GSS/GSR Disclosure Control Guidance for Microdata Produced from Social Surveys – Case Studies

These Case Studies are designed to show how statistical disclosure control is applied in practice when producing microdata. It is intended that these studies can be updated when later versions of the data are released. New studies (ideally from departments other than ONS) can also be added to ensure that this document covers as wide a range of examples as possible.
If there are any queries on these studies or if there are examples to add please email sdc.queries@ons.gsi.gov.uk

Case Study 1

The Wealth and Assets Survey (WAS) is an ONS longitudinal survey that interviewed across Great Britain; England, Wales and Scotland (excluding North of the Caledonian Canal and the Isles of Scilly). Respondents to wave one (July 2006 – June 2008) of the survey were invited to take part in a follow up interview two years later (July 2008 – June 2010) to see how household assets change over the life course.
These data were released in 2012 initially to the UKDA and now the UK Data Service for use by Approved Researchers under an SL arrangement. There was additional demand for access to these data especially from international users who are unable to be defined as approved researchers. Therefore a EUL was created.

Risk Assessment.

The following information was used in the risk assessment.

- The most likely intruder scenarios were considered to be ‘Use of public datasets’ and ‘Spontaneous recognition’.
- Sample size (Wave one achieved approximately 30,000 household interviews; wave two achieved approximately 20,000 household interviews i.e. a relatively small sample.
- The survey is longitudinal.
- The survey is a household survey
- The potential for extreme outliers on wealth variables

The above information was used when determining the key variables, those variables which in combination might enable identification of an individual or household or an attribute relating to the individual or household. A partial list of key variables is:
Geography
Country of Birth
Ethnicity
Religion
Sexual identity
Age
Household Size
Occupation

Applying Disclosure Control

- Remove households of size 10 and above from the EUL dataset
- Top code Age at 80
- Give special consideration to the Wealth variable

Following these guidelines, the EUL microdata would require all variables relating to wealth and finance to be top-coded. However, the primary purpose of WAS is to provide statistical detail of those variables, and to have as much detail as possible. A compromise was sought whereby some variables of lesser research importance were removed in order to reduce the risk of identification. In order to retain the full detail of the financial variables although some rounding at the top level was still required.

Additional disclosure control was thus required, especially as this is also a longitudinal dataset. The combination of the longitudinal aspect and the wealth requirements lead to the following decisions.

- Remove Geography from the EUL dataset
- Remove Sensitive and 'observable' socio-demographic variables (country of birth, ethnic group, religion and sexual identity from EUL dataset
- Code Age into 5 year age bands
- Limit SOC (Occupation) to 2 digits
- Remove any flags that can identify births
- Suppress Wealth to three significant figures.
- The number of cars for which value could be supplied was also limited to the first three with the values of any remaining cars grouped to a total.

- As a rule treat outliers on a case by case basis. A recent wave of the WAS dataset contained an outlier and as a result much detail was removed for both the EUL and SL. Top level components of wealth (physical, property, financial and pensions), total wealth and total income remained in the published data.

Comments on these data

- The removal of geography **significantly** reduces the risk of disclosure. It was confirmed that there are no country level questions that would reveal Scotland or Wales cases.
- The data are longitudinal but will not be pre-linked. Analysts will need to link the data themselves, this extra step is likely to reduce the likelihood of identifying split households and disclosing information about new household (possibly temporary household) members.
  The disclosure risk is also decreased due to the age of the data (Wave 1 are data are up to 6 years old and Wave 2 data are up to 4 years old).Older data are clearly less disclosive. It will be difficult to identify positively an individual from data over 10 years old. An element of protection is also given by the gap between collection and publication.
- The data are reviewed on a wave by wave basis to ensure the rules are still appropriate with 'evolving' data

Case Study 2

The Measures of Subjective Well-being Project is one of four projects which together form the Measuring National Well-being (MNW) Programme. The aim of the Programme is to provide a fuller picture of 'how society is doing' than is given by economic indicators such as GDP.
In April 2011 ONS introduced four overall monitoring questions to the IHS core: one on life satisfaction, two on experienced based measures and one on purpose and meaning of life.

The data to be released to the UK Data Service as a EUL are well-being estimates from a 6 month APS dataset.

Risk Assessment.

The following information was used in the risk assessment...

- The most likely intruder scenarios were considered to be ‘Use of public datasets’ and ‘Spontaneous recognition’.
- Sample size is about 80,000. The whole UK is sampled.
- The survey is not longitudinal
- The survey is an individual survey

As in Case Study 1, the above information was used when determining the key variables. Possible Key variables are
Income / Salary
Ethnicity
Religion
Educational Qualifications
Variables potentially relating to sexual identity
Household Size

Applying Disclosure Control

Initial changes recommended by the supplier were

- Geography level to be Region
- Age to be aggregated into 5 year age groups
- Country of birth to be two categories, EU/non EU
- Family size top coded at 9
- Some recoding of marital status
- Case number anonymised using random numbering

The following additional changes to the microdata were recommended by the ONS SDC branch prior to the data being released to the UK Data Service.

- Households of size 10 above are to be removed from the EUL data
- All income and salary variables to be top coded
- The maximum number of categories for religion is to be 10
- The maximum number of categories for ethnicity is to be 16 (This is number of standard categories in Census 2001. If this dataset was requested now the maximum number of categories could be increased to 18 to be consistent with Census 2011.
- The main health problem variable will not be included
- The general health variable will only have 4 categories
- Variables which potentially reveal sexual identity will be removed
- Top coding applied for personal monetary information.

Comments on these data

- This is a high profile dataset and caution has been applied in deciding which variables can be released. The changes recommended by the SDC branch were very detailed when compared with the initial disclosure procedures.
- In particular data relating to health has been reduced in detail. This may be an issue with researchers as this is a major part of the wellbeing project.
- Variables relating to health are subjective when answered by the respondent and maybe of low quality. There may be a possibility of the data being disclosive if these are combined with the respondent’s medical records. Certain variable combinations may encourage an intruder to look at the data in more detail.

Case Study 3

The Living Costs and Food Survey (LCFS) collects information on spending patterns and the cost of living that reflects household budgets across the country. The primary uses of the survey are to provide information about spending patterns for the Consumer Price Indices, and about food consumption and nutrition.

Many requests are made for data from this survey. This specific request was for a 'teaching dataset', that is, one that would be made available 'publicly' under Open Government License (OGL). The OGL has few conditions other than not to misrepresent the data, as opposed to the End User License (EUL) where users have to undertake not to try to identify an individual respondent, or to claim to have identified a respondent (see Appendix F). A request for a teaching dataset thus required considerable disclosure control as this would be available without restriction.

Risk Assessment

The following information was used in the risk assessment...

- The most likely intruder scenarios were considered to be 'Use of public datasets' and 'Spontaneous recognition'.
- Sample size is about 0.04% of the UK population.
- The survey is not longitudinal
- The survey is a household survey

Applying Disclosure Control

- This is a household dataset with individual-level variables for only the HRP.
- Sample size is 0.04% households - achieved sample is around 6,000. Coverage is UK
- Few socio-demographic variables are included - sex, 3-class NS-SEC, 4-category economic activity.
- Household size variables at top-coded at 4+ adults, 2+ children
- Gross Household income is not banded, but is top-coded satisfactorily at approx £62k p.a. Main source is 'earned' or 'other'.
- Only geographic variable is 'modified' Region

Data include case number - this should not relate to case number in any other version of the data.

Comments on these data

- This is a training dataset and hence only a relatively small amount of information is released. Only a handful of socio-demographic variables are included.
- A relatively low value for top coded household income
- Household size is top coded at 6
- The data will be of use to researchers and students to develop code which could then be run on a more detailed extract under an EUL.