ENHANCING AND ADDING VALUE TO RESEARCH DATA

LAURENCE HORTON
Research Data Management Senior Officer
UK DATA ARCHIVE
UNIVERSITY OF ESSEX

NCRM RESEARCH METHODS FESTIVAL 2010
6 JULY, 2010
KINDS OF DATA UK DATA ARCHIVE DEALS WITH

- **Quantitative**
  - micro data
  - macro data
  - formats: SPSS, Stata and tab delimited formats

- **Qualitative**
  - in-depth interviews, diaries, anthropological field notes
  - complete answers to survey questions
  - formats: Excel, Word and RTF

- **Multimedia**
  - image files, such as photographs, and audio files
  - formats: WAV, Mp3, JPEG, TIFF

- **Non-digital material**
  - photographs, reports, questionnaires and transcriptions
  - formats: paper, analogue audio or audio-visual recordings
WHAT DO USERS DO WITH DATA?

- Use descriptive material
- Comparative research, restudy or follow-up study
- Re-analysis/secondary analysis
- Research design and methodological advancement
- Replication of published statistics
- Teaching and learning
“The costs of archiving activities (archival storage and preservation planning and actions) are consistently a very small proportion of the overall costs and significantly lower than the costs of acquisition/ingest or access activities…”

N. Beagrie et.al *Keeping Research Data Safe 2*, p.4
PROVIDING GOOD DOCUMENTATION

• Quality of the information provided by the data creator determines ease of discovery and appropriate re-use
  • comprehensive and comprehensible documentation and metadata
  • complete the deposit form as fully as possible
EXAMPLES

- Quantitative dataset
  - Questionnaire, variable list, codebook

- Qualitative dataset – depends on size and scale
  - data listing, interview schedules

End of Award reports, publications list, methodology
QUANTITATIVE DOCUMENTATION

- smaller-scale study - user guide may just contain survey questionnaire, methodology information

<table>
<thead>
<tr>
<th>Documentation:</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Format</strong></td>
</tr>
<tr>
<td>PDF</td>
</tr>
<tr>
<td>PDF</td>
</tr>
<tr>
<td>PDF</td>
</tr>
<tr>
<td>PDF</td>
</tr>
<tr>
<td>HTML</td>
</tr>
</tbody>
</table>
Introduction

This document is the most sensible starting point to analysing the HSE data, as it categorises all the variables across the dataset into two levels, and it is therefore easier to see the coverage of questions asked at this summary level, rather than ploughing straight into the documentation of the questionnaires and self-completion booklets.

Once you have found the appropriate variables that you want to analyse, you then need to look at the other documentation to see in more detail exactly how the question was asked in the study, or how a derived variable has been defined.

The source of each variable is indicated in the final column of each table of variables with abbreviations as follows:

<table>
<thead>
<tr>
<th>Source</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>HHid</td>
<td>Household CAPI Questionnaire</td>
</tr>
<tr>
<td>Indiv</td>
<td>Individual CAPI Questionnaire</td>
</tr>
<tr>
<td>Nurse</td>
<td>Nurse CAPI Questionnaire</td>
</tr>
<tr>
<td>SC</td>
<td>Self-Completion Booklet: SC 6-10, SC 11-22, SC 23-45, SC 46-7, SC Adult, SC Glycyr, SC FT 4.15, SC Rating Habits or whose a question appears in more than one booklet the range is widened (eg SC8-13, SC 16+)</td>
</tr>
<tr>
<td>Lab</td>
<td>Results from laboratory, ie from saliva, serum or urine testing</td>
</tr>
<tr>
<td>ABE</td>
<td>Address Booklet completed for each issued solution</td>
</tr>
<tr>
<td>NSF</td>
<td>Nurse Booklet completed for each household where at least one person had agreed to a nurse interview</td>
</tr>
<tr>
<td>Derived</td>
<td>A variable derived from other variables, and detailed in the Derived Variable Specification document</td>
</tr>
</tbody>
</table>
The Health Survey for England 2007

Program Documentation

Household Questionnaire

First

SAMPLE POINT NUMBER
Zange: 1.097

Address

ADDRESS NUMBER
Zange: 1.97

Household

HOUSEHOLD NUMBER
Zange: 1.9

Ask Field

PLEASE ENTER THE FIRST TEN CHARACTERS OF THE FIRST LINE OF THE ADDRESS
JUNK FROM A.P. ADDRESS LABEL. MAKE SURE TO TYPE EXACTLY AS IT IS
PRINTED.
Text: Maximum 10 characters

First

INTERVIEWER FOR INFORMATION ....... You are in the Questionnaire for
Point no: (Point number)
Address no: (Address number)
Household no: (Household number)

End Date

PLEASE ENTER THE DATE OF THIS INTERVIEW. ENTER DAY OF MONTH IN NUMBERS,
NAME OF MONTH IN WORDS (FIRST THREE LETTERS), YEAR IN NUMBERS. EG: 1 Jan 98.

Who Lives Here

INTERVIEWER: COLLECT THE NAMES OF THE PEOPLE IN THIS HOUSEHOLD.
1 Continue

IF First person in household OR Move-Yes THEN
Name
What is the name of person number (1-12)?
ENTER PERSON'S FORENAME
Move
Is there anyone else in this household?
1 Yes
2 No
ENDIF
1. Background

The data files contain data from Health Survey for England 2007 (HSE), the seventeenth year of a series of surveys designed to monitor trends in the nation’s health. The 2007 Health Survey was commissioned by the Information Centre and carried out by the Joint Health Surveys Unit of the National Centre for Social Research and the Department of Epidemiology and Public Health at Royal Free and University College Medical School.

The aims of the Health Survey series are:
- to provide annual data about the nation’s health;
- to estimate the proportion of people in England with specified health conditions;
- to estimate the prevalence of certain risk factors associated with these conditions;
- to examine differences between population subgroups in their likelihood of having specific conditions or risk factors;
- to assess the frequency with which particular combinations of risk factors are found, and which groups these combinations most commonly occur;
- to monitor progress towards selected health targets;
- since 1995, to measure the height of children at different ages, replacing the National Study of Health and Growth;
- since 1995, monitor the prevalence of overweight and obesity in children
- to monitor the impact of the national smoking ban 2007.

2. Survey Design

The Health Survey for England 2007 was designed to provide data at both national and regional level about the population living in private households in England. The sample for the HSE 2007 comprised of two components: the core (general population) sample and a boost sample of children aged 2-15. The core sample was designed to be representative of the population living in private households in England and should be used for analyses at the national level. The core sample was split in two for some modules of the 2007 survey, further details are shown in Appendix A.

A random sample of 720 PSUs (Primary Sampling Units) was selected for the core and the boost sample, an additional 180 PSUs were used to supplement the child boost sample. The PSUs were selected with probability proportional to the total number of addresses within them. Once selected, the PSUs were randomly allocated to the 12 months of the year (60 per month in the core sample, 15 per month in the additional child boost) so that each
The Health Survey for England

Errata Note

Author: The NHS Information Centre, Lifestyle Statistics

Responsible Statistician: Paul Eastwood, Lifestyles Section Head

Version: 1

Original Date of Publication: 16th December 2008

Date of Publication of Errata Note: 19th November 2009

As a result of a detailed validation carried out during the production of the Health Survey for England (HSE), 2008, the following error has been identified in the Health Survey for England series. The error pertains to childhood obesity data for the years 1995 to 2007 inclusive and will affect all publications during these years.

Between 1995 and 2007 there was an error which meant that small numbers of children that should have been classified as either overweight or obese were omitted from these categories because of rounding of age and Body Mass Index (BMI) thresholds. The revised percentages of those who were overweight each year differ by less than 0.1 to 0.8 percentage points and for those who were obese each year they differ by less than 0.1 to 1.1 percentage points from those originally published, and 0.3% to 1.2% of children in each year were misclassified. In no cases were results significantly different from those presented previously.

An updated version of the following latest trend table for 1995 to 2007 has therefore been published:

Children Table 4, Children’s overweight and obesity prevalence, by survey year, age-group and sex.

This table gives revised figures for children’s overweight and obesity prevalence for 1995 to 2007, as included both in the main reports and separate trend table publications.

References to these figures within the publications have not been corrected, as the changes are small and will be corrected in the next HSE report which is expected to be published in December 2009. Figures within the reports, referring to the children’s categories of overweight or obese may be slightly different, but none are significantly different. All figures in the Health Survey for England 2008 (expected to be published on 17th December 2009), main report and trend tables will be calculated using the correct methodology and the time series data for 1995 to 2007 will all be updated using the correct methodology.

This note has been agreed with the HSE contractors, the National Centre for Social Research.
UK DATA ARCHIVE METADATA

- UK Data Archive collects and creates structured metadata for each archived dataset

- created during ingest data processing (Data-level metadata)
  - data dictionaries, format transfer, data listing, ingest processing details and information gathered in ‘readme’ file for users

- Catalogue record and keyword index
  (mix of Study-/Data-level metadata - ‘Catalogue metadata’. Also contains ‘Administrative metadata’, such as access conditions, date of publication, etc.)
  - data deposit form
  - keyword index covers data elements and concepts
  - international standards: DDI, METS, ISAD(G), TEI
  - standardised elements + controlled vocabularies = consistent search and retrieval
  - sufficient information for users to decide if the data suitable
  - information on the provenance of a dataset
  - record of publications
SN 6112 - Health Survey for England, 2007

* Internal User *

Documentation Variable List Keywords Publications Related Studies

**Title:**
Health Survey for England, 2007 (HSE)

**Subject Categories:**
Health Survey for England - Major studies
General - Health
Nutrition - Health
Physical fitness and exercise - Health
Health services and medical care - Health

**Depositor(s):**
National Centre for Social Research

**Principal Investigator(s):**
National Centre for Social Research
University College London, Department of Epidemiology and Public Health

**Sponsor(s):**
Information Centre for Health and Social Care

**Abstract:**
The Health Survey for England (HSE) is part of a wider programme of surveys commissioned by the Department of Health, and is designed to monitor trends in the nation's health. The aims of the HSE series are:

- to provide annual data about the nation's health
- to estimate the proportion of people in England with specified health conditions
- to estimate the prevalence of certain risk factors associated with these conditions
- to examine differences between population subgroups in their likelihood of having specific conditions or risk factors
- to assess the frequency with which particular combinations of risk factors are found, and which groups these combinations most commonly occur
- to monitor progress towards selected health targets
- since 1990, to measure the height of children at different ages, replacing the National Study of Health and Growth
- since 1995, to monitor the prevalence of overweight and obesity in children

The survey focuses on different health issues each year, although a number of core questions are included every year. Topics are revisited at appropriate intervals in order to monitor change:

- in 1993 and 1996, the major focus of the survey was atopic disease (including asthma and eczema), accidents and disability
- in 1997 the survey report focused on the health of young people aged 2-24 years, and combined data (in report format only) from the 1995-1997 surveys
- in 1999 and again in 2004, the focus was on the health of minority ethnic groups. The boost sample was designed solely to yield additional interviews with members of the most populous minority ethnic groups: Black Caribbean, Black African (covered in 2004 only), Indian, Pakistani, Bangladeshis, Chinese and Irish
- in 2000 and 2005, the survey focused on the health of older people, as the HSE 2000 included a sample of care home residents, whereas HSE 2005 included a boost sample of people aged 65 and over residents in private households
- in 2001, the survey included the first time questions on fruit and vegetable consumption
- the 2002 survey focused on children, young people and menstrual health, and included a boost sample of young people aged 6 to 24 years and mothers of infants aged under one year
- the 2007 survey focused on knowledge and attitudes towards key aspects of lifestyle, and monitoring any impact from the 2007 smoking legislation

Further information about the HSE may be found on the Department of Health, the Information Centre for Health and Social Care and National Centre for Social Research.
Main Topics:
For adult respondents, the HSE 2007 focused on knowledge, attitudes and behaviour on key aspects of lifestyle - smoking, drinking, eating and physical activity. Both adults and children were asked about their views on what constitutes healthy behaviour, their knowledge of government recommendations and the factors that may encourage or discourage healthy behaviour.

A secondary focus was the impact of the smokefree legislation. The 2007 survey allowed an initial examination of the effect of the legislation by looking at adults’ and children’s smoking behaviour and their exposure to other people’s smoke, pre and post 1 July 2007 (the date the legislation came into effect). As well as questions being asked, saliva samples were taken and tested for nicotine, an indicator of recent nicotine exposure.

Questionnaire interviews were followed by a nurse visit, both using computer assisted interviewing (CAPI) and self completion booklets. Parents of children aged 0-12 years were asked about their children, with the child interview including questions on physical activity and fruit and vegetable consumption. Children aged 13-15 were interviewed themselves.

Standard Measures:
- Strengths and Difficulties Questionnaire (SDQ)

Coverage:
- Time Period Covered: January 2007-December 2007
- Dates of Fieldwork: January 2007-April 2008
- Field date: 2007
- Country: England
- Spatial Units: Government Office Regions (GORs)
- Observation Units: Individuals
- Kind of Data: Numeric data; Individual (micro) level

Universe Sampled:
- Location of Units of Observation: National
- Population: Adults (aged 16 and over) and children (aged 0-15 years) in the general population, living in private households in England during 2007. A boost sample of children aged 2-15 was also included.

Methodology:
- Time Dimensions: Repeated cross-sectional study
- The survey is conducted annually.
- Sampling Procedure: Multi-stage stratified random sample
- Number of Units: 14,306 cases (individual level), 24,910 cases (household level).
- Method of Data Collection: Face-to-face interview; Self-completion; Clinical measurements; Physical measurements; CAPI
- Weighting: Several types of weighting variables have been used. See User Guide for details.

Language(s) of Written Materials:
- Study Description: English
- Study Documentation: English

Documentation:
Permission for commercial orders from Patrick Tucker email: Patrick.Tucker@cshl.gov.uk. No royalty requests so far.
(28/2/2007) permission from Andy Sutherland: andy.sutherland@cshl.gov.uk
Operations Directorate
The information centre for health and Social care
2nd Floor South
1 Trelivian square
Bear lane
Leeds LS1 0AE
0113 254 7200

Andy Sutherland would like to receive a copy of the signed commercial agreement.

Andy Sutherland is no longer responsible for the HSE. Permission for commercial requests now needs to be obtained by emailing surveys.queries@cshl.gov.uk. A copy of the signed commercial agreement should be attached. 84 June 2000
SN 6112 - Health Survey for England, 2007

Variables

**GROUP 1: hs672ah**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>First outcome code</td>
</tr>
<tr>
<td>1</td>
<td>Individual outcome codes</td>
</tr>
<tr>
<td>2</td>
<td>Sex</td>
</tr>
<tr>
<td>3</td>
<td>Age (past birthday)</td>
</tr>
<tr>
<td>4</td>
<td>Number of adults aged 16+</td>
</tr>
<tr>
<td>5</td>
<td>Number of children</td>
</tr>
<tr>
<td>6</td>
<td>Number of infants</td>
</tr>
<tr>
<td>7</td>
<td>Enter number of households covered by this address</td>
</tr>
<tr>
<td>8</td>
<td>Address number</td>
</tr>
<tr>
<td>9</td>
<td>Household</td>
</tr>
<tr>
<td>10</td>
<td>Relationship to other persons in household</td>
</tr>
<tr>
<td>11</td>
<td>Relationship to other persons in household</td>
</tr>
<tr>
<td>12</td>
<td>Relationship to other persons in household</td>
</tr>
<tr>
<td>13</td>
<td>Relationship to other persons in household</td>
</tr>
</tbody>
</table>

Show variable

**GROUP 2: hs672ae**
SN 6112 — Health Survey for England, 2007

Variable detail

GROUP 1: Isen07eh

<table>
<thead>
<tr>
<th>Variable Id</th>
<th>Value</th>
<th>Unweighted Frequency</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Value 1 (Men)</td>
<td>12084</td>
<td>48.51%</td>
</tr>
<tr>
<td></td>
<td>Value 2 (Women)</td>
<td>12090</td>
<td>51.49%</td>
</tr>
</tbody>
</table>

Full list of variables
UK Data Archive

SN.6112: Health Survey for England, 2007

Publications

By Principal Investigator(s):


Related studies

**Group Constituents:**

- 6112 - Health Survey for England, 2007
- 5806 - Health Survey for England, 2006
- 5678 - Health Survey for England, 2005
- 5439 - Health Survey for England, 2004
- 5008 - Health Survey for England, 2003
- 4912 - Health Survey for England, 2002
- 4608 - Health Survey for England, 2001
- 4472 - Health Survey for England, 2000
- 4365 - Health Survey for England, 1999
- 4253 - Health Survey for England, 1998
- 3979 - Health Survey for England, 1997
- 3886 - Health Survey for England, 1996
- 3796 - Health Survey for England, 1995
- 3640 - Health Survey for England, 1994
- 3238 - Health Survey for England, 1993

**Other Related:**

Andy Sutherland would like to receive a copy of the signed commercial agreement.

Andy Sutherland is no longer responsible for the HSE. Permission for commercial requests now needs to be obtained by emailing surveys@esds.ac.uk. A copy of the signed commercial agreement should be attached. 8 June 2009

If user requires additional variables (e.g. strategic health authorities), user needs to complete a request form from NatCan. Contact Rachel Craig - r.craig@natcen.ac.uk

In response to a request from the Information Centre for Health and Social Care, who note that they took over sponsorship of the HSE from 2004, ‘Department of Health’ has been removed from the generic catalogue record and added to the checks from 1991-2003. See red folder correspondence for details. SB 12/07.

Access:
Access Conditions: The depositor has specified that registration is required and standard conditions of use apply. The depositor may be informed about usage. See terms and conditions for further information.
Availability: ESDD Government, UK: Data Archive
Contact: Help desk: govsurveys@esds.ac.uk
Access Code: 2003A

Date of Release:
First Edition: 26 February 2009

Copyright:
Crown copyright material is reproduced with the permission of the Controller of HMSO and the Queen's Printer for Scotland

Documentation:

<table>
<thead>
<tr>
<th>Format</th>
<th>Name</th>
<th>Size in Kilobytes</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDF</td>
<td>6112Sampledata.pdf</td>
<td>808</td>
<td>Lists of Variables and Derived Variables</td>
</tr>
<tr>
<td>PDF</td>
<td>6112InterviewQuestions.pdf</td>
<td>1655</td>
<td>Questionnaires, showcards, coding frames and consent booklet</td>
</tr>
<tr>
<td>PDF</td>
<td>6112SupportMaterial.pdf</td>
<td>1573</td>
<td>Interviewer, Nurse, Coding and Editing Instructions</td>
</tr>
<tr>
<td>PDF</td>
<td>6112UserGuide.pdf</td>
<td>208</td>
<td>User Guide</td>
</tr>
<tr>
<td>HTML</td>
<td>bmi_enropa_note.pdf</td>
<td>143</td>
<td>BMI Enropa Note</td>
</tr>
<tr>
<td>HTML</td>
<td>UKDA Study 6112 Information.htm</td>
<td>26</td>
<td>Study information and citation</td>
</tr>
</tbody>
</table>

MORE File

READ File

Download Acrobat PDF Reader
Catalogue record last updated:
27 April 2010
UK DATA ARCHIVE: IMPORTANT STUDY INFORMATION

Study Number 6112 - Health Survey for England, 2007

Details of Study Received
First edition received January 2009 and second edition received March 2010.
Anna Strauss, Manager National Centre for Social Research 35 Northampton Square London EC1V 0AX Email: a.strauss@natcen.ac.uk

Details of data files received
SPPS formats:
1st edition:
HSE07h raw HSE07h.sav
2nd edition:
bm07a raw

Details of doc files received
1st edition: DOC
Adobe PDF format:
bmu_errata_note.pdf

Details of hard copy received
All editions:

Filename changes: files for issue
All editions:
Spaces were removed and capitals changed to lower case.

Filename changes: files in /muisue
All editions:
Spaces were removed and capitals changed to lower case.

Level of processing
A*

Data file conversion: methods, formats created and validation
All editions:
SPSS files were converted to Stata and Tab via scripts.

During conversion of these data files within the Nectarch Publisher (V3.56) it was found that the variable measure definitions were set incorrectly i.e. scale, nominal etc. The definitions were fixed using the Publisher software for the Nectarch version of the study but the original SPSS file remains unchanged.
Details of hard copy received
All editions
Note

Filename changes: files for issue
All editions
Spaces were removed and capitals changed to lower case.

Filename changes: files in pipeline
All editions
Spaces were removed and capitals changed to lower case.

Level of processing
A*

Data file conversion: methods, formats created and validation
All editions
Spat files were converted to Stata and Tab via scripts.

During conversion of these data files within the Neorstar Publisher (V3.56) it was found that the variable measure definitions were not correctly i.e. scale, nominal etc. The definitions were fixed using the Publisher software for the Neorstar version of the study but the original SPSS file remains unchanged.

Doc file conversion: methods, formats created and validation
1st edition
Word files were converted to PDF, amalgamated and bookmarks added in line with HSE conventions.
2nd edition
PDF added as a separate note. Bookmarks and standard UKDA header added.

Dataset validation and confidentiality checks and solutions to problems raised
1st edition
All frequencies were checked. Variable labels were added and value labels added where necessary. Revised queries (see emails for full details). Variable XEDRUG was removed after consultation with the depositor. It had no labelling and turned out to be a scratch variable. Appendix D was not supplied and not needed so this page was removed.
2nd edition
Both data files were sorted by personal and the new variables were matched onto the individual file. Scripts were run again.

Data and documentation problems
None

New Edition Information
For the second edition (April 2010), three new children’s Body Mass Index (BMI) variables have been added to the individual data file (bmiinc1, bmicnt2, bmicnt3). The original variables (bmicnt1, bmicnt2, bmicnt3) are unreliable and should not be used. Further information is available in the documentation and on the Information Centre for Health and Social Care Health Survey for England web page.

Notes to data delivery

Notes from data delivery and post order corrections

Data and documentation processed by Kathryn Shu: April 2010
NEW EDITION INFORMATION

For the second edition (April 2010), some new child's body mass index (BMI) variables have been added to the individual data file (BMIcat, BMIcentile). The original variables (BMIcat, BMI2centile) are unreliable and should not be used. Further information is available in the documentation and on the Information Centre for Health and Social Care website (see link on page.

DATA PROCESSING NOTES

**Data Archival Processing Standards**

The data were processed by the UK Data Archive's A* standard. This is the Archive's highest standard, and means that an extremely rigorous and comprehensive series of checks was carried out to ensure the quality of the data and documentation. (Strictly, checks were made that the number of cases and variables matched the depositor's records. Secondly, checks were made that all variables had comprehensible variable labels and all nominal (categorical) variables had comprehensible value labels. Where possible, either with reference to the documentation and/or in common discussion with the depositor, labels were accordingly edited or created. Thirdly, logical checks were performed to ensure that nominal (categorical) variables had values within the range defined either by value labels or in the depositor's documentation. Lastly, any data or documentation that breached confidentiality rules were altered or suppressed to preserve anonymity.)

**Data and Documentation Problems**

Note.

**Data conversion information**

From January 2000 onwards, almost all data conversions have been performed using software developed by the UKDA. This enables standardisation of the conversion methods and ensures optimal data quality. In addition, to its own data processing/conversion code, this software uses the SPSS and Stata/Transfer command processors to perform certain transformations. Although data conversion is automated, all data files are also subject to visual review by a UKDA data processing officer.

With some format conversions, data and more especially internal metadata (i.e. variable labels, value labels, missing value definitions, data type information), will necessarily be lost or truncated owing to the different limits of the proprietary formats/UKDA/UKDA Data Dictionary files (enrich both formats) corresponding to such data files, is usually provided for viewing and searching the internal metadata, as it exists in the original format. These files are called (data filename).UKDA_Data_Dictionary.rtf

**Important information about the data format supplied**

The links below provide important information about the format in which the data is supplied. Some of this information is specific to the impact format of the data, that is the format in which the UKDA supplied the data in. The impact format for this study was SPSS.

Please click below to find out more about the format in which you have been supplied the data in:

**SPSS** (.sav)

SPSS portable (.sav file)

If SPSS portable was not the impact format, this format will generally either have been created via the SPSS command processor (e.g. if the input format is .sav, SAS, Excel, or Stata), or if the impact format was STATA, the SPSS version will be created via the Stata/Transfer command processor. If the input format was undeletable text, the data will have been read into SPSS using an SPSS encoded file.

Issues: There is very seldom any loss of data or internal metadata when importing data files into SPSS. Any problems will have been listed above in the Data and Documentation Problems section of this file.

**STATA** (.dta)

STATA can use the import format, all STATA files will have been created from SPSS data file via the Stata/Transfer command processor. Importantly, Stata's optimization routine is run on data variables with SPSS wide format meaning that the data (e.g. numeric variables with 10 decimal places of data formatted to Fix2) are not restricted upon conversion to STATA because they are converted to doubles rather than floats. Unrecogizable values are copied across into STATA as opposed to being collapsed into a single missing data code.

Issues: There is a number of data and internal metadata that have been lost or truncated, this will have been automatically logged in this file (fix_3). The Data Dictionary file (.ukdata) and (data filename).UKDA_Data_Dictionary.rtf contains the complete internal metadata that has been supplied with the UKDA Data Dictionary files (ukdata) and (data filename).UKDA_Data_Dictionary.rtf.

**Tab-delimited text (.dat)**

If tab-delimited text was not the report format, tab-delimited files have been created from SPSS Portable files via the SPSS command processor, and also from Excel and MS Access files. When importing Access data tables to tab-delimited text, they have potential problems with special characters (tabs, carriage returns, line feeds, etc.) allowed by Access viewers and Excel files. Tab and/or carriage returns in cells are stripped out by the UKDA.

Issues: Data formats in SPSS are always created as 32-bit floating point values, in tab-delimited text format, these values are converted to 16-bit floating point values. Any time information is critical, a new variable will have been created in the tab-delimited data file by the UKDA. All rows of the data in the tab-delimited format should consult the UKDA Data Dictionary file (.ukdata) and (data filename).UKDA_Data_Dictionary.rtf.

If the data was sourced from MS Access; more limited [data documentation information is supplied within the file's (data filename).ukdata file(s). These files may also contain SQL update information.

**Excel** (.csv)

If MS Excel was not the report format, Excel files have been created via the SPSS command processor. The data will then be issued outputted via the tab-delimited format to provide SPSS to Excel conversion via the SPSS command processor.

**SAS** (.sas7bdat)

If SAS was not the report format, all SAS files have been created from SPSS Portable files via the Stata/Transfer command processor. The data files are provided as a fixed width text file (.dat) and a SAS command file (.sas), which when run will create a SAS dataset. This enables the user to re-create the SAS dataset and formats library in almost all versions of SAS and all operating systems.

Issues: The main loss of information when converting from SPSS to SAS is user-assigned variable or value definitions. By editing the .sas file, the user can choose whether to collapse all user-assigned values into system missing or preserve them/both and lose the user-assigned definition. To achieve the latter, note the following section of the .sas file should be removed before running it:

```
/* User Written Value Specification */
```

Note that the complete internal metadata has been supplied with the UKDA Data Dictionary file (.ukdata) and (data filename).UKDA_Data_Dictionary.rtf.

**MS Access** (.mdb)

Due to the substantial incompatibilities between versions of MS Access, the UKDA only make data available in MS Access format if this is the impact format and the database contains important information in addition to the data tables (coding information, forms, queries, etc.).
HOW?

• Researchers can embed certain annotations in data files:
  • quantitative data: variable/value labels; worksheet information; table relationships and queries in relational database; GIS data layers/tables

Examples:
  SPSS: variable attributes documented in Variable View (label, code, data type, missing values)
  MS-Access: variable descriptions and attributes documented in Design View; relationships
  ArcGIS: shapefiles (layers) and tables in geodatabase; metadata created in ArcCatalog.
  Excel: base worksheet data-related documentation.
HOW?

• qualitative data/text documents: interview transcript speech demarcation (speaker tags); document header with brief details of interview date, place, interviewer name, interviewee details, context

• Additional documents: final report, published reports, user guide, working paper, publications, lab books
“to identify national scale survey data for England that have potential to be used for rural and urban analyses by the Commission for Rural Communities.”
# ONS RURAL AND URBAN DEFINITION, 2004

<table>
<thead>
<tr>
<th>Settlement type</th>
<th>Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>Urban</td>
<td></td>
</tr>
<tr>
<td>Urban &gt;10k</td>
<td>Less sparse</td>
</tr>
<tr>
<td></td>
<td>Sparse</td>
</tr>
<tr>
<td>Rural</td>
<td></td>
</tr>
<tr>
<td>Town and fringe</td>
<td>Less sparse</td>
</tr>
<tr>
<td></td>
<td>Sparse</td>
</tr>
<tr>
<td>Village</td>
<td>Less sparse</td>
</tr>
<tr>
<td></td>
<td>Sparse</td>
</tr>
<tr>
<td>Hamlet and isolated dwellings</td>
<td>Less sparse</td>
</tr>
<tr>
<td></td>
<td>Sparse</td>
</tr>
</tbody>
</table>
1. Datasets with rural variables

- ONS 2004 rural/urban area classification or contain Output Areas as spatial units
- Post 2000 national datasets for England with representative population sample.

<table>
<thead>
<tr>
<th>Rural marker type</th>
<th>Number of datasets where variable type used</th>
</tr>
</thead>
<tbody>
<tr>
<td>ONS 2004 rural/urban definition, distinguishing either 8 or 6 categories</td>
<td>23</td>
</tr>
<tr>
<td>Binary rural/urban variable</td>
<td>6</td>
</tr>
<tr>
<td>Variable distinguishing rural/urban/other category</td>
<td>2</td>
</tr>
<tr>
<td>Other variable indicating ‘rural areas’</td>
<td>8</td>
</tr>
<tr>
<td>Respondent evaluation of rural/urban area</td>
<td>7</td>
</tr>
<tr>
<td>Indirect indication through settlement size or population density</td>
<td>3</td>
</tr>
</tbody>
</table>
2. National survey datasets for England (post 2000), with representative population sample that do not contain a rural/urban variable. Data owners contacted to investigate

- whether a rural/urban variable may exist for the dataset and can be made available
- dataset can be retrospectively coded to include a rural/urban variable, based on the presence of postcodes in the dataset
- nineteen datasets identified
- retrospective coding dependent on willingness of depositor
WHY ENHANCEMENT MATTERS

- Good metadata enables easy discovery of data providing meaningful information (titles, descriptions, abstract, keywords)
- Submitting good documentation and metadata enables data archives/data centres to create rich and detailed metadata
- Good data documentation leads to good data re-use
- Detailed yet comprehensive documentation
- Where possible integrate within dataset
• Contact the UK Data Archive if not sure what to produce or provide:
  • Advice on our Managing and Sharing web pages: http://www.data-archive.ac.uk/sharing/metadata.asp
  • The UK Data Archive Data Management Team: datashar@essex.ac.uk