



OPERATIONAL SELECTION POLICY OSP0

THE SELECTION OF CASE FILES: SAMPLING TECHNIQUES

Revised October 2005

1 Authority

The National Archives' Acquisition and Disposition Policies announced the Archives' intention of developing Operational Selection Policies across government. These would apply the collection themes described in the overall policy to the records of individual departments and agencies. Although this guidance is not an Operational Selection Policy *per se*, it is issued as part of the Operational Selection Policy programme. It should be used by government departments in conjunction with the Acquisition and Disposition Policies and relevant Operational Selection Policies when selecting case records for permanent preservation.

This guidance is a working tool for those involved in the selection of public records and may, therefore, be reviewed and revised in the light of comments received from the users of the records, from archive professionals, or from government departments. There is no formal cycle of review but we would welcome comments at any time. The extent of any review and revision exercise will be determined according to the nature of the comments received.

If you have any comments on this policy, please e-mail records-management@nationalarchives.gov.uk or write to:

Acquisition and Disposition Policy Manager
Records Management Department
The National Archives
Kew
Richmond
Surrey TW9 4DU

2 Scope

Operational Selection Policies apply to both case and policy records. In some instances it may be appropriate to select some but not all case files for permanent preservation. This document provides guidance on how such a selection is to be made. It explains when a set of records is suitable for sampling, what type of sampling is most appropriate and how to create a sample.

This guidance applies to all case files, whether electronic or paper based. However, with electronic case files, since the major costs of long-term preservation are not storage, it will often be feasible to select the whole series. This contrasts with paper where cost per metre tends to increase slowly year-on-year.

The guidance is especially applicable to particular instance papers. These files relate to a single event or transaction usually concerning one individual or organisation. They are characterised by a high degree of subject or document similarity and a low degree of variation between files. For example, a series of benefit application forms would be particular instance papers.

3 Suitability for Sampling

The types of file series suitable for sampling are case files where the information contained is mostly routine and voluminous in nature. The greater the variety in the nature of the files' contents, the less suitable they are for sampling. To some extent this can be addressed by varying the type of sample taken (eg a stratified random sample instead of a pure random sample), or by taking a larger sample size, but highly varied case files may not be suitable for sampling.

The structure and size of a series might also affect its suitability for sampling by making it impractical to take a meaningful sample of the series. This will dictate the appropriate methods to use; in some cases non-random sampling methods may be the only practicable techniques.

4 Principles of selection for case files

The National Archives' principles of selection are set out in the Acquisition and Disposition Policies. Case files present a particular challenge for appraisal because they create significant storage pressures and yet may be a major source for records which document 'the economic, social and demographic condition of the UK' or the 'impact of the state on the physical environment' (*Acquisition Policy 2.2.2*). They may play an important role in creating an archive which supports government policies to overcome social exclusion and which implements The National Archives' Diversity Strategy. The National Archives is developing principles for the selection of case files to ensure the relevant and balanced acquisition and disposition of social, economic and scientific data.

The National Archives believes that its collection provides an avenue for everyone to develop an interest in history and hence a broader understanding of their personal and cultural identity and of the government decisions which shape their lives. To help promote social inclusion, and as a partner in A2A: Access to Archives, The National Archives' selection of case files will consider the need for its collection to:

- Help to promote interest in the history of the individual, of a community or of British society as whole
- Be of direct value to educational organisations
- Reflect the social diversity of Britain

In line with the Acquisition Policy The National Archives will make the cost of selection and storage of such case files explicit in its decisions (*Acquisition Policy 2.3.4*). The National Archives will consider the selection of electronic datasets in the same way, and will actively seek out appropriate additions to its collections at the National Digital Archive of Datasets (NDAD). See Operational Selection Policy OSP 30 *The State and the Citizen* for guidance on the selection of electronic datasets containing information relating to individuals.

5 Terminology

The following terms have particular meanings when applied to sampling procedures which may differ from normal usage.

Bias: occurs when a method of selection yields a proportionally unrepresentative sample of the whole file series.

Cluster: a group of files within a series whose contents have moderate homogeneity but which are similar to other groups within the series. Groups are often formed for administrative convenience. For example, consider an investigation into the use of pesticides by farmers in England: the different counties are identified as clusters. A sample of these counties (clusters) would then be selected so all farmers in those chosen counties are included. It is more efficient in terms of resources to visit several farmers in the same county than it is to travel to each farm in a random sample.

Homogeneous: of uniform structure or composition throughout.

Heterogeneous: consisting of dissimilar or diverse files or constituents.

Random: a method of selection where every file has some positive probability of being selected.

Strata: a homogenous group of files within a heterogeneous file series. For example, if we consider taking a sample of hospitals, these can be stratified into general, maternity, research, psychiatric, etc. Samples can be drawn from each strata in the correct proportions if desired.

6 Sampling Methods

This section describes the main techniques of sampling and gives guidance as to when they are appropriate. The table below gives a summary of the most appropriate sampling methods for particular series structures. It is not exhaustive and combinations of different methods may be required to derive the most suitable sample.

Table of Sampling Methods		
Type of sample required	Structure of file series	Appropriate sampling techniques
Non-random	Homogenous, internal structure may be present	Convenience selection
		Exemplary selection
	Heterogeneous groups	Quota sampling ¹

¹ Whilst quota sampling is in theory a form of random sampling, in practice it is rarely random as the selection of files is subjective.

Random	Homogenous, no internal structure	Simple random sample Systematic sample
	Homogenous groups	Cluster sample
	Heterogeneous groups	Stratified random sample

6.1 Representation

If only a portion of a file series is to be selected, this should be representative of the series as a whole.

Before sampling it is important to know:

- The contents of the records
- The structure of the records
- The type of use to which the sample will be put, eg for longitudinal studies
- The way in which we wish it to be representative

This will determine the type of sample that is most appropriate for each situation.

If we wish to be able to reconstruct the main features of the data from the sample, a random sample must be used. If this is not the case - for example we may be interested in some predetermined characteristics, such as a particular place or a particular section of the population - or if the cost of taking a random sample is prohibitive, we can use a non-random method. However, these have no statistical validity and this will limit the uses to which the records can be put.

6.2 Partially Destroyed File Series

If a file series has already in part been destroyed, it may still be possible to take a meaningful sample. If the files have been destroyed randomly, then we can just treat the remaining series as usual. If criteria for their destruction are known and are not random, any sample taken will be biased. It will be representative of the remaining files but not the whole series. If the destruction criteria are not known, the remaining collection of files is already biased. In this case a non-random method should be used as it would be inefficient to devise a random sampling scheme which is invalid from the outset.

6.3 Accruing File Series

When a file series is still accruing it is difficult to determine how many files belong to the series or whether they are homogeneous or otherwise. In order to obtain a sample, the file series can be "artificially closed" every few years. However there is no guarantee that separate samples taken at different time periods will add up to one taken of the eventual whole series. Each sample

must be treated as representative of the time period in which it was drawn and not used to make inferences from the whole series.

6.4 Non-Random Methods

These should be used when the aim is to preserve particular characteristics only or when a random sample is not possible or practical.

6.4.1 Convenience Selection

Suitable for:	Homogeneous file series.
Example:	Particular instance papers such as benefit claims
Method:	Select the most convenient sample to hand, eg the first 20 files, the middle shelf etc.
Pros:	Simple to carry out.
Cons:	No statistical validity unless we know the files are in a random order.

6.4.2 Exemplary Selection

Suitable for:	File series where a particular grouping of files are thought to be representative of the whole series and there is homogeneity both within and between groups.
Example:	Particular instance papers such as benefit claims where a convenient grouping could be all the files dealt with by a particular person or office, or all files opened within a defined time period. The group would be chosen because it is seen to be representative of the series as a whole.
Method:	Select a grouping of case files.
Pros:	Simple to carry out. May appear less biased than convenience sampling.
Cons:	The common characteristic by which a group of files is selected must be chosen with care. (A Canadian study whereby all surnames starting with "F" were selected yielded the desired 3.5% sample but completely missed some ethnic groups). No statistical validity.

6.4.3 Exceptional Selection

Suitable for:	Mostly homogenous file series with a few interesting cases.
----------------------	---

Example:	Death duties where files concerning persons of note are deemed worthy of preservation.
Method:	Select the individual cases judged to have value using predetermined criteria such as precedent-setting cases or cases which attracted national press interest. The "fat" file selection procedure is of this type since the <i>fattest</i> files are selected because they relate to cases which generate more correspondence and which are, therefore, likely to be more interesting. Similarly files could be selected which have a high number of movements, as this shows they have been consulted frequently.
Pros:	Simple to carry out. Preserves interesting cases.
Cons:	No statistical validity as it over-represents the "interesting" cases compared to the routine work.

6.5 Random Sampling

The aim of sampling is to preserve the main features of the population so that inferences can be made about the characteristics of the whole population by examining a small proportion of it. To do this with statistical validity, it is essential that a random sample is taken in order to avoid bias.

6.5.1 Sample Sizes

An important issue is sample size - how large a sample we need to be able to reconstruct the characteristics of the whole file series with confidence. The statistical validity depends on the size of the sample as well as on its being randomly chosen. There is a trade-off between resources required in terms of effort and storage and precision/size. The greater the variability a file series displays, the greater the sample size required.

A percentage sample size should be avoided because it will often give too large or too small a number. The accepted maximum sample size is 1,400 files regardless of the size of the file series. An industry adopted standard, developed by Bell Laboratories, which takes into account the variability of file series is given below. This should be used to determine the size of your sample.

The method used to construct this table is given in J Carvalho, "Archival application of mathematical sampling techniques", *Records Management Quarterly* 18:63 (1984).

Table for Determining Sample Size
--

Population ¹	Sample Size ²		
	Low	Medium	High
51-90	5	13	20
91-150	8	20	32
151-280	13	32	50
281-500	20	50	80
501-1,200	32	80	125
1,201-3,200	50	125	200
3,201-10,000	80	200	315
10,001-35,000	125	315	500
35,001-150,000	200	500	800

6.5.2 Simple Random Sampling

Suitable for:	<ul style="list-style-type: none"> • This is the most appropriate technique to use if a file series is largely homogenous with no meaningful internal structure • With simple random sampling, every file has an equal chance of being selected • This method of selection is only really practicable if there is a convenient listing of records either on paper or on computer, otherwise the allocation of random numbers and identification of files can be extremely time consuming. Where a simple random sample is difficult to achieve, or we need to preserve the significance of the internal structure of the file series, there are several other ways in which to obtain pseudo-random samples which are adequate for most purpose. These are detailed below (5.5.3-5.5.5)
Example:	Particular instance papers such as tax claims
Method:	<ul style="list-style-type: none"> • In principle a simple random sample is straightforward to achieve

¹ Population denotes the total number of items to be sampled from.

² The sample size will depend on the homogeneity of the records. Low sample sizes are taken for files which are very similar in terms of content and subject matter while high sample sizes are needed for more diverse series.

	<ul style="list-style-type: none"> • First establish how many files are in the series • Decide the sample size required, n (see 5.5.1) • Generate n random numbers between 1 and the total number of files (see Annex for guidance on generating random numbers) • If a duplicate number is generated, generate another number • Select the files which correspond to the random numbers generated • For example, to obtain a sample of size n=100 from a file series of 1,000 files, generate 100 random numbers between 1 and 1,000. If the numbers generated are {71, 263, ...,106}, select the files with the corresponding positions in the list, i.e. the 71st, 263rd, ...106th files • This would be a practical method for computer and paper lists. Depending on the software used for storing the computer-based lists, it may be more practical to assign a random number to each file if this can be done automatically, and then select the n files with the smallest (or largest) associated random numbers
Pros:	The ideal form of sample from a statistical point of view. It is completely random: every file in the series has an equal chance of being picked.
Cons:	<ul style="list-style-type: none"> • Simple random sampling can be very hard to achieve in practice because all the files in the files series must be clearly identified and accessible with an equal chance of being selected • The allocation of random numbers can be time-consuming unless the records or a catalogue of the records is held on computer • A randomly selected sample might not include small, but important, sub-groups: a stratified random sample would be more appropriate here

6.5.3 Systematic Samples

Suitable for:	File series which are sufficiently large with no internal structure
Example:	Large volumes of particular instance papers such as claim forms
Method:	<ul style="list-style-type: none"> • Establish the number of files in the series. Decide the sample size (see 5.5.1)

	<ul style="list-style-type: none"> • Divide the total number of files by the required sample size to obtain a number, k • Select every kth file • The first file in the sample should be a randomly chosen number between 1 and k • For example: if we have a series of 10,000 files and desire a sample of 200, then we take every $10,000/200=50$th file. A random number between 1 and 50 is generated to determine the first file
Pros:	Easier to implement than simple random sampling.
Cons:	<ul style="list-style-type: none"> • Not statistically valid as the method is open to bias if files are systematically arranged • The sample might not include small, but important, subgroups

6.5.4 Stratified Random Sampling

Suitable for:	Records series which have a meaningful internal structure and fall into distinct groups (strata). Records within each strata are largely homogeneous, but the strata are heterogeneous. This method ensures that the representative sample has a proportional number of files from each strata.
Example:	A series of case records where the records could be classified according to land usage eg northern industrial region, southern industrial region, northern rural region, etc.
Method:	<ul style="list-style-type: none"> • Identify the strata to which each file belongs • Use either simple random or systematic sampling to draw a sample from each strata • It is usual to have proportionally representative samples from each strata to avoid bias in the resulting complete sample. However, some strata may be too small for this to be possible. In this case either select the whole of the strata concerned or amalgamate it with similar strata. The identity of the files concerned should be highlighted in the accompanying documentation so that subsequent analyses can be adjusted to allow for this bias in selection
Pros:	Guarantees representation of each strata, regardless of strata size. In some circumstances this means it produces a more meaningful sample than simple random sampling or systematic sampling. For example, if a series of 10,000 files includes a strata of 500 files, a 10% simple random sample has a probability of slightly less than 6% of actually including

	a 10% sample (50 files) of this particular strata. A stratified can ensure exact proportional representation. Statistically valid when carried out properly.
Cons:	Requires prior identification of all strata.

6.5.5 Cluster Sampling

Suitable for:	Record series that fall into convenient groupings(clusters). The records are broadly alike both between and within the clusters.
Example:	Benefit claims made to different offices. The different offices can be considered as clusters. Since the claims are broadly similar, it is more efficient to randomly choose a few offices and select all their records than to randomly select files regardless of the office.
Method:	Use simple random sampling to choose whole clusters. Every record within a selected cluster is preserved as part of the sample.
Pros:	Relatively simple and efficient to carry out. Statistically valid if carried out carefully.
Cons:	There must be little between-cluster variation; otherwise the sample becomes a convenience sample with no valid statistical properties.

6.5.6 Multi-Stage Sampling

Suitable for:	File series with a complex internal structure consisting of a combination of clusters and strata.
Example:	Where a cluster sample may give too many records, we can use simple random sampling to further reduce the files selected within the cluster. In the Benefits Office example in section 6.5.5, each office may still produce a large amount of records despite only a few offices being selected. Random sampling may be used to reduce the volume to a manageable number.
Method:	A combination of the above methods can be used to form samples. The structure of the records will dictate which are the most appropriate sampling techniques to use.
Pros:	May be more efficient because it makes use of the internal series structure.
Cons:	Can be complicated to carry out. Requires the identification of suitable groupings of files.

6.5.7 Quota Sampling

Suitable for:	File series organised into internally homogenous groups where the identity of the groups and their proportions within the whole series are known. The aim is to ensure proportional representation of these groups.
Example:	This is most often used in surveys such as opinion polls where the identification of the sampling units is crucial to obtain reliable results eg for political polls we need to ensure middle-class suburban housewives, or engineering students, say, have the correct proportional representation.
Method:	<ul style="list-style-type: none">• Identify the groupings we are especially interested in• Pre-determine the numbers to take from each group• Take a simple random sample from each group of interest
Pros:	This method can provide an efficient mechanism to ensure all groups are represented if used with care.
Cons:	<ul style="list-style-type: none">• It is not really random. Attempts to mimic stratified random sampling often fail because random selection is not used, often just the first few files are selected or some other such convenient selection. At worst it is a form of stratified convenience sampling• Results are very often unreliable. For example, many general election forecasts obtained from such sampling are inaccurate• This approach relies much more on subjective judgement than other approaches

7 Implementation

Before starting to sample records, government departments should consult their National Archives client manager for assistance in choosing the most appropriate form of sampling.

The method of sampling used must be documented and this information should be supplied to The National Archives when a new series is raised as part of the appraisal information. It should include:

- The size of the original file series
- The number of files selected
- The method used to select them eg random, exceptional etc
- Where stratified sampling has been used, the strata should be identified along with each associated sample size. Similarly for cluster or quota sampling

This information is important for researchers as it affects how they can use the records. If a sample is not random then any statistical analyses performed will be unreliable. Different types of random sample will affect the types of analyses used, for example, if a stratified random analysis has been performed and the strata used are not proportionally representative, then statistical analyses will have to incorporate this information to produce valid results.

ANNEX

Generating Random Numbers

For sampling purposes, random numbers must be uniformly distributed between certain limits such that any number is equally likely to be chosen. Spreadsheets and scientific calculators usually have a facility for generating random numbers between 0 and 1 which may need re-scaling.

Generating Random Numbers in Microsoft Excel

To generate a column of random values between 0 and 1 in Excel, use the RAND() command. To generate a value in a different range use the command RANDBETWEEN(a, b) where a is the bottom of the range and b is the top of the range.

Caution must be used with Excel as it re-calculates random values each time any operation is performed. This means that *any* random values already generated will be changed. Pressing F9 in the formula bar preserves the number but this has to be done cell by cell.

Excel will also produce a random sample of a specified size from a list of values using the Sampling Analysis Tool. (If this function is not available, run the Setup program to install the Analysis ToolPak. After you install the Analysis ToolPak, you must select and enable it in the Add-In Manager.) The output is not in any order but results can be subsequently sorted. Whilst this may seem an ideal solution if data are held on Excel there are drawbacks. The function can only be used on numeric data, and so cannot deal with file series with alpha-numeric elements and a particular item can be chosen more than once.

Scaling of Random Numbers

Most spreadsheets and scientific calculators generate numbers between 0 and 1 eg 0.523, 0.185, 0.349. However, you may need a random number in a different range, such as 1 to 20. Some spreadsheets are able to generate random numbers within any range. If you do not have a spreadsheet package that can do this, generate a series of random numbers between 0 and 1. To convert them, perform the following calculation for each number that you have generated: $(b-a) \times x + a$

where:

a is the bottom of the range in which you need a number
b is the top of the range in which you need a number
x is the random number

The resulting number will need to be rounded to the nearest whole number.
The simplest way of doing this is to add 0.5 and then ignore the decimal part.

Example:

Suppose we wish to have numbers between 1 and 100 have generated a set of 5 random numbers between 0 and 1: {0.398, 0.792, 0.581, 0.004, 0.998}.

For the first value we have $(100-1) \cdot 0.398 + 1 = 40.402$ Add 0.5 to obtain 40.902, discard the decimal part to obtain 40. Do this for each number generated.

The resulting set is thus {40, 79, 58, 1, 100}.