SMALL AREA ESTIMATION STRATEGY FOR THE 2011 UK CENSUS

Denise Britz do N. Silva (IBGE / ENCE)
Bernard Baffour (University of Southampton)
Christine Sexton and Alan Taylor (ONS-UK)
Aline Veiga (UERJ)

Outline

• Brief description of UK Census Coverage Adjustment Strategy
• Specify need for Small Area Estimation
• Description of Simulation Study
• Presentation of Results
• Future Work
UK Census Coverage Assessment Process

• **Stage 1**: Census – complete enumeration of whole population

• **Stage 2**: CCS – intensive re-enumeration of sample of population

• **Stage 3**: Matching of the Census and CCS database to estimate level of undercount (at Estimation Area level)

• **Stage 4**: Obtain population estimates for Local Authorities, also adjusted for undercount (*small area modelling*)
2011 Census

Census Coverage Survey (CCS)

Matching Census and CCS

Dual System Estimation

Ratio estimation

Small Area Estimation

EA popn estimates

National population estimates

LA population estimates

Impute individual and household records, controlled to LA estimates

Demographic, Survey and Administrative data

Quality Assurance using Demographic Analysis, Survey data, Qualitative information, Administrative records and Expert Analysis

The need for small area modelling

• Census will have undercount, so Coverage Strategy has to adjust for this

• At Estimation Area level (popn > 500,000), direct estimation using CCS possible

• Local Authority estimates, however, less precise under direct estimation

• Indirect Estimation more precise, but can be biased

• Small Area Strategy is to strike **balance** between the (imprecise) direct and (biased) indirect estimates
Model based Small Area Estimation in the Census

- Accurate population estimates required, stratified by age-sex and other key demographic variables

- The CCS is not designed to provide suitably precise estimates for such small domains

- Ideally want to exploit similarity between groups of ‘small’ areas

- Techniques based on regression models, looking at

\[ E[Y_{kadlg} | X_{kadlg}] = \theta_{adlg} X_{kadlg} \]

for postcode \( k \), age-sex group \( a \), HtC stratum \( d \), Local Authority \( l \) in Estimation Area \( g \)
Small Area Models considered (1)

- **Direct Estimator** (ratio estimation)
  - only uses data from postcodes in the specific LA
  - exploit the similarity within the LA
  - Calculate ratio factor based on census and DSE count
  - Explicit use of information (so less bias)

\[
\hat{T}_{ad1g} = \sum_{k \in s_{dl}} \frac{\sum_{a \in c} \sum_{k \in s_{dl}} Y_{kad1g}}{\sum_{k \in s_{dl}} \sum_{a \in c} X_{kad1g}} X_{kad1g} = \hat{\theta}_{cd1g} X_{ad1g}
\]
Small Area Models considered (2)

• **Synthetic estimator**
  
  – uses data from all the LAs within a specific EA
  
  – explores similarity of sampled postcodes within the specific EA
  
  – fits a regression model with *age-sex group effects*

\[
Y_{kadl} = \theta_{ad} X_{kadl} + \varepsilon_{kadl} \sqrt{X_{kadl}}
\]

*Age-sex effect*
Small Area Models considered (3)

- **Local fixed model**
  - uses data from postcodes of all LAs in a specific EA
  - exploits the similarity of sampled postcodes within an EA
  - collapsed age-sex group effects used to further improve precision
  - used in the 2001 Census
  - similar to the synthetic estimator, but difference is that an effect for each LA included

\[
Y_{kadl} = \left( \theta_{cd} + \gamma_{dl} \right) X_{kadl} + \varepsilon_{kadl} \sqrt{X_{kadl}}
\]
Small Area Models considered (4)

- Investigate whether by considering more broader areas than EA, effective sample size can be increased
- Explore similarities between groups of LAs within the same GOR
- Use all the postcodes in all the LAs within a given GOR
- More complex models can be fitted (i.e. with random effects)

- **Regional Fixed Effects model**
  - Fixed LA effects
  - Fixed age-sex effects

- **Regional Random Effects model**
  - Random LA effects
  - Fixed age-sex effects
Simulation Framework

- Small Area work follows the framework described in earlier presentation
- Simulated 400 ‘Censuses’ and 400 ‘CCSs’
- Estimation of EA population totals found for each Census-CCS combination
- Objective is to estimate LA population totals by HtC and age-sex groups
- Results are for 8 specially chosen EAs and GORs (South East and North East)
Performance measures

• Total Population Estimates derived for 105 domains (3 HtC x 35 age-sex groups)

• Relative Root Mean Squared Error (RRMSE)

\[
RRMSE(\hat{T}_{adl}) = \frac{1}{\hat{T}_{adl}} \sqrt{\frac{400}{\sum_{j=1}^{400} (\hat{T}_{adlj} - T_{adl})^2}}
\]

• Relative Bias (RB)

\[
RB(\hat{T}_{adl}) = \frac{1}{T_{adl}} \frac{400}{\sum_{j=1}^{400} (\hat{T}_{adlj} - T_{adl})}^2
\]

Local Authority Population Total, by age-sex and HtC group
An alternative measure of precision

\[ P_m = \frac{1}{400} \sum_{j=1}^{400} I \left( \min \left[ \left| \hat{T}_{adljm} - T_{adl} \right| \right] \right) \]

\( I(\cdot) \) is an indicator showing whether or not small area model \( m \) produces the estimate closest to the true population.
Results

- Results presented for the different SAE methods
- Distribution of the RRMSEs and RBs for the different small area models investigated
- Boxplots: good SAE method has lower medians, smaller spread, few outliers
- Also looking at the ‘alternative precision’ of the estimates
## Coverage Rates

<table>
<thead>
<tr>
<th>Estimation Area</th>
<th>Local Authority</th>
<th>Description</th>
<th>LA Coverage Rate in 2001 (%)</th>
<th>EA Coverage Rate in 2001 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>KK</td>
<td>00FY</td>
<td>Nottinghamshire</td>
<td>91.42</td>
<td>95.5</td>
</tr>
<tr>
<td></td>
<td>37UD</td>
<td></td>
<td>98.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>37UJ</td>
<td></td>
<td>97.17</td>
<td></td>
</tr>
<tr>
<td>KO</td>
<td>00CQ</td>
<td>West Midlands</td>
<td>92.39</td>
<td>95.2</td>
</tr>
<tr>
<td></td>
<td>00CT</td>
<td></td>
<td>98.02</td>
<td></td>
</tr>
<tr>
<td>LB</td>
<td>00AM</td>
<td>Inner London</td>
<td>73.28</td>
<td>76.5</td>
</tr>
<tr>
<td></td>
<td>00AU</td>
<td></td>
<td>79.32</td>
<td></td>
</tr>
<tr>
<td></td>
<td>00BG</td>
<td></td>
<td>76.93</td>
<td></td>
</tr>
<tr>
<td>LJ</td>
<td>00AB</td>
<td>Outer London</td>
<td>87.80</td>
<td></td>
</tr>
<tr>
<td></td>
<td>00BC</td>
<td></td>
<td>88.38</td>
<td></td>
</tr>
<tr>
<td></td>
<td>00BH</td>
<td></td>
<td>88.93</td>
<td></td>
</tr>
</tbody>
</table>
Estimation Area KK – Nottinghamshire

Local Fixed

Synthetic

Relative Bias (%)

Relative Mean Square Error (%)
Estimation Area KO – West Midlands

Relative Bias (%)  Relative Mean Square Error (%)
Estimation Area LB – Inner London

Relative Bias (%)  Relative Mean Square Error (%)
Estimation Area LJ – Outer London

Relative Bias (%)  Relative Mean Square Error (%)
# Alternative Measure of Precision

Proportion of times SAE method gets closest to ‘truth’

<table>
<thead>
<tr>
<th>Estimation Area</th>
<th>Local Authority</th>
<th>Small Area Estimation Models</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Direct</td>
</tr>
<tr>
<td>KK</td>
<td>00FY</td>
<td>0.35</td>
</tr>
<tr>
<td></td>
<td>37UD</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>37UJ</td>
<td>0.31</td>
</tr>
<tr>
<td>KO</td>
<td>00CQ</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>00CT</td>
<td>0.36</td>
</tr>
<tr>
<td>LB</td>
<td>00AM</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td>00AU</td>
<td>0.31</td>
</tr>
<tr>
<td></td>
<td>00BG</td>
<td>0.30</td>
</tr>
<tr>
<td>LJ</td>
<td>00AB</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td>00BC</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>00BH</td>
<td>0.19</td>
</tr>
</tbody>
</table>
Conclusions

• Synthetic Estimator and Local Fixed Model are reasonable options to produce LA population estimates

• In general accounting for the age-sex differentials captures most of the differences in small areas, so synthetic estimation is adequate

• However, in cases where specific LAs behave differently to the EA, the local fixed model is needed

• Regional Models do not improve upon the precision
Number of times (out of 400 simulations) the Local Authority effects are significant (p-value < 0.05)

<table>
<thead>
<tr>
<th>Estimation Area</th>
<th>Hard to Count Stratum</th>
<th>Number of times</th>
<th>Proportion of times</th>
</tr>
</thead>
<tbody>
<tr>
<td>KK</td>
<td>1</td>
<td>298</td>
<td>0.745</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>270</td>
<td>0.675</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>236</td>
<td>0.590</td>
</tr>
<tr>
<td>KO</td>
<td>1</td>
<td>196</td>
<td>0.490</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>174</td>
<td>0.435</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>50</td>
<td>0.125</td>
</tr>
<tr>
<td>LB</td>
<td>2</td>
<td>142</td>
<td>0.355</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>311</td>
<td>0.778</td>
</tr>
<tr>
<td>LJ</td>
<td>1</td>
<td>134</td>
<td>0.335</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>223</td>
<td>0.558</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>208</td>
<td>0.520</td>
</tr>
</tbody>
</table>
Strategy in 2011 Census

• The synthetic model is taken to be the default
• However, if LA effects are significant then the LA fixed effects model is implemented
• Similar to 2001 but more robust to (better) cope with differences in coverage
References


• Abbott, O., Brown, J., Chambers, R. and Cruddas, M., (2000a). One Number Census Local Authority Estimation. Paper submitted to the One Number Census Steering Committee numbered as ONS(ONC(SC)00/03B)). (available at the ONS website)
