Data collection methods with geospatial tools

One key area of statistical production refers to data collection, including survey data and sample design. Developing procedures and tools that make it easier for survey respondents to provide information, whit at the same time, trying to reduce response burden, is an important goal. Also, working on solutions that make it for interviewers to conduct their work in the field in the most productive way possible is also a fundamental dimension to increase efficiency and quality in data collection.

In the context of sample design and under the scope of a strategy to increase the modernisation and efficiency of statistical production, through its methodological and technological development, Statistics Portugal has also put into practice a new methodology to define sampling frame and sample design taking advantage of the Spatial Data Infrastructure (SDI) of Statistics Portugal, including geocoded microdata.

Using spatial sampling frame

Taking advantage of the georeferenced information (x,y geographic coordinates) for all the 2011 Census buildings, a National Buildings and Dwellings Register has been defined to support the sampling process for household surveys, regularly updated through administrative data and survey data. An important geospatial data has also been integrated in this process, the European 1 km2 grid (INSPIRE grid net ETRS89-LAEA-1K) as a new reference for Primary Sampling Unit (SPU) selection. It is important to highlight that 2011 Census buildings as a point-based database (x,y geographic coordinates) was crucial input to produce the 2011 Portuguese population grid, as part also of the European Statistical System (ESS) project GEOSTAT 2 and the dissemination of the 2011 European grid to the 1 km2 INSPIRE grid net, above mentioned.

Usually, sampling selection follows a stratified and multi-stage sampling scheme, in which the PSUs, geographically constituted by one or more contiguous cells of the 1 km2 (Figure 1), are systematically selected with a probability proportional to the size of the number of dwellings of usual residence. The Secondary Sampling Units (SSUs) are systematically selected within the units for the first step. All the PSU of sampling frames for surveys with rotations must include roads.

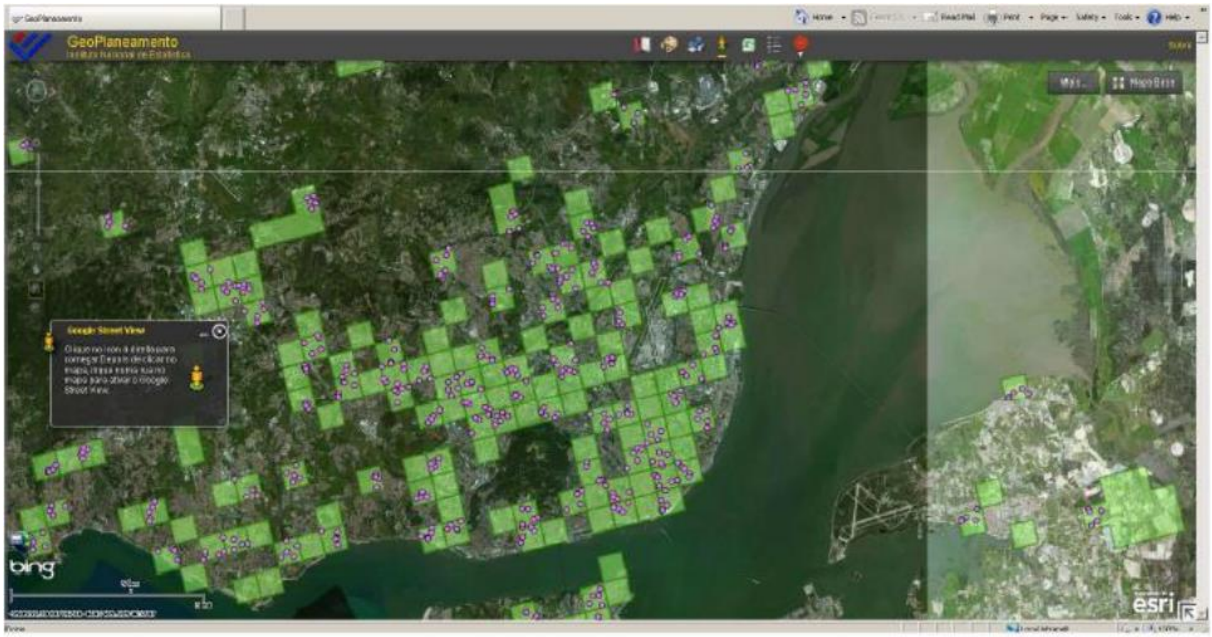

Figure 1. Example of grid cells selection to define PSUs

Using this spatial sampling design has allowed to reduce the intra-cluster correlation coefficient (which measures the similarity of statistical units) associated with selecting dwellings in ”segments”.

A georeferenced sampling frame has shown to improve the accuracy of estimates. The most the sampling design selects individuals geographically distant from one another, the more the estimation will be precise for a spatially auto-correlated variable. Additionally, in case of face-to-face interviews (CAPI mode collection) knowing the location of the statistical units sampled makes it easier to identify them in the field and to manage interviewers' locations during the fiedwork. Maintaining the underlying point-based data update is crucial to increase the efficiency of the spatial sample design process, as well as of data collection.

Increasing efficiency in data collection management with geospatial tools

In the past, interviewers regularly faced difficulties in locating their sample housing units in household surveys as they could only rely on tables with address information, name and contacts of the household representative. A geospatial web tool, custom-designed to respond to the needs of statistical data production was implemented within the scope of integrating geospatial data into the official statistics' production model.

The GeoINQ web application was develop by Statistics Portugal in partnership with ESRI using an API for ArcGIS environment. The tool integrates point-based data for households of sampling frames and a set of relevant background geospatial layers (NUTS, Administrative Units, 1 km2, BGRI) and base maps, including the orthophoto maps from the Portuguese National Mapping and Cadastral Agency (NMCA), Directorate-General of Territory.

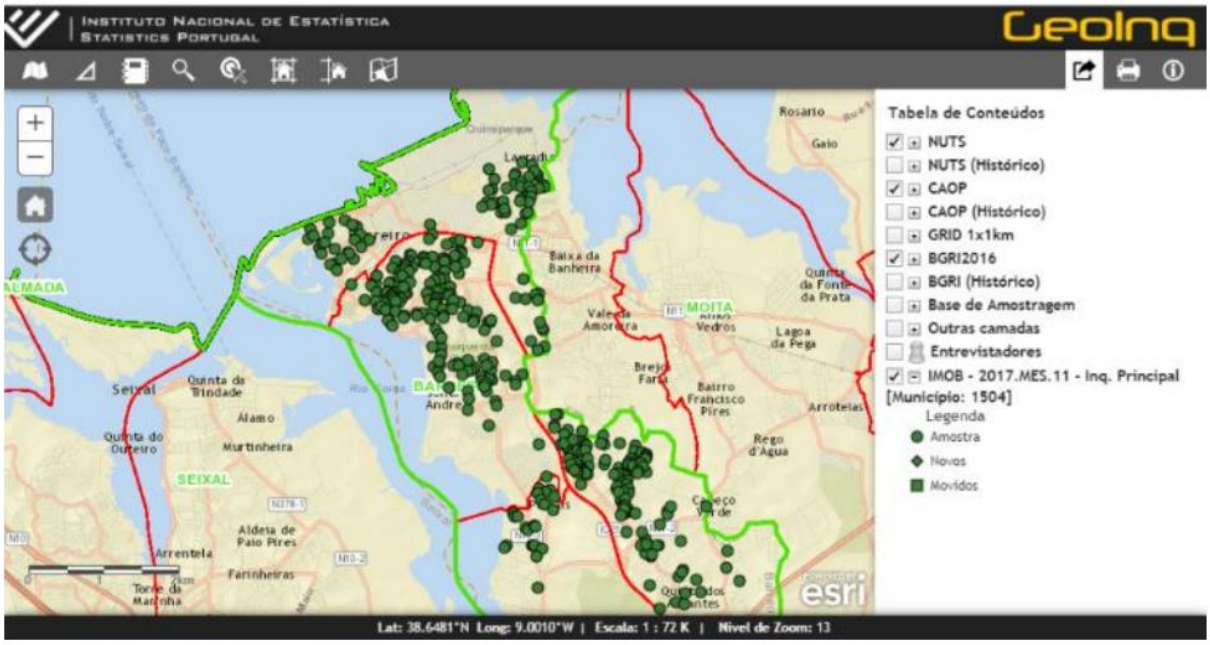

Figure 2. GeoINQ web application

With GeoINQ interviewers can easily identify the precise (x,y geographic coordinates) location of dwellings and have access to associated data. GeoINQ runs on a desktop environment and users can only access those features and geographical layers compatible with their user profile.

GeoINQ is fully integrated with other systems developed at Statistics Portugal, with the global interview survey management system (SIGINQ-IE) and support the collecting data of all social surveys. Thus, besides interviewers, other internal users make use of this web application to meet their needs on data management and analysis, namely to analyse the geographical dispersion and

overlap of samples on national territory within the process of spatial sampling design. Maintaining the underlying geospatial data updated is, in this context, fundamental to keep benefiting from the useful features associated with this type of geospatial tools supporting statistical data production.

Implementing geo-solutions to capture challenging variables

In 2017, Statistics Portugal conducted a survey on mobility in the two Portuguese metropolitan areas – the Metropolitan Area of Porto and the Metropolitan Area of Lisbon. Based on a stratified and multiphasic random sample, which considered homogeneous areas of accessibility to transport, a mix-mode data collection approach was followed, by combining Computer Assisted Web Interview (CAWI) and CAPI (Computer Assisted Personal Interview).

The aim of the survey was to characterise the movements (not limited to commuting) of the resident population (6 to 84 years old) in the two metropolitan areas. This involved being able to capture points of origin and destination for each trip during a specific weekday, as well as other dimensions to understand how people move, how often they travel, how much time they spend moving, where they go and to do what. The main challenge associated with designing a web-survey to meet this aim was to come up with a way people could easily register their movements during the day and find/pinpoint the locations where they went to.

Instead of descriptive reports, an innovative solution was implemented using Google Maps. The maps were used to capture travel destinations with the same functions people are used to finding in Google Maps, as well as location circles based on the centroid of the municipality (LAU 1) to the farthest point to help people navigate the different locations.

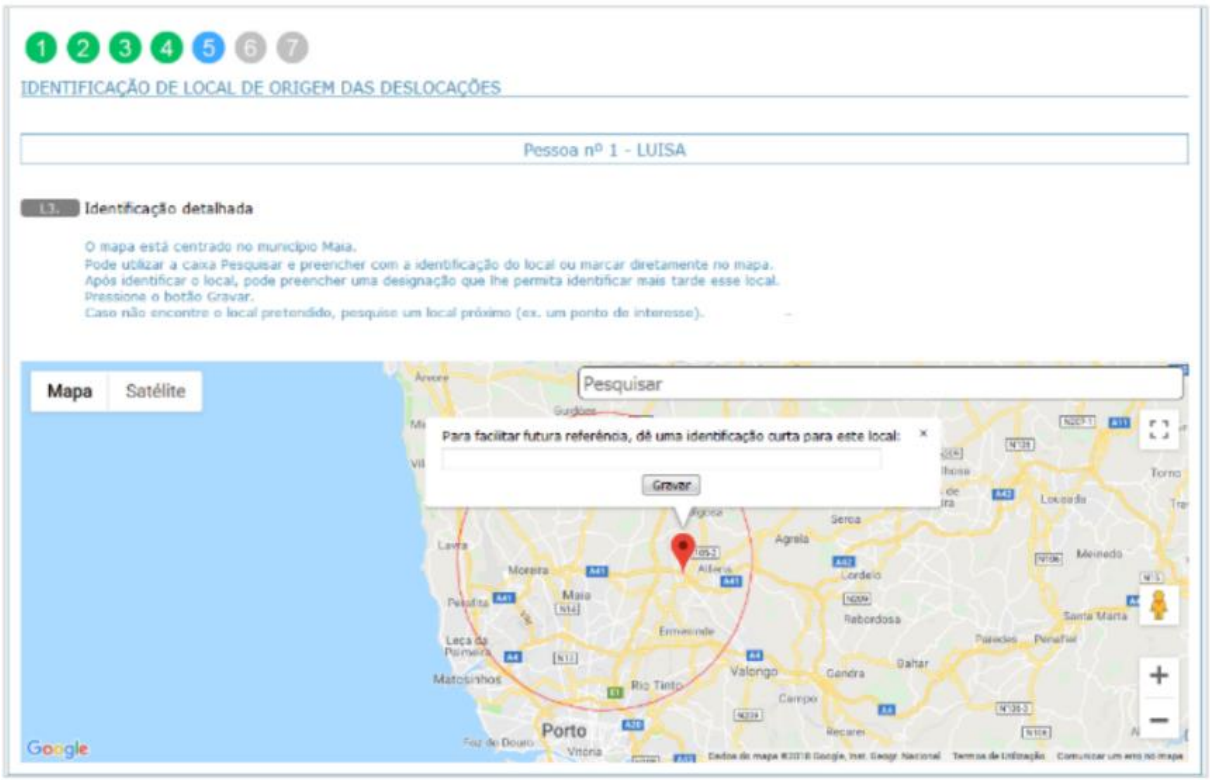

Figure 3. Example of the response screen for identifying locations on the Survey on mobility in metropolitan areas

Nevertheless, outsourcing services in the geo-information area are quite uncommon and subjects to an evaluation to ensure that they meet the quality criteria and assurance for statistical production. This assessment may be more limited for commercial bases and products. In addition, it also implies that statistical institutes may be dependent on external services with limited capacity for interventions and subject to changes that may direct or indirectly affect implemented statistical production processes.

Contact information

Ana Santos, Statistics Portugal, ana.msantos@ine.pt or ine@ine.pt
Rossano Figueiredo, Statistics Portugal, rossano.figueiredo@ine.pt or ine@ine.pt