Mass imputation for census estimation

Jacco Daalmans

Funded by Eurostat

The views expressed in this presentation are those of the author(s) and do not necessarily reflect the policies of the European Commission or of Statistics Netherlands. Neither the European Commission nor Statistics Netherlands are responsible for any use that may be made of the information this presentation contains.
Dutch Census
Virtual census: linking available sources

- Employment registers
- Social security registers
- Pension registers
- Municipal population registers
- Tax registers
- Labour force surveys
- Public Employment service register
Census 2011

All census variables, except
- Educational attainment
- Occupation

Registers

Sample Surveys

- Educational attainment
- Occupation

16,655,799
ca 300,000
Census 2011

All census variables, except
- Educational attainment
- Occupation

Registers

Sample Surveys

Educational attainment
Occupation

Units

Variables

16,655,799
ca 300,000
Educational attainment File (EAF)

- All available information on educational attainment
  (Registers and sample surveys)

- Coverage of register data is selective

- Educational levels known > 10,000,000 people (2011)
**Educational attainment file 2011**

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Registerpart</td>
<td>Remaining part – LFS data unavailable</td>
<td>Remaining part - LFS data available</td>
</tr>
<tr>
<td>N=6,456,834</td>
<td>N= 6,951,418</td>
<td>N= 340,472</td>
</tr>
<tr>
<td>[SELECTIVE!]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N=13,748,724 (&gt;=15 age)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Estimation of remaining part: (mass) imputation versus weighting
Educational attainment file 2011

A

Registerpart
N=6,456,834

[SELECTIVE!]

N=13,748,724 (>=15 age)

B

Remaining part –
LFS data unavailable
N= 6,951,418

C

Remaining part - LFS data available
N= 340,472

Mass imputation
Mass imputation

Logistic regression (continuation ratio model)
Census variables and income as covariates

Procedure:
1) For each person estimate probabilities for all educational categories (given his/her characteristics).

Example: none: 0.20; prim: 0.10; secL: 0.50; secH: 0.10
post: 0.05; tert1: 0.03; tert2: 0.02
Mass imputation

Logistic regression (continuation ratio model)
Census variables and income as covariates

Procedure:
1) For each person estimate probabilities for all educational categories
   (given his/her characteristics).

2) Assign one category to each person
   (using the estimated probabilities).
Results

- Imputation of 6,951,418 records technically feasible (using R)

- Differences in results with official EAF publications and Census 2011:

  small at aggregate level and larger at detailed level
## EDU level in % - population 15 years and older

<table>
<thead>
<tr>
<th>Education Level</th>
<th>Mass imputation (based on EAF)</th>
<th>Official EAF publication</th>
<th>EU Census 2011 (based on LFS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. No formal education</td>
<td>2.4</td>
<td>2.1</td>
<td>1.6</td>
</tr>
<tr>
<td>2. Primary education</td>
<td>9.0</td>
<td>8.5</td>
<td>8.4</td>
</tr>
<tr>
<td>3. Lower secondary education</td>
<td>24.1</td>
<td>23.4</td>
<td>24.9</td>
</tr>
<tr>
<td>4. Upper secondary education</td>
<td>37.3</td>
<td>37.6</td>
<td>34.7</td>
</tr>
<tr>
<td>5. Post secondary education</td>
<td>3.6</td>
<td>3.6</td>
<td>2.8</td>
</tr>
<tr>
<td>6. First stage tertiary education</td>
<td>23.5</td>
<td>24.6</td>
<td>25.8</td>
</tr>
<tr>
<td>7. Second stage tertiary education</td>
<td>0.2</td>
<td>0.2</td>
<td>0.5</td>
</tr>
<tr>
<td>8. Unknown</td>
<td>--</td>
<td>--</td>
<td>1.3</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>100.0</strong></td>
<td><strong>100.0</strong></td>
<td><strong>100.0</strong></td>
</tr>
</tbody>
</table>
CENSUS hypercube 24.2 (6 dimensional table); (GEO.L x LPW.N x SEX x POB.M x AGE.L x EDU) relative difference mass imputation versus EU Census 2011; for different cell sizes
Cross validation

Estimate educational attainment levels that are actually known

Based on 340,472 observations

<table>
<thead>
<tr>
<th>Education Level</th>
<th>Mass imputed</th>
<th>Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. No formal education</td>
<td>1.4</td>
<td>1.4</td>
</tr>
<tr>
<td>2. Primary education</td>
<td>8.1</td>
<td>8.2</td>
</tr>
<tr>
<td>3. Lower secondary education</td>
<td>25.0</td>
<td>25.1</td>
</tr>
<tr>
<td>4. Upper secondary education</td>
<td>38.5</td>
<td>38.6</td>
</tr>
<tr>
<td>5. Post secondary education</td>
<td>5.4</td>
<td>5.4</td>
</tr>
<tr>
<td>6. First stage tertiary education</td>
<td>21.2</td>
<td>21.0</td>
</tr>
<tr>
<td>7. Second stage tertiary educat.</td>
<td>0.4</td>
<td>0.4</td>
</tr>
<tr>
<td>8. Unknown</td>
<td>1.4</td>
<td>1.4</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>100.0</strong></td>
<td><strong>100.0</strong></td>
</tr>
</tbody>
</table>
Future activities

- Explain differences in results

- Criterion ‘sufficiently reliable’ for publication

- What to do with occupation?
  Use new source? (in addition to LFS)
  (linkedin data, occupation registers for certain professions)
Thank you

Very much!