Introduction to the Longitudinal Study and census microdata

Dr Oliver Duke-Williams

@oliver_dw

Student Research Using the ONS LS and Census Microdata (aka SARs/Samples of Anonymised Records)
Census data in the UK / Great Britain

• Censuses held since 1801
• Current arrangement is three separate but coordinated censuses (E&W, S, NI)
• Users are generally most familiar with aggregate data
Types of census output

<table>
<thead>
<tr>
<th>ID</th>
<th>Age</th>
<th>Gender</th>
<th>Occupation</th>
</tr>
</thead>
<tbody>
<tr>
<td>0001</td>
<td>23</td>
<td>F</td>
<td>Teacher</td>
</tr>
<tr>
<td>0002</td>
<td>17</td>
<td>M</td>
<td>Student</td>
</tr>
<tr>
<td>0003</td>
<td>45</td>
<td>F</td>
<td>Doctor</td>
</tr>
</tbody>
</table>

Individual data (microdata)

Area data (aggregate data)

<table>
<thead>
<tr>
<th>Age</th>
<th>M</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-4</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>5-9</td>
<td>8</td>
<td>9</td>
</tr>
</tbody>
</table>

Flow data (O/D data)

<table>
<thead>
<tr>
<th>Age</th>
<th>M</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-15</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

Boundary data
Census microdata

- Often referred to as ‘individual data’
- Two types
  - Cross-sectional microdata
  - Longitudinal microdata
Census microdata in the UK

• First produced from the 1991 Census; labelled as ‘Sample of Anonymised Records’ or SARs
  • Separate files for Great Britain and for Northern Ireland
  • Separate individual and household samples
• Changes in sample sizes and license arrangements over time
‘Samples of Anonymised Records’

<table>
<thead>
<tr>
<th></th>
<th>Open</th>
<th>Safeguarded</th>
<th>Secure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1991</td>
<td>Individuals</td>
<td>2% (GB,NI)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Households</td>
<td>1% (GB,NI)</td>
<td></td>
</tr>
<tr>
<td>2001</td>
<td>Individuals</td>
<td>3% (UK)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Households</td>
<td></td>
<td>1% (EW)</td>
</tr>
<tr>
<td></td>
<td>‘Small Area Microdata’</td>
<td></td>
<td>5% (UK)</td>
</tr>
</tbody>
</table>
‘Census Microdata’

<table>
<thead>
<tr>
<th>Year</th>
<th>Category</th>
<th>Open</th>
<th>Safeguarded</th>
<th>Secure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1961</td>
<td>Individuals</td>
<td>1% (GB)</td>
<td>5% (GB)</td>
<td>9% (GB)</td>
</tr>
<tr>
<td></td>
<td>Households</td>
<td></td>
<td>0.95% (GB)</td>
<td></td>
</tr>
<tr>
<td>1971</td>
<td>Individuals</td>
<td>1% (GB)</td>
<td>5% (GB)</td>
<td>9% (GB)</td>
</tr>
<tr>
<td></td>
<td>Households</td>
<td></td>
<td>0.95% (GB)</td>
<td></td>
</tr>
<tr>
<td>1981</td>
<td>Individuals</td>
<td>1% (GB)</td>
<td>5% (GB)</td>
<td>9% (GB)</td>
</tr>
<tr>
<td></td>
<td>Households</td>
<td></td>
<td>0.95% (GB)</td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>Individual</td>
<td></td>
<td>5% (EW,S,NI)</td>
<td>10% (EW,S,NI)</td>
</tr>
<tr>
<td></td>
<td>Households</td>
<td></td>
<td></td>
<td>10% (EW,S,NI)</td>
</tr>
<tr>
<td></td>
<td>Individual (‘Teaching’)</td>
<td></td>
<td>1% (EW,S,NI)</td>
<td></td>
</tr>
</tbody>
</table>
What is in the microdata files?

• Broadly, all data, except personal identifiers

But!

• Trade-offs in level of detail, completeness of data etc with access arrangements and sample size
What can we do with microdata

- Regression modelling etc on individual characteristics
- Multi-level modelling on individual / household / local area characteristics
- Cross tabulation of fields in combinations not found in the standard outputs, e.g....
2011, England and Wales
Travel to work by general health

Longitudinal data

- There are three longitudinal studies in the UK
- They have differ sample sizes and cover different time periods
- They also differ in the range and amount of linked data
- All have secure access arrangements
# UK Census Longitudinal Studies

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ONS Longitudinal Study</td>
<td>4/365.25</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Scottish Longitudinal Study</td>
<td>20/365.25</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Northern Ireland Longitudinal Study</td>
<td>104/365.25</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
What is in the LS?

- Similarly to census microdata, all variables apart from individual identifiers
  - More detail than the safeguarded and open microdata files
  - Some (detailed) variables have additionally restricted access
- Sample members and others in the household
Census Data

From each census
- LS sample members
- Other household members
- Age, sex, marital status, country of birth
- Family and household types, communal establishments
- Housing: tenure, rooms and amenities
- Qualifications, economic activity, occupation, industry and social class
- Travel to work and one-year migration
- Geographical information

More recent censuses
- Ethnicity (1991-2011)
- National identity (2011)
- Year of arrival (2011)
- Care-giving (2001, 2011)
- Short-term migration (2011)
- Main language (2011)
LS structure: England & Wales

![Diagram showing linked events and sample members for various censuses between 1971 and 2012.]

**Entrants (1971-2012)**
- Births: 303,000
- Immigrants: 202,000
- Re-entries: 25,000

**Exits (1971-2012)**
- Deaths: 262,000
- Embarks: 45,000
- Enlistments: 7,600

**Linked Events (1971-2012)**
- Live births to sample women: 292,000
- Still births to sample women: 1,700
- Infant deaths: 2,400
- Widow(er)hoods: 90,000
- Cancer registrations: 140,000

**1971 Census**
- Sample members selected: 530,000
- Traced: 513,000

**1981 Census**
- Sample members selected: 536,000
- Traced: 530,000

**1991 Census**
- Sample members selected: 544,000
- Traced: 535,000

**2001 Census**
- Sample members selected: 540,000
- Traced: 537,000

**2011 Census**
- Sample members selected: 590,000
- Traced: 582,000
Who uses the LS?

Source: Cox, F (2017) CALLS Hub Citation Analysis
https://calls.ac.uk/research-blog/
Travel to work by general health 2001 v 2011
Comparing cross-sectional microdata

2001

2011

Sources:

• Previous slide shows travel to work in 2001 and in 2011
• We can compare cross-sections easily enough
• But: we don’t know whether those that used (mode x) in 2001 were the same people that used (mode x) in 2011, unless we use longitudinal data...
### Travel to work 2001 v 2011: longitudinal transitions

For each 2001 mode

![Most common 2011 outcome](image1)

![Second most common 2011 outcome](image2)

<table>
<thead>
<tr>
<th>Mode in 2001</th>
<th>Train</th>
<th>Tube</th>
<th>Bus</th>
<th>Driving a car or van</th>
<th>Passenger in a car or van</th>
<th>Motorcycle</th>
<th>Bicycle</th>
<th>On foot</th>
<th>Work mainly at or from home</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>41%</td>
<td>6%</td>
<td>4%</td>
<td>32%</td>
<td>2%</td>
<td>1%</td>
<td>2%</td>
<td>5%</td>
<td>7%</td>
<td>6324</td>
</tr>
<tr>
<td>Tube/Metro/Light rail/tram (E&amp;W)</td>
<td>17%</td>
<td>33%</td>
<td>8%</td>
<td>24%</td>
<td>2%</td>
<td>1%</td>
<td>3%</td>
<td>5%</td>
<td>7%</td>
<td>3849</td>
</tr>
<tr>
<td>Bus/Minibus/Coach</td>
<td>5%</td>
<td>3%</td>
<td>27%</td>
<td>39%</td>
<td>8%</td>
<td>0%</td>
<td>2%</td>
<td>12%</td>
<td>3%</td>
<td>10638</td>
</tr>
<tr>
<td>Drive car/van</td>
<td>2%</td>
<td>1%</td>
<td>2%</td>
<td>82%</td>
<td>2%</td>
<td>1%</td>
<td>1%</td>
<td>4%</td>
<td>5%</td>
<td>93087</td>
</tr>
<tr>
<td>Passenger car/van</td>
<td>3%</td>
<td>1%</td>
<td>8%</td>
<td>53%</td>
<td>20%</td>
<td>1%</td>
<td>2%</td>
<td>10%</td>
<td>3%</td>
<td>10522</td>
</tr>
<tr>
<td>Motor cycle/Scooter/Moped</td>
<td>4%</td>
<td>2%</td>
<td>3%</td>
<td>56%</td>
<td>4%</td>
<td>19%</td>
<td>6%</td>
<td>4%</td>
<td>3%</td>
<td>1840</td>
</tr>
<tr>
<td>Bicycle</td>
<td>3%</td>
<td>1%</td>
<td>4%</td>
<td>43%</td>
<td>4%</td>
<td>2%</td>
<td>30%</td>
<td>10%</td>
<td>3%</td>
<td>4415</td>
</tr>
<tr>
<td>On foot</td>
<td>3%</td>
<td>2%</td>
<td>7%</td>
<td>40%</td>
<td>6%</td>
<td>1%</td>
<td>3%</td>
<td>34%</td>
<td>4%</td>
<td>14621</td>
</tr>
<tr>
<td>Works at/from home</td>
<td>3%</td>
<td>1%</td>
<td>2%</td>
<td>54%</td>
<td>3%</td>
<td>0%</td>
<td>1%</td>
<td>7%</td>
<td>28%</td>
<td>14035</td>
</tr>
<tr>
<td>Total</td>
<td>7381</td>
<td>3428</td>
<td>7323</td>
<td>105507</td>
<td>6879</td>
<td>1248</td>
<td>4053</td>
<td>13075</td>
<td>10437</td>
<td>159331</td>
</tr>
</tbody>
</table>

- 30% retention rate for cyclists
- 34% retention rate for walkers

Source: ONS Longitudinal Study
Comparison with birth cohorts

- People unfamiliar with the LSes are often more familiar with the idea of birth cohort studies
  - Birth cohorts draw a sample of persons born in a particular year
  - Census longitudinal studies draw a sample across all persons regardless of age
Comparison with birth cohorts

• Sample sizes
  • LS total samples are much bigger
  • LS individual year-of-age samples are smaller in England and Wales, but less subject to attrition
    • Starting cohort sizes 17-19K
    • ONS LS, 2011 per single year of age < 65 c. 5-8K

• Content
  – Cohort studies have much broader content
  – Sample size allows LS to have more detailed geography etc
Acknowledgements

• The permission of the Office for National Statistics (ONS) to use the Longitudinal Study is gratefully acknowledged, as is the help provided by staff of the Centre for Longitudinal Study Information & User Support (CeLSIUS). CeLSIUS is supported by the ESRC Census of Population Programme (Award Ref: ES/K000365/1). The authors alone are responsible for the interpretation of the data.