2011 UK Census Aggregate Outputs

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 disseminate tabular data to a public audience.

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 data: Office for National Statistics (ONS).
Data: 2011 UK Census for England and Wales.
Data type: Tabular data. Public dataset.
Summary of sharing, disseminatig, publication practices: Accessible online.
Risk type: Attribute disclosure.
Features of risk: This is population data given at low levels of geography. Whole population data means that there is near certainty as to who is in the dataset.
Disclosure Control Methods (SDC):
Targeted record swapping applied to the microdata before tabulation**.**
The removal of direct identifiers**.**
For tabular data further methods include reducing the detail of variables and/or omitting variables.
Disclosure risk checks: Assessment of doubt and Intruder testing exercise.

- **Background**

ONS runs the UK Census which happens every ten years. Census statistics help paint a picture of the nation and how people live. They provide a detailed snapshot of the population and its characteristics, and underpin funding allocation to provide public services.

The information provided in the 2011 Census is **confidential and protected in law by the Statistics and Registration Service Act 2007 (SRSA)**. ONS also has practical and ethical responsibilities for not disclosing personal information. Respondents trust that their data will be kept safe and this in turn impacts on response rates.

The 2011 Census is a vast source of information and covers the whole of England and Wales, at all levels, from the country as a whole down to small areas, for

example, wards. Therefore the potential for disclosure is much higher than it would be for a small sample survey because there is no doubt about who is in the data. The main aim of ONS' disclosure control strategy is to reduce disclosure risk to an acceptable level whilst retaining as much data utility as possible.

- **The Anonymisation Problem**

The Registrars General highlighted the key risk for 2011 Census outputs to be **attribute disclosure** i.e. learning something from the census data about an individual or group of individuals that was not previously known. This is highly associated, although not exclusively, with low numbers in tables. ONS' **aim was thus to introduce sufficient uncertainty into the data so that it cannot be said with certain whether any cell count is the true value, particularly for small cell counts**.

The exact threshold of uncertainty required was not decided until a later stage. This judgement was made at a later stage in the context of results from methodological research into the balance of protection afforded, and the damage caused. This was supported by evidence from an 'intruder test' described below.

- **Anonymisation Practices**

Statistical disclosure control has been applied to the data before production of outputs. ONS carried out a review of potential methods evaluated extensively taking into account user requirements. The final shortlisted **method of targeted record swapping was applied to the microdata** as a final stage before tabulation. Targeted record swapping involves swapping the records of a small percentage of households between geographic areas. While all households in the census had a chance of being swapped, the swapping was targeted towards individuals and households with unique or rare characteristics at small geographies. By targeting in this way the protection has been achieved by swapping fewer households than if it had been done entirely randomly. Most swapping took place at Middle Super Output Area (MSOA) level or below. A similar targeted approach was used to protect residents in communal establishments, but individuals rather than households were swapped. Some protection is already in the data due to non-response and the uncertainty created by imputed records. As the level of imputation varied between areas, higher rates of swapping were applied in areas with low imputation to ensure adequate protection was achieved.

The appearance of small values (zeros, ones, twos) in tables may give the impression of 'attribute disclosure' i.e. that information has been disclosed about an individual. However as every household had a chance of being selected for swapping with one another area, there is doubt whether the value of one in any cell in a table is a 'real' one. **Further measures were then applied** such as changes to the design of the

tables, either by reducing the detail or excluding variables to manage the disclosure risk.

Data utility was always considered in parallel to the design of our SDC strategy. The impact on utility of key characteristics was monitored at low geographies during the swapping process. Great care was taken to achieve a balance between disclosure risk and data utility**.**

Assessing re-identification risk in protected outputs

Before releasing any multivariates (tables of two or more variables), two further precautions were followed.

First an assessment of proportions of true identity and attribute disclosure was carried out across releases subsequent to first release, and in particular on third and fourth releases to ensure it was sufficiently low. Disclosures introduced by processes of imputation and swapping would be considered *not true* disclosure. Tables were further modified where the disclosure risk was unacceptably high, for example further collapsing categories or removing variables.

Secondly an 'intruder testing' exercise was carried out on a sample representation of the most detailed tables which involved asking staff to act as intruders in attempting to claim disclosure. They were supplied with tables for their local area as well as public sources of information gathered from the web. In line with official guidance from the Information Commissioner's Office on a motivated intruder test, ONS looked at whether disclosure of personal information - without prior knowledge of a person - was reasonably likely, and the intruder's level of confidence in any claim and its likely impact. The positive results from these two assessments gave ONS the extra justification to release tabular outputs given that there was sufficient uncertainty.

- **Data sharing, Dissemination, Publication**

Outputs for the 2011 UK Census are being released over a number of months. The first release involved **estimates of the usual resident population broken down by age and sex**. As these counts were large, no further disclosure control was applied. Subsequent releases required careful table design. **Key and Quick Statistics** are univariate tables only, so were much less of a concern as compared to later releases which contain multivariate distributions down to the local area level.

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