

SCOTTISH HOUSEHOLD SURVEY

Year 3 User Guide

Contents

1.	Introduction	3
2.	Survey methodology	4
2.1	Background to the survey	4
2.2	Sampling	4
2.3	Survey content	5
2.4	Mode of interview	5
2.5	Number of interviews	5
2.6	Response rates	6
2.7	Technical report	6
3.	Overview of the SHS dataset	7
3.1	Main units of analysis	7
3.2	General variable naming conventions	8
3.3	Missing values	9
3.4	Variables with a high number of missing values	10
3.5	Weights	10
3.6	Variables useful for splitting and linking records	10
3.7	Other useful variables	11
3.8	Anonymised data	11
3.9	Other data issues to note	12
4.	Derived variables	13
4.1	Descriptions of derived variables	13
4.2	Other points to note on using the SHS datasets	17
5.	Income imputation	19
5.1	Income from earnings	19
5.2	Income from benefits	20
5.3	Income from other sources	20
5.4	Summary variables	21
6.	Example programming syntax	22
6.1	Analysing 'looped' variables (e.g. information on people, vehicles, journeys)	22
6.2	Analysing multiple response variables	23
6.3	Analysing random adult data	23

1. Introduction

The Scottish Household Survey (SHS) is a major cross-sectional survey, first commissioned by the Scottish Executive in 1998 to provide reliable and up-to-date information on the composition, characteristics and behaviour of Scottish households, both nationally and at a sub-Scotland level.

From the outset, it was intended that the data should be made publicly available for secondary analysis. The aim of this document is to provide potential users and other interested parties with an overview of the structure of the dataset and guidance on how to undertake basic analysis.

This document has the following structure. Section 2 provides a brief account of the background to the survey and the survey methodology. Section 3 provides an overview of the dataset, Section 4 describes the other derived variables in the dataset and Section 5 explains the way in which missing income data are imputed. Section 6 provides some example SPSS programming syntax to guide analysis. A more detailed account (and a copy of the questionnaire) can be found in 'Scotland's People: Results from the 2001 Scottish Household Survey', Volume 6, Technical Report (Hope, 2002), ISBN 0-7559-0568-7.¹

¹ This can be downloaded from the SHS web site at <http://www.scotland.gov.uk/library5/society/spv6-00.asp>

2. Survey methodology

2.1 Background to the survey

The Scottish Household Survey (SHS) is a continuous survey based on a sample of the general population in private residences in Scotland. It is financed by the Scottish Executive and undertaken by a partnership of NFO System Three Social Research and MORI Scotland.

The aim of the survey is to provide representative information about the composition, characteristics and behaviours of Scottish households, both nationally and at a more local level. The survey covers a wide range of topics to allow links to be made between different policy areas. There is a particular focus on information to inform policy on Transport and Social Inclusion. Results are reported in a series of SHS Bulletins, an annual report (see www.scotland.gov.uk/shs) and some other Scottish Executive publications.

2.2 Sampling

The sample for the survey is designed to meet a number of criteria. It is designed to provide nationally representative samples of private households and of the adult population in private households. This is achieved by splitting the interview between a household respondent and an adult selected at random from the permanent residents of the household.

In order to meet the reporting requirements, the sample is structured to be nationally representative each quarter and to provide a representative sample for larger local authorities each year (those with an achieved sample of over 750 per year).

The sample is also designed to provide data for each local authority over a two-year period. This is achieved by disproportionately sampling in each local authority to achieve a target of at least 550 interviews over two years (equivalent to a simple random sample of 500, with a 95% confidence interval for an estimate of 50% of ± 4 percentage points).

Two different sampling approaches are used. In areas of high population density (Glasgow, Edinburgh, Aberdeen, Dundee, etc.) a simple random sample of dwellings is selected covering the entire two-year sample period. These are subsequently clustered into interviewer allocations. The areas in which a simple random sample is selected are the council areas of Edinburgh, Glasgow, Aberdeen, Dundee, East Dunbartonshire, West Dunbartonshire, Renfrewshire, East Renfrewshire and Inverclyde.

In areas of lower population density, census enumeration districts (EDs) are selected with probability proportionate to population. In each ED, 18 addresses are selected, based on achieving 11 interviews from each sampling point.

The sample is selected from the small user Postal Address File (PAF) for Scotland, expanded to take account of addresses which might only be listed once but actually contain multiple dwellings, such as tenement blocks and multi-storey flats. Although the small user PAF excludes many institutional addresses such as student halls of residence or nurses' homes, there are no geographical exclusions from the survey.

2.3 Survey content

The survey questionnaire is in two parts. A householder or the spouse/partner of a householder completes Part 1 of the interview. Once the composition of the household has been established, one of the adults in the household is randomly selected to complete Part 2.² In all households with a single adult the same person completes both parts but as the number of adults in the household increases, the probability of the random adult being the same as the household respondent declines.³

The section of the interview completed by the household respondent deals with topics such as household composition, housing and tenure, health, the vehicles available to the household, the occupation and industry of the highest income householder, household income and housing costs. The random adult section deals with individuals' housing change, tenure change, neighbourhood problems, transport and use of public transport, public services, income and employment.

2.4 Mode of interview

Interviewing is conducted using Computer Assisted Personal Interviewing (CAPI). Instead of using pen and paper to record responses, data is collected on handheld computers.

2.5 Number of interviews

In 2001, a total of 15,566 valid interviews were carried out with householders. Of these 15,566 households, interviews were achieved with 14,643 random adults (94%).

² Adults who are household members but have been living away for the previous six months are excluded from the selection of the random adult. Children and students living away during term time are counted as household members but are excluded from the random adult and random child selection.

³ Where the same person completes both parts one and two (i.e. they are both the household respondent and selected as the random adult) the CAPI script does not repeat the questions common to both sections so these respondents are not asked for the same information twice.

2.6 Response rates

After excluding addresses that were outwith the scope of the survey⁴, the overall response rate for the third year of the survey was 67%. The nature of the SHS as a survey with continuous fieldwork means that separating the data into fieldwork periods and data files for analysis involves simply establishing cut-off dates and any data submitted to the CAPI servers by a particular date is assigned to one data period or another. In each dataset there are a few interviews that strictly speaking belong to a different data period. For example, in the 2001 data 74 of the 15,566 interviews were carried out on sample issued in 2000. Similarly, there are 93 interviews carried out on 2001 sample that were not on the CAPI servers when the 2001 data file was closed. These interviews are carried into the file for 2002.

When calculating response rates based on the sample issued in 2001, we need to exclude the 74 interviews from the previous sample and include the 93 interviews that will be in the 2002 data. There is, therefore, a difference of 19 interviews between the records in the 2001 data file and the number of interviews achieved on 2001 sample, as shown below.

Interviews in 2001 data file	15,566
minus 74 carried out on 2000 sample	15,492
plus 93 carried into 2002 data	15,585 interviews on 2001 sample
net difference	+19 interviews

The response rate for the 2001 data is based on all addresses issued in 2001 even though some mopping up on 2001 addresses is continuing and all addresses remain valid for two years. The final response rate will be calculated at the end of the 2002 fieldwork period. There was significant variation in response between local authorities. The highest response rate was achieved in the Western Isles (88%) and the lowest response was achieved in Edinburgh and Fife (62%). Post-survey weighting takes account of both the disproportionate sampling between local authorities and the differential response between authorities.

2.7 Technical report

Full details of the survey can be obtained from the companion Technical Report of the survey. This contains full details of the sampling, questionnaire, fieldwork and response rates.

⁴ These are mainly addresses without any private dwellings (such as businesses) and vacant or derelict addresses.

3. Overview of the SHS dataset

3.1 Main units of analysis

The 2001 dataset has been provided as a single flat file, containing 15,566 cases and 2,012 variables. This file was created on 18 November 2002. Any significant changes to the data are noted on the SHS web site and all changes are recorded in the document *Change History.doc* that is supplied as part of the survey documentation.

The basic unit of sampling and analysis for the SHS is the *HOUSEHOLD* and data are collected at this level in relation to the type of property, amenities available, transport available, household composition, working status of the highest income householder, household income, and credit and debt, among other things. When weighted, these data are representative of Scottish households.

The survey also collects information in relation to various other units of analysis. The most important of these are as follows.

HOUSEHOLD MEMBER – Data is collected about each member of the household (up to a maximum of 10) including sex, age, ethnic origin, religious affiliation, economic status, some transport related details, general health, disabilities and care-related facts. Information on all household members is generally derived from analysis of variables such as HA7_1 to HA7_10. These data need to be weighted by LA_WT (see below). The un-weighted base for analysis of all household member data is 35,854.

RANDOM ADULT data (i.e. collected for one adult member of the household, selected at random)⁵ include information on age, sex, other personal characteristics, housing, personal views on local area/ community safety, education, transport, services and local government, health, economic activity and income. When weighted, these data are representative of the Scottish *adult* (i.e. aged 16 plus) population. The un-weighted base for analysis of random adult data is 14,643.

RANDOM SCHOOLCHILD (i.e. collected for one school-aged member of the household, selected at random) data include information on schooling and transport to/from school. When weighted, these data are representative of the Scottish *schoolchild* (i.e. currently attending school) population. The un-weighted base for analysis of random schoolchild data is 3,467.

⁵ 'Random adult' data is missing in records where there was 'non-response' to the 'random adult' part of the interview.

VEHICLE data include information on the type of vehicle, age, annual mileage and fuel costs is available. When weighted, these data are representative of all vehicles owned or regularly used by Scottish households. The un-weighted base for analysis of vehicle data is 14,313.

3.2 General variable naming conventions

Most of the variables on the dataset are derived directly from the answer given to a question in the questionnaire. In many cases – and especially for information relating to either the household, the random adult or the random schoolchild – the variable name will be the same as the question 'number'. So, for example, question HA2 asks who the Highest Income Householder is. Other examples are:

HC4 – Number of bedrooms

HD8 – Number of motor vehicles

RA1 – How long have lived at current address

For questions asked in 'loops' for each household member, vehicle, journey, etc., the variable name is based on the question number, plus an additional number to indicate the household member referred to. For example, question HA5 asks the age of up to 10 household members – the resulting variables are, therefore, named HA5_1 to HA5_10. Other examples are:

HA7_1 to HA7_10 – Economic status

HA9_1 to HA9_10 – Ethnic Origin

There are some questions where the respondent can give more than one answer and the interviewer is instructed to 'code all that apply' or 'code up to (say) three answers'. In such cases there will be a variable for every possible answer. The variable names in these cases will consist of the question number plus a letter to represent the answer – i.e. 'a' for the first possible answer, 'b' for the second possible answer.

An example of where this situation occurs is with question RD6 which asks for main reasons for using method of travel to work/education. There are a range of reasons and each reason translates into a variable. The variables are named RD6A, RD6B, RD6C etc.

Some questions are asked in loops for each household member *and* respondents can give more than one answer. The variable names in these cases will consist of the question number plus a number to indicate the household member referred to plus a letter to represent the answer.

An example of where this situation occurs is with question HF2, which asks what types of health problems or disabilities each household member. In all there are 19 possible codes (1: 'a speech impairment', 2: 'chest or breathing problems', through to 18 'some other health problem or disability' and 19 'Refused') and the interviewer is asked to 'code all that apply'. Each code generates a variable which will be coded '1' for yes and '0' for no. In the case of HF2, the responses for all household members can range across 190 variables – from HF2_1A to HF2_10S.

The dataset does not follow a standard procedure for coding 'Yes/No' variables. For example, for some variables 'no' is coded 0 and for others 'no' is coded 2. However *in all cases, a '1' will represent 'Yes'*. Users should always be sure as to the coding being used.

In addition to the variables that relate directly to questions, there are a number of supplementary and derived variables on the dataset. For example, variables such as household type, property type, and annual net household income have been derived from the information collected by the questionnaire variables. There are also some variables which have been generated by the survey's administrative processes – e.g. each household has a unique identifier and each person has a 'number of person within household' identifier. Such variables are named in a different way, and are listed in Section 4.

3.3 Missing values

In the majority of cases, missing values will be represented by the SPSS 'system missing' identifier – i.e. '.' for numeric variables and blank for character (string) variables. However, there are instances where missing values are coded differently. In some questions, if the respondent doesn't know the answer or can't remember then rather than a system missing being input, a code representing the reason for a missing value is inserted. In the majority of cases where this can happen, the questionnaire will show that the option exists. An example of where this happens regularly is with the income questions, where respondents may refuse to answer, or can't remember how much benefit, say, is received. The following list details all of the missing value codes.

999986	Full housing benefit
999987	No housing benefit
999988	Not able to walk
999989	No trips abroad/ no business mileage
999990	Less than 1000 miles
999991	Less than one year
999992	None
999993	Can't say
999994	No usual pay
999995	Not enough information
999996	Not stated
999997	Can't remember
999998	Don't know
999999	Refused

Note that some of the codes are valid only for certain questions, for example 999990 is a code used for questions such as RE1 and RE3 – number of miles travelled by car in a year.

3.4 Variables with a high number of missing values

Social Class variables have missing values in a large percentage of cases because the 'occupational' information that is needed to derive Social Class is only collected if the person is in employment or has been employed in the past five years.

Similarly, educational qualifications are only collected of people aged 16-65 years.

3.5 Weights

All analyses require weighting, but the correct weight to be used varies for different types of data.

LA_WT – is the weight that adjusts for differences in sampling fractions and response rates between local authorities. This should be used when analysing 'household', 'household member' or 'vehicle' variables. This includes all variables beginning in H (except those from HE6 to HE17) and derived household variables about the household, the HIH or the spouse of the HIH.

IND_WT – contains the individual weight to be used when analysing the Random Adult data (all variables beginning in R and the derived random adult variables).

KID_WT – contains the individual weight to be used when analysing the Random Schoolchild data (variables from questions HE6 to HE17 and the derived random school-child variables).

Where there are no random adult or random child data, the value of the weight will be zero.

3.6 Variables useful for splitting and linking records

The variables which can be used for sorting and splitting the main dataset (for example, to create files based on 'households', 'random adults', etc. are shown below.

	Identified using ...	
Household	UNIQID	
Highest Income Householder	UNIQID	HA2
Random Adult	UNIQID	RANDPEO
Random Schoolchild	UNIQID	KIDPNO

UNIQID – This is the unique reference number given to all households taking part in the SHS.

HA2 – This is the person number of the Highest Income Householder and can be used to generate variables related to the HIH. For example, if HA2 = 3 (person 3 is the HIH) then variables such as HA5_3 and HA7_3 would hold information about, respectively, the age and economic status of the HIH. In many cases this data has already been extracted into a derived variable for ease of use. Thus, HIHAGE and HIHECON can be used instead. These variables are listed in Section 4.

RANDPEO – This is the person number of the Random Adult within the household. As with the HIH, the value of RANDPEO indicates which variables hold the relevant data for the random adult. Similarly, there are a number of derived variables for the random adult.

KIDPNO – This is the person number of the Random Child within the household.

Linking information

Link data from the HIH (Highest Income Householder) using HA2 – the person number of the HIH.

Link data from the HIH's spouse using 'SPNUMO' – the person number of the spouse/partner of the HIH.

Link data from the random person to the household interview using 'RANDPEO' – the person number of the random person.

Link data from the random child section to the household interview using 'KIDPNO' – the person number of the random child.

3.7 Other useful variables

RAND_OK: Indicates whether the record has a valid random adult interview (1='Yes', 2='No').

KID_OK: Indicates whether there is valid random child information (1='Yes', 0='No').

3.8 Anonymised data

In some cases, data for potentially identifying variables are collected during the interview but are not included in the dataset, or are only provided in a broad or summary form. For example, HA4 (date of birth), RD1 (postcode of workplace), HD11 (vehicle registration number) all appear in the questionnaire but are not included in the data set. Similarly, information relating to occupation (from HG21

and RH19) are provided in a broad format in HSOC and RSOC. Data relating to ethnicity and religion are also provided in summary variables instead of the detailed codes collected in the survey.

3.9 Other data issues to note

All questions which ask for reasons (e.g. for liking or disliking a neighbourhood) were originally 'open text' throughout February, March, April, May and June 1999. Answers were listed and coded during these months and then a list of pre-codes drawn up to enable interviewers to enter a code rather than open text. On average, using the pre-code method leads to more reasons being recorded. This should be taken into account when comparing figures with previous datasets. The relevant variables are: HE12, HE13, HE14, HE17, RB2, RB3, RD6, RD8, RD9, RE11, RE12, RG10, RH10 AND RH12.

4. Derived variables

4.1 Descriptions of derived variables

Derived variables are those which are derived from questionnaire variables to enable easier and more meaningful analysis of the datasets. Some relate to the household as a whole, and others to individuals. Definitions of the main classifications used in the SHS can be found in the Glossary document that forms part of the survey documentation. The details of the derivation for individual variables can be provided on request. The following is a full list of the 'administrative' and derived variables currently on the datasets. For information on income variables, see section 5.

Variable Name	Description
<i>Survey Administration Variables</i>	
Uniqid	Unique Household Identifier
Main_ok	Indicates a valid record. Should have a value of 1 in all cases
Day	Day of interview
Month	Month of interview
Quarter	Quarter interview took place in
Year	Year interview took place in
Smonth	Sample month. Addresses are issued on a monthly basis, where an interview did not take place in the month of issue, 'smonth' will not be the same as 'month'.
Syear	Sample year. If an interview did not take place in the year of issue, syear will not be the same as 'year'.
Dyear	Data file the record belongs to. Useful for identifying data files when separate years have been combined.
Dateint	Date of interview with householder (random adult interview might have different date)
R_day	Day of random adult interview
R_month	Month of random adult interview
R_year	Year of random adult interview
Ind_wt	Random Adult weight
La_wt	Local Authority weight
Kid_wt	Random Schoolchild weight
<i>HH variables</i>	
H_SIC	Standard industrial classification for Highest Income Householder

Variable Name	Description
HCLASS	HIH Social class
HIH_eth1	Ethnic group of the HIH (White / non-white)
HIH_rel	Religious affiliation of the HIH
HIH_stat	Marital status of the HIH
Hihage	Age of highest income householder (years)
Hihagebd	Age of highest income householder (banded as 16 to 24; 25 to 34; 35 to 44; 45 to 59; 60 to 74; 75 plus)
Hihecon	Economic status of highest income householder
Hihret	Whether HIH is of retirement age
Hihsex	Sex of highest income householder
HSEG	Socio Economic grouping for HIH
HSOC	Standard occupational classification (1990) for Highest Income Householder
<i>Random adult variables</i>	
Agerband	Banded age of random adult (banded as 16 to 24; 25 to 34; 35 to 44; 45 to 59; 60 to 74; 75 plus)
Fredriv	Frequency of random adult driving
Licence	Whether random adult has driving licence
Rand_ok	Whether a random adult interview was completed successfully
Rand_rel	Religious affiliation of the random adult
Randage	Age of Random adult
Randecon	Economic status of Random Adult
Randeth1	Ethnic group of the random adult (white / non-white)
Randpeo	Random adult person number
Randsex	Sex of random adult
Randstat	Marital status of the random adult
Rclass	Standard industrial classification for Random Adult
RCLASS	Random adult Social Class
RSEG	Socio Economic grouping for Random Adult
RSOC	Standard occupational classification (1990) for Random Adult
Yrsres	Length of residence at current address
<i>Household and other derived variables</i>	
Agerank	Age of the random school child
Area	Local Authority Grouping. Only the five local authorities where there is an achieved sample of 750 interviews or more can be analysed separately in a single year's data. These five authorities (coded 1 to 5) are City of Edinburgh, City of Glasgow, Fife, North Lanarkshire and South Lanarkshire. The groupings of the

Variable Name	Description
	<p>remaining authorities are:</p> <p>Highlands & Islands: Eilean Siar, Argyll & Bute, Highland, Moray, Orkney and Shetland.</p> <p>Grampian: City of Aberdeen, Aberdeenshire</p> <p>Tayside: Angus, Dundee City, Perth & Kinross</p> <p>Central: Stirling, Clackmannanshire, Falkirk</p> <p>Dunbartonshire: West Dunbartonshire, East Dunbartonshire</p> <p>Renfrewshire and Inverclyde: East Renfrewshire, Inverclyde, Renfrewshire</p> <p>Ayrshire: South Ayrshire, East Ayrshire, North Ayrshire</p> <p>Lothian: West Lothian, East Lothian, Midlothian</p> <p>Southern Scotland: Scottish Borders, Dumfries & Galloway</p>
Bedstand	<p>Whether housing fails or meets bedroom standard. The bedroom standard is a measure of occupation density and is used to calculate the minimum number of bedrooms that might be expected to be required by the people resident in a dwelling, taking into account their ages and the nature of their relationships as far as possible. It then compares this number with the number of bedrooms available in the dwellings. The calculation of the number of bedrooms required is based on the assumption that a separate bedroom is required for:</p> <ul style="list-style-type: none"> • each cohabiting couple • any other person aged 21 years or over • each pair of young persons of the same sex aged 10-20 years, and • each pair of children under 10 year (regardless of sex). <p>Unpaired young persons aged 10-20 are paired with a child under 10 of the same sex if possible or allocated a separate bedroom. Any remaining unpaired children under 10 are also allocated a separate bedroom.</p>
Dtime_mi	Drivetime (in minutes) to nearest population centre (with a population of 10,000 or more).
Entarea	Enterprise area (Highlands and Islands Enterprise or Scottish Enterprise areas)
Fyear	Financial year of interview
Hhtype	Household type
Hhwork	Household working status
Kid_ok	Whether successful interview for random child

Variable Name	Description
Kidage	Banded age of Random Schoolchild (4-6, 7-9, 10-12, 13 and above)
Kidecon	Economic status of Random Schoolchild (these are all 'at school')
Kidpno	Person number of Random Schoolchild
Kidsex	Sex of the random school child
Mos00_47	Scottish MOSAIC type – Narrow mosaic codes. Coding changed in 2001
Mosaic00	Scottish MOSAIC grouping - Broad mosaic codes. Coding changed in 2001
Newrural	<p>8-fold urban/rural classification of address.H</p> <p>Using respondents' home postcodes, households have been classified as follows:</p> <p>Large urban areas - households in the city conurbations of Edinburgh, Aberdeen, Dundee, and Glasgow (settlements over 125,000 population).</p> <p>Other urban areas – households in settlements of 10,000 to 125,000 people.</p> <p>Accessible small towns – households in settlements of between 3,000 and 10,000 people and within 30 minutes drive of a settlement of 10,000 or more.</p> <p>*Remote small towns – small towns (between 3,000 and 10,000 people) within a drive time of between 30 and 60 minutes of a settlement of 10,000 or more.</p> <p>*Very remote small towns – small towns (between 3,000 and 10,000 people) over 60 minutes drive of a settlement of 10,000 or more.</p> <p>Accessible rural - households in settlements of less than 3,000 people and within 30 minutes drive of a settlement of 10,000 or more.</p> <p>**Remote rural – households in settlements of less than 3,000 people and within a drive time of between 30 and 60 minutes of a settlement of 10,000 or more.</p> <p>**Very remote rural - households in settlements of less than 3,000 people, over 60 minutes drive of a settlement of 10,000 or more.</p> <p>Isolated houses and hamlets are included in settlements of less than 3,000 people.</p>
Numads	Number of eligible adults (aged over 16 and meet residence criteria)
Numbhh	Total number of people in household from ha1
Numcars	Number of cars household has access to
Numkids	Number of eligible schoolchildren (live at home, at school and the responsibility of the respondent)
Numret	Number of retired people in household
Numveh	Number of motor vehicles available to the household
Proptype	Property type

Variable Name	Description
Respcho	Number of children rep/partner responsible for
Shs_6cla	6-fold urban/rural classification of address. See Newrural. Collapsing the categories remote and very remote small towns and remote and very remote rural areas would provide the 6-fold classification.
SIP	Whether address is in a SIP area
Sp_eth1	Ethnic group of the HIH spouse / partner (white / non-white)
Sp_rel	Religious affiliation of the HIH spouse / partner
Sp_stat	Marital status of the HIH spouse / partner
Spage	Age of HIH spouse / partner
Specon	Economic status of the HIH spouse / partner
Spnumo	Person number of spouse
Spret	Whether spouse/partner is of retirement age
Ssex	Sex of spouse/partner
Tenure	Tenure
Totads	Total number of adults (whether eligible to be random adult or not)
Totkids	Total number of children (whether eligible to be the random child or not)

4.2 Other points to note on using the SHS datasets

The remainder of this Section covers a number of other points that should be kept in mind when using the SHS dataset.

4.2.1 Unemployment rates, average earnings figures, and other statistics

The SHS was *not* designed to collect reliable statistics on topics such as unemployment rates, average earnings, income and benefits. The SHS has questions on such topics *only* for selecting the data for particular groups of people (such as the unemployed or the low-paid) for further analysis, or for use as 'background' variables when analysing other topics (such as the means of travel or the frequency of driving).

4.2.2 Analysis of data for a particular period, for Scotland and for areas within Scotland

The SHS's design is such that

1. the sample for a *quarter* should be representative of Scotland as a whole
2. the samples for a *calendar year* are representative for certain Council areas (those with more than 750 interviews: Edinburgh, Fife, Glasgow, North Lanarkshire and South Lanarkshire) – see entry on 'Area' in Section 4.1.

3. the samples for a *two-year sweep* (eg 1999-2000, 2001-2002) are representative for all Councils, regardless of size.

Therefore, one should *not* use monthly figures for Scotland. Also, because of the sensitivity of providing Council-level data of an inadequate size to allow analysis in most cases, the Council identifier variables have been removed. Only the five Councils with more than 750 interviews can be analysed separately by using the AREA variable.

4.2.3 Statistics that can be produced from different variables within the SHS

Some statistics can be produced from either the data on all household members or from the 'Random Adult' data (e.g. information about driving licences). In such cases, the figures could well differ because of sampling variability (the 'all household member' data has a response for every adult in the interviewed households, whereas the 'Random Adult' data has one response per household) and the additional effects of non-response by some of the randomly-chosen adults. The figures from the 'household member' data might be considered the more reliable, but some people may prefer to use the 'Random Adult' data, for consistency with the other statistics about adults that are only available from that source. Therefore, it is important to specify which data were used when providing figures that could have come from either source.

Similarly, some statistics about schoolchildren could be produced from either the 'household member' data (everyone coded 7 at HA7_1 to HA7_10) or the 'Random Schoolchild' data. In such cases, one should specify which unit of analysis was used.

Finally, statistics produced from the data for different periods may differ as a result of, for example, sampling variability, seasonal variation and other changes with time. It is therefore important to specify which quarters' data were used when citing data from the SHS.

4.2.4 Sample numbers

It is usually best to give the sample numbers which are the basis of the results that you are reporting (see, for example, the 'Base' lines in the tables in the bulletins or Annual Report). In such cases, the normal practice is to give the *unweighted* sample numbers.

5. Income imputation

In the SHS, total net income is the primary indicator of household income. Total net income is defined as the total income from earnings, benefits, and a variety of miscellaneous sources of the Highest Income Householder and their spouse, where applicable. Each component - income from earnings, from benefits and from other sources – is collected separately.

Income data is also collected from the random adult, where this has not already been collected as part of the HIH or spouse income. However, no imputation is carried out for the random adult because of the difficulty of accurately attributing benefit data to individuals. Where the random adult is the HIH or the spouse of the HIH, summary variables are created for the random adult's income from their main job (RINCMINC) and from other sources (RINCOINC) after imputation of household income.

Incomplete data resulted in around 34% of households having no computed total net income. Moreover, missing income data was not distributed evenly through the SHS. Imputation was carried out for the individual components of income in order that total net household income could be calculated. The principal methods used were 'hot deck imputation', where the sample is divided into subgroups (imputation classes) based on the relevant characteristics, and 'predictive mean modelling', where a statistical model is constructed to provide an estimate.

After imputation, 2.5% of cases are still missing information on income. This residue comprises two groups which are roughly equal in size:

- Households where the HIH stated that both themselves and their partners were neither working, nor receiving any benefits, and were not receiving any miscellaneous source of income.
- Households that, after imputation, had total net annual income of less than £25 a week. In these households, it seems likely that the SHS is not picking up their sources of income or that their income at the time of interview was atypical.

5.1 Income from earnings

Income from earnings was collected for the main job and for other jobs of the highest income household and their partner where applicable. Income was imputed separated for each of these components. For the imputation of income from main jobs, predictive mean modelling was employed. The models used the following:

- Age, sex and SEG of head of household

- Whether the work was full or part-time, self-employed or not,
- Car ownership, whether living in rented accommodation, computer ownership, receipt of means-tested benefits
- Whether respondent lives in a remote location.

A smaller number of individuals lack information on their second and subsequent jobs. These were imputed using hot deck imputation with the imputation classes based on, age of HIH/partner of HIH, sex, and whether self-employed or employed.

5.2 Income from benefits

Imputation was carried out on each benefit separately, where possible.

For earnings top-up, maternity allowance, statutory maternity pay, widow's pension benefit, disability working allowance, industrial injury benefit, invalid care allowance, statutory sick pay, war disablement pension, IS/HB disability premium, other disability benefit, and other state benefit, the median amount received for the benefit was imputed. These benefits are either flat-rate benefits or were received by too few people to allow modelling.

Child benefit, state retirement pension, disability living allowance, severe disablement benefit, attendance allowance, incapacity benefit, jobseeker's allowance, family credit, and council tax benefit were imputed using hot deck imputation with the imputation classes based on relevant household characteristics.

Imputation of housing benefit was done using a multi-stage approach. For a large minority of those missing information on amount received, the respondent provided the rent which they paid after housing benefit, and the housing benefit was derived as the difference between that and the imputed gross rent. The remaining cases had housing benefit imputed using tenure, receipt of income support, number of bedrooms, banded income from earnings, and age of the HIH.

Missing data for income support was imputed last of all in order to utilise the other imputed information. The hot deck imputation classes were based on level of income and eligibility for income support. This was based on a simplified model of how much respondents were eligible to receive and on income they received from earnings and other benefits.

5.3 Income from other sources

For income from maintenance, annuity/trust, rent from property, dig money, sickness pay, student loan, student grant, and other regular non-work sources, the median amount received was imputed. These sources of income were received by too few people to allow modelling.

Investment income and income from a non-state pension were imputed using hot deck imputation with the imputation classes based on characteristics which were correlated to the amount received.

5.4 Summary variables

For each individual component of income, there is a variable detailing the amount received and an associated summary variable, indicating if the value was given by the respondent or if it was imputed. Similar pairs of summary variables have also been calculated for income from earnings, income from benefits, income from other sources, and total household income. These imputation flags can be used to exclude imputed data from analyses that might be sensitive to the imputation procedures.

Variable	Summary variable	Description
HINCMINC	HINCMSUM	Income from earnings of HIH's main job
HINCOINC	HINCOSUM	Income from earnings of HIH's other jobs
SINCMINC	SINCMSUM	Income from earnings of partner's main job
SINCOINC	SINCOSUM	Income from earnings of partner's other jobs
EARNINC	EARNSUM	Total income from earnings (HIH and partner)
BENINC	BENSUM	Total income from benefits
MSCINC	MSCSUM	Total income from other sources
ANNETINC	INCSUM	Total net annual income
WKNETINC	-	Weekly net income
BANDINC	-	Total net annual income banded

The summary variable values are as follows:

1 = Yes, correct – Income is received from this source and an amount was given by the respondent

2 = Imputed – Income was received from this source, but an amount was not given so the amount was imputed.

3 = Yes but did not use – Income was received but the amount given was very high or low so this was not used in the imputation.

4 = No, not correct – There was an amount given for this income even though the respondent said they did not receive it. This value was set to zero.

5 = No, missing – The respondent received income from this source but gave no amount.

6 = No, correct – The respondent did not receive income from this source.

6. Example programming syntax

In this section, we provide some example programming syntax to guide some of the more common but complex forms of analysis that users may wish to undertake.

6.1 Analysing 'looped' variables (e.g. information on people, vehicles, journeys)

As indicated earlier, the basic unit of analysis in the SHS is the household, but it is also possible to carry out analysis based, for example, on all people within the household, vehicles belonging to the household, etc. The easiest way to do this within the 'flat file' structure of the data is to use a 'loop' within SPSS. An example of this – which counts the number of cars the household has access to – is shown below.

```
*****number of cars*****.
vector cars = hd9_1 to hd9_10 .
compute numcars = 0 .
loop #i = 1 to 10 .
    if cars(#i) = 1 numcars = numcars+1 .
end loop .
recode numcars (3 thru 7=3) .
variable labels numcars 'Number of cars household has access to'.
value labels numcars 0 'None' 1 'One' 2 'Two' 3 'Three or more' .
```

An example of programming syntax to examine the characteristics of all household members is shown below.

```
****household population characteristics****

recode ha5_1 to ha5_10 (0 thru 15=1) (16 thru 24=2) (25 thru 34=3) (35 thru 44=4) (45 thru
54=5)
(55 thru 64=6) (65 thru 74=7) (75 thru hi=8) into agebd1 to agebd10.
value labels agebd1 to agebd10 1 "0-15" 2 "16-24" 3 "25-34" 4 "35-44" 5 "45-54" 6 "55-64"
7 "65-74" 8 "75+".

weight by la_wt.
MULT RESPONSE
GROUPS=$sex 'Sex of household members' (ha6_1 ha6_10 ha6_2 ha6_3 ha6_4
ha6_5 ha6_6 ha6_7 ha6_8 ha6_9 (1,2)) $ethnic 'Ethnicity of household'+
' members' (ha9_1 ha9_10 ha9_2 ha9_3 ha9_4 ha9_5 ha9_6 ha9_7 ha9_8 ha9_9 (1
,99)) $ages 'Age of household members' (agebd1 agebd10 agebd2 agebd3
agebd4 agebd5 agebd6 agebd7 agebd8 agebd9 (1,99))
```

```
/FREQUENCIES=$sex $ages $ethnic.
```

6.2 Analysing multiple response variables

To look at multiple response variables, you need to specify all the variables in the multiple response set (e.g. HC5A to HC5D) in the 'define sets' menu (through Analyse/Multiple Response in SPSS menus). You need to specify a code for the responses (values of 1), a variable label and description.

The newly defined variable can then be used in the Multiple Response/ 'Frequencies' or 'Crosstabs' command. An example of programming syntax defining the multiple response and for running commands is shown below:

```
MULT RESPONSE GROUPS=$hc5all 'HC5 - goods owned by household' (hc5a hc5b hc5c hc5d (1)) /FREQUENCIES=$hc5all .
```

```
MULT RESPONSE GROUPS=$hc5all 'HC5 - goods owned by household' (hc5a hc5b hc5c hc5d (1)) /VARIABLES=hhtype(1 99) /TABLES=$hc5all BY hhtype /BASE=CASES .
```

6.3 Analysing random adult data

An example of programming syntax to examine the marital status of adults in the random adult data, is shown below. You need to specify that only cases with complete random adult data are included in the analysis (if rand_ok=1).

```
compute randwed = 0.  
if (randpeo = 1) randwed = ha8_1.  
if (randpeo = 2) randwed = ha8_2.  
if (randpeo = 3) randwed = ha8_3.  
if (randpeo = 4) randwed = ha8_4.  
if (randpeo = 5) randwed = ha8_5.  
if (randpeo = 6) randwed = ha8_6.  
if (randpeo = 7) randwed = ha8_7.  
if (randpeo = 8) randwed = ha8_8.  
if (randpeo = 9) randwed = ha8_9.  
if (randpeo = 10) randwed = ha8_10.
```

```
variable labels randwed 'Marital status of random adult'.
```

```
value labels randwed 1 'Married'  
                2 'Cohabiting'  
                3 'Single'  
                4 'Widowed'  
                5 'Divorced'  
                6 'Separated'.
```

```
temp.  
sel if rand_ok=1.  
fre randwed .
```