FAMILY RESOURCES SURVEY 2004-05

SUMMARY OF EDITING AND IMPUTATION PROCEDURES CARRIED OUT BY DWP

For the 2004-05 dataset, the following tasks were carried out by DWP.

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 (e.g. 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 period code.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 DWP, 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 minmaxmean.xls and search on maximum values of 95 or 97. The link between period codes and monetary amounts is given in period code.xls

1.3 Note that in the dataset period codes shown as –1 (skipped) have an imputed weekly amount attached.

2 Validation, editing and imputation

Information about procedures carried out by DWP is contained in the Methodology chapter of the latest FRS publication.

.

Family Resources Survey

RELEASES 2004-05

RELEASE	CHANGES SINCE LAST RELEASE	RELEASE DATE
Frs0405a	**RESTRICTED RELEASE**	21/11/05
Frs0405b	**PUBLIC RELEASE** TEA corrected to include the number of years anyone over the age of 18 has spent in full time education. CHLOOK 01-09 corrected to include unpaid childcare Adjustment made to re-mortgage (RMAMT) for two cases. Added Fixed Savings Bonds and 'Other deductions from PENPAY' to flat-file Edits to tax credit values for ten cases Correction made to the previously corrupted ACORNEW data ADULTS who may become eligible for Child Benefit under the new 2006-07 rules (CHBFLG) has been revised to include "NDDP" and "Any other training schemes" Cost of weekly travel to work (TTWCOST) amended to include taxi fares Edits following introduction of two checks to validate the data against the metadata; SIC variables for three cases to valid values; updated formats	30/03/06
Frs0405c	Correction made to the previous NIRATE , NINRV data Rural/ Urban indicators added to the dataset Missing "Follow-up" data now incorporated	30/03/2007

Frs0405d	Correction made to the control totals of 16 to 19 yrs old Adult and Child population	17/04/2007
Frs0405e	Correction to HHRENT DV (and its dependants) due to miscalculation of NI households (rates being deducted from rent more than once in multi-adult households).	27/06/07
Frs0405f	New Grossing regime (GROSS4) introduced - See paper for more details https://www.gov.uk/government/publications/family-resources-survey-grossing-methodology-review-and-2011-census-updates	01/07/14

IMPUTATION OF MISSING VALUES IN THE 2004-05 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 is carried out in three 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 that 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).

Missing Values

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

- **.A** – denotes a 'skipped' response. Such a response 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 corrected prior to user release.

Imputation Checking

Checks are carried out 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 t han 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.

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 the previous year. It should be noted that hotdecking is the most common method of imputation, followed by bulk edits.

- 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.

Table 2 shows the extent of imputation on the BENEFITS table. Each benefit type is listed by variable, showing the number of expected responses, the number and percentage imputed and the number left missing. Each benefit is listed on the first sheet although this is repeated for each benefit on the subsequent sheets.

Table 3 shows the extent of imputation on all tables. Each variable is listed with the number of expected responses, the number and percentage imputed and the number left missing. Apart from the BENEFITS table, where any variable has had a missing imputed all missing values for that variable will have been imputed. There is a zero in cell A1 on each sheet that will display as a positive number if this were not the case (as on the BENEFITS table).