Specific Grant Agreement (SGA)

Harmonised protection of census data in the ESS

Contract N° 11112.2016.005-2016.367
under
FPA N° 11112.2014.005-2014.533

Date
19/05/2017

Work Package 3
Development and testing of recommendations; identification of best practices

Deliverable D3.1 Part I
Statistical disclosure control methods for harmonised protection of census data

Authors
Laszlo Antal, Tobias Enderle, Sarah Giessing

Sensitivity
Available to NSIs
1. Introduction

Censuses are arguably the most important data collections member states conduct. A Census provides comprehensive data about the population of the country. Member states have to provide census data for Eurostat, where European-level census data are compiled. In most member states, Census data can only be released if measures have been taken to prevent the data from being disclosed. Therefore, statistical disclosure control (SDC) is an important aspect before the data are released.

This document serves as a deliverable of the Harmonised Protection of Census Data in the ESS project, which aims at testing disclosure control methods that might be suitable for protection of Census data in the ESS for the Census 2021 round, suggests statistical disclosure control solutions for census hypercubes and grid data and invites member states to experiment with alternative ways to do the confidentiality treatment, eventually leading to a more harmonised approach and less suppressed cells.

The core of the document is constituted by its second part (part II). In part II we explain in detail what and how to test, while the present part (part I) of the document serves as an introduction to the proposed methodologies and explains why in this project we have selected certain methods for the testing.

Section 2 briefly discusses the need for harmonised statistical disclosure control at European level. The relevant disclosure risk concepts and scenarios are outlined in section 3. Two perturbative SDC methods have been selected for testing within the framework of the project. Section 4 explains why the two methods have been chosen and introduces their basic properties. Finally, section 5 suggests how to evaluate information loss in order to compare different methods or different variants of the same method.

Different partners in this SGA are testing the methods proposed. All European countries not taking part in this SGA are invited to test the methods proposed on their own census data. The project team welcomes member states wishing to get involved in the test. Any feedback from the member states is important for the project and the European Census 2021. If you have any problem, comment, remark during the test, please contact the project team.

2. Census hypercubes and grid data – current situation

Member states produce and provide census hypercubes for Eurostat. The variables of each hypercube and their categories are harmonised across the countries. Therefore, the data of the member states can be combined into European-level data. However, member states can apply SDC methods of their own choice and the differences of the methods might have a negative effect on the quality of European-level data. Eurostat aims at harmonising the SDC methods across the member states in order to increase data quality. For this reason the Harmonised Protection of Census Data in the ESS project was launched in September 2016. The project aims to recommend SDC solutions to the member states and to encourage them to experiment with the proposed SDC methods. The more member states apply the recommended SDC methods the more harmonised European-level data can become.
Besides hypercubes, grid data will also be important outputs of the 2021 Census data. Grid data are calculated on 1km x 1km squares. This new geographical variable also needs to be considered from the viewpoint of statistical disclosure control, especially with regard to already existing and used geographical variables. The importance of grid data lies in their easily understandable interpretation.

Many grid data will presumably contain zero frequencies. A statistical disclosure control solution cannot alter the spatial distribution of grid data too much. It means that if few grid cells contain non-zero frequencies in a certain geographical area, then they should not be changed much and not too many zero grid frequencies should be changed to positive frequencies.

3. Disclosure risk, disclosure risk scenarios

The disclosure risk of statistical data can be quantified by disclosure risk measures. Disclosure risk measures make notions and concepts operational and help to make decisions about the data release. If the disclosure risk is low, a statistical institute might release the data without any change. However, if the disclosure risk is unacceptably high, the statistical institute has to protect the data carefully. Member states might be reluctant to release their data for example in the following cases.

- Many countries do not release their hypercubes/grid data with small counts. The existence of, for example, a count “1” in a hypercube or grid cell means that the respective individual unit is at risk of being identified from the table. Often, it is nevertheless not be possible to infer something about this individual. However, the risk of attribute disclosure (see below) may be higher. Also, self-identifications may raise concerns on the disclosure protection.
- Member states may also protect against attribute disclosure. Attribute disclosure might happen if an attribute of an individual/household (or more individuals/households) can be learnt from the hypercube/grid data. Any non-zero frequency presented in a hypercube (or grid cell) discloses that at least one individual/household exists in the data set with the attributes defining the cell. This might be considered especially critical if a group defined by the cross-combination of a subset of the variables defining the non-zero cell is already very small. An intruder might identify a person/household in this group, and learn from the data that some (at least one) of these people also have properties defined by the remaining part of the variable-combination of the non-zero cells. Imagine a small group of people with a specific age/sex combination in a small municipality where the data exhibit that one/some of them fall into a category of the place-of-birth variable considered sensitive. Obviously, risks of attribute disclosure are more direct risks, compared to the identification risks associated with the publication of small counts.

---

1 For a general discussion of disclosure risks in frequency tables, see for example Hundepool et al. (2012)
In a rather strict understanding of attribute disclosure even the information that a non-zero group of people/households with a certain attribute combination exists is considered risky. In this document we refer to this as risk of “disclosure of existence”.

- **Disclosure by differencing** might happen if we take the difference of two tables (or some parts of two tables) and the resulting table is disclosive. Different geographical variables, such as grid squares and the NUTS classification, potentially increase the risk of disclosure by differencing. Especially grid data, combined with other traditional geographical variables might be susceptible of disclosure by differencing.

4. **Selected protection methods for the project**

If the disclosure risk is high, an SDC protection method needs to be applied to the data. A protection method changes the data in order to reduce the disclosure risk and make the data release possible. SDC methods may be pre-tabular or post-tabular. A pre-tabular method is applied to the microdata. Hypercubes/grid data are generated from the protected microdata. A post-tabular method is applied directly to the hypercubes/grid data. Protection methods can be non-perturbative or perturbative, see a short discussion below.

4.1. **Non-perturbative and perturbative SDC methods from the perspective of the project**

**Non-perturbative methods** do not alter the data. They often suppress table cells/data or collapse some categories of variables into a single one. The disclosure risk of data can be lowered by applying a non-perturbative method. However, they might reduce the information content of the data drastically, leading to unnecessarily high information loss. The data structure might also change by the application of a non-perturbative method. Non-perturbative methods can be carried out consistently and tables might remain additive but it is very difficult to maintain consistency among huge sets of many tables, like e.g. the EU Census hypercubes. Lack of consistency may however lead to disclosure risks. Typical non-perturbative methods are cell suppression and global recoding. Cell suppression also leads to missing data in (census) time series, because especially secondary suppressions are likely to be different at different times.

A **cell suppression** method consists of two steps. For the first step, a rule is required to determine the cells of unacceptably high disclosure risk. These cells are subject to primary suppression. In the second step some cells are selected as secondary suppressions to protect the primary cells from disclosure. Defining primary suppressions based on a minimum frequency rule to protect small counts is simple. Defining a rule for primary suppression that takes care of attribute disclosure risks and at the same time does not cause a lot of overprotection (because it also avoids disclosure of rather trivial / obvious / non-sensitive attributes), is more difficult. This may explain why according to the survey carried out in the project, 18 member states do not protect against the risk of attribute disclosure, even though it is the more obvious and hence more critical (cf. sec. 3). Also, when the goal is to suppress all low-frequency cells (according to the questionnaire this is a requirement in 9 countries), the
number of secondary suppressions can rise almost exponentially with the number of hypercube dimensions. In this situation cell suppression leads to huge information loss.

*Global recoding* means that several categories of a categorical variable are collapsed into a single one, i.e. the “format” of the protected hypercubes may differ from that of the original hypercubes. As for the EU hypercubes, the format is fixed in advance, global recoding is virtually the same as suppressing certain lower levels of the hypercubes. This implies a huge information loss.

**Perturbative methods** deliberately change the data slightly. The information loss caused by a perturbative method can often be kept at a lower level than that caused by a non-perturbative method. It implies that users might find the data more useful. Many perturbative methods do not change the data structure. Examples for perturbative methods are record swapping or adding random noise. A general requirement for a perturbative method is that its effects on the data do not harm data quality, or at least only slightly. In particular, effects should be much smaller than the effect of the changes in a population that happen during a decade. Hence the effect of a suitable perturbative method on a Census time series should be negligible. More details on these two methods can be found in sections 5. Below we summarize some general properties of non-perturbative and perturbative methods.

<table>
<thead>
<tr>
<th>Non-perturbative methods</th>
<th>Perturbative methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Data are not changed but suppressed, recoded, collapsed, etc.</td>
<td>• Data are changed slightly</td>
</tr>
<tr>
<td>• Information loss can become significant</td>
<td>• Information loss controlled by parameters</td>
</tr>
<tr>
<td>• Structure of data might change</td>
<td>• Structure of data can be preserved</td>
</tr>
<tr>
<td>• Can be consistent, but consistency of secondary suppressions is sometimes hard to achieve</td>
<td>• Consistency and/or additivity might be lost</td>
</tr>
<tr>
<td>• Can preserve additivity</td>
<td>• More flexible than non-perturbative methods</td>
</tr>
<tr>
<td>• Missing data points for time series</td>
<td>• Effects on time series are negligible</td>
</tr>
</tbody>
</table>

Specific advantages and disadvantages of the non-perturbative and perturbative methods mentioned above can be found in the table below and in more detail in Annex A.
out consistently when higher dimensional and linked tables are considered

<table>
<thead>
<tr>
<th>Additivity?</th>
<th>Consistency</th>
<th>Additivity</th>
<th>Consistency or additivity can be preserved</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Consistent</td>
<td>Table</td>
<td>Generated from the same protected microdata set are consistent or additivity can be preserved</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Can be additive</td>
<td>Generated from the same protected microdata set are additive</td>
</tr>
</tbody>
</table>

The project team considered the following aspects the most important:

1. A harmonised SDC method should not change the structure of hypercubes.
2. A harmonised SDC method should be able to take care of attribute disclosure risk in a sensible way.
3. The method should keep the information loss minimal, also for detailed hypercube data.
4. Many member states should be able to apply the same method, perhaps with slightly different parameters.

Based on aspects 1-4 above, it can be seen that perturbative methods are superior to non-perturbative methods from the perspective of the project. We support this argument by the reasons below.

Global recoding is not a feasible method, because the design of the EU Census hypercubes is fixed. On the other hand, designing a cell suppression concept to protect the census hypercubes in a consistent way can be quite difficult. Complexity and feasibility depends very much on the disclosure risk concepts and rules used. The survey carried out at the beginning of the project has proven that there are major differences between the countries in this respect. When disclosure risk rules are rather strict, the information loss due to suppressed cells (in particular: secondary suppressions) will be very high. Another problem is the management of differing risk between hypercube and grid level data, adding further to the complexity of cell suppression based protection concepts.

For these reasons, the conclusion was that cell suppression is not the best candidate for a harmonised approach, even though it is an established, good practice method for countries with less strict disclosure risk concepts, supported by well-known software tools like for instance τ-Argus.

A harmonised method should offer some flexibility so that countries can easily adapt it to their specific needs and expectations regarding an acceptable level of residual disclosure risk on the one hand and an acceptable level of information loss on the other hand. The method should be adaptable simply by changing parameters and should consist of separate modules...
that can be used in combination, but may as well be “deactivated” if a country thinks that a specific component is not necessary. The idea is to include modules for pre-tabular perturbation, as well as modules for post-tabular perturbation.

Therefore, the project team decided to select the pre-tabular method of targeted record swapping and the post-tabular random noise method for the test. The parameters of both methods are not fixed, the member states can decide on them. Both methods do not lead to suppressed data, therefore the member states’ data, if treated by these methods, can be combined into European-level data. See Annex A for a comprehensive overview of pros and cons of the methods discussed.

If many member states use the same method – though perhaps in different flavours - this will help to prepare European-level data in a more straightforward way. It will also be an advantage when it comes to tool development, i.e. tools not just suitable for a small testing round in a short project, but for actual data production. It will ease teaching and documentation of the method, as well as preparation of user communication. Unlike with cell suppression, with the perturbative methods proposed, usually data will be available for all hypercube cells. This will be a great advantage for all Census Hub users and will greatly improve the comparability of the data between countries.

In order to maintain consistency between European and national data releases, member states are encouraged to apply the same SDC method to all kinds of data releases. If, however, another method is employed to protect the national release data, member states should check and eventually develop variants that avoid residual disclosure risks that might arise when users compare the EU-hypercube data to national releases.

The following two sub-sections briefly introduce the suggested methods. Detailed descriptions and explanations regarding how to use and test the methods are provided in Part II of the deliverable.

4.2. Record swapping

Record swapping is a pre-tabular SDC method, and as such, it is applied to microdata. Some pairs of records are selected in the microdata set. The paired individuals/households match on some variables in order to maintain the analytical properties and to minimize the bias of the perturbed microdata set as much as possible. Record swapping exchanges some of the non-equal variable-values between paired individuals/households. The exchanged variables are often geographical variables. Since this exchange introduces uncertainty to the microdata, an intruder’s inference about a certain individual/household might not be correct. Record swapping can for example be random or targeted. In case of random record swapping the individuals/households to be swapped are selected with equal probability, while in case of targeted record swapping records of high disclosure risk are determined and a pair to each of these records is selected. It is important to note that record swapping is applied to the microdata. Therefore, at least one of the variables of each hypercube needs to be swapped in order to obtain a perturbed hypercube that is actually different from the original one.
The project follows the targeted record swapping approach of the UK Office for National Statistics (ONS). The most important points of the method are summarized below.

- The method swaps households. (Individuals are not swapped separately.)
- Three main geographical variables of the UK are used: LAD (Local Authority District), MSOA (Middle Layer Super Output Area) and OA (Output Area). The variables are nested with LAD being the highest, while OA the lowest level of geography.
- For each level of geography individuals of high risk are determined.
- For each level of geography households of high risk are determined.
- A sample of households is selected. High risk households are selected in the sample with high probability.
- A household in the sample is paired with another one that matches on some variables. (Note that two sampled households can be paired with each other. However, a sampled household might be paired with another one outside the sample.)
- The geographical variables of paired households are swapped.

4.3. Random noise

Random noise, as a post-tabular method, is defined by noise probability distributions and by a mechanism to draw from the noise distributions. The variant used by the Australian Bureau of Statistics (ABS) builds on so called “cell keys” to ensure that the random noise added to a specific cell will always be exactly the same, even if the cell appears in e.g. different hypercubes. An implementation of random noise as outlined in the following may involve three “modules”:

1. Cell key module
2. Module to determine noise based on cell key and noise distribution parameter matrix
3. Module to restore additivity

The first module enforces consistency of the perturbation, the second determines the statistical properties of the perturbation. The third module achieves hypercube additivity, but might spoil consistency. The following three subsections briefly introduce the basic concepts.

4.3.1. Cell Keys

Cell keys should be drawn from a discrete uniform distribution, defined on some integer values (for example integers between 1 and 100). The process that defines the cell keys has to be consistent, i.e. it must guarantee that the same cell always gets the same key in any hypercube or grid cell or tabulation. For each cell, its cell key and its frequency are used to determine the noise applied to the cell. This step is actually deterministic and can be implemented in such a way that the distributions of the noise match almost exactly the predefined distributions to be specified as parameters of the method. The randomness of the process lies entirely in the part that leads to the cell keys:
1) Assign a random number to each record in the microdata (so called “record keys”). Record keys should be evenly distributed and defined on some given interval, say between 1 and 100.

2) When computing the hypercube data, i.e. counting the number of records having the particular variable combinations of the hypercube cell, do the “same” with the record keys, i.e. take the sum of the record keys for those microdata records having the particular variable combinations of the hypercube cell. Take Modulo 100 (remainder when divided by 100) of the sum of these record keys. The result, referred to as “cell key” obviously lies then also between 1 and 100. Less obvious, but established mathematically, the cell keys will then also be evenly distributed on this interval.

4.3.2. Noise distribution parameters
The performance of a random noise method can easily be controlled in a very flexible way via parameter settings defining the probability distributions. Even a random rounding approach can be considered as random noise with specific noise distributions. In a typical implementation, the following properties will be required and/or controlled by parameters:

- Noise expectation/Unbiasedness property;
- Noise Variance;
- Property that certain frequencies (viz., 1s and 2s) should not appear in the perturbed data;
- Property that (structural) zero cells will never be perturbed.

In the project, there should be the option for the testers to try a number of different variants, leading to noise with different properties.

4.3.3. Additivity Module
As a result of using consistent cell keys the perturbation step leads to consistently perturbed hypercubes. The random noise added to a specific cell will always be exactly the same even if the cell appears in different hypercubes.

However, according to this concept the perturbation is applied to each cell “independently”. As a consequence, the perturbed hypercube data will generally not add up exactly. The ABS has therefore extended the concept by an “additivity module”. After the application of such a module, perturbed hypercubes add up in the usual way. On the other hand, adjustments performed by the additivity module may create “small” inconsistencies between figures appearing in different hypercubes, even though according to their definition they should be identical.

See part II of the document for a discussion of options for implementation of an “additivity module”.

9
5. Information loss measures suggested for the project

SDC methods change the data. The role of information loss measures is to assess the divergence between the data before and after a protection method has been applied. Notably, disclosure risk and information loss are two conflicting concepts. The lower the disclosure risk, the higher the information loss. The ‘best’ SDC method reduces the disclosure risk to an acceptable level, whilst keeping the information loss minimal.

Information loss measures can help to select between variants (e.g. different parameter settings) of the same protection method, or also to choose between different protection methods. Information loss measures can distinguish between protection methods that yield about the same “acceptable” level of protection.

Traditionally, when the goal is to compare different variants of cell suppression, usually the number of suppressions will be compared, but what if the protection methods perturb data?

**Absolute and Relative Differences, Differences between square roots**

For perturbative methods, we typically measure the maxima, means, medians and some percentiles of

A. absolute differences (AD), and

B. relative differences (RAD)

between original and altered counts in a table (or hypercube). Shlomo and Young (2006) suggest to consider also

C. (squared) differences of the square roots between original and altered counts.

Counts may be altered because a perturbative protection method has been applied to the data, or due to the effect of cell suppression. The most straightforward way to take into account suppression (for example when comparing effects of cell suppression in actually released 2011 Census hypercube data to an alternative perturbative method) is to impute zeros for suppressed count.

Interesting is the **cumulative distribution of the absolute differences**, \( F_{AD}(d) \), e.g. proportion of cells with absolute difference less than \( d = 1, 2, 3, 4, \ldots \). For cells on higher aggregate levels of the hypercubes also the **cumulative distribution of the relative absolute differences** \( F_{RAD}(r) \) i.e. the proportion of cells with relative absolute difference less than \( r = 2\%, 5\%, 10\%, \ldots \) is interesting.

**Means** of absolute differences and relative absolute differences may be taken by summing across all cells (of, e.g. a census hypercube) and dividing through the total number of (hypercube) cells. Except for means, we should also look at other descriptive statistics, like **medians, percentiles** (for instance \( p60, p70, p90, p95, p99 \)).
Instead of taking means across all cells of a hypercube, Shlomo and Young (2006) suggest to calculate for certain aggregates \( k \) (such aggregates might be defined by geography) summary statistics from absolute and relative distances observed for the cell counts \( D^k(c) \) contributing to these aggregates in a univariate distribution (viz. a table row), or in a multivariate distribution (like in a hypercube). Summary statistics suggested are means of the absolute distances and sums of the relative distances (cf. table 1, (1) and (2)), and Hellinger’s distance, which is based on the differences between the square roots (table 1, (3)).

<table>
<thead>
<tr>
<th>Distance (per cell ( c ) of aggregate ( k ))</th>
<th>Definition</th>
<th>Summary Statistic suggested per aggregate ( k )</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Absolute Distance</td>
<td>( AD(k,c) :=</td>
<td>D^k_{pert}(c) - D^k_{orig}(c)</td>
</tr>
<tr>
<td>(2) Relative Absolute distance</td>
<td>( RAD(k,c) := \frac{</td>
<td>D^k_{pert}(c) - D^k_{orig}(c)</td>
</tr>
<tr>
<td>(3) Distance of Square-Roots</td>
<td>( D_{-R}(k,c) := \sqrt{D^k_{pert}(c)} - \sqrt{D^k_{orig}(c)} )</td>
<td>Hellinger’s Distance: ( HD(k) := \frac{1}{\sqrt{2}} |D_{-R}(k)|<em>2 = \frac{1}{\sqrt{2}} \sum</em>{c \in k} (D_{-R}(k,c))^2 )</td>
</tr>
</tbody>
</table>

Table 1: Information loss metrics

To further condense the summary statistics of table 1, we can average across the rows (or aggregates) \( k \), e.g. taking means for each metric:

\[
\left( \frac{1}{n_r} \sum_{k=1}^{n_r} AD(k) \right), \left( \frac{1}{n_r} \sum_{k=1}^{n_r} \Sigma_{RAD}(k) \right), \left( \frac{1}{n_r} \sum_{k=1}^{n_r} HD(k) \right),
\]

where \( n_r \) denotes the number of rows (or aggregates).

Especially when comparing additive to non-additive protection methods, it is interesting to calculate also the distance metrics for sub-totals. Otherwise, cumulative effects of the perturbation on sub-totals computed by users who sum up perturbed counts of a census hyper-

\[ ^2 \text{Notably, when a comparison to cell suppression is intended, a technical advantage of a univariate, “row-wise” approach is that it also offers a slightly advanced, natural way to compare to cell suppression, not by simply imputing zeroes (as suggested above), but by imputing for each suppressed cell the mean original count of all suppressed cells in a row.} \]
cube are not taken into account in the comparison. Cf. Shlomo and Young (2006) for respective formulas.

**Acceptance Limits**

It may also be interesting to count the number of cells where the perturbation exceeds a certain limit assumed to be acceptable. On the one hand, for small counts even small changes can be extremely large in the ‘relative’ perspective. On the other hand, for large cells, even large absolute changes do not matter much. So one might consider large relative changes acceptable in small counts, but not large absolute changes. *Vice versa*, in large counts one might accept larger absolute changes, but no large relative changes. This notion can be collapsed into a single criterion, by defining acceptance rules like the following:

“A change is acceptable if (A OR B)", where

- rule A says: “absolute difference is less than \( c \);”
- rule B says: “relative absolute difference is less than \( r\% \).”

An information loss measure based on this consideration would then compare the number (or rate) of cells with acceptable changes according to this concept.

It is interesting to compare these concepts to the concept based on the (squared) differences of the square roots between original and altered counts.

Table 2 illustrates that for small counts very large relative changes lead to the same (square) distance of the square roots as large absolute, but small relative changes in larger cells:

<table>
<thead>
<tr>
<th>Original count</th>
<th>Absolute distance</th>
<th>Relative Absolute distance</th>
<th>(Squared) Distance of square roots</th>
</tr>
</thead>
<tbody>
<tr>
<td>( D^k_{orig}(c) )</td>
<td>(</td>
<td>D^k_{pert}(c) - D^k_{orig}(c)</td>
<td>)</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>167%</td>
<td>1.2</td>
</tr>
<tr>
<td>13</td>
<td>9</td>
<td>69%</td>
<td>1.2</td>
</tr>
<tr>
<td>103</td>
<td>23</td>
<td>22%</td>
<td>1.2</td>
</tr>
<tr>
<td>1003</td>
<td>71</td>
<td>7%</td>
<td>1.2</td>
</tr>
<tr>
<td>10003</td>
<td>222</td>
<td>2%</td>
<td>1.2</td>
</tr>
</tbody>
</table>

**Table 2**: Absolute and relative changes for increasing original counts \( n_{orig}(c) \) with constant \( D\_R(k,c)^2 \).

Compared to the “or”-combination of absolute and relative distance criteria discussed above, a criterion based on distances between the square roots will tend to be stricter on

---

3 In the project, this relates to the suggested random noise method which is non-additive: for all cells explicitly provided in the Census hypercubes, the effects of the method are controlled strictly by its design parameters (up to the NSI to choose). However, when users build own sub-totals, there is much less control. If, for example the design noise variance parameter is given by \( V \) (\( V \) is constant for all hypercube cells), the noise variance of a sub-total \( S \) computed by a user by adding \( n(S) \) perturbed hypercube cells is: \( \text{Var}(S) = n(S) \times V \).
very large and very small cells, and less strict on medium sized cells. It may therefore also make sense to combine all three criteria and define an acceptance rule like:

“\( A \) change is acceptable, if \((A \text{ OR } B) \text{ AND } C\), where

- rule A says: “absolute difference is less than \(a\)”;  
- rule B says: “relative absolute difference is less than \(r\%\)”;  
- rule C says: “srd-distance is less than \(s\”).

Again, an information loss measure based on this combination rule would compare the number (or rate) of hypercube cells with acceptable changes according to the combined concept of the A, B and C rule.

These types of results will of course depend on the rule-parameters (e.g. \(a\), \(r\), \(s\)). The definition of such parameters requires some experience with the protection methods used. The approach has therefore not been taken into account in part II of the document.

**References**


### Advantages/Disadvantages of Protection Methods

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Cell Suppression / Global recoding</th>
<th>Random Noise</th>
<th>Record Swapping</th>
</tr>
</thead>
<tbody>
<tr>
<td>Risks of Identification</td>
<td>No common understanding which small counts (1s, 2s,...) have to be protected</td>
<td>All frequencies possibly changed → protection provided</td>
<td></td>
</tr>
<tr>
<td>Risks of Attribute Disclosure</td>
<td>No common understanding to which extent protection against risks of attribute disclosure (&quot;all 40-50 year old males married&quot;) is needed</td>
<td>All &quot;0&quot; frequencies(^4) may result from perturbation → protection provided</td>
<td></td>
</tr>
<tr>
<td>Secondary Protection/ Differencing risks</td>
<td>Inconsistencies in suppressions across hypercubes may cause disclosure risks</td>
<td>Perturbation protects also against differencing disclosure risks</td>
<td></td>
</tr>
<tr>
<td>Protection obvious?</td>
<td>Method could be implemented to avoid all small counts (1’s, 2’s) in protected hypercubes → Obvious form of protection</td>
<td>Small counts presented in the hypercubes may result from swapping</td>
<td></td>
</tr>
<tr>
<td>Additivity and Consistency</td>
<td>Results additive (unless there are suppressed entries) and consistent</td>
<td>Result presented in the hypercubes to be consistent, but not additive; After additivity module: Results are additive, but not fully consistent</td>
<td>Results additive and consistent</td>
</tr>
<tr>
<td>Information loss</td>
<td>Information loss due to suppressed cells might be high</td>
<td>Information loss low, controlled to a large extent by parameters selected by the NSI</td>
<td>Information loss depends on parameters selected by NSI. Cumulative effects possible.</td>
</tr>
</tbody>
</table>

\(^4\) Attribute disclosure risks arise because of “too many” “0” frequencies in table rows.