Introduction to Census Microdata

Jo Wathan
UK Data Service

Fiona Cox
CALLS Hub

3 May 2017
The headline

- The census collects a wide range of topics at the individual and household level
- By bringing together a number of characteristics you can create powerful flexible analyses
- Census microdata contain individual records to allow you to do this
- BUT this has to be done with care to ensure that census records are kept confidential, so there constraints
What we will be covering today

• Where the data come from?
• What are the census microdata cross-sectional files? How might you use them?
• What are the longitudinal studies? How are these used, how can you access them?
What is a census?

• Main function to count the *population*
  • At one or more location

• Obtain some characteristics about the population

• Outputs at small geographies

• Informs public spending

• Used as a basis for other statistical systems
How were the data collected?

Self-completion

Post-out post back

Internet option

Census offices could focus resources on follow-up and hard to count areas
What questions were asked?

<table>
<thead>
<tr>
<th>Demographics</th>
<th>Household</th>
<th>Socio-economic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Tenure</td>
<td>Health</td>
</tr>
<tr>
<td>Sex</td>
<td>Accommodation type</td>
<td>Unpaid care</td>
</tr>
<tr>
<td>Country of Birth</td>
<td>Cars</td>
<td>Qualifications</td>
</tr>
<tr>
<td>Short-term residence</td>
<td>Central heating</td>
<td>Economic activity</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>No. of bedrooms</td>
<td>Occupation</td>
</tr>
<tr>
<td>Religion</td>
<td></td>
<td>Industry</td>
</tr>
<tr>
<td>Passports</td>
<td></td>
<td>Supervision</td>
</tr>
<tr>
<td>Language</td>
<td></td>
<td>Travel to work</td>
</tr>
<tr>
<td>National identity</td>
<td></td>
<td>FT / PT</td>
</tr>
<tr>
<td>Household relationships</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marital status</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Second residence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Migration</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The promise of confidentiality

The information you provided to us in the 2011 Census is confidential and protected by law.

The confidentiality of personal information is a top priority for the census. Your personal census information is not shared with any other government department, local councils or marketing companies.

Information collected in the 2011 Census will be used solely to produce statistics and for statistical research. These statistics will not reveal any personal information.

The paper questionnaires are scanned, then shredded, pulped and recycled. Census records are kept confidential for 100 years before being made available to the public. Census records remain closed while they are in the custody of the census offices. Records from the 2011 Census for England and Wales are not scheduled for public release before January 2112.

Office for National Statistics,
Census output types

Supported by the UK Data Service http://ukdataservice.ac.uk

- Microdata
  - Samples of census records
  - Detail limited to protect confidentiality
- Census area statistics
  - Counts of combinations of characteristics for areas
- Flow data
  - Counts of migrants from origin to destination
  - Or, between first and second address
  - Or home and place of work
- Shape files to enable these to be mapped

Supported by CALLS http://calls.ac.uk/

- Longitudinal microdata for each country
Individual level detail

- Individual records contain person and household characteristics
- Sample data only

<table>
<thead>
<tr>
<th>agegot</th>
<th>aprxsoer</th>
<th>cofbt</th>
<th>econoct</th>
<th>ethwlt</th>
<th>famcomp</th>
<th>health1</th>
<th>indust1</th>
<th>marstat1</th>
<th>numbhrs</th>
<th>occupa1</th>
</tr>
</thead>
<tbody>
<tr>
<td>55 to 6</td>
<td>DE</td>
<td>UK</td>
<td>Econom1</td>
<td>White</td>
<td>Married</td>
<td>Good he</td>
<td>Mining/</td>
<td>Married</td>
<td>No Code</td>
<td>Process</td>
</tr>
<tr>
<td>35 to 4</td>
<td>C2</td>
<td>UK</td>
<td>Econom1</td>
<td>White</td>
<td>Lone pa</td>
<td>Very go</td>
<td>Transpo</td>
<td>Single</td>
<td>Full-ti</td>
<td>Career</td>
</tr>
<tr>
<td>35 to 4</td>
<td>DE</td>
<td>UK</td>
<td>Econom1</td>
<td>White</td>
<td>Cohabit</td>
<td>Very go</td>
<td>Human h</td>
<td>Single</td>
<td>Full-ti</td>
<td>Career</td>
</tr>
<tr>
<td>16 to 2</td>
<td>C1</td>
<td>UK</td>
<td>Econom1</td>
<td>White</td>
<td>Cohabit</td>
<td>Good he</td>
<td>Financi</td>
<td>Single</td>
<td>Full-ti</td>
<td>Sales</td>
</tr>
<tr>
<td>45 to 5</td>
<td>C1</td>
<td>UK</td>
<td>Econom1</td>
<td>White</td>
<td>Cohabit</td>
<td>Very go</td>
<td>Wholesa</td>
<td>Divorce</td>
<td>Full-ti</td>
<td>Manage</td>
</tr>
<tr>
<td>55 to 6</td>
<td>C2</td>
<td>UK</td>
<td>Econom1</td>
<td>White</td>
<td>Married</td>
<td>Good he</td>
<td>Mining/</td>
<td>Married</td>
<td>Full-ti</td>
<td>Element</td>
</tr>
<tr>
<td>35 to 4</td>
<td>C2</td>
<td>UK</td>
<td>Econom1</td>
<td>White</td>
<td>Lone pa</td>
<td>Good he</td>
<td>Human h</td>
<td>Separat</td>
<td>Part-ti</td>
<td>Career</td>
</tr>
<tr>
<td>0 to 15</td>
<td>No code</td>
<td>UK</td>
<td>No code</td>
<td>White</td>
<td>Cohabit</td>
<td>Good he</td>
<td>No Code</td>
<td>Single</td>
<td>No Code</td>
<td>No Code</td>
</tr>
<tr>
<td>65 to 7</td>
<td>C1</td>
<td>UK</td>
<td>Econom1</td>
<td>White</td>
<td>Married</td>
<td>Very go</td>
<td>Mining/</td>
<td>Married</td>
<td>No Code</td>
<td>Process</td>
</tr>
<tr>
<td>55 to 6</td>
<td>DE</td>
<td>UK</td>
<td>Econom1</td>
<td>White</td>
<td>Not in</td>
<td>Fair he</td>
<td>Accommo</td>
<td>Divorce</td>
<td>No Code</td>
<td>Element</td>
</tr>
<tr>
<td>25 to 3</td>
<td>DE</td>
<td>UK</td>
<td>Econom1</td>
<td>White</td>
<td>Not in</td>
<td>Fair he</td>
<td>Mining/</td>
<td>Single</td>
<td>Full-ti</td>
<td>Process</td>
</tr>
<tr>
<td>0 to 15</td>
<td>No code</td>
<td>UK</td>
<td>No code</td>
<td>White</td>
<td>Married</td>
<td>Very go</td>
<td>Mining/</td>
<td>Single</td>
<td>No Code</td>
<td>No Code</td>
</tr>
<tr>
<td>16 to 2</td>
<td>C1</td>
<td>UK</td>
<td>Econom1</td>
<td>White</td>
<td>Not in</td>
<td>Very go</td>
<td>Mining/</td>
<td>Single</td>
<td>Full-ti</td>
<td>Sales</td>
</tr>
<tr>
<td>16 to 2</td>
<td>C1</td>
<td>UK</td>
<td>Econom1</td>
<td>White</td>
<td>Not in</td>
<td>Good he</td>
<td>No Code</td>
<td>Single</td>
<td>No Code</td>
<td>No Code</td>
</tr>
<tr>
<td>16 to 2</td>
<td>C1</td>
<td>UK</td>
<td>Econom1</td>
<td>White</td>
<td>Not in</td>
<td>Good he</td>
<td>No Code</td>
<td>Single</td>
<td>No Code</td>
<td>No Code</td>
</tr>
<tr>
<td>16 to 2</td>
<td>C1</td>
<td>UK</td>
<td>Econom1</td>
<td>White</td>
<td>Not in</td>
<td>Good he</td>
<td>No Code</td>
<td>Single</td>
<td>No Code</td>
<td>No Code</td>
</tr>
<tr>
<td>17 to 2</td>
<td>C1</td>
<td>UK</td>
<td>Econom1</td>
<td>White</td>
<td>Cohabit</td>
<td>Very go</td>
<td>Constru</td>
<td>Single</td>
<td>Full-ti</td>
<td>Process</td>
</tr>
<tr>
<td>16 to 2</td>
<td>C1</td>
<td>UK</td>
<td>Econom1</td>
<td>White</td>
<td>Cohabit</td>
<td>Very go</td>
<td>No code</td>
<td>Single</td>
<td>No Code</td>
<td>No Code</td>
</tr>
<tr>
<td>35 to 4</td>
<td>C1</td>
<td>UK</td>
<td>Econom1</td>
<td>White</td>
<td>Not in</td>
<td>Fair he</td>
<td>Mining/</td>
<td>Single</td>
<td>Full-ti</td>
<td>Process</td>
</tr>
<tr>
<td>0 to 15</td>
<td>No code</td>
<td>UK</td>
<td>No code</td>
<td>White</td>
<td>Married</td>
<td>Very go</td>
<td>Mining/</td>
<td>Single</td>
<td>No Code</td>
<td>No Code</td>
</tr>
<tr>
<td>55 to 6</td>
<td>DE</td>
<td>UK</td>
<td>Econom1</td>
<td>White</td>
<td>Married</td>
<td>Good he</td>
<td>Mining/</td>
<td>Married</td>
<td>No Code</td>
<td>Process</td>
</tr>
</tbody>
</table>
Why use this type of data?

• Very flexible
  • Can create your own tables
  • Can combine characteristics to create new ones
  • Can define sub populations
  • Can undertake multivariate analysis
• Bonus that can be used alongside count data
  • Microdata to design custom table design
  • Microdata to expand area-level characteristics
  • Microdata with area level characteristics added in
• But
  • Sample data, means that results are estimates
  • Geographical detail is limited
Reflecting confidentiality in microdata

Lots of geography

Safeguarded Local authority

Safeguarded regional

Teaching file

Lots of Socio-economic detail

Controlled data
Types of 2011 census microdata

• Being released by each census office separately

• Teaching files (Open) – available from census offices for some time (1% regional taster file)
• Safeguarded regional files (5%)
• Safeguarded grouped local files (5%)
• Controlled access secure files – individual and household (10% each) England and Wales
<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Teaching</th>
<th>Safeguarded Regional</th>
<th>Safeguarded grouped LA</th>
<th>Secure individual</th>
<th>Secure household</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of variables</td>
<td>18</td>
<td>121</td>
<td>121</td>
<td>258</td>
<td>245</td>
</tr>
<tr>
<td>Smallest geography</td>
<td>Region</td>
<td>Region</td>
<td>LAD &gt; 120,000</td>
<td>LAD</td>
<td>LAD</td>
</tr>
<tr>
<td>Licence</td>
<td>OGL</td>
<td>End User Licence</td>
<td>End User Licence</td>
<td>Approved Researcher</td>
<td>Approved Researcher</td>
</tr>
<tr>
<td>Sample size</td>
<td>n=569,741 1%</td>
<td>n=2,858,155 5%</td>
<td>n=2,795,020 5%</td>
<td>10% n&gt;5 million</td>
<td>10% n&gt;5 million</td>
</tr>
<tr>
<td>Age detail</td>
<td>8 age groups</td>
<td>Individual yrs to 70, then 5 groups</td>
<td>Individual yrs to 94, then 95-96, 97+</td>
<td>Individual yrs to 94, then 95-96, 97+</td>
<td></td>
</tr>
<tr>
<td>Occupation detail</td>
<td>9 classes</td>
<td>25 classes</td>
<td>25 classes</td>
<td>Full: 369 classes of SOC plus ISCO</td>
<td>Full: 369 classes of SOC plus ISCO</td>
</tr>
<tr>
<td>Ethnic Group</td>
<td>5 classes</td>
<td>18 classes</td>
<td>13 classes</td>
<td>Full: 151 categories</td>
<td>Full: 151 categories</td>
</tr>
</tbody>
</table>
Producing maps using the grouped LA files

Percentage of residents who are have lived in GB since 2006 or earlier who do not speak English well

Census Microdata Individual Safeguarded Sample 2011 (Local Authority): England and Wales
### Population Base – Short term migrants

“A STR was defined in the 2011 Census as anyone living in England and Wales who was born outside the UK and who intended to stay in the UK for a period of between three and 12 months, for any reason.”  
ONS 2013

### Full-time students/schoolchildren by population base and age
Census Microdata Teaching File for England and Wales 2011

<table>
<thead>
<tr>
<th>Population Base</th>
<th>Age group</th>
<th>Less than 16</th>
<th>16-24</th>
<th>25-34</th>
<th>35+</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Usual resident</strong></td>
<td>Student</td>
<td>Yes</td>
<td>81737</td>
<td>31166</td>
<td>3608</td>
</tr>
<tr>
<td></td>
<td>Student</td>
<td>No</td>
<td>23993</td>
<td>35122</td>
<td>71608</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td>105730</td>
<td>66288</td>
<td>75216</td>
</tr>
<tr>
<td><strong>Student living away from home during term-time</strong></td>
<td>Student</td>
<td>Yes</td>
<td>944</td>
<td>5572</td>
<td>185</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td>944</td>
<td>5572</td>
<td>185</td>
</tr>
<tr>
<td><strong>Short-term resident</strong></td>
<td>Student</td>
<td>Yes</td>
<td>115</td>
<td>711</td>
<td>233</td>
</tr>
<tr>
<td></td>
<td>Student</td>
<td>No</td>
<td>43</td>
<td>214</td>
<td>314</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td>158</td>
<td>925</td>
<td>547</td>
</tr>
</tbody>
</table>
Research: e.g. on caring

Kinship care re-visited: using census 2011 microdata to examine and map the extent and nature of kinship care households in the UK
Accessing data

- Full information on the census microdata web page on UK Data Service
- Open data – available in spreadsheet format from census offices (also available as SPSS/Stata/Nesstar format from UK Data Service without registration)
- Safeguarded data: 1991, most 2001, Safeguarded 2011 UK Data Service, online registration
- Controlled data via census offices only
Census microdata

We provide access to cross-sectional census microdata, these are anonymised individual level records sampled from a single census. They are large individual level files which resemble the sort of data that might be collected directly from a questionnaire. Because individual records contain a wide range of individual and household characteristics, the data enable multivariate analyses for a wide range or purposes.

The UK Data Service holds contemporary data from the 1991 to 2011 Censuses and microdata samples from 1961 onwards are planned.

- Guide to census microdata
- 2011 progress

SAR files are available at four levels, two of which can currently be accessed through the UK Data Service in the same manner as survey data.
Also of interest & on the horizon

Related sources
• I-CeM: Integrated Census Microdata 1851-1911  
  http://icem.data-archive.ac.uk/
• IPUMS: International Public Use Microdata Sample collection from 79 countries, starting 1960  
  https://international.ipums.org/international/
• EEHCM: Historic Census Microdata 1961-1981 Microdata files

Census Research Conference 22nd June
   London, RSS
   Booking:
   https://www.ukdataservice.ac.uk/news-and-events/eventsitem/?id=4945
An introduction to the UK Census-based Longitudinal Studies

Dr Fiona Cox
Project Manager, CALLS Hub
Structure of the LSs

http://calls.ac.uk/ls-units/

Northern Ireland Longitudinal Study
Maintained by NISRA
Supported by NILS-RSU

Scottish Longitudinal Study
Maintained & Supported by SLS-DSU

ONS LS (England & Wales)
Maintained by ONS
Supported by CeLSIUS
<table>
<thead>
<tr>
<th>Study sample</th>
<th>ONS Longitudinal Study</th>
<th>Scottish Longitudinal Study</th>
<th>Northern Ireland Longitudinal Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study sample</td>
<td>1% of the population of England &amp; Wales</td>
<td>5% of the population of Scotland</td>
<td>28% of the population of Northern Ireland</td>
</tr>
<tr>
<td>Records from 2011 Census</td>
<td>580,000</td>
<td>~270,000</td>
<td>~500,000</td>
</tr>
<tr>
<td>Academic user support</td>
<td>Centre for Longitudinal Study Information &amp; User Support, University College London</td>
<td>Scottish Longitudinal Study Development &amp; Support Unit, University of St. Andrews &amp; University of Edinburgh</td>
<td>Northern Ireland Longitudinal Study Research Support Unit, Queens University Belfast</td>
</tr>
<tr>
<td>Census data available</td>
<td>Complete census data for study members and for people living in the same household as a study member</td>
<td>Complete census data for study members and for people living in the same household as a study member</td>
<td>Complete census data for study members and for people living in the same household as a study member</td>
</tr>
</tbody>
</table>
| Event data available | Civil registration system  
- Births of sample members  
- Births to sample mothers  
- Stillbirths / Infant deaths  
- Deaths of sample members  
- Widow(er)hoods  
NHS Central Register  
- Immigration  
- Emigration  
- Minor events  
Cancer registries  
- Cancer data | Civil registration system  
- Births of sample members  
- Births to sample mothers  
- Births to sample fathers  
- Stillbirths / Infant deaths  
- Marriages  
- Deaths of sample members  
- Widow(er)hoods  
NHS Central Register  
- Immigration  
- Emigration  
Scottish Govt. Education Directorate  
- School level education data including attendance, exclusions, attainment and qualifications  
Other data available, subject to approval  
- Hospital episodes  
- Maternity data  
- Cancer data | Civil registration system  
- Births of sample members  
- Births to sample mothers  
- Births to sample fathers  
- Infant mortality  
- Marriages  
- Deaths of sample members  
Health card registration system  
- Immigration  
- Emigration  
- Internal migration  
Land & Property Services  
- Housing data  
Health & Social Care  
- Health data linked in one-off distinct linkage projects (e.g. breast screening, dental treatments) subject to approval
Geographies in the LSs

- **Super Output Area (approx 2,000 persons)**
- **Output Area**
- **County District level (or Ward groupings to equivalent size)**

Lowest geographies allowed (lower levels may be used by RSU staff, eg for linking data or creating derived variables)
Longitudinal analysis

- Adds the dimension of time to the analysis
- Allows examination of the effect of policy, personal or environmental changes
- Allows researcher to better establish causality
Longitudinal Analysis

- Comparison of same group over time (Age effects)
- Comparisons over time (Period comparisons)
- Comparisons between cohorts over time (Cohort effects)

(Source: Findlay, McCollum et al., 2015)
New mobilities across the lifecourse, *Population Space and Place*, 21)
Uses of LS data

Citation analysis of LS-based papers 2010-16
(more detail on CALLS Hub blog: http://calls.ac.uk/research-blog/)

- Medicine 48%
- Social Sciences 26%
- No SCImago entry 3%
- Psychology 1%
- Nursing 2%
- Environmental Science 3%
- Economics, Econometrics and Finance 4%
- Business, Management & Accounting 1%
- Dentistry 1%
- Biochemistry, Genetics & Molecular Biology 4%
- Arts & Humanities 5%
- Agricultural & Biological Sciences 2%
Impact examples from the LSs

- ONS LS has made important contributions to policy, e.g., Dilnot Commission (2011), Marmot Review (2010)
- SLS research on changes in patterns of tenure type informed housing policy
- NILS research on patterns of uptake of breast cancer screening informed public health strategy
Accessing the data

- The LSs are free to use
- Researchers welcome from:
  - academia, including research students
  - Government and policy groups
  - Third sector organisations
- Researchers from other organisations or overseas should contact CALLS to discuss access
Accessing the data

Because of the sensitive nature of the data, the LSs may only be accessed within designated ‘safe-settings’

NILS-RSU, Belfast
SLS-DSU, Edinburgh
ONS VMLs, London, Titchfield & Newport
Accessing the data

- Researchers are required to complete an application process, and to undergo training before access to the LSs.
- Information on the steps involved are available in the Guides & Resources section of our website: http://calls.ac.uk/guides-resources/
New developments

(more info at http://calls.ac.uk/guides-resources/)

- It is possible to use more than one LS in your analyses through the **eDatashield** methodology developed at SLS-DSU

- **Synthetic** versions of core LS variables are available to download from the CALLS website

- For the SLS (hopefully soon for ONS LS and NILS) it may be possible to receive a **synthetic** version of your project dataset to allow development of syntax and models prior to using the real data
More information at calls.ac.uk
Get in touch:

website: calls.ac.uk

twitter: @CALLS_Hub

email: info@calls.ac.uk