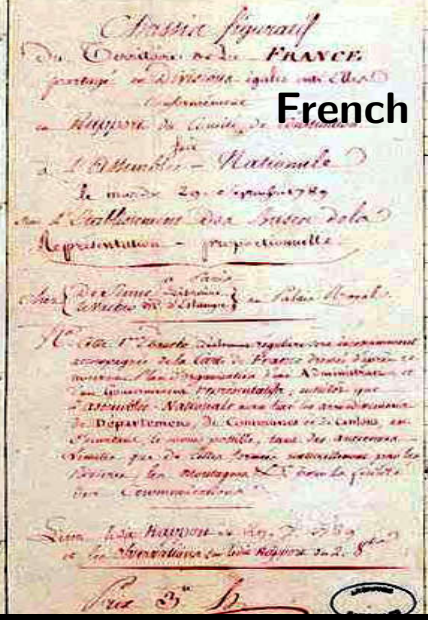
Dissemination of Statistical information in a (200m)2 Grid Dataset

Managing the confidentiality

Vincent Loonis

Geographical repositories and methods division
French national institute of statistics and economic studies (Insee)

Thursday, 12 November



Introduction

- INSEE decided to explore disseminating in a 200m x 200m national grid datasets the bulk of statistical information, including tax files.
- These tax files contain very sensitive variables, such as tax incomes, which led INSEE to pay careful attention to disclosure problems.
- The purpose of this presentation is to present the solution adopted by INSEE to release these variables.
- This presentation is in line with the principle 5 of the GSGF and the current work of the EG-ISGI Task Team on confidentiality.

Insee
Mesurer pour comprendre

Agenda

1. Presenting the French tax files,
2. Dealing with the primary statistical secrecy
3. Dealing with secondary statistical secrecy (geo differencing)

Insee
Mesurer pour comprendre

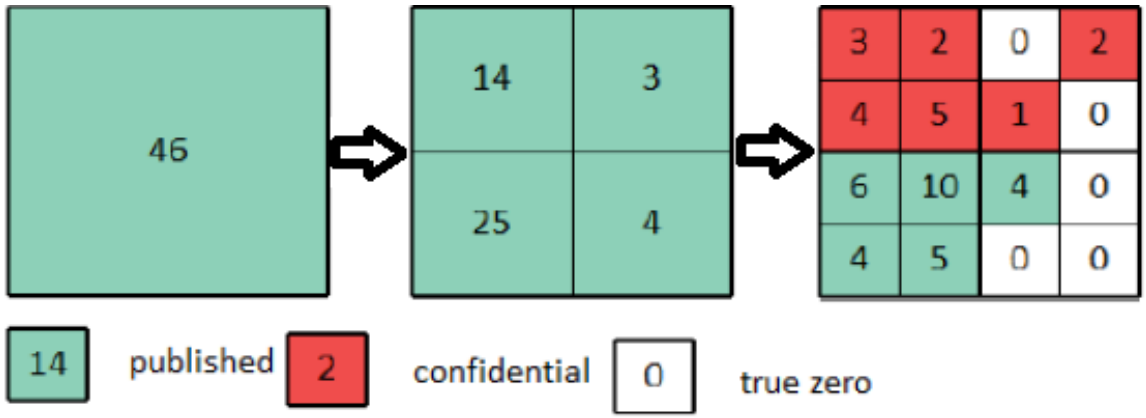
The tax files

- The tax files are georeferenced but also a comprehensive statistical source on dwellings, households, individuals and incomes.
- Any tax variable, other than the number of individuals, is a sensitive variable
- For any sensitive variable, according to fiscal secrecy rules, no statistical results must be released in a grid or a table cell having less than 11 households (primary secret).
- To comply with this rule, we were asked not to use pertubative methods.
- We explored various aggregations (or desaggregation) processes relying on the quadtree method.

Insee
Mesurer pour comprendre

Dealing with the primary statistical secret

Figure: Quadtree method leads to a non optimal dissemination process with cells of different size (Threshold is 3 here)



- Nevertheless, it is possible to find methods that
  - meet the primary secret,
  - allow the dissemination of the true information on a nested system of cells,
  - allow the dissemination of more information that the quadtree,
  - include an imputation process, for the cells below the threshold, that is compliant with the previous nested system and, as such, preserves the spatial pattern of the underlying phenomenon.

Dealing with the primary statistical secret

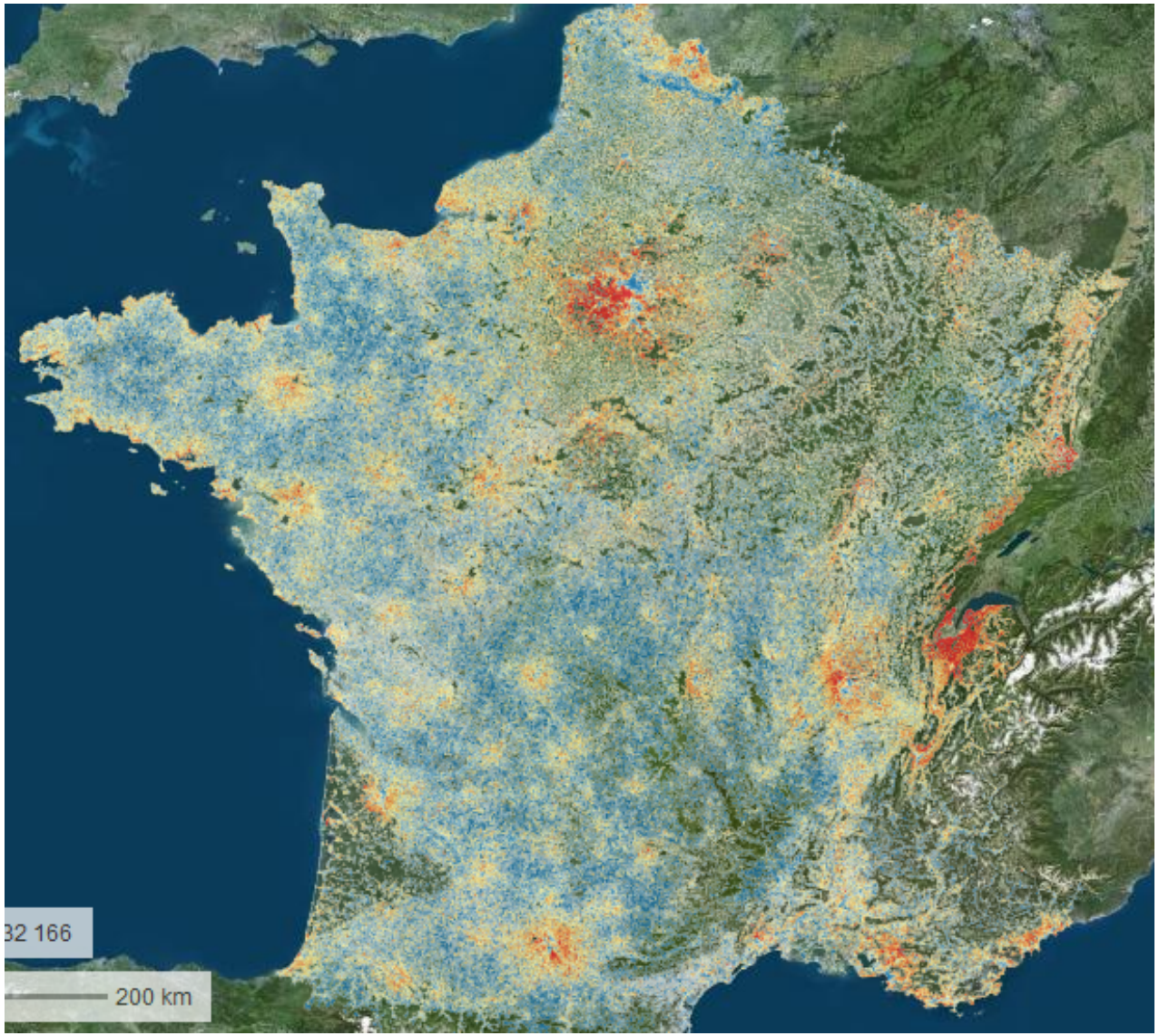


Figure: Results at a small scale (In collaboration with IGN)

Dealing with the primary statistical secret

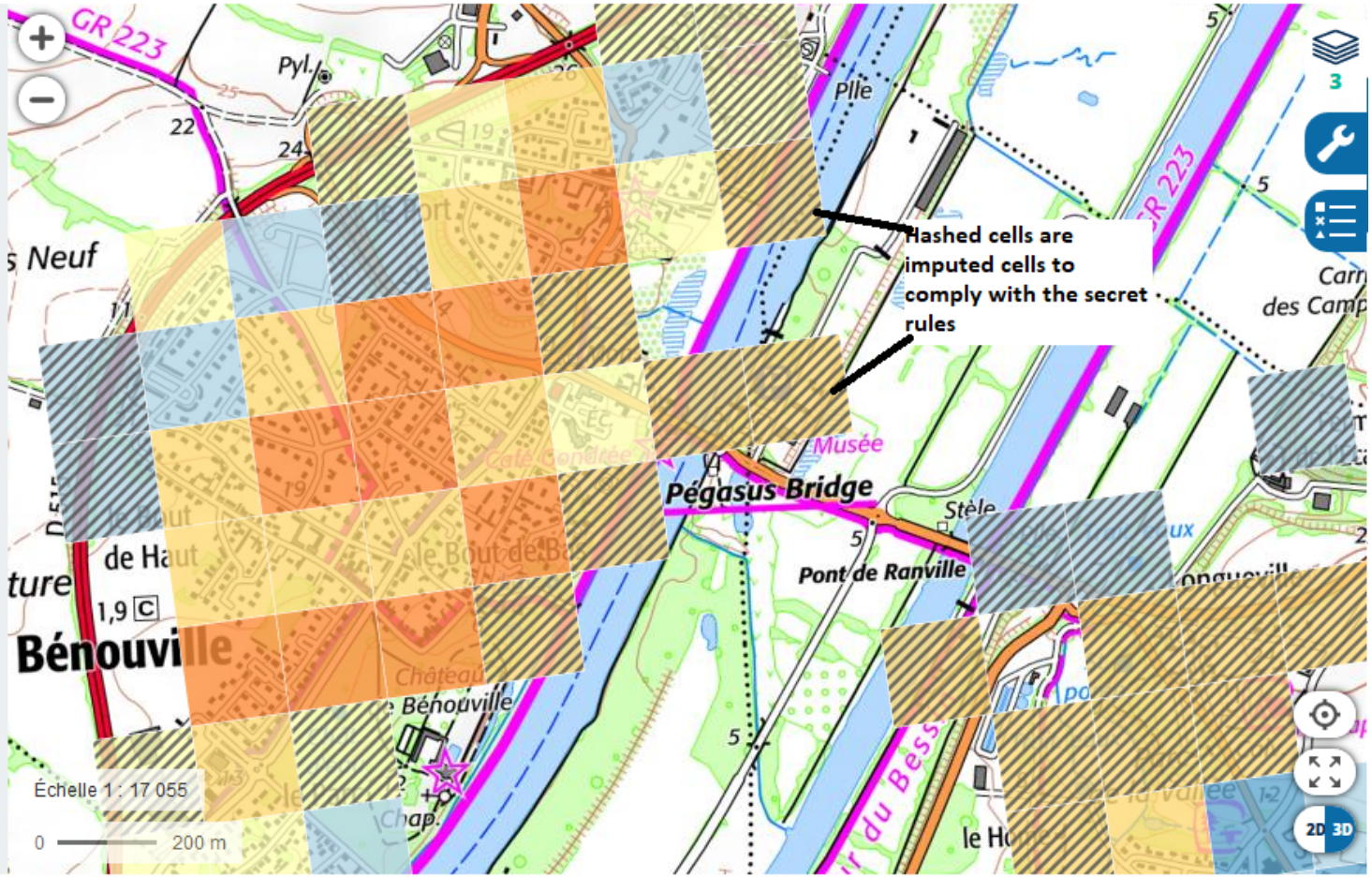
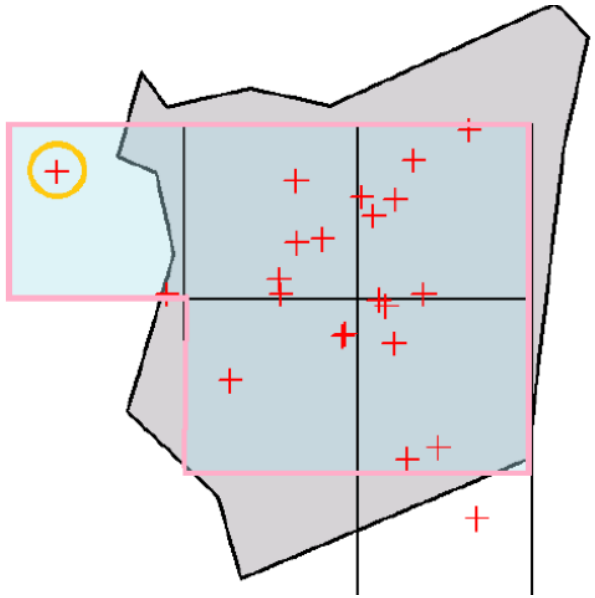


Figure: Results at a large scale (In collaboration with IGN)

Dealing with secondary statistical secrecy (geo differencing)

Insee keeps on disseminating the same statistical information at the municipality level, that might lead to a breach of confidentiality, called geo-differencing.

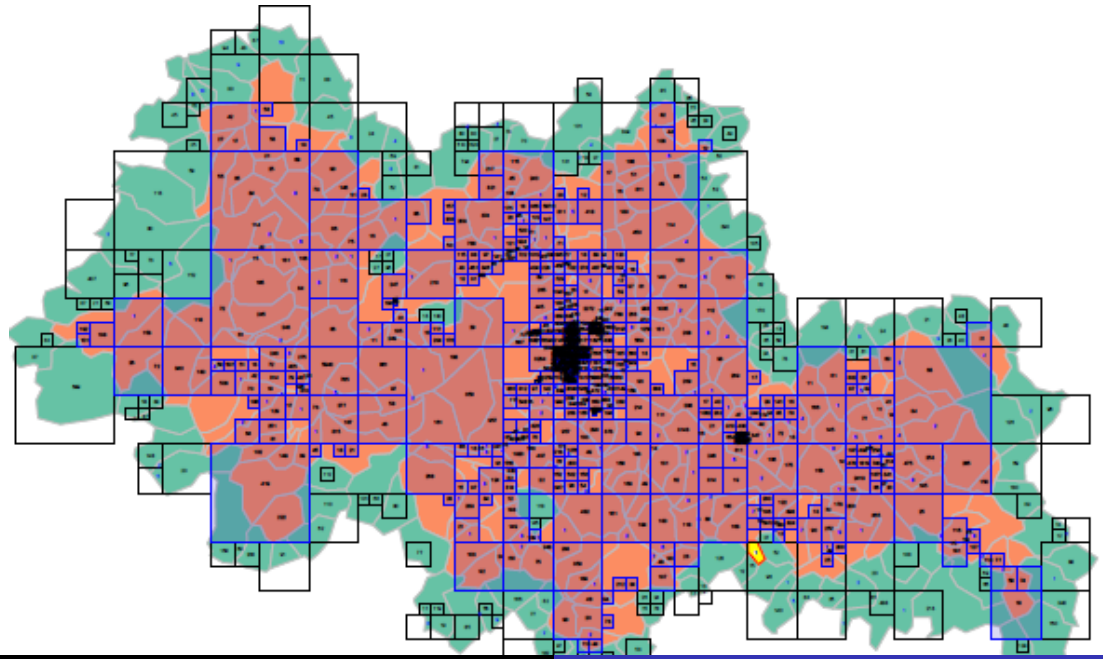
Figure: Example of geographical differentiation based on grid data. One can identify the cross surrounded by a yellow circle by geographical differentiation.



Dealing with secondary statistical secrecy (geo differencing)

- Geo differencing issues may occur with any arbitrary P cells and any abritrary Q municipalities.
- Insee set up a methodology relying on Graph theory, to identify 10 000 households among 30 million living in 35 000 municipalities and 2 million 200-meter cells.

Figure: A complex example



References

- Martin Behnisch, Meineln Gotthard, Sebastian Tramsen, and Disselmann. Using Quadtree representations in building stock visualization and analysis. Erdkunde, 2013.
- Vianney Costemalle. Detecting geographical differencing problems in the context of spatial data dissemination. Statistical Journal of the IAOS, pages 559–568, 2019.

!! THANK YOU FOR YOUR ATTENTION !!

Insee
Mesurer pour comprendre