Grided data from the French census 2007

Aggregation without coordinates, coordinates but disaggregation

JL.LIPATZ 23/11/2011

For a long time, the French census has been organized in different ways according to the size of the municipalities or their belonging to the European continent of not. Overseas territories are part of France and for some of them share the same administrative organization than the European French territories implying the same statistical obligations for the INSEE. Nevertheless these overseas territories are not in the scope of the Geostat project so their case will not be described here: let just say that what follows doesn't apply to them.
For the European part of France one unique distinction applies: according to the 10 000 inhabitant threshold two different methods must be used to produce grided data from the census.

Challenge 1 – Municipalities below 10 000 inhabitants

The standard process

Within municipalities below 10 000 inhabitants, the census data collection is still a traditional one. But it now obeys to the standard feature of a rolling census whose data collection occur every year for a part of the territory in order to produce results aggregating the data collection of five years running. For smaller municipalities, this means that data collection takes place during only one year for a given municipality, but two different municipalities will generally have different date of data collection. For producing results and some of them have legal effects, this difficulty is overcome by a correction applied to census results that takes into account the evolution of the number of dwellings as measured in fiscal sources.

Making spatial data with the census data collection for small municipalities looks like an impossible mission because of the traditional field organization: data processing uses census districts as the smallest geographical unit and as each municipality is responsible for its data collection; The tools used to define the census districts vary from sophisticated GIS to paper plans or just textual definitions for municipalities under 500 inhabitants (there are a lot! Moe than 20 000). So, 35 000 is a good approximate for the number of practical situations : that makes apparently impossible to get a better solution than disaggregation of municipal population counts.

The process applied for grid cells

For a long time, the only perspective for localizing census data was keyboarding address texts that are collected together with the census variables by the census enumerators. Unfortunately, it would imply a cost that is too high in the context of the Geostat project and would cause additional difficulties to translate these texts into geographical coordinates.
Trying to solve these later difficulties for other data sources where addresses were already keyboarded, INSEE has been developing in parallel the use of the fiscal data, together with cadastral data. This path gave the opportunity to draw the first grided fiscal population map at the beginning of Summer 2010 that still contained a part of estimations. At the beginning of Summer 2011 these estimation were eventually made no more necessary.
Beyond the huge progress that is being able to exploit the entire fiscal data at any geographical level, the availability of a data set containing fiscal dwellings together with their

geographical coordinates offers an opportunity of geolocalizing the census data too. Individual match between census data and fiscal data based on some kind of identifier is not allowed either for legal or for technical reason. But the two data sets contain information describing the same people living in the same physical buildings and the improvement of an address register is fully allowed. Common information is not a lot: essentially the municipality of the residence, the date of birth and the place of birth (with some restrictions), but for each person in the census data set it is possible to find a limited number of possible locations taken from the fiscal source and to choose the most likely one. This was done the following way:

Step 1: Match the census data set and the fiscal data set (at same date) at individual level, using the municipality of residence, the date and place of birth. The match with the place of birth is allowed to be fuzzy: it can be no match at all or just a match on the d*épartment* (NUTS3).

In a lot of cases there will be several possible matches. Each of one is given a score according to the quality of the match of the place of birth and according to the characteristics of the residence as known in the fiscal data (principal/secondary…).

Step 2: For each building in the census data set, aggregate the individual scores at cadastral parcel and street level as they are provided by the fiscal data set.

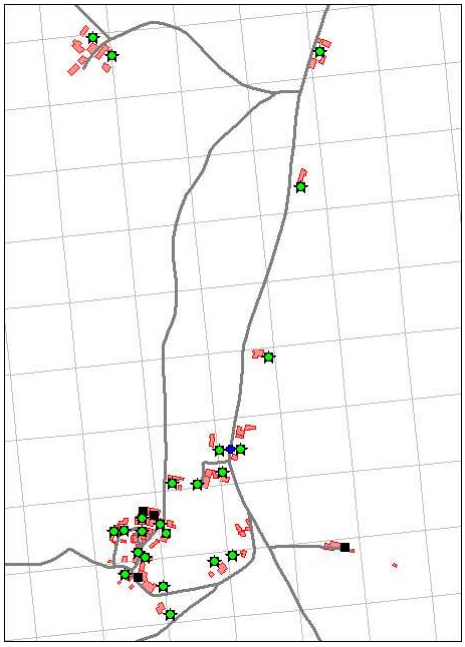
Step 3: Choose the best possibility. High scores provide better candidates, but this is not enough to decide between two equal score values and is sometimes misleading. An important feature of the census is used to control the selection: a census enumerator doesn't collect his census district in a pure random way. He has to search all over the district in some systematic way to be sure he is not forgetting some place. So records that are close in the data set should be close geographically too. This rule is applied to get the resulting census data set with locations as clustered as possible. In some cases it invalidates a match even if there is no other alternative match.

Step 4: Interpolate the unmatched census buildings. Again the proximity rule is used. An unmatched building that is between two that are matched and are located in the same street is very likely to be physically between the later two. But in some cases more rough location are computed: location of the neighbouring record in the census dataset or an average position. But these later cases are rather rare and actually more frequent in very small municipalities where the generated absolute error will be negligible. Actually, aggregation at the building level provides the conclusive trick because only one person matched locates the whole building he lives in. So the overall results are generally excellent. Most of the problems occur because the collected data for the census is incorrect or missing.

Match at building level	97.4 %
Located using interpolation	2 %
Located using approximate location	0.6 %

Département de la Vienne
Results of geolocalizing the census.

Step 5: Convert coordinates to the LAEA projection system, truncate coordinates to multiple of 1000m and aggregate the individuals into population for each grid cell of each municipality



Example 1: Hamlet of Lombardie, small part from the municipality of Curzay sur Vonne (65 inhabitants)
Census dwellings are green and blue (when interpolated) spots. Fiscal locations are the black squares, mostly under the coloured spots. Grid cells are 100 m wide; the red building shapes come from the cadastral register.
Almost every one is correctly located, including the writer of this paper whose place of birth was not keyboarded correctly and a student who is recorded in different municipalities in the two data sources

Note: The previous process applies to every municipality below 10 000 inhabitants except two ones: *Ile de Sein* et *Ile Molène*. These ones are not covered by fiscal data for historical reasons: they were exempted from the housing tax because they are small islands in Brittany difficult to access. Their grid cells are left blank in the resulting data set. Both of them have a population of about 200 inhabitants.

The time issue

When making results for the year 2007, five years of data collection are actually used : from 2005 to 2009 in order to gather every municipality. Associated fiscal data comes from the same reference year to improve the match and the overall result will keep on being heterogeneous relatively to the time question. But it is the same in the standard results of the French census which fortunately contain very few things at sub city level for small municipalities. It means that any census based population estimation is unable to estimate correctly a sub municipality area built in 2006 for a municipality collected in 2005 and will wrongly take into account an area built in 2008 for a municipality collected in 2009.

Challenge 2 – Municipalities above 10 000 inhabitants

The standard process

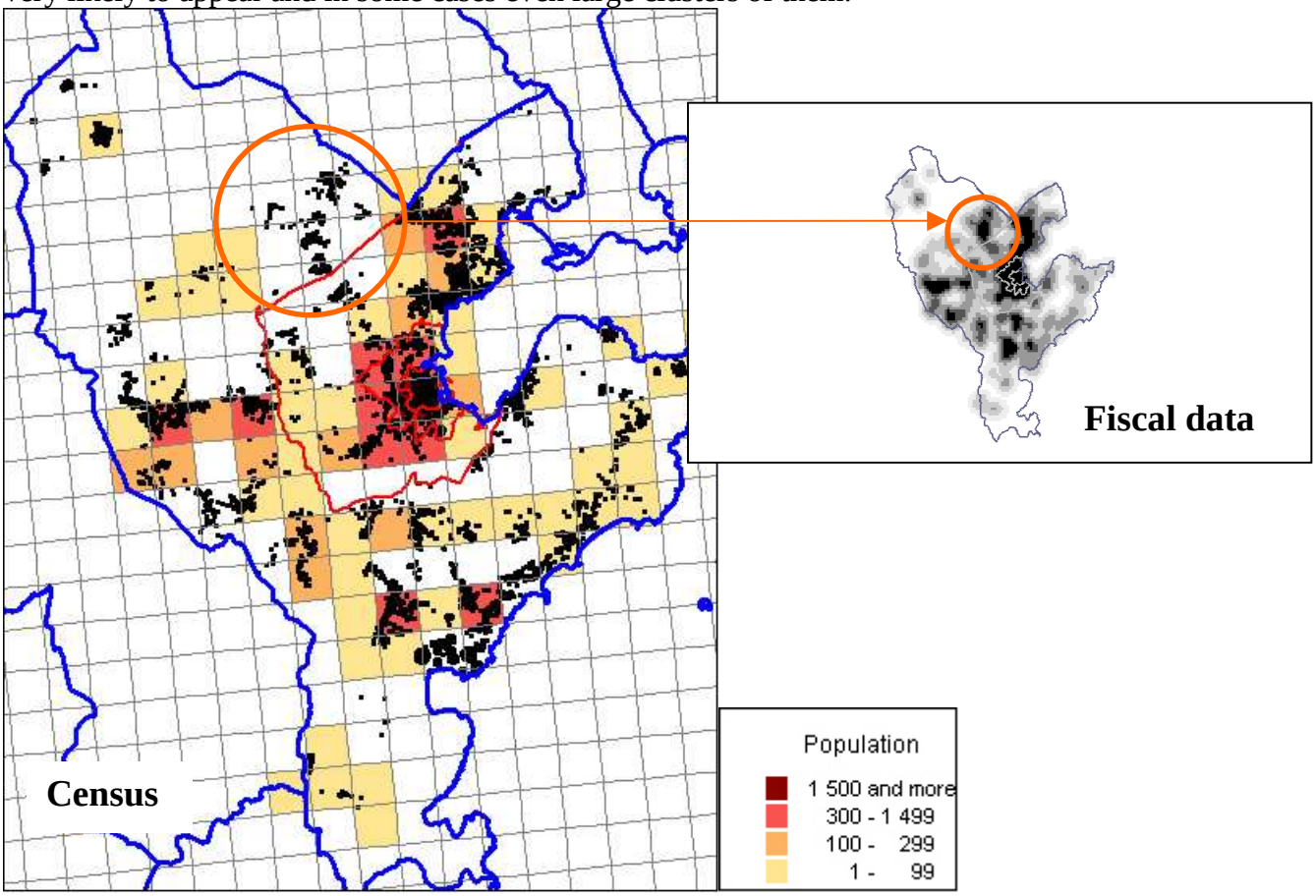
Within municipalities above 10 000 inhabitants, INSEE is fully responsible for the physical organization of the data collection. This unifies the geographical infrastructures that are used, but it is now necessary to tackle two difficulties: data collection occurs every year but involves a sample of the population (roughly 8 %) and standard results are made gathering the collection of five years running.

There are two kinds of population: people living in ordinary dwellings and people living within communities. For the later kind there is no special difficulty for estimations because they don't obey to a sample. The question of coping with the sampling only applies to ordinary dwellings.

The basis for this sampling is an address register maintained by the INSEE itself that doesn't contain so much information: only a number of apartments and spatial coordinates. In the standard processing of the census data collection, the first information is used to correct the result of the data collection. The final number of apartments is forced to fit to the one contained in the register for any zone starting from the smallest standard output areas used for census dissemination (IRIS).
The coordinates that are in the address register are written back in the resulting census file together with the weights resulting both from the sampling and from the correction by the number of apartments in the register. These variables could be used to estimate any quantity from the census provided that the area where the estimation is done is larger enough to contain enough collected addresses to guarantee the accuracy of the computation.

The process applied for grid cells

Unfortunately, the design of the sample used in data collection only guarantees an accurate computation for areas larger than the standard output area (IRIS) and 1km wide grid cells are smaller in most of the cases. Grid cells where no data collection has occurred are very likely to appear and in some cases even large clusters of them.



Example 2a: Municipality of Porte Vecchio (Corsica).
Large map: blue boundaries are those of municipalities, red ones those of standard output areas (IRIS). Black dots are known addresses in the census register and 1km wide grid cells were painted according to the "naïve" census estimation (sum of the number of people, weighted by the corrected sampling weight).
A large zone (orange) is not covered by the data collection and thus estimated to null. That goes against what says the fiscal data for population distribution (small map : population density estimate using kernel methods in order to highlight the most populated places).

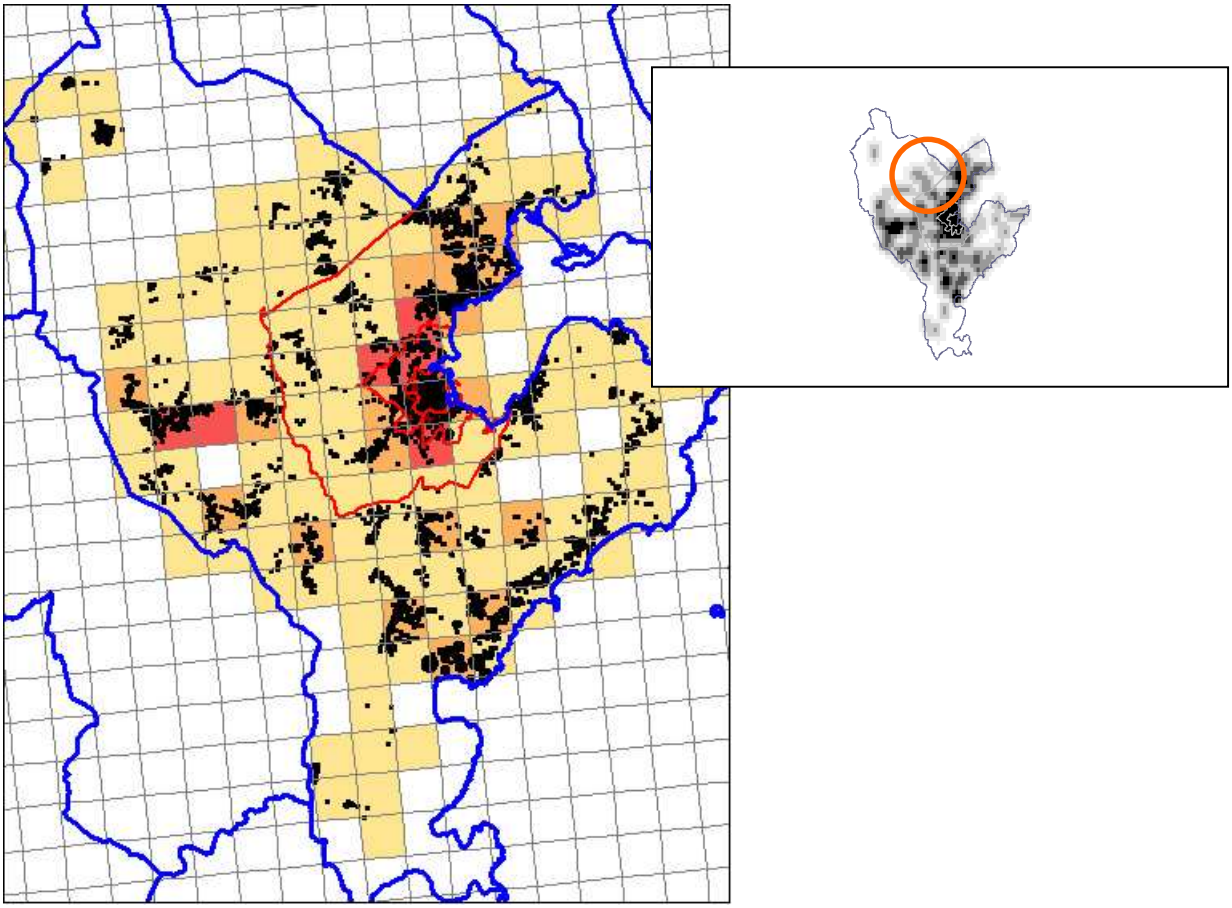
To produce correct results for the 1km wide grid a special process has to be applied that gives a more important part to space. Basically the idea is to go through each populated place as know in the census address register and to estimate census variables looking at what happens in the physical neighbourhood of the address. In the general case, INSEE uses an auxiliary data source other than the address register to strengthen the estimation of specific variables. In the case of simple population counts this unnecessary and census data was used alone, for each municipality in turn, in the way described here. That method simplifies the answer to the questions of consistency with census result produced by the standard non-grided process.

Step 1: for each address contained in the census address register look at the 50 addresses that are physically nearer (in the sense of the simple Cartesian distance between the two points) and were collected by the census. Compute a mean number of persons by apartment for this set and applies it to the number of apartments of the target address to get an estimated number of persons.

Step 2: the estimated number of persons per address gives an opportunity to compute a population for each existing zoning, starting from the standard output areas and even for the municipality itself. It is very likely that the produced figure will differ from the one that is produced using the standard non-grided process for which a steady dissemination takes place. To avoid any problem of consistency between the two approaches a correction is applied to every figure computed at step 1. The sum of the estimates for each standard output area is forced to fit the figure obtained with the standard non grided process (sum of the number of people for the collected address only, weighted by the corrected sampling weights).

Step 3: Previous steps only apply to ordinary dwellings. People living in communities are simply added to the previous estimates. At the end of the step a data set is available that contains population count for each populated place: address of ordinary apartments or houses and locations of communities.

Step 4: Convert coordinates to the LAEA projection system, truncate coordinates to multiple of 1000m and aggregate the population counts of the populated places into population for each grid cell of each municipality.



Example 2b : Municipality of Porte Vecchio (Corsica).
Grid cells were painted according to the new population estimation for ordinary dwellings. The small map was obtained with the same computation as for the fiscal data before. The two small maps don't perfectly match but several explanations could be put forward: fiscal data refers to 2009, census estimation refers to 2007 and more over location of people in the two sources may differ for taxation reasons especially in region with a high rate of residences for holidays.

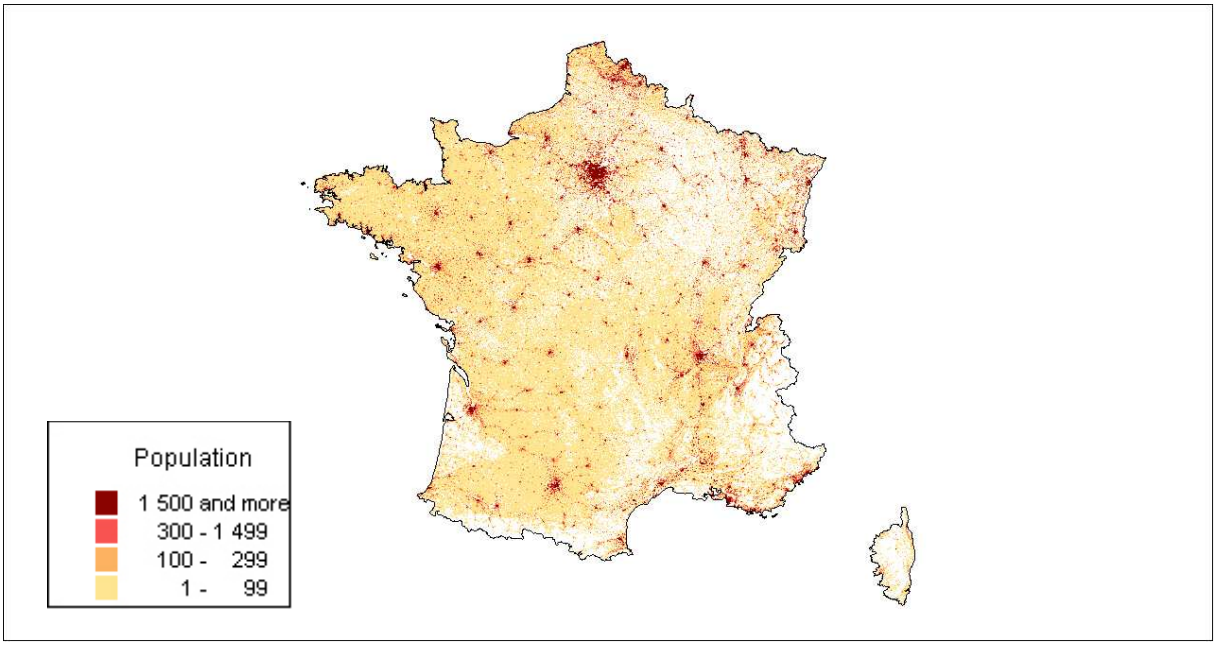
The time issue

When making results for the year 2007, five years of data collection are actually used : from 2005 to 2009, because the data collection of only one year is too small to ensure correct accuracy. The time issue is implicitly handled in step 1 by selecting the address register relative to the target year for the estimation. The process only produces population for addresses known as existing during this reference year in the register. Also the addresses collected by the census that contributes to the computation are those present in the register and exclude buildings built and collected after the reference year or collected and demolished before the reference year. This approach is driven by the use of an address register as reference starting point but is not similar to the way estimates are done in the standard non-grided method. Nevertheless it is not supposed to have effects on the estimates bigger than the effect itself of having to estimate from a sample.

The data set

The final data set was build by aggregating the results of the two phases for small and large municipalities, summing population counts for grid cells relative to the two cases and removing the reference to the municipality for which they were computed. Actually the both cases give population counts that are not integers : for the second one because of the estimation that was carried out and for the first one because of the standard correction that is made to adapt the data collected to the target reference year. Rounding was done on the figures immediately after the last aggregation. The result is in the only numeric variable '**ind**'. It also must be noted that empty cells are not included in the data set. Those whose null value comes from rounding were excluded for the sake of consistency.

The identifiers of the grid cells (variable '**idLAEA**") were built to comply the suggestion made by the INSPIRE specification for grids. They have the form: 1KMNyyyyExxxx, where yyyy and xxxx are respectively the South-North coordinate and the West-East coordinate of the Southwest corner of each cell, expressed in kilometres in the LAEA projection system.



France (European part), 2007 population distribution by 1km grid cells.
Source: INSEE, Grided Census Estimations.

File history

12/10/2011 (FR_GCE_v1_0_1) : Grid cell identifier field was scratched in original release.
19/10/2011 (FR_GCE_v1_0_2) : Population counts rounded to the unity.
23/11/2011 (FR_GCE_v1_1_0) : Correction of an error in the location of some institutions whose coordinates were missing in the census result file and incorrectly estimated.
Coordinates were removed from the data file, because they were redundant with the identifier.
24/11/2011 (FR_GCE_v1_1_1) : Counts of populations not located back to the file (id='1km'), identifier back with heading lowercase characters.