GeoStat

GEOSTAT 4:

Geo-enabling statistical sampling frames

Geostat 4

eurostat

Title: Geo-enabling statistical sampling frames

Project: Eurostat ESSnet grant project GEOSTAT 4

Grant agreement number: 945503 - 2019-FI-GEOSTAT4

Author: GEOSTAT 4

Table of Contents

1. Introduction

Spatial data such as addresses, cadastral parcels and GPS coordinates are key information when designing surveys and collecting data, for they can be used to (see de Bellefon & Loonis,2018):

1. Reduce the costs of collecting data: the costs induced by collecting data in the field increase with the spatial coverage. In order to keep those costs under control, we use two-stage sampling. In the first stage, we sample small areas over the territory to be the collecting areas for the survey (they are called primary sampling units - PSU). The second stage consists in sampling units inside each PSU.
2. Improve the accuracy of the estimates: if two statistical units are spatially close together, they are most likely to look alike and follow the same pattern (Tobler's First Law of Geography related to the concept of spatial autocorrelation and foundation for spatial analysis). Thus, picking units that are spatially distant makes the sample more representative of the whole pattern. This is more important when the number of sampled units is small. In other words, balanced sampling should include the spatial dimension as often as possible, for it carries information that are related to the variable of interest but not necessarily carried by other variables. This paper discusses the impact of geo-enabling statistical frames on the data collection process and on the accuracy and precision of the results of the Labour French Survey (LFS).

2. Sampling principles of the French household surveys and of the LFS

The French household surveys include personal, face-to-face interviews, which are conducted by professional interviewers who belong to the permanent staff of INSEE and who are located throughout the country. The sampling is based on the so-called "master sample" of household-PSUs. Yet the LFS is an exception to that general pattern, for its PSUs are spatial groups of 120 dwellings called "sectors".

Since the 1960s, the population census is the common sampling frame for all household surveys in which the master sample is drawn every 10 years after the census. Furthermore, in the early 2000's the French population census moved from a ten-yearly exhaustive census to a more complex pattern, combining a five-yearly exhaustive census in the municipalities having less than 10,000 inhabitants and a yearly sample survey with rotating samples over 5 years in the municipalities having more than 10,000 inhabitants. While the sampling of the household-PSUs constituting the master sample was adapted to fit the new census, the LFS changed frames to a new source: the tax database.

Since 2017, the LFS sampling has been coordinated with the sampling of the master sample. The coordination of the two samples of PSUs makes it easier to organise the work of the interviewers and to control that a same household is not interviewed several times. Moreover, the household surveys now use the tax database as sampling frame, which fits better their annual frequency.

3. How is geospatial data used to build PSUs?

The method for defining the PSUs for both types of surveys is detailed below (see Costa et al.,2018). This method is close to the one previously used in the LFS (see Loonis, (2009).

PSUs for the Labour French Survey: the "sectors"

Since 2003, the LFS has been conducted on a continuous mode (it is carried out throughout the year with results as an annual or quarterly average). The primary statistical units are "sectors" gathering approximately 120 dwellings. Each sector is divided in 6 clusters of 20 dwellings and each cluster is surveyed during 6 consecutive quarters. Then, each sampled sector has a 9-years lifetime (36 quarters), and so has the sample itself. The sectors are built in a bottom-up approach: the clusters are built first, then the sectors are defined as a set of clusters. The clusters must be spatially concentrated to avoid time loss between two interviews. Clusters are built in two steps:

- Drawing a path of dwellings: the algorithm based on solving the travelling salesman problem, and the data used are the geospatial coordinates of the dwellings (Figure 1).
- Cutting the clusters: the path is cut every 20 dwellings.

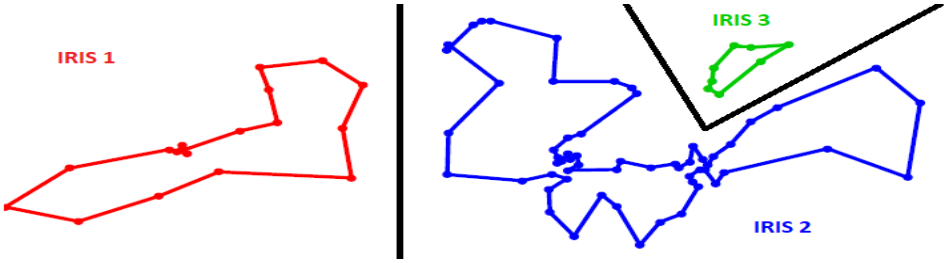

Figure 1: Figure 1. Create path of dwellings

To keep the extent of the clusters under control, several sub-paths are drawn inside subsets of the municipality called IRIS. This was not the case in 2009 (Loonis,2009) where a single path was built by ordering the dwellings by cadastral section number and by address. Cadastral sections, which area superset of the cadastral parcels, can however be very large, which led to unoptimized clusters. In this regard, some sectors contain more than 6 clusters and since it is impossible to ensure a perfect division of the number of sectors by 6we allow sectors to be grouped into 7 clusters. If some of these sectors are selected in the sample, 6 clusters among the 7 grouped will be effectively surveyed. The choice to draw paths inside the IRIS increases the number of 7-clusters sectors.

To limit this effect, the paths of some IRIS are combined into larger, cross-IRIS paths (Figure 2). To connect two IRIS, I1 and I2, you need to:

1. Look for the two dwellings inI1, with coordinates A1, B1, which are closest to I2 and the two dwellings in I2, with coordinates A2, B2, which are closest to I1

2. Connect the paths of I1 and I2 by connecting A1 to A2 and B1 to B2.

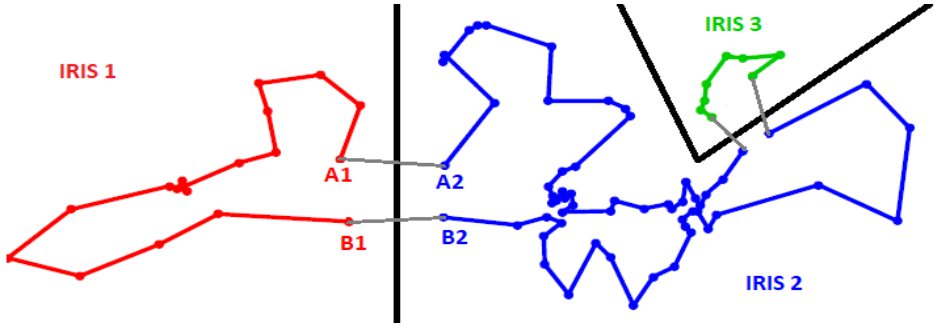

Figure 2. Expand the created path

PSUs for common household surveys: the master sample

Since 2004, the new method to carry out the population census consists in:

- Splitting the set of municipalities having less than 10,000 inhabitants into 5 rotation groups.
- Splitting the addresses within the municipalities having more than 10,000 inhabitants into 5 rotation groups as well.
- Assigning one rotation group to each year to be collected.

All inhabitants of the small municipalities belonging to the rotation group of the year are interviewed, so that over 5 years, the census is exhaustive within the municipalities having less than 10,000 inhabitants. In large municipalities, only 8% of the addresses belonging to the rotation group of the year are picked for their inhabitants to be interviewed, so that over 5 years, the census coverage in each municipality having more than 10,000 inhabitants is about 40%.

The official population of each municipality, whether it be large or small, is then computed over 5 years to make it robust. Until 2017, the PSUs for household surveys were drawn from the census. This primary sampling was balanced over the 5 rotation groups of the census (see Christine & Faivre,2009). From 2017 on, they have been drawn from the tax database, using the same method as the one described above, used in the primary sampling of the LFS:

- In each department, municipalities are linked by a path. The path connects the municipalities' centroids using an algorithm for travelling salesman problem.
- Then, each path is cut every 2,500 inhabitants (Figure 3)

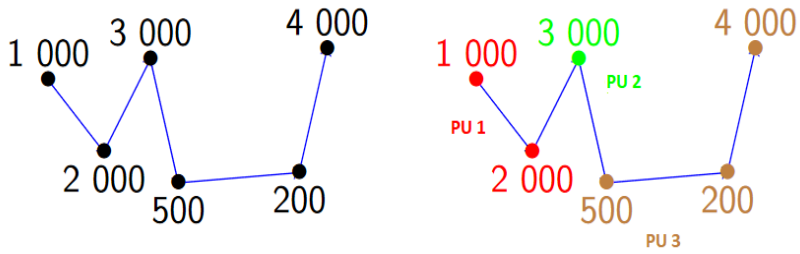

Figure 3. Build primary units with a path of municipalities

A PSU is then a geographical area delimited by the boundaries of municipalities in 2017. The PSUs defined depend on the starting point of the path. Therefore, several starting points are tested to select the path that minimizes the spatial extent of the PSUs. Figure 4 shows an example of a path in the department of Morbihan (Northwest of France), which results in 5,000 PSUs.

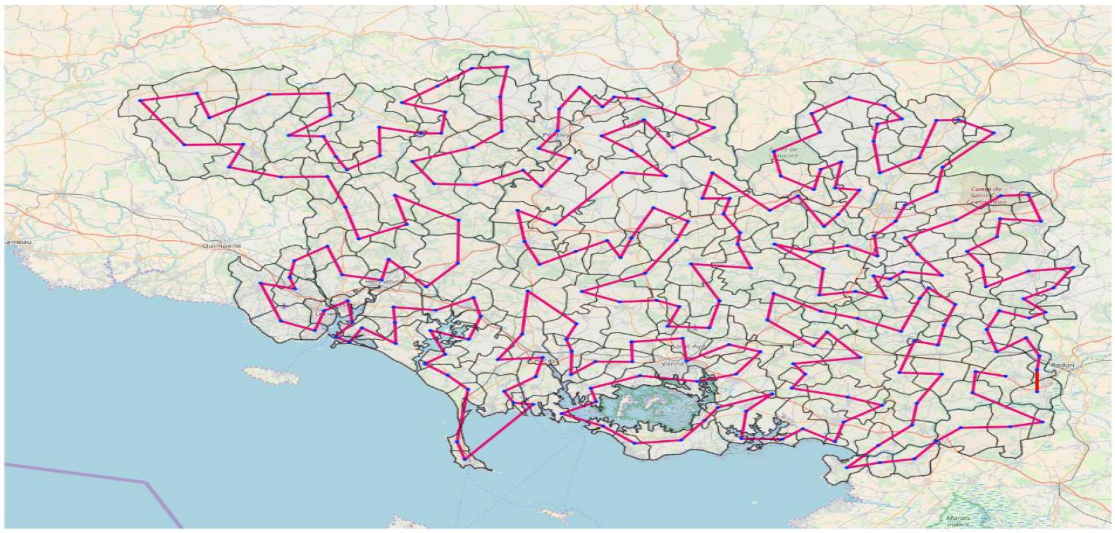

Figure 4. Example of path in Morbihan

4. Coordinated Sampling

As the PSUs for household surveys and LFS are two different, not interlocked sets of spatial areas, we have to define a superset to coordinate the samplings. This spatial superset for both PSUs is called "coordination units" (CUs) and is defined as a group of 4 spatially contiguous household-PSUs. CUs are thus spatial areas of 10,000 inhabitants (see Costa (2018)). Coordinating the primary samplings of LFS and the common household surveys then consists in:

- Drawing the master sample as usual.
- Drawing the LFS's sectors within the CUs containing the master sample. We could draw a CU sample first, but this would affect the balancing of the master sample and make it less representative of the population.

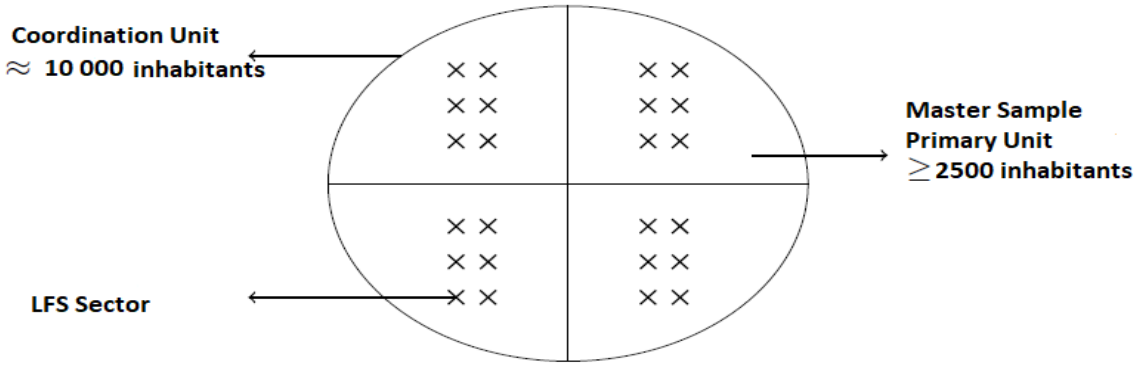

Figure 5. Coordinated sampling

Furthermore, the primary sampling for the LFS is spatially balanced (see details about spatial balancing in Grafström et al.,2012), which allows more distant units to be drawn. The Figure 6 shows on the left side a spatially-unbalanced sample and on the right side a spatially-balanced sample.

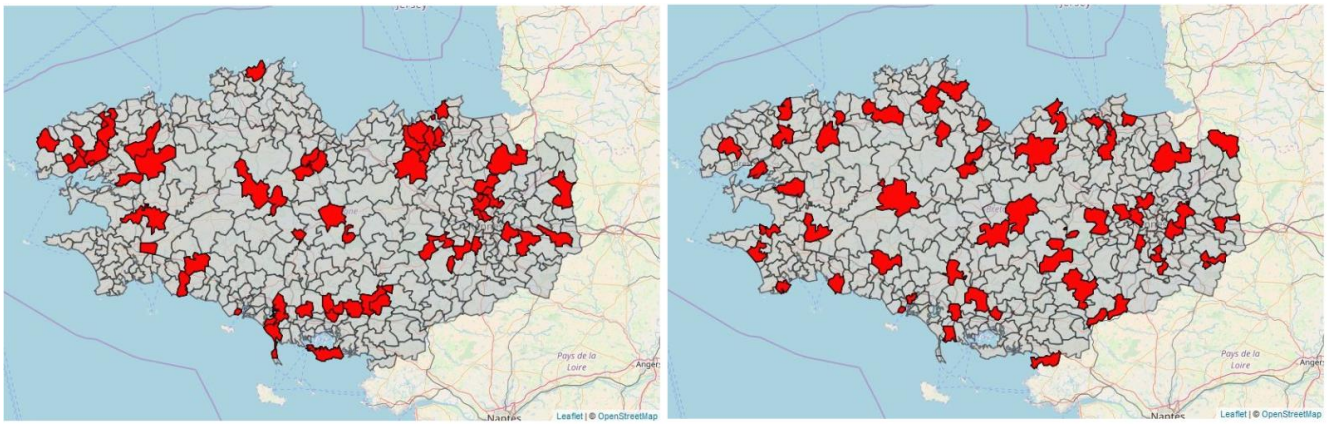

Figure 6. Spatially Balanced Sampling

5. Input data and tools

The constitution of the PUs for the master sample as well as the spatially balanced sampling of the master sample PUs only requires the municipality boundaries as geographical information. For the LFS, however, it is necessary that the tax source contains the coordinates of the dwellings. Administrative contours (IRIS and municipalities) must also be assigned to all dwellings. This is done through the georeferencing process of the tax source.

This geo-referencing process is carried out by matching the tax source to the cadastral data via the cadastral parcel identifier available in the tax source file. However, the cadastral parcels were not available in vector format and their location was approximated by a point. In 2020 the geo-referencing process of the tax source was redesigned. The new version of the process considers the fact that 99.9 % of cadastral parcels in France are now available in vector polygonal format. Therefore, in order to allocate a dwelling to an IRIS (municipalities' subdivision), the following procedure is used:

1. The municipalities that intersect with the vector cadastral are listed.
2. The IRIS on which the cadastral parcel spreads the most in surface is selected for this dwelling (Figure 7).

In theory, intersections between parcels and IRIS should not exist since the cadastral parcels form a partition of the IRIS. However, the geometric contours of the communes and the geometric contours of the parcels are not consistent. This "surface" method thus makes it possible to carry out this lack of spatial consistency. One of the main drawbacks of this method is that the parcel centroid calculated from the polygon may point to a different IRIS than the one obtained with the "surface" method. On the other hand, the IRIS geocoding process is more robust to small contour changes.

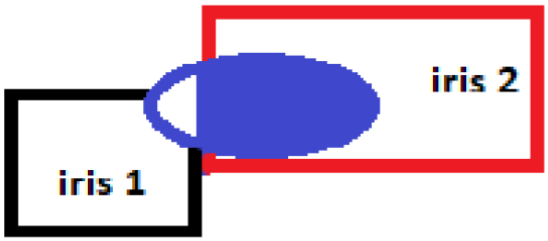

Figure 7: Spatially Balanced Sampling

The use of data in polygon form is also facilitated by the combination of the use of a PostGIS database server (in particular via the use of geometric indexes) interfaced with the R programming software in which the *sf* package makes it possible to easily combine database and calculation on geometric data.

6. Conclusion

The example of the LFS's primary sampling shows how geospatial data can be used to improve sampling. The definition of sectors of dwellings for the Labour Force Survey and of areas of households for common household surveys requires information on the coordinates of the dwellings. Those primary, spatial-based sampling, make the data collection process more efficient, particularly regarding resources management. The spatially balanced sampling increases the heterogeneity within the sampled units, so that estimates are more accurate.

7. References

Christine, M. and Faivre, S. (2009). Le projet octopusse du nouvel échantillon maître de l'insee.

Costa, L. (2017-2018). Le nouvel échantillon de l'enquête emploi. Master's thesis.

Costa, L., Guillo, C., Paliod, N., Merly-alpa, T., Vincent, L., Chevalier, M., and Deroyon, T. (2018).Le tirage coordonné du nouvel Échantillon-maître nautile – Journées de méthodologie statistique.

de Bellefon, M.-P. and Loonis, V. (2018). Handbook of Spatial Analysis Insee.

Grafström, A., Lundström, N. L., and Schelin, L. (2012). Spatially balanced sampling through the pivotal method. Biometrics, 68(2):514–520.

Loonis, V. (2009). La construction du nouvel échantillon de l'enquête emploi en continue à partir des fichiers de la taxe d'habitation