Insee
Measuring, understanding

About releasing protected grid data on multiple levels

Clément Guillo [1] Julien Jamme [1]

[1]Insee (French National Statistical Institute)

Context

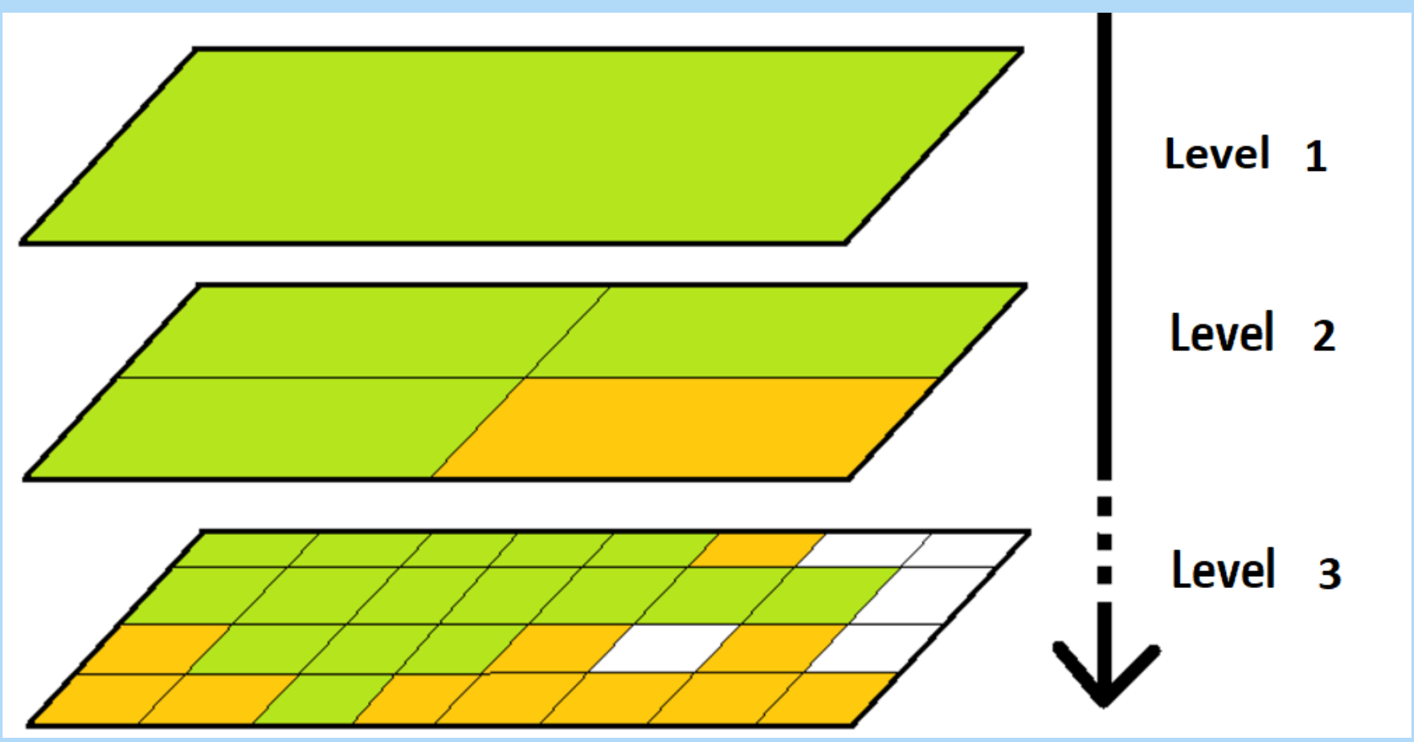
The French national institute of statistics (Insee) is about to release Census data on two grids, measuring respectively 1km and 200m sideways. At the same time, these data are also available on administrative areas, such as municipalities. In this work, we focus only on the Metropolitan France, *i.e.* without outermost regions.

The release has to ensure the following **confidentiality rules**:

- **Threshold rule**: Original data in a square can not be released if less than 11 households live in it.
- **Sensitive attributes**: Users can not infer with more than 80% confidence that all inhabitants in a given square are born abroad or have arrived from a foreign country .
- **Exception**: the count of people in a square is not a confidential piece of information.

Confidentiality issues and objectives

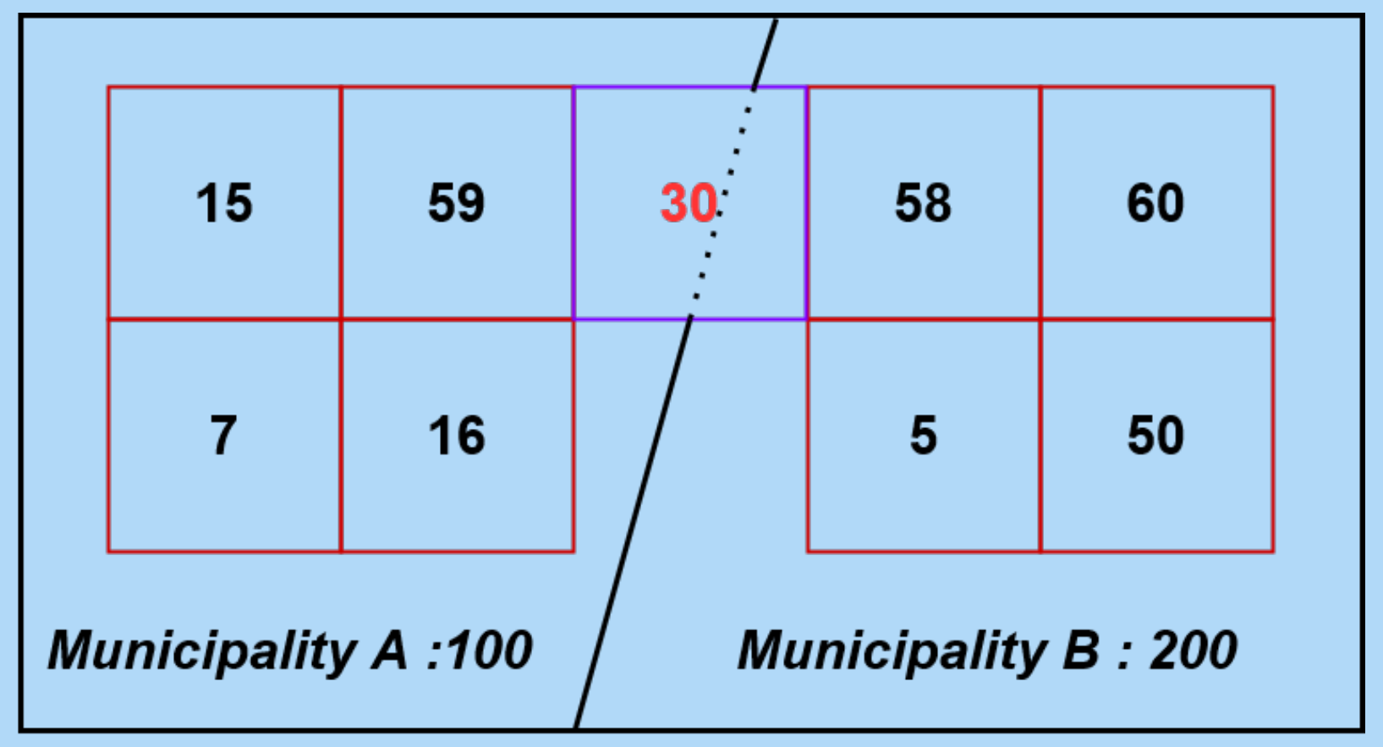
- Disclosure by inter-level differentiation: The release of two (or more) levels of grid data has to be taken into account during the protection process.



- 1km level: 375 159 non empty tiles - 52.6% are below the threshold and 31 tiles are sensitive.
- 200m level: 2 224 377 non empty tiles - 78.1% below the threshold and 232 tiles are sensitive.

Figure 1. Disclosure by inter-level diff.

- Disclosure by overlapping differentiation: This issue comes from the concomitant release of the same information on grids and on administrative areas. These two kinds of zonings are not nested.



There are always two ways to make a differentiation:

- **Disclosure by internal differentiation**: $100 - (15 + 59 + 7 + 16) = 3$
- **Disclosure by external differentiation**: $(58 + 60 + 5 + 50 + 30) - 200 = 3$

Figure 2. Disclosure by non-nesting diff.

Other constraints

- Inference disclosure risk of sensitive attributes
- Original population counts are released as they are, While other counts may be perturbed.

Additional objectives

- Preserve additivity within squares
- Actual empty squares have to remain empty
- Preserve spatial structures

An alternative: the Cell Key Method

The Cell Key Method (CKM) is an efficient method designed to protect tabular data by adding noise into cells. As grids can be viewed as tables, the CKM was proposed as a relevant method to publish Census grid data ([2]).

But, it shows some limitations in handling all our confidentiality issues and objectives:

- In adding a key to each individual to produce consistent tables, CKM can't at the same time release the original count of a cell while perturbing the counts of attributes.
- The CKM doesn't let us monitor how the spatial structures are perturbed.
- Despite the perturbation of squares by CKM, there are still differentiation issues with the municipality level. Indeed, some of these issues may concern big squares that have a high probability of being unperturbed.

1-Detect cells at risk of overlapping differentiation

The detection of intersections between squares and municipalities at risk of differentiation relies on a method implemented in the R package called `diffman`. This method is based on a **graph modeling** in which two municipalities are connected if there is a square whose population is located in both of them.

This representation allows to drastically limit the space of research of the differentiation problems and to extract quasi exhaustively the risky intersections in a reasonable time ([3]).

Figure 3. Geographic representation

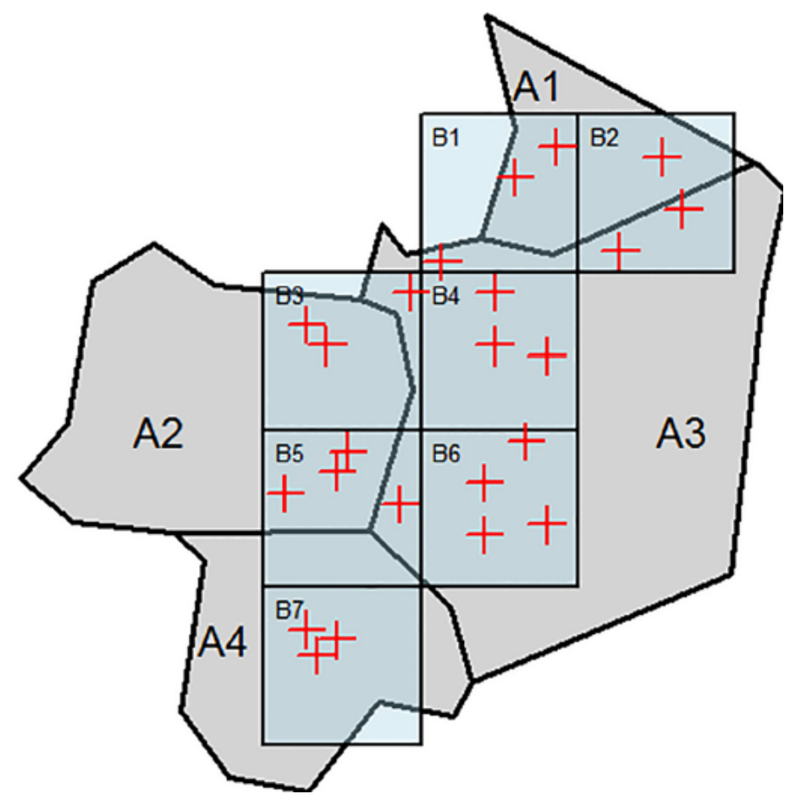
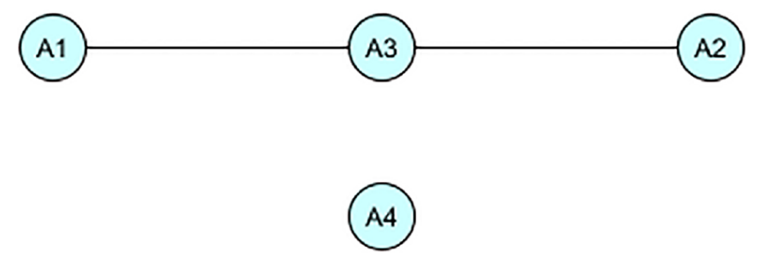


Figure 4. Graph modeling



Remark: The attacker doesn't have such detailed information. ⇒ The detection of risks is easier for us (the producer) than for him/her (the attacker).

2-Protect the cells at risk of overlapping differentiation

Figure 5. Start

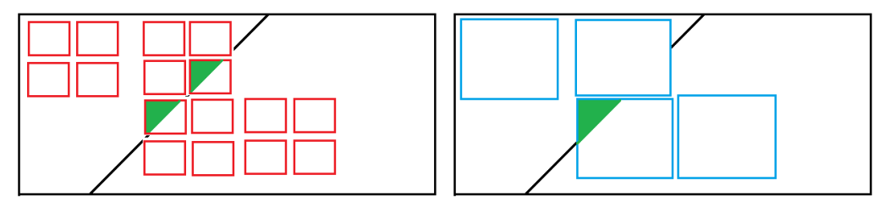


Figure 6. Step 1

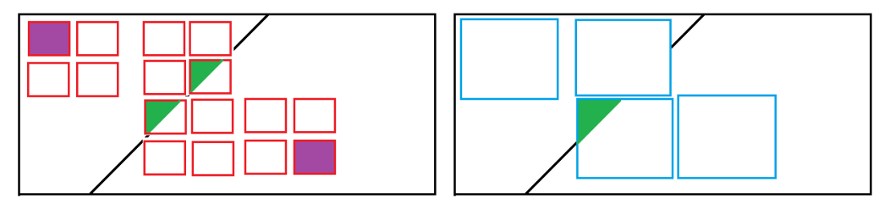
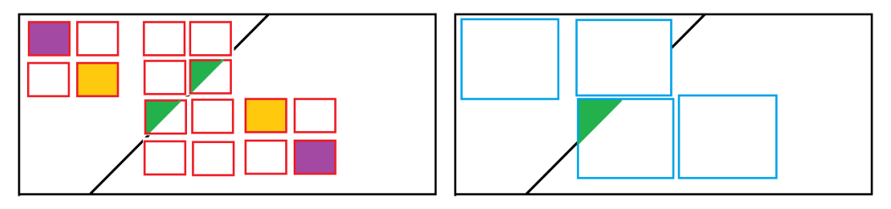


Figure 7. Step 2



Legend:

- 🟩 cells: risky intersections
- 🟪 cells: direct protection of the risky intersections
- 🟨 cells or 🟧 cells: additional suppression to prepare inter-level diff.

Figure 8. Step 3

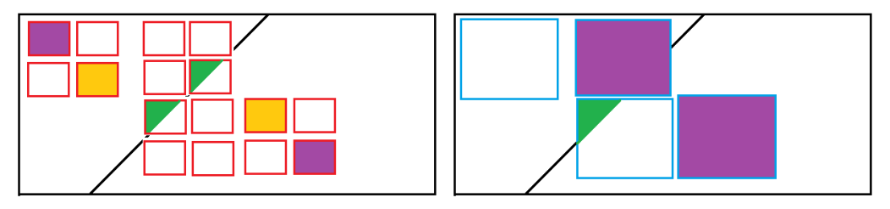
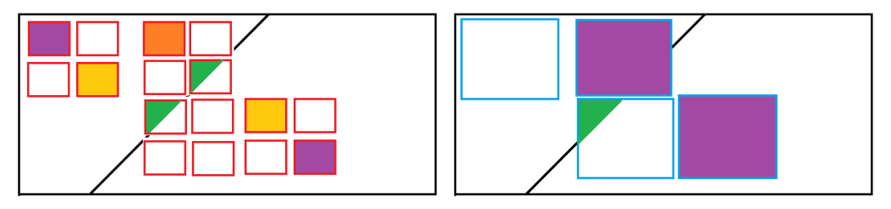


Figure 9. Step 4



3-Protect all the risky cells from inter-level differentiation

The method has been implemented in a `R` package - optimized in `C++` - called `gridy` and applied, the first time, on the fiscal grid data released in 2019 ([1]). It was inspired by the quadtree representation. Some adjustments were made to release all the data on the same size of squares.

The idea is to gather risky tiles with other ones (and also with some non risky tiles) to make groups that contain more than the required threshold. The process begins from the coarsest level (64km) to the finest one (200m). At each step, the inter-level diff. is avoided by **gathering suppressed cells in groups**. These groups are inherited from a level to the next one.

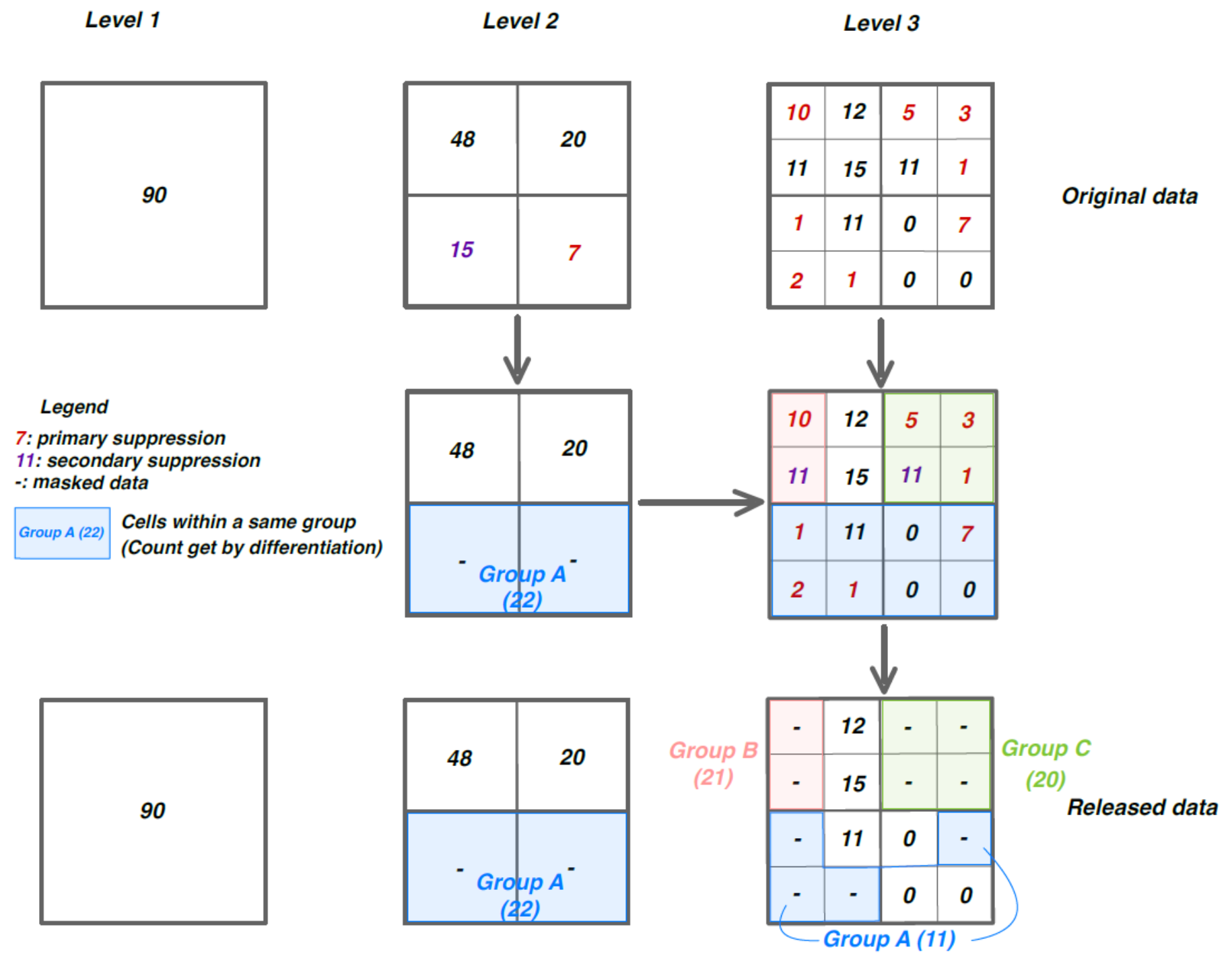


Figure 10. Handling inter-level diff.

Which grid levels to use ?

To proceed, we need the released grids and at least one another which does not contain any confidential tiles (64km squares grid in France). Using other intermediary grids will generate more secondary suppression while better preserving the spatial structure. For example:

- Using 64km - 1km - 200m levels will generate less cell suppression but perturb more the spatial structure;
- Using 64km - 32km - 16km - 8km - 4km - 2km - 1km - 200m levels will generate more cell suppression and perturb less the spatial structure.

4-Protect cells at risk of attributes disclosure

At this point, we haven't treated the cells at risk of ***inference disclosure***. The idea is the following one : once detected, each of these risky cells among the previously non suppressed cells is affected to an existing group of suppressed cells. Two criteria help to choose a good candidate:

- **A distance criterium**: we look for a group not too far from a given risky cell.
- **A risk criterium**: we look for a group such that inference won't be possible.

5-Impute data within the groups of suppressed cells

The suppression process is just temporary: the idea is to release values for each non-empty square. This is the moment **to use the groups of cells** that have been built during the previous process. The idea is to ***breakdown the group total*** (for example the number of males) ***in the tiles of the group, proportionally to the population of each tile***.

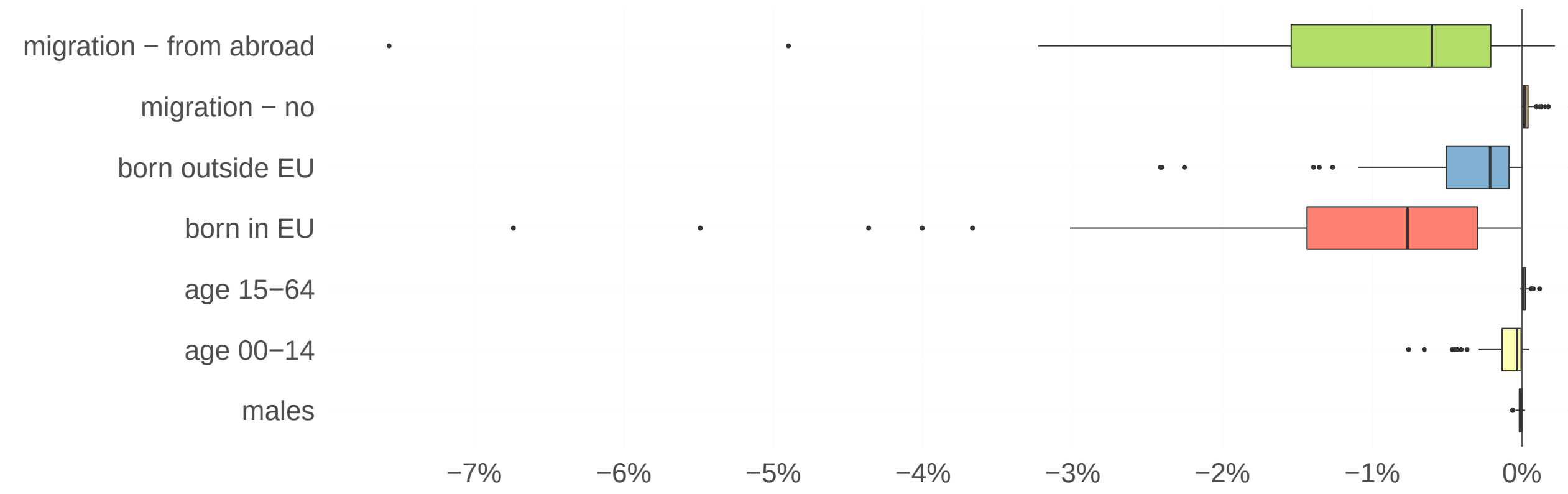
Let c be a tile belonging to a group g, with P^c and P^g their respective actual populations, P_m^g and P_m^c, their respective actual numbers of males. Then, the imputed counts of males is given by $\hat{p}_m^c = P_m^g \times \frac{P^c}{P^g}$.

⇒ the totals within each tile are consistent: the sum of the numbers of males and females is equal to the population count.

A limited loss of information

The protection process is very efficient to minimize the loss of information of the main counts at a fine level (French departments in the figure 11), for breakdown by gender or by age. The loss is more important for small counts as population born abroad.

Figure 11. Differences in % between the released and the original aggregates by French Department



Note: A department, here, is approximated by all the 1km squares that intersect its effective administrative area.

Advantages and Drawbacks

Advantages

- A rather general method able to take several confidential issues into account.
- The detection of overlapping differentiation is actually very complex for an attacker. With our tool, the complexity level of an attack can be set depending on the scenario of attack which was designed.
- Possibility to monitor the amount of secondary suppression and the perturbation of spatial structures in choosing different levels of grids.

Drawbacks

- Some perturbed cells are not actually perturbed, if an attribute is evenly distributed in every square of a group.
  - As close squares tend to be similar, the hypothesis is a serious one.
  - But:
    - The user doesn't know which cells belong to the same group.
    - There's no way for the user to know for which cells the hypothesis is really true.
    - In reality, only the distribution of the gender is really stable among squares. And gender isn't a sensitive attribute.
- The re-identification disclosure risk is not directly handled.
  - The original population counts (rounded) are released and used to dispatch attributes counts into a group.
  - But, in our case, the re-identification is only possible with a prior information about the location of a person...

References

[1] Fontaine M. Branchu M., Costemalle V. Données carroyées et confidentialité. *Journées de Méthodologie statistique*, 2018. URL `http://www.jms-insee.fr/2018/S23_3_ACTE_BRANCHU_JMS2018.pdf`.

[2] S. Giessing L.Antal, T. Enderle. Statistical disclosure control methods for harmonised protection of census data. *Harmonised protection of census data in the ESS*, 2017. URL `https://ec.europa.eu/eurostat/cros/content/harmonised-protection-census-data_en`.

[3] Costemalle V. Detecting geographical differencing problems in the context of spatial data dissemination. *Statistical Journal of the IAOS*, 35(4):559–568, 2019. doi:10.3233/SJI-190564.