A FRAMEWORK FOR ASSESSING PERTURBATIVE METHODS FOR PROTECTION OF CENSUS 2021 DATA AT STATISTICS PORTUGAL

Inês Rodrigues (ines.rodrigues@ine.pt)
Paula Paulino, Pedro Campos, Teresa Fragoso

Work Session on Statistical Data Confidentiality
The Hague, Netherlands
29 - 31 October 2019
Which SDC methods to apply to Census 2021 data at Statistics Portugal?

- Data products from Census 2021 - types of disclosure risk
- Candidate methods according to data product
- Parameters for each candidate method
- Risk and utility measures to compare methods/parameters
Background

Framework

- Data products from Census 2021 - types of disclosure risk
- Candidate methods
Background

Mainly tabular data

Framework

- Data products from Census 2021 - types of disclosure risk
- Candidate methods
Background

- Data products from Census 2021 - types of disclosure risk
- Candidate methods
Framework

- Data products from Census 2021 - types of disclosure risk
- Candidate methods

EU-project “Harmonized Protection of Census Data in the ESS”
  - Targeted Record Swapping
  - Cell Key Method
Framework

- Data products from Census 2021 - types of disclosure risk

- **Candidate methods**
  - EU-project “Harmonized Protection of Census Data in the ESS”
    - Targeted Record Swapping
    - Cell Key Method

- Parameters for each candidate method

- Risk and utility measures to compare methods/parameters
Background

Framework

- Data products from Census 2021 - types of disclosure risk

- Candidate methods
  - EU-project “Harmonized Protection of Census Data in the ESS”
    - Targeted Record Swapping
    - Cell Key Method

- Parameters for each candidate method

- Risk and utility measures to compare methods/parameters
Methods

Data

- 10,000 dwellings sampled from PT Census 2011 data

Tables

- Groups 2 and 11 of the EU-hypercubes
  Commission Regulation (EU) 2017/712, of 20 April 2017
  Commission Implementing Regulation (EU) 2017/543, of 22 March 2017

- identifying variables (region, sex, age group or size of the locality)
- variables that might be considered sensitive (marital status, household and family status or country of citizenship)
Targeted Record Swapping

(Shlomo et al., 2010)
Targeted Record Swapping

(Shlomo et al., 2010)

1. Identify high risk individuals/households
2. Sample high risk households for data swapping
Targeted Record Swapping

(Shlomo et al., 2010)

1. Identify high risk individuals/households
2. Sample high risk households for data swapping
3. Pair the selected households with other households having the same values for the matching variables
Targeted Record Swapping
(Shlomo et al., 2010)

1. Identify high risk individuals/households
2. Sample high risk households for data swapping
3. Pair the selected households with other households having the same values for the matching variables
4. Swap the geographical information
### RS – variants

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Variant 1</th>
<th>Variant 2</th>
<th>Variant 3</th>
<th>Variant 4</th>
<th>Variant 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Swap rate</td>
<td>5%</td>
<td>5%</td>
<td>5%</td>
<td>10%</td>
<td>5%</td>
</tr>
<tr>
<td>Variables to define high risk (risk)</td>
<td>age.m</td>
<td></td>
<td>sex</td>
<td></td>
<td>geo.m</td>
</tr>
<tr>
<td>Threshold for defining high risk (th)</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Profiles of matching variables (similar)</td>
<td>ageg1</td>
<td></td>
<td>ageg2</td>
<td></td>
<td>ageg3</td>
</tr>
</tbody>
</table>

Where:

- ageg1 = number of people under 20 years old
- ageg2 = number of men aged 20 to 59
- ageg3 = number of men aged 60 and over
- ageg4 = number of women aged 20 to 59
- ageg5 = number of women aged 60 and over
- ethc = number of people not born in the country
- person = number of individuals in the household
Cell Key Method
(Marley & Leaver, 2011; Enderle et al., 2018)

Microdata

<table>
<thead>
<tr>
<th>ID</th>
<th>Sex</th>
<th>Age</th>
<th>Record key</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>45</td>
<td>0.13</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>32</td>
<td>0.78</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Cell Key Method
(Marley & Leaver, 2011; Enderle et al., 2018)

Microdata

<table>
<thead>
<tr>
<th>ID</th>
<th>Sex</th>
<th>Age</th>
<th>...</th>
<th>Record key</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>45</td>
<td>...</td>
<td>0.13</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>32</td>
<td>...</td>
<td>0.78</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Frequency table

<table>
<thead>
<tr>
<th>Age</th>
<th>Count</th>
<th>Sex</th>
<th>Count</th>
<th>Count</th>
<th>Count</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>15-24</td>
<td></td>
<td>1</td>
<td>354</td>
<td>0.89</td>
<td>786</td>
<td>0.24</td>
</tr>
<tr>
<td>25-29</td>
<td></td>
<td>2</td>
<td>632</td>
<td>0.68</td>
<td>485</td>
<td>0.76</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Total</td>
<td>986</td>
<td></td>
<td>1271</td>
<td>0.53</td>
<td></td>
<td>...</td>
</tr>
</tbody>
</table>
## Cell Key Method

(Marley & Leaver, 2011; Enderle et al., 2018)

### Microdata

<table>
<thead>
<tr>
<th>ID</th>
<th>Sex</th>
<th>Age</th>
<th>...</th>
<th>Record key</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>45</td>
<td>...</td>
<td>0.13</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>32</td>
<td>...</td>
<td>0.78</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

### Frequency table

<table>
<thead>
<tr>
<th>Age</th>
<th>Count</th>
<th>Cell key</th>
<th>Count</th>
<th>Cell key</th>
<th>Count</th>
<th>Cell key</th>
</tr>
</thead>
<tbody>
<tr>
<td>15-24</td>
<td></td>
<td></td>
<td>25-29</td>
<td></td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>Count</td>
<td>Cell key</td>
<td>Count</td>
<td>Cell key</td>
<td>Count</td>
<td>Cell key</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>354</td>
<td>0.89</td>
<td>786</td>
<td>0.24</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>2</td>
<td>632</td>
<td>0.68</td>
<td>485</td>
<td>0.76</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Total</td>
<td>986</td>
<td>0.31</td>
<td>1271</td>
<td>0.53</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

---

Microdata (Marley & Leaver, 2011; Enderle et al., 2018)
**Cell Key Method**

(Marley & Leaver, 2011; Enderle et al., 2018)

### Microdata

<table>
<thead>
<tr>
<th>ID</th>
<th>Sex</th>
<th>Age</th>
<th>...</th>
<th>Record key</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>45</td>
<td>...</td>
<td>0.13</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>32</td>
<td>...</td>
<td>0.78</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

### Frequency Table

<table>
<thead>
<tr>
<th>Age</th>
<th>Count</th>
<th>Cell key</th>
<th>Count</th>
<th>Cell key</th>
<th>Count</th>
<th>Cell key</th>
</tr>
</thead>
<tbody>
<tr>
<td>15-24</td>
<td></td>
<td></td>
<td>25-29</td>
<td></td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>354</td>
<td>0.89</td>
<td>786</td>
<td>0.24</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>2</td>
<td>632</td>
<td>0.68</td>
<td>485</td>
<td>0.76</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Total</td>
<td>986</td>
<td>0.31</td>
<td>1271</td>
<td>0.53</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

### Perturbation Table

<table>
<thead>
<tr>
<th>Target frequency</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original frequency</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0.59</td>
<td>0</td>
<td>0.41</td>
<td>...</td>
</tr>
<tr>
<td>2</td>
<td>0.18</td>
<td>0</td>
<td>0.29</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Cell Key Method
(Marley & Leaver, 2011; Enderle et al., 2018)

### Microdata

<table>
<thead>
<tr>
<th>ID</th>
<th>Sex</th>
<th>Age</th>
<th>...</th>
<th>Record key</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>45</td>
<td>...</td>
<td>0.13</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>32</td>
<td>...</td>
<td>0.78</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

### Frequency table

<table>
<thead>
<tr>
<th>Age</th>
<th>15-24</th>
<th>25-29</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cell key</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sex</th>
<th>Count</th>
<th>Cell key</th>
<th>Count</th>
<th>Cell key</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>354</td>
<td>0.89</td>
<td>786</td>
<td>0.24</td>
</tr>
<tr>
<td>2</td>
<td>632</td>
<td>0.68</td>
<td>485</td>
<td>0.76</td>
</tr>
<tr>
<td>Total</td>
<td>986</td>
<td>0.31</td>
<td>1271</td>
<td>0.53</td>
</tr>
</tbody>
</table>

### Perturbed frequency table

<table>
<thead>
<tr>
<th>Age</th>
<th>15-24</th>
<th>25-29</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>353</td>
<td>783</td>
<td>...</td>
</tr>
<tr>
<td>2</td>
<td>635</td>
<td>487</td>
<td>...</td>
</tr>
<tr>
<td>Total</td>
<td>987</td>
<td>1270</td>
<td></td>
</tr>
</tbody>
</table>

### Perturbation table

<table>
<thead>
<tr>
<th>Target frequency</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original frequency</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0.59</td>
<td>0</td>
<td>0.41</td>
<td>...</td>
</tr>
<tr>
<td>2</td>
<td>0.18</td>
<td>0</td>
<td>0.29</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
### CKM – variants

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Variant 1</th>
<th>Variant 2</th>
<th>Variant 3</th>
<th>Variant 4</th>
<th>Variant 5</th>
<th>Variant 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum perturbation (D)</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Perturbation variance (V)</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Threshold value for small frequencies (js)</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Probability of an original frequency to remain unperturbed (pstay)</td>
<td>NA (produces the max. entropy solution)</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>0.5</td>
<td>NA</td>
</tr>
<tr>
<td>Monotony condition (mono)</td>
<td>TRUE</td>
<td>TRUE</td>
<td>TRUE</td>
<td>TRUE</td>
<td>TRUE</td>
<td>FALSE</td>
</tr>
</tbody>
</table>
### RS+CKM – variant

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Variant 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Swap rate</td>
<td>5%</td>
</tr>
<tr>
<td>Variables to define high risk (risk)</td>
<td>age.m</td>
</tr>
<tr>
<td>Threshold for defining high risk (th)</td>
<td>2</td>
</tr>
<tr>
<td>Profiles of matching variables (similar)</td>
<td>person</td>
</tr>
<tr>
<td>Maximum perturbation (D)</td>
<td>4</td>
</tr>
<tr>
<td>Perturbation variance (V)</td>
<td>3</td>
</tr>
<tr>
<td>Threshold value for small frequencies (js)</td>
<td>2</td>
</tr>
<tr>
<td>Probability of an original frequency to remain unperturbed (pstay)</td>
<td>NA</td>
</tr>
<tr>
<td>Monotony condition (mono)</td>
<td>TRUE</td>
</tr>
</tbody>
</table>
Risk measures

- Let:
  - $n_c$ : number of units that fall into cell $c$ in the original table $T$
  - $n'_c$ : number of units that fall into cell $c$ in the protected table $T'$
  - $K$ : total number of cells in table $T$ (or $T'$)

**RM 1**

- Relative change of the number of cells with frequency lower than 3 (change in low frequencies)

$$CLF = \left( \frac{\sum_{c=1}^{K} I(n'_c < 3)}{\sum_{c=1}^{K} I(n_c < 3)} - 1 \right) \times 100\%$$
Risk measures

RM 2

- Proportion of cells with frequency lower than 3 both in the original and the perturbed table (real low frequencies)

\[
RLF = \frac{\sum_{c=1}^{K} I(n_c < 3 \land n'_c < 3)}{K} \times 100\%
\]

RM 3

- Relative change in the number of cells where there can be attribute disclosure
  - Group disclosure: cell frequency = row or column total
  - Group disclosure by element: cell frequency = row or column total - 1
  - Inferential disclosure: cell frequency = row or column total * \((1\pm p\%\))
Utility measures

UM 1

Absolute distance (AD) between the original and the perturbed counts

\[ AD_c = |n'_c - n_c| \]

UM 2

Relative distance (RD) between the original and the perturbed counts

\[ RD_c = \frac{|n'_c - n_c|}{n_c} \]

Simple descriptive statistics (max, mean, sd, median) across all cells
Utility measures

UM 3 (Buron et al., 2017)
- Proportion of false zeros

\[ FZ = \frac{\sum_{c=1}^{K} I(n'_c = 0 \land n_c \neq 0)}{\sum_{c=1}^{K} I(n'_c = 0)} \times 100\% \]

UM 4 (Buron et al., 2017)
- Proportion of false positives

\[ FP = \frac{\sum_{c=1}^{K} I(n'_c > 0 \land n_c = 0)}{\sum_{c=1}^{K} I(n'_c > 0)} \times 100\% \]

UM 5
- Proportion of unchanged cells

\[ UC = \frac{\sum_{c=1}^{K} I(n'_c = n_c)}{K} \times 100\% \]
Utility measures

**UM 6** (Shlomo, 2007; Marley & Leaver, 2011)

- Relative change in Cramer’s V for each pair of variables \((i, j)\)

\[
RCV_{ij} = \left( \frac{CV'_{ij}}{CV_{ij}} - 1 \right) \times 100\%
\]

where \(CV_{ij} = \sqrt{\frac{\chi^2}{n} \min(I-1, J-1)}\), \(CV'_{ij} = \sqrt{\frac{\chi^2}{n'} \min(I-1, J-1)}\)

\(\chi^2\) is the Pearson's chi-squared statistic

\(I\) is the number of rows and \(J\) is the number of columns of the two-way table
Software

- EU-project **Open source tools for perturbative confidentiality methods**
  - R packages `recordSwapping` (version 0.1.0), `ptable` (version 0.2.0) and `cellKey` (version 0.16.3)
Results

Change in low frequencies (CLF)

- HC2 RS
- HC2 CKM
- HC2 RSCKM

Change in group disclosure (CGD)

- HC2 RS
- HC2 CKM
- HC2 RSCKM
Results

Mean of absolute distance (mean_ad)

Proportion of false zeros (FZ)

HC2
RS
CKM
RSCKM

HC11
RS
CKM
RSCKM

HC
2_1
2_2
2_3

HC
11_1
11_2

Mean of absolute distance (mean_ad)

Proportion of false zeros (FZ)
Discussion and conclusions

- RS can be used with the aim of increasing uncertainty, namely regarding the lowest geographic levels
- RS should be used together with CKM, to effectively reduce risk
- CKM results in the loss of table additivity
Discussion and conclusions

- RS can be used with the aim of increasing uncertainty, namely regarding the lowest geographic levels.
- RS should be used together with CKM, to effectively reduce risk.
- CKM results in the loss of table additivity.

Challenge: Communicating to the users
Discussion and conclusions

- RS can be used with the aim of increasing uncertainty, namely regarding the lowest geographic levels
- RS should be used together with CKM, to effectively reduce risk
- CKM results in the loss of table additivity

**Challenge: Communicating to the users**

- Users need to be aware that perturbative SDC methods were used
Discussion and conclusions

- RS can be used with the aim of increasing uncertainty, namely regarding the lowest geographic levels.
- RS should be used together with CKM, to effectively reduce risk.
- CKM results in the loss of table additivity.

**Challenge: Communicating to the users**

- Users need to be aware that perturbative SDC methods were used.
- Selected disclosure risk and utility indicators might be published, possibly in the quality report.
Discussion and conclusions

- RS can be used with the aim of increasing uncertainty, namely regarding the lowest geographic levels
- RS should be used together with CKM, to effectively reduce risk
- CKM results in the loss of table additivity

**Challenge: Communicating to the users**

- Users need to be aware that perturbative SDC methods were used
- Selected disclosure risk and utility indicators might be published, possibly in the quality report
- The loss of table additivity due to confidentiality protection should be clearly stated
References


Thank you!

Inês Rodrigues (ines.rodrigues@ine.pt)