GeoStat

GSGF Europe:

Geospatial statistics according to data collection method

SURVEY DATA

GEOSTAT 4

eurostat

Title: GSGF Europe: Geospatial statistics according to data collection method - Survey Data

Author: GEOSTAT 4

Project: Eurostat ESSnet grant project GEOSTAT 4

Grant agreement number: 945503 - 2019-FI-GEOSTAT4

Summary

Most statistical production processes rely on a survey data to collect, process, analyse and disseminate information thus constituting a traditional data acquisition method for most statistical authorities. However, there is an increasing awareness that conventional surveys, including censuses and questionnaires, with longer periods, less data frequency and fixed output areas, such as enumeration areas or other small statistical areas, do not meet the user demands on data availability, usefulness, timeliness, spatial resolution and territorial flexibility. These data gaps and constraints are related to basic units of collection, alignment of geographic entities and temporal cycles of data collection and updates, and sometimes institutional preferences, national and international context. In addition, survey operations are a very resource and time-consuming, especially when exhaustive surveys are conducted and statistical representativeness for small areas is achieved.

The above issues pose major challenges to the statistical authorities meanwhile responding to the growing need to add the spatial dimension in traditional data sources, processes and derived outputs in order to geo-enabling the statistical production and produce geospatial statistics. In this regard, approaches of geospatial integration should be considered when designing the future of official statistics, particularly through geospatial data and tools implementation on data collection management and coordination, updating the traditional survey lifecycle. These challenges may also encompass the combination with non-traditional data sources and data types as a strategic pathway for the modernisation of official statistics. These alternative and emerging data sources and data types such as administrative data and Big Data should be integrated within the statistical production process, while exploiting their strengths to tackle data gaps and obstacles.

Attending these concerns, a document focused on survey data to support the production of geospatial statistics was developed within the GSGF Europe's User Guide. This document also aims to promote a long-term processes and objectives through short-term actions in data collection management, using geospatial approaches, activities and tools to geo-enabling statistical production, particularly processes based on survey data, and bringing the geospatial perspective into statistical pipelines which may depend on survey data as input data. The document describes the data collection method, such as types of surveys, contextualise survey data within the GSGF Europe and European statistical-geospatial operating environment, including GSBPM and Geospatial View of GSBPM, focusing on the Design, Build and Collect Phases, phases that manage more directly with survey data and encompass related sub-processes and activities. This type of analysis is suitable since despite GSBPM is applicable to a wide range of data sources such as surveys and censuses, the framework does not focus so much on survey data when describing the phases and sub-processes. A brief overview to spatial sampling and geo-sampling to survey data management is presented and lastly, specific Requirements and Recommendations within the GSGF implementation in Europe addressed to survey data are proposed.

Content

1. Introduction

In the current context of rapid urbanisation and migration of population with the consequent change in the spatial distribution of people and business, it is crucial for the National Statistical Institutes (NSIs) to collect updated data with greater spatiotemporal resolution, completeness and reliability in a landscape of digital transformation and an open data environment.

Since data collection is one of the key areas of the statistical production process, the development of methodologies and tools that lead to reduction of the statistical burden and promote more flexible and productive collection processes, are relevant ambitions for the statistical community.

The importance of geospatial information is being widely recognised. It plays an increased integrative and strategic role in the statistical production process and in the design of new solutions for efficient data collection methods in a more systematic approach. Furthermore, the potential of location for full integration into the statistical production process has been acknowledged by the European Union and its Member States for almost two decades towards a society empowered by data.

NSIs face challenges in data acquisition related to common standards and frameworks, basic units of collection, consistent geographies, and temporal cycles of data collection and updates. Also, data collection methods within the statistical production process are still conventional in many cases and have a self-centred perspective from the statistical domain, not being open and permeable to other types of data, production processes and frameworks. These challenges become critical concerns to produce geospatial statistics when methodologies of geospatial integration and other geospatial activities are implemented towards a common framework for a geospatial-statistical data acquisition context. This data integration process will uncover differences in the data collected requiring a consensus on basic statistical and geospatial content, related to metadata content (e.g., spatial characteristics of data). In this regard, survey data need to have a more comprehensive understanding of the geographical space mainly through a point-based geocoding infrastructure for statistical production, a key condition supported by the National Spatial Data Infrastructure (NSDI).

The inclusion of the geospatial dimension in the collect phase within the statistical production process, including surveys' operations, provides a more accurate location of the statistical units, supporting the construction of geographical primary sample units, thereby improving the survey sampling design and producing geospatial statistics on very small areas.

Furthermore, the integration of statistical and geospatial data can contribute to all phases of the statistical production process and geospatial data and related activities may be part and play a relevant role in all phases and most of the processes, including survey data and related data collection methods and in all dimensions of statistical outputs. Statistical and geospatial data integration in data collection activities provides the necessary information for data collectors and survey interviewers to evaluate compliance between traditional and non-traditional data collection methods, encompassing the geospatial perspective. For that matter, this document aims to: i) broadly describe the survey data collection method, its features and specific characteristics; ii) contextualise survey data within the GSGF Europe framework and in the European statistical-geospatial operating environment; iii) describe survey data and related data collection methods within the statistical pipelines and overall data production model, including references to GSBPM and its geospatial perspective, GeoGSBPM; and iv) identify and briefly describe new specific requirements and recommendations in the context of the GSGF Europe and its Principles.

2. Data collection method description

Survey data is defined as output data from a survey used as a research method for collecting data from a pre-defined group of respondents to gain information and insights into various topics of interest. In general, survey data could be obtained and requested manually and electronically from citizens and organisations by direct interview (CAPI - Computer Assisted Personal Interview), by filling in the questionnaires (with or without the support of an interviewer) and by telephone (CATI – Computer Assisted Telephone Interview) or Internet (CAWI - Computer Assisted Web Interview). In most surveys carried out by direct interview (face-to-face), multi-stage sampling designs are used to lower data collection costs through geographically concentrated interviews and are constrained by initially defined parameters and estimators. This happens because survey data collection is time-consuming and requires a great amount of effort and expense demanding the minimum sample size in a statistically valid manner. In the case of survey sampling on continuous data flows (e.g., social or business surveys) is very often firstly to have a direct interview and consequently to use telephone interviews to conduct surveys and household surveys for survey sampling housing units.

Survey data comes from Primary Data Sources, such as Census and Surveys, and unlike Secondary Data Sources (e.g., administrative data) and Big Data sources, they are designed to be used in statistical production, including concepts, definitions and classifications that are stable and known and having a well-established body of knowledge. Like administrative data, survey data are structured, interest variables are directly available and do not need “heavy” pre-processing to be used in statistical production.

Survey data are also representative - or the lack of representativeness is intentional and/or can be adjusted for in analysis – and auxiliary variables are directly available, which often happens with administrative data and never with Big Data. Lastly, survey data has a target (sub-)population defined and always refers to units of the population of interest, one of the main differences in comparison to administrative data and Big Data. In addition, administrative data can often define a target population and usually refer to units of the population of interest, for instance, sectorial public bodies that only collect administrative data regarding their field of application, covering only a part of the population. Survey data also presents a temporal concern related to the frequency of data availability in which the longer the period, the more dated the data become unless supplemented by other data sources, namely administrative data to support the sampling frame.

In general, survey data is obtained through exhaustive or sample surveys. The exhaustive surveys are Census statistical operations in which every item from a given population is subject to observation (universe) and assumes a complete coverage of a territory, a door-to-door count resulting in a full enumeration of the country. In other words, a Census is a survey conducted on the full set of observation objects belonging to a given population or universe. In some countries, statistical sampling in a decennial census was implemented. Sampling made it possible to ask additional detailed questions of the population without unduly increasing cost or respondent burden. Enumerators asked a random sample of the population (in some cases approximately 5 percent) a set of extra questions. Then, the NSIs used the sample to extrapolate demographic data for the entire country.

Otherwise, sample survey data are collected from a sample representing the population under observation and usually collected through standardised and traditional methods and procedures from a selected population. Common sources for a sampling frame are administrative and statistical registers, Census data and information from other sample surveys. Concerning both types of survey data, Census are the main statistical data collection method gathering more detailed statistics and analysis of demographic and socioeconomic variables in various countries. During a Census, most statistical institutes are now able to capture data based on geocoded data collection points such as addresses or buildings. Census data are usually much more detailed than any other official data published by the NSIs.

In a national and global context, the address as basic location information is becoming increasingly important and the most elemental geospatial unit, usually represented in some form of geographic coordinates (x,y) and/or descriptive information using a unique identifier system. Despite producing complete coverage of the country, Census' lower frequency on collecting data – most countries either have a census every ten years or five years – and higher costs and time-consuming from a face-to-face model that results in a full enumeration of the country may not be sustainable for many of the countries.

According to the United Nations (UN), Census are among the most complex and massive exercises that a country undertakes. They require the mapping of the entire territory, the mobilisation and training of many professionals, the carrying out of a vast public campaign, the adhesion of the entire population, the collection of individual information, the compilation of large amounts of information and the analysis and dissemination of a vast amount of data.

The UN classifies the census models in three groups: i) traditional model that uses exhaustive surveys, short to long term questionnaires or rolling census (this last one is the case of France); ii) combined model which combines administrative data with surveys from a sampling or exhaustive approach to complement administrative data; and iii) administrative model that uses exclusively administrative data without conducting any specific survey (the case of Norway, Sweden and Finland).

In recent decades, there has been a trajectory of change in the census model used in many countries around the world, especially in the European Nordic countries and progressively extending to a considerable number of other countries. Census carried out using the traditional method (the case of Portugal) have been progressively and consistently replaced by a census model partially or totally based on administrative data (e.g., Germany, Poland, Spain and Italy). The change in the census model, on a global scale, aims at responding to society in a more efficient and effective way, allowing for a reduction of costs and the burden on respondents, ensuring an increased and more frequent dissemination of information meanwhile producing new relevant and reliable statistical data.

In the context of a Traditional or Register-based Census based exclusively on administrative records, without the use of surveys or even a Mixed Census (use of administrative data, complemented with results obtained by sample or exhaustive surveys) NSIs need to have an updated addresses list or dataset supported by and linked to an integrated geospatial database and common geographic references. This framework is essential to send online questionnaires and use some form of field follow-up for any non-respondents.

NSIs should be increasingly familiar with confidentiality requirements when collecting data using an address list for assuring that the identity of a respondent is protected from identification and exposure of personal information. In this regard, proceedings to avoid the disclosure of an address and protect the privacy of the respondents must be established and documented within a national and/or regional legal and policy framework specialised in data protection and confidentiality issues.

The survey sample is typically selected to represent some larger population of interest – the group of people or institutions that are the subject of research. Effective survey sampling requires that this population of interest be clearly defined, and well-established statistical methods are used to extrapolate the data resulting from the survey sample to the universe in the process of calculating the results. A multi-stage sampling design appears to be a classic solution wherein first primary units are selected, and then statistical units are selected within each primary unit that must be as small as possible.

In the classical approach, the sampling frame is not geocoded, and primary units are restricted to the smallest geographical units available (e.g., administrative areas or enumeration areas for Census purposes) or need to be beforehand manually constructed and require higher costs. In addition, the selection of a sampling scheme depends on the situation and purpose of an application.

Use of the location of statistical units in the sampling design, has gained more interest within the statistical community in recent years. Using a geocoded survey sampling frame is decisive for constructing the Primary Sampling Units (PSUs) in which this geospatial data can also be used at the selection stage to improve the statistical efficiency of the survey sample when the variables of interest are positively correlated. PSUs are a sub-division of the population often based on geographical criteria, in which firstly the PSUs are selected and then the individuals in those units. It is also appropriate for neighbourhood studies.

Spatial sampling appears to be a promising alternative to the classic sampling design and frame that could support statistical evidence of the survey sample and enrich its schema, ultimately improving the sampling design context. With a geocoded sampling frame, the PSUs can be constructed automatically at a lower cost and improve estimations of the population characteristics through deeper territorial knowledge about the sampling units, as well as other benefits to the overall sampling design process.

Statistical and geospatial data are obtained through sampling with scales of measurement, in which the scales of variation observable in spatial data are inextricably linked to the scales of measurement through which they were obtained. Since most environmental processes are scale-dependent, the observed spatial variation is likely to differ as the spatial scale of measurement varies. This means that there is a need to identify a sampling strategy that enables the identification of spatial variation of interest, enabling territorial flexibility at different geographic levels when collecting survey data. In addition, to facilitate the acquisition of suitable data and integration of data at different spatial scales or integration of different variables, the scaling properties of spatial variables should be used.

3. Context within the GSGF Europe and European statistical-geospatial operating environment

GSGF Europe and European statistical-geospatial operating environment

In the context of the GSGF[1] implementation in Europe[2], data collection activities are included in both Principle 1 (Use of fundamental geospatial infrastructure and geocoding of statistical information) and Principle 2 (Geocoded unit record data in a data management environment). Nevertheless, data acquisition procedures have effects on the feasibility of Principle 3 (Common geographies for production and dissemination of statistics), Principle 4 (Statistical and geospatial interoperability – Data, Standards and Processes) and Principle 5 (Accessible and usable geospatially enabled statistics). In general, specific requirements and recommendations for geospatial statistics based on survey data, should be focused mainly on geocoded data at the unit record level and on the geocoding process to have accurate location data for most statistical unit records. Hence, particular focus will be given to Principle 1 and Principle 2, making brief references to impacts and implications of other Principles.

Principle 1 aims that all statistical unit records should be collected or associated with a location reference through geocoding to produce geospatially enabled statistics, encompassing requirements and recommendations regarding data specification elements of common use. These data specification elements include the spatial and temporal schema, unique identifier management, object referencing and consistent coding systems, which are topics that are crucial components when collecting survey data towards a statistical-geospatial data environment

In this regard, the geocoding of statistical and administrative data should be incorporated during data collection procedures through direct or indirect capture of geographic coordinates or another type of precise location information from survey fieldwork, considering quality concerns on location or spatial precision.

The location information should therefore have a common and authoritative basis supported by the NSDI to ensure a consistent data structure when using authoritative and NSDI's compliant geospatial data and services for survey purposes. Up-to-date high-quality location information, including standardised physical address, property or building identifier, or other comprehensive location description through more accurate geographic coordinates and/or small geographic areas or standard grid reference to each statistical unit will promote a more fit-to-purpose approach in data collection.

[1] Global Statistical Geospatial Framework
[2] GSGF Implementation Guide, GEOSTAT 3 (2019)

In addition, having a constantly updated national addresses register linked to a geocoding infrastructure will also support the survey sampling strategy and design, ensuring a higher spatial resolution and location precision in the level of geography that determines the smallest geographic unit to report the data. Also concerning this Principle, the geospatial and statistical communities should work closely together and focus on defining common standards and consistent requirements for data collection methods and approaches that use geospatial data and services through the NSDI. Both communities need to achieve agreements at the lowest common geographic denominator for statistical data collection in a statistically valid manner, incorporating spatial autocorrelation principle and spatial characteristics and properties of data when designing the data collection methods.

Principle 2 proposes that the elementary units of statistical data (e.g., a person, household, business, building or parcel/unit of land) – microdata -, link to highly accurate geographic references or precise location information (e.g., geographic coordinates, small area codes, etc.) resulting in geocoded unit record data in a data management environment (prerequisite supported by Principle 1). The aim of Principle 2 is that statistical data can be used later in any geographic context through persistent storage of high precision geocodes (e.g., location data tables with references to administrative and statistical geographies), thereby minimising the risks derived from new geographies or changes to existing ones and ensuring timeliness of data.

By linking unit record data to highly precise geocodes at the unit record level, further data aggregation processes can be easily conducted for the administrative and statistical geographies by use of standard database tools and tabulation software. However, linking statistical and spatial objects at the unit record level must not compromise the privacy and confidentiality of microdata. That is why an efficient data management environment with data warehouse solutions within the data architecture could be a part of the solution to combine a widespread use of geocodes with confidence that prevents privacy breaches. Principle 2 is particularly addressed to the statistical and administrative communities but requires the participation and involvement of the geospatial community.

Geocoding plays a crucial role when designing data collection methods and sample frameworks for survey processes in which standardised and systematic approaches and guidelines must be formulated and implemented in the geocoding activities to ensure the same results regardless of the person or the institution conducting the survey. Thereby, consistent frameworks for location data and related technical guidelines should be applied in national practices within the statistical production process by the NSIs, with close cooperation with the National Mapping and Cadastral Agencies (NMCAs) and administrative data providers.

The geospatial community has a relevant role in three main tasks: i) design location data objects' full integration in the general data architecture of NSDI to facilitate workflows for data integration and geocoding; ii) build geocodes storage and location data repositories (geocoding databases) holding timely references to several common administrative and statistical geographies in a harmonised and usable way for non-geospatial experts; and iii) develop and provide geocoding services to fully support the use of life-cycle attributes and versioning through metadata procedures[3].

NMCAs should assist NSIs concerning those frameworks for location data making available various types of location to be used for various cases, assisting when locations are missing (for instance, using interpolation techniques), improving matching between records, and helping implement metadata procedures related to versioning and historical changes in geographies. In the long run, it is relevant to ensure that geocoding metadata is stored at object level. In other words, NMCAs should ensure the quality of geocoding and geocoded data to make sure that data collection methods for statistical purposes conducted by NSIs are effective, repeatable, and methodologically valid.

On the other hand, administrative data providers should take into consideration the established geocoding guidelines and follow the frameworks for location data as well as geocoding metadata specifications. Likewise, these concerns should prompt the workflows and avoid additional work by the NSIs when collecting and analysing survey data. In addition, these concerns need to be addressed at the point object level and in a long-term perspective since surveys (and other data collection methods) are increasingly being conducted for each observation with greater spatial accuracy of the assigned location and with a shorter period of frequency (usually determined by factors such as country's constitution, legislation, practice, and specific needs). Ensuring fully standardised and operating geocoding workflows for data collection and survey needs will reduce costs of production which sometimes affects plans for the extent and scope of data which is important in integrating statistical and geospatial data.

In general terms, location information of the statistical units is very useful to organise the fieldwork for face-to-face surveys (for instance, minimising travel time and costs), produce more accurate estimations (for instance, using spatial sampling methods) make the data collection process more efficient and enable small area estimations or flexible output geographies other than the initial one.

[3] See "INSPIRE Data Specification on Address – Technical Guidelines" (2014), INSPIRE Knowledge Base, European Commission (EC)

In the survey sampling context, precise location information of all statistical units allows the creation of PSUs in a systematic and repeatable approach, fitting several purposes of different surveys. Also, having more detailed location information of the statistical units will facilitate the interviewer's work field management while keeping the statistical qualities of the sampling survey. Also, during the collect phase, having an accurate location or geographic references for the statistical units sampled will make them easier to be properly identified and further georeferenced in digital format and enable more accurate spatial analysis.

In summary, the GEOSTAT 2 report presented three recommendations on building a geocoded sampling frame: i) constructing geographical primary units, preferably in an automatic way as opposed to the classic solution of multi-stage sampling design; ii) improving the sampling design ensuring better estimations of the population characteristics and deeper knowledge of the sampling units; and iii) from a theoretical point of view, disseminating statistics on any small area even if the accuracy of such estimations might be very poor. To address these recommendations, new statistical methods available with a point-based geocoding infrastructure need to be assessed, including spatial sampling methodologies[4].

The geospatial object for statistical content (e.g., building) has a value for survey data that goes beyond its location. Location of the statistical units may be used to improve the design of the sampling, for example, to ensure that the sample is spatially distributed well, for example through building density, or to address missing objects through spatial estimation methods (e.g., measurements in neighbouring points). These issues are more difficult to address when using an area-based approach related to traditional surveys and censuses where the survey data is assigned to a fixed output area, such as enumeration areas for statistical purposes. That is why a geocoding infrastructure supported by a point-based approach is essential to achieve a new paradigm in data collection and a long-term and sustainable data providing strategy despite encompassing some constraints and challenges related to legal, technical, organisational, and financial issues. Nevertheless, in the past few years, many NSIs have started to geocode their sampling frames according to geographical codes and by adding the precise or estimated spatial location of each record.

[4] See "Handbook of Spatial Analysis" (2018), INSEE

The use of high-quality detailed location information, preferably at point-based data updated with timestamps, supports survey operations and sampling design but entails some confidentiality constraints and spatial disclosure risks. The first ones are usually taking the form of adequate thresholds, and the second ones are related to the identity of the statistical individual and its attributes or sensitive variables.

Thereby, it is essential to maintain and ensure safeguarding and privacy concerning the personal information and geographic location of the survey respondents when collecting data meanwhile avoiding any reduction in the response rate. In this sense, further research in combined statistical and geospatial confidentiality and privacy protection must be conducted and efforts for an international legal framework supported by regulation of statistical-geospatial confidentiality should be made.

In addition, since the temporal and location information and time-space geographies play an increasingly crucial role in survey data, it is important to foster efficient strategies to manage the geospatial data streams from the technical infrastructure perspective, avoiding data duplication and developing routines that deal with temporal aspects of data to be geocoded. The instability of administrative geographies for further time-series analysis and the inconsistency of grid populations for comparison of populations are also two challenges that need to be considered and tackled when designing data collection operations. Also, improving geocoding quality assessment, such as address validation tools or interpolation of address location points, as well as concerning geospatial data obtained for statistical purposes is relevant to secure the quality of geospatial statistics. Although this recommendation is recognised, methods and procedures for geospatial data quality assessment are poorly described in guidelines for statistical production, otherwise being traditionally included in international standards or standards from other sectors (e.g., ISO 19157:2013 – Data quality from the International Organisation for Standardisation and other standards from the Open Geospatial Consortium).

GSBPM and GeoGSBPM context

The integration of geospatial aspects into the different phases of the statistical production model has shown to increase the value of the statistical information being produced and disseminated wherein statistical institutes and geospatial agencies have been moving towards an integrated production approach. In this regard, the Global Statistical Business Process Model (GSBPM) and the Geospatial View of the GSBPM - GeoGSBPM (released in 2021) constitute starting points to help the user on how to set up or enhance geospatial information, services, workflows and capabilities for their data collection methods, particularly related to development and maintenance of the point-based geocoding infrastructure.

In addition, the first version of quality indicators focused on surveys was added to the quality management process of the GSBPM in 2016, and a second version of the quality indicators for statistical processes based on surveys and administrative data sources was released in 2017.

Furthermore, it is important to highlight the meaningful contributions of the GeoGSBPM to statistical institutes on understanding and describing the geospatial-related activities using the framework of the GSBPM towards the production of geospatially enabled statistics consistently and systematically, and a greater capacity for data integration and interoperability.

The GeoGSBPM constitutes a significant development and methodological extension from a stand-alone approach wherein GSBPM is implemented in many NSIs but typically not applied to describe and implement processes involving geospatial data and services. Therefore, GeoGSBPM addresses the previously identified need to see where and how geospatial data might appear, or should appear, in various phases and sub-processes of the statistical production process and clarifies the role of geospatial data according to statistical domain and requirements, either as a data source or an outcome. In addition, the description of these geospatial-related actions and considerations despite identifying common activities and workflows required to produce geospatially enabled statistics also explores in a deeper way some geospatial considerations presented in the GSBPM, such as related to geocoding during the Collect phase.

The activities related to survey data are intrinsically operationalised in the Collect phase which occurs each time new statistical and/or geospatial data is obtained or collected and exclusively centred on information collection, including data and metadata. Nevertheless, this phase also requires outcomes of the sub-processes in the Design and Build phases. These phases are related to the setup and improvement of the geocoding infrastructure based on recommended systematic assessment regarding the data sources, technical conditions and human resources that sustain the infrastructure. They are extremely important when trying to construct frequent and stable data collection methods that support a regular production system requiring an industrial production setting and demonstrate how the geospatial dimension may be carried out in the same way or with a common tool in different processes. The following contents will describe the three GSBPM phases in which survey data concerns and geospatial-related activities and considerations should be considered:

GSBPM Phase 2. Design

The Design phase describes the development and design activities and is very important for data collection since it includes all the research work needed to define the collection methods and instruments and related operational processes, as well as possible candidate data sources recognition and assessment, as recommended to set up and/or improve the geocoding infrastructure. These data sources may include geocoded address data, building data, dwelling data and cadastral parcel data which provide very useful geocoding information and location data frameworks that support the collection processes. The sub-processes that are most relevant for survey data are the Design Collection (2.3 sub-process) and the Design Frame and Sample (2.4 sub-process).

The first sub-process determines the most appropriate collection instruments and methods which may depend on the type of data collection (e.g., census, sample survey, etc.), the collection unit type (e.g., person, business, etc.) and the available sources of data. This sub-process includes several activities according to the type of collection method which may require geospatial data and services, for instance, to support a web questionnaire, and direct or indirect use of administrative data for either controlling survey data or assisting the survey data collection operation. Also, since most NSI do not directly collect geospatial data this sub-process may include the design of mechanisms to monitor this type of data (or any other type of data not collected directly by statistical institutes) and metadata to assess the impact of any change made by the geospatial data producers and providers. This is particularly important for administrative units that change over time or other types of output geographies that usually have variations in their boundaries. That is why it is so important that NMCA, as the main producers and providers of authoritative geospatial data, should establish a close and ongoing institutional cooperation and technical agreement with NSIs.

Concerning geospatial aspects in this sub-process, geospatial data can be collected along with statistical data in different ways depending on the collection mode (e.g., sensors and GPS coordinates). When survey data is collected in the field, there are two possible scenarios to be considered: i) if the sampling frame is geocoded (e.g., geocoded address or geographic coordinates of dwellings) the interviewers already have the geospatial data of the statistical unit; and ii) if this type of geospatial data is not available or validated, either implemented in the sampling frame or have geometric inaccuracies, they need to be captured in the field operations employing GPS (sensor data) via a collection device or manually on a digital map by surveyors supported by a GIS service.

The second scenario is particularly important when collecting mobility survey data or any data that has a relevant geospatial dimension and needs to be registered continuously and automatically, in which the collection devices should record necessary metadata (e.g., timestamp and geodesic reference system). When survey data is collected alongside the geospatial data it is recommended to include a point-of-entry validation tool when designing the collection method, especially when using direct or indirect geocoded administrative data that could either control survey data or assist it during the collection.

The second sub-process only applies when data collection is based on sampling, for instance for social or business surveys, and mainly constitutes the sampling plan to be further followed. This sub-process includes activities such as identification and specification of the population of interest, the definition of the sampling frame and the most appropriate sampling criteria and methodology. It may include geospatial data and geographic classifications, combined with statistical and administrative data sources if needed. Geospatial data and services play a relevant role in designing frames and samples by geo-enabling the sampling activities.

The design of a geocoded sampling frame, to be further built supported by geospatial data and services, can help reduce survey costs, for instance, related to interviewers' routes by optimising them through routing services and network analysis for measuring the territorial coverage of each interviewer. It also ensures geographical representativeness in the sample, avoiding the selection of two, too different sampling areas based on geographic characteristics (e.g., urban areas *vs* rural areas) with inhomogeneous error, and increasing the efficiency of estimates in spatial analysis.

During the Design phase and closely related to the GSGF Principle 1 (Use of fundamental geospatial infrastructure and geocoding), geographies should be designed for the statistical unit level preferably using point-based location as the base geospatial variable and enabling a point-based geocoding infrastructure that will support the statistical production process. As previously mentioned, regarding required geospatial data and metadata for the design activities, this recommendation will also provide considerable adaptability to changes over time and territorial flexibility to aggregate up to various sampling and dissemination level geographies. In addition, when using grid geographies to design the frame and sample, the choice of the grid system should consider existing regional and global systems - in the case of the ESS Member States, the ETRS89 1km^2 grid cell.

National and international standards on geospatial data and metadata should also be considered in the design activities to reduce the duration and cost of the design process and further data collection activities and enhance the comparability and usability of outputs. This recommendation is relevant for all phases of the statistical production since standards on geospatial data and metadata may be quite different from those of the statistical community and considering both communities' standards on data and services will greatly ensure interoperability and usability of the statistical data.

GSBPM Phase 3. Build

The Build phase is related to the setup and configuration of the production environment, for instance, the sampling design to support the survey operation. In this phase, the infrastructure where the geospatial data and geocodes will be stored and organised will be developed or improved, preferably following a centralised model wherein an unlimited number of spatial references – unique identifiers - can be stored and kept separately from unit record data. This infrastructure setting contributes to a smother spatial reference maintenance process and allows more flexible and customized solutions when designing data collection methods and tools, mainly spatial sampling frames. It also facilitates the data aggregation and geographic classifications in the further Process phase and includes national and European output geographies.

In this phase, the sampling frame is built-in which many NSIs have recently started to geocode their sampling frame according to geographical codes by adding the exact or estimated geographic location of each record. Location information associated with the statistical units adds value to the survey sampling by helping to better organise the fieldwork and spread the selected units over the population surveyed ensuring more accurate estimations for more efficiency. Related to later phases, especially concerning Analyse and Disseminate phases, it may enable small areas estimations or flexible output geographies other than the initial one, bringing more territorial flexibility to the survey sampling frame.

Since geocoded data (e.g., addresses and geographic coordinates' buildings) are increasingly auxiliary to data collection methods and instruments, as well as collected and/or transferred during surveys, geospatial services are becoming critical building components. Geospatial services may include geospatial data quality management such as web-based applications to validate and verify addresses and georeferencing addresses lists into points, or innovative modes to receive survey data such as automatic location collection tools. These types of geo-solutions, increasingly in an open environment and free access, can be used by statistical institutions and other providers of geospatial data required to build the collection instruments, such as NMCAs and administrative data providers.

Also, these geo-solutions should be built to strengthen collaboration between the NSIs and the NMCAs and other relevant stakeholders, attending to their needs and capabilities and avoiding duplication of efforts towards a common building environment and collaborative strategy for the use of geospatial data and services in the statistical production process. In addition, when using the same application or tool to check the quality of geospatial data from the different data sources and providers a methodological consistency is assured as the same identified errors and inaccuracies will have the same quality measurement criteria and will be handled by a common approach.

GSBPM Phase 4. Collect

In the Collect phase, all necessary survey data (and metadata) are captured using different collection methods, planned and designed in the previous Design phase, and loaded into the appropriate environment, developed in the Build phase taking into consideration a long-term production, for further processing and analysing. The Collect phase also includes the strategy, planning and training activities on setting up the collection, which for survey data encompass several tasks concerning collection staff, assess the collection resources and configure the collection systems. These activities include preparing the collection strategy, in which geospatial data and services can help for instance, by optimising routes for data collection.

It is in this phase that location data (e.g., buildings data at a point or polygon format) associated to survey data (e.g., statistical variables) is acquired, usually from external producers such as NMCAsna and other local government bodies. Procedures for the acquisition of location data may vary in terms of the frequency and mode of data transfer, from manual and annual deliveries to automatic transfers at monthly, weekly or daily intervals with direct access to the source and cloud storage that promotes availability and easy access to existing data and metadata. During data collection geocoding procedures (assigning codes to statistical units related to geographical places, preferably by unique or standardised identifiers/geocodes) may also be conducted, preferably for each statistical unit and at the most detailed level such as point-based geocoding as opposed to area-based geocoding – as it was proposed in the GEOSTAT 2 final project report. Also, during field collection when capturing geospatial data and related inaccuracies are detected (e.g., positional inaccuracy or coding error), the identified inaccuracies should be documented and transferred to the central geospatial system or database for maintenance, update and validation. This process should preferably be conducted under statistical confidentiality terms assured by the specialised service on data protection within the NSI and/or according to the authority of the national or European data protection entity that controls and supervises the compliance of the national or European legal and regulatory provisions on the protection of personal data.

The fieldwork on geospatial data quality assessment made by the interviewers should be oriented by previously developed technical and methodological guidelines, for instance, when preparing the collection strategy or even earlier when designing a collection in the Design Phase. This preparation requirement will allow interviewers to identify, manage and report inaccuracies and errors and assure a standardised approach to data collection.

Additionally, GIS tools within the collection instruments used by the interviewers should be previously tested and configured appropriately for the specific collection method and considering extreme situations (e.g., technological breakdown due to high volume data being collected at the same time and considerable delays between expected and actual sign-off of collection systems and materials). To address these issues training sessions before data collection and testing series of collection operations in a specific or several survey areas are relevant requirements to assess the effectiveness of data collection and adequacy of the resources. These initiatives will also promote the benefits of geospatial data and services in data collection methods within the statistical institutes and give some basic geospatial expertise to the interviewers.

In the long run, these new forms of modernising the statistical production process will (and ideally should) mainstream some geospatial concepts and techniques strongly recognising geospatial terminology and methodologies in the statistical model as well as common tools and methods.

The spatial sampling frame at the microdata level also enables to connect and aggregate to other types of geographies for the definition of the geographical primary units and address data comparability purposes for surveys at target areas of different sizes (e.g., using grid cells as primary units to select a sample of dwellings in different sized neighbourhoods in terms of a social survey). The advantage of using grid cells as geographical primary units is to assure a consistent representation of the registers to be sampled regardless of the administrative division and independently from any future changes in the boundaries of the administrative geographies (LAU 1 and LAU2) and statistical geographies (enumerations areas and census blocks). This advantage is also relevant for data comparability across time series ensuring temporal stability of data avoiding errors from combination effects when data representing different times are combined.

Survey data, especially concerning fixed data collection such as Census data, represents a significant challenge to NSIs in terms of the temporal dimension of data to ensure the smaller gap between data collection, analysis and dissemination and therefore, timely data that respond to user needs. This is particularly important to tackle temporal uncertainty aspects of data related to their currency and timing that describes and measures the time gaps or periods between the event’s occurrence, data collection and finally their use and broadcasting.

For instance, Census data directly obtained from exhaustive surveys with regular and large time intervals are less current and temporally precise and relevant than administrative data that is continuously updated such as population registration. This is the most representative example of population data considering the type of data sources and related temporal advantages and limitations.

In a paradigm wherein most geospatial data are not explicitly temporal as geographic objects are rapidly changing between the time of collection and the time of database management and spatial analysis leading to errors, metadata guidance that manages the issue of temporality is crucial. Filling requirements for consistent time references in survey design (e.g., date fields in the survey questionary or the geospatial dataset) will help to describe and track the status of the collected survey data – with or without location data.

During the data entry, all time references of the input data should be checked ensuring a standard format, correct notation and right spelling as well as spatially validated (e.g., spatial accuracy measurement and geocoding assessment), and ideally only accept valid input data. These quality management procedures should also be conducted in more depth in the next Process phase when preparing the collected data for analyses.

These good practices on consistent historical recording and managing temporality allow validating further processing activities related to data (geographical) classification and data aggregation by using data's current version and extracting data for any desired point in time. It is also relevant to ensure timely data according to needs previously identified and further processing while guaranteeing updates in the appropriate time and therefore avoiding duplicate time references and geocodes or mismatches errors with damaging effects in further processing activities and results. In addition, this topic of survey data's temporal dimension presents higher complexity when data being collected are more temporally dynamic such as moving unit record data objects.

Spatial sampling and geo-solutions to survey data management

Geography has long been understood as a fundamental component of the statistical production process by statistical organisations, concerning geographical classification for designing sampling, a tool to support and plan field operations and processing input data. The modernisation and efficiency of statistical production is accomplished through methodological and technological developments, including new methodologies to define sampling design and frame that take advantage of the point-based geocoded data, such as buildings and dwellings. Spatial autocorrelation and spatial heterogeneity concepts, as well as other geographic-related concepts, should be considered in order to enrich survey sampling techniques, which are particularly important for statistical institutes whose data are often obtained through surveys and enable spatial analysis through more comprehensive data.

On the other hand, geospatial data, workflows and capabilities can provide more accurate location information and other geographic characteristics regarding PSUs and individual observations which may improve the design of spatial sampling and other data collection methods reducing costs without compromising the statistical validity.

The spatial sampling (also designated as geocoded sampling or geosampling) is typically used to estimate the total or mean for a parameter in a geographic area, to optimise parameters estimations for unsampled locations, including non-allocated observations, or to predict the location of a movable object, which is particularly relevant nowadays with increased movement of people and businesses (Wang et al., 2012). Spatial sampling differs from conventional sampling whereas this one is usually random and not repeatable assuming that data are independent and identically distributed.

Data collected from spatial sampling are usually spatially autocorrelated and heterogeneous considering spatial variation and defining the number of sampling units based on spatial autocorrelation and spatial heterogeneity principles. Thereby, considers that two statistical units that are spatially close to one another are more likely to look alike and follow the same pattern and behaviour (Tobler's First Law of Geography).

Selecting spatially distant units, increases the spatial coverage of the sample frame and will make the sample more representative of the population of interest and ultimately improve the accuracy of the estimates. Spatial heterogeneity is increasingly relevant for spatial analysis, especially when analysing environmental events, and its importance has been recognised in designing and evaluating sample designs, representing the more practical side of spatial statistics. Including the spatial dimension in the sample design with authoritative geospatial data and scientifically based, well-documented, and publicly available geospatial methods will address more effectively a certain number of non-allocated observations that typically remain undetected due to their geographic location characteristics, such as homeless people or people without a permanent place of dwelling.

The process of designing the spatial sampling may be supported by statistical datasets or files properly geocoded in a standardised way (e.g., dwellings for household surveys) and geospatial information, such as buildings (x,y geographic coordinates). The statistical information linked to the geospatial data, preferably at a point-based level, should be regularly updated through administrative data, which is why maintaining the fundamental point-based data infrastructure update is crucial to increase the efficiency of the spatial sample design process and overall data collection. Thereby, the process of designing the spatial sampling may rely on assigning geographic coordinates to statistical units such as dwellings, however other approaches could be taken.

That is the case of INSEE (France), which during the redesign of the georeferencing process of the tax source in the country, questioned the concept of point associated with a dwelling since it remains methodologically rather imprecise. In the past, INSEE used the centroids of cadastral parcels to geocode dwellings within a municipality. However, cadastral parcels can have high spatial variation being very spread out in space, which is why INSEE recently started to prefer to geocode by selecting the municipality in which the cadastral parcels spread out the most. It is thus a concept of "containing surface" or more commonly known as "spatial coverage" that INSEE retains as geolocating data of the dwellings in order to have a 1:1 spatial relation rather than a 1:n spatial relation. Furthermore, in some cases a building will be associated with several addresses, or several cadastral parcels and the spatial delimitation of these cases are sometimes difficult, but a conceptual definition of the location can be obtained by combining all the links between the geolocating information (parcels, addresses, points) available for the dwellings concerned. That is why, having detailed, multi-layered, multi-format and comparable location information supports the NSDI enabling geocode statistical units in multiple interconnected approaches and performing spatial sampling in a more accurate way possible.

Geographic units, such as 1 km^2 grid cells, should be defined as the reference for the selection of PSUs. Using spatial sampling design may allow reduction of the intra-cluster correlation coefficient (measures the similarity of statistical units) associated with selecting statistical units (e.g., dwellings) in "segments", and improve the accuracy of estimates based on the spatial autocorrelation principle applied to the variables. Additionally, as it was previously mentioned, spatial sampling design is very helpful in the case of face-to-face interviews, knowing the precise location of the statistical units sampled, making it easier to identify them in the field and to manage interviewers' locations during the fieldwork. This makes the overall data collection process more efficient by improving the logistics of operations, minimising travel costs and consequently increasing the time efficiency and productivity of each interviewer. In addition, when selecting the samples from point-based geocoded data (e.g., buildings and dwellings), the homogeneity requirement should be fulfilled in order to extract multiple subsets of data with the same homogenous error spatial distribution and uniform geographic characteristics. This requirement should be applied despite the size of the data, so that it does not affect the significance of the statistical quality control.

Geospatial tools may also be developed to increase efficiency in survey data collection management to respond to regular needs and challenges faced in survey operations such as difficulties from the interviewers in precisely locating their sampled statistical units in a specific survey as traditionally they only rely on tables or descriptive information (e.g., addresses, attributes, and other sensitive information).

Geospatial-based solutions that use GPS tracking systems and location-based services supported by mobile devices allow interviewers to easily identify the precise (x,y) location of dwellings and have access to associated data. Nevertheless, users' permissions constraints and confidentiality and privacy requirements must be considered and secured when developing these tools for restricting access to sensitive data and protecting data. Developing these types of tools and applications, mostly custom-designed according to specific needs and requirements, is a good example of integrating geospatial data into the official statistics' production model.

Implementing innovative geo-solutions to capture new relevant territorial variables and developing mix-mode data collection approaches that combine conventional survey data collection methods, such as CAWI and CAPI with geospatial data and services, takes advantage of the location dimension. Furthermore, when using open geospatial data and related outsourcing services (e.g., Google Maps, web geocoding tools or routing services) it is important to ensure quality criteria and assessment for statistical production, making it more difficult and limited when using commercial bases and products.

These types of non-official data sources and commercial products are increasingly relevant when designing new solutions for the statistical production processes. However, they also imply dependency on external services and operators where NSIs may have limited capacity for intervention and customised design, as well as being subject to changes that may directly and indirectly affect the processes and compromise their viability over time. Nevertheless, the convenience of having well-documented and certified official and authoritative geospatial data and tools for the statistical production process, including the collect phase and data collection methods, is clear.

4. Specific Recommendations

Based on the previously described contents, the following specific Recommendations are presented as an additional support to the GSGF Europe's Recommendations and may be useful to produce and disseminate geospatial statistics using survey data while geospatially enabling the statistical production process of the organisation. The proposed Recommendations are described and assigned to the Principle.

- The NSDIs should integrate a geocoded/location data repository built on relevant, authoritative location data and geocoding services to support statistical and administrative data collection activities in which NMCAs should promote the development and maintenance of the repository. This repository should include standardised identifiers/geocodes to link unit record data with location data.

This Recommendation may be included in the Principle 1 and an addition to the Recommendations 1.1.2, 1.1.3 and 1.1.4.

- Survey data and administrative data should be geocoded to the same reference and map projection system according to INSPIRE data specifications and metadata standards making this geospatial reference data mandatory at national and European levels. The data providers or custodians, such as public stakeholders at all government and administrative levels and other public data and administrative providers, should make sure that they have a consistent and systematic scheme of maintenance to keep the data updated and properly described. This recommendation constitutes an important aspect in the geocoding infrastructure setup, maintenance and improvement. Metadata should also include temporal characteristics of the data (timestamp), as a basic form of information, to enable routines that deal with the temporal aspects of data.

  This Recommendation may be included in the Principle 1 and Principle 2 and an addition to the Recommendations 1.2.2 and 2.5.2 respectively.

- Establish and disseminate national technical standards and formulate and implement guidelines for geocoded data infrastructure to develop a common and consistent approach on linking statistical or administrative data to geocodes and/or geographic references at the unit record level.

  This Recommendation may be included in the Principle 2 and an addition to the Recommendation 2.1.1.

- Apply an Open Data Policy through Statistical Laws Regulation and/or national geospatial data agreements for access to geocoded/location data and geocoding services, free-of-charge or for a small fee. This recommendation aims to ensure regular and long-term access to relevant and authoritative geospatial data, and consequently avoid duplication of data collection and inconsistencies in geocoding processes using the same reference and standardised spatial objects. This recommendation also recognises the relevance of administrative data providers, including private data providers, in the geocoding process and data management environment attending specific survey needs that sometimes, public bodies – statistical offices and geospatial agencies – do not respond to.

  This Recommendation may be included in the Principle 2 and an addition to the Recommendations 2.2.3 and 2.3.1.

- Develop geo-visualisation tools with authoritative geospatial data for survey purposes (e.g., reference basemaps) and make them available for data collection units and teams within the NSIs allowing statistical experts to visualise, consult and select the geographic scope (with associated geocodes and metadata) according to the population of interest or extent of the sample design. These tools with high-resolution data and tailor-made design should be provided to internal users of the NSIs and attend to the needs and specifications of the survey configuration, data collection methods and instruments, as well as survey sampling design and frame based on geospatial data and services.

  This Recommendation may be included in the Principle 2 and Principle 5 and an addition to the Recommendations 2.5.1, 5.2.2, 5.2.3 and 5.2.5 respectively.

5. References

Favre-Martinoz, C., Fontaine, M., Le Gleut, R. & Loonis, V. (2018). Spatial sampling. In V. Loonis and M.P Bellefon, Handbook of Spatial Analysis: Theory and Application with R. Montrouge: INSEE.

Wang, J. F., Stein, A., Gao, B. B., & Ge, Y. (2012). A review of spatial sampling. Spatial Statistics, 2(1), 1–14.