6: Data handling and analysis

Data management

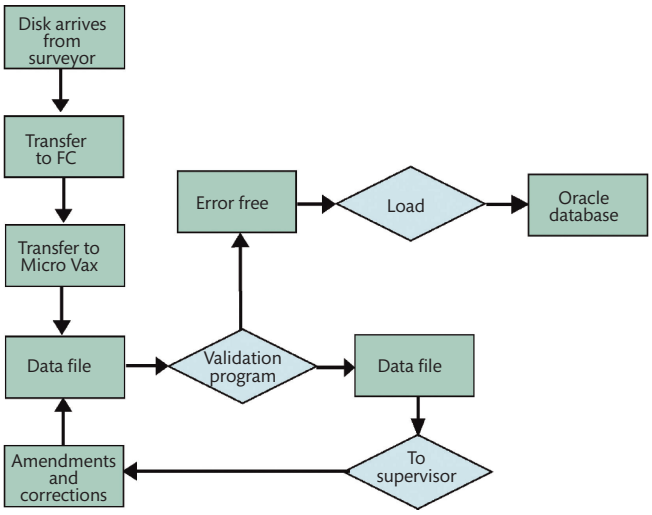
Transmission of data

The fieldwork data were submitted by the field teams on floppy disk, by post or by e-mail. A disadvantage of submitting disks by post was that the data processing team did not know when to expect them. Consequently, it was the field team's responsibility to follow up on a disk if no response from headquarters was received within a reasonable time. A backup disk held at the field station meant that any disks lost in the post could be copied and re-submitted. Unfortunately, the e-mail system, with read-receipt confirmation to the sender, came into use for the transfer of data only towards the end of the fieldwork.

Processing data

Figure 48 summarises the main stages of the procedure used for handling the data from their arrival on floppy disk until they were loaded onto the Oracle database. The arrival of the disk was recorded on a spreadsheet so that it was possible to follow progress on several sets of data at the same time. The validation procedure actually consisted of more than one stage; it was not a single program.

Figure 48 Summary of data handling procedures.



The programs performed the following tasks:

- *locates missing data items* – looked at supplied information and compared that with a table which showed which items should be present, e.g. species present for a section described as 'conifer' forest type. If missing then this was flagged as an error
- *inconsistent data* – the validation checked the data entered and checked associated values against each other, e.g. if forest type was 'coppice' then the thinning category must correspond, i.e. 'coppiced'
- *checks for data that exceeded pre-set limits* – for instance a tree height of more than 65m or too many sections for a square
- *empty records* – empty records were ones that contained no data and were usually deleted by the surveyor
- *checks on the structure of the data* – for example, elements entered for a section that had been inadvertently deleted by the surveyor
- *cross-checks with the database* – could generate messages such as 'This woodland already present on the database'.

Checks on progress were also made, by comparing grid references of squares in the incoming data with those on the list of selected samples. This program provided:

- lists of grid references that did not match selected samples
- lists of squares not yet received for incomplete woodlands
- lists of woodlands for which all squares had been received and work was therefore complete.

The validation system developed as knowledge of the data and their inter-relationships grew. A print of the data was returned to the surveyor with the error listings.

Most errors were caused by the surveyor neglecting to enter values when the data required an input, despite the checks already built into the data collection software. These errors would have been very difficult to correct if the data collection had been entirely electronic, the only other resource being the surveyor's memory of a particular square. However, from an early stage it had proved very useful for the surveyor to sketch map the square and sections as a way of allocating area. This was a valuable resource for correct-

ing the data. Other information, such as timber potential, would not normally have been manually recorded but the Husky had proved insufficiently reliable in the early stages of the survey and surveyors felt more confident if they also had a paper record of the data.

If the errors were relatively few and of a minor nature then the required information could be gathered by a phone call. The data file was corrected by editing the data; there was potential here to create 'new' errors that had nothing to do with the surveyor if the editing was not done with sufficient care.

Loading the data onto the database

When work for a particular area was completed and all the data files were free of errors, the information was loaded onto the Oracle7 database. The procedure for loading the data onto the database rejected any record that appeared to be the same as one that already existed on the database. One way of simplifying the loading of the data for an area was to compile all the data files into one large file; this was easier to work with as it meant that overall there were fewer files to handle and it also gave the opportunity to detect and resolve any duplicated information.

Organisation of data on the database

The description of the field data was set out in six levels as shown in Figure 49. Where surveyors recorded more than one value against a particular item then the structure of data on the database had to cope with the extra information. This was done by establishing another store, or 'table', of information designed to store the values for these particular data items. An example of this would be the breakdown of the area of a woodland by the Interpreted Forest Types, each of which has an area, in contrast to the grid reference for the wood which is a single value. In total, what were 6 levels in the field data become 13 in the database for the Main Woodland Survey. As an addition there were 'look-ups', which stored the codes for all the data and their meanings. A set of 14 tables stored the field data for the Small Woodland Survey (Figure 50).

The database was a good way of storing and organising the data and worked well in conjunction with the validation programs. It was not so good for statistical analysis of the data. A program was written to extract the information in the woodland blocks required for analysis, for example the main woodland data for a particular county. This was then transferred to a PC for further work as a file with one record per element, drawing data from all levels of the hierarchy.

Figure 49 Main Woodland Survey – data table structure.

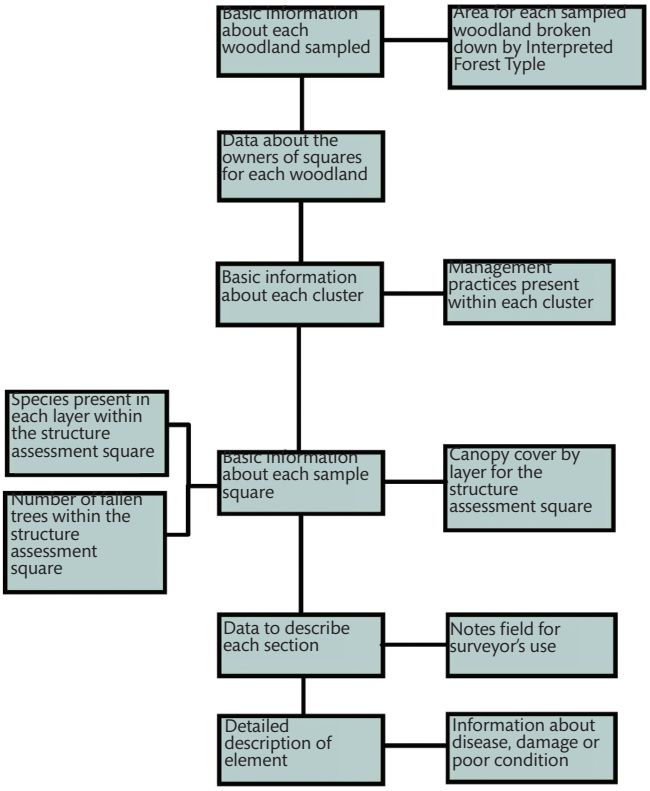
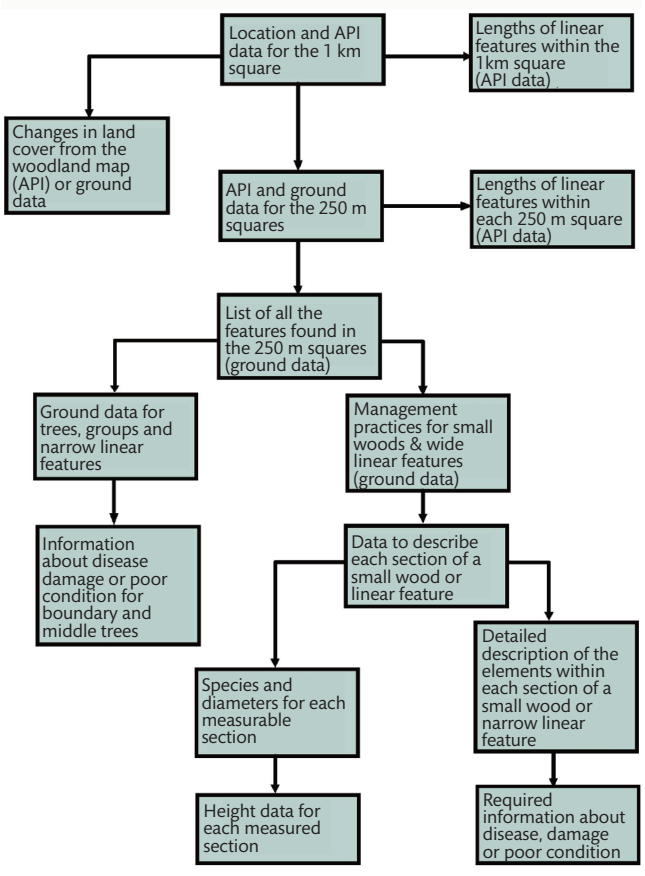


Figure 50 Survey of Small Woodland and Trees – data table structure.



Prior to the information being extracted, checks were made to ensure that:

- the data contained only woods relevant to the appropriate counties or districts
- all the woods for the county or district were accounted for
- each wood had an area on the database that corresponded with the map area
- adjustments were made to 'replacement' woodlands as required for analysis.

Data archiving

The data have been lodged with The National Archive for long-term archiving.

Analysis and production of the results

This section considers a typical Forestry Commission Inventory Report (a list of maps, tables and charts from a typical report is illustrated in Appendix 7), and comments on the analysis of map and field data required to achieve the published results. The headings below follow the Inventory Report headings.

Introduction

Highlights the background of the report and survey methodologies, and notes the main points of the report. Maps of the area are also given.

Summary of results

The first section of a report presents the headline figures for the reporting unit, e.g. region or county, bringing together area data from the Main Woodland Survey and the Survey of Small Woodland and Trees. This enabled the reporting of woodland area to a minimum size of 0.1 hectare. It gave a brief breakdown of the composition of this woodland and a summary of the other tree features in the landscape, i.e. linear features, groups and individual trees.

Assessment of the Main Woodland Survey

The digital map of woodland provided the primary estimate of the woodland area of ≥2 hectares. However, the woodland map was based on aerial photo interpretation and inevitably included some areas that had been incorrectly classified (such as gorse) and were not woodland, and other areas where woodland had appeared on the aerial photographs but had been subsequently converted to other land uses. The field data included information within the sample about these differences, enabling adjustment of the estimate of total woodland area by an appropriate amount. Because the information used to adjust the area was based on a sample it was not possible to adjust the map in the same way. While we knew the location of the samples, this only accounted for 1% (the approximate sampling fraction) of these areas and there was no information on the other 99%.

Standard reports were produced using a reporting and analysis package (SAS Version 6.12) designed to give outputs by 'Forestry Commission', 'other' and 'all ownerships'. These programs, which were written in-house, provided information on the approximate standard errors of the estimates given in the report. The reporting also developed with the progress of the NIWT: early reports for the regions of Scotland were produced when the data for each region was completed. Later, the reports were produced from information for whole countries, giving the opportunity to reconcile the data effectively, prior to publication.

Restricting the data to only those forest types that were included within High Forest (conifer, broadleaved, mixed and windblown plus felled) enabled the production of the species breakdown. Including data on timber potential from the elements meant that the species could be grouped into Category 1 and Category 2 High Forest. An alternative data analysis for Category 1 High Forest produced tables by planting year class.

The principal species for all High Forest were given next. This presentation lists the top three species as defined by the proportion of the area occupied in the planting year class.

The data held within the NIWT for ownership type are distinct from the broad categories of simply 'Forestry Commission' and 'other' and give a more useful breakdown of the different types of owners in the 'other' category. This more detailed information came from a questionnaire that owners or their representatives completed on a voluntary basis.

Assessment of the Survey of Small Woodland and Trees

The sampling scheme broadly aimed to sample 1% of land area in each county of England and Wales, and each region in Scotland. It was hoped that the sampling scheme would generate enough data to support the production of statistics for all the features identified within the survey. This was largely the case except for the small woodlands themselves, and the geographic areas had to be grouped when there were insufficient samples. Within each grouped area, the individual counties or regions were allocated values as a proportion of the whole and there was no distinction between counties in these groups.

Within the Small Woodland Survey, data were summarised in two ways: features were represented as either small areas of woodland or as the number of trees contained in a feature. Linear features would be in both categories, the wide linear features having much in common with woodland. The minimum width used to separate the two categories was 16 m, and the length was collected for both wide linear features and narrow ones (e.g. an avenue).

The analysis of the data was again in two parts and produced information in either hectares or in terms of number of trees. Both produced an estimate of the number of features. The pattern of the data for small woodlands and wide linear features follows the reporting for the main woods with forest type and species.

The features were represented in terms of numbers of trees (boundary trees, middle trees, groups and narrow linear features). Equivalent figures for dead trees recorded in the course of the survey were also given.

For these 'tree' features the emphasis has shifted since the 1980 Census, which regarded them as contributing to the standing volume. However, such timber was often of poor quality with heavy branching and the risk of metal in the butts as a result of use as informal fence posts. To reflect their other values, such as their importance to the landscape, these features are grouped according to size, starting at 2m height, in the tables for individual trees, groups and narrow linear features.

Comparison of results with the 1980 Census and previous surveys

This section of the inventory reports is described in greater detail in Chapter 7.