the variance of our total area estimate is identically given by var $(\hat{Y}_f)$ since the known sample values have zero variance.

CHAPTER 7

Mathematical details of the non-woodland survey

Chapter 4 presented the design structure and described the method in which sample sizes were determined. The details of how this was done are set out in Appendix 5A on constrained optimal allocation and the use of such methods in the woodland survey. We therefore do not further explicitly consider the sample size determination problem for the non-woodland survey but proceed directly to the details of the Mode 1, 2 and 3 estimators.

Suppose that there are M secondary units in each primary unit and that in the h^{th} stratum there are N_h primaries, $(h=1 \ldots H)$. Suppose ν_h primaries are selected for aerial measurement of the variate 'X' on each of the component secondaries yielding data $\{x_{hji}\}$; $h=1 \ldots H$, $j=1 \ldots \nu_h$, $i=1 \ldots M$. Of these ν_h phase–1 aerial samples n_h are selected for phase–2 ground visit and without loss of generality we assume that these are the first n_h from the ν_h aerial samples. On each such ground assessed cluster we measure the variate 'y' to yield ground data $\{y_{hji}\}$; $h=1 \ldots H$, $j=1 \ldots n_h$, $i=1 \ldots m$ where m secondaries per primary are assessed. Again, without loss of generality, we suppose a numbering of secondaries in primaries, which ensures that the assessed secondaries are the first m from the M possible secondaries.

MODE 1 Estimation

This would be appropriate when making an estimate of a quantity which is closely related to the 'x' variate (or *is* the x–variate) but which is not necessarily related to 'y'.

Define the sampled cluster totals by

$$z_{hj} = \sum_{i=1}^{M} x_{hji} \tag{7.1}$$

and hence the mean total z–value per cluster in the h^{th} stratum by

$$\hat{z}_h = \frac{1}{\nu_h} \sum_{j=1}^{\nu_h} z_{hj} \tag{7.2}$$

then the expansion population estimator for the total of the x–variate is

$$\hat{X} = \sum_{h=1}^{H} N_h \bar{z}_h \tag{7.3}$$

$$\text{and var } (\hat{X}) = \sum_{h=1}^{H} N_h^2 \,\text{var}\,(\hat{z}_h) \tag{7.4}$$

$$= \sum_{h=1}^{H} N_h^2 \left(1 - \frac{\nu_h}{N_h}\right) \frac{s_{zh}^2}{\nu_h}$$

$$\text{where } s_{zh}^2 = \frac{1}{(\nu_h - 1)} \sum_{j=1}^{\nu_h} (z_{hj} - \bar{z}_h)^2$$

since we have a stratified random sample of clusters.

MODE 2 Estimation

This method of estimation is of particular relevance when y is the feature of interest and there is little relationship with the corresponding x–values. It makes use only of the second stage data which is collected from a stratified two-stage design. The methods of estimation are standard and may be found in general terms in texts of sampling theory such as Sukhatme (1954), Cochran (1963), Des Raj (1968) and Kish and Frankel (1974). However, we present the explicit formulae in the form they have been used in this survey.

We adopt the following definitions;

$$y_{hj} = \sum_{i=1}^{m} Y_{hji} \tag{7.5}$$

$$\bar{y}_{hj\cdot} = \frac{1}{m} \sum_{i=1}^{m} y_{hji} \tag{7.6}$$

$$y_{h\,\cdot\cdot} = \sum_{j=1}^{n} \sum_{i=1}^{m} y_{hji} \tag{7.7}$$

$$\bar{y}_{h\,\cdot\cdot} = \frac{1}{n_h m} \sum_{j=1}^{n_h} \sum_{i=1}^{m} y_{hji} \tag{7.8}$$

$\bar{y}_{h\,\cdot\cdot}$ gives an unbiased estimate of the population mean in stratum h. Hence an unbiased estimate of the population total in stratum h is

$$\hat{Y}_h = MN_h\bar{y}_{h\,\cdot\cdot} \tag{7.9}$$

with a variance given by

$$\text{var}\,(\hat{Y}) = \sum_{h=1}^{H} M^2 N_h^2\,\text{var}\,(\bar{y}_{h\,\cdot\cdot}) \tag{7.10}$$

where

$$\text{var}\,(\bar{y}_{h\,\cdot\cdot}) = \left(\frac{1}{n_h} - \frac{1}{N_h}\right)s_{bh}^2 + \frac{1}{N_h}\left(\frac{1}{m} - \frac{1}{M}\right)s_{wh}^2 \tag{7.11}$$

and s_{bh}^2 and s_{wh}^2 are the sample estimates of the between and within cluster variances for stratum h. These are given by

$$s_{bh}^2 = \frac{\sum_{j=1}^{n_h} (y_{hj\cdot} - y_{h\,\cdot\cdot})^2}{(n_h - 1)} \tag{7.12}$$

$$\text{and } s_{wh}^2 = \frac{1}{n_h(m-1)} \sum_{j=1}^{n_h} \sum_{i=1}^{m} (y_{hji} - \bar{y}_{hj\cdot})^2 \tag{7.13}$$

These formulae reduce to MODE 1 estimators when m is set equal to M.

MODE 3 Estimation

This mode of estimation makes full use of the data at both phases by basing estimation upon an observed close relationship between the x and y values on secondaries. Figure 6 shows some data illustrating the relationship between the number of trees per secondary estimated from a photograph and the actual number observed on a ground visit. As expected there is a reasonably close relationship and it will be recalled that the sample size determination, being based upon MODE 1 estimation, tacitly assumes that x and y are identically equal. Pairs of points are shown joined in Figure 6 indicating that the two points are from the same primary unit. Clearly there are a number of essentially different possible patterns which this data can take. The extremes are illustrated in Figure 7.

Within cluster regression coefficient = 0.91
Between cluster regression coefficient = 0.70
Overall regression coefficient = 0.895

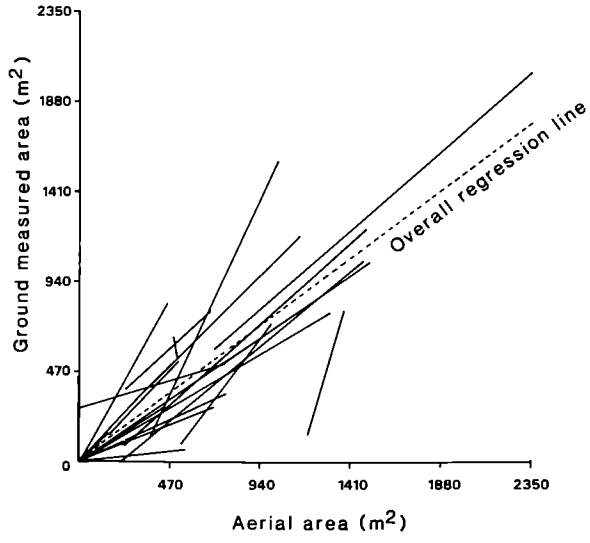


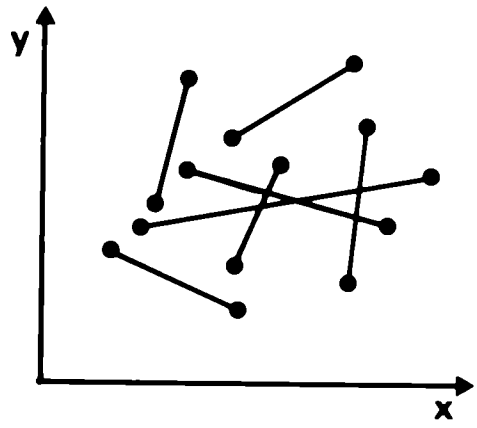
Figure 6. Example of relationship between x and y data.

Since there are $n_h' = (\nu_h - n_h)$ primaries in the h^{th} stratum for which the x–variates have been measured, but not the y–variate, we would expect that in certain cases the extra x–data might be used to improve the estimates which could be obtained by using MODE 2 on the x–data alone. If the relationship is as shown in Figure 7 (a) then it is clear that there is no relationship between the x and y variates and MODE 2 is the best estimator.

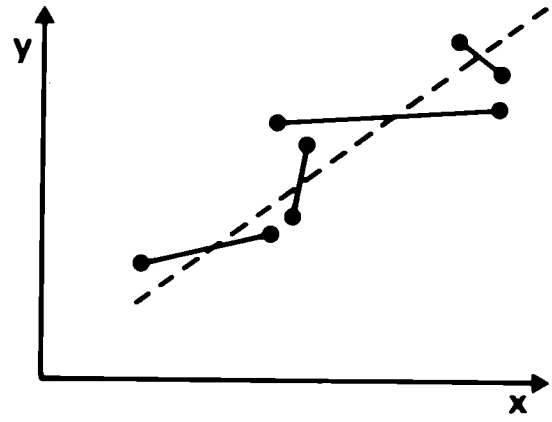
However, if the data display the pattern shown in Figure 7 (b) then there is a clear between-cluster x–y relationship whilst no within cluster relationship is apparent. The mean x–value on the n_h' clusters may therefore be used to improve the precision of estimation of the population value of y.

Figure 7 (c) on the other hand illustrates a strong within-cluster relationship but no between cluster relationship. In such a case precise x–estimates are possible on those secondaries which are in ground-assessed clusters, but which are not themselves assessed. However, the n'_k primaries on which only x is assessed provide no aid to the estimation. Figure (d) illustrates a perfect linear relationship between x and y and in such a case all of the secondaries on which x is assessed may be used to improve the precision of the estimates. It is necessary to formulate an estimator, termed MODE 3, which will make use of the 'within' and 'between' relationships but in the case of no relationship between x and y (Figure 7(a)) will reduce to MODE 2, whilst in the case of a perfect relationship will be essentially equivalent to MODE 1 estimation. We develop such an estimator based upon the assumption that a linear relationship of y on x will be an adequate representation in most cases. With such an assumption implicit in MODE 3 estimation it can, in the cases when the assumption is unjustified, lead to a biased estimate. We therefore did not use MODE 3 in all cases but rather made a judgement in each case as to whether the relationship was sufficiently strong and linear for MODE 3 to be used. If this was not the case then either MODE 1 or MODE 2 was used, as appropriate, these estimators being unbiased in all circumstances.

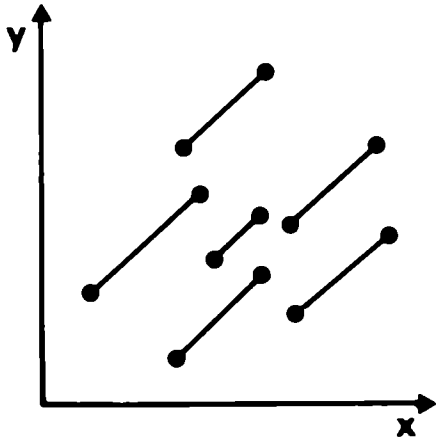
As discussed in Appendix 5B there are two distinct interpretive approaches to survey estimation. The first method is the model based predictive approach in which the sample data, such as that shown in Figure 6, is used to estimate a relationship and this estimated relationship is then used to predict the population total. The other approach is via the classical randomisation approach of hypothetical repeated sampling. It has been shown (Appendix 5B and Rennolls,



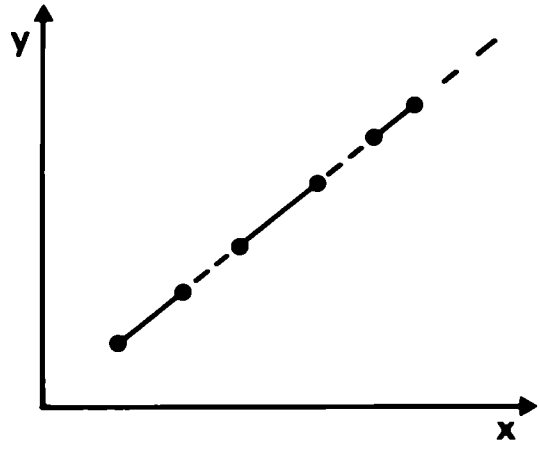
(a) No relationship.



(b) Good between primary relationship, no within primary relationship.



(c) Good within primary relationship, no between primary relationship.



(d) Perfect linear relationship.

Figure 7. Illustration of the possible extreme relationships between the data from the two phases of the non-woodland survey.

1981) that the resultant formulations are equivalent for a stratified random sample and are almost identical for a regression estimator from a simple random sample. Both approaches have been followed through but the model-based-predictive approach met some difficulties at the programming stage and hence is not presented further here. We therefore restrict our presentation to a set of classically based estimators, which were actually used in practice.

The symbolism for the estimators is specified in the context of *three* phases of sampling. First the selection of ν_h primaries, secondly the sub-selection of n_h primaries followed by the third phase of subsampling within primaries. The variance of our estimators will therefore have three terms corresponding to these randomisation phases. The estimator used to estimate the mean y–value per secondary was

$$\hat{\bar{Y}}_h \,..\, = \bar{y}_h \,..\, + b_{bh}\,(\bar{x}'_h \,..\, - \bar{x}'_h) + b_{wh}\,(\bar{x}'_h \,..\, - \bar{x}_h \,..) \qquad \textbf{(7.14)}$$

where b_{bh} and b_{wh} are the estimated between-cluster and within-cluster regression coefficients for the observed x–y data. $\bar{x}'_h{}'$.. is the mean x–value per h^{th} stratum secondary calculated from the ν_h first phase sample primaries.
$\bar{x}'_h$ .. is similar but calculated on the n_h sampled primaries. $\bar{x}_h$ .. is the mean x–value per h^{th} stratum secondary calculated from the $n_h m$ ground visited secondaries. Similarly for $\bar{y}_h$ ... The least squares estimates for b_{bh} and b_{wh} are given by

$$b_{bh} = \frac{s_{bxyh}}{s_{bxh}{}^2} \text{ and } b_{wh} = \frac{s_{wxyh}}{s_{wxh}{}^2} \tag{7.15}$$

where $s_{bxh}{}^2$ and $s_{wxh}{}^2$ are the sample variances given by

$$\left.\begin{aligned} s_{bxh}{}^2 &= \frac{1}{(n_h-1)} \sum_{j}^{n_h} (\bar{x}_{hj.} - \bar{x}_{h..})^2 \\ \text{and} \\ s_{wzh}{}^2 &= \frac{1}{(n_h-1)} \sum_{j=1}^{n_h} \frac{1}{(m-1)} \sum_{i=1}^{m} (y_{hji} - \bar{y}_{hj.})^2 \end{aligned}\right\} \tag{7.16}$$

The sample covariances s_{bxyh} and s_{wxyh} are given by

$$s_{bxyh} = \frac{1}{(n_h-1)} \sum_{j=1}^{n_h} (\bar{y}_{hj.} - \bar{y}_{h..})(\bar{x}_{hj.} - \bar{x}_{h..}) \tag{7.17}$$

and

$$s_{wxyh} = \frac{1}{n_h\,(m\,-\,1)} \sum_{j=1}^{n_h} \sum_{i=1}^{m} (y_{hji} - \bar{y}_{hj.})(x_{hji} - \bar{x}_{hj.}) \tag{7.18}$$

The estimated variance of the estimator given in (7.14) is obtained by taking expectations over the three phases of randomisation to yield,

$$\begin{aligned} \text{var}\,(\overset{\triangle}{Y}_h\,..) = {} & \frac{1}{\nu_h}\left(1 - \frac{\nu_h}{N_h}\right) s_{byh}{}^2 \\ & + \frac{1}{n_h}\left(1 - \frac{n_h}{\nu_h}\right)\{s_{byh}{}^2 - 2b_{bh}\, s_{bxyh} + b_{bh}{}^2\, s_{bxh}{}^2\} \\ & + \frac{1}{n_h m}\left(1 - \frac{m}{M}\right)\{s_{wyh}{}^2 - 2b_{wh}\, s_{wxyh} + b_{wh}{}^2\, s_{wxh}{}^2\} \end{aligned} \tag{7.19}$$

Under suitable conditions on the within and between cluster correlations this can be shown to reduce the variance estimators for MODE 1 and MODE 2.

References

CASSEL, C.M., SARNDAL, C.E. and WRETMAN, J.A. (1977). *Foundations of inference is survey sampling*. John Wiley, New York.

COCHRAN, W.G. (1963). *Sampling techniques*. John Wiley, New York.

COX, D.R. and HINKLEY, D.V. (1974). *Theoretical statistics*. Chapman and Hall, London.

HOLT, D., SMITH, T.M.F. and WINTER, P.D. (1980). Regression analysis of data from complex surveys. *Journal of the Royal Statistical Society* A, **143**(4), 474–487.

HUGHES, E. and RAO, J.N.K. (1979). Some problems of optimal allocation in sample surveys involving inequality constraints. *Communications in Statistics Theory and Methods* A8(15), 1551–1574.

KALTON, G. (1976). Contribution to the discussion of Smith (1976).

KISH, L. (1965). *Survey sampling*. John Wiley, New York.

KISH, L. and FRANKEL, M.R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society* B, **36**, 1–37.

LOCKE, G.M.L. (1987). *Census of woodlands and trees 1979–82*. Forestry Commission Bulletin 63. HMSO, London.

NAG (1982). *Fortran libraries*. Numerical Algorithms Group, Mayfield House, 256, Banbury Road, Oxford.

RAJ, D. (1968). *Sampling theory*. McGraw Hill, New York.

RENNOLLS, K. (1981). The use of superpopulation – prediction methods in survey analysis, with applications to the British national census of woodlands and trees. In, *Place resource inventories; principles and practice*. Proceedings of a national workshop, Orono, Maine. Society of American Foresters.

ROYALL, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika* **57**(2), 377–387.

ROYALL, R.M. (1976). The linear least-squares prediction approach to two-stage sampling. *Journal of the American Statistical Association* **71**(355), 657–664.

SEARLE, S.R.R. (1971). *Linear models*. John Wiley, New York.

SMITH, T.M.F. (1976). The foundations of survey sampling: a review. *Journal of the Royal Statistical Society* A, 1389(2), 183–204.

SUKHATME, P.V. (1954). *Sampling theory of surveys and applications*. Bangalore Press.

Printed in the United Kingdom for Her Majesty's Stationery Office
Dd 8998052 C8 2/89 3983/2 12521