Dissemination of sensitive variables in a 200m x 200m national grid datasets

Loonis Vincent[1]

Insee, France

The GEOSTAT project's goal was to create a 1km^2 grid dataset of the 2011 censuses that were carried out by the various European countries. As a follow up to this successful project, the French National Statistics Office : National Institute of Statistics and Economic Studies (INSEE) decided to further explore building and dissemination of 200m x 200m national grid datasets from other statistical sources. Among them, the 2010 national property 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 paper is to present the solution adopted by INSEE to release those variables. The first part presents an overview of the French spatial reference system, while the second part deals with the methodology established by INSEE. The third part presents the use of the files and the precautions that are required to be taken in their use. The last part is made up of the conclusions and the challenges for the future.

1 The French spatial reference system

1.1 Nomenclature of territorial units for statistics

1.1.1 Levels including municipalities

France has a population of 65.8-million and is 650 000 km^2. It consists of 27 regions, 22 in metropolitan France (including the territorial collectivity of Corsica), and 5 overseas regions (Guadeloupe, Martinique in the West Indies, French Guyana in South America, The Reunion and Mayotte islands in the Indian Ocean).

In addition to the 27 regions, the French Republic has five overseas collectivities: (French Polynesia, Saint Barthélemy, Saint Martin, Saint Pierre and Miquelon, and Wallis and Futuna), one sui generis collectivity (New Caledonia), one overseas territory (French Southern and Antarctic Lands), and one island possession in the Pacific Ocean (Clipperton Island). These territories do not fall in the scope of this presentation.

Each region consists of departments, the metropolitan departments were created during the French Revolution in order to meet certain geographical criteria[2]. Currently, there are 101 departments. Each department consists of municipalities, which are successors of the parishes from the Middle Ages. Of all the European countries, France is the country with the highest number of municipalities: 36 600 or 40 % of the European municipalities are French.

Each level of this administrative division has its specific functions, which need statistics to be undertaken efficiently. As a result, each of these divisions are a structural component of the French nomenclature of territorial units for statistics.

In addition to this purely administrative territorial classification, other geographic classifications with a more functional focus have been introduced by INSEE. These are:

- *French agglomeration*–is made up of an aggregation of municipalities, which meet criteria linked to the continuity of built-up areas and to the number of inhabitants.

[1] The author would like to thank Martin Brady (Australian Bureau of Statistics) for his valuable comments which helped to improve the paper.

[2] The main town of a department should be accessible by horse in less than a day for any inhabitants of the department.

- *French urban area* – is a more economic approach of the city and is made of up of an aggregation 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.

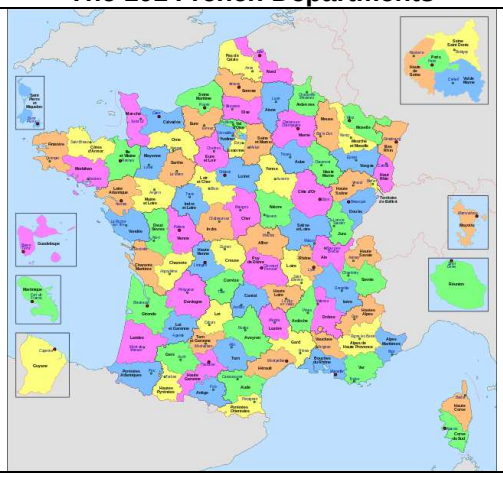
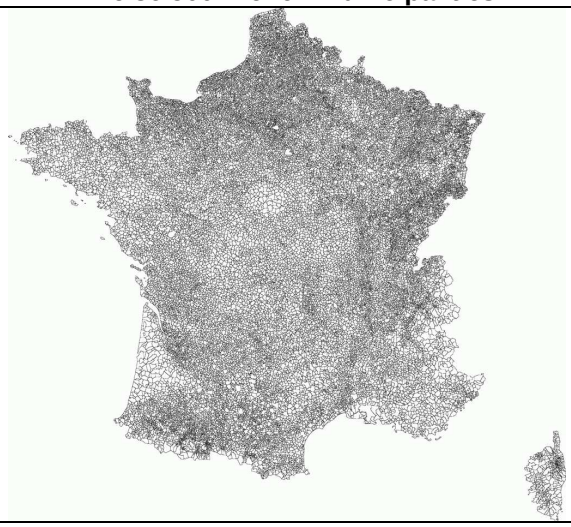
1.1.2 Municipal sub-areas

As a result of a historical process, municipalities differ greatly in terms of both population and size. More than 50% of them have less than 500 inhabitants, whereas only 2 000 have more than 5 000 inhabitants.

INSEE has therefore further divided the large municipalities, using specific criterion, for the dissemination of the statistical information, this geographic classification is called the IRIS (from the French Ilots Regroupés pour l'Information Statistique). For the municipalities with more than 10 000 inhabitants and most of those with more than 5 000, the IRIS were set up according to the results of the 1999 census, with areas classified into one of three categories :

- 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' containing the remaining areas, not included above.

For some years, INSEE has also been providing grid1 Km^2) or 200 m x 200 m (200 m^2) grid cells, established according to European Terrestrial Reference System (ETRS89) with a common origin and an unambiguous identity code for each grid cell.

The 27 French Regions	The 101 French Departments
	
The 36 500 French municipalities	
	

1.1.3 Boundaries

INSEE maintains the nomenclature of the administrative divisions of France and of its own statistical divisions. The French National Mapping Agency (the IGN) is responsible for all of the associated boundaries, including IRIS.

1.2 Capabilities for geocoding

The French statistical system is not a point based system. Unit records are not linked to location coordinates, they are coded to small geographic areas and aggregated to larger geographic units. The integration of spatial information and statistical information is easy for the municipalities and for the levels above, but more difficult for the municipal sub-areas. There are 3 main methods for the integration at these levels.

For municipalities under 10 000 inhabitants, the census data collection is still a traditional one. The census districts, assigned to the various enumerators, are supposed to be aligned to the IRIS boundaries. Within municipalities over 10 000 inhabitants, data collection involves a sample of the population (8 %) and standard results are compiled from accumulating the sample over five years. The sampling frame, from which the addresses to be surveyed are drawn, contains the IRIS as information. This IRIS information is used to apply the relevant IRIS code to each unit record (i.e. geocoding each unit record with an IRIS code). As a result, the integration of the IRIS at the very beginning of the statistical process enables dissemination at the infra-municipal level at the end of the process, without using the coordinates.

For any other files, that have a "*location street address*" in the set of unit record variables, the integration of the coordinates is done by linking the file with an address register. Currently, there is no authoritative national address register in France. INSEE built its own register by compiling various sources, including the national land register.

In France, each resident is either a tax reference person (i.e. Family Head, Main Income Earner, Property Owner)), or linked to a tax reference person in a tax household (i.e. Spouse, Children, Dependants). Each tax household must present a tax return and declare his or her place of residence, whether it is taxable or not. A statistical household is therefore formed, by INSEE, by grouping of tax households declaring their place of residence in a single dwelling. The tax files are thus a comprehensive source for studies of population, dwellings, households and incomes. Moreover, the French tax administration also manages the cadastre (i.e. property boundaries). Therefore, the various statistical units in the tax files can be linked together and *easily* geocoded.

The quality of the location information in the tax files for Guadeloupe, Guiana and Mayotte make it impossible to reconstitute the statistical households, as occurs for metropolitan France, the Island of Réunion and Martinique. Therefore, no income statistics are produced for these three Overseas Regions.

2 Dissemination of 200 m^2 grid tax datasets

The rules governing the dissemination of the tax files, do not allow the dissemination of statistical information *(except for the total number of persons*) on geographical levels, or for cells of a table, that have less than 11 statistical households contributing to the statistic. To comply with this rule, INSEE has established a 3-stage methodology for enabling the dissemination of a 200 m^2 grid tax dataset:

- The low number grid cells are grouped together into larger area rectangles (or squares – generally referred to as rectangles from this point) so that each has at least 11 statistical households.
- Certain variables considered as havening "high sensitivity" have undergone additional processing to avoid any risk of breach of privacy.
  Release of the new grid tax dataset was identified as creating risks to privacy as a result of the "differencing problem"; some data was suppressed to manage this issue.

2.1 Constitution of the rectangles

To maintain confidentiality within the statistics published, in particular the taxable income statistics, INSEE implemented the following process. Each building containing the households and individuals in the tax files were assigned to the standard European grid of 200 m^2 cells.

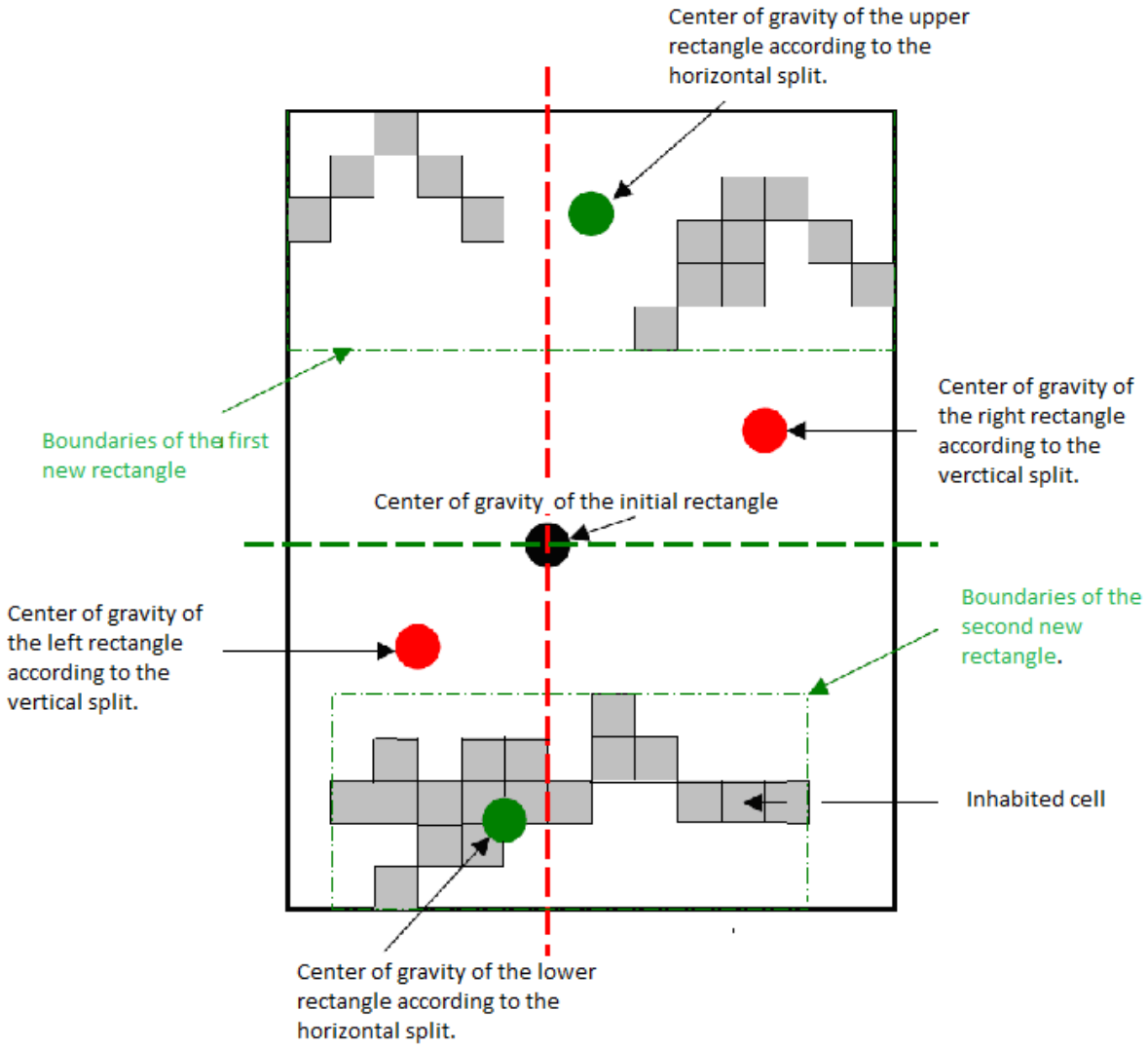
In addition, the metropolitan territory was also divided into 36 squares of similar size, which were made up of the 200 m^2 cells. Each large square was then cut horizontally or vertically to form 2 new rectangles, according to specified rules. These new rectangles were then split horizontally or vertically, and so one. At each step, the location of the splits and the choice between a horizontal or vertical split, or the absence of any split, was made as follows:

1. The location of the two horizontal and vertical splits was determined by the centre of gravity of the populated 200 m^2 grid cells contained within the rectangle being split (see diagram x below). The centre of gravity is a mean centre weighted by population. This split along the centre of gravity was done to assist in maintaining areas of approximately equal population size.
2. Within each of these split areas a 'new rectangle' was defined. This was defined by a rectangle that contained only the populated 200m grid cells with the split area of the previous rectangle (see diagram x below).
3. If the two horizontal and vertical splits each produced a 'new rectangle' of less than 11 households, the split was not made. This was done to ensure that statistical confidentiality requirements were maintained.
4. If only one of the two splits resulted in a breakdown where at least one new of the 'new rectangles' contained less than 11 households, then the other split was performed. Again, this ensured confidentiality was maintained.
5. If the two splits each produce two 'new rectangles' of over 11 households, the choice between the horizontal and vertical split was based on the split which produced two 'new rectangles' where the populated grid cells showed the least geographic spread. The geographic spread of a 'new rectangle' was measured by the square of the sum of the distance between the centre of gravity of the new rectangle and the populated grid cells within it, weighted by the population (see diagram x below). This decision rule minimised the spread of the two new rectangles created, which helped to maintain a compact shape for the rectangles.

<u>Three important comments</u>:

- The procedure for splitting rectangles does not take into account the continuity of the built-up area and can "separate" spatially and socio-economically related populations, just as it can combine together populations from different areas (e.g. on each side of a mountain, across the sea, etc.).
- The final rectangles created are of varying sizes; ranging from a single 200 m^2 grid cell (of at least 11 households) to a rectangle with an area equal to 3,000 grid cells. In metropolitan France, the rectangles produced include 38 households on average.
- Direct mapping of the numbers in the rectangles is definitely not to be recommended, due to the wide variety of the areas of the rectangles. The statistics per rectangle should be disaggregated (using statistical allocation methods) into the inhabited 200 m^2 grid cells with each rectangles and transformed into densities or proportions of the population.

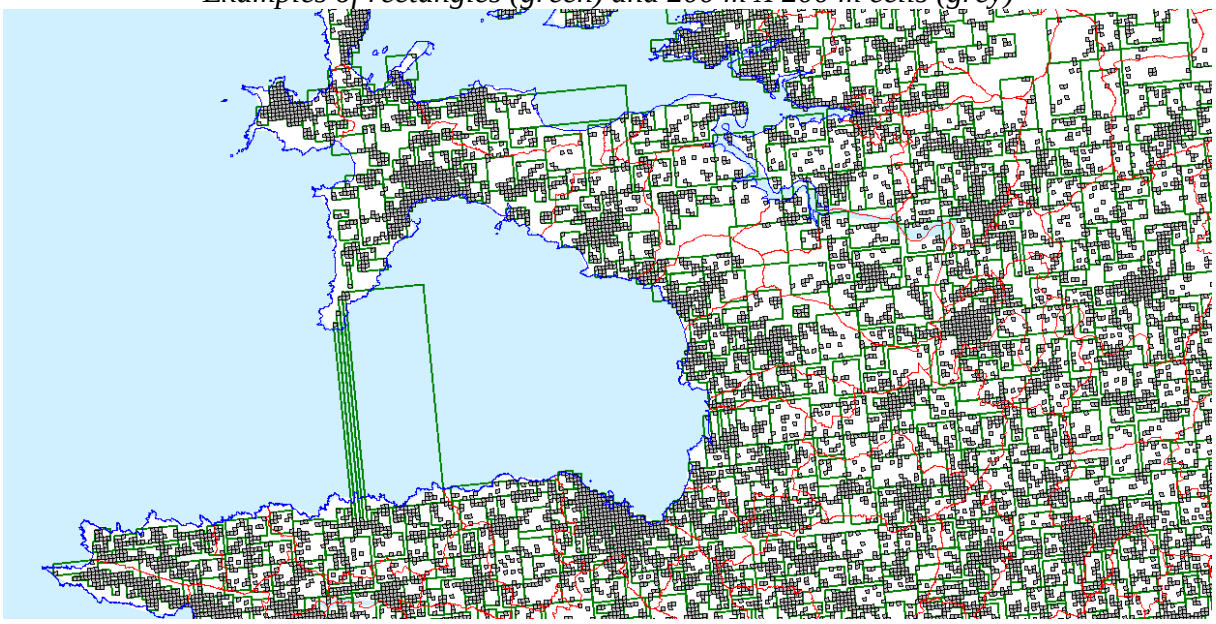
Example of the choice between a horizontal and a vertical split:



All-in-all, 698 659 rectangles were formed from this process. One-half of these had 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 distributed 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	

Examples of rectangles (green) and 200 m X 200 m cells (grey)



2.2 Processing of the rectangle statistics

2.2.1 Prior winsorisation of tax incomes per consumer unit

Winsorisation is a statistical technique for the reprocessing of extreme values of a distribution. It consists of moving the values above or below a given threshold to that threshold. The thresholds can be specific quantiles within the distribution.

INSEE has exhaustive files of income tax returns and of residence taxes, from which the localised tax incomes are calculated. These incomes were winsorised for the distribution of the statistics in the rectangle. More precisely, two types of winsorisation are performed.

For metropolitan France, two windsorisation methods were applied:

- A “fixed winsorisation” of the taxable income of households per consumption unit (c.u). If the tax income per consumption unit of the household is above the 8th decile of the distribution, i.e. 29,336 euros, the taxable income the household per consumption unit is then lowered to that threshold. Inversely, if the taxable income of the household per consumption unit is 40% below the median of the distribution of income, i.e. 7,499.60 euros, the taxable income of the household is raised up to that threshold.

- A “floating winsorisation” of the taxable income of households per consumption unit. If the taxable income of the household is above the 8th decile of the distribution, the taxable income of that household is then lowered to a random value[3] of between 28,836 euros and 29,336 euros. Inversely, if the tax income of the household per consumption unit is 40% below the median of the distribution, the income of that household per consumption unit is raised to a random value of between 7,499.00 euros and 7 999.00 euros.

Fixed winsorisation is used in the vast majority of cases. Floating winsorisation is used only in a few, very specific cases, detailed in §2.2.2.

For Martinique and the Island of Réunion, the departmental 8th deciles have been used, i.e. 25,098 and 23,131 euros respectively, to take account of the differences in income with metropolitan France. Similarly, incomes corresponding to 40% of the median in Martinique and the Island of Réunion have been used, i.e. 5,209.20 euros and 4,136 euros respectively. Floating winsorisation in these two

[3] Taken according to uniform rules

overseas departments also consists of calculating a random income within a window of 500 euros below the 8th departmental decile and above 40% of the median of the departmental distribution.

One should note that, because of lesser quality matching within the tax files for Guiana, Guadeloupe and Mayotte, it is not possible to calculate the taxable income of households in those three overseas departments.

2.2.2 Processing of the rectangle sum of the winsorised income of persons

After fixed windsorisation had been applied, the sum of the income per c.u.[3] of the persons is calculated for each rectangle.
When it is possible to deduce from that sum, that all the persons have an income per c.u. above the 8th decile, floating winsorisation is applied to the original data for that rectangle. This was the case for 66 rectangles in metropolitan France, for 1 in Martinique and for 9 in the Island of Réunion.

Similarly, floating winsorisation is used when it can be deduced from that sum that all the persons in a rectangle have an income 40% below the median or close to that value. This was the case for 3 rectangles in metropolitan France, but none in Martinique or in the Island of Réunion.

2.2.3 Processing of other sensitive variables

The following variable categories are considered to be sensitive with respect to statistical confidentiality:

- the number of people aged over 65
- the number of households of just one person
- the number of households who are home owners
- the number of households where the tax income per c.u. is below the “low income threshold”, equal to 60% of the median of the distribution of incomes per c.u. (i.e. 11,249.40 euros in metropolitan France, 7,813.80 euros in Martinique and 6,204 euros in the Island of Réunion).

The data released for the numbers of persons or households in these specific categories must not be greater than or equal to 80% of the overall total for this variable. Additionally, the number of home-owning households must not correspond to a proportion that is less than or equal to 20%. Where these threshold proportions were exceeded, it was not possible to release the actual number. The data was suppressed and is indicated to the user as follows: each of these sensitive variables in the file distributed carries an additional indicative variable (or field) with a value of 1 if the number “n” as distributed is to be understood as “n or more” and otherwise 0.

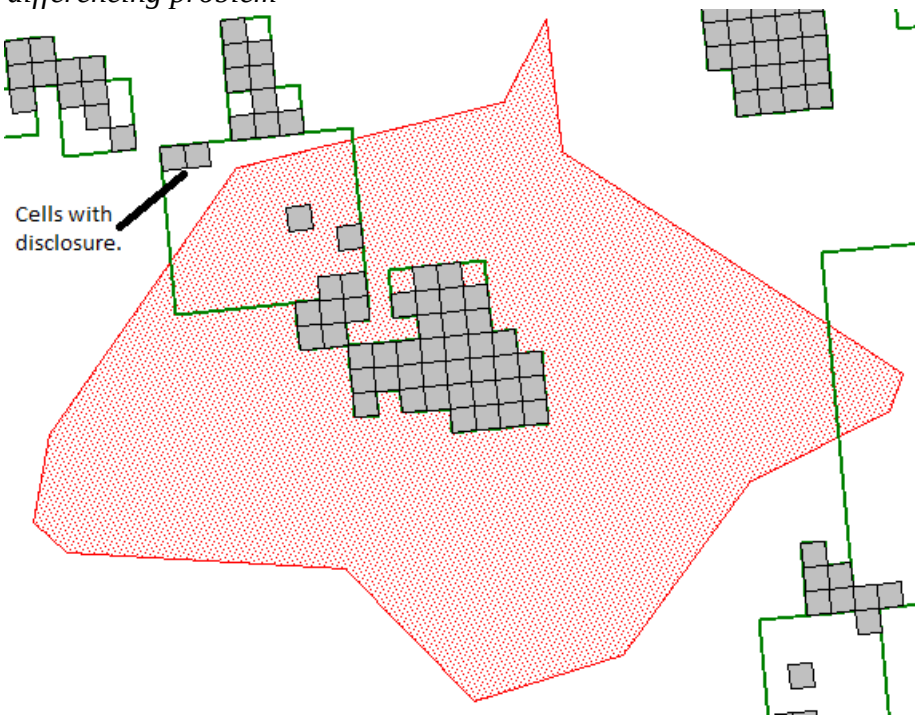
In the particular case of home-owning households, the indicative variable takes the value of 2, if the number “n” of the home-owning households distributed is to be understood as “n or less”.

2.3 Dealing with the “differencing problem”

INSEE had also released 2010 taxable income data at the municipal level on its Web site. This created the possibility of a breach of confidentiality by comparing two similar geographic areas with very small coverage differences using the taxable income data for the municipal levels and the data produced for the rectangles: this is the differencing problem, see diagram below.

Where this differencing problem occurred, data was suppressed for the smallest rectangle (by number of households). In the case of the data for the sum of the winsorised incomes, a blank was used in rectangles in 153 towns in metropolitan France, one town in Martinique and of towns in the Island of Réunion. Those rectangles represented a total of 1,843 households.

Example of differencing problem



The total income, be it winsorized or not, is released on the web for the municipality (in red) and for the rectangles (in green), producing overlapping geographic coverage. In addition, the total number of persons is released for each cell (in grey) constituting the various rectangles. The indicated cells can be have their data differenced through subtracting the data obtained for the municipality from the rectangles covering the same area.

In cases where the number of people aged 65 is processed as above, it may prove necessary, for the sake of coherence between the data distributed, to indicate that the number “n” of persons aged 75 and over is to be understood as “n or more”. The corresponding indicative variable then has a value of 1.

3 Use of the files and precautions to be taken

INSEE disseminated on its website a cells file and a file of rectangles.
Variables distributed in the cells file included the following.

For each inhabited cell:

- The geographic identifiers of **the cell**
- The number of inhabited squares of the **rectangle** to which the square belongs
- The geographic identifiers of the rectangle to which the **cell** belongs
- The number of persons in the **cell**

Unprocessed variables distributed in the rectangles file:

- The total number of persons
- The total number of households (i.e. main homes)
- The total number of persons 0 to 3 years old
- The total number of persons 4 to 5 years old
- The total number of persons 6 to 10 years old
- The total number of persons 11 to 14 years old
- The total number of persons 15 to 17 years old
- The total number of persons aged 25 and over
- The total number of households of 5 persons and over
- The total number of households who have been living in their current home for 5 years or more

- The total number of households in multifamily housing
- The total area of the main homes in square metres
- The sum of the winsorised taxable incomes per c.u. of the persons (the tax income per c.u. of a person is the income per c.i. of the household to which the person belong)[4]

Variables processed and distributed in the rectangles file:

- The total number of persons aged 65 and over + the corresponding indicative variable
- The total number of persons aged 75 and over + the corresponding indicative variable
- The total number of households of one person + the corresponding indicative variable
- The total number of home-owning households + the corresponding indicative variable
- The total number of households whose tax income per c.u. is below the low income threshold (60% of the median of the distribution) + the corresponding indicative variable

3.1 Use of the files

The rectangles file is an intermediary file. It must not be used as such, in particular for mapping. The areas of the rectangles are variable, and do direct mapping of raw numbers would be erroneous. Furthermore, for the relatively large rectangles, the overall value may mask considerable internal spatial variation. Any analysis work should be undertaken at the level of the squares.

For mapping of the total number of persons, the cells file can be used as supplied as the size of the inhabited cells are the same.

For mapping of the other variables, data for the cells must be built from the two files delivered (rectangles with the variables and cells without the variables). Data for the cells can be derived by distributing the total numbers for the variable with each rectangle into each of its inhabited squares on a pro rata basis of the total population.

Thus, for a rectangle composed of n inhabited squares:

Let:

$Ptot_r$	the total population of the rectangle
$Ptot_{c1}…Ptot_{cn}$,	the total population of each inhabited square of the rectangle
V_r,	the count in the rectangle for the variable V

This will then give:

V_{c1} = count for the cell c1 for the variable V

$= (Ptot_{c1} \times V_r) / Ptot_r$

Fictitious example:

Total population

10		18	
	4	2	

34 persons in the rectangle

Children 0 to 3 years old

12

12 children 0 to 3 years old in the rectangle

How many children 0 to 3 years old are there in each of the inhabited squares of the rectangle?

[4] This sum must be divided by the total number of persons in the rectangle to obtain the mean tax income per c.u. of the persons.

Vc1		Vc2	
	Vc3	VC4	

Vc1 = (10x12)/34 = 3,53 Vc2 = (18x12)/34=6,35

Vc3 = (4x12))34=1,41 Vc4 = (2x12)/34 = 0,71

Children 0 to 3 years old – distributed into inhabited cells by population

4		6	
	1	1	

3.2 Precautions to be taken

In the file of cells, all the cells are distributed without any threshold on the number of persons or of households. This requires a high degree of **caution, in particular when it comes to analysing low density zones**. Half of the squares of metropolitan France contain at least 2 households or less and half contain 6 persons or less.
In urban areas, given the high densities, the data is considered to be reliable over a relatively small set of squares. However, in rural areas, work at any level below the size of a township is not recommended and smoothing should be applied to the data.

Any timeseries comparison would not be valid. No analysis should therefore be made regarding any timeseries between the population data resulting for the 2009 LTI, previously distributed, and the 2010 LTI data presented in the current release.

Under no circumstances must a comparison be made between the sum of the population of the squares making up a town and the population of the town at the time of the population census. Those two numbers will necessarily be different, due to the difference in source (see above, point 1) and the processing of the variables.

Given the winsorisation procedure of the incomes, **the sum of the incomes per c.u. of the persons in the rectangles of a town will be different from the sum of the incomes per c.u. of the persons as shown on insee.fr for the whole town in the municipalities dataset** (the difference can be as great as 40%).

4 Conclusion and challenges for the future

Despite its restrictions, the dissemination of 200 m x 200m grid tax dataset was very successful among official bodies, the geo-statistics community, citizens and participative websites. , For instance, the results were used by the Ministry for Urban Areas, Regional Planning and the Environment to create functional areas eligible for assistance offered in the context of urban policy.

The solution adopted by INSEE is a compromise between the theoretical benefits and shortcomings of grid statistics as stated by Eurostat[5] during the first meeting of the UN-GGIM expert group. On the one hand, the rectangles have a suitable size, corresponding to the extent of the phenomena, and avoid the dilution effect. They can also be joined to create virtually any functional output area for statistics, based on objective criteria. The method applied also manages the disclosure issues and rules correctly. On the other hand, the rectangles are not stable over time. They also do not have the same size, complicating the comparisons over time and space.

[5] ESA/STAT/AC.279/P6

Despite these issues, INSEE expects to further explore geolocalization and grid datasets. It will then have to face the following two challenges :

- Firstly, establish an authoritative address register, preferably with the IGN, and extend the geo-referencing to other sources, such as business register
- In the end, establish a point based statistical system built on registers.