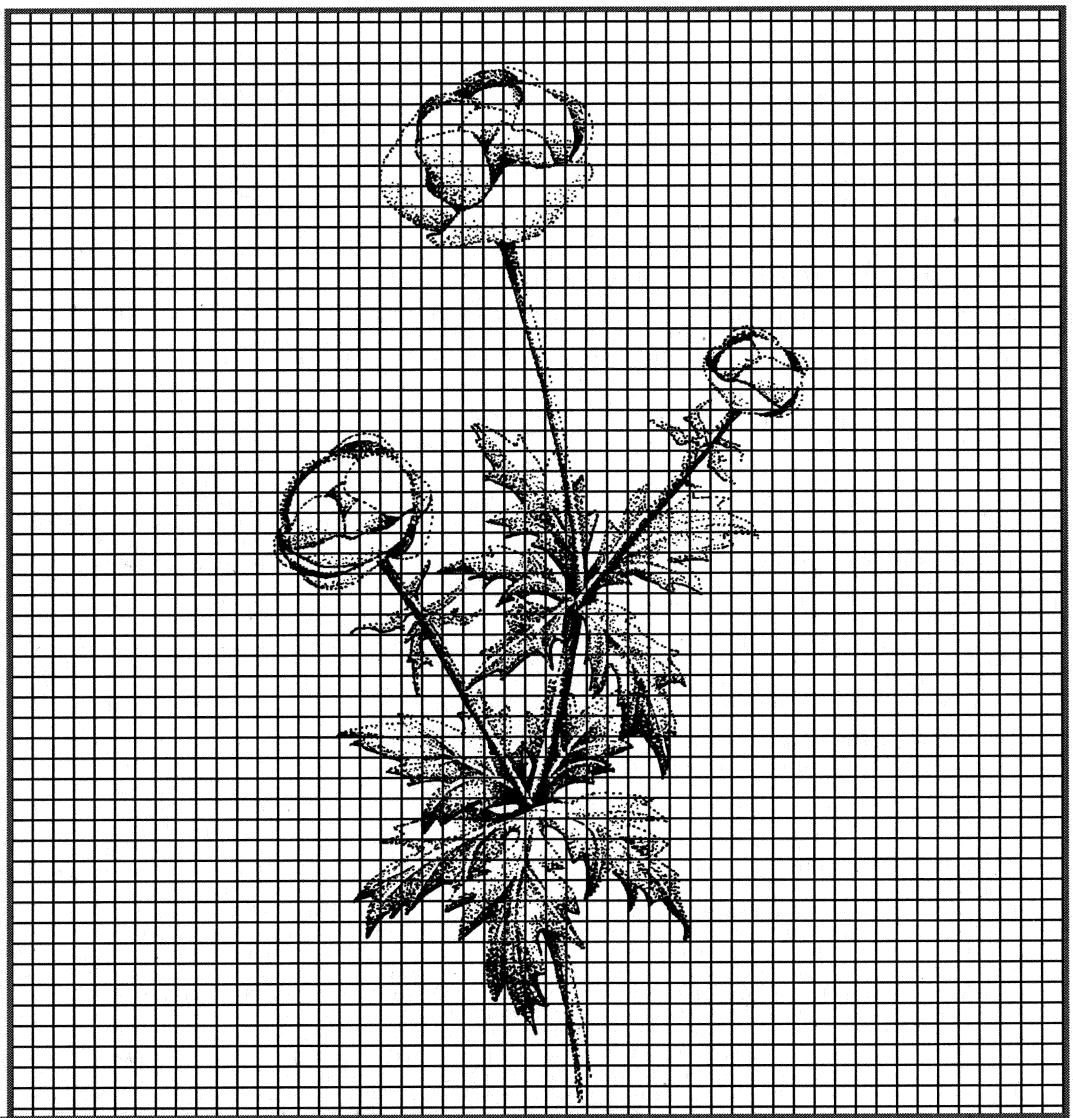
ENGLISH NATURE

English Nature Research Reports

Sampling framework for monitoring of grassland Biodiversity Action Plan targets



working today for nature tomorrow

No. 388

English Nature Research Reports

Number 388

Sampling framework for monitoring of grassland Biodiversity Action Plan targets

S M C Poulton
(Biological and Ecological Statistical Services)

ISSN 0967-876X

Contents

1. Executive summary and conclusions

The remit for this contract was to design a sampling framework for monitoring grassland Biodiversity Action Plan targets (UK Steering Group 1995; UK Biodiversity Group 1998). I have considered this contract to incorporate two main tasks:

- Firstly, to explore the existing databases of known grassland sites and appraise their utility and limitations. From this, it has then been possible to make recommendations about sampling strategy, including issues such as sample size and stratification.

- The second task has been actually to create a computer database holding the complete sampling framework. This has been implemented as a Microsoft Access 2000 Run-time System. In addition, it incorporates a "Ready-Reckoner" to assist choice of sample sizes and a mechanism to draw the required random samples. These can then be output either as printed lists or data files for further manipulation. The database has been supplied to the UK country agencies and does not form part of this report.

Consequently, this report deals with both the theoretical aspects of sampling strategy and the practical considerations imposed by the real data, along with instructions for the use of the database. Rather than produce two separate documents, I have combined these two aspects and integrated the theoretical discussion with practical instruction on using the database where it is useful to illustrate a point.

Five separate datasets were provided in electronic format for England, Scotland and Wales. Considerable manipulation was required to classify and restructure the data. In particular, different habitat classifications were provided with each dataset and these had to be transformed into the five Habitat Action Plan and two Habitat Directive types. In some cases, there was insufficient information to define the types unequivocally, so there are gaps in the sampling framework. Furthermore, the inventories from which these datasets were compiled represented different degrees of coverage of the real populations. Coverage in Wales is comprehensive, but it is thought that certain habitats (such as Lowland Dry Acid Grassland) are under-represented in the English dataset (Robertson, *pers. comm.*). There is clear evidence of incomplete geographical coverage in the Scottish non-SSSI dataset, largely because this has been compiled from several disparate surveys undertaken for research aims.

The main consequence of this is that separate samples will have to be drawn from the English, Scottish and Welsh datasets, and within them, from the SSSI and non-SSSI datasets. Furthermore, the sample units do not belong uniquely to a particular habitat type, but may contain several different habitats. Thus, samples will also have to be drawn separately from the subsets of units known to contain the required habitat types. This means that many different samples will have to be drawn, although sites may be chosen more than once if they contain more than one habitat type. This will clearly provide a saving in time if two or more habitats can be assessed on a single visit.

Considerations such as the geographical location of sites and the highly variable site size and sampling intensity were investigated. The influences which these factors have on the sampling strategy are discussed. In general I do not recommend stratification as the sampling frame already comprises separate populations which will have to be sampled independently.

Furthermore, a pilot analysis is probably too costly to undertake, but suggestions are given for a process of sequential sampling which might help to optimise the sample sizes.

The types of response variables which are likely to be derived from the monitoring program are explored. It is proposed that sites will be classified using a four-point condition score. This will allow change in condition to be categorised on subsequent visits. Consequently, the main response variable will consist of binomial data representing the proportions of sites falling into particular condition or change categories.

The choice of sample size is explored in some detail. Examples are given of the effect of sample size on the confidence intervals of population estimates. The relative merits of proportional, equal and optimal sample sizes are discussed and illustrated using the ready-reckoner. I recommend that due to the very unequal sample sizes, the best strategy would be to draw optimal samples which provide equivalent precision in calculating confidence intervals. In some cases this will result in a small proportion of the available sites being sampled but in other cases it may represent a 100% sample. The actual precision chosen will inevitably depend on the resources available for the monitoring program.

Instructions are given for drawing the samples and the use of primary and backup samples is explained. Due to the rather broad habitat classifications in some of the datasets, I expect that many of the units in the sample will fail to contain the required habitats. Furthermore, some of the sites have not been visited since the 1980s or before, so loss-rates may be high in some cases. For both these reasons, it is important to undertake an objective procedure for site replacement, using the backup samples. I have also discussed procedures for dealing with fragmented and large sites which will have to be sub-sampled in some way.

In conclusion, the fragmented nature of this sampling frame makes it difficult to draw national conclusions about the status or fate of BAP or Habitat Directive sites in the UK. In particular, I would urge caution in analysing comparisons between countries or between statutory status. It would be statistically possible to compared loss rates, for example, and maybe show significantly lower loss rates in SSSI sites. However, as the populations from which the samples have been drawn are different, it is not clear what conclusions can be made about the situation in country as a whole. Nevertheless, this is the best inventory that is currently available so providing that sufficient caveats (regarding known populations and representativeness) are clearly stated, I believe national conclusions could be drawn with caution.

Finally, no equivalent electronic dataset was provided for Northern Ireland, so I was unable to include it in the sampling database. However, all the considerations and recommendations made for the other three countries are applicable to any dataset which may subsequently by available for Northern Ireland. There is some information based on a sample survey of grassland vegetation (Cooper *et al*; 1997 and Cooper & McCann; 1994) although this is not equivalent to the grassland inventories available in the other countries.

2. Defining the sampling frame

The raw datasets supplied had two types of structure; linear or matrix. In the linear datasets, each record (row) represented a unique combination of site and habitat classification. In these cases one (or more) columns defined the site name (and code), with one column listing the habitat classification. In the matrix datasets, each record uniquely defined a site, with several columns representing the known habitat classifications within the site. All datasets contained three standard columns – Site Name, Grid Reference and County. In addition, some datasets included a Site Code and various site characteristics such as area and altitude.

Defining the sampling frame from the raw datasets was, therefore, undertaken in two stages. Firstly, the various habitat classifications which were used had to be transformed into the desired Habitat Action Plan and Habitat Directive classifications. Secondly, the linear datasets had to be "unstacked" so that each record represented one site with a series of columns representing each habitat classification.

2.1 Deriving "Habitat Action Plan" and "Habitat Directive" classifications

Two main sources of these classifications were available. Firstly, and most precisely, NVC communities specifically defined these categories. Table 1 shows the lists of communities defined for each of the Habitat Action Plan and Habitat Directive types. However, in some cases only the broad "NVC groups" of CG, MG,M & U were provided.

Table 1. NVC communities transformed into Habitat Action Plan and Habitat Directive types.

Habitat Action Plan types					Habitat Directive types	
Lowland Calcareous Grassland (LCG)	**Upland Hay Meadows (UHM)**	**Lowland Meadows (LM)**	**Purple Moor-grass & Rush Pasture (PMGRP)**	**Lowland Dry Acid Grassland (LDAG)**	**Flood Meadows (MG4)**	**Purple Moor-grass Mire/fen meadow (M24/6)**
CG1	MG3	MG1*	M22	U1	MG4	M24
CG2		MG2	M23	U2		M26
CG3		MG4	M24	U3		
CG4		MG5	M25	U4		
CG5		MG8	M26			
CG6		MG11*	M27			
CG7		MG13	M28			
CG8		CA*	AK*			
CG9						

* These communities were only included in the Scottish dataset.
CA – Provisional *Carex nigra-Agrostis stolonifera* type
AK – Provisional *Aira praecox-Koeleria macranta* type (after Cooper & Mackintosh 1996)

Secondly, NCC Phase I "Feature Classes" were available in some datasets. These comprised a four-character alpha-numeric coding as listed in Appendix 2 of the *Handbook for Phase 1 Habitat Survey* (England Field Unit 1990). The first character (alpha) defined major habitat categories – in this case all sites were coded "B" for grassland. The second character (numeric) defined pH status – 1: acidic, 2: neutral, 3: basic. The third character defined management status – 0: no information, 1: unimproved, 2: semi-improved. The fourth character defined 1 for upland (above 300m) and 2 for lowland. These feature classes were used simply to classify sites as containing Lowland Dry Acid Grassland (B1), Lowland Meadows (B2) or Lowland Calcareous Grassland (B3). All records classified as upland were excluded.

2.2 Constructing the final dataset

Five separate datasets were supplied in electronic format:

- the English Non-SSSI dataset consisted of the "Lowland Grassland Inventory" held at English Nature headquarters.
- the English SSSI dataset comprised listings from the "Ensis" database from English Nature headquarters;
- the Scottish non-SSSI dataset consisted of the "Lowland Grassland Inventory" held at SNH Headquarters;
- the Scottish SSSI dataset consisted of a compilation from SNH Area offices (although not up-to-date);
- the Welsh Dataset comprised a complete listing of both SSSI & non-SSSI sites held at CCW headquarters in Bangor.

All had different structures or habitat classifications so their transformations are described separately below.

2.2.1 English non-SSSI

The raw dataset comprised a matrix structure of 7960 records from the "Lowland Grassland Inventory". In addition the standard codes, it contained an alphanumeric Site Code, date of visit, site area and some areas of grassland. Each record represented a site or "part-site" which probably equated to a field or separate unit. This dataset also included several statutory designations such as SSSI, AONB, ESA or SNCI. All sites containing the designation SSSI were filtered out on the basis that they would be included in the SSSI dataset. This left a total of 5783 sites.

The data for habitat classification included a four-level coding (MG, CG, M, U) for each record. Some records had more that one code. They were "mapped" to Lowland Meadow, Lowland Calcareous Grassland, Purple Moor-grass & Rough Pasture and Lowland Dry Acid Grassland respectively. However, the known localities of MG3 grassland sites on a 2 x 2km grid basis were given in Cooper *et al* (1996). 127 tetrads were defined in total. So in addition, Upland Hay Meadows (UHM) were defined as all sites known to contain Lowland

Meadows and with grid references in the list of tetrads. It is important to note that in this dataset not all the sites will contain the required habitat types.

2.2.2 English SSSI

The raw data for this dataset comprised two related tables which were output from the "Ensis" database. The first consisted of a linear listing of all SSSI sites with Phase II NVC community codes specified in the list in Table 1. Each of the 4150 records represented a unique combination of the Site, "Level 1 Feature" and NVC community. The second table consisted of a linear listing of all 6386 sub-units of the sites in the first table.

The first stage was to transform the NVC communities to the required habitat types as shown in Table 1. These were then "unstacked" to give a matrix format with each row representing a unique Level 1 Feature Code. These were then linked to the table of sub-units using this field as the common key-field.

The structure of the resulting table requires some explanation. The table comprised 6386 records, each of which was a single unit within a site. There were 1642 unique sites, but 37 of these were duplicated across counties. (The duplicates were discarded, leaving the site allocated to the first county in alphabetic order.) Each unit was classified as containing one or more of the seven required habitat types. However, as the NVC information was only available at the site level, this was replicated across all the units within the site. Thus, all that can be said about the relationship between units and habitat types is that the unit belongs to a site which is known to contain a particular habitat, but that does not mean that the unit itself contains the habitat. The rationale for including units in the sampling frame is discussed further in section 2.4.

2.2.3 Scottish non-SSSI

This dataset was in matrix format and comprised 594 unique site records. There were 24 columns of NVC communities including three previously unidentified types (labelled CA: neutral grassland, FH: not used & AK: acid grassland). From these it was possible to identify Lowland Meadows, Upland Hay Meadows, Lowland Dry Acid Grassland and Purple Moor-grass & Rough Pasture. However, the only calcareous community listed was CG10 so no Lowland Calcareous Grassland was identified. Both Habitat Directive types were identified, although no M24 was present. 525 sites were left after the required habitat types were defined.

2.2.4 Scottish SSSI

This dataset comprised a linear table of 1,414 records. 1,365 of these used Phase I Feature Classes and 49 used Phase II NVC Classification. As there were no records classified as MG3, no Upland Hay Meadows could be classified for this dataset. Lowland Calcareous Grassland, Lowland Meadows and Lowland Dry Acid grassland were identified as described in section 2.1. Purple Moor-grass & Rush Pastures and the two Habitat Directive types could only be defined for the 49 records with NVC classifications. The table was then "unstacked" to give 710 unique site records, only 559 of which contained one or more of the desired habitat types.

2.2.5 Wales

The Welsh dataset was the most comprehensive and complete supplied. It comprised a matrix of 954 site records with a total of 56 NVC communities identified. No MG3 was present. However, all the other four BAP types and the two Habitat directive types could be identified. A total of 897 sites contained one or more of these habitat types.

2.3 Summary of the sampling frame database

The structure of the database holding the final sampling frame is a flat-file table with 14,150 rows and 14 columns. These fields are defined in Table 2 – most of them are self-explanatory, but the numeric codings for country and county use related tables to return their text values.

The combined sampling frame for all three countries, including statutory and non-statutory sites, comprises 14,150 units (Table 3). There is a fairly equitable split between SSSI and non-SSSI sites overall and in each of the three countries independently. However, the bulk of the units were found in England (86%). It should also be remembered that not all English SSSI units are certain to contain the BAP types (see sections 2.2.2 and 2.4). Furthermore, the distribution of habitat types between countries and statutory designation was very inequitable. This has serious consequences for the subsequent sampling protocol.

The database has been supplied as an Microsoft Access 2000 run-time system. On opening the database, the population displayed in the user interface is the whole dataset; 14,150 records. However, this form allows the selection of any combination of Country, SSSI Status and the seven habitat types. For example, if you select England, you will see that the population now defined in the sampling frame is 12,169. Further selection of, for example, SSSI and Upland Hay Meadows defines a sampling frame of 267 (c.f. Table 3). Selection of samples from this population is dealt with in section 4 below.

Table 2. Definitions of fields in final database.

Field Name	Description
ID	Unique numeric identified
Country	Numeric coding for country
County	Numeric coding for county
SiteCode	Numeric or alpha-numeric code
SiteName	Site Name
GridRef	Six-figure OS grid reference
SSSI	SSSI status (Boolean)
UHM	Presence of Upland Hay Meadows (Boolean)
LM	Presence of Lowland Meadows (Boolean)
LCG	Presence of Lowland Calcareous Grassland (Boolean)
PMRGP	Presence of Purple Moor-grass & Rough Pasture (Boolean)
LDAG	Presence of Lowland Dry Acid Grassland (Boolean)
MG4	Presence of MG4 Flood Meadows (Boolean)
M24/6	Presence of Purple Moor-grass mires and fens. (Boolean)

Table 3. Summary of the sampling frame

Country	Statutory designation	Habitat Action Plan types					Habitat Directive types		Total+
		LCG	UHM	LM	PMGRP	LDAG	MG4	M24/6	
England	SSSI	1,980	267	2,485	2,491	1,278	454	600	6,386
	non-SSSI	1,957	406	3,633	993	688	-	-	5,783
	Total	3,937	673	6,118	3,484	1,966	454	600	12,169
Scotland	SSSI	72	-	373	13*	337	1*	2*	559
	non-SSSI	0	38	355	234	209	0	10	525
	Total	72	38	728	247	546	1	12	1,084
Wales	SSSI	38	0	196	221	147	5	123	290
	non-SSSI	32	0	413	440	283	2	211	607
	Total	70	0	609	661	430	7	334	897
Total	SSSI	2,143	268	3,170	2,993	2,205	462	729	7,235
	non-SSSI	1,989	444	4,401	1,667	1,180	2	221	6,915
Grand total		4,079	711	7,455	4,392	2,942	462	946	14,150

- Impossible to identify these habitat types from the data provided
* Based on limited recordset of 49 cases
\+ May contain more than one habitat type

Two other major factors need to be considered before a sampling strategy can be finalised; geographical location and site size. Both these factors could have important implications for stratification or within-site procedures discussed below (sections 3.5 and 5.2).

2.3.1 Geographical location

This factor is represented by the variable county. Appendix I summarises the number of units found in each county. (Note that for the non-SSSI sites in Scotland only the SNH Areas were provided in the original dataset so these have been summarised separately.) The important point here is that there is a very inequitable split across counties which, especially in England and Scotland, seems to bear little relationship to county size.

2.3.2 Site size

Firstly, no data was provided on the size of the units from SSSIs in England, so this summary is based on site areas. In England & Wales site size was very highly right-skewed. Consequently, there was a very high proportion of sites less than 5ha in area (55% and 51% respectively), with only 14% and 13% of sites respectively being greater than 20ha (Figures 1 & 2). However, in Scotland, larger sites were far more abundant with 485 sites being greater than 20 ha and only 27% less than 5 ha (Figure 3).

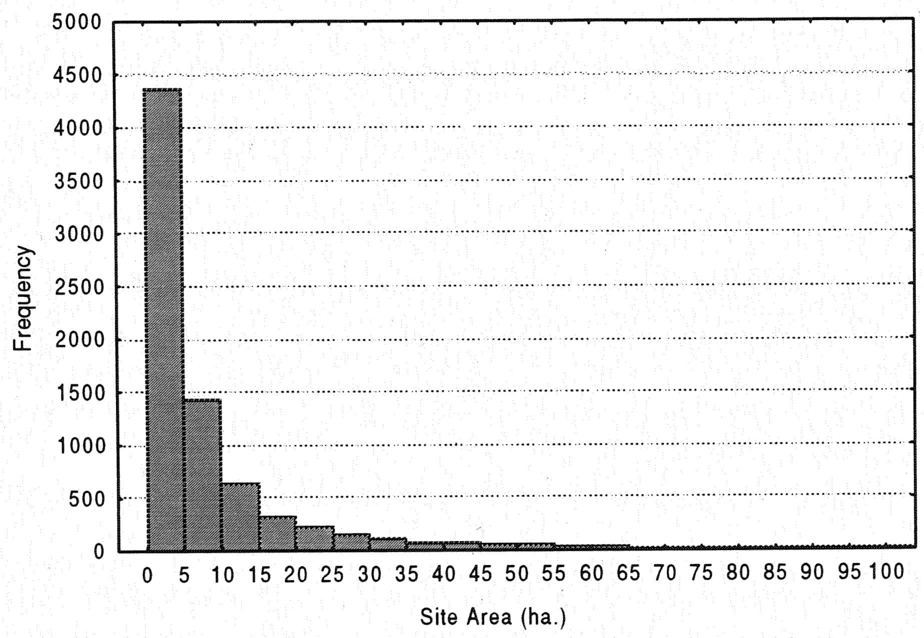


Figure 1 Frequency distribution of site areas in England

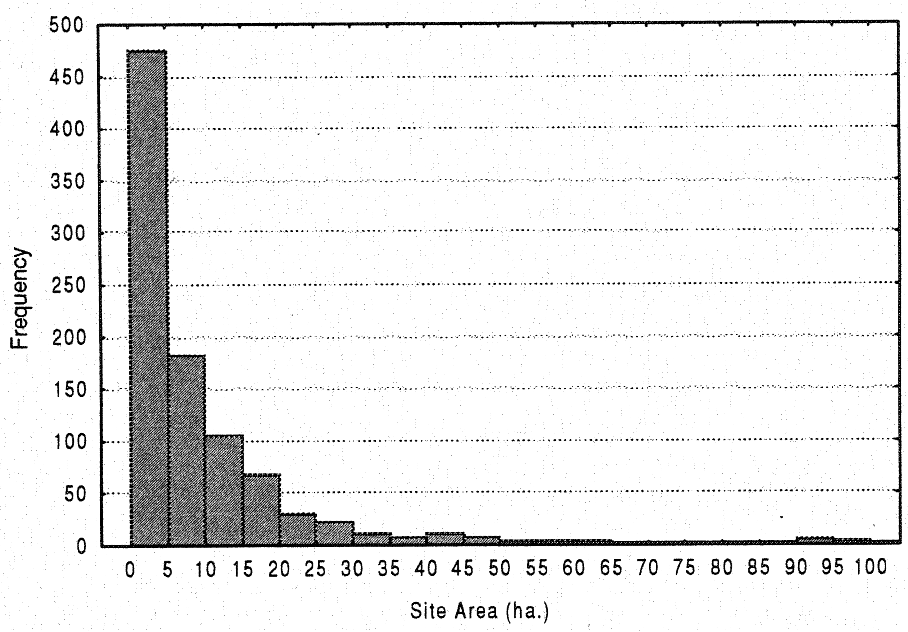


Figure 2. Frequency distribution of sites areas in Wales

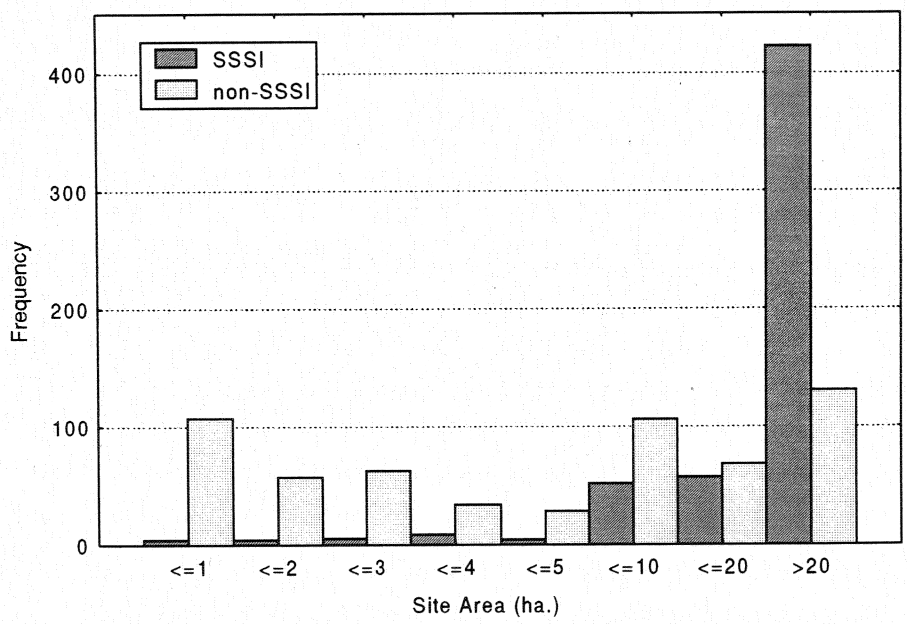


Figure 3. Frequency distribution of site areas in Scotland

Furthermore, there was an especially large discripancy between site areas in Scotland between SSSI and non-SSSI sites, with over 75% of SSSI sites being larger than 20 ha. This effect was also highly significant in England and Wales, though much less marked, with 32% and 22% respectively ($\chi^2_{(7)}$, p close to zero in all cases).

2.4 Constraints of the sampling frame

It is important at this stage to highlight the many constraints which this sampling frame will impose on the sampling strategy.

- Firstly, the overall sampling frame was derived from five separate datasets. Many of the characteristics of these datasets varied, such as the source of data, the coverage and the data used for the habitat classifications.

- The relationship between the statistical population and biological population was not always known and probably not the same in each dataset. In other words, how did the known set of sites in the sampling frame compare with the real set of lowland grassland sites? In the case of SSSIs the coverage was probably fairly good and should allow some conclusions to be drawn from the sample. However, the non-SSSI sampling frames were probably more patchy. In England, for example, it is thought that the Lowland Dry Acid Grassland was highly under-represented. In Scotland there were some geographical regions, such as Orkney, Caithness, Inverness and the south-western islands, which were not covered.

- The data provided for the definitions of habitat classifications was highly variable. In some cases (all Welsh data, the English SSSIs and the Scottish non-SSSIs) NVC communities were available, allowing all required habitat types to be identified. In the

English non-SSSIs, only broad NVC groups were available, allowing four of the BAP types to be identified. The fifth type (Upland Hay Meadows) could only be inferred from their known distributions within 127 tetrads. Finally, the Scottish SSSIs only had NCC Phase I classifications, with a small subset (3.5%) having NVC communities.

- The resolutions of the sampling-units within the sampling frame is different for England compared to Scotland & Wales. In the latter two cases, the sampling-units were essentially whole sites, either designated SSSIs or sites defined in the Lowland Grassland Inventories. In Scotland especially, this resulted in some sampling-units being very large (> 1000 ha.) which still only had the same probability of selection in a sample as very much smaller sites. In contrast, in England the non-SSSI dataset, derived from the Lowland Grassland Inventory, clearly had a combination of whole sites and part-sites as its records. The size of this dataset (5783 records) precluded any attempt manually to combine individual fields into sites based on site names or codes. Consequently, this dataset is a hybrid of discrete whole sites, single sites representing a number of separate sub-units and individually identified sub-units or fields. The English SSSI dataset had a relational link between sites and units which, ideally, would have allowed a two-stage sampling strategy – first select sites and then for sites containing more than one unit, randomly select units within the site. However, to maintain consistency with the non-SSSI dataset, I decided to expand the 1605 SSSI sites into 6386 records representing individual units.

3. Choosing the sample size

The process of choosing the sample size is usually dictated by a number of factors. Foremost of these is the required precision of the estimate derived from the sample. If parametric data are gathered, then the sample mean and variance can be calculated to estimate the population mean with required confidence limits. Ideally, a small pilot survey would be undertaken first to indicate the approximate sample variance and then a power analysis carried out to calculate the sample size required to achieve sufficient precision of the estimate of the population parameter. At this stage, it might be possible to reduce the sample variance by using stratification or increasing sampling efficiency by using multi-stage sampling.

Although it appears that the data derived from this monitoring program will be qualitative rather than quantitative, I shall follow this procedure in developing the rationale for choosing the sample size.

3.1 Defining the response variables

Within each site it is intended to use the English Nature SURF Condition Assessment Protocols (Robertson & Jefferson 1999). This procedure uses a subjective (semi-quantitative) assessment of a number of pre-defined attributes within a maximum area of 15ha. The site is then assessed on a four-point scale; Favourable, Unfavourable, Partially destroyed or Destroyed.

For the purposes of this monitoring program, it is assumed that after the first visit any sites which are partially or wholly destroyed will be replaced with new sampling units to achieve the required sample size. Consequently, on subsequent visits the sites would be assessed on the same four-point scale to give eight possible combinations (Table 4). These can be combined to give three change categories; Improving, Unchanged or Declining, although only one of the eight combinations can be defined as improving. These categories are slightly different from those defined in Field Guidance Notes, but both essentially define the sites on the basis of initial condition and subsequent change.

Table 4. Combinations of SURF Assessments Between Two Site Visits

		Outcome of second visit			
		Favourable	**Unfavourable**	**Partially destroyed**	**Destroyed**
Outcome of first visit	Favourable				
	Unfavourable				

Change categories: Improving | Unchanged | Declining

This suggests that the basic response variable will be a summary of these qualitative categories. For example, the proportion of sites which fell into the favourable category on the first visit, or the proportion of sites which were found to be declining after the second visit.

3.2 Defining the hypotheses to be tested

Although not hypothesis testing *per se*, the first use of these response variables would be to calculate confidence limits for the proportions. This would allow conclusions to be drawn about the population as a whole. For example, if the proportion of destroyed sites in the sample was 10%, this would equate to 50 sites in a population of 500 sites. Depending on the sample size and required precision, the confidence limits might then be calculated as, say, between 45 and 55 sites (see section 3.5 below).

Bearing in mind the caveats mentioned in the Executive Summary it might be possible to test certain restricted hypotheses such as "Is the rate of site destruction different between SSSI and non-SSSI sites?" This would ideally be tested using either Chi-squared analysis or Fisher's Exact test on the resultant 2 x 2 contingency table (see *e.g.* Zar: 1984). However, this does not take into account the relative sizes of the samples in terms of their populations. It might, therefore, be advantageous to compare the proportions directly, with their confidence limits derived using the "finite population correction" which takes this into account. Testing of larger contingency tables could be undertaken (with caution) using Chi-squared or Log-Linear analysis (Fienberg; 1970)

3.3 Pilot survey

Considering the intensive nature of the fieldwork for this monitoring program, I do not propose a specific pilot survey. However, it would be possible to use the technique of "Sequential Sampling" to optimise sampling effort.

For example, the question of interest might be the rate of destruction of sites, as described above. All sites currently in the sampling frame were known to exist at some time in the past, and the date of last visit was recorded for most sites. Consequently, it would be possible to visit a proportion (say 25%) of the sample and derive a relationship between the time since last visit and the loss-rate of sites. This could be used as an approximate indication of the rate of lost to be expected between the first and subsequent visits in the monitoring program. This, in turn, would allow the final sample size to be altered according to the absolute loss-rate, or the difference between loss-rates in different categories of sites. Most importantly, the confidence limits for proportions depend largely on the size of the proportions themselves sites (see section 3.5 below).

3.4 Cost-benefit analysis

In the original tender for this contract I proposed carrying out a cost-benefit analysis to determine the optimum balance between "within-site" and "between-site" costs. However, this was based largely on the assumption that some parametric data would be obtained. In the event, as sites are only being classified into categories, it is not possible to vary the within-site effort – it takes as long as necessary to assess the site – between approx. 30 and 100 minutes (Robertson & Jefferson; 1999). The only way to improve the precision of the estimate of the proportions of sites falling into various categories is to visit a larger sample of sites. Consequently, I have carried out no cost-benefit analysis *per se,* as the power analysis provides the main consideration for sample size.

3.5 Stratification

Similarly, I do not propose that the sampling strategy should include stratification, for a number of reasons. Firstly, stratification is usually employed within a single large sampling frame, where there is *a priori* evidence that its sub-division would reduce the variance of any sample statistic. Secondly, stratification might be employed to ensure a sufficiently large sample from a small category of particular interest within the sample frame.

As shown in Table 3, the sampling frame already consists of five separate populations, derived from different sources and using different habitat classifications. For this reason, it will be essential to draw independent samples from these populations. Furthermore, within each of these populations, there are seven pseudo-populations representing the subsets of sites known (or expected) to contain each of the required habitat types. I see no practical alternative to drawing separate samples from each of these subsets.

The mechanics of stratified sampling require that separate samples be drawn from within each stratum. As this is essentially the procedure imposed by the structure of the sampling frame, I do not propose any further stratification of the dataset, with two provisos. Firstly, as there was a single Welsh dataset comprising both SSSI and non-SSSI sites, which is thought to be equally representative (and for the sake of consistency) I would recommend stratifying by Status within the Welsh dataset.

Secondly, the disparity in the number of sites found between counties in all three countries means that a number of counties may not be represented in a randomly selected sample. If it is important to have good regional representation, then it may be necessary to amalgamate counties into regions and then stratify on them. Due to the number of counties I would not recommend stratifying on them individually and I have made no provision to do this in the Sampling database.

3.6 Power analysis

A power analysis would normally be carried out to determine the sample size required to test a specific hypothesis with probabilities α and β of not committing Type I and Type II errors respectively. However, as described above in section 3.2 the primary requirement of this monitoring program will be to calculate proportions from the sample to estimate population proportions with confidence limits. To facilitate this, the Sampling Database includes a "ready-reckoner" in the lower part of the main form. This calculates confidence limits based on various population and sample sizes.

To give the calculations some realism, the ready-reckoner is based on the population selected as the sampling frame in the upper part of the form. By default, when the form is opened, the population includes all 14,150 records in the database. The sample size can be defined either as a percentage of this population or as an absolute number. By default, this is set to 1%, which equates to a sample size of 142. Two further parameters are required. Firstly, p represents the proportion of the sample which falls into a particular category, eg the proportion of sites destroyed. By default this is set to 50%, which represents 71 in the current sample of 142. The second parameter is the confidence interval. By default this is set to 95% ($\alpha = 0.05$) but 90% or 80% intervals can be chosen. The 95% confidence intervals for a p of 50% with these sample and population sizes are given as 43.0% to 57.0%. A p of 50% in the sample gives a population estimate of P of 7,075 sites. In other words, if the rate of site destruction in the sample was 50%, this would equate to an estimated loss of 7,075 sites in the whole population. However the sample confidence intervals would mean that you could be 95% certain that the actual number of sites lost would lie between 6,084 and 8,066 [1].

3.6.1 Proportional, equal or optimal sample sizes?

One of the major questions to address when defining sample sizes from separate populations or strata is whether to draw proportional, equal or optimal sample sizes. The former has the advantage that when parametric data are recorded, overall means across strata can be calculated without needing to weight the individual stratum means, thereby considerably easing computation. However, if the strata or populations differ greatly in size then proportional sampling may be inefficient, with the larger populations being effectively over-sampled. Conversely, if equal sample sizes are taken from each population then the smaller populations may be over- (or completely) sampled, whilst the larger populations may be under-sampled. The best option, especially in the current scenario, is to draw optimal samples, which allow similar degrees of confidence to be calculated for statistics from each sample.

A simple example should illustrate this approach. Using the Sampling Database, select Purple Moor-grass sites in Wales to give a population of 661. By selecting SSSI and non-SSSI alternately, you can see that there are almost exactly twice as many sites in the non-SSSI population(440) as the SSSI population (221). Set the expected p to 5% (to represent, say, a 5% destruction of sites) and leave the confidence intervals as 95%.

Now set the sample size to 10%. For the SSSI population this equates to a sample of 22. The confidence intervals for $p = 5\%$ from a sample of this size are approximately 0.4% to 19.9% - fairly wide. Now select the 440 sites in the non-SSSI population; this equates to a sample of 44. Note that the confidence intervals around the same 5% estimate are considerably narrowed at 1.0% to 13.9%. This is because the population is twice the size and so a proportional sample will be twice as large too.

An alternative approach would be to draw equal samples from each population. Click the "Absolute" radio-button which now highlights the sample size of 44 for the non-SSSI population. Now select the SSSI population – note that the sample remains at 44 but this now represents an approximate 20% sample. As this is a larger proportion of the population, the Finite Population Correction reduces the confidence intervals to 1.2% to 13.4%. This is not a large difference, but it does suggest that a sample of 44 from a population of 221 is slightly over-sampled compared to a population of 440. By trial and error, it is easy to see that a sample of 41 from a population of 221 would give approximately the same precision as a sample of 44 from a population of 440, thereby saving 3 sites compared to the equal sampling strategy.

It is worth experimenting with a range of different values, but a number of important points should be noted when choosing the final sample sizes for the monitoring program:

---

[1] To calculate these confidence limits, values from the *F*-distribution are required. Please note that for simplicity, an approximation to this distribution has been incorporated in the ready-reckoner, so the confidence intervals provided are slightly optimistic. For final analysis of the monitoring data, a proprietary statistics program with an algorithm for calculating correct probability distribution functions should be employed.

- As proportions follow the binomial distribution, their confidence intervals are not symmetrical, always being truncated towards zero and extended towards the mid-value of 50%.

- For any given sample size confidence intervals are widest around the mid-value of 50%. In other words, if it is expected that the rate of site destruction will be near 50%, then a larger sample with be required to give equivalent confidence intervals than if expected loss-rate was only 5%.

- Sample size is the primary determinant of confidence intervals. However, the relationship is geometric rather than linear. For example, reset the population to 14,150 and set the sample to 50. Now set *p* to 50% to make the confidence intervals symmetrical – they are now ±12.1%. Double the sample size to 100 and note that the confidence intervals are now reduced to ±8.5%, and a further doubling to 200 gives confidence intervals of ±5.9%. In other words, you need to quadruple the sample size to achieve a halving of confidence intervals.

- The Finite Population Correction has little influence on confidence intervals for samples less than about 25% of the population. For example, reset the population to 14,150 and set the sample to 1000 (7.1%). Note that the confidence intervals around a *p* of 50% are ±2.56%. Now reduce the population to 6,915 by selecting non-SSSI. The sample now represents 14.5%, but the confidence intervals have only reduced to ±2.45%. Reduce the population again to 3,633 by select England and Lowland Meadows. Now the sample represents 27.5% of the population but the confidence intervals are still only ±2.26%.

4. Drawing the sample

The Sampling Database provides a mechanism for drawing a random sample from the required population. In fact, the program randomly selects two samples from the population; a "Primary" and a "Backup" sample

4.1 Using the primary and backup samples

The primary sample is selected first with the size specified in the database. These sites should be used first and the backup sample only utilised if any primary sites prove not to be suitable.

The backup sample is a further random selection from the remaining sites in the population. It is usually half the size of the Primary Sample unless the latter is 50% or more, in which case all the remaining sites are selected. This sample is used for the replacement of sites under two different scenarios. Firstly, on checking the site records, it is found that the habitat required is not actually present in the site. This might have occurred, for example, in a record from the English SSSI, where only the site, not the unit, is known to contain the specified habitat (see section 2.2.2 above). Secondly, on visiting a site, it might be found that the habitat of interest has been partially or totally destroyed. Although this is itself of historical interest (and see comments about selecting optimal sample size in section 3.3 above), to provide the desired sample for revisits, a replacement site would be required.

In principle, replacement sites should be used from the Backup Sample in the order they were original selected. This ensures that final the sample remains truly random. In the first instance, where sites need to be replaced as part of the initial checking procedure, which is presumably office-based, this should provide no problems. However, the logistics of site visits might encourage the selection of replacement sites on a county basis, at least. Although this is not recommended, I recognise that the practical considerations of having a local replacement might outweigh the statistical objections.

After the population has been specified and the sample size chosen, click the "Draw Sample" button. Two options are then provided:

- the samples can be printed to a paper report, or
- they can be output as a comma-separated-value (CSV) file

Note that each time the "Draw Sample" button is clicked a new sample is selected. So, if you print out a paper report, then click the "Draw Sample" button and then produce a CSV file, these data will not be the same as those in the report.

4.2 Paper reports

The paper reports are provided primarily for ease of use in the initial checking of the sites for the required habitats. To facilitate this, the primary sample is printed first, sub-divided into counties and ordered by site name. In addition, the site code (if present) and the grid reference are also listed. Each county is printed on a separate page (in case distribution to area offices is required). These are followed by the backup sample, listing the same information but in the original random order selected. Any replacement sites required should be read off the backup sample in the order listed to retain the random nature of the overall sample.

4.3 CSV files

This option has be provided to enable export to a spreadsheet or other database for further manipulation. The data are not sorted in any way, but each record is coded as primary or backup sample (1 or 2), and has the random number used for its selection. All the fields from the sampling database are exported, so samples drawn from different populations can be combined, but still identified. This would allow sites which contain more than one required habitat to be identified easily, so that repeat visits are not made accidentally.

5. Within-site procedures

Objective procedures need to be specified for dealing with two within-site situations.

5.1 Compound sites

Firstly, many records in the database represent compound sites. These may be in the form of separate physical units such as fields, or the required habitat may exist in separate blocks within the site. In the case of small sites it may be possible to survey all the units but this might result in somewhat heterogeneous assessments – one unit may be in perfect condition whilst another has been destroyed. In the case of large sites, where the units themselves approach the maximum recommended size of 15ha (Robertson & Jefferson 1999), a sub-sampling procedure could be employed. In this event, it would be necessary to number each unit or block and then randomly select one of these. The drawback with this approach is that the single unit may not appear to be representative of the site as a whole. However, as the site itself is only one of a random sample, this effect should not invalidate the assessment. For smaller sites, it may be practical to select a number of units and survey these to give an aggregate assessment.

5.2 Large sites

A number of records in the database represent sites which are hundreds of hectares in size. This is especially the case for the Scottish SSSIs (see section 2.2.4). Given that the required habitats are likely only to be a part of this area, there may still be cases where these extend over areas much greater than 15ha. In the case of very large areas, say over 150ha, then it would be possible to divide the area into a rectangular 15ha grid (16ha would be easier!) and randomly select one unit. The same considerations regarding representativeness discussed above apply here.

A more difficult problem is how to sub-sample sites which are larger than 15ha but not big enough to warrant superimposing a grid. Suppose, for example, a site was 25ha in area with a trapezoid shape (Figure 4). It would be possible to divide it into five radial sectors of approximately equal size (5ha). To do this it would be necessary to identify the "centre-of-gravity" of the site and extend the radii out from here. This would also tend to equilibrate the lengths of the boundary to reduce edge-effects. However, for linear sites or those with extreme shapes, some other subdivision might be necessary. The principle is to divide the site into sectors with equal areas which are the highest common denominator of the site area and 15ha, and then randomly select the required number of sectors. So, for example, a 32ha site could be subdivided into four 8ha blocks and two chosen (or eight 4ha blocks with four chosen) to give 16ha.

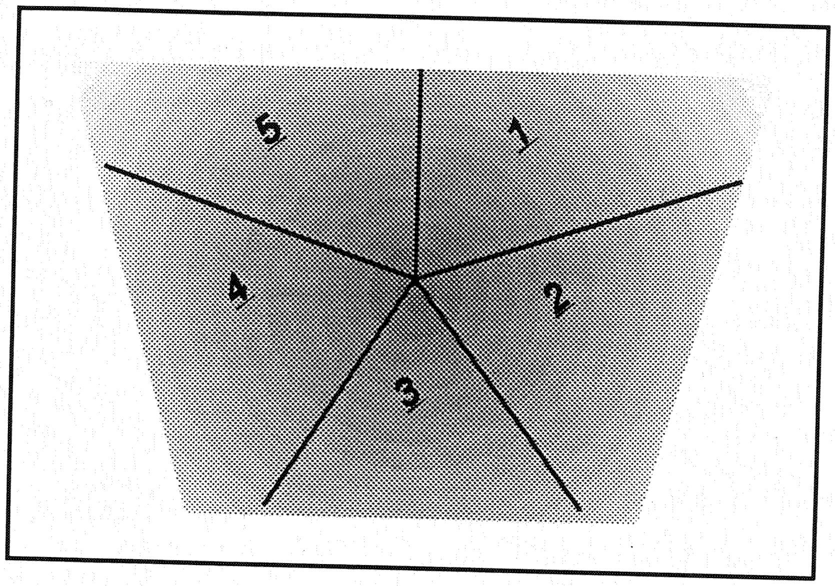


Figure 4. Schematic example of sub-division of large site into approximately equal areas

References

COOPER, A., MURRAY, R. & MCCANN, T. 1997. *The Northern Ireland Countryside Survey*. Belfast: Environment & Heritage Service.

COOPER, A. & MCCANN, T. 1994. *The Botanical Composition of Grassland Land Cover Types in Northern Ireland.* DOE, Northern Ireland.

COOPER, E., COOCH, S. & RODWELL, J. 1996. *Habitat Quality & Value in the Countryside: Northern Hay Meadows.* Lancaster: Unit of Vegetation Science.

COOPER, E. & MACKINTOSH, J. 1996. *NVC review of Scottish grassland surveys.* Scottish Natural Heritage Review No 65.

ENGLAND FIELD UNIT. 1990. *Handbook for Phase 1 Habitat Survey – A Technique for Environmental Audit.* Peterborough: Nature Conservancy Council.

FIENBERG, S.E. 1970. The analysis of multidimensional contingency tables. *Ecology*, **51:** 419-433.

ROBERTSON, H, & JEFFERSON, R. 1999. *Lowland Grassland Condition Assessment Protocols: Interim Field Guidance Version 1.1* Peterborough: English Nature.

UK BIODIVERSITY GROUP. 1998. *Tranche 2 Action Plans: Volume II - terrestrial and freshwater habitats.* Peterborough: English Nature.

UK STEERING GROUP. 1995. *Biodiversity: the UK Steering Group Report.* London: HMSO.

ZAR, J.H. 1984. *Biostatistical Analysis.* New Jersey: Prentice Hall.

Appendix I Breakdown of units by county

England

North Yorkshire	1310
Cumbria	822
Derbyshire	742
Dorset	728
Somerset	720
Devon	655
Somerset	720
Devon	655
Wiltshire	547
Hampshire	479
Hereford & Worcester	426
Durham	368
Gloucestershire	350
Cornwall	338
Norfolk	330
Staffordshire	328
Cheshire	323
Suffolk	276
Shropshire	275
Leicestershire	260
Northumberland	239
Humberside	232
Kent	213
East Sussex	207
Lincolnshire	194
Lancashire	192
Avon	185
Nottinghamshire	178
Hertfordshire	165
Essex	150
Oxfordshire	147
West Sussex	141
Isle of Wight	140
Cambridgeshire	137
Berkshire	131
Greater Manchester	129
Northamptonshire	118
West Yorkshire	107
Warwickshire	105
Surrey	96
Buckinghamshire	71
Greater London	64
Tyne and Wear	62
Cleveland	59
West Midlands	55
Merseyside	51
Bedfordshire	44
South Yorkshire	36

Scotland

Forth & Borders	138
Argyll & Stirling	103
Tayside & Clackmannanshire	71
Strathclyde & Ayrshire	64
Dumfries & Galloway	46
Grampian	36
North Highland	30
Northern Isles	29
East Highland	18
West Highland	16
Western Isles	

non-SSSI (SNH Areas)

Lochaber	98
NE Fife	49
Western Isles	48
Renfrew	34
Stirling	34
West Lothian	30
Shetland	27
Kyle & Carrick	24
Perth & Kinross	24
Roxburgh	20
Skye & Lochalsh	19
Moray	16
Ettrick & Lauderdale	13
Kincardine & Deeside	12
Cunninghame	11
Dunfermline	9
Inverclyde	9
Cumnock & Doon	8
Berwickshire	7
Nithsdale	6
Stewartry	6
Falkirk	5
Kircaldy	5
Kilmarnock & Loudoun	3
Clackmannan	2
Tweedale	2
Aberdeen	1
Wigtownshire	1
Sutherland	1
Gordon	1

Wales

Dyfed	260
Powys	232
Gwynedd	133
Clwyd	73
Mid Glamorgan	68
Gwent	67
West Glamorgan	41
South Glamorgan	