Family Resources Survey

RELEASES 1999-00

RELEASE	CHANGES SINCE LAST RELEASE	RELEASE DATE
frs990a	**FIRST RELEASE**	20/09/2000
frs990b	**SECOND RELEASE** Due to problems pointed out during user testing the following DV's have been rerun: • TOTHOURS • JOBHOURS • CARE DV's • DEPBAND (Still based on 1998/9 deprivation ranks) • UGRSPAY • MORTINT • CWATAMTD • NDDCTB • GROSSCT Plus any dependant DV's.	09/02/2001
frs990c	**THIRD RELEASE** The hierarchical data has not changed, this release is a change to the flatfile only. Payments from annuity pensions were incorrectly mapped to trust funds or missed from the flatfile. 88% (£12,603) of the payments from annuity pensions were incorrectly mapped to trust funds. The remaining 12% (£1707) plus all trust fund payments (£7,375) were previously missed from the flatfile.	02/03/01
frs9900d	**FOURTH RELEASE** The changes are in the following areas: **All investment/total income derived variables due to ISA.** - Interest from ISAs has been added to the two investment income DVs (ININV and NININV). **Grossing factors.** - Revised grossing factors.	27/03/01
frs9900e	**FIFTH RELEASE** **Interim Grossing factor added**	25/11/02

	- See paper for more details	
frs9900f	**SIXTH RELEASE** Misleading Deprivation Band Indicator (DEPBAND) removed for non-English Local Authorities. See 2002-03 Changes documentation for full details. Correction made to usual Gross pay to stop double counting deductions. (UGRSPAY). See 2002-03 Changes documentation for full details. In 1996-97 the derived variable for specific household tenure types (TENTYPE) was changed from ten categories to eight. The format attached to this variable was not updated. This has now been corrected for all affected years. Family Type (FAMTHBAI) definition adjusted to be in line with HBAI definition introduced in 2001-02. See 2002-03 Changes documentation for full details.	24/11/03
Frs9900g	**SEVENTH RELEASE** New Grossing regime (GROSS3) introduced - See paper for more details DV ININV – Income from Investments – has been corrected for where the code was excluding income from GILT Edged Stock when income reported after tax.	22/11/04
Frs9900h	Revised weights issued for the new Grossing regime (GROSS3).	27/01/05
Frs9900i	Revised weights issued for the new Grossing regime (GROSS3) to correct for overestimation of the Lone Parent population control.	09/02/2005

FAMILY RESOURCES SURVEY 1999/00:

SUMMARY OF EDITING AND IMPUTATION PROCEDURES CARRIED OUT BY DSS

For the 1999/00 data set, the following tasks were carried out by DSS.

1 **Conversion of monetary amounts to weekly values**

Many of the questions on the FRS ask for amounts received/paid and to what period they relate (eg benefit receipt, council tax payments). In these cases, amounts were converted to weekly equivalents. More information on which period code relates to which value is given in the Excel spreadsheet period35.xls.

1.1 During the conversion process amounts were not converted where:

1.1.1 payments were one off or lump sum payments (period code 95)
1.1.2 "none of the above" (period code 97)
1.1.3 period code missing
1.1.4 payments were less than 1 week (period code 90)

1.2 However, for those items of income and expenditure which feed in to derived variables used by the DSS, missing, 90, 95 and 97 period code payments were scrutinised and edited to a weekly value. Remaining 90, 95 and 97 period codes will appear in analyses as outliers. Users will need to consider whether to edit or delete these cases. The easiest way to identify such variables is to consult minmaxan.xls and search on maximum values of 95 or 97. The link between period codes and monetary amounts is given in period34.xls.

2 **Validation, editing and imputation**

Information about procedures carried out by DSS are contained in the file methodology chapter of the latest FRS publication.

3 **Anonymisation**

1.2 ONS/National Centre for Social Research have their own procedures to ensure the confidentiality of respondents. Names and addresses are kept separately from the data and are not supplied to the DSS.

1.3 Additional steps have been taken by the DSS prior to release of the data outside the department. These are:

1.3.1 The following variables have been removed from the data set:

Variable	**Table**
Acorn	Househol
Grossct	Househol
Lac	Househol
Nindinc	Adult
Ninearns	Adult
Nininv	Adult
Ninpenin	Adult
DOB	Adult
DOB	Child

1.3.2 Monetary amounts relating to council tax variables have been rounded to whole pounds. Variables affected are:

Variable	**Description**	**Table**
ctamt	last CT payment	househol
ctrebamt	amount of CT rebate	househol
ctredamt	amount of transitional reduction	househol

cwatamt	amount included in rent for CT water charge	househol
indinc	Derived Variable (DV) for adult income	adult
inrpinc	DV for adult RP/IS income	adult
indisben	DV for adult disability benefit income	adult
inirben	DV for adult income related benefit income	adult
innirben	DV for adult non-income related benefit income	adult
inothben	DV for adult other benefits	adult
buinc	DV for benefit unit income	benunit
burpinc	DV for benefit unit RP/IS income	benunit
budisben	DV for benefit unit disability benefit income	benunit
buirben	DV for benefit unit income related benefit income	benunit
bunirben	DV for BU non-income related benefit income	benunit
buothben	DV for BU other benefit income	benunit
hhinc	DV for household income	househol
hhrpinc	DV for HH RP/IS income	househol
hhdisben	DV for HH disability benefit income	househol
hhirben	DV for HH income related benefit income	househol
hhnirben	DV for HH non-income related benefit income	househol
hhothben	DV for HH other benefit income	househol
hbeninc	DV for HH benefit income	househol
cwatamtd	DV for council tax water charge	househol
burent	DV for BU rent	benunit
hhrent	DV for HH rent	househol
hscosthh	DV for HH housing costs	househol

1.4 However, assurances given to interviewees allow DSS to provide unanonymised data in very restricted circumstances. For more information, please contact Angela White at the address given below.

ASD3E
Analytical Services Division
Department of Social Security
4th Floor
The Adelphi
1-11 John Adam Street
London WC2N 6HT

IMPUTATION OF MISSING VALUES IN THE 1999/2000 FAMILY RESOURCES SURVEY

INTRODUCTION

Imputation is the process in which missing values in a data set are converted to non-missing values.

When a respondent answers a particular question in a survey they can state that they don't know the answer to a question, or simply refuse to give a response. Such responses are recorded and are referred to as *'missing values'*.

These values can either be left as missing, in which case you would have gaps in your data set, or replaced (*imputed*) with an estimate of the answer that the respondent would have given if they had actually answered the question.

User requirements have deemed the latter process necessary in the Family Resources Survey (FRS). The main objective of imputation is to maximise the information available to users for analysis. Furthermore, the imputation carried out simplifies the analysis for users and helps to secure the uniformity of analysis created from the FRS data sets.

It should be noted that none of the variables in the admin, benefits and care data sets are imputed and that benefit editing is carried out separately to the rest of imputation.

Methodology

Imputation on the FRS has traditionally been carried out in four different ways. A brief overview of these methods is given here:

- **Bulk edits** – converting en masse a batch of cases with missing values that satisfy a particular characteristic to an identical value. This is a very crude method of imputation and can only be used in certain circumstances. For example, for people who don't know if they are in receipt of a particular benefit, we could:

    i) edit the answers to yes, in which case we would have to open up a record for the particular benefit and impute answers for it

    ii) edit such answers to no – which is known as *closing down routes* and is the default principle adopted in the imputation of such *routing* variables in the FRS.

- **Hotdecks** – examining the data set for non-missing cases which have similar characteristics to that with the missing value, and substituting one of these non-missing values for the missing case at random. It is usual for the characteristics to bear some

relationship to the variable to be imputed, the theory being that all cases matching the chosen characteristics will have similar values for the variable we are concerned with. For example we could impute rent for a household by randomly selecting a non-missing value from a case with the same number of rooms, council tax band, type of landlord and region as the case in question.

- **Algorithms** – a process in which one can predict the missing value for a particular case by looking at other relevant characteristics and applying a pre determined set of rules (e.g. modelling council tax payments based on council tax band, local authority and entitlement to discount).

- **Neural Networks** - Neural networks are information processing systems that learn by example, recognising patterns in data. Their main advantage over standard statistical techniques is that they can extract and model non-linear relationships without assuming any particular underlying distribution. This method was previously used in the FRS, but its use has been discontinued this year (see below).

Missing Values

There are four possible types of missing values in the FRS:

- **.A** – denotes a 'skipped' response. Such a reponse occurs where a respondent has not been routed to this particular question and an answer is not therefore required and imputation is not normally necessary.
- **.B** – denotes the fact that the respondent *'doesn't know'* the answer to the question and imputation will normally be required.
- **.C** – denotes a refusal to answer a question and, again, imputation is normally required.
- **.D** - is only output in the production of derived variables, and denotes either a mistake in the imputation process or faulty logic in the DV code. All .Ds in income and expenditure data are investigated and removed from the data set prior to user release.

Imputation Checking

Checks are carried to ensure that the imputation process has not changed the distribution of the data. Examples of these are as follows:

- A comparison of the means, standard deviations and minimum/maximum values for each variable is undertaken both post and prior imputation. Any large discrepancies (indicating that imputation is potentially biasing the data) are investigated.

- There can be cases in Hotdecks where we impute a large number of cases to a particular value, which is taken from one particular 'donor' case. This is a source of potential bias, and checks exist within hotdecks to monitor this. Where these checks show this to be a problem, remedial action, in the form of adjusting either the imputed value or the hotdeck, is taken.

- Finally credibility checks are run, which ensure that the data within individual cases is consistent, and feasible values have been imputed. Examples of these include:

    i) Checking that housing costs are generally less than income for cases in which components of either have been imputed.

    ii) Checking that gross income is greater than or equal to net income.

    iii) Checking that personal pension contributions are generally less than income for cases where components of either have been imputed.

Changes to Imputation 1999/2000

The Neural Networks method imputes the distribution mean when it can't successfully extract and model non-linear relationships. As Neural networks take a lot of time to set up and run, and it was discovered that the mean was in fact being imputed in the majority of cases there has been a strategy of gradually phasing out imputation by neural networks over the past few years. In 1999-2000 they were phased out completely and have been replaced by other hotdeck methods. Standard checks of imputation (see above) have not indicated that this policy has adversely affected the data in any way.

Hotdecking has been improved by the introduction of diagnostic checks and by introducing increased precision in the process that matches the characteristics of target (missing) and donor (non missing) cases for the variable in question.

The imputation process has also been almost fully automated, which has resulted in a considerable improvement in the timeliness of data delivery as well as reducing the risk of errors.

Tables of Results

Table 1 provides an overall summary of imputation outlining the number of missing values initially and how many were imputed by each method. It also provides a comparison with previous years. It should be noted that hotdecking is the most common method of imputation, followed by bulk edits, the use of both these methods has increased over the last three years as neural network imputation was phased out.

Table 1 also shows a slight increase in the number of missing values and the explanation for this is as follows:

- As with any questionnaire, a typical feature of the FRS is the gatekeeper question positioned at the top of a block of further questions, at which a particular response will open up the block. If the gatekeeper question itself is answered as 'don't know' or 'refused', the block contains skipped values for all variables within it.

- A missing gatekeeper variable could be imputed such that a further series of answers would be expected. However, these answers will not appear because a whole new route has been opened. For example, if the amount of rent is missing for a record and has since been imputed, any further questions about rent would not have been asked. From the post-imputed database, it will appear that these questions should have been asked because a value is there for rent.

- The above has been the case for most blocks of questions in previous years, however in 1999-2000 the decision was taken to further open up routes for some variables (such as did rent include housing benefit, or did your last mortgage interest payment include the endowment premium). This has resulted in around 8000 cases being converted from skipped to missing and then imputed in the normal way. These cases have been included in the analysis in Table 1 creating a one-off step increase in the number of missing values recorded.

Table 1: Summary of imputation in FRS 1997-98 to 1999-2000

	1997-98		1998-99		1999-2000	
	Values	Percentage of values	Values	Percentage of values	Values	Percentage of values
Responses						
Expected number of responses	11,671,803	100	11,496,549	100	11,938,060	100
Valid responses	11,620,954	99.6	11,451,417	99.6	11,880,641	99.5
Missing values (don't know / refused)	50,849	0.4	45,132	0.4	57,419	0.5
Treatment of missing values						
Hotdeck	25,977	51	27,782	62	41,220	72
Neural Networks	10,640	21	5,129	11	0	0
Bulk Edits	2,651	5	3,364	7	9,638	17
Other imputation method	5,510	11	2,995	7	582	1
Benefit editing	1,288	3	1,540	3	1,701	3
Left as Missing	4,783	9	4,322	10	4,278	7

Table 2 shows the 30 variables with the highest number of imputed values. As usual, the worst offenders are assets and employment/self-employment income.

Table 2: The 30 variables with the most imputed values

Table	Variable	Number of Imputed values
ACCOUNTS	ACCINT	12097
ACCOUNTS	ACCTAX	6757
ASSETS	HOWMUCH	2303
BENEFITS	BENAMT	1559
ADULT	KEEPPEN	1479
JOB	PAYAMT	1185
JOB	NATINS	1137
JOB	PAYE	1132
HOUSEHOL	STRAMT1	1022
HOUSEHOL	STRAMT2	900
HOUSEHOL	WSEWAMT	766
RENTER	WSINCAMT	762
ASSETS	HOWMANY	722
JOB	PROFTAX	661
JOB	PROFIT2	649
JOB	PROFIT1	648
INSURANC	POLAMT	576
ENDOWMNT	MENPOLAM	553
PENSION	PENPAY	508
MORTGAGE	MORTLEFT	506
ADULT	EPCUR	458
JOB	SETAXAMT	416
ADULT	PPPAY1	351
JOB	DEDUC1	347
JOB	SEINCAMT	332
PENSION	PTAMT	326
ADULT	PPREBDSS	313
JOB	SENIIAMT	300
ADULT	PPREB	295
JOB	OTHDED1	275