Wealth and Assets Survey

This is a best practice example of data dissemination by the Office for National Statistics. This particular case study is most relevant to anyone who wants or needs to safely disseminate anonymised individual level survey data. It details the use of both restricted data access methods and restricted data methods.

When reading this case study it is important to remember that the anonymisation assessment, recommendations and actions are specific to the data examined here although they will have relevance to other similar data and data context.

- **Summary**

Organisation disseminating the data: Office for National Statistics (ONS).
Data: Longitudinal survey for England, Wales and Scotland.
Data type: Microdata.
Features of risk: Longitudinal, hierarchical data - Data recorded across time as well as hierarchal data can potentially produce unusual patterns that could facilitate re-identification.
Risk type: re-identification and the disclosure of information not previously in the public domain.
Disclosure Control Methods (SDC):
Direct identifiers removed.
Targeted SDC on rare cases based on intruder scenarios and identification of key variables.
Restricted data access methods i.e. Special Licence and End User Licence agreements. Restricted data methods - depending on the type of license: (i) removal of variables such as geography, ethnic group, country of birth and sexual identity, (ii) reduction in variable detail such as banding, top coding and rounding.
Disclosure risk checks: Intruder scenarios 'use of public datasets' and 'spontaneous recognition'.

- **Background**

The economic well-being of households is sometimes measured by their income; this ignores the fact that a household's resources can be influenced by their stock of wealth. The increase in home ownership, the move from traditional roles and working patterns, a higher proportion of the population now owning shares and contributing to investment schemes as well as the accumulation of wealth over

the life cycle, particularly through pension participation, have all contributed to the changing composition of wealth. **The Wealth and Assets Survey (WAS)** aims to address gaps identified in data about the economic well-being of households gathering information on level of assets, savings and debt. It also looks at how wealth is distributed among households and factors affecting financial planning.

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). Private households were sampled for the survey (meaning that people in residential institutions, such as retirement homes, nursing homes, prisons, barracks or university halls of residence, and also homeless people were not included). The WAS commenced in July 2006. 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 identify whether their circumstances had changed. Interviews in waves one and two were conducted using Computer Assisted Personal Interviewing (CAPI). Wave achieved approximately 30,000 household interviews; wave two achieved approximately 20,000 household interviews. The aim was to release the Wealth and Assets Survey as microdata.

- **The Anonymisation Problem and Assessment**

The risk assessment of the WAS considered the fact that the survey is longitudinal and unusual patterns across time might help to identify people or households. The hierarchical nature of the data also significantly increases the risk of disclosure since unusual combinations of residents, especially in large households, may facilitate identification. However the sample size is relatively small overall. In addition, extremely wealthy households (forming only a tiny minority of the total population) were initially oversampled to ensure inclusion in the final dataset; however weights were later used to recalibrate the sample back to the normal population. The fact that these extremely wealthy households are present means there is a substantial risk in the dataset.

ONS considered likely 'intruder scenarios' as specified in the GSS SDC policy for Microdata produced from Social Surveys to help assess the risk of disclosure. The most likely intruder scenarios were considered to be 'use of public datasets' and 'spontaneous recognition'. **Key variables were determined:** those variables in each intruder scenario which in combination might enable identification of an individual or household or an attribute relating to the individual or household. The key variables included geography, country of birth, ethnicity, religion, sexual identity, age, household size and occupation. Where there were few

records with a particular combination of key variable characteristics, disclosure control was applied by removing variables or reducing detail.

- **Data sharing, dissemination and Publication**

These data were released in 2012 to the UK Data Archive for use by Approved Researchers under a special licence arrangement (http://ukdataservice.ac.uk/get-data/how-to-access/conditions.aspx). This means researchers have to apply to access the data for valid research purposes and have to agree to a set of terms and conditions. The data have to be stored appropriately. There was additional demand for access to these data especially from international users who are unable to be defined as approved researchers. As a result a revised (and further disclosure controlled) dataset was made available under an End User Licence (EUL) agreement which has fewer restrictions on who can use the data and how they are stored (see previous weblink).

Differences between the Special Licence (SL) data and End User Licence (EUL) data:

The main difference between the End User Licence (EUL) data and the Special Licence (SL) data is that the SL includes geography variables: Local Authority and Government Office Region. **Both datasets have direct identifiers omitted** such as names, addresses, other contact details, National Insurance number, Court case number, etc. The industrial classification SIC is included on both datasets at 2 and 5 digits, whereas Standard Occupational Classification is limited to 2 digits only on EUL but available at 1,2,3,4 digits on the SL. On the SL a small number of households have more than 10 people. There is no suppression of financial variables although many are still banded to disguise outliers. Full information is provided on number of cars and the value of each and no cap on number of bedrooms. Furthermore the data contain no other variable which could allow a respondent to be identified without considerable effort on the part of the data user.

End User Licence (EUL) data:

It was with the consideration that the wealth variables are of paramount importance (as they are of most interest to users) that the EUL was created. The application of disclosure methods or removal of these wealth variables would therefore greatly reduce the analytical utility of the data and indeed, the primary purpose of the survey itself. Therefore, **variables of 'secondary importance' were removed to preserve the analytical value of wealth variables** as the

primary objective to ensure data utility is maximised whilst complying with acceptance risk threshold for an EUL dataset.

First steps involved removing all households of size 10 and above. Age was coded into five year bands and top coded at 80. The number of vehicles has been top coded at 3 and number of bedrooms capped at 6. Sensitive and observable socio-demographic variables (country of birth, ethnic group, religion and sexual identity) were removed from the dataset. Any flags that can identify births were removed.

Additional disclosure control was then required as this is a longitudinal dataset. All geography variables were removed to create GB files as this **significantly** reduces the risk of disclosure – this includes urban/rural indicators. As a compromise, the Output Area Classifications (OAC) were introduced to the EUL file (which do not exist on the SL). There were no country level questions which would reveal cases from Scotland or Wales.

In order to retain the full detail of the financial variables some rounding at the top level was still required. Special consideration was given to the income variable. All variables relating to wealth and finance were top-coded in accordance with the GSS SDC policy. Many salaries are published and so could allow an intruder to be pretty certain of whom they think they have found. The rounding of 3 significant figures (equivalent of nearest thousand for £100k-£999k, for example) would create some uncertainty, with the absence of detailed occupation, geography, and no ethnic group information, but still retain most (if not all) of the data utility. This is not the traditional approach recommended by the GSS policy but in this case it allowed the variables of paramount importance to be preserved. Furthermore, the WAS data, although longitudinal is not pre-linked. Analysts will need to link the data themselves, thus reducing the likelihood of 'Joe Bloggs' identifying split households and disclosing information about new household (temporary household) members. The disclosure risk is also decreased due to the age of the data (Wave 1 data are up to 6 years old and Wave 2 data are up to 4 years old).

For further information please contact: Elaine Mackey at admin@ukanon.net