Use of multilevel grid to release protected grid data from the French 2021 Census

Clément Guillo
Julien Jamme
Nathanaël Rastout

20/09/2022

Context

- Release of Census 2021 data
- Cube and grid data
- Grid data releases :
  - Eurostat : one grid of 1km2
  - Insee : several grids at **different resolutions** will be released ("natural" level, and at least grids of 1km and 2km).
- Suggested SDC methods (CKM, TRS) to protect the release, by a SDC group of experts.

Context

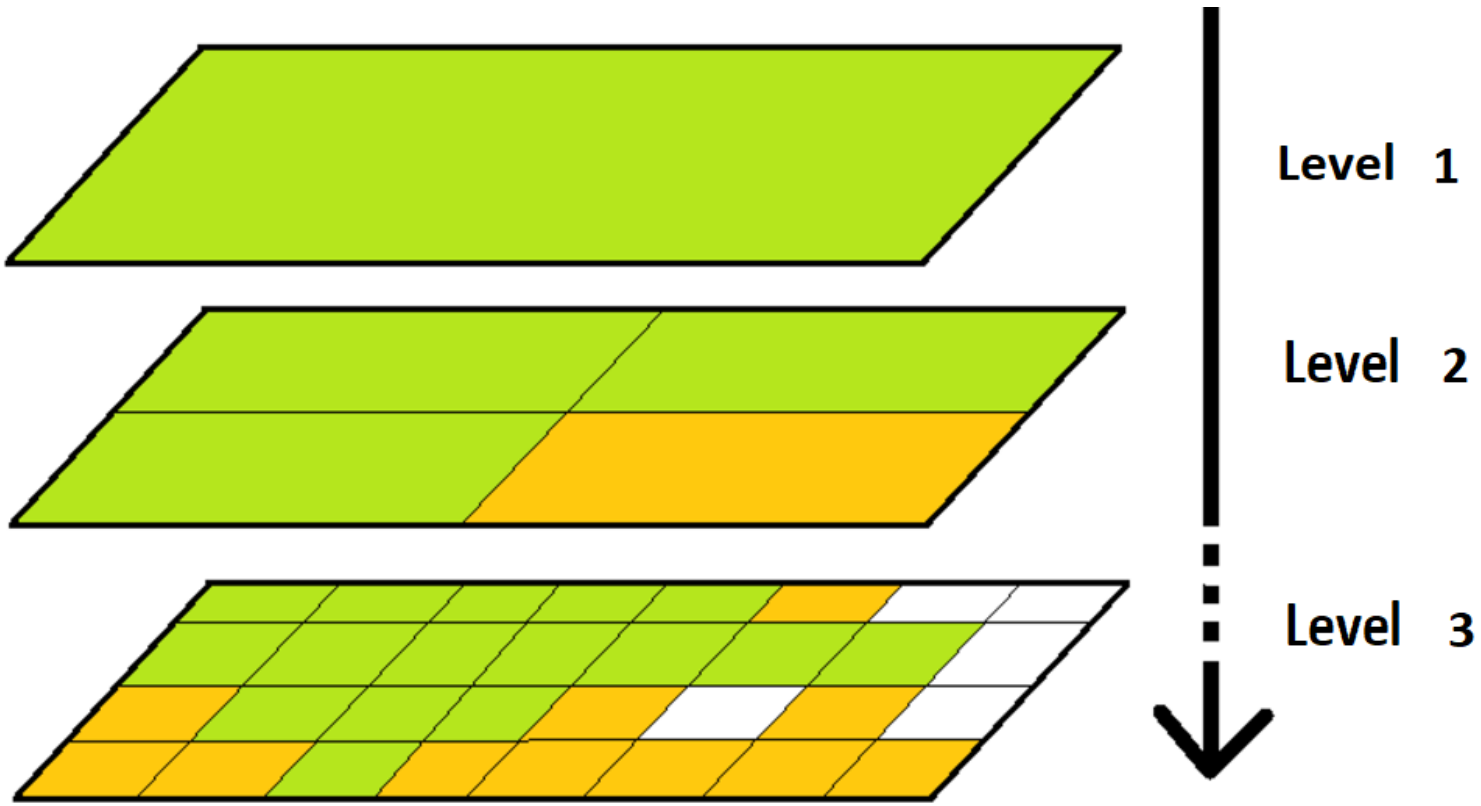
- INSEE has a painful history with the use of swapping in protecting grid data (leak in 2015) ;
- Since, to disseminate the tax data it was therefore decided to use a mixed method ;
- For the sake of consistency, France has decided to protect all gridded data, and especially its census, the same way.
- Confidentiality rules for grid data :
  - No release of tile with less than 11 households
  - The population counts are not confidential.

Table of Contents

Multilevels

- All the reported methods are based on a multilevel grid
- From the coarsest resolution :
  - tiles of 32km of side ;
  - resolution at which the number of households in any tile is not below the threshold (11).
- To the finest one :
  - 200m (for tax data)
  - 1km for Census 2021
- Between : all the intermediate resolutions are used (not necessarily released) : 16km, 8km, 4km and 2km.
- Each tile at level N is divided in 4 tiles at level N+1

MULTILEVEL GRID RELEASE



About releasing the finest unconfidential data

The ***natural* grid** is built to release the finest unconfidential data.
Method :

- based on the *quadtree* algorithm
- From the coarsest level (32km) to the finest one ;
- Let be a tile at level N with a count above the threshold S
- The tile is divided in 4 tiles at level N+1, if none of them have a count below the threshold
- otherwise the tile at level N is the released one and the process is stopped.

The *natural* grid

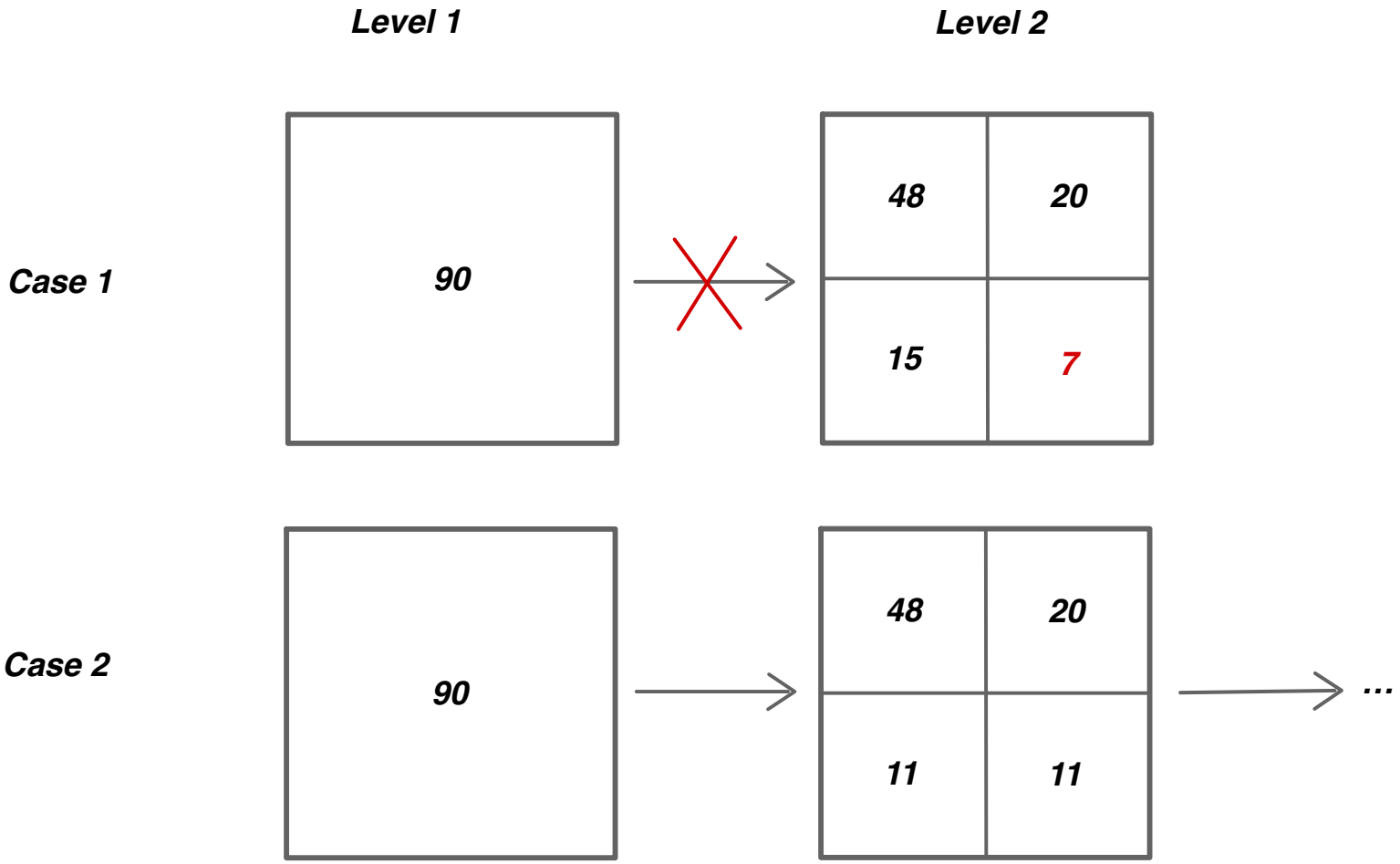


Figure – Two types of cases during the quadtree process

The *natural* grid

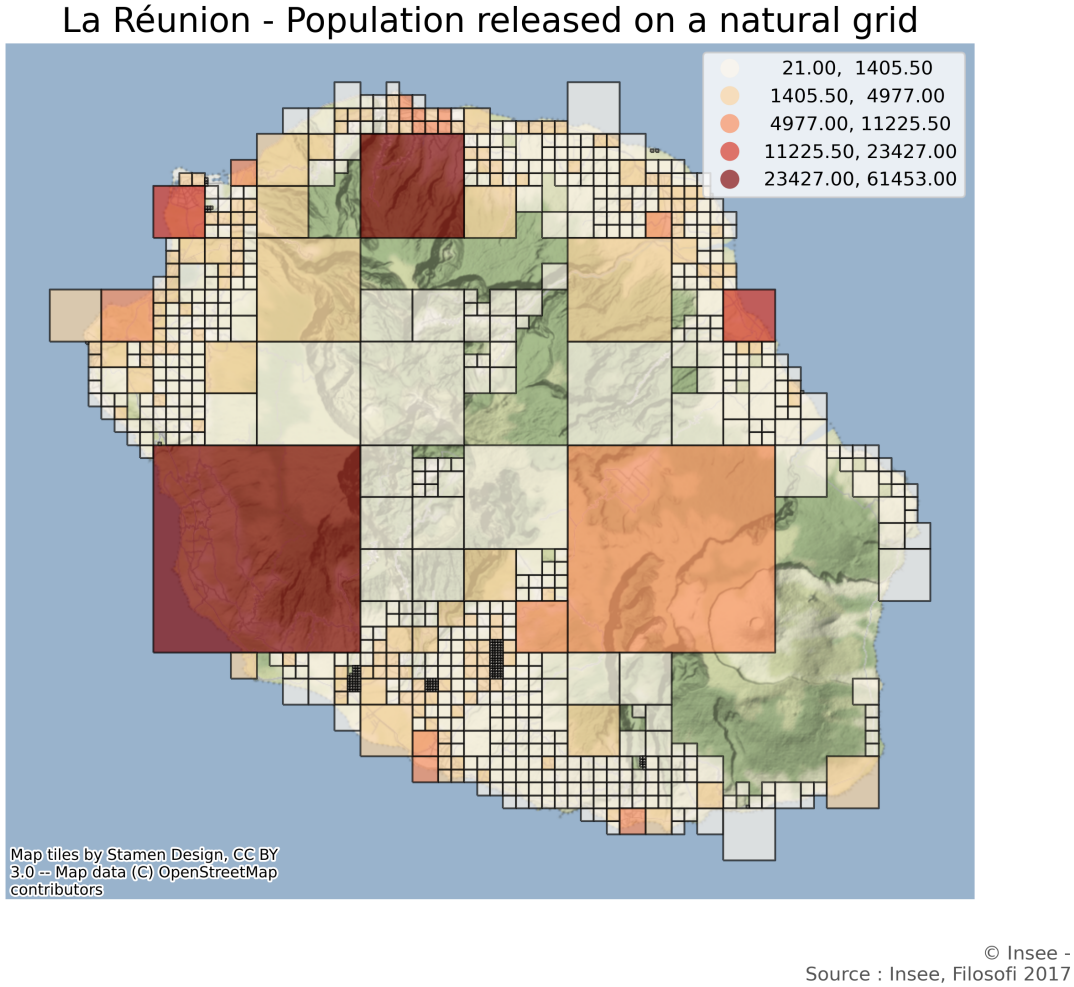


Figure – Quadtree algorithm applied in the island of La Réunion

Advantages :

- None confidential data is released ;
- Easy to implement
- Denser is an area, finer is the resolution (in general)

Drawbacks :

- Tiles with different sizes
- Dense areas can be released on a coarse level only. Case of La Réunion
  - two 16km tiles with at least more than 20 000 people
  - due to the specificity of the topology (volcanoes on the center of the island).
- Not so easy to handle for users
- An homogeneous grid is required by Insee and Eurostat

Multilevel grid release

How to handle confidentiality between levels ?

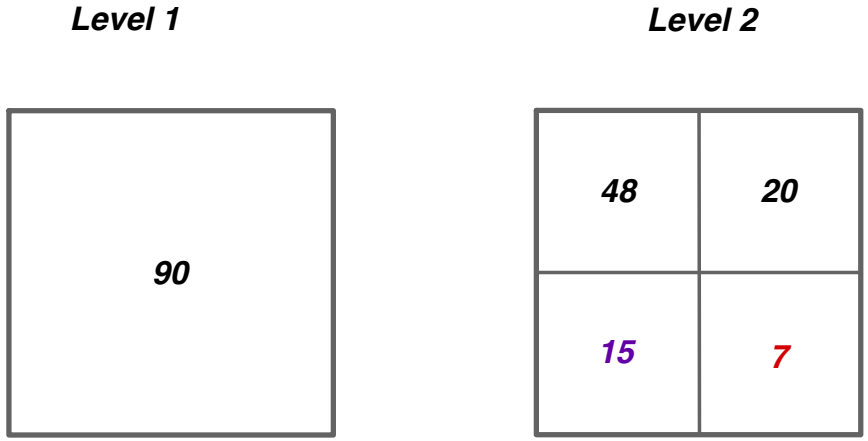
Objectives :

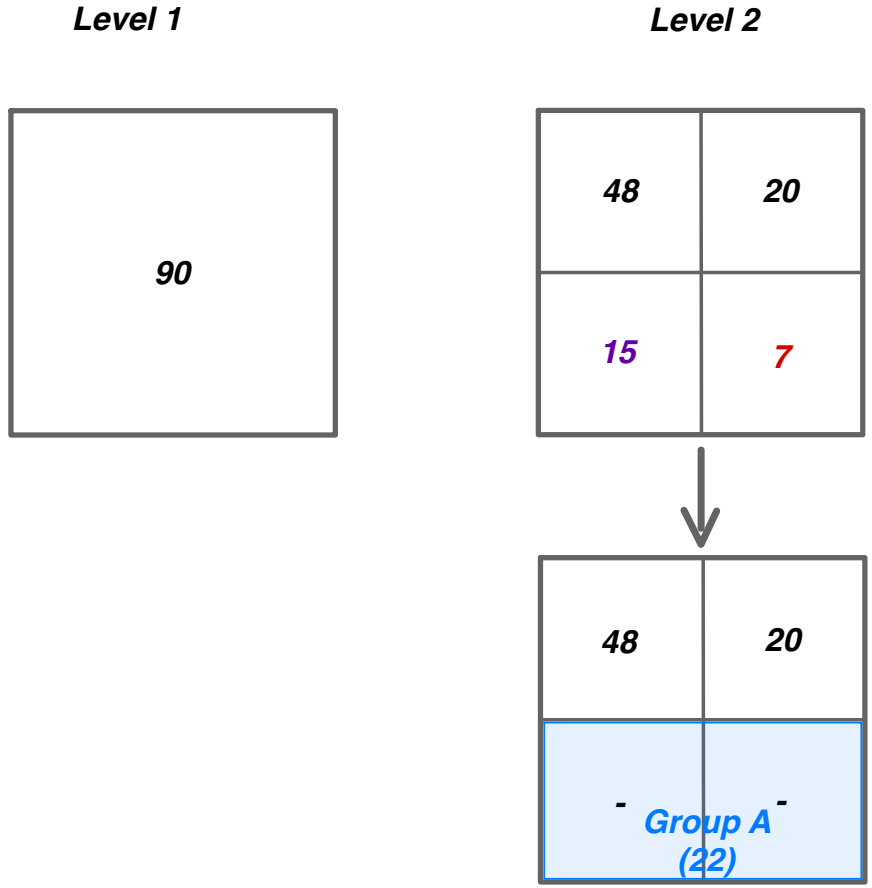
- Release several grids with different resolution ;
- Protect data from disclosure

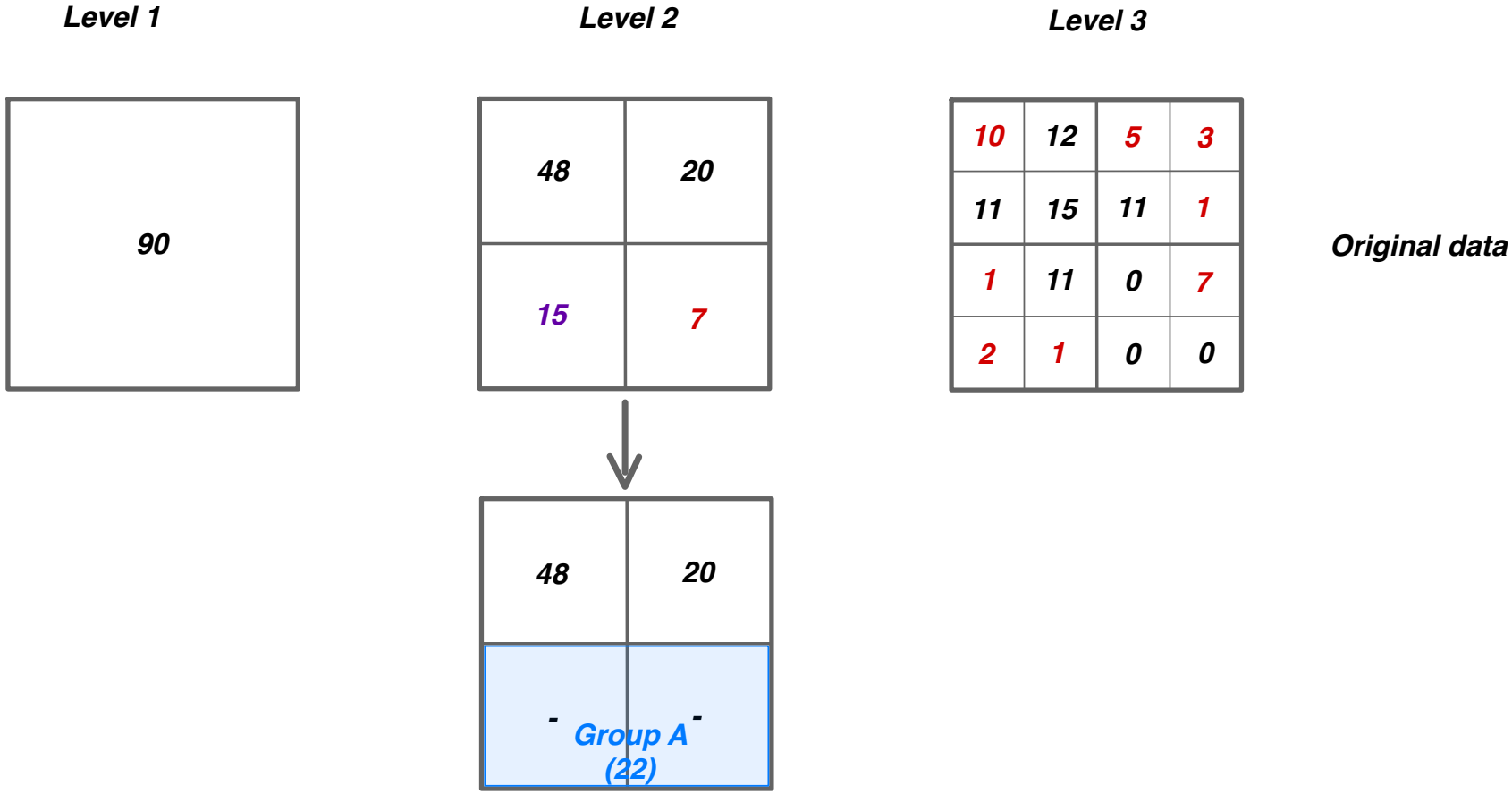
Principle :

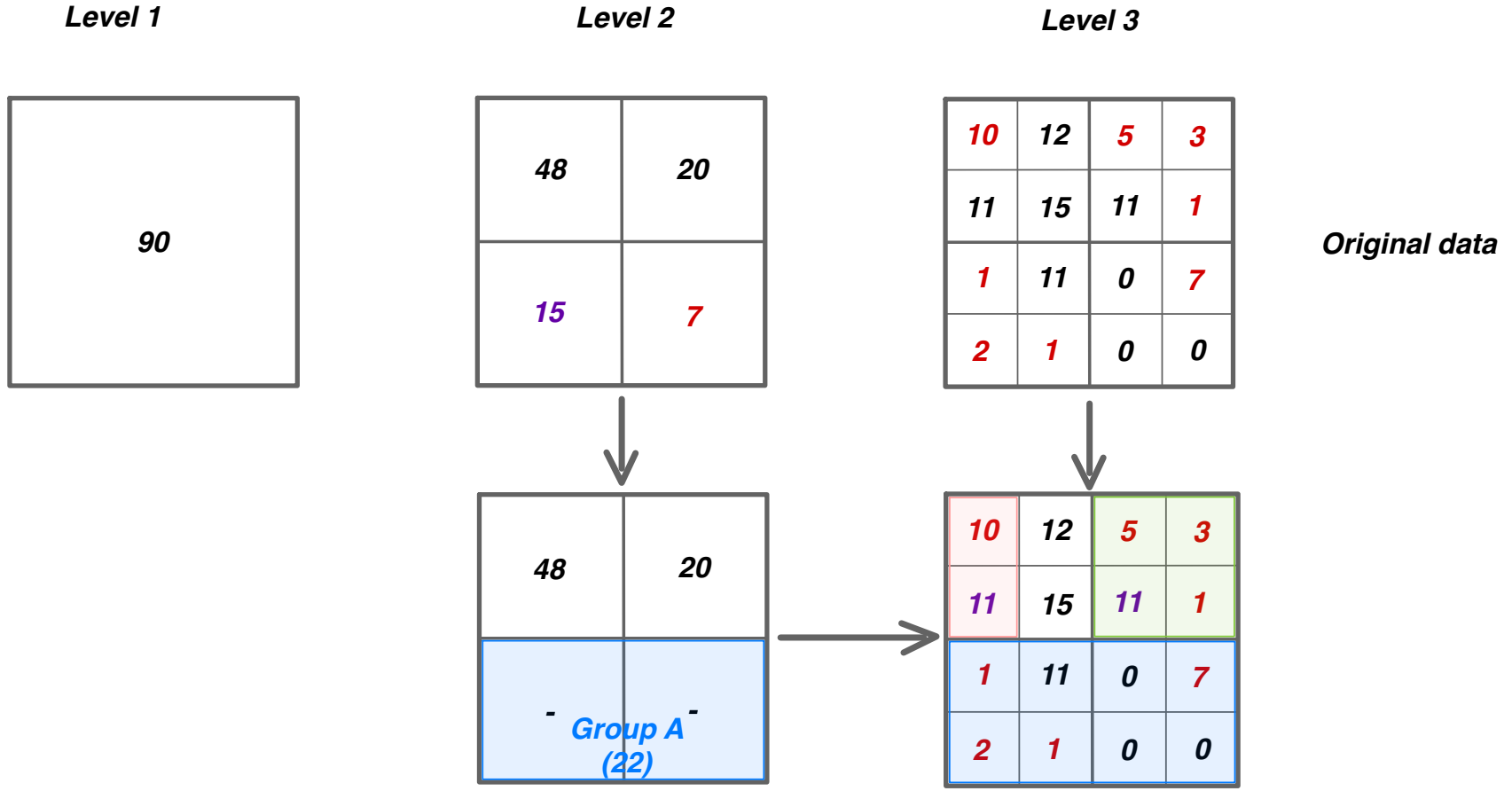
- Suppress tiles below the threshold (Primary suppression)
- Suppress additional tiles to protect the first ones (Secondary suppression)

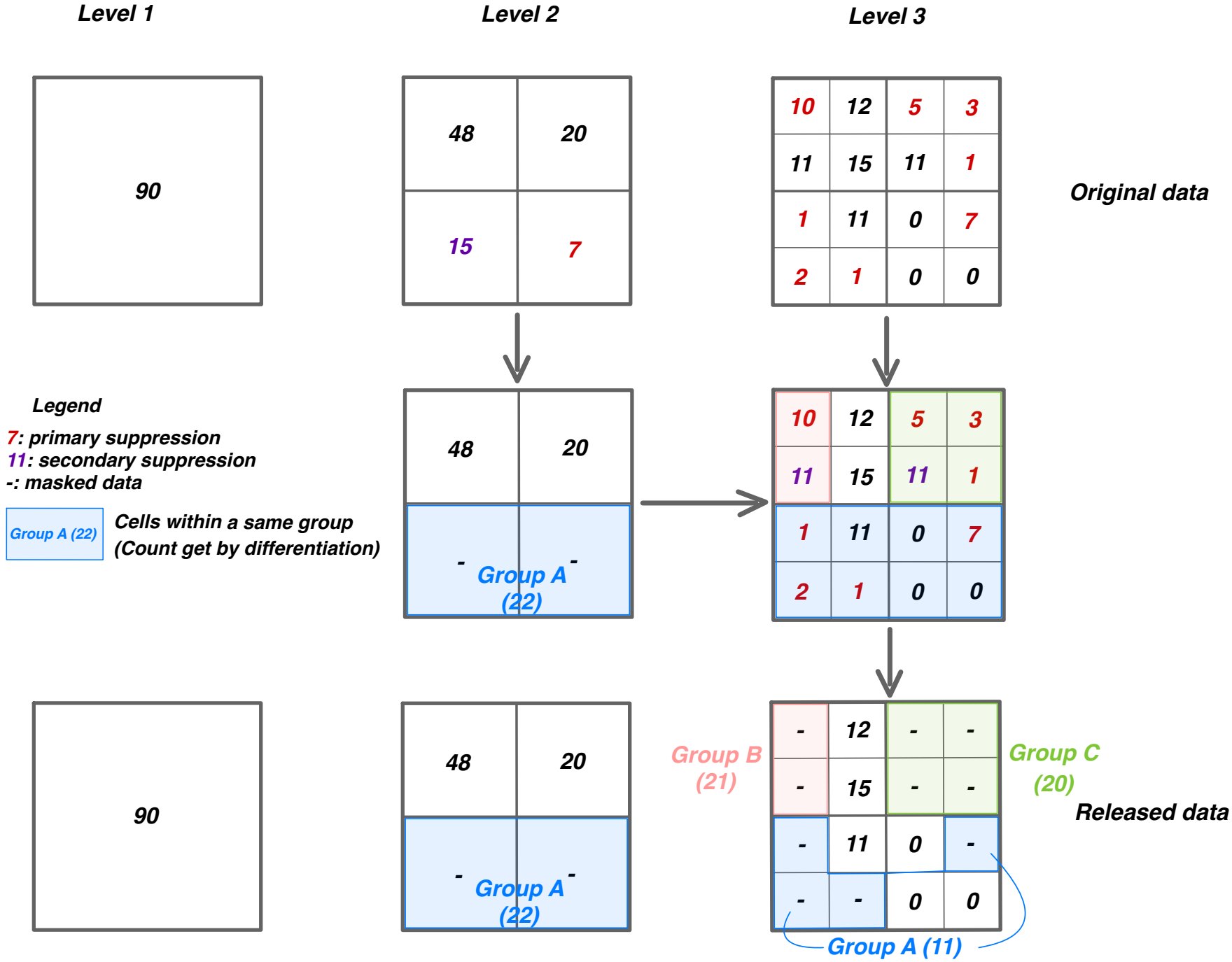
- Suppressive method ;
- At a first glance, the problem is like a problem of suppression in a table with one hierarchical variable (nested grid levels) ;
- But :
  - With more than 500 000 tiles of 1km, the classical optimisation's program is not fitted for this problem ;
  - We need to take into account a specific interlevel differentiation.


Level 1
90
Level 2
48
20
15
7


Level 1
Level 2
90
48
20
15
7
48
20
- Group A -
(22)


Level 1
Level 2
Level 3
90
48
20
15
7
10
12
5
3
11
15
11
1
1
11
0
7
2
1
0
0
Original data
48
20
- Group A -
(22)


Level 1
Level 2
Level 3
90
48
20
15
7
10
12
5
3
11
15
11
1
1
11
0
7
2
1
0
0
Original data
48
20
- Group A -
(22)


Level 1
Level 2
Level 3
90
48
20
15
7
10
12
5
3
11
15
11
1
1
11
0
7
2
1
0
0
Original data
Legend
7: primary suppression
11: secondary suppression
-: masked data
Group A (22)
Cells within a same group
(Count get by differentiation)
Group A
(22)
Group A (11)
Group B
(21)
Group C
(20)
Released data

- The interlevel differentiation is avoided by gathering suppressed cells in groups at each level ;
- These groups have to reach the required threshold ;
- Suppressed tiles within an unsuppressed coarser tile belong to a new group ;
- A group at one level is inherited by its tiles at the next level except if the tile is above the threshold and there is no need to suppress it to reach the threshold with the (or some) other ones.
- Then, some tiles can be released even if, at the previous (coarser) level, the tile is suppressed.

Advantages :

- Information is disseminated at the finest level
- Easier to read and interpret
- Implemented in an R package called *gridy* (not published yet)

Drawbacks :

- Much information is still unpublished
- We'd like to release an information for populated tiles (required by Insee and Eurostat)

Multilevel grid release without suppression

Refill the suppressed cells

To provide an information into all suppressed tiles :

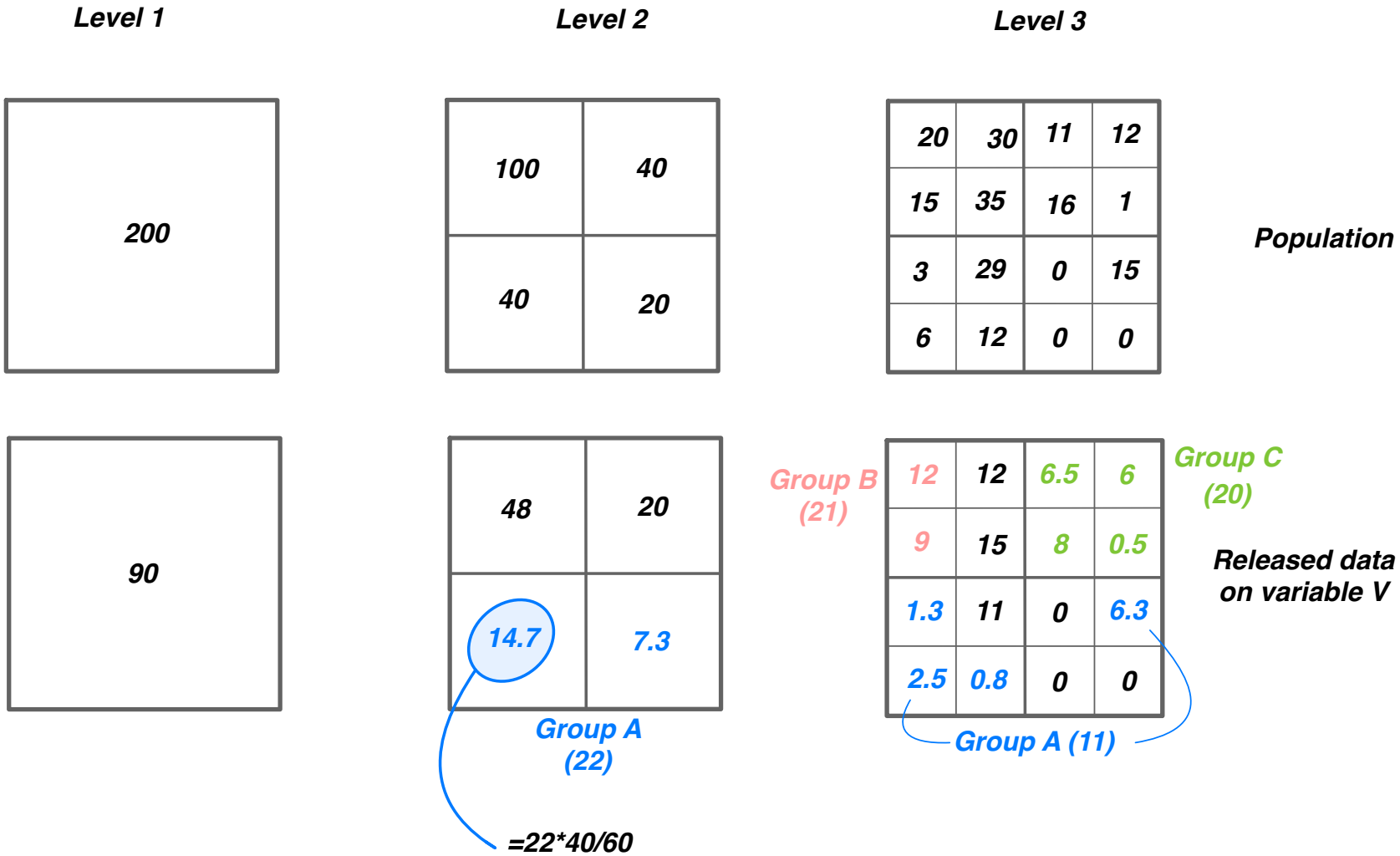
- we use the tiles' groups set up during the suppression process ;
- Suppressed values are imputed ;
- Imputation process done at each level of the grid based ;
- on a distribution in proportion to the cell population in the group.

- Let a group G of suppressed tiles $= (c_1, .., c_n)$ at a given level ;
  - Let $(p_1, .., p_n)$ be their populations (unconfidential information)
  - Let $(Y_1, .., Y_n)$ be their original values (then suppressed) on the released variable V
  - Let $(\hat{Y}_1, .., \hat{Y}_n)$ be their imputed values (then released) on V
- We can then define $Y_G = \sum_{i=1}^{n} Y_i$ and $P_G = \sum_{i=1}^{n} p_i$
- The suppressed value of the cell c_i is replaced by :

$$\hat{Y}_i = Y_G \frac{p_i}{P_G}$$

Multilevel grid release without suppression

Refill the suppressed cells

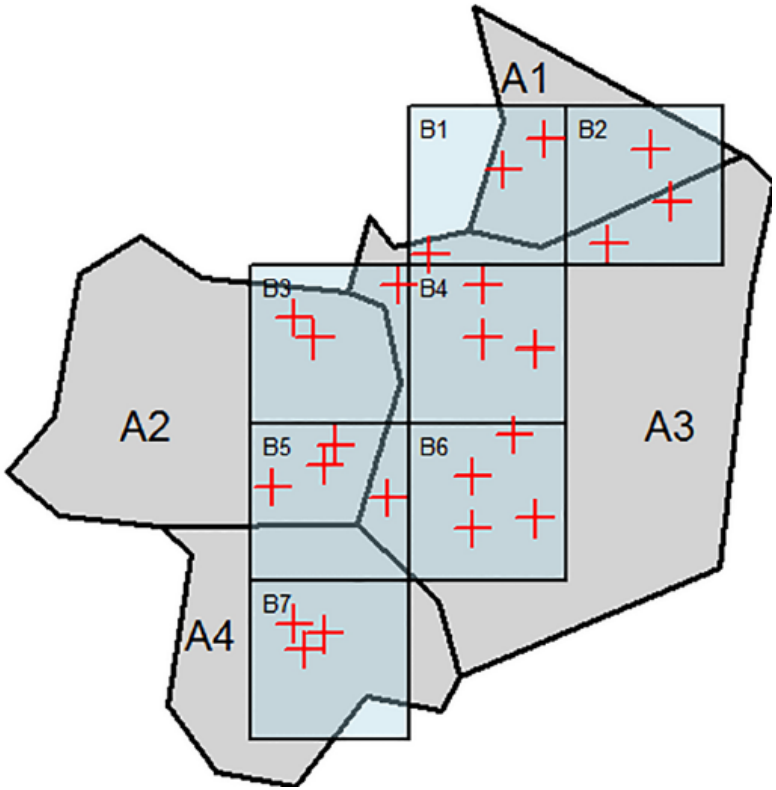


Multilevel grid release

A review

- The suppressive method has become a perturbative one ;
- The perturbation is quite local : tiles within a group are close from each other ;
- But, the method doesn't ensure the additivity ;
- Non integer values are released as counts (assumed by Insee but not by Eurostat).
- Differentiation with non-nested areas such as administrative ones is not yet handled => topic of the second part

CANONICAL EXAMPLE

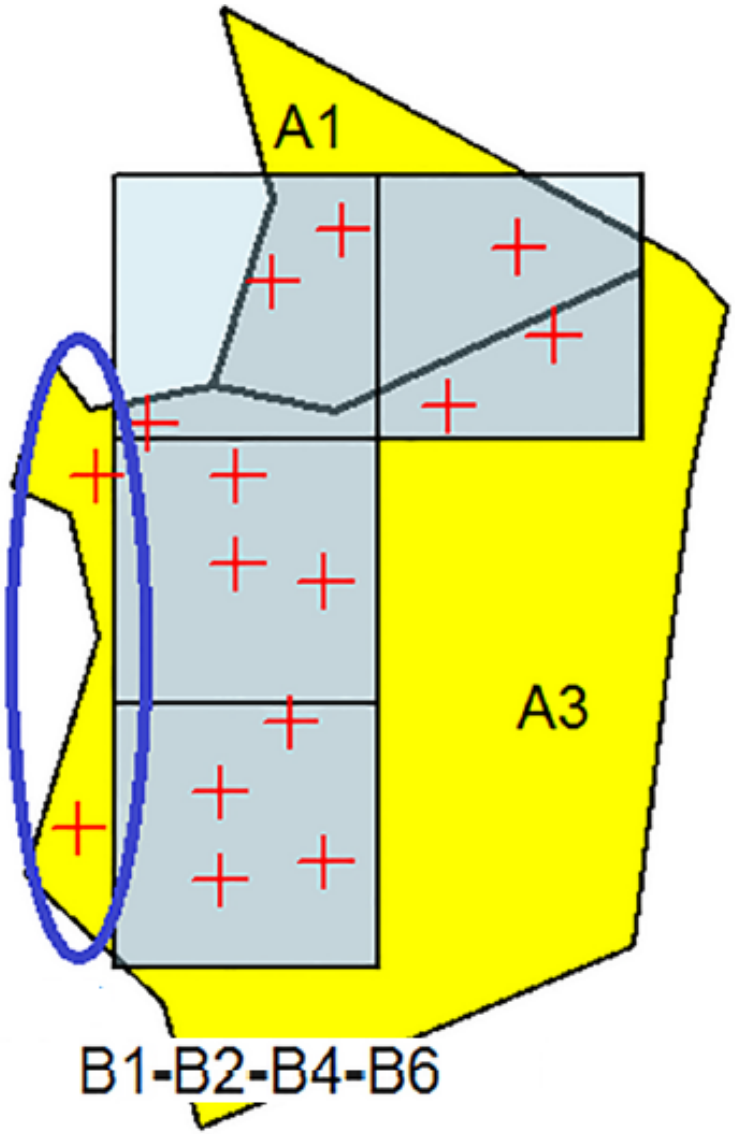


- 23 statistical observations located on 4 administrative units $(A_1, .., A_4)$ and 7 tile cells $(B_1, .., B_7)$
- threshold $t = 3$ observations, no problem if the adminstrative units and the tile cells are taken separately

Internal and External differenciation

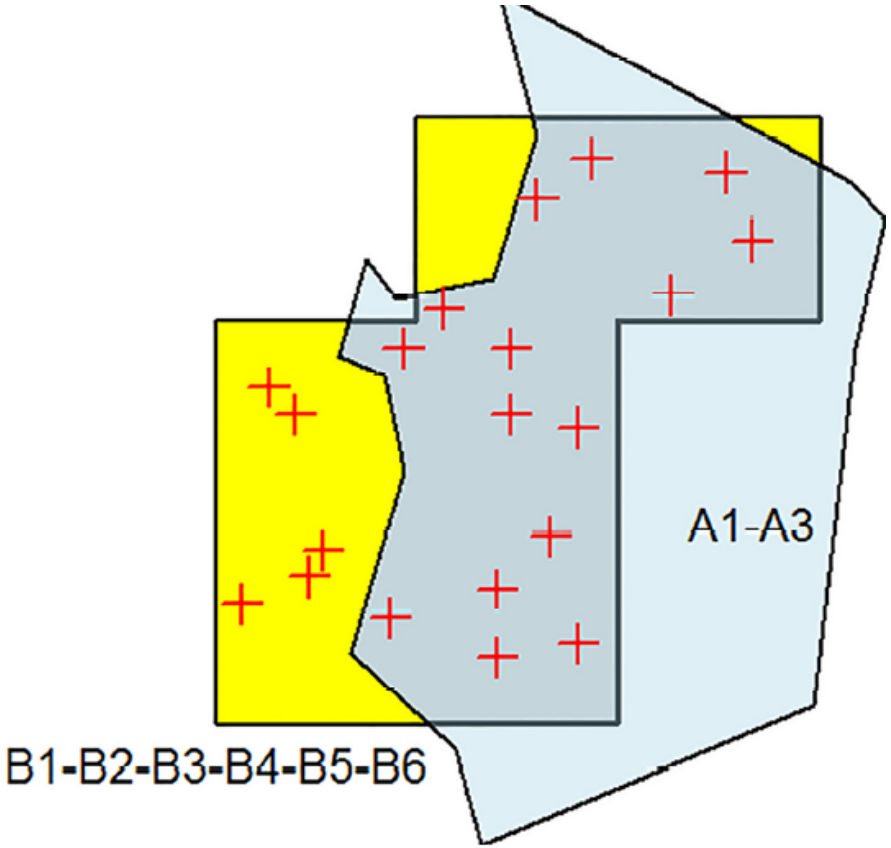
- Let $M_A = (A_1, .., A_N)$ be a territorial classification and $M_B = (B_1, .., B_M)$ another (non-nested with A) territorial classification with $N < M$. In the toy example, $N = 4$ and $M = 7$
- The A_i (resp. B_i) are called the elementary zones
- Let a *region* A be an area composed of P elementary zones of M_A

Internal and External differentiation



- Internal differentiation of Area $A_1 \cup A_3$

Internal and External differentiation

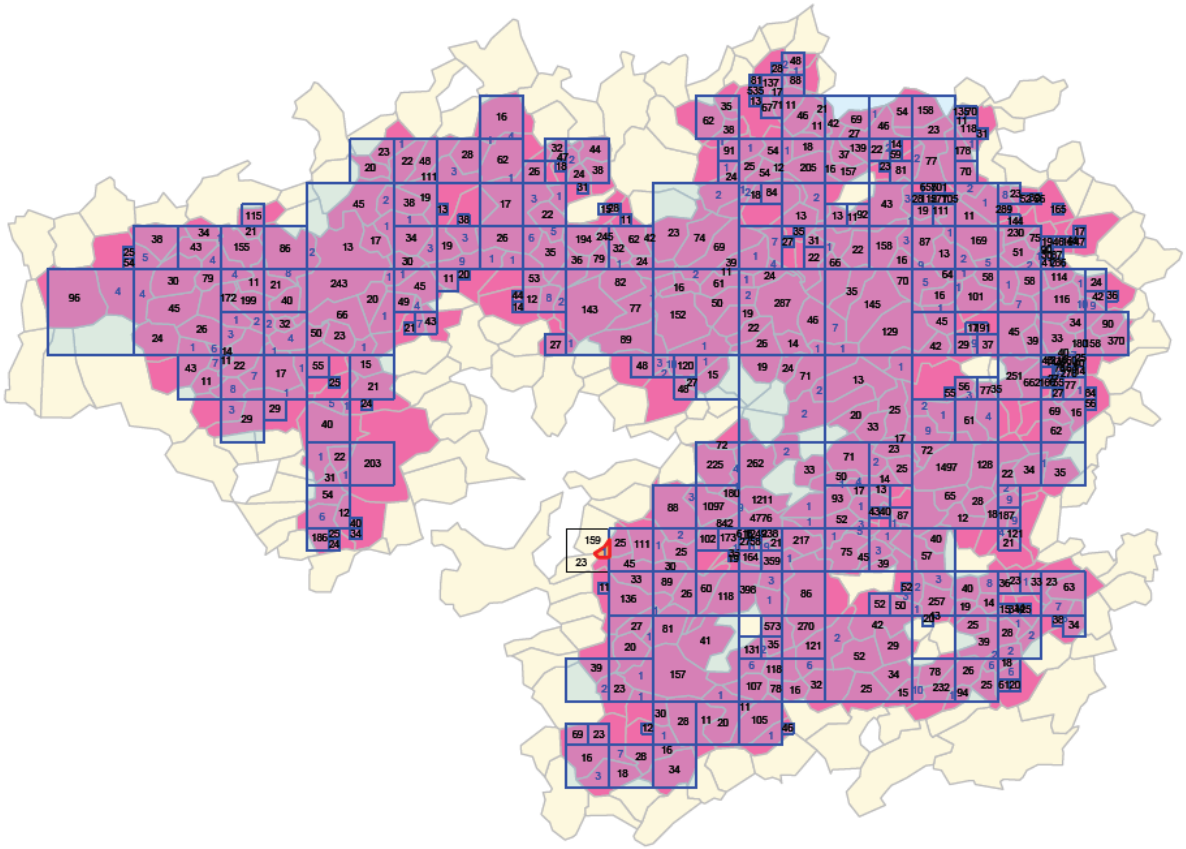


- External differentiation of Area $A_1 \cup A_3$

Complexity

- In the toy example, **both internal and external differentiation** for group $A_1 \cup A_3$ of M_A elementary zones have been checked $\longrightarrow$ this has to be done **for all the possible groups**
- M_B and M_A are **symmetrical**, so checking all M_A groups (2^N groups) is equivalent to checking all M_B groups (2^M)
- Since $N < M$, it is better to check the differentiations on the groups constituted with the elementary zones of M_A

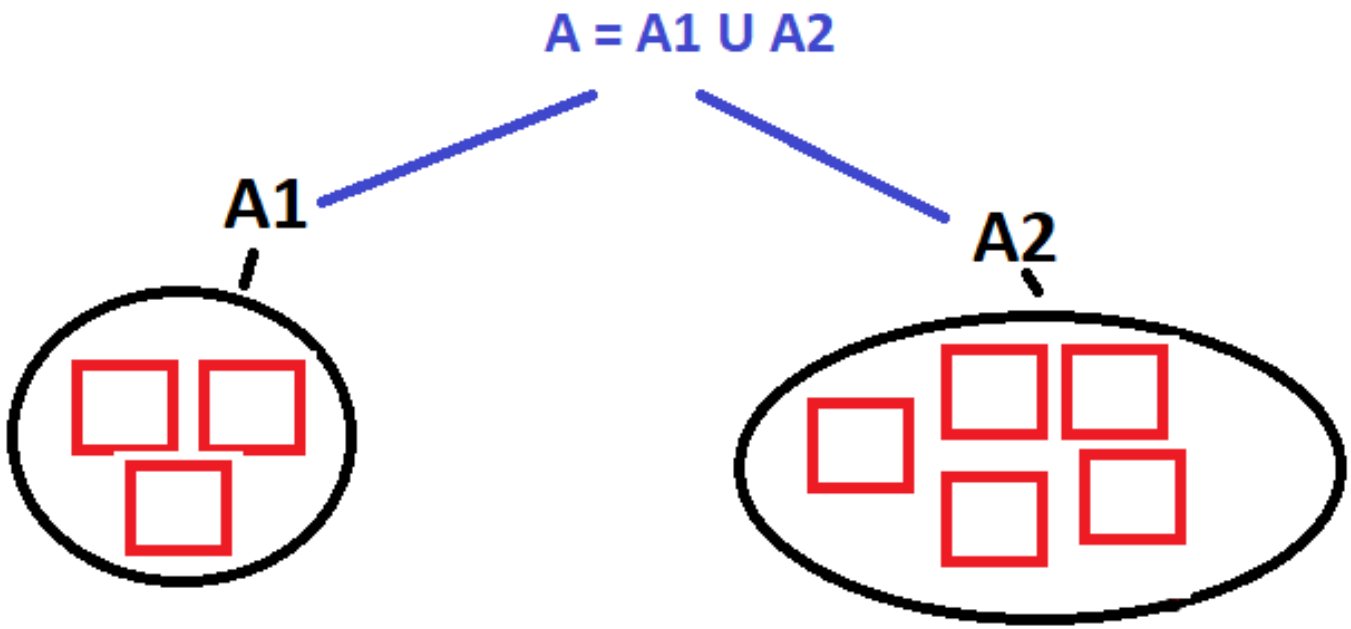
Complexity



- The number of groups to check increases exponentially with the number of elementary zones in M_A !

How to Get rid of the complexity ?

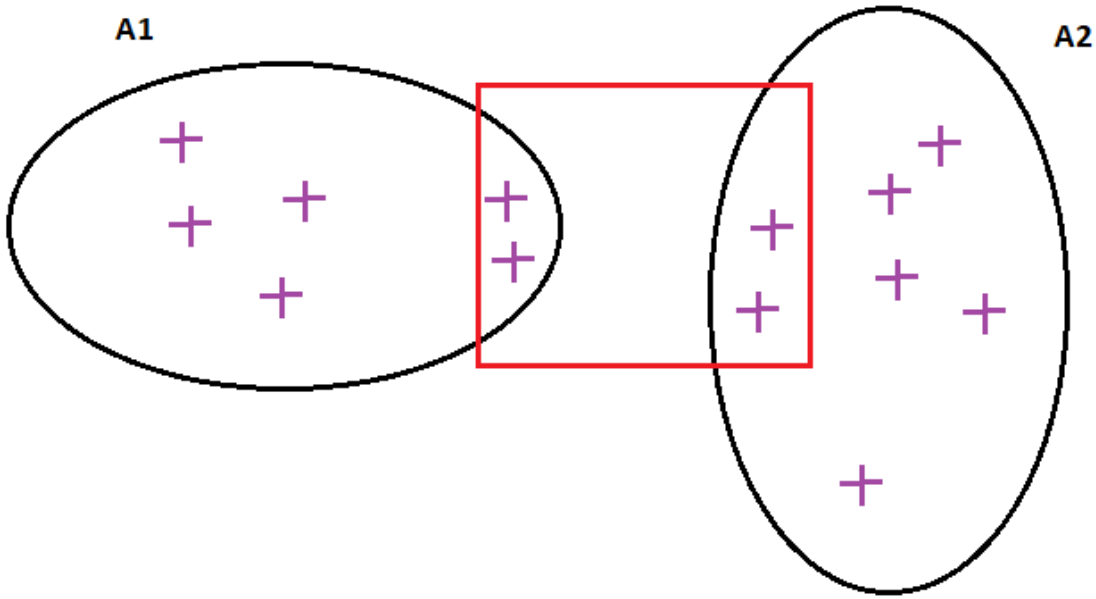
- Idea : restrict search on encapsulating regions A whose elementary zones A_i are connected



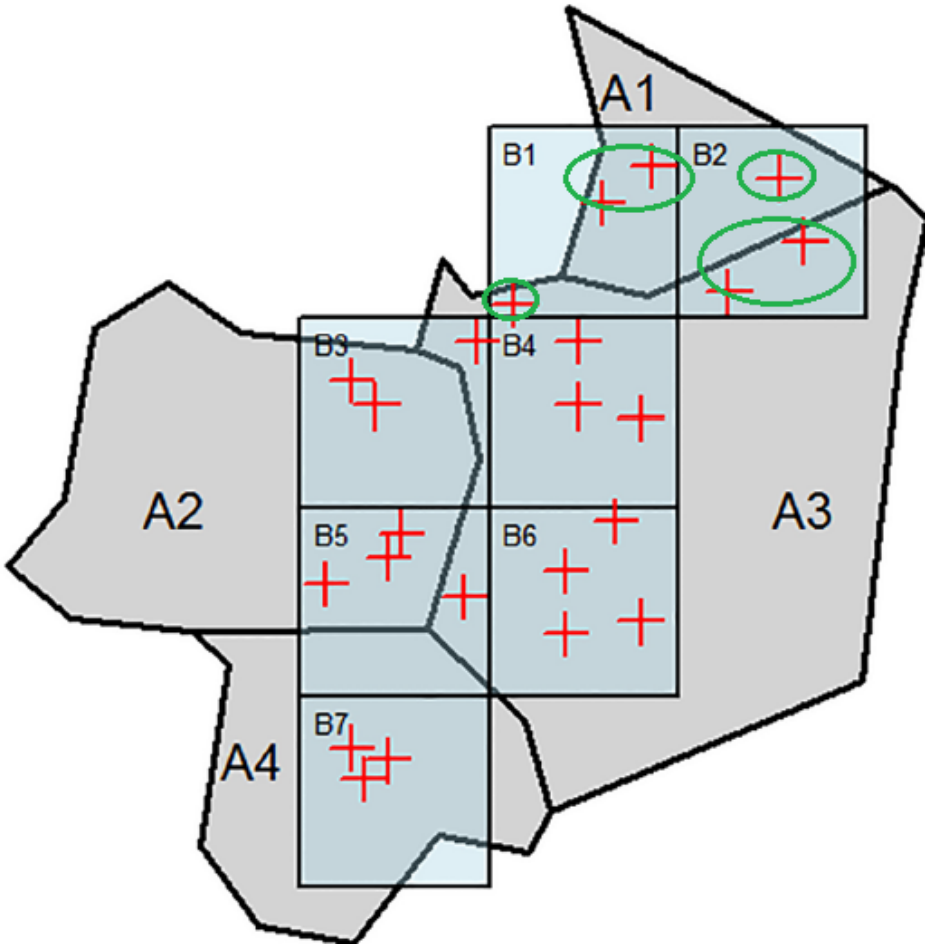
- If group A1 and A2 have already been checked, no need to check A !
- justify a graph representation of the problem, let's go back to the toy example

Graph representation

- Let the elementary zones of M_A be the nodes of our graph
- Two nodes are connected by an edge if the two corresponding elementary zones are connected



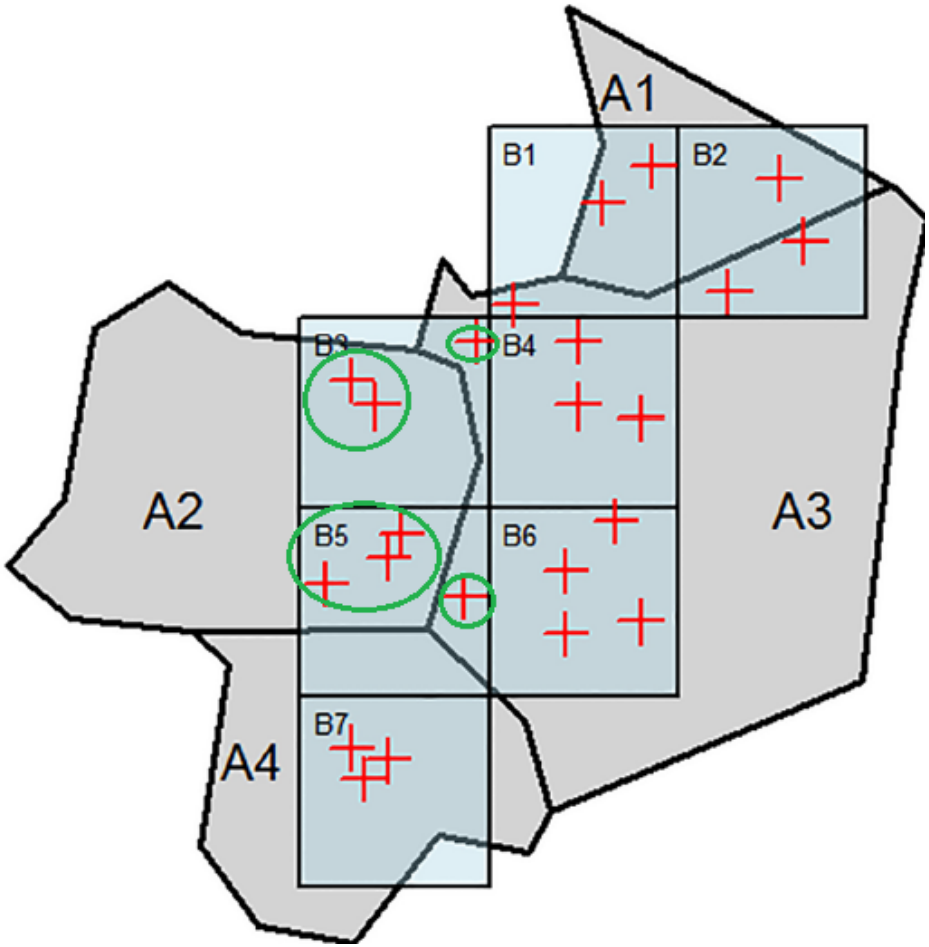
Graph representation



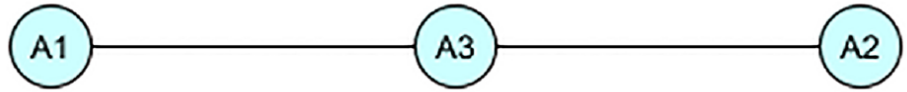
geo representation

graph representation

Graph representation

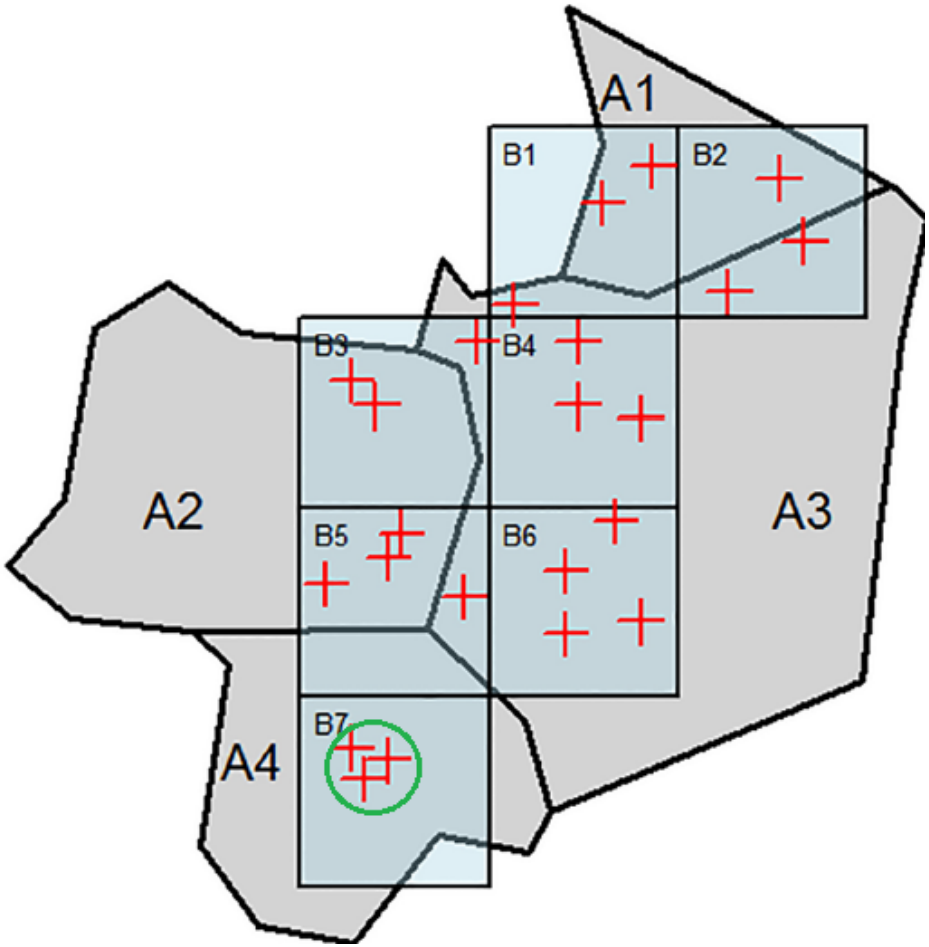


geo representation

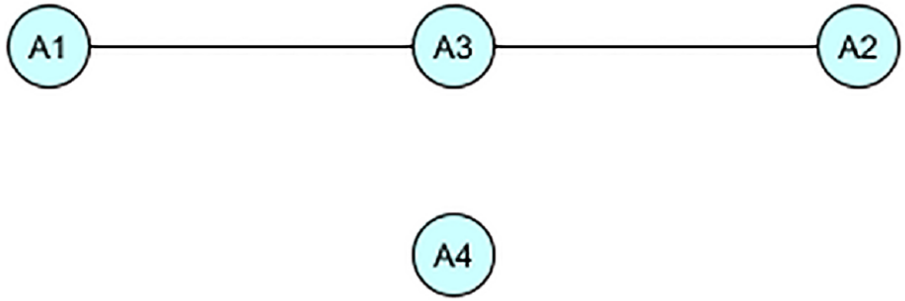


graph representation

Graph representation

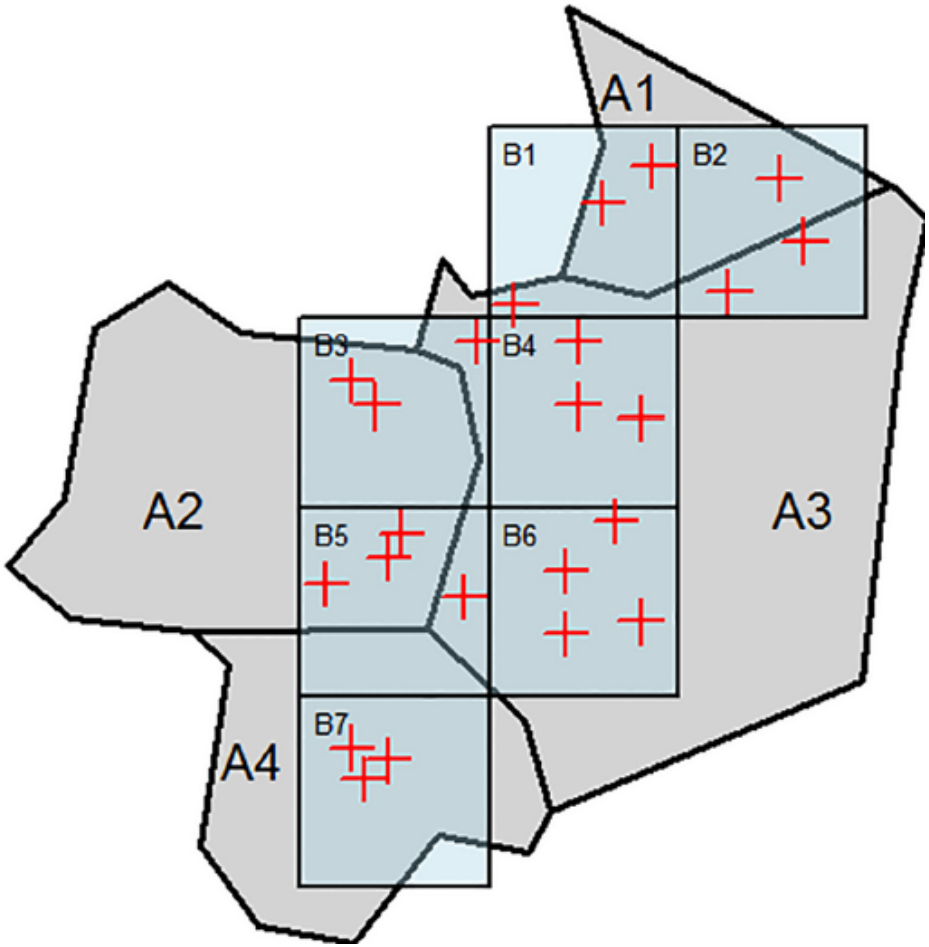


geo representation

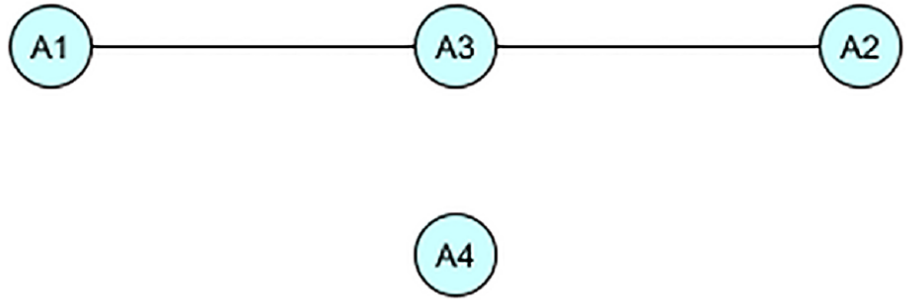


graph representation

Graph representation



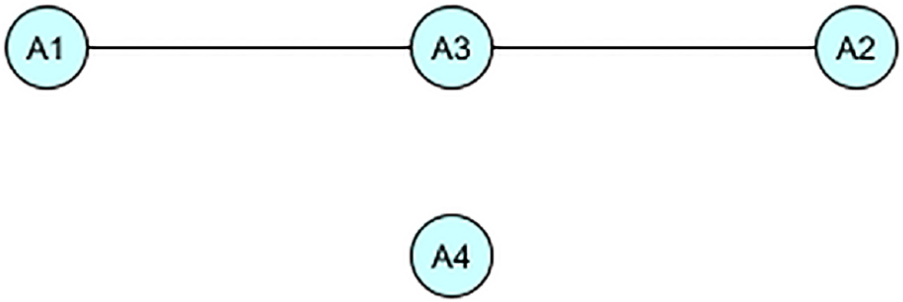
geo representation

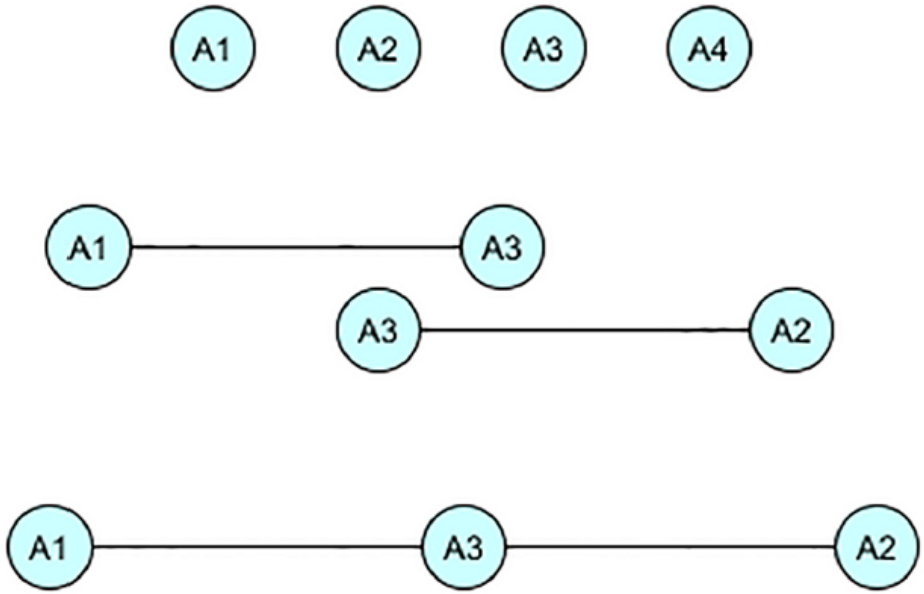


graph representation

SUB GRAPHS

- We can limit the disclosure search to the set of subgraphs

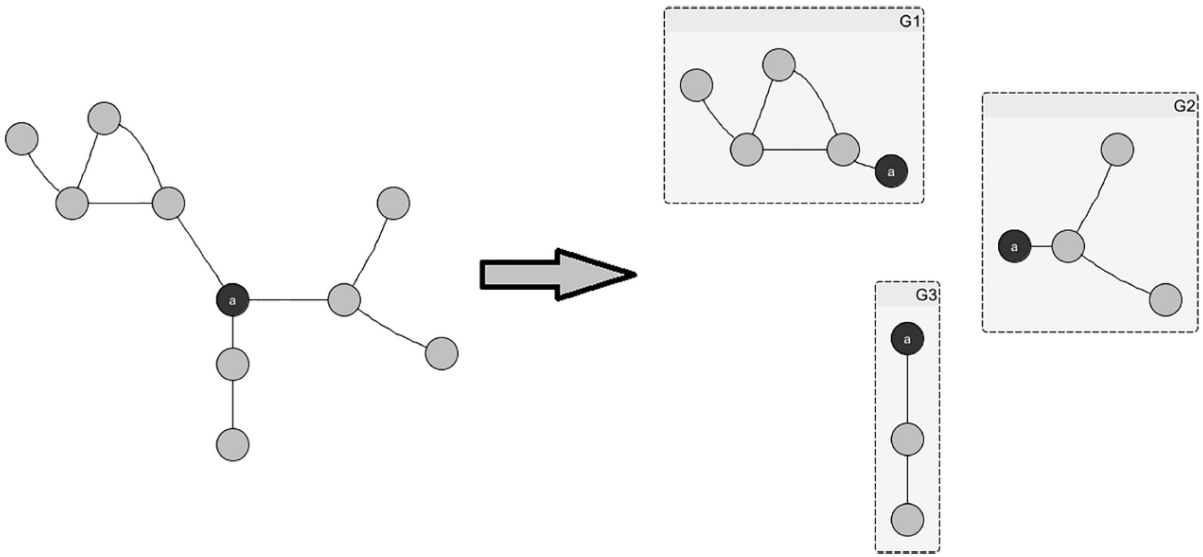




- 7 differences to compute instead of 2^4
- N = 35,000 in France $\longrightarrow$ billions of subgraphs : need to simplify

Reducing the graph

- Idea 1 : merge the non-risky nodes
- Idea 2 : split the graph into independent components



How to solve differentiation issues ?

- Detecting differentiation issues is a big challenge
- An R package that implements the presented method has been developed at INSEE : **diffman**

Summary

Two issues to deal with :

- Protection of census grid data is based on a method implemented for tax grid data.
- As same data are released on municipality level, we have to deal with differentiation issues.

Then, the steps are the following ones :

1. Identification of cells that can lead to **differentiation** ;
2. **Primary suppression** in every grid level, with two sources :
   - Frequency rule (threshold set to 11 households)
   - Risk of differentiation (risky cells from 1^{st} step)
3. **Secondary suppression**
4. **Filling the blanks** :
   - To release information in all *populated* tiles
   - Blanks replaced by imputed values
   - Imputed values = proportional distribution of the sum within a group of suppressed cells

Bibliography

- Confidentiality of gridded data
  - Branchu, Costemalle, Fontaine, *Données carroyées et confidentialité*, 2018, http://www.jms-insee.fr/2018/S23_2_ACTE_COSTEMALLE_JMS2018.pdf, (FR) ;
  - Lagonigro, Oller, Martori, *A quadtree approach based on European geographic grids : reconciling data privacy and accuracy*, SORT, n°41, 217, pp139-158
- Differentiation issues
  - Costemalle, *Detecting Geographical Differencing Problems in the Context of Spatial Data Dissemination*, Statistical journal of the IAOS, 2019, pp559 – 568. https://content.iospress.com/articles/statistical-journal-of-the-iaos/sji190564

Thank you for your attention !