GeoStat

GSGF Europe:

A point-based foundation for statistics

Geostat 4

eurostat

Title: GSGF Europe: A point-based foundation for statistics.

Project: Eurostat ESSnet grant project GEOSTAT 4

Grant agreement number: 945503 - 2019-FI-GEOSTAT4

Author: GEOSTAT 4

Executive summary

This paper outlines the fundamentals of a point-based foundation for statistics. A point-based foundation for statistics is a technical and methodological framework to enable assignment of a precise geographical reference to a statistical observation. It is also known as a point-based geocoding infrastructure.

GSGF Europe assumes that a point-based geocoding infrastructure is far more flexible in terms of production and maintenance than a traditional area-based infrastructure with fixed output areas, such as enumeration areas or other small area geographies. A point-based geocoding infrastructure is also better equipped to integrate data in order to better exploit the spatial dimension of statistics (e.g., spatial analysis).

The paper also discusses constraints and challenges and related to establishment of a point-based foundation. Furthermore, the paper discusses approaches for implementing a point-based geocoding infrastructure. The content of this paper is a revised extract of the content from the GEOSTAT 2 final report.

Content

1. What is a point-based foundation for statistics?

In its most fundamental sense, a point-based foundation for statistics is a technical and methodological framework to enable assignment of a precise geographical reference to a statistical observation.

The process of geospatially enabling statistical unit records, or other non-spatial data, by adding x- and y- (and potentially z) coordinates[1] is generally known as "geocoding". More specifically, geocoding is the process of linking unreferenced location information (e.g., an address) associated with a statistical unit (e.g., individual, household, local unit etc.) to a set of coordinates. The resulting coordinates are the geocode.

In essence, a point-based foundation for statistics is a point-based *geocoding infrastructure*. The term *point-based* should be understood in contrast to *area-based*, which is an approach used in traditional surveys and census operations where the population surveyed is assigned to a fixed output *area*, such as an enumeration area. It should be stressed that the proposed shift from an area-based to a point-based approach, as described in this paper, only refers to the process of collection, geocoding and processing of statistics, not the process of dissemination. For dissemination of geospatial statistics, the use of area-based reference data (statistical or administrative geographies) is and will continue to be, the primary method.

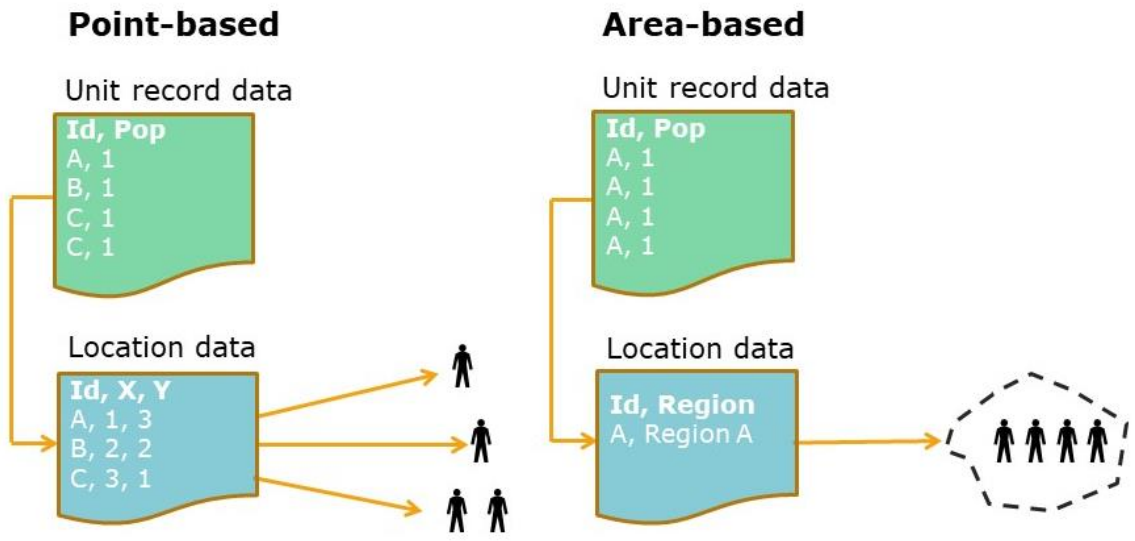

Figure 1: The conceptual difference between point-based and area-based geocoding infrastructures.

[1] x- and y- coordinates referring to a Latitude and Longitude or an Eastings and Northings, with the z- coordinate referring to elevation.

In Figure 1 above, the conceptual differences between a point-based and an area-based foundation for statistics are illustrated. In both cases, there is a record with statistical data comprising four individuals, and a corresponding record containing geospatial reference objects. In the point-based approach, shown on the left, each individual portrayed in the statistical record is linked to a unique dwelling location (id) represented by three discrete point locations. Two individuals have been assigned to the same location as they are linked to the same geospatial reference data object (dwelling id = C). In the area-based approach on the right, all four individuals are linked to the same geospatial reference object (Region A), as the area-based approach does not support spatial discrimination of their individual dwelling locations *within* Region A.

GSGF Europe assumes that a *point-based geocoding infrastructure* is far more flexible in terms of production and maintenance than a traditional area-based infrastructure with fixed output areas, such as enumeration areas or other small area geographies. A point-based geocoding infrastructure is also better equipped to integrate data in order to better exploit the spatial dimension of statistics (e.g., spatial analysis).

A point-based geocoding infrastructure does not presuppose a specific mode of data collection. It can be implemented in the context of traditional census data collection, as well as in an administrative data and register based context. However, as one of the key goals of the European Statistical System (ESS) is to better exploit new data sources for statistics, National Statistical Institutes (NSIs) should opt for a point-based infrastructure based on authoritative geospatial reference data, along with use of administrative data as this also allows easier integration with other data sources sharing the same location.

The Global Statistical Geospatial Framework (GSGF) recognises the geocoding infrastructure as a fundamental feature encapsulated in Principle 1. However, the GSGF does not explicitly require the use of a *point-based* geocoding infrastructure (UN-GGIM, 2021). Considering the substantial progress in the ESS over the last decade, in terms of accessibility of fundamental geospatial data and establishment of National Spatial Data Infrastructures (NSDIs), GSGF Europe goes beyond the global implementation guidance of the GSGF and recommends Member States to implement a *point-based geocoding infrastructure* if they have not already done so.

2. Constraints and challenges

In theory, a point-based foundation is a fairly simple and clear concept. However, in practice it may encompass several challenges. The first, and by far most important, precondition for the successful implementation of a point-based foundation is access to high-quality geospatial reference data, such as address or building data. Data sources qualifying for high quality point-based geospatial reference data exist in the majority of the ESS countries. However, this does not necessarily imply that these data are used as a point-based infrastructure to geocode statistical data. There are several reasons behind this:

- The geospatial reference data coverage may be incomplete (e.g., address or building data exist only in urban areas or in certain regions of the country) or is heterogeneous in terms of data models due to lack of national standards.
- Access to geospatial reference data is restricted for legal or financial reasons. Data can simply be too expensive.
- The quality of geospatial reference data is too poor or is outdated due to lack of maintenance.
- No consistent legal, technical, semantic and organisational framework for authoritative geospatial reference data exists. The role of authoritative data is crucial, as production of statistics needs to be able to rely on a long-term data provision strategy.

Altogether, these conditions restrain the development of a point-based infrastructure for statistics. At the same time, area-based census frameworks have been successfully used for decades in many countries. Changing the approach and systematically geocoding all census information to point-location requires substantial investments and can only be expected if the basic conditions are sound and safe.

The concept of a point-based geocoding infrastructure may not be fully embraced also for other reasons. Although high-quality geospatial reference data may exist, and is both sound and accessible, legal restrictions on collection and/or storing of non-aggregated statistical micro data may prevent establishment of infrastructures to geocode individuals to coordinate level, even in internal production databases of NSIs.

3. What is geospatial reference data?

From a statistical production point of view, it is necessary to distinguish between geospatial data needed as part of the *infrastructure for geocoding* and geospatial data needed to *create statistical content*. The GEOSTAT 2 project (GEOSTAT, 2017) developed a three-tier model to describe these different types of geospatial data in accordance with its purpose in the statistical production:

- Tier 1: geospatial data used *exclusively* for the purpose of geocoding, dissemination or display of statistical or other data. Such geospatial data is instrumental in the sense that it does not have any intrinsic value to statistics. It simply does not make sense to use this kind of data in statistical production unless it is integrated with other data. Examples of data in Tier 1 include address data, census enumeration areas, postal code areas, statistical grids or other statistical or administrative geographies.
- Tier 2: geospatial data which is used *both* to geocode, disseminate or display statistical or other data *and* to create statistical content. Typical information found in Tier 2 includes buildings, cadastral parcels and transport networks, but also new data sources such as traffic sensor information. The mixed purpose can be illustrated by building data, which is a key dataset for geocoding of census data, but it is also used to retrieve statistics on building density and building footprints, to assess the degree of urbanisation, etc. The geospatial object of a building has a value for statistics that goes beyond its role as a geospatial reference object.
- Tier 3: geospatial data which is used *only* to produce statistical content. This category of data cannot be used directly to geocode statistical or other data. As such, data in Tier 3 can be regarded as *complementary* to, and *independent from,* data in Tier 1 and 2. Examples of data found in Tier 3 include digital elevation models (DEMs), land use or land cover maps, topographic data, ortho photo or satellite imagery, or other products derived from earth observation data. Typically, data in Tier 3 need to be combined with data from tier 1 or 2 in order to be transformed into statistical information. The calculation of land area within a Nomenclature of territorial units for statistics (NUTS) region can serve as an accurate example. In this case authoritative data on land mass and topographic maps (Tier 3) are combined with a dataset containing NUTS regions (Tier 1).

In conclusion Tier 1 and 2 comprise *geospatial reference data* that are required to geospatially enable all relevant statistical information with the main purpose to geocode, integrate, disseminate or display (e.g., on maps) statistical or other information.

Point-based geospatial reference data forms a sub-section of the data sources mentioned in Tier 1 and 2.

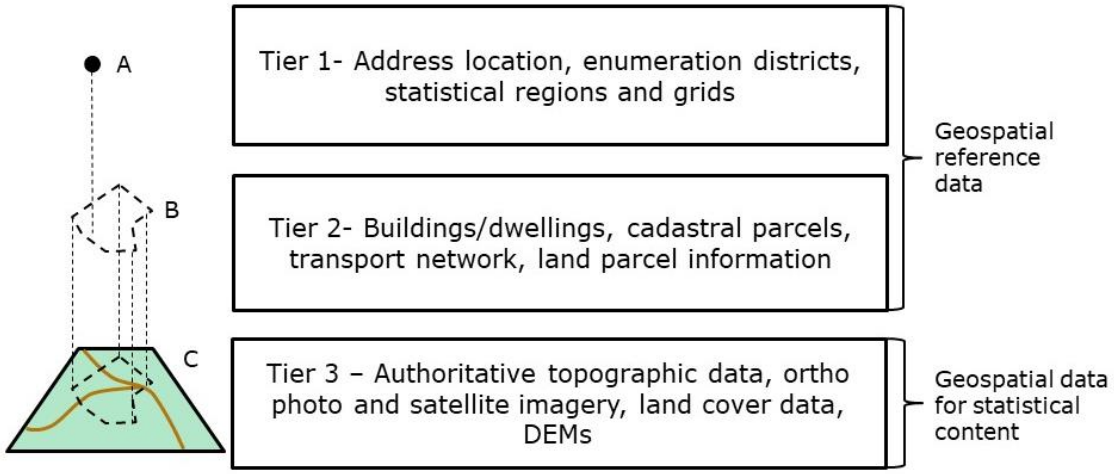

Figure 2: Tiers of data for production and dissemination of statistics. A workplace geocoded to an address location (A) can be linked to a cadastral parcel (B) from which land use can be assessed by combining the parcel with a land use map (C). The more consistent the system, the more opportunities for flexible linking of data.

4. Characteristics of a point-based geocoding infrastructure

The characteristics of a point-based geocoding infrastructure encompass the following characteristics:

- Use of high-quality point-based geospatial reference data (typically address or building registers), regularly updated and time-stamped
- Geocoding of statistical unit, and related statistical information, at unit record level
- Use of standardised and authoritative identifiers to link unit record data with geospatial reference data

High-quality point-based geospatial reference data should be understood as geospatial data that accurately represents the geographic location of a given phenomenon. The accurate point-based representation of an individual or a dwelling typically requires the use of address or building data. Depending on various traditions throughout Europe, the rationale for choosing one of the categories over another may vary between countries. The GEOSTAT 2 project concluded that it is of less importance whether the geocoding infrastructure is built on address data, building data or cadastral parcel information, as long as it can produce harmonised output with equal quality across countries (GEOSTAT 2, 2017).

Hence, the choice of geospatial reference data objects should rather be guided by the principles of authoritativeness, maturity and the prospect for long-term temporal maintenance of data sources. According to the GEOSTAT 2 survey, temporal accuracy and well-managed maintenance policies are rated even higher than the spatial accuracy of the geospatial reference data (GEOSTAT 2, 2016). Yet, topological and geometrical accuracy play an important role as well. As an example, the geographical location needs to be on the correct side of a street, and fall into the correct postal code areas, enumeration areas, electoral areas, etc. However, spatial accuracy within a couple of meters is typically good enough for most statistical purposes.

In some Member States, geospatial reference data frameworks comprise integrated combinations of address data, building or dwelling data and cadastral parcel data. Ideally, the objects in these frameworks are consistently and hierarchically linked to each other, enabling a flexible choice of the location data objects to be used, depending on the purpose of the task and the scope of output data. UN-GGIM: Europe has recognised address, building and cadastral parcel information as Core Data (UN-GGIM: Europe, 2017, 2018a, 2018b).

Geocoding of statistical information at the unit record level means that each statistical unit record included in a dataset should be assigned a high accuracy geocode, i.e. without previous data aggregation or grouping. The geocodes assigned to each statistical unit record need to match the address or building identifier found in the corresponding geospatial reference data. Also, records from different statistical or administrative data sources need to use the same identifiers to allow for cross-domain data integration.

The use of standardised identifiers to connect statistical information with location data means that the identifiers used should be based on a nationally and officially agreed system that is unambiguous. The latter requirement means that postal addresses are only the second-best choice as they are prone to inconsistencies, incompleteness, updates and duplicates. However not all countries have implemented national standards for properly managing address information.

Finally, it is worth noting that, in addition to the three spatial dimensions of the geocoding infrastructure (x, y, z), time as the fourth dimension is equally important to consistently and unambiguously geocode statistical unit record data. All features making up the infrastructure need to be time-aware and have a start- and end-date, or at least very good metadata to know when and how the geocode was derived.

5. Approaches for implementing a point-based geocoding infrastructure

The concept of a point-based geocoding infrastructure is generic and the underlying principles will be the same across countries, however the production setting may differ between countries due to various traditions in data collection and governance between the institutions involved.

Based on the survey conducted in the GEOSTAT 2 project (GEOSTAT 2, 2016), three different approaches were broadly identified among those countries that already have a point-based infrastructure in place. As further elaborated below, each approach has its own set of benefits and challenges in terms of performance and maintenance.

5.1 "In-house" – both location data and statistical data are collected and managed completely within the NSI

The "in-house" approach is a pragmatic solution used by NSIs when authoritative geospatial reference data does not exist, is not accessible or not fit-for-purpose. In such a context, the NSI needs to build its own location data in order to set up the infrastructure needed for a fully geocoded census. The benefit of this approach is that the NSI has full control of the entire data collection pipeline, including the coding systems (identifiers to link a location at the unit record level) and quality of both location data and statistical data. As the entire pipeline of data is in-house, the NSI does not have to deal with inconsistency problems resulting from changes to the address data standards, etc., at least as long as no third-party data, e.g. from tax authorities, needs to be geocoded.

The drawback of this approach is high data collection and maintenance costs as the NSI needs to bear the costs on its own. In addition, this approach may also imply the lack of synergies regarding geospatial reference data. Due to confidentiality legislations, the NSI may not be allowed to share location data with other public institutions and vice versa. As a result, each public institution responsible for data collection may need to create its own reference data. This limits the implementation of harmonised identifiers in records within public administrations, and eventually hampers data integration and/or the coherence for those datasets which are not produced by the NSI.

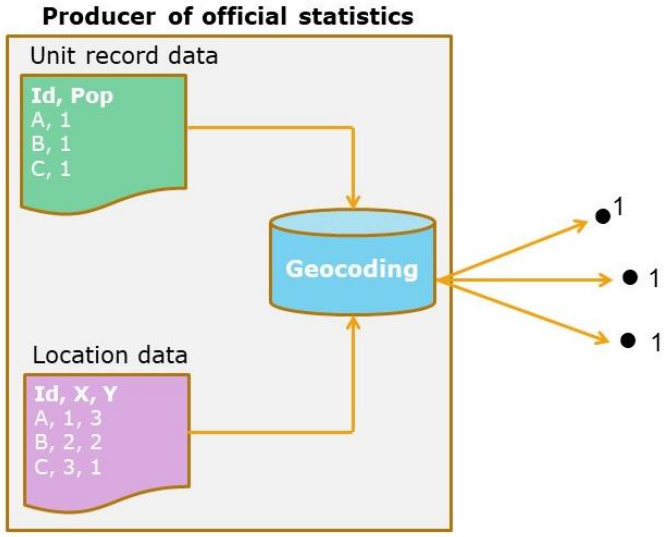

Figure 3: A conceptual illustration of the "in-house" approach

5.2 The "Hybrid" – location data is collected and managed outside the NSI and statistical data within

The "hybrid" approach is quite common in countries with a well-developed framework of authoritative geospatial reference data in combination with traditional census data collection modalities. Typically, the responsibility for the collection and maintenance of geospatial reference data rests with the National Geospatial Information Agency (NGIA) or with a consortium consisting of the NGIA and the regions and/or municipalities.

An obvious benefit of the "hybrid" approach is shared costs for the collection and maintenance of the geospatial reference data. Potential synergies in the use of location data constitutes another benefit. Where a framework of authoritative geospatial reference data exists, other public administrations are more prone to use the same data, which provides better prospects for an increased use of administrative data as the source for statistics and better data integration.

The challenge of this approach is that NSIs have to consider themselves one of many stakeholders to the national geospatial data policy. Policies regarding the maintenance of geospatial data will have a strong impact on the internal statistical production processes. NSIs need to spend more time on monitoring and interfering with policies regarding the geospatial reference data in order to safeguard consistent coding and data modelling. Typically, some kind of a formal agreement is required between the NSI and the producers of geospatial reference data to ensure long-term data access. If no such agreements are in place, there is a risk that changing business models or priorities among data providers will endanger timely access or consistency of geospatial information.

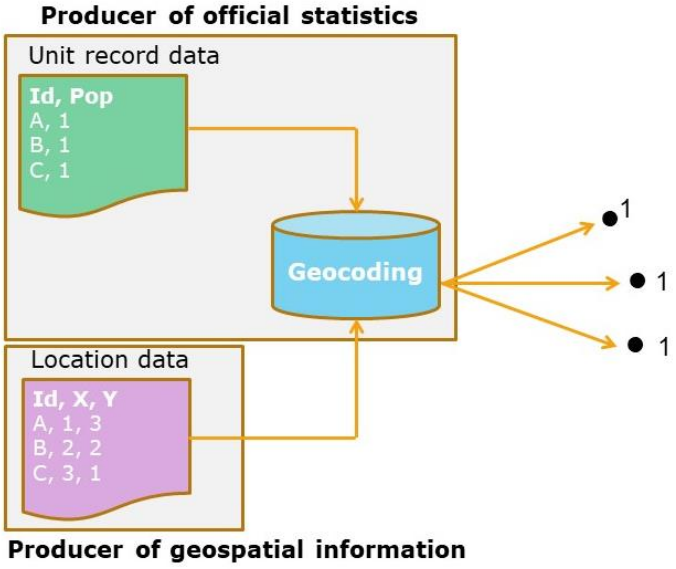

Figure 4: A conceptual illustration of the "hybrid" approach

5.3 The "Data broker" – both location data and sources for statistical data are collected and managed outside the NSI

The "Data broker" implies that the main approach of the NSI is to pull data together from a number of other data custodians. The NSI obtains geospatial reference data from NGIAs, or other providers of geospatial information, whereas data for statistical content are mainly obtained from other public administrations (tax administration, etc). Direct collection of data for statistics is at a minimum. Typically, access to administrative data is regulated in the statistical legislation, whereas access to geospatial data is governed by separate agreements.

As in the case of the "hybrid" approach, shared costs for data collection and maintenance are an obvious benefit of this approach. In the "Data broker" approach, shared costs also apply to data for statistical content, e.g. administrative data sources. There is also a great flexibility in terms of production as administrative data sources are typically updated monthly, weekly or even daily. The use of authoritative geospatial reference data with officially agreed identifiers forms the basis for registers/records in most public administrations and ensures consistency of all data, which creates almost unlimited data linkage opportunities.

The downside of the "data-broker" approach is no or little direct control of the data collection. As the collection of administrative data is not conducted primarily for statistical purposes, NSIs also have to deal with the fact that data may be structured in a way that requires substantial restructuring procedures to make it fit for statistical purposes. In this production mode, NSIs typically have to spend a considerable amount of time and resources to negotiate with other data custodians. This can be a cumbersome task as there may be legislative restraints on the collection of data which is not needed to perform the administrative task, for which it is collected, but which proves crucial for statistical purposes. In some cases, administrative data can come geocoded from the custodian. This can pose problems if the geocoding method is unknown or different to the methods used and expected by NSIs.

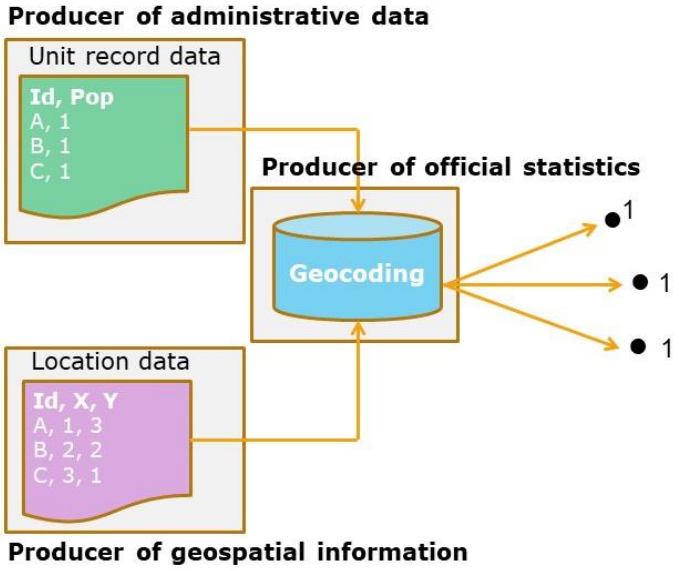

Figure 5: A conceptual illustration of the "data broker" approach.

In principle, all NSIs in the ESS can be assigned to some of the approaches broadly described above. Yet, the approaches are generalised as the mode of production may comprise elements from different approaches. Mixed use cases can also be identified where different approaches apply to the same NSI, depending on the statistical domain (e.g. social or business statistics), the task to be conducted and the information and specific infrastructure used for this purpose. Nevertheless, the three different approaches conceptualise some fundamental differences that need to be understood in order to provide the relevant guidance on how to build and maintain a point-based geocoding infrastructure.

6. Can and should all kinds of information be geocoded to point location?

Without questioning the fundamental benefits of a point-based geocoding infrastructure, a relevant issue to address is whether all information can or should be geocoded to a point location? Another relevant question is whether the official address and building location data is sufficient to build a complete point-based geocoding infrastructure. To what extent is there a need to geocode data to the level of point coordinates where address or building location data may not be appropriate or sufficient sources?

According to the respondents to the GEOSTAT 2 survey, with some few exceptions, a well-managed authoritative geospatial reference data infrastructure with address and/or building information as its backbone fits the needs for geocoding most of the information found in NSIs or in other public administrations. The exceptions mentioned by the respondents can be divided into two categories:

- Cases where the statistical object does not have a clear or easily discernible location (e.g. mobile or internet transactions, population in mobile households or certain types of transport statistics).
- Cases where the use of address or building locations will produce a non-appropriate spatial representation of the phenomena.

The latter category includes for example agricultural holdings, discharge points for water and pollution from industries. Agricultural holdings may be inappropriately geocoded when using address information if the address location refers to the dwelling of the farmer, rather than to the actual farm site of the holding. Typically, the farm site and the place of residence of the farmer coincide, but if the location of the farm site is different from the residence of the farmer, address geocoding of the agricultural holding may produce erroneous results.

Another problem with agricultural holdings is that they represent the site of the holding rather than the spatial envelope of the farmstead. Area features, such as the agricultural land that belongs to the holding, will be linked to a single point location. This may potentially cause erroneous outputs if, for example, administrative data on agricultural land linked to point locations is aggregated to grid cells, as a result of which the whole land will be assigned to the grid cell of the holding. The geocoding of agricultural holdings may therefore require alternative location strategies, using Land Parcel Identification System (LPIS) (European Court of Auditors (2016), or the collection of location data specifically for the purpose of geocoding the holdings.

In many countries, economic units or premises of industries or other enterprises can be geocoded by means of address data. In most cases, this will result in a decent spatial representation of the production site of the industry. However, the water discharge point of the industry may deviate significantly from the location of the production site, and as such cannot be properly represented by the address location. Using the address location may potentially cause erroneous outputs in the cases where water discharges should be aggregated from point-location to watersheds or river basin districts. Typically, locations of discharge points have to be retrieved from administrative registers on environmental permits or from monitoring systems.

In conclusion, all kinds of data have to be properly assessed with regard to the application purpose before geocoding. The possibility to assign data to a point location does not necessary mean that this option is advisable. The examples above demonstrate that using address locations to spatially represent an agricultural holding or industrial premises may be accurate in one context but not in another. This is the main reason why considerations regarding geospatial information must form an integral part of the design and production of statistics.

7. References

European Court of Auditors (2016). Special Report No 25. The Land Parcel Identification System.

GEOSTAT 2 (2016). Spatialising statistics in the ESS. Results from the 2015 GEOSTAT 2 survey on geocoding practices in European NSIs (https://www.efgs.info/wp-content/uploads/2017/03/GEOSTAT2-Spatialising-statistics-in-the-ESS-2015.pdf).

GEOSTAT 2 (2017). A Point-based Foundation for Statistics. Final report from the GEOSTAT 2 project (https://www.efgs.info/wp-content/uploads/2017/03/GEOSTAT2ReportMain.pdf).

UN-GGIM (2021). The Global Statistical Geospatial Framework: Draft Implementation Guide. GGIM 11th Session. (https://ggim.un.org/meetings/GGIM-committee/11th-Session/documents/GSGF%20-%20Implementation%20Guidance.pdf)

UN GGIM: Europe (2017). Core spatial data theme 'Cadastral parcels' – Recommendations for content – Final version 1.1 (https://un-ggim-europe.org/wp-content/uploads/2018/11/UN-GGIM-Europe_WGA_Recommandation_Content_CP-v1.1.pdf).

UN GGIM: Europe (2018a). Core spatial data theme 'Address' – Recommendations for content – Draft version 1.0 (https://un-ggim-europe.org/wp-content/uploads/2018/11/UN-GGIM-Europe_WGA_Recommendation_Content_AD-v1.0_0.pdf).

UN GGIM: Europe (2018b). Core spatial data theme 'Buildings' – Recommendations for content – Draft version 1.0 (https://un-ggim-europe.org/wp-content/uploads/2018/11/UN-GGIM-Europe_WGA_Recommandation_Content_theme-BU-v1.0.pdf).