Using the European Grid "ETRS89/LAEA_PT_1K" as the foundation for the new Portuguese Sampling Infrastructure

Author: Ana SANTOS
Statistics Portugal, Portugal
Author 2: Bart-Jan SCHOENMAKERS
Statistics Portugal, Portugal

Operative case study report of GEOSTAT 1B

Introduction

Since 1979 household Surveys are based on a large sample of housing units, commonly referred to as "Master Sample" (MS). The Master Sample is selected after the completion of each Census and is maintained over a decade updated to a greater or lesser degree.

Failure to update the MS is a contributing factor to an increase of the non-response to surveys and to the consequent loss of quality of the produced estimates. Some of the reasons that contribute to this downgrade in quality are the inclusion of new accommodation without eliminating others that do not exist anymore (demolished or changed form of occupation). On the other hand, the increase in household surveys causes depletion of the MS, therefore in 2006 Statistics Portugal decided to start an undergoing update process.

New developments like, the availability of the georeferenced buildings of the 2011 Census, together with the access to the data of different administrative sources (with different contents and designs) and the implementation of European initiatives as the project EURADIN (European Address Infrastructure) and INSPIRE (Infrastructure for Spatial Information in the European Community), gave Statistics Portugal an important opportunity to change their sampling infrastructure for household surveys.

Objective

The implementation of a new approach for the sampling infrastructure of Statistics Portugal.

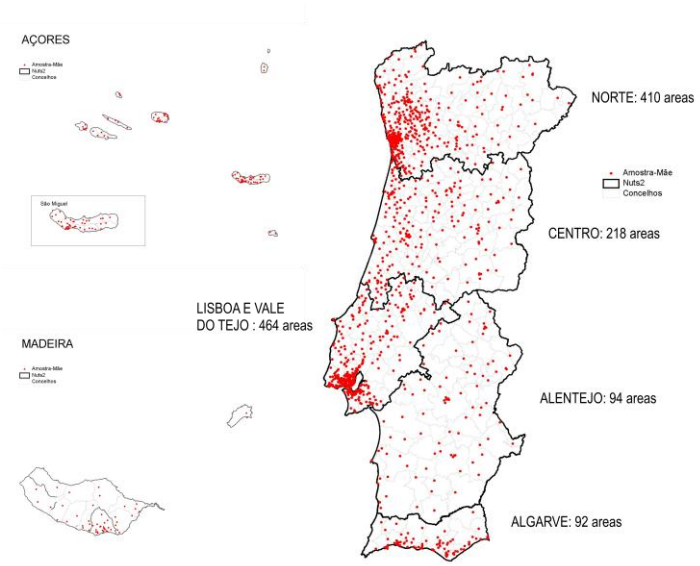
Framework

MASTER SAMPLE 2001

The 2001 Master Sample (MS2001) is a stratified, one-stage cluster sample selected with probability proportional to size and it was designed taking into account the provisional results of the 2001 Population and Housing Census. The MS2001 was initially planned for a 5 year period. The primary sampling units (clusters of statistical sections) were selected with the help of a GIS; contiguous sections of the 2001 Census were grouped taking into consideration a minimum number of dwellings.

The MS2001, like its predecessors, had to be updated in order to maintain the quality of the frame. Thus, between 2006 and 2010 Statistics Portugal conducted a fieldwork to guarantee that the MS2001 could be used until 2013, the year of the transition to the new sampling frame.

Distribution of the MS by NUTS II (1408 areas)



The MS2001 has the following variables:

- dwelling identification code
- name of the household head
- address
- locality
- postal code
- telephone
- occupancy status: usual residence, seasonal or secondary use, vacant for sale , for rent, for demolition or other cases
- type: conventional or non-conventional dwelling (shanty, rudimentary wood house, mobile house unit, others)

THE NEW SAMPLING FRAME FOR HOUSEHOLD SURVEY PURPOSES (NATIONAL BUILDINGS AND DWELLINGS REGISTER)

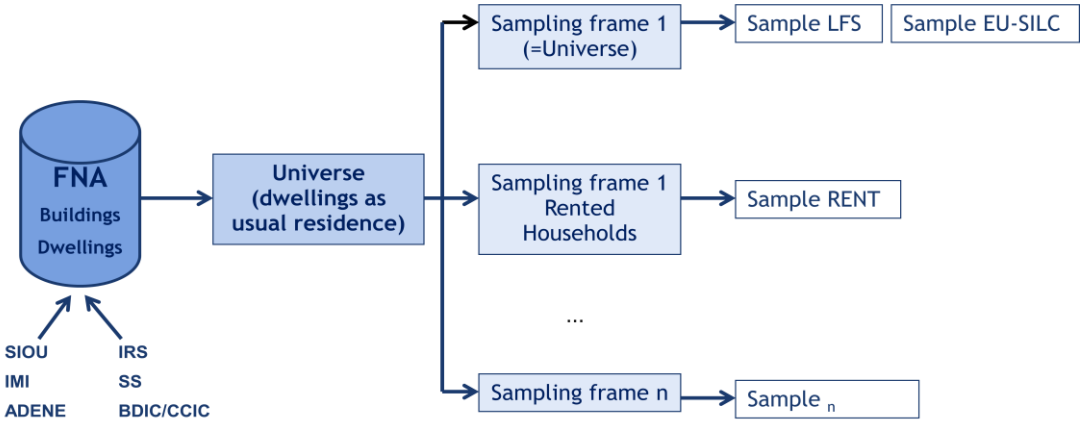
In 2011, Statistics Portugal obtained a national database comprising all the georeferenced buildings from the 2011 Population and Housing Census. This geographical base has been used to reference census data at point level and to support the creation of a National Dwellings Register (FNA).

Once the necessary conditions are fulfilled, this extensive file can be updated by taking advantage of data available in different sources, like statistical surveys conducted by Statistics Portugal and also the following administrative sources:

Source	Entity	Register	Frequency
ADENE	Agency for Energy	Efficiency and Indoor Air Certification	Monthly
IMI	Tax Authority	Property tax	Annual / Monthly
SS	Ministry of Labor and Social Security	Social Security File	Annual
IRS	Tax Authority	Singular Income Tax	Quarterly
BDIC/CCIC	Institute of Registries and Notaries, IP	Civil Identification Base	Annual

The model to be adopted for the purpose of the household surveys will be as follows:

Architecture of the proposed model for Household Surveys



a) Annually a Universe\Population will be extracted from the FNA comprising all the dwellings as usual residence;
b) The sampling frames will be extracted from this Universe, the differences between the sampling frames are related to the specific scope of the survey. For example, the Labour Force Survey (LFS) will have a sampling frame equal to the population as the scope extends to all people residing in Portugal, in the case of the Rent Survey the scope is limited to rented accommodations;
c) Samples will be selected from the corresponding sampling frame;
d) Any information obtained from fieldwork will be included to update the information of the population\universe and the different sampling frames.
e) We plan to update FNA in two different ways:
   – Dynamic updates from different sources:
      i. External: IMI and ADENE sources concerning the population characteristics, which include new and/or demolished buildings or dwellings and the update of some specific attributes which are relevant. Other essential information for the sampling process regards the changes in the occupation of the dwellings;
      ii. Internal: The "System of Indicators in Urbanistic Operations" (SIOU) assures the geocoding of all the new construction and

eventual demolishments for all the municipalities of the Portuguese territory.

The Geography of the new Sampling Frame for household survey purposes

In order to support FNA it is essential that there is a strong geographical component which supports the sampling process. Some reasons can be mentioned for the increased importance of the geography in the sampling infrastructure.

Taking into account the characteristics of household surveys, namely the fact that the collection of information is still dependent on Computer-assisted personal interviewing (CAPI), the interview is conducted at the selected housing unit by an interviewer, therefore the correct geographical localization is essential.

Furthermore, in 2013 there is a major revision of the Portuguese administrative division (at parish level), along with the possible revisions of the NUTS areas every three years.

The promotion of the use of the European GRID with the 1 Km^2 cells is another important aspect. Using this European GRID, consisting of rectangular cells, allows the organization to represent uniformly the buildings regardless of their administrative division. This also creates the advantage that the system will be independent of changes in the boundaries of the statistical sections and subsections, enabling the harmonized and interoperable geographical location of buildings.

Input Data

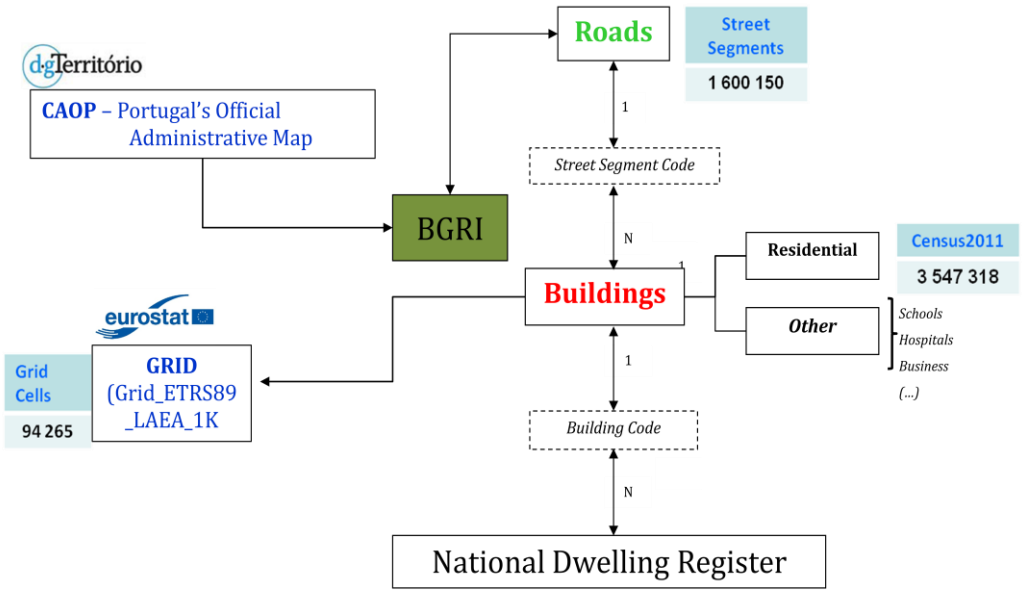
The system is composed of several geographical datasets and the 2011 census population data, all organised within the Datawarehouse of Statistics Portugal and an ArcSDE geographical database.

- **Geographical features**
  - Centroids of the georeferenced buildings
  - The European "ETRS89_LAEA_1K" GRID
  - The administrative division at parish level (NUTS5)
  - Geographical reference data like road information, digital imaginary and the statistical division

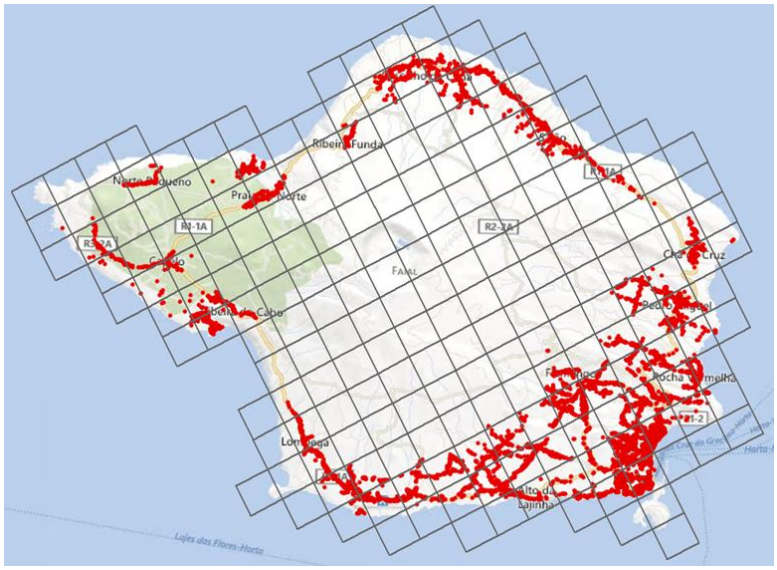
- **The Portuguese part of the "ETRS89_LAEA_1K" GRID – 94 265 Grid Cells of 1x1 Km**

  GeoStat 's 1Km^2 Grid and georeferenced Census building points are being used for the creation of the sampling frame and the extraction of optimized samples regarding the type of survey or certain social, economic and territorial characteristics in a certain geographical area.

Geographical features of the national dwelling register



GRID and Census Buildings at Faial Island

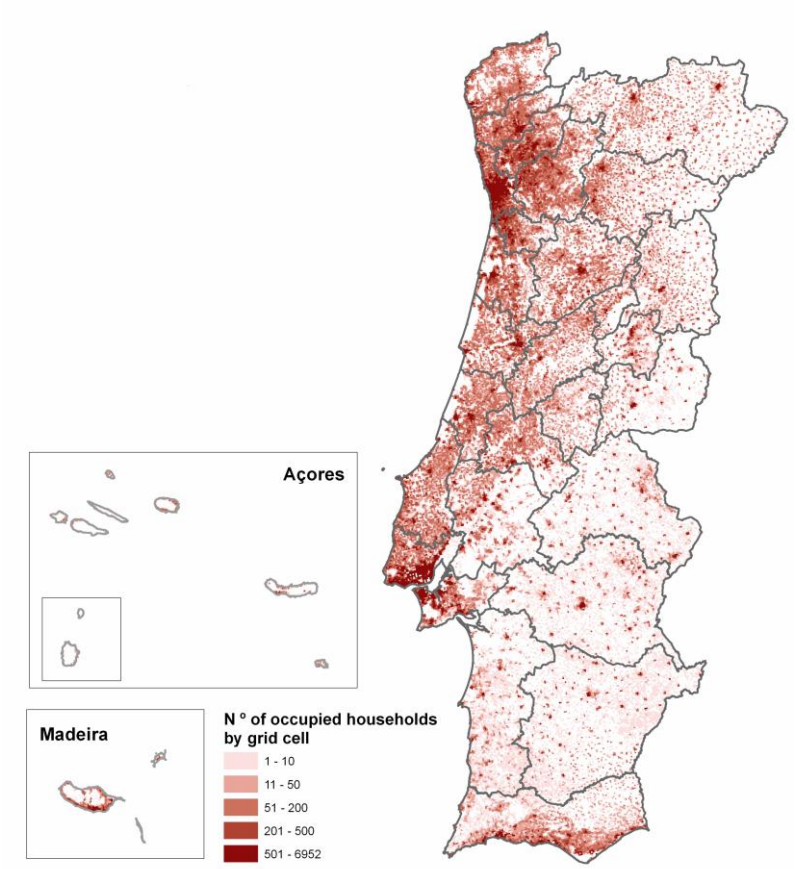


Methodology

The methodology was the following:

- Relating the Census Buildings to the European GRID
- A spatial overlay to obtain the corresponding GRID Cell of each building
- The European GRID is in the LAE-ETRS projection and has been projected for the 4 different projections used for the Portuguese territory.
- Only 43% of the cells have occupied homes

N º of occupied households by grid cell



Classification of the Cells in High and Low Density

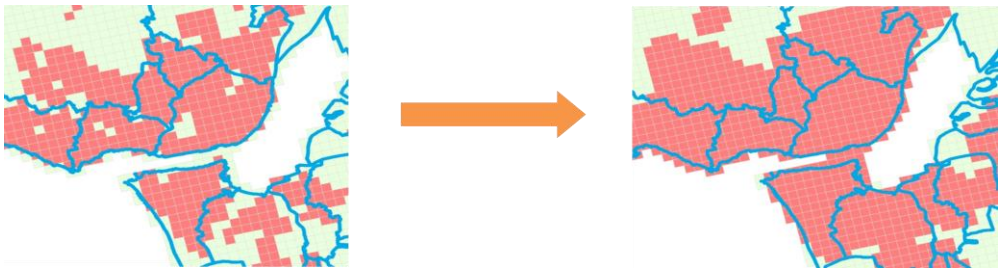
For some surveys it might be useful to distinguish between urban and rural cells, therefore a classification has been made in High Density or Low Density. This classification is based on the population density of each cell and its adjacent cells. This is an adaptation of the revised typology for the Degree of Urbanization of Eurostat.

A cell is classified as High density if they fulfil all of the following conditions:

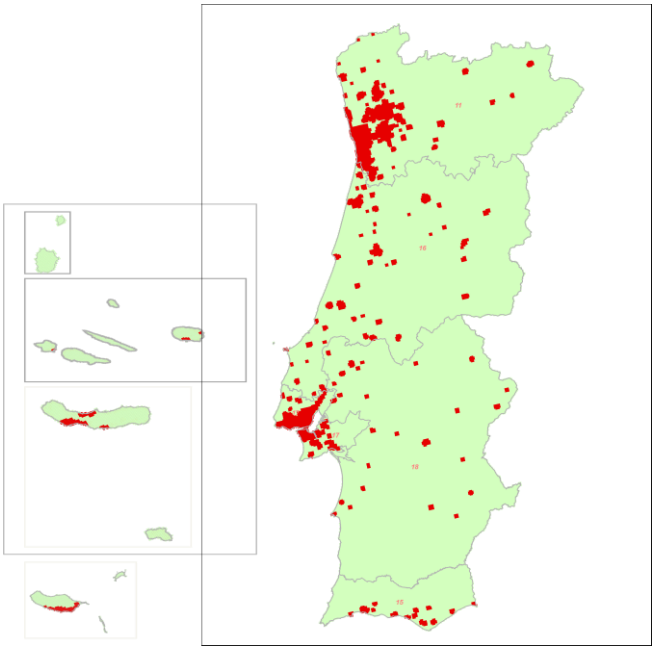
- Each cell must have population densities exceeding 300 inhabitants/Km2
- The set of contiguous cells must contain at least 5,000 inhabitants

All the other cells are classified as low density cells.

To avoid the fragmentation of the areas of high density cells and the influence of physical borders (such as sea or river shore) all the adjacent low density cells of high density cells are classified as high density. The figure below shows the effect of this procedure:



Distribution of high and low density cells



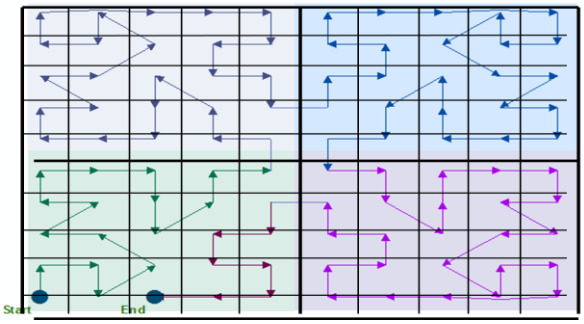
Ordering of cells in each NUTS III area

For the sampling process of the household surveys NUTS3 areas (or aggregation of several areas) are used as strata. Within each stratum, cells are selected in an order which should assure their contiguity/proximity. For this process each cell within the NUTS 3 area is assigned a sequential number (other than the ID). To be able to assure the spatial contiguity of the cells

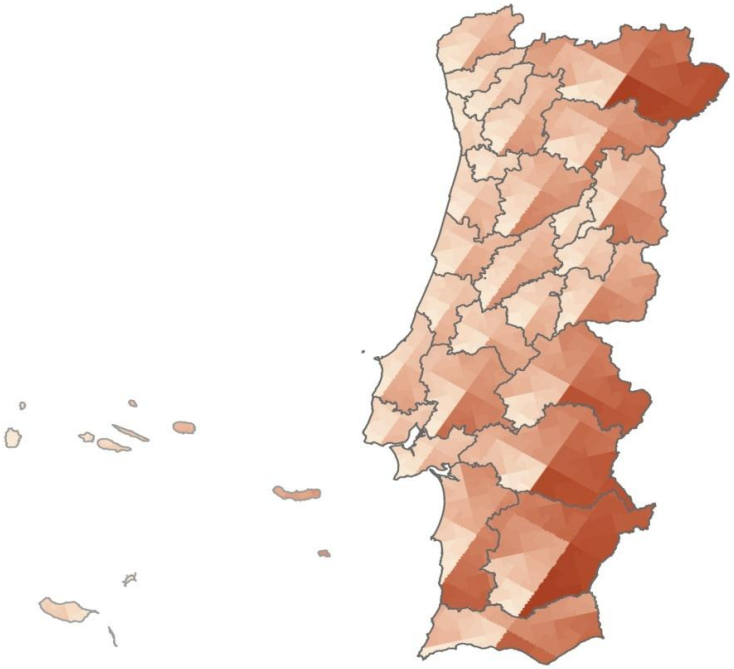
within the sequential order the cells have been sorted geographically regardless of the classification of the cell.

The methodology used for this spatial sort is the "PEANO" method (using a space filling curve algorithm, also known as the Peano curve), a functionality implemented within the ArcGIS software.

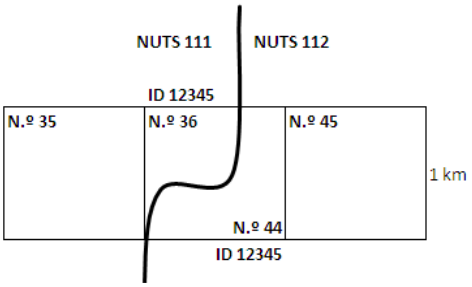
Example Methodology Ordering Method "PEANO"



Spatial sort within each NUTS3 Area



- Darker Cells have a higher Rank order.
- One cell can be represented more than once, when it is located at a NUTS 3 boundary. The same cell has a different sequential numbers in each distinct NUTS 3 area (see figure).



Conclusions

This study, to evaluate the feasibility of using the European GRID for sampling purposes is an important development to support the sampling infrastructure of the household surveys conducted by Statistics Portugal.

One of the main problems with the former sampling infrastructure was the exhausting of the sampling frame and the absence of a continuous updating process. Creating a national dwelling register will allow statistics Portugal to be able to have an updated sampling frame where administrative sources are an important element of this infrastructure. This can lead to a cost reduction for the carrying out and management of the different surveys at Statistics Portugal. However some challenges still exist for the integration of this information in the national dwelling register. Right now Statistics Portugal is studying and implementing methodologies to assure the proper integration of the administrative information.

The use of the European GRID as one of the geographical components of the sampling infrastructure is a major development, since the former sampling process was dependent on the administrative division. Using this GRID can incorporate a degree of flexibility in the sampling process.