CONFERENCE OF EUROPEAN STATISTICIANS
Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality Session 6
29-31 October 2019, the Hague, the Netherlands 09 October 2019

A framework for assessing perturbative methods for protection of Census 2021 data at Statistics Portugal

Inês Rodrigues, Paula Paulino, Pedro Campos, Teresa Fragoso (Statistics Portugal)
ines.rodrigues@ine.pt, paula.paulino@ine.pt, pedro.campos@ine.pt, teresa.fragoso@ine.pt

Abstract and Paper

Statistics Portugal is developing a framework to compare different techniques so as to support the decision on the statistical disclosure control methods to apply to Census 2021 data. This will by itself represent an important development with respect to previous Census rounds. We aim at presenting the main issues to be surpassed for the purpose of providing useful and low-risk products, including the tabular data that is foreseen to be published, as well microdata for research purposes. In methodological terms, work will include performing a risk-utility analysis, in order to compare different methods and corresponding parameters. In this regard, quantitative risk measures will be considered concerning the different disclosure risks involved. Methods to compare will be those proposed by the EU project "Harmonized Protection of Census Data in the ESS", namely targeted record swapping and the cell key method, separately and in combination. Data from Census 2011 will be used in this analysis. Selected tabular outputs (specific census hypercubes as well as national tables) will serve as a basis for the first risk-utility analysis, while the protection of microdata for research use will be taken into account in a second stage. A thorough analysis of the challenges regarding communication of such a perturbative approach to Census confidentiality protection will also be performed.

A framework for assessing perturbative methods for protection of Census 2021 data at Statistics Portugal

Inês Rodrigues[*], Paula Paulino[**], Pedro Campos[***] and Teresa Fragoso[****]

[*] Statistics Portugal, ines.rodrigues@ine.pt
[**] Statistics Portugal, paula.paulino@ine.pt
[***] Statistics Portugal, pedro.campos@ine.pt
[****] Statistics Portugal, teresa.fragoso@ine.pt

Abstract: Statistics Portugal is developing a framework to compare different techniques so as to support the decision on the statistical disclosure control methods to apply to Census 2021 data. This will by itself represent an important development with respect to previous Census rounds. The main issues involved in the release of useful and low-risk products – including the tabular data that is foreseen to be published, as well microdata for research purposes – will be considered. In methodological terms, work will include performing a risk-utility analysis, in order to compare different methods and corresponding parameters. In this regard, quantitative risk measures will be considered concerning the different disclosure risks involved. Methods to compare will be those proposed by the EU project "Harmonized Protection of Census Data in the ESS", namely targeted record swapping and the cell key method, separately and in combination. Data from Census 2011 will be used in this analysis. Selected tabular outputs will serve as a basis for the first risk-utility analysis, while the protection of microdata for research use will be taken into account in a second stage. An analysis of the challenges regarding communication of such a perturbative approach to Census confidentiality protection will also be performed.

1 Introduction

Protecting confidentiality of Census data is as essential as it is challenging. Citizens are increasingly aware of data privacy (Special Eurobarometer 487a, 2019); also, data is growing in diversity and volume in different sectors of society (Daas et al., 2013) and this increases the risk that external data sources are used by an ill-intentioned user to disclose confidential information from Census data. By applying statistical disclosure methods (SDC) to the data, the statistical agency is not only effectively reducing such risk, but increasing trust from the respondents as well, having an impact on response and ultimately on the quality of Census outputs (Regulation (EC) No 223/2009 of the European Parliament and of the Council, of 11 March 2009; Hundepool et al., 2010). On the other hand, the implementation of SDC methods necessarily comes with some degree of information loss – either through its reduction or perturbation – which reduces data utility (Duncan et al., 2001; Shlomo & Young, 2006). Since Census is the main source of statistical information on population, households and housing at the national level, a large amount of information is released, while confidentiality protection must be assured. In this paper, a framework

for assessing the SDC methods proposed by Eurostat's Specific Grant Agreement (SGA) "Harmonized Protection of Census Data in the ESS"[1] (Giessing & Schulte Nordholt, 2017) – targeted record swapping and the cell key method – to protect Census data at Statistics Portugal is presented. A first step consisted in identifying the risks related to the release of Census data. Thereafter, a set of scenarios and variants within each scenario were defined, that differ with respect to the applied method(s) and corresponding parameter values. A number of measures to quantify disclosure risk and data utility were specified. A sample of data from Census 2011 was used as test data and two sets of hypercubes established by EU regulation were considered as the output of interest.

2 Data products and types of disclosure risk

One of the first steps when establishing a disclosure control approach is to identify the products that are to be released from the data that is to be protected. The type of dissemination products determines, on the one hand, what are the disclosure risks that have to be taken into account and, on the other hand, the choices regarding the protection techniques to apply (Hundepool et al., 2010; Hundepool et al., 2012). The products released in previous Census rounds, and that are currently expected to be released based on Census 2021 data, include:

- Essentially aggregate data (tabular data), namely: an extensive set of tables to present indicators concerning buildings, dwellings, households and population data up to the LAU 2 level (parish); georeferenced indicators up to the statistical subsection level (named synthetic file); maps showing the distribution of buildings, dwellings and of the population (disaggregated by a set of selected attributes) geocoded to the 1x1km grid; hypercubes available through Eurostat's Census Hub; tables created to answer specific requests made by the users; and
- A sample of microdata subject to a strict SDC procedure, made available as a Public Use File.

The Census is characterized by an exhaustive data collection; therefore, the protection that comes from observing only a sample of the population (which generates uncertainty since the user cannot usually be sure if a given individual or household is included in the data or not) is absent in this case (Hundepool et al., 2012). In addition, outputs are to be released to the public (in general, for free), which implies a lack of control over the intentions that users might have and the purposes for which they can be used. The disclosure risk should therefore be reduced as much as possible, while retaining the usefulness of the data.

[1] Information on this SGA can be found in the CROS portal: https://ec.europa.eu/eurostat/cros/content/harmonised-protection-census-data_en.

Disclosure happens when a user gets to know, from the data, something about a specific unit he/she did not know beforehand. Two main types of disclosure risks need to be accounted for, given the data products that will be released: attribute disclosure and identity disclosure. Attribute disclosure consists in associating the value of some attribute in the data, or a value estimated from the data, to a given unit. Identity disclosure happens when a given record is identified as corresponding to a specific unit in the population (Hundepool et al., 2010; Hundepool et al., 2012).

Attribute disclosure can happen if some characteristic about a unit is disclosed based on a set of other attributes (Hundepool et al., 2010; Hundepool et al., 2012). This can occur whenever a row or column of the table has a unique non-zero value, being such value equal to the corresponding marginal sum. This means that a user would learn the value of the attribute being disaggregated in that row/column, based on the values that define the marginal sum, for those units (and irrespectively of the size of this non-zero frequency). If there are two non-zero counts in a row/column, and one of these equals one, that specific unit (e.g. person) would be able to do attribute disclosure on the remaining units, by subtracting its own count to the table (therefore obtaining a unique non-zero frequency in that row/column as in the previous setting). Table with low-frequency counts are therefore more susceptible to this type of disclosure. Additionally, even when there are more than two non-zero frequencies in a row/column, if the distribution of units among cells is largely concentrated in one cell, an attribute disclosure 'by inference' can provide enough information to the user (e.g. knowing that 90% of those units have a given characteristic can satisfy his/her goals). As is the case of the example above, any subtractions between table counts that might disclose information (e.g. subtracting counts regarding subgroups of a given categorical/geographical variable and disclosing information on the remaining units) constitute the so-called disclosure by differencing. This can occur when combining different classifications for the same attribute (Hundepool et al., 2010; Hundepool et al., 2012).

In tabular data, a user can locate a given unit in a table cell, if he/she knows some of its attributes. This identification (with the establishment of a link between the unit in the population and a cell in the table), however, can only lead to information disclosure if it happens alongside attribute disclosure (as discussed above). On the other hand, the risk that a user attempts to identify some unit(s) based on a specific combination of attributes, e.g. by going to that geographical area and trying to locate the unit(s) with such characteristics, depends more on the attributes/values that define the table (and the interest or motivation that the user has to try an identification), than on the corresponding frequencies (a high frequency can even make this kind of identification easier). Besides, the information that a user would be able to disclose by successfully performing such identification would be obtained by additional efforts and sources of information, and not directly by the data provided by the statistical agency.

When dealing with microdata files, the risk of identity disclosure is the most relevant: by accessing the file, the user gets access to a set of identifying variables that he/she can combine in order to establish a link between a record in the file and a unit in the population. The establishment of such link depends on the external information the user has access to and on the counts associated with each combination of values (lower frequencies in the file and in the population facilitates a potential link) (Hundepool et al., 2010; Hundepool et al., 2012). As a consequence of being successful in identifying a given unit, the user gets to know all the additional attributes contained in the file regarding this unit.

In this paper, we focus on the protection against attribute disclosure, since most of the products to be released based on Census data are tabular data. The protection of microdata for research use is outside the scope of this paper.

3 Candidate methods

Methods used to protect data confidentiality through its modification (masking methods) can be classified into non-perturbative, if they reduce the released information by means of data aggregation and/or suppression, or perturbative, if data is modified by purposely adding an element of error (Hundepool et al., 2012). The risk of disclosing confidential information is already taken into account when designing the tables to be disseminated (by deciding on how many and which categories to present, data aggregation can be applied at this step to protect privacy). A huge loss of utility would nevertheless be required in order to protect data only based on data aggregation, besides the fact that the format of some tables may be fixed, cases in which table redesign is not an alternative (this is the case with the hypercubes that are defined at the European level and whose structure is therefore not modifiable at the national level) (Antal et al., 2017). Data suppression is not considered to be a feasible alternative, given that many of the tables are linked (share common cells) and successfully suppressing confidential cells would be a very complex task and would also result in excessive information loss; also, the suppression pattern applied to a given table would always have to be the same to avoid disclosure (e.g. if a table is produced in different times to answer different user requests). Additional methods, namely perturbative methods, need therefore to be considered (Antal et al., 2017).

Perturbative methods can be classified either as pre-tabular, since they are applied to the microdata before producing the table(s), or as post-tabular, if they act only over the table(s) to be released (Hundepool et al., 2012). Pre-tabular methods can be considered as candidates also for protecting the microdata to be released. Two perturbative methods have been pointed out by the SGA “Harmonized Protection of Census Data in the ESS” as appropriate choices for protecting confidentiality in Census data: (targeted) record swapping (RS) and the so-called cell key method (CKM) (Giessing & Schulte Nordholt, 2017).

RS is a pre-tabular method that consists in exchanging geographical data between pairs of households with equal values for a set of matching variables. It includes four steps: 1. identifying high risk households; 2. sampling high risk households for data swapping; 3. pairing the selected households with other households having the same values for the matching variables; and 4. swapping the geographical information between the paired households. In the first step, risk is calculated based on the combination of a set of identifying variables (also called risk variables) over the established geographic hierarchy (e.g. NUTS I > NUTS II > NUTS III). Records whose counts for the respective combination of risk variables are below a given threshold are at risk. The reciprocal values of these counts are used as sampling probabilities for selecting records to swap (both risk and donor households). Additional records can be swapped in order to attain a given swap rate (Shlomo et al., 2010).

CKM is a post-tabular method in which unbiased random noise is added to each table cell, according to a given perturbation table and in such a way that the same cell is consistently perturbed by the same value. The perturbation table is composed by the probabilities of transitioning from the original to the perturbed frequencies. It is computed based on a specific random noise distribution; this is, in turn, defined by parameters such as the maximum perturbation value and the perturbation variance. A threshold value for which there will be no smaller (or equal) perturbed frequencies and the probability of an original frequency to remain unperturbed can also be set. Additionally, a monotony condition can additionally be imposed, according to which the transition probabilities decrease monotonously when the distance between the two frequencies (original and perturbed) increases. A random record key is assigned to each record of the data file; the keys for records in a given cell, together with the original count frequency, are used to get the perturbed value from the perturbation table (Marley & Leaver, 2011; Enderle et al., 2018).

4 Assessment of methods

4.1 Data

Data from Census 2011 was used in this study. A sample of 10.000 dwellings was selected and treated as the total population, so as to simplify the computational requirements of the study. The hypercubes to be transmitted to Eurostat by the member states are defined by Commission Regulation (EU) 2017/712, of 20 April 2017. In addition, Commission Implementing Regulation (EU) 2017/543, of 22 March 2017, defines the rules regarding the technical specifications of the topics that constitute the hypercubes and their breakdowns. Two groups of EU hypercubes were set as the tables to be protected: groups 2 and 11 (tables 1 and 2). These are defined as cross combinations of several indirect identifying variables (like region, sex, age group and size of the locality, depending on the table) and variables that express

information that might be considered as being sensitive (as marital status, household and family status or country of citizenship). Details on each variable breakdown are presented in table 3.

	Group 2 Total population	**GEO.M.**	**SEX.**	**AGE.M.**	**LMS.L.**	**HST.H.**	**FST.H.**	**HAR.**	**LOC.**
2.1		GEO.M.	SEX.	AGE.L.	LMS.L.		FST.H.		
2.2		GEO.M.	SEX.	AGE.L.		HST.H.		HAR.	
2.3		GEO.M.	SEX.	AGE.M.				HAR.	LOC.

	Group 11 Total population	**GEO.M.**	**SEX.**	**AGE.M.**	**COC.H.**	**YAE.L.**
11.1		GEO.M.	SEX.	AGE.M.	COC.H.	
11.2		GEO.M.	SEX.		COC.H.	YAE.L.

Tables 1 and 2 Hypercubes from groups 2 and 11.

Code	Description	Breakdown	Categories
GEO	Geographical area	GEO.M.	All NUTS 3 regions in the Member State
SEX	Sex	SEX.	1. Male; 2. Female
AGE	Age	AGE.L.	1. < 15; 2. 15-29; 3. 30-49; 4. 50-64; 5. 65-84; 6. ≥ 85
		AGE.M.	5-year age groups; 100+
LMS	Legal marital status	LMS.L.	1. Never married and never in a registered partnership; 2. Married or in registered partnership; 3. Widowed or registered partnership ended with the death of partner (and not remarried or in a registered partnership); 4. Divorced or registered partnership legally dissolved (and not remarried or in a registered partnership); 5. Not stated
HST	Household status	HST.H.	1. Persons living in a private household; 1.1. Persons in a family nucleus; 1.2. Persons not in a family nucleus; 1.2.1. Living alone; 1.2.2. Not living alone; 1.3. Persons living in a private household, but category not stated; 2. Persons not living in a private household; 2.1. Persons in an institutional household; 2.2. Persons not living in a private household (including homeless persons), but category not stated

FST	Family status	FST.H.	1. Partners; 1.1. Persons in a married couple or registered partnership; 1.1.1. Persons in an opposite-sex married couple or registered partnership; 1.1.2. Persons in a same-sex married couple or registered partnership; 1.2. Partners in a consensual union; 2. Partners in a consensual union; 3. Sons/daughters; 3.1. Not of lone parent; 3.2. Of lone parent; 4. Not stated; 5. Not applicable — not in a family nucleus
HAR	Housing arrangements	HAR.	1. Occupants living in a conventional dwelling or in a collective living quarter; 1.1. Occupants living in a conventional dwelling; 1.2. Occupants living in a collective living quarter; 2. Occupants living in another housing unit and the homeless; 3. Not stated
LOC	Size of the locality	LOC.	1. ≥ 1 000 000 persons; 2. 500 000 – 999 999; 3. 200 000 — 499 999; 4. 100 000 — 199 999; 5. 50 000 — 99 999; 6. 20 000 — 49 999; 7. 10 000 — 19 999; 8. 5 000 — 9 999; 9. 2 000 — 4 999; 10. 1 000 — 1 999; 11. 500 — 999; 12. 200 — 499; 13. < 200 persons
COC	Country of citizenship	COC.H.	Country level
YAE	Year of arrival in the country since 1980	YAE.L.	1.1. 2010-2011 1.2. to 1.7. 5-year groups between 1980-2009 2. Resided abroad and arrived 1979 or before, or never resided abroad 3. Not stated

Table 3 Variables in hypercubes from groups 2 and 11.

4.2 Methods

4.2.1 Scenarios and variants

Targeted record swapping (RS) and the cell key method (CKM) were both applied, independently and in combination. Three scenarios were therefore considered: scenario 1 (only RS); scenario 2 (only CKM) and scenario 3 (both methods). For each scenario, a number of variants were defined depending on the values assigned to the corresponding parameters (tables 4 to 6). In each scenario, variant 1 corresponds to a baseline setting, defined by parameters thought to be adequate. The remaining variants derive from variant 1 by changing one parameter at a time (highlighted in grey). In RS, the geographic hierarchy was defined by the NUTS 2, NUTS 3 and LAU 1 regions (municipalities).

Parameter	Variant 1	Variant 2	Variant 3	Variant 4	Variant 5
Swap rate	5%	5%	5%	10%	5%
Variables to define high risk (risk)	age.m || sex || geo.m	age.m || sex || geo.m || person || lms.l || har	age.m || sex || geo.m	age.m || sex || geo.m	age.m || sex || geo.m
Threshold for defining high risk (th)	2	2	0	2	2
Profiles of matching variables (similar)	ageg1 || ageg2 || ageg3 || ageg4 || ageg5 || ethc	ageg1 || ageg2 || ageg3 || ageg4 || ageg5 || ethc	ageg1 || ageg2 || ageg3 || ageg4 || ageg5 || ethc	ageg1 || ageg2 || ageg3 || ageg4 || ageg5 || ethc	person || ethc

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

Table 4 Variants for scenario 1: record swapping.

Parameter	Variant 1	Variant 2	Variant 3	Variant 4	Variant 5	Variant 6
Maximum perturbation (D)	4	5	4	4	4	4
Perturbation variance (V)	3	3	6	3	3	3
Threshold value for small frequencies (js)	2	2	2	0	2	2
Probability of an original frequency to remain unperturbed (pstay)	NA (produces the max. entropy solution)	NA	NA	NA	0.5	NA
Monotony condition (mono)	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE

Table 5 Variants for scenario 2: cell key method

Parameter	Variant 1
Swap rate	5%
Variables to define high risk (risk)	age.m || sex || geo.m
Threshold for defining high risk (th)	2
Profiles of matching variables (similar)	person || ethc
Maximum perturbation (D)	4
Perturbation variance (V)	5
Threshold value for small frequencies (js)	3
Probability of an original frequency to remain unperturbed (pstay)	NA
Monotony condition (mono)	TRUE

Table 6 Variant for scenario 3: record swapping + cell key method.

4.2.2 Risk measures

In order to compare the different scenarios and variants, a set of risk and utility measures were computed. Let T be a frequency table obtained from the original microdata set D. Let n_c represent the number of units that fall into cell c in table T. Let also T' be the perturbed frequency table with the same structure of T, computed from the microdata set D' ($D' = D$ if only post-tabular perturbative methods were employed and $D' \neq D$, otherwise). We represent by n'_c the number of units that fall into cell c in table T'. Let additionally K be the total number of cells in table T (including marginal cells and the grand total), which equals the total number of cells in table T'.

Being I the indicator function that receives a value of 1 if the condition is true and 0 otherwise, risk measures used to evaluate each scenario/variant included the:

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

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

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

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

- Relative change of the number of cases where there can be attribute disclosure; three types of disclosure are distinguished:
    - Group disclosure, when there is a cell whose frequency is equal to the corresponding row or column total;
    - Group disclosure by element, when there is a cell whose frequency is equal to the corresponding row or column total minus 1;
    - Inferential disclosure, when there is a cell whose frequency is within a given percentage p of the corresponding row or column total[2].

4.2.3 Utility measures

In regard to utility of the perturbed tables, the following measures were computed for each cell c ($c = 1, \ldots, K$) (Shlomo & Young, 2005; Shlomo, 2007):

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

$$AD_c = |n'_c - n_c|$$

[2] Percentage p was set to 10%.

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

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

Simple descriptive statistics for these distances across all cells of the table – maximum, mean, standard deviation and median – were computed. Besides these measures, utility measures included (Buron et al., 2017):

- Proportion of false zeros:

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

- Proportion of false positives:

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

- Proportion of unchanged cells:

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

- Relative change in Cramer's V (Shlomo, 2007; Marley & Leaver, 2011) for each pair of variables (i, j) in the table[3]:

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

where $CV_{ij} = \sqrt{\frac{\chi^2/n}{\min(I-1,J-1)}}$, $CV'_{ij} = \sqrt{\frac{\chi^2/n'}{\min(I-1,J-1)}}$, χ^2 is the Pearson's chi-squared statistic, n and n' are, respectively, the number of units in D and D', I is the number of rows and J is the number of columns in the table.

Additivity of the perturbed table is also verified.

4.2.4 Software

Both methods were applied using the new implementations being developed under Eurostat's SGA "Open source tools for perturbative confidentiality methods"[4]. In particular, the `R` packages `recordSwapping` (version 0.1.0), `ptable` (version 0.2.0) and `cellKey` (version 0.16.3) were used[5] in R version 3.5.2 (R Core Team, 2018). Computation time was recorded for each scenario/variant.

[3] Cramer's V is based on Pearson's chi-squared statistic (χ^2) and measures the association between nominal variables, varying from 0 (no association) to 1 (complete association).

[4] Information on this SGA can be found in the CROS portal: https://ec.europa.eu/eurostat/cros/content/perturbative-confidentiality-methods_en.

[5] Packages were downloaded from GitHub: https://github.com/sdcTools/protoTestCensus.

4.3 Results

Results obtained from this study suggest that, in general, CKM results in higher protection (lower disclosure risk), while still outperforming RS in most utility measures. The exception is the proportion of false zeros, which tends to be higher in CKM, namely in variants where js = 2. Figures regarding selected risk and utility measures are shown in figures 1 and 2, while the results concerning the remaining measures are plotted in the Annex.

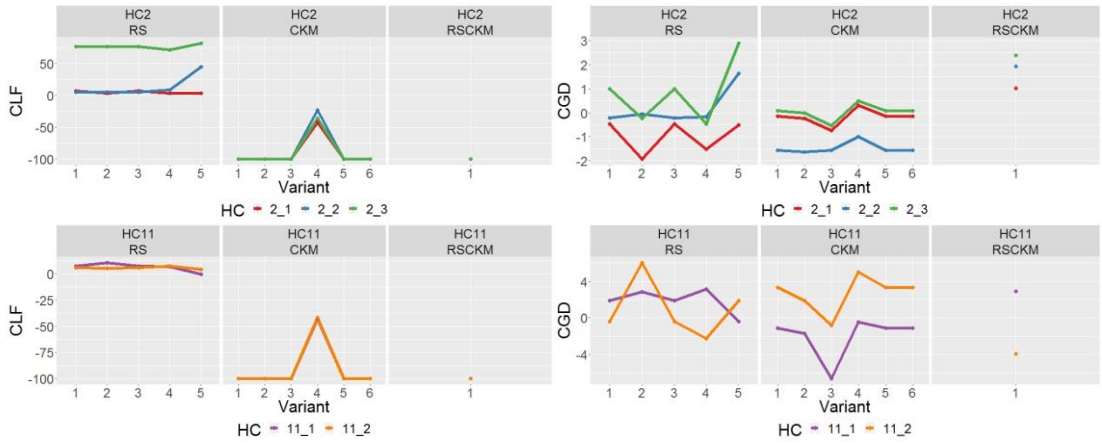

Figure 1 Change in low frequencies (CLF) and Change in group disclosure (CGD), by hypercube, scenario and variant.

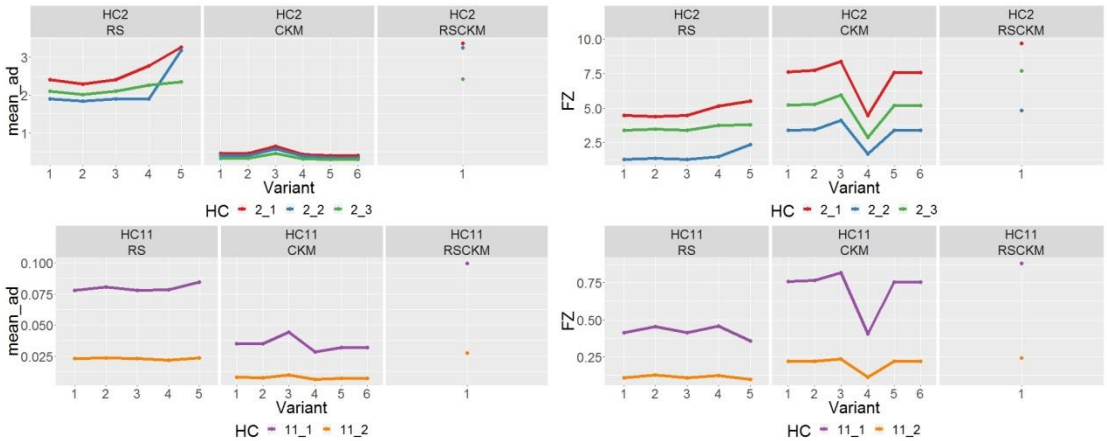

Figure 2 Mean of absolute distance (mean_ad) and Proportion of false zeros (FZ), by hypercube, scenario and variant.

When applying only RS (scenario 1), variants 4 (increase in the swap rate) and 5 (reduction of the number of matching variables) seem to slightly differ from the remaining alternatives by resulting in lower utility. This comes without a substantial decrease in disclosure risk: only the proportion of real low frequencies (RLF) seems to

decrease in these variants, although the number of cells with frequency lower than 3 (CLF) actually increases, particularly in HC 2.

As regards scenario 2 (CKM), variant 4 (no threshold value for small frequencies) results in the more expressive differences in the values for several indicators, both in risk and utility terms (suggesting higher risk and higher utility). Variant 3 (higher perturbation variance) seems to slightly decrease utility when compared to the other variants, particularly when analysing the statistical measures regarding absolute and relative distances between the perturbed and the original cells.

Scenario 3 (RS+CKM) has to be compared with variant 5 from scenario 1 and variant 1 from scenario 2, since these reflect the separate use of the parameters defining the unique variant from this scenario. Particularly in regard to risk measures that concern low frequencies, CKM counteracts the results from RS (since js = 2), therefore decreasing disclosure risk. The effect of CKM in the remaining measures is not that evident.

Table 7 shows the results concerning the change in Cramer's V for the associations with the highest five Cramer's V original values. RS only changes Cramer's V for associations that include geography, given that the method is based on the exchange of geographic information between households. Variant 4 from CKM tends to result in the lowest changes in Cramer's V.

		HC 2.1						**HC 2.2**		**HC 2.3**	
		AGE.L_LMS.L		**AGE.L_FST.H**		**LMS.L_FST.H**		**HST.H_HAR**		**GEO.M_LOC**	
Scenario	**Variant**	CV_{ij}	RCV_{ij}	CV_{ij}	RCV_{ij}	CV_{ij}	RCV_{ij}	CV_{ij}	RCV_{ij}	CV_{ij}	RCV_{ij}
RS	1	0.51	0	0.4	0	0.68	0	0.99	0	0.32	-71.2
RS	2	0.51	0	0.4	0	0.68	0	0.99	0	0.32	-67.1
RS	3	0.51	0	0.4	0	0.68	0	0.99	0	0.32	-71.2
RS	4	0.51	0	0.4	0	0.68	0	0.99	0	0.32	-68.9
RS	5	0.51	0	0.4	0	0.68	0	0.99	0	0.32	-76.6
CKM	1	0.51	0.9	0.4	0.1	0.68	0.7	0.99	0.5	0.32	0.9
CKM	2	0.51	0.8	0.4	0.0	0.68	0.6	0.99	0.5	0.32	1.0
CKM	3	0.51	1.0	0.4	0.1	0.68	0.9	0.99	0.5	0.32	1.1
CKM	4	0.51	0.5	0.4	0.1	0.68	0.6	0.99	0.3	0.32	0.6
CKM	5	0.51	1.0	0.4	0.0	0.68	0.7	0.99	0.5	0.32	0.8
CKM	6	0.51	1.0	0.4	0.0	0.68	0.7	0.99	0.5	0.32	0.8
RSCKM	1	0.51	0.0	0.4	0.2	0.68	0.2	0.99	0.5	0.32	-73.1

Table 7 Original Cramer's V (CV) and Relative change in Cramer's V (RCV) for each pair of variables, by hypercube, scenario and variant.

As expected, table additivity is only retained in scenario 1 (since RS is a pre-tabular method, additivity of the perturbed tables is guaranteed). Applying CKM always results in losing table additivity. Although there are some differences in the results concerning each group of hypercubes (2 and 11), the general conclusions obtained from comparing the two methods (RS and CKM) tend to apply in both cases.

The mean computation time is similar between the three scenarios and the variants of each scenario (table 8)[6]. On the other hand, time required to apply all scenarios greatly depends on the hypercube: the higher the number of cells in the table, the higher is the computation time.

Scenario	Mean computation time (minutes)	Variant	Mean computation time (minutes)	Hypercube	Mean computation time (minutes)
1: RS	30.1	RS - v.1	30.1	2.1	5.0
2: CKM	32.9	RS - v.2	30.0	2.2	4.2
3: RSCKM	29.9	RS - v.3	30.1	2.3	18.2
		RS - v.4	30.2	11.1	70.6
		RS - v.5	30.2	11.2	59.4
		CKM - v. 1	35.3		
		CKM - v. 2	34.1		
		CKM - v. 3	34.2		
		CKM - v. 4	34.9		
		CKM - v. 5	29.1		
		CKM - v. 6	29.6		
		RSCKM - v.1	29.9		

Table 8 Mean computation time (minutes) by scenario and by hypercube.

5 Discussion and conclusions

A quantitative assessment of the perturbative methods proposed by the SGA "Harmonized Protection of Census Data in the ESS" to be applied to Census data – targeted record swapping and the cell key method – was performed. Measures of disclosure risk and utility were used to compare scenarios defined by the method(s) and corresponding parameters to implement. Such measures are adequate to the main type of data products to release – tabular frequency data.

Results showed that RS alone results in high disclosure risk, which comes from the fact that only the geographical information is swapped between matching households; therefore, all frequencies regarding the geographic level(s) above the geographic hierarchy that is used (in our study, NUTS 1 or national level) are not perturbed (and hence not protected). It has however the advantage of having to be applied only once (since it is a pre-tabular method) (Shlomo et al., 2010). Applying CKM by itself results in values suggesting lower disclosure risk, while still outperforming or being equivalent to RS in some utility indicators. RS can therefore be used with the aim of lowering disclosure risk through an increase in uncertainty, namely regarding the lowest geographic levels; it should nevertheless be used together with CKM, so as to effectively reduce disclosure risk. Also, CKM has the advantage of being able to control the presence of low-frequency cells (based on parameter js). CKM might be applied alone, without disadvantage regarding most risk and utility indicators (as

[6] Time presented in table 2 includes that required to applying the method and computing the correspondent risk and utility indicators.

compared to the joint use of RS and CKM). In practice, the record keys in CKM should be generated only once or by setting a specific seed in order to guarantee table consistency, also throughout time. One important effect of CKM is the loss of table additivity (Marley & Leaver, 2011; Giessing & Schulte Nordholt, 2017). This increases one of the great challenges that arise from disseminating tables subject to perturbative SDC methods: communicating the results to the users (Enderle et al., 2018). Users need to be aware that perturbative SDC methods were applied in order to protect data privacy; 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 indicated. Results also showed that computation time is more dependent on the table than on the method (although CKM turns out to be more resource-expensive since it is a post-tabular method and needs to be applied to each table).

This study allowed a better understanding of RS and CKM, and of the `R` packages under development to implement the two methods, which increase flexibility in implementing and adapting them to any national requirements. Besides, it provided some methodological insight, from the methodological point of view, in order to support the decision on the SDC methods to apply to Census 2021 data at the national level.

References

Antal, L., Enderle, T. & Giessing, S. (2017) Statistical disclosure control methods for harmonised protection of census data. SGA Harmonised protection of census data in the ESS, Work Package 3, Deliverable D3.1 Part I

Buron, M.L., Cabrera, A. & Lukan, J. (2017) Results of the tests on census hypercube and grid data and information loss analysis. SGA Harmonised protection of census data in the ESS, Work Package 3, Deliverable D3.2

Commission Implementing Regulation (EU) 2017/543 of 22 March 2017 laying down rules for the application of Regulation (EC) No 763/2008 of the European Parliament and of the Council on population and housing censuses as regards the technical specifications of the topics and of their breakdowns

Commission Regulation (EU) 2017/712 of 20 April 2017 establishing the reference year and the programme of the statistical data and metadata for population and housing censuses provided for by Regulation (EC) No 763/2008 of the European Parliament and of the Council

Daas P.J.H., Puts M.J., Buelens B. & van den Hurk P.A.M. (2013) Big Data and Official Statistics. Proceedings of the NTTS - Conferences on New Techniques and Technologies for Statistics

Duncan, G., Keller-McNulty, S. & Stokes, S. (2001) Disclosure Risk vs. Data Utility: the R-U Confidentiality Map. Technical Report LA-UR-01-6428. Statistical Sciences Group. Los Alamos National Laboratory, Los Alamos

Enderle T., Giessing S. & Tent R. (2018) Designing Confidentiality on the Fly Methodology – Three Aspects. In: Domingo-Ferrer J., Montes F. (eds) Privacy in Statistical Databases. PSD 2018. Lecture Notes in Computer Science, vol. 11126. Springer, Cham

Giessing, S. & Schulte Nordholt, E. (2017) Recommendations for best practices to protect the Census 2021 hypercubes. SGA Harmonised protection of census data in the ESS, Work Package 3, Deliverable D3.3

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Naylor, J., Schulte Nordholt, E., Seri, G. & de Wolf, P.P. (2010) Handbook on Statistical Disclosure Control. Version 1.2. ESSNet SDC

Hundepool A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Schulte Nordholt, E., Spicer, K. & de Wolf, P.P. (2012) Statistical Disclosure Control. Wiley, Chichester, UK.

Marley, J.K. & Leaver, V.L. (2011) A method for confidentialising user-defined tables: statistical properties and a risk-utility analysis. In: Proceedings of 58th World Statistical Congress, pp. 1072–1081

R Core Team (2018). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Regulation (EC) No 223/2009 of the European Parliament and of the Council of 11 March 2009 on European statistics

Shlomo, N. & Young, C. (2005) Information Loss Measures for Frequency Tables. Joint UNECE/Eurostat work session on statistical data confidentiality

Shlomo, N. & Young, C. (2006). Statistical Disclosure Control Methods Through a Risk-Utility Framework. In J. Domingo-Ferrer and L. Franconi, eds., Privacy in Statistical Databases, 68-81. New York: Springer-Verlag, LNCS, vol. 4302.

Shlomo, N. (2007). Assessing the Impact of SDC Methods on Census Frequency Tables. Joint UNECE/Eurostat work session on statistical data confidentiality

Shlomo N., Tudor C. & Groom P. (2010) Data Swapping for Protecting Census Tables. In: Domingo-Ferrer J., Magkos E. (eds) Privacy in Statistical Databases. PSD 2010. Lecture Notes in Computer Science, vol 6344. Springer, Berlin, Heidelberg

Special Eurobarometer 487a (2019) The General Data Protection Regulation, Report, https://ec.europa.eu/commfrontoffice/publicopinion/index.cfm/ResultDoc/download/DocumentKy/86886

Disclaimer

The views expressed in this paper are those of the authors and do not necessarily reflect the official position of Statistics Portugal.

Annex – Results concerning additional risk and utility measures

Risk measures

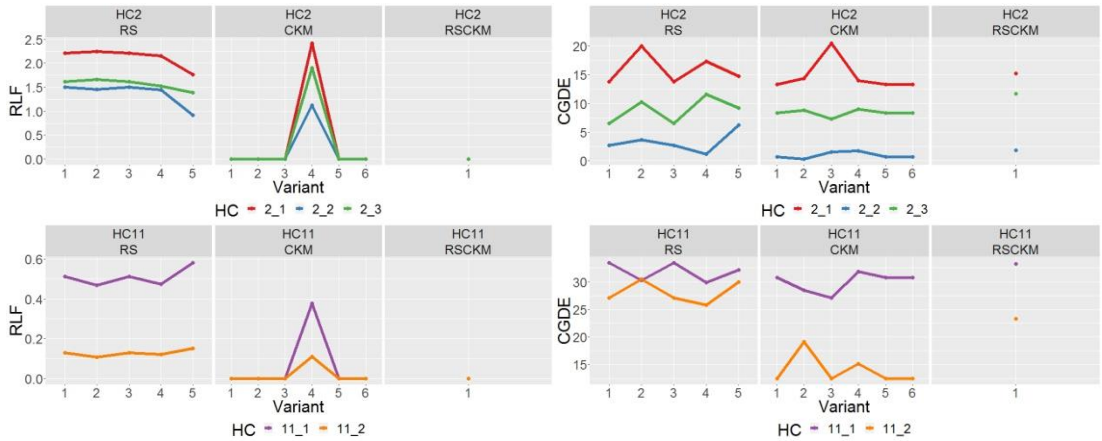

Figure A.1 Real low frequencies (RLF) and Change in group disclosure by element (CGDE), by hypercube, scenario and variant.

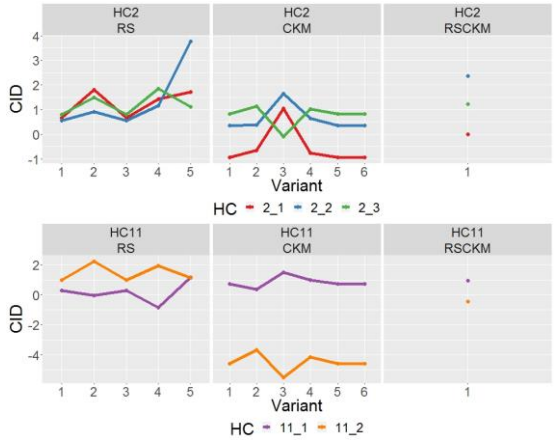

Figure A.2 Change in inferential disclosure (CID), by hypercube, scenario and variant.

Utility measures

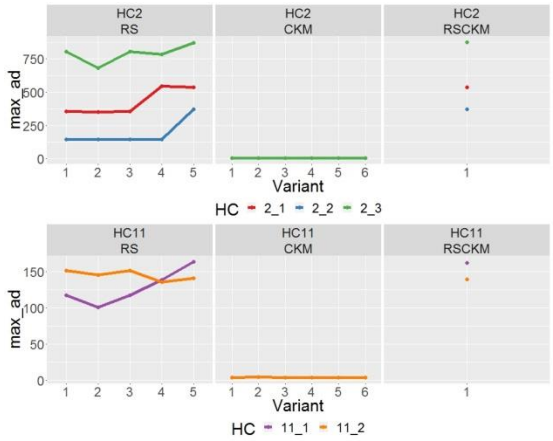

Figure A.3 Maximum of absolute distance (max_ad), by hypercube, scenario and variant.

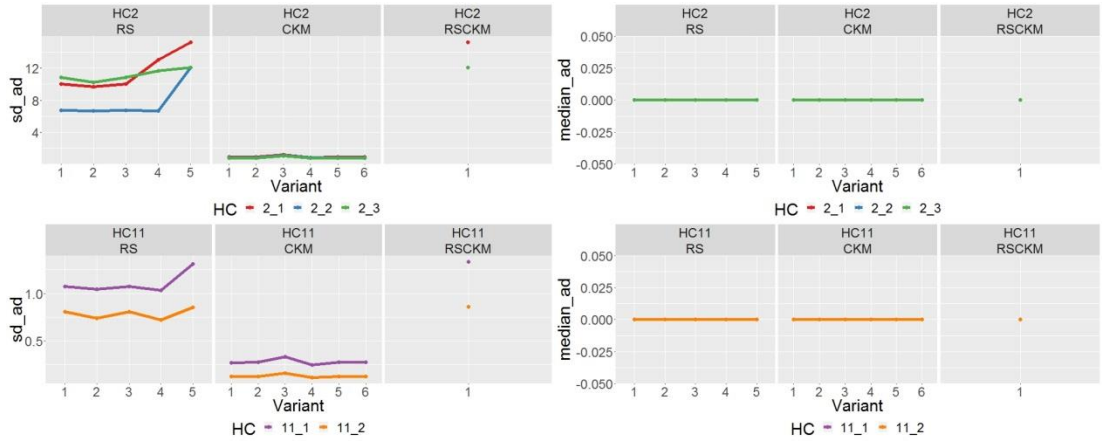

Figure A.4 Standard deviation of absolute distance (sd_ad) and Median of absolute distance (median_ad), by hypercube, scenario and variant.

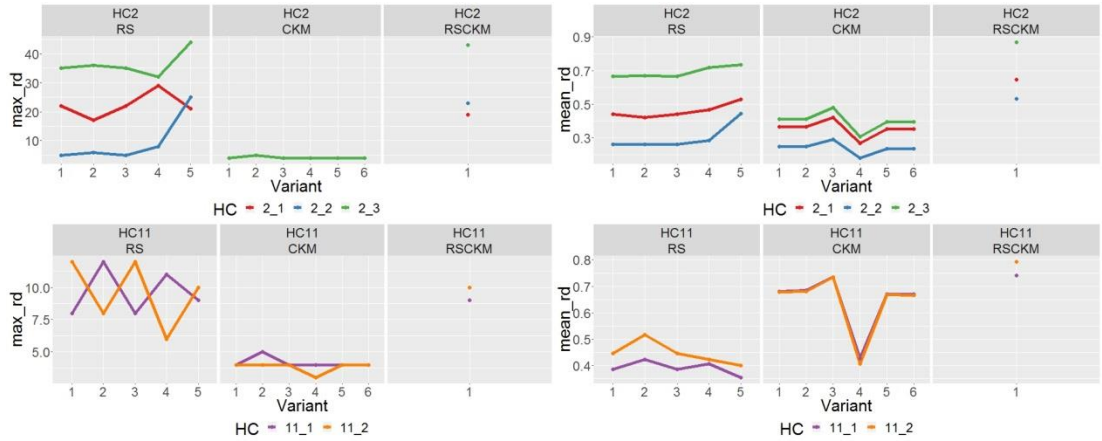

Figure A.5 Maximum of relative distance (max_rd) and Mean of relative distance

(mean_rd), by hypercube, scenario and variant.

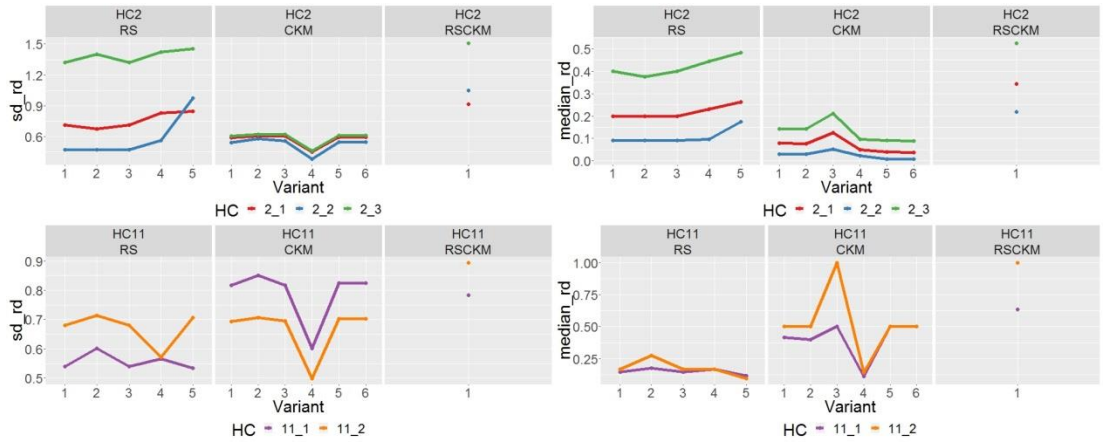

Figure A.6 Standard deviation of relative distance (sd_rd) and Median of relative distance (median_rd), by hypercube, scenario and variant.

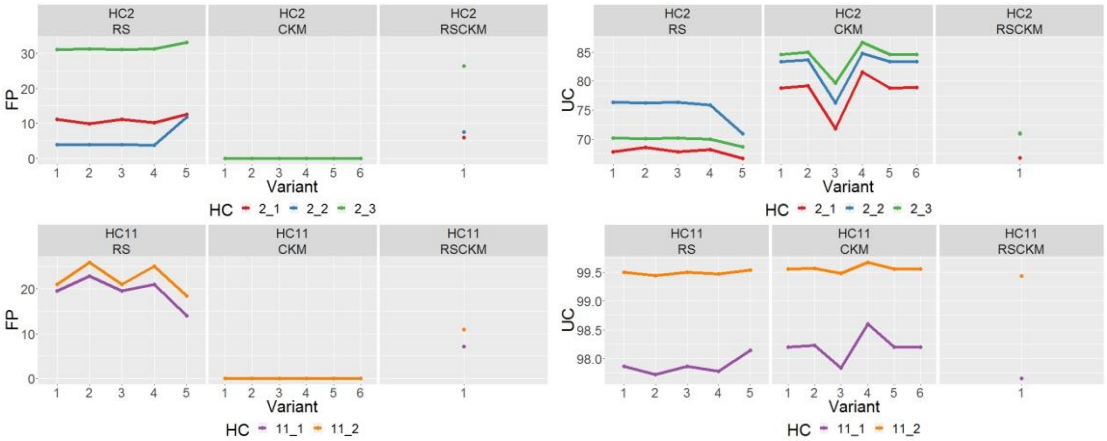

Figure A.7 Proportion of false positives (FP) and Proportion of unchanged cells (UC), by hypercube, scenario and variant.