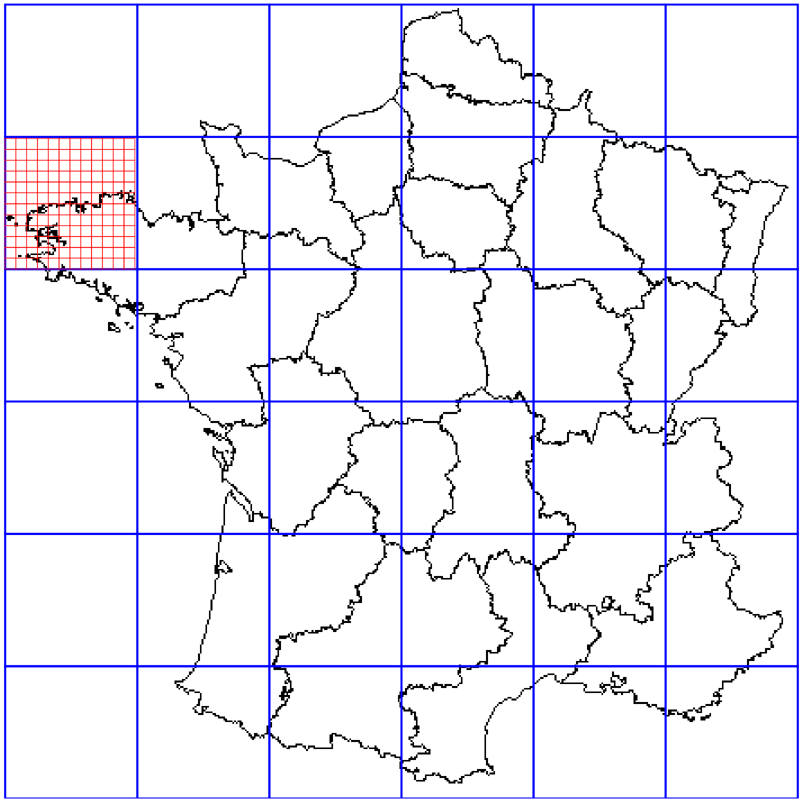
Dissemination of Sensitive Variables in a $(200m)^2$ Grid Dataset :

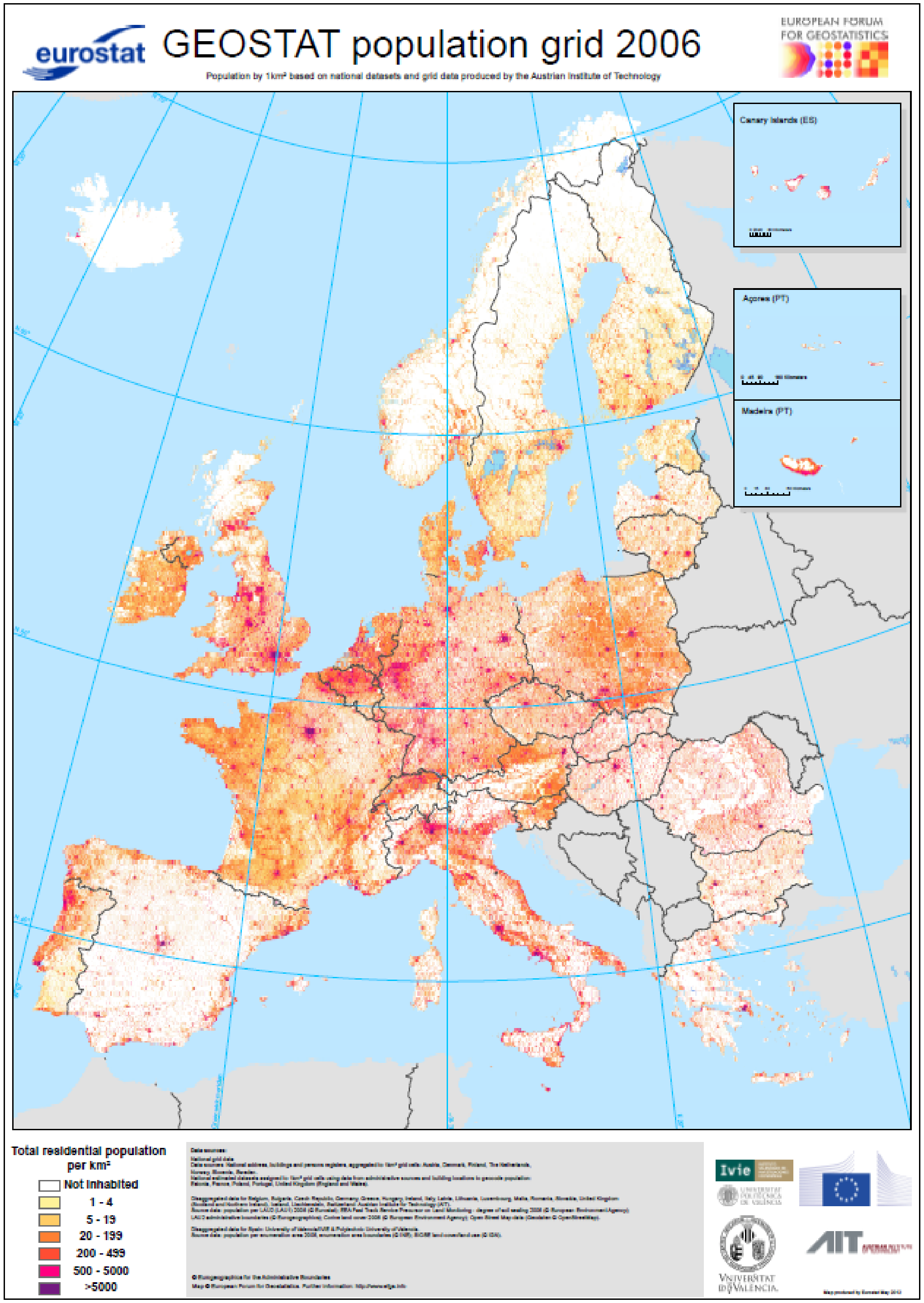
The French Experience

Vincent Loonis

Insee

Beijing, 9-12 june 2014




eurostat
GEOSTAT population grid 2006
Population by 1km² based on national datasets and grid data produced by the Austrian Institute of Technology
EUROPEAN FORUM FOR GEOSTATISTICS
Canary Islands (ES)
Açores (PT)
Madeira (PT)
Total residential population per km²
□ Not inhabited
1 - 4
5 - 19
20 - 199
200 - 499
500 - 5000
>5000
Data sources:
National grid data
© Eurogeographics for the Administrative Boundaries
Map © European Forum for Geostatistics. Further information: http://www.efgs.info
Ivie
AIT

Introduction

The GEOSTAT project’s goal was to create a 1km^2 grid dataset for the dissemination of results of the censuses carried out by the various European countries.

Introduction

As a follow up to this successful project, the French National Statistics Office : the Insee, decided to further delve into the building and the dissemination **of 200m x 200m** national grid datatsets of other statistical sources. Among them, tax files.

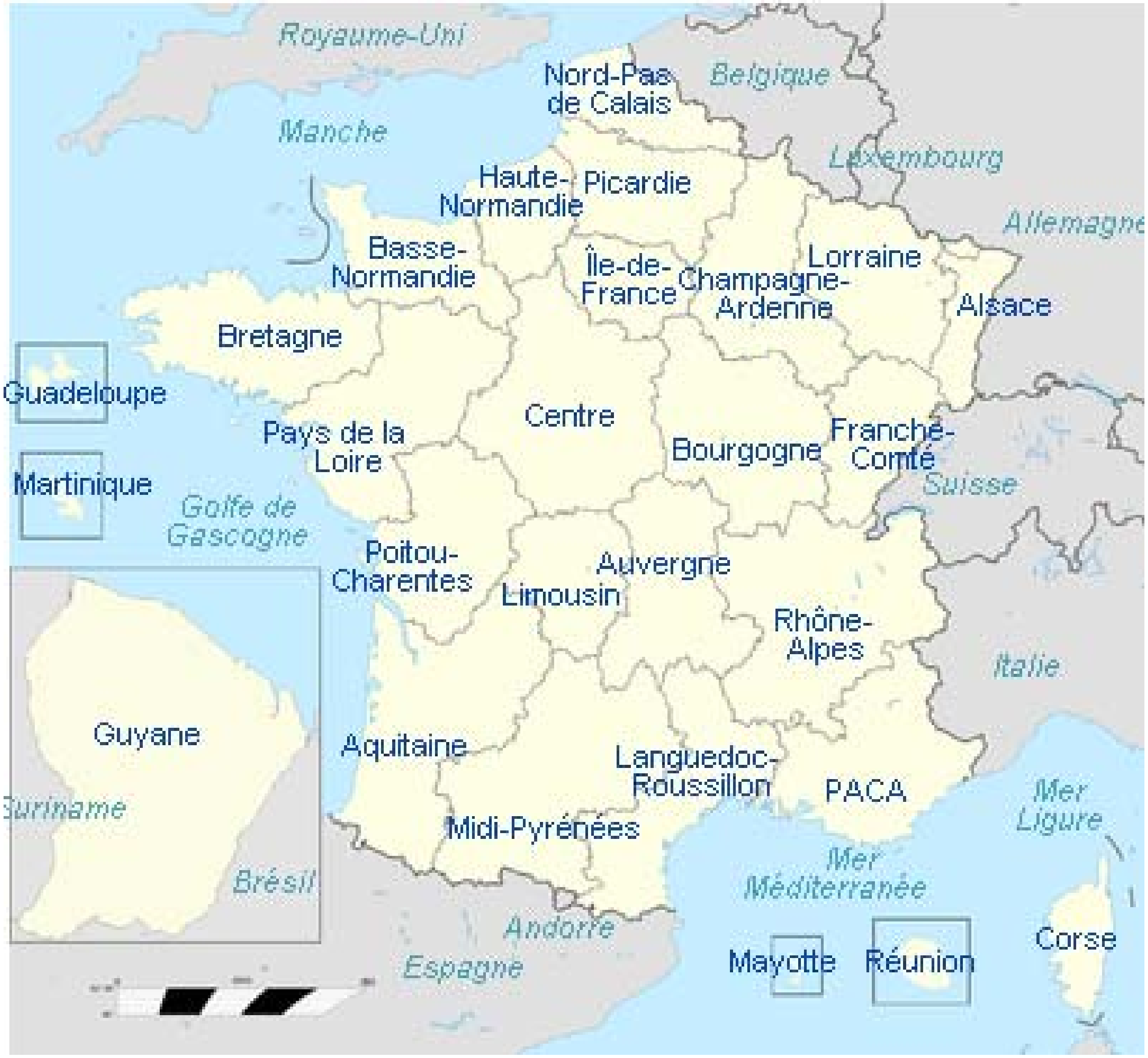
The tax files contain very sensitive variables, such as taxable income, which led the Insee to **pay careful attention to disclosure problems.**

Outline

- Background
  - **The French spatial reference system**
  - **Capabilities for geocoding**

- Methodology
  - **<u>Aggregation of cells in rectangles</u> to comply with some fiscal secrecy rules**
  - **processing of « *risky* » variables**
  - **Tackle the differencing problem**
- Dissemination
- Challenges and conclusions

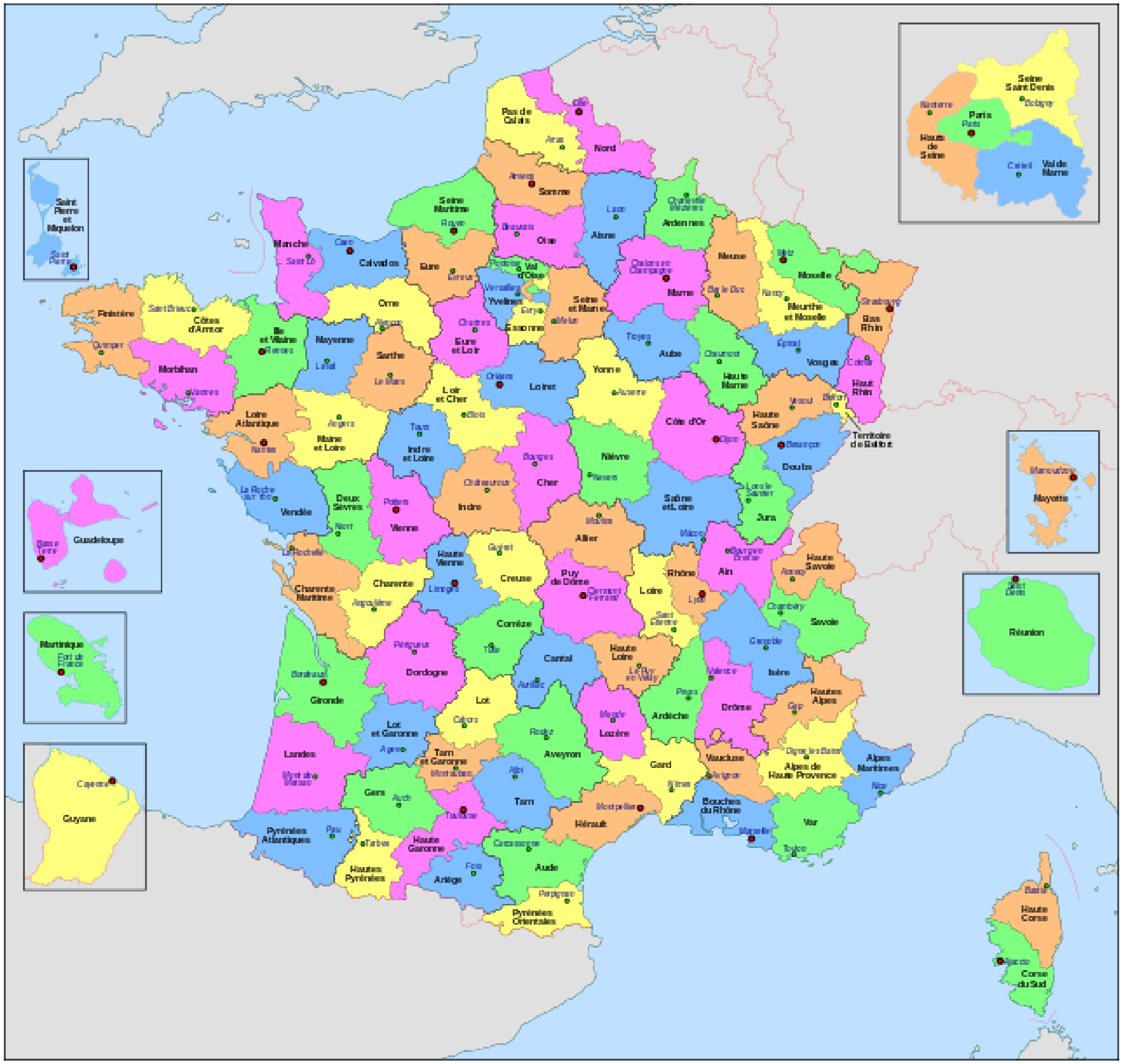
The Currently French spatial reference system



27 Administrative regions including
5 overseas regions

There is an ongoing debate to reduce the number of administrative regions.

The French spatial reference system

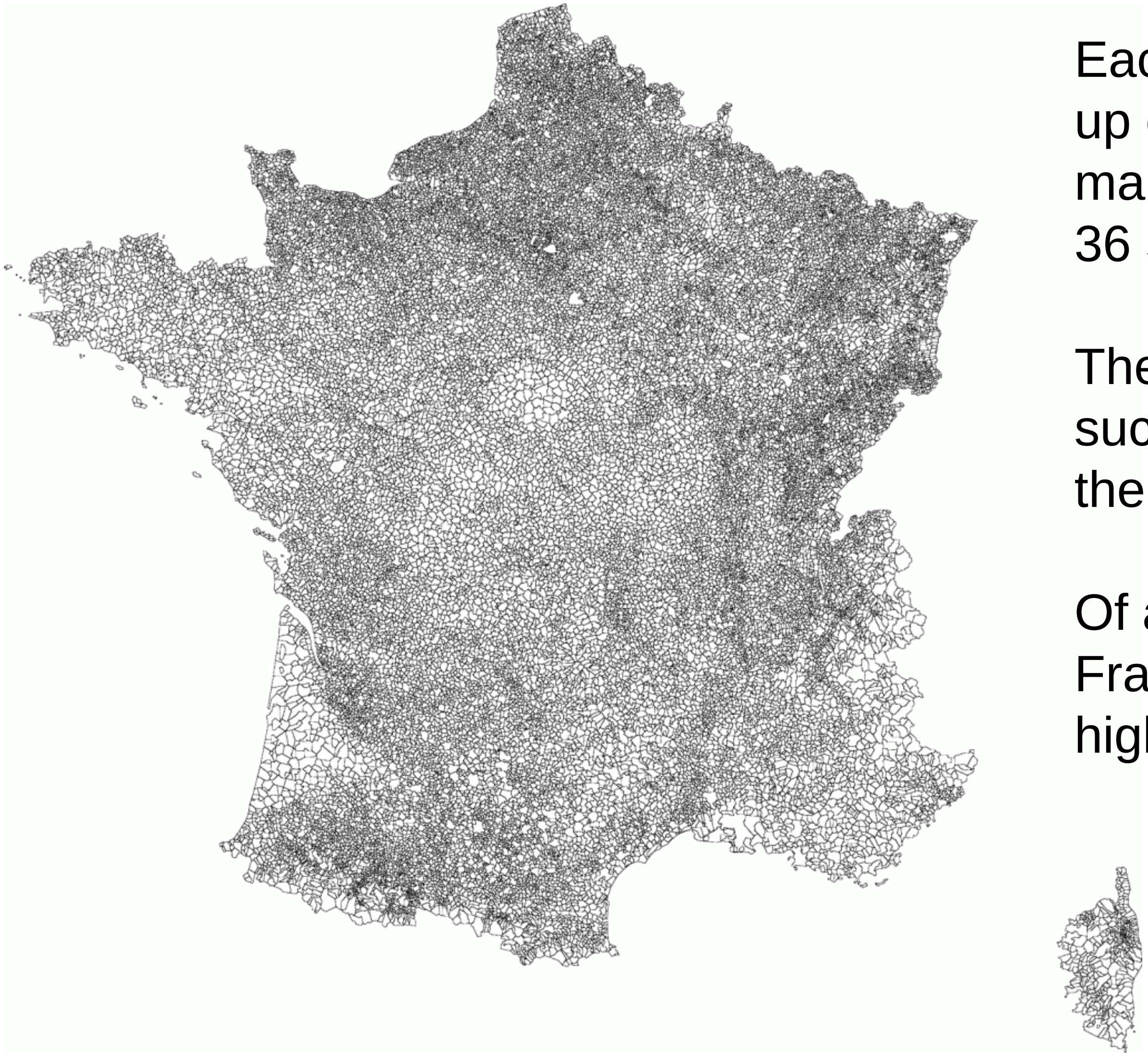


Each region is made up of departments, making a total of 101 departments.

The departments were set up in 1790 to meet a geographical criteria :

The main town of a department should be accessible, for any inhabitants, in less than two days, by horse.

The French spatial reference system

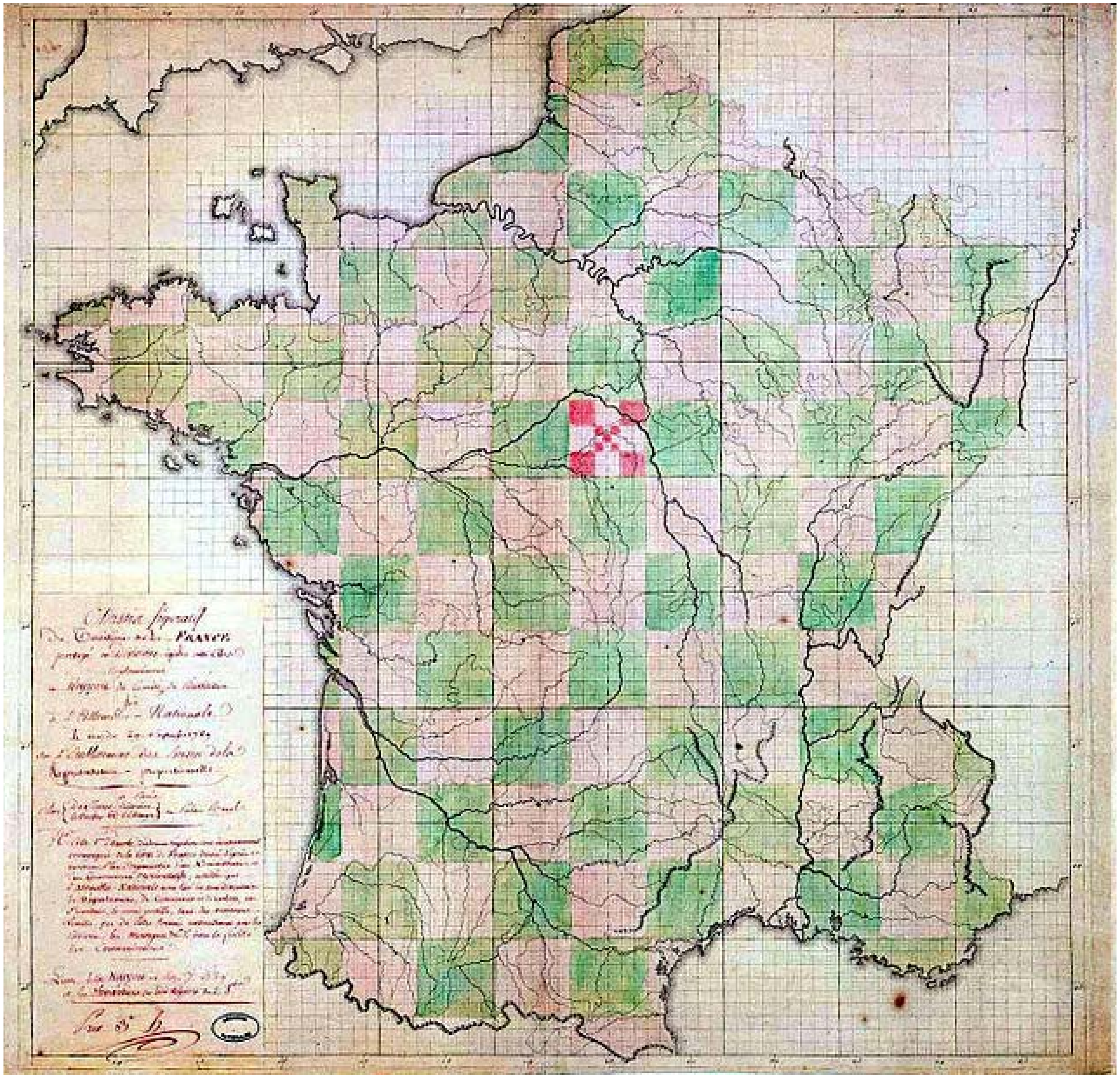


Each department is made
up of municipalities,
making a total of
36 500 muncipalities.

The municipalities are the
successors of the parishes of
the Middle Ages..

Of all european countries,
France is the country with the
highest number of municipalities.

The French spatial reference system



A proposal from 1789 to set up the departments and municipalities as the cells of a grid.

The French spatial reference system

Derived classifications with a more functional focus :

French agglomerations **:** aggregations of municipalities which meet criteria linked to the continuity of built-up areas and to the number of inhabitants.

French urban areas : aggregations of agglomerations and municipalities which meet criteria linked to the number of jobs located in the area or to the number of commuters within the area.

The French spatial reference system
Sub municipal areas

More than 50 % of the municipalities have less than 500 inhabitants, whereas **only 2 000 have more than 5 000 inhabitants**.

For most of the latter, the Insee has established a specific level for the dissemination of statistical results : the IRIS :

The living IRIS with a population between 1800 and 5 000 inhabitants,
The IRIS of activity with more than 1 000 employees, and twice as many employees than inhabitants
Other IRIS.

1 Km^2 and (200m)^2 of grid are other sub municpal areas used by the Insee.

The French spatial reference system

The Insee maintains the nonemclature for the administrative levels and its own levels.

The Insee maintains the nomenclature and the boundaries of the grids.

The French National Mapping Agency (IGN) is responsible for the boundaries of the administrative levels

The Insee and the IGN are working together for the boundaries of the IRIS.

POSSIBILITIES FOR GEOREFERENCING : 4 METHODS

The French statistical system is currently not a point based system, which needs methods to integrate sub municipal spatial information and statistical information.

1) The enumeration areas, assigned to the various enumerators, are supposed to be compliant with the IRIS.
2) Any other file, that has the *address* in the set of its variables, the integration of the coordinates is done by matching with an *address register*
3) The French tax administration also manages the cadastra, **making the georeferencing of any tax file much easier.**
4) The *address* is not in the set of the census variables. The integration of coordinates can be done but is difficult.

Dissemination of Sensitive Tax Variables in a 200m x 200m Grid Dataset :

The tax files are georeferenced but also a comprehensive statistical source on dwellings, households, individuals and incomes.

1) **Any tax variable, but the number of individuals, is a sensitive variable**
2) 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**.

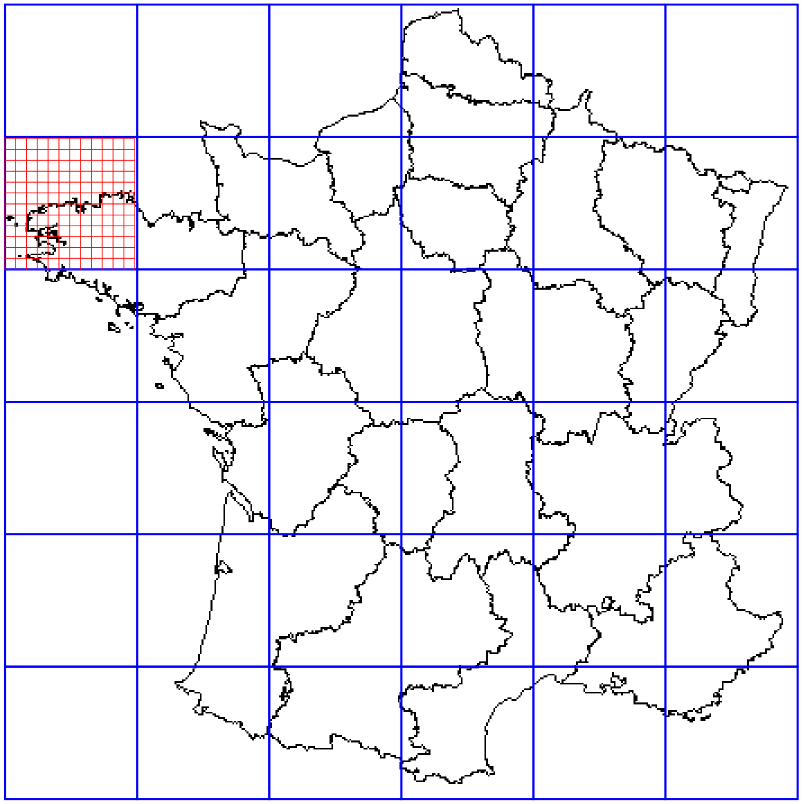
To comply with those rules, the Insee has established a 3-stage methodology

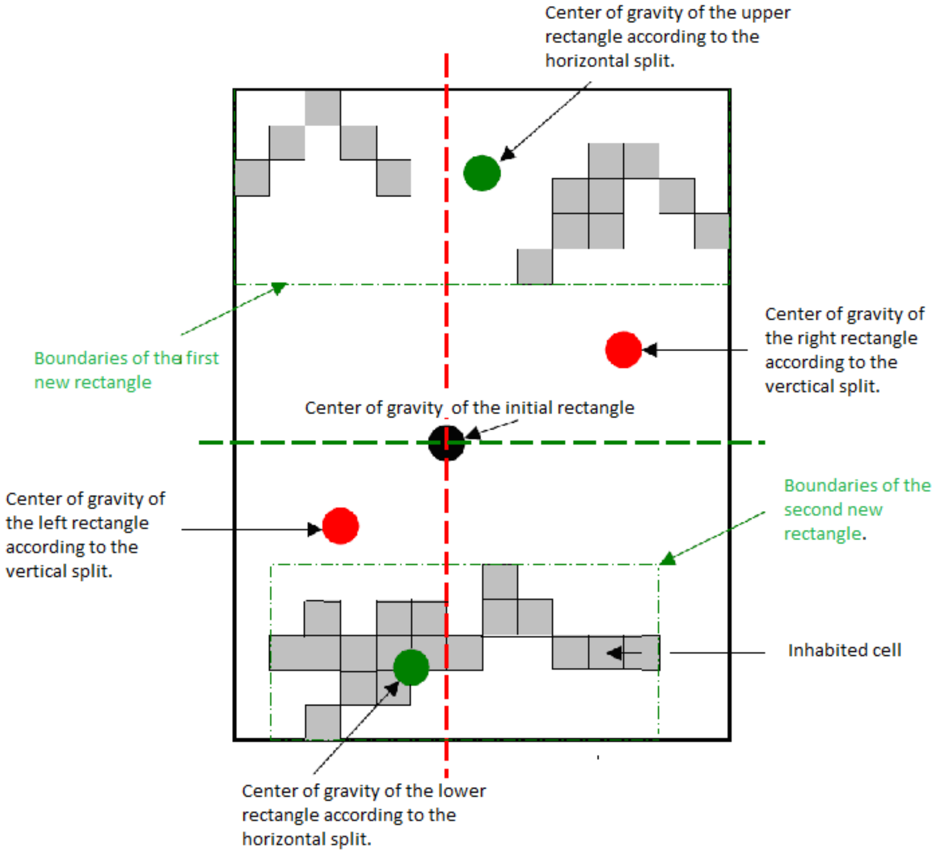
1) The low number cells were grouped in rectangles with more than 10 households,
2) A certain number of variables considered as being “at risk” were processed to avoid any risk of breach of confidentiality.
3) The differencing problem had to be tackled.

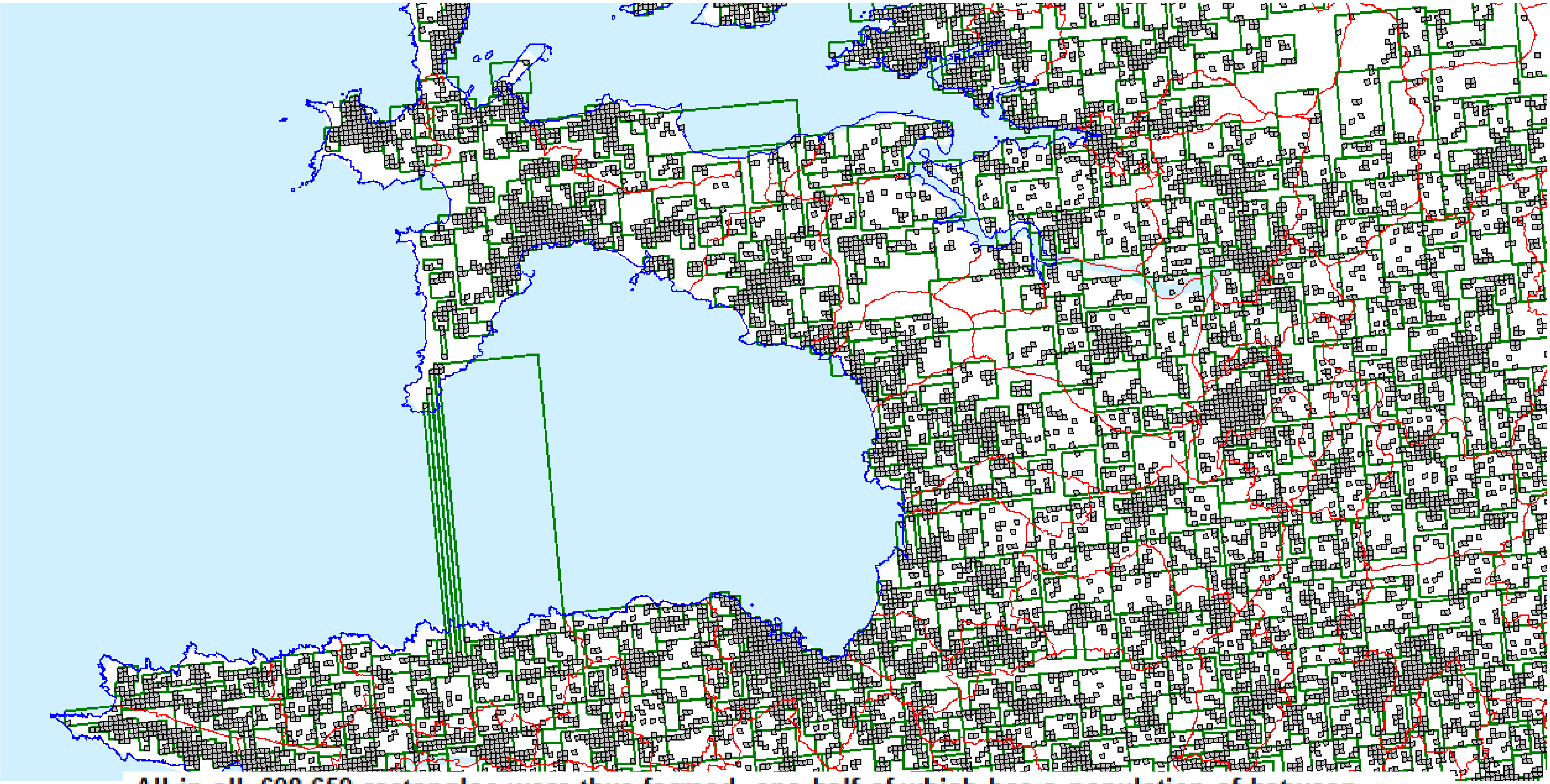
Aggregation of the low number cells

Because of computing capacity issues, France is split up into 36 equal size squares made up of 200 m x 200 m cells.

Each large square is cut horizontally or vertically to form 2 rectangles. These rectangles are then split horizontally or vertically, and so on




Center of gravity of the upper rectangle according to the horizontal split.
Center of gravity of the right rectangle according to the verctical split.
Boundaries of the first new rectangle
Center of gravity of the initial rectangle
Center of gravity of the left rectangle according to the vertical split.
Boundaries of the second new rectangle.
Inhabited cell
Center of gravity of the lower rectangle according to the horizontal split.



All in all, 698,659 rectangles were thus formed, one half of which has a population of between 11 and 21 households.

	numbers	% squares
1. Total inhabited cells in metropolitan France	2 278 213	
2. Cells with 11 households or more	462 413	20.3
Including :		
3. Cells with 11 households or more dissiminated as such	273 459	12.0
4. Cells with 11 households or more grouped with other cells in rectangles	188 954	8.3
5. Cells of less than 11 households grouped with other cells in rectangles	1 815 800	79.7
Total number of rectangles in metropolitan France (including 3.)	698 659	

Processing of variables being at risk

Winsorisation consists in moving the values above or below a given threshold to that threshold. The thresholds can be specific quantiles of the distribution.

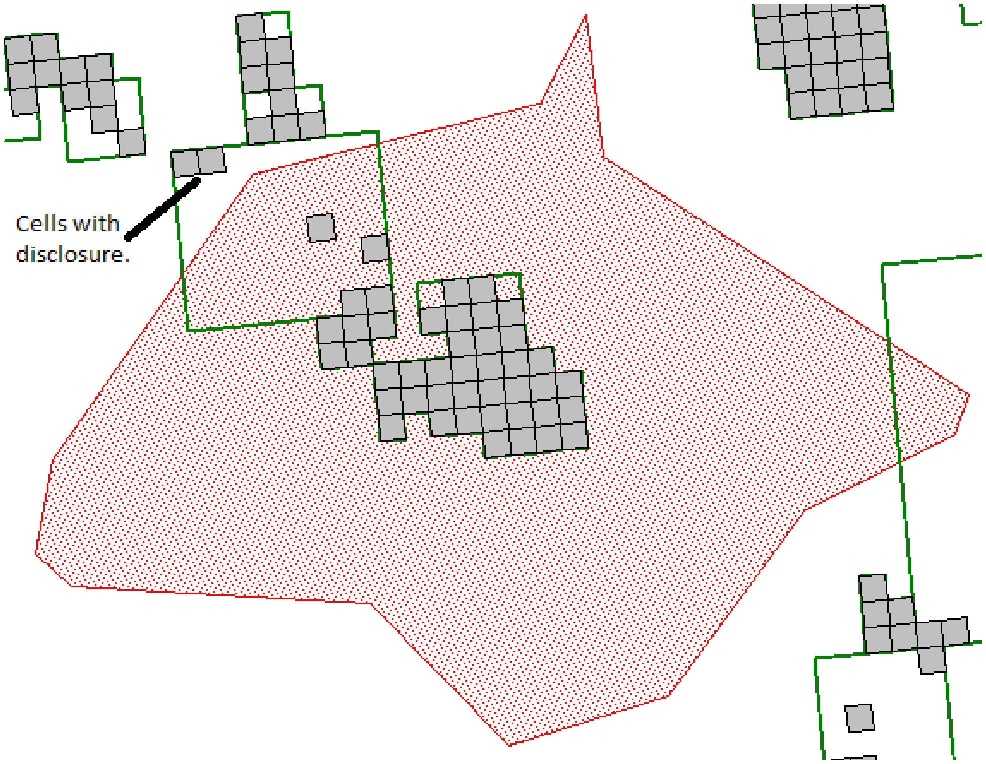
The taxable incomes of the households were previously winsorised for the distribution of the statistics in the rectangle.

The upper threshold is the 8th decile of the distribution : 30 000 euros
The lower threshold is 40 % of the median of the distribution : 7 500 euros

The following variables are also considered to be sensitive with respect to statistical confidentiality, and are therefore processed :

the number of people aged over 65
the number of households of just one person
· the number of households who are home owners

Tackling the differencing problem



To avoid any breach of confidentiality, a blank was used for the sum of the winsorised incomes of the smallest rectangle (by number of households)

Two files for the dissemination on the Insee website

- **A file for each inhabited cell :**
  - –The geographic identifiers of **the cell**
  - –The geographic identifiers of the rectangle to which the **cell** belongs
  - –The number of persons in the **cell**
- **A file for each inhabited rectangle**
  - –**Unprocessed variables**
    - Total number of persons, households, persons 0-3, 4-5, 6-10, 11-14, 15-17, 25 and over - years old,
  - –**Processed variables**
    - the sum of the winsorised taxable incomes
    - The total number of persons aged 65 and over, aged 75 and over, of households of one person, of home owning households,

Two files for the dissemination on the Insee website

The rectangles file is an intermediary file. It must not be used as such, in particular for mapping.

For the mapping of the total number of persons, the cells file can be used as such.

For the mapping of the other variables, a new file of cells must be built as from the two files delivered (rectangles with the variables and cells without the variables, by distributing the total numbers for each variable of the rectangle in each of its inhabited cells on a pro rata basis of the total population).

Challenges and conclusions

A very successful dissemination with the official bodies, the participative Websites, the community of geo statisticians, the citizens, the written or Televised press.

Dissemination in 700 000 cells or rectangles, whereas the previous dissemination was in a few thousands IRIS or municipalities.

To extend the experience to other sources, the Insee would need

- an authoritative address register, to be built with our NMA ?
- a reshaping of its system to go towards a point based system.
- to foster cooperation with our nma.

THANK
YOU!!