The Global Statistical Geospatial Framework: Draft Implementation Guide

Summary

In its decision 10/106, UN-GGIM urged Member States to continue efforts towards the adoption and implementation of the Global Statistical Geospatial Framework (GSGF) and to support institutional coordination and collaboration between national statistical offices, national geospatial information agencies and other relevant stakeholders to support the ongoing implementation of the Framework, especially in the context of the global coronavirus disease (COVID-19) pandemic.

Further, the Committee also requested the EG-ISGI to continue the development of key statistical standards and processes that would strengthen the integration of statistical and geospatial information, to provide practical guidance in the production and use of integrated geospatial information, and to develop the interlinkages between the GSGF and the Integrated Geospatial Information Framework (IGIF) to further support the implementation and operationalization of both Frameworks, including through the regional commissions and the United Nations Global Geospatial Information Management regional committees.

To practically respond to this mandate, the EG-ISGI has developed this GSGF Implementation Guide to assist countries with implementing the GSGF and enable them to produce geospatially enabled statistical data for national to global decision-making.

Through its Task Team on Privacy and Confidentiality and its Task Team on Principles of the GSGF (composed of three Work Streams: Geocoding, Common Geographies, and Interoperability), the following guidance has been developed. The EG-ISGI aims to finalise this guidance in advance of the upcoming 53rd Session of the Statistical Commission in March 2022 and seeks the input of the global geospatial information management community as it works to finalise its implementation guidance.

This document aims to be an advanced draft that can be used by countries to implement and operationalise the GSGF. By highlighting the relevant GSGF Principle(s), the relevant sections elaborate the importance of relevant sections and identify key resources and further reading. Further, to further improve the understandability of the concepts discussed within this document, the GSGF and other relevant bodies of work, this document seeks to update and agree on common definitions, within its section on “the Terminology of the Integration of Statistical and Geospatial Information”.

Table of Contents

Implementing Geocoding

Relevant Principles of the Global Statistical Geospatial Framework

Principle 1: Use of **fundamental geospatial infrastructure** and **geocoding**
Principle 2: **Geocoding unit record data** in a data management environment

What is Geocoding?

Generally, people prefer to use descriptions of locations instead of coordinates to navigate their environment. As an example, for the delivery of goods, we supply an address instead of the coordinates of our doorstep. However, modern geospatial technologies depend on absolute position data coordinates within a specific reference system, rather than relative location descriptions. The process of bridging this divide is referred to as geocoding.

Geocoding is the method of linking a description of a location to the location's measurable position in space. Geocoding links unreferenced location information (e.g., an address, or other location description) associated with a statistical unit (e.g., housing unit or business) to a set of coordinates within a coordinate system[1]. These resulting coordinates are the geocode. More formally stated, geocoding is generally defined as the process of geospatially enabling statistical unit records or other nonspatial data (such as address lists or housing unit records) by creating x- and y- (and potentially z) coordinates[2] and linking them to each record. Once geocoding is performed on individual statistical unit records, they (or the associated data) can be aggregated into larger geographic units (e.g., states, provinces, or municipalities) for statistical analysis. The records are ready for further applications such as methodologies to ensure confidentiality and avoid data disclosure.

Geospatial science is a rapidly growing field of study, with an evolving body of geospatial terms and concepts as the technologies are adopted in a wide variety of applications. Geocoding is often referred to with the terms geoenabling, geolocalising, or simply "linking" in some implementations. Frequently geocoding is associated with, and is considered as, a subset of georeferencing. Georeferencing, in its broadest definition, is understood to be the process of linking geospatially enabled data to a common geospatial reference frame that allows geospatial presentation and analysis of those data, usually in Geographic Information System (GIS) software. Georeferencing requires linking coordinates to a defined geospatial reference frame (i.e. a geospatial datum, ellipsoid, coordinate system, and often a projection). Georeferencing may refer to the alignment of orthoimagery or digital copies of paper maps with their inherent geographic coordinates (i.e., geocodes); or the transformation of geospatial data from a one defined geospatial reference frame to another.

Why is Geocoding needed?

Appropriately geocoding statistical unit records to a specific geospatial location fosters the greatest opportunity to reuse and aggregate statistical data. The GSGF states that *"all statistical unit records should include or be linked to a precise geographic reference (an x- and y- coordinate), and if not, the*

[1] Also referred to as a spatial reference systems.
[2] x- and y- coordinates referring to a Latitude and Longitude or an Eastings and Northings, with the z- coordinate referring to elevation are the most commonly used, but other references are in use.

smallest geographic area possible". This recommendation for using an x- and y- coordinate for geocoding was first issued by the Expert Group in 2018[3] and is reiterated by the Expert Group in 2021. By geocoding each statistical unit record in a consistent, accurate, and precise manner, aggregation and disaggregation of their associated statistical data by geospatial location becomes possible. In this manner, the dissemination of statistical data using common geographic areas is enabled, promoting the reuse and comparability of data throughout time. Subsequent principles and key elements of the GSGF guide the dissemination of geospatially enabled statistical data in-line with prevailing privacy and confidentiality concerns, and national and international norms, standards, and policies regarding data disclosure.

How can records be geocoded?

Modern geocoding processes are largely automated, involving matching captured data with a reference database with some in-built spatial intelligence to improve the matching process. The efficiency of geocoding relies on having a comprehensive reference database of addresses and locations. This is a component of a mature national spatial infrastructure. Geocoding is also helped by having a standardised, structured description of a location. For example, a street address contains a number of specific elements with formatting requirements that are used in geocoding.

Geocodes can be generated directly (i.e., coordinates accepted as being specific for the statistical unit record) or indirectly when they use an internal point of a geographic area. Conceptually the most accurate geocodes are the x- and y- coordinates that were assigned to a statistical unit record at the time it was collected by using a Global Navigation Satellite System (GNSS), such as the Global Positioning System (GPS) or coordinates from the nationally agreed geodetic reference frame. Equally specific are geocodes assigned using specific standardised structure IDs or even within structure IDs (e.g., one apartment within an apartment building). The next most specific geocodes are for addresses or standardised parcel IDs.

If this level of precision is not possible, geocodes can be generated using an internal point (e.g., a centroid) for any functional area which represents a specific geography (e.g. an enumeration unit, a small building block geographic area, or a small area grid cell up to localities, postal code areas, or even second-level administrative units). Regardless of the geographic level used to geocode a statistical unit record, the manner of geocoding must be consistently documented for each statistical unit record in a dataset along with a corresponding record of a time and date for each record when each record was geocoded.

To identify available data to geocode statistical unit records or help identify their absence (and in turn, identify gaps where capacity can be developed), the 14 Global Fundamental Geospatial Data Themes[4] may be useful. These are 14 Themes considered fundamental to strengthening a national geospatial

[3] E/CN.3/2018/33 https://unstats.un.org/unsd/statcom/49th-session/documents/2018-33-GeoInfo-E.pdf
[4] E/C.20/2020/14/Add.1 http://ggim.un.org/meetings/GGIM-committee/9th-Session/documents/E_C.20_2020_14_GFGDT.pdf and http://ggim.un.org/documents/Fundamental%20Data%20Publication.pdf

information infrastructure. Specifically, the Global Geodetic Reference Frame (x- and y- coordinates), Addresses, and Functional Areas are directly relevant to geocode statistical unit records.

x- and y- coordinates

Addresses

Functional Areas

The Expert Group urges geocoding in the most accurate way possible to allow the most flexibility in combining various geocoded data and reiterates its previous recommendation to geocode statistical unit records with specific x- and y- coordinates. If this is not possible, it recommends geocoding (creating x- and y- coordinates) by using Addresses, or lastly by using the smallest Functional Areas possible.

Further Reading and Associated Resources

- The Global Statistical Geospatial Framework E/CN.3/2020/25 http://ggim.un.org/meetings/GGIM-committee/9th-Session/documents/The_GSGF.pdf
- Integrated Geospatial Information Framework Strategic Pathway 4: Data https://igif.un.org/
- Australian Bureau of Statistics: Geocoding Unit Record Data Using Address and Location: https://www.abs.gov.au/websitedbs/d3310114.nsf/home/Statistical+Spatial+Framework+Guidance+Material/$File/Geocoding+Unit+Record+Data.pdf
- Academic resources: (such as Texas A&M's geocoding resources https://geoservices.tamu.edu/Services/Geocode/
- The Global Fundamental Geospatial Data Themes: http://ggim.un.org/documents/Fundamental%20Data%20Publication.pdf

Implementing Common Geographies

Relevant GSGF Principles

Principle 3: **Common geographies** for the dissemination of statistics

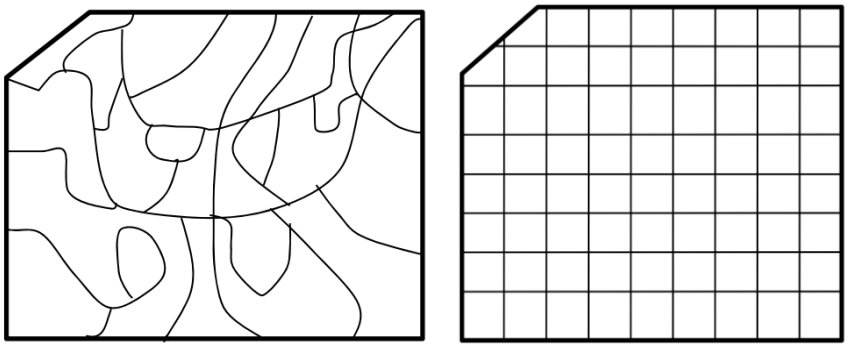
What are Common Geographies?

Common geographies are an agreed set of geographic areas for the display, storage, reporting, and analysis of social, economic and environmental comparisons across statistical datasets from different sources. They enable the production and dissemination of integrated statistics and geospatial information within a country to support informed decision-making.

Principle 3 of the GSGF recognizes and acknowledges the continuing need for country-specific dissemination geographies. New or proposed common dissemination geographies should be viewed as congruent and adjunct to the existing administrative and statistical geographies maintained by National Statistical Offices (NSOs), National Mapping Agencies (NMAs) and National Geospatial Information Authorities (NGIAs).

Why are Common Geographies needed?

The GSGF calls for *"a common set of geographies [to] ensure that statistical data is geospatially enabled in a consistent manner and is capable of being integrated at the aggregate level; and also ensures that users can discover, access, integrate, analyse, and visualise statistical information seamlessly into geographies of interest".*



Administrative Geographies Gridded Geographies

Figure 1 Administrative and Gridded Geographies

Further, through using consistent common geographies within the statistical production process ensures that statistical data is geospatially enabled in a consistent manner, whether in gridded form or using administrative or statistical boundaries. With this foundation, it is then possible to define methodologies that enable the transformation of geospatially enabled statistics amongst administrative, statistical and gridded geographies (as exemplified in Figure 4).

NSOs, NMAs and NGIAs that adopt a common dissemination geography are encouraged to move forward and begin producing social, economic and environmental data, indicators and other information

from the current and future integrated statistical geospatial infrastructure. Three objectives may be attained:

1. enhanced capacity to produce data and indicators for domestic purposes;
2. to meet emerging monitoring and reporting in support of thematic and global indicator framework requirements of international and regional initiatives (e.g. the 2020 Round of Population Censuses, the monitoring of and reporting on 2030 Sustainable Development Goals); and,
3. new emergent challenges that become immediate priorities for countries, and regional and international agencies.

Data produced by common dissemination geographies is a key facet to achieving basic comparable and interpretable statistical reporting, enabling advanced and exploratory geospatial analysis and a corner stone for producing clear and comparable data visualizations. This has become evident during the COVID-19 pandemic where the integration of statistical data with national common geographies was implemented to build web mapping tools and dashboards.

How can Common Geographies be realised?

What are the criteria needed for a country to adequately implement Principle 3?
Including:

- How to establish a geographic hierarchy?
    - Stocktaking of what geographies are currently being used? Who is maintaining them etc.? The IGIF Data Inventory Questionnaire / Dataset Profile Template[5]?
    - Identifying which existing geographies need to be used and new geographic areas created to support informed decision making for national priorities and global development agendas.
- Agreeing on the methodology for the translation of geospatially enabled statistical data from one form of geography to another?
    - Stocktaking of various methodologies and approaches (dependent on Geocode (if x- and y coordinate – point in polygon to capture the aggregation – if aggregated, is there a level of detail loss acceptable?

What sources of data can support the development of Common Geographies?

A common dissemination geography should be viewed as congruent and an adjunct to existing administrative and statistical geographies maintained by NSOs, NGIAs and other data stakeholders.

Existing national dissemination geographies, including administrative, electoral, census, and postal areas are often the foundation of common geographies in NSOs. Regional geographies such as the 1km^2 Population Grid within the European Union by Eurostat's GEOSTAT project or the Nomenclature of

[5] Appendices 4.2 and 4.3 of Strategic Pathway 4: Data of the Integrated Geospatial Information Framework https://ggim.un.org/IGIF/documents/SP4-Appendices-26Feb2020-GLOBAL-CONSULTATION.pdf

Territorial Units for Statistics (NUTS) 2 and NUTS 3 classification systems are the basis for regional and international comparison.

The nature of common geographies means that there are many potential stakeholders involved in their production, analysis or use including the National Statistical Office (NSO), National Geospatial Information Agency (NGIAs), international and regional organisations and other institutions (e.g. NGOs, the Open Geospatial Consortium (OGC) and the private sector etc.).

Further Reading and Associated Resources

- The Global Statistical Geospatial Framework https://unstats.un.org/unsd/statcom/51st-session/documents/The_GSGF-E.pdf, specifically "Comparison of the Advantages and Disadvantages of Administrative and Gridded Geographies"
- Defining the Global Statistical Geospatial Framework (prior work of the EG-ISGI): https://unstats.un.org/wiki/pages/viewpage.action?spaceKey=ISGI&title=United+Nations+Expert+Group+on+the+Integration+of+Statistical+and+Geospatial+Information&preview=/36143407/36143410/GSGF%20Principle%20Three%202018-07-25%20v2_EUgrids_added.docx
- The Integrated Geospatial Information Framework http://igif.un.org
- Tim Trainor, Challenges for Data Dissemination: Small Geographic Areas and Statistical Grids, http://ggim.un.org/meetings/2014-IGSI_Beijing/documents/04_USA_UN_Grid_Admin_Trainor_6_5_14.pdf
- Australian Bureau of Statistics, Location Index Social Architecture - Design of Institutional Arrangements https://doi.org/10.25919/5f32eab7c7d66
- Office for National Statistics, Hierarchical Representation of UK Statistical Geographies, https://geoportal.statistics.gov.uk/datasets/9c04ff58854040d09a5a7ce146ab59b4
- Eurostat, NUTS Classification https://ec.europa.eu/eurostat/web/nuts/background
- SDSN, Leaving No-One Off the Map: A Guide for Gridded Population Data for Sustainable Development https://bit.ly/35CHeqd
- OECD (2020), *Delineating Functional Areas in All Territories*, OECD Territorial Reviews, OECD Publishing, Paris https://doi.org/10.1787/07970966-en
- Eurostat et al, Applying the Degree of Urbanisation: A methodological manual to define cities, towns and rural areas for international comparisons – 2021 edition https://ec.europa.eu/eurostat/documents/3859598/12519999/KS-02-20-499-EN-N.pdf/0d412b58-046f-750b-0f48-7134f1a3a4c2?t=1615477801160

Fostering Interoperability

Relevant Principles of the Global Statistical Geospatial Framework

Principle 4: Statistical and geospatial interoperability: **Data, Standards, Processes, and Organisations**

The Importance of Interoperability

Interoperability is a crucial yet often underserved concept which touches every part of work within statistical, geospatial information, as well as the overarching pillars of the 2030 Agenda for Sustainable Development – Society, Economy, and the Environment. Interoperability concerns how data travels from the source to the end-user and as such is critical to the successful implementation of the GSGF.

As a result, interoperability issues in most cases cut across the other Principles of the GSGF rather than belonging to one Principle only. It is this cross-cutting, interlinked nature that this guidance aims to inform, highlighting resources that can support the ideals of Principle 4 and the broader GSGF.

Implementing Principle 4 is less prescriptive than the other Principles of the GSGF, and is more focused on fostering a conducive environment that enables the ability for statistical and geospatial data to be integrated, overcoming the (often) deep structural, semantic, and syntactic barriers between data and metadata from different communities and providers. This also improves the discovery, access, and use of geospatially enabled statistical data. The full implementation of interoperability described in this Principle is particularly important for Principle 5, as failure to achieve interoperability in any of the other Principles will often result in incomplete or less useful information for the end-user.

What is Interoperability?

Interoperability at a basic level ensures that different 'groups' (such as agencies within a national or global context, technology, or frameworks) can exchange data and share resources in the common interest. NSOs and NGIAs are augmented by administrative data custodians, which also act as providers of statistical data, but which are often not interoperable with statistics and geospatial information (e.g. why implementing a suitable set of common geographies is important).

There are several components to ensuring interoperability, but principally, leveraging and implementing open standards are key. Standards provide the critical architecture by which data can be discovered, collected, published, shared, stored, combined and applied and run through the four dimensions of interoperability – Legal, Organisational, Semantic, and Technical, as considered by the GSGF:

1. **Legal Interoperability** enables organisations operating under different national legal frameworks, policies and strategies to work together. National laws and policies should not block cooperation and there should be clear agreements about how to deal with differences in legislation across borders. As an example, national laws and policies on statistics should include the right of NSOs to have access to essential geospatial information with defined quality and ideally without charging;
2. **Organisational Interoperability** refers to the way in which public administrations (i.e. government agencies and organisations) align their business processes, responsibilities and

expectations to achieve commonly agreed goals. In practice organisational Interoperability means documenting and integrating or aligning business processes and relevant information exchanged. This also covers meeting the requirements from the user community and the NSS;

3. **Semantic Interoperability** ensures that the precise format and meaning of exchanged data and information is preserved and understood: "What is sent is understood". This includes syntactic aspects, such as the terminology used to describe concepts, as well describing the exact format of the information; and,
4. **Technical Interoperability** covers the linking systems and services of applications and infrastructures. Aspects include interface and services specifications, and data and metadata standards and formats.

Future Work by the Work Stream on Interoperability includes

The future work of the EG-ISGI in this area aims to demonstrate how there is no "one" way to foster Interoperability by:

- Developing guidance for developed and developing nations to promote and foster Interoperability to help implement the GSGF?
- Highlighting specific resources to advance development against for the four dimensions of interoperability;
- Examining Maturity Models and other key concepts to target and strengthen the capability of countries to foster interoperability and integrate statistical and geospatial information; and,
- Highlighting examples of Interoperability (COVID-19, needs of exchange of geospatially enabled statistical data etc.), within the National Examples of GSGF Implementation.

Further Reading and Associated Resources

- The Global Statistical Geospatial Framework E/CN.3/2020/25 http://ggim.un.org/meetings/GGIM-committee/9th-Session/documents/The_GSGF.pdf
- Integrated Geospatial Information Framework Strategic Pathway 4: Data https://igif.un.org/
- Integrated Geospatial Information Framework Strategic Pathway 6: Standards and Appendix *ibid* *and* https://ggim.un.org/IGIF/documents/SP6-Appendices-7Apr2020-GLOBAL-CONSULTATION.pdf
- The Global Fundamental Geospatial Data Themes: http://ggim.un.org/documents/Fundamental%20Data%20Publication.pdf
- Data Interoperability: A Practitioners Guide to Joining Up Data in the Development Sector https://unstats.un.org/wiki/display/InteropGuide/Home
- Geo Generic Statistical Business Process Model: https://statswiki.unece.org/display/GSBPM/GeoGSBPM
- Statistical Data and Metadata eXchange 3.0: https://sdmx.org/
- Connecting Geographic and Statistical Information Standards: http://ggim.un.org/meetings/2015-2nd_Mtg_EG-ISGI-Portugal/documents/Connecting%20Geographic%20and%20Statistical%20Information%20Standards%20EG-ISGI%202015.pdf
- A Guide to the Role of Standards in Geospatial Information Management:

https://ggim.un.org/meetings/GGIM-committee/8th-Session/documents/Standards_Guide_2018.pdf

- Companion document on Standards Recommendations by Tier Introduction: https://ggim.un.org/meetings/GGIM-committee/8th-Session/documents/Standards-by-Tier-2018.pdf

Ensuring Privacy and Confidentiality

Relevant GSGF Principles

Principle 2: **Geocoding unit record data** in a data management environment
Principle 5: **Accessible and usable geospatially enabled statistics**

Introduction

To meet the growing needs of geospatial and regional analysis, an ever increasing amount of precisely geocoded data is being created by the National Statistical Offices (NSO), in collaboration with National Geospatial Information (NGIAs) and National Mapping Agencies (NMAs) or other national, regional or global official bodies. This is in turn is leading to an increased amount of geospatial data (aggregate statistics released for geographic areas), often for small area geographies.

This wealth of this data ultimately leads to crucial challenges concerning data confidentiality, since the number of features necessary to uniquely identify a statistical unit (i.e. a person, household or business) decreases as the population of a cell or geographic area decreases within which information is released. This risk is even higher in an era of big data, artificial intelligence and proliferation of open access geographical visualization and analysis tools.

NSOs have been facing a conflict between two great principles of public statistics release (VanWey, et al. 2005). On one hand, they aim at offering as much data as possible with a high quality level, and on the other, they have to manage with strong constraints to guarantee and enforce the confidentiality of information providers. We live in an age where administrative and commercial third-party data can reveal more about individuals than national census and surveys, but we should also be aware of the tensions inherent where disclosure in some scenarios is in the public interest, for example in situations where personal safety becomes a more important concern to individuals and agencies than data safety.

At the global level, the EG-ISGI is well aware of these challenges. As a consequence, in its Work Plan 2020 – 2022[6] developed at its sixth meeting held in Manchester in November 2019, it agreed to establish guidelines and recommendations with which to address the emergent statistical and geospatial privacy and confidentiality issues. To achieve this goal, the EG-ISGI set up a task team whose work had to be, as much as possible, in line with the overarching Integrated Geospatial Information Framework (IGIF). Therefore, the Task Team, developed this guidance based on rich existing academic literature, good practices shared by the EG-ISGI, and, the results of the Global Survey on Readiness to Implement the GSGF.

This work is organized as follows: The first section provides a global overview on how to manage confidentiality and privacy within NSOs with any kind of data, including current and emerging methods; the second presents the theoretical or practical specific issues of geospatially enabled statistical data within this framework; the third discusses ideas on how to deal with these issues; and the last section provides conclusions and elaborates a list of recommendations.

[6] http://ggim.un.org/meetings/GGIM-committee/10th-Session/documents/EG-ISGI_Work%20Plan_2020-2022.pdf

This document does not aim to be a comprehensive technical handbook on statistical disclosure control methods for geospatially enabled statistical data. While being hopefully educational and easy-to-read, it nevertheless requires some basic statistical knowledge. Its main goal is to increase the level of awareness of the specific issues regarding the management of confidentiality in geospatially enabled statistical data, and to foster new initiatives to reduce, mitigate and raise-awareness of the inherent risks that emanate from breaches of confidentiality.

Contextualizing Privacy and Confidentiality

For NSOs, maintaining privacy is essential to retaining the trust of data providers and meeting the prevailing nationa legal obligations, whether it be households or businesses responding to a survey, or official bodies or companies making their administrative data available for official statistical purposes. In order to safeguard personal information and to preserve confidentiality, the statistical processes have to comply with global, regional or national regulations.

The constraints inherent with ensuring the confidentiality of a defined set of data usually consist in meeting a given threshold while releasing data. No information can be disclosed if it concerns less than a given number of statistical units. The thresholds depend on various parameters such as sparsity of the area, risk aversion and sensitivity of variables.

Disclosure occurs when an user uses released data to learn some information they do not already know. As such, the user is not a « hacker » who attempts to break into a security system. They process the data in order to find how to breach of confidentiality. Disclosure may also occur in innocent use of the data, for example when the data is very detailed and/or the person knows a lot about the population. Literature often identifies three kinds of data disclosure:

- **Identity disclosure**[7] refers to finding a direct identifier of a statistical unit from the data (for example, name or address);
- **Attribute disclosure** refers to revealing an association between a statistical unit and its sensitive features. For example, the user knows someone is living in an area, while the data show that all the inhabitants of this area share a common characteristic, such as income; and,

[7] Identification as a disclosure risk involves finding yourself or another individual or group within a table. Many NSOs will not consider that self-identification alone poses a disclosure risk. An individual that can recall their circumstances at the time of data collection will likely be able to deduct which cell in a published table their information contributes to. In other words, they will be able to identify themselves but only because they know what attributes were provided in the data collection, along with any other information about themselves which may assist in this detection. However, identification or self-identification can lead to the discovery of rareness, or even uniqueness, in the population of the statistic, which is something an individual might not have known about themselves before. This is most likely to occur where a cell has a small value, e.g. a 1, or where it becomes in effect a population of 1 through subtraction or deduction using other available information. Often, with a small cell, it may not be possible to find an attribute disclosure (learning something new about an individual) but the individual who has self-identified may perceive a risk that someone else might be able to find out something about them. Identification itself poses a relatively low disclosure risk, but its tendency to lead to other types of disclosure, together with the perception issues it raises makes several NSIs choose to protect against identification disclosure. See (Hundepool, et al. 2012)

- **Inferential disclosure** refers to inferring some attribute with a high confidence level, where increasing confidence levels is a desirable outcome for statistical data users.

To comply with regulations, one approach is to consider different kinds of users, an example is discussed in Figure 2, provided by New Zealand. General purpose users will only have access to less information (aggregation methods, masking), while specific audiences will have restricted access to more data in secure centers. A complementary approach is to introduce perturbation in the data in order to reach an acceptable level of disclosure risk. Applying a Statistical Disclosure Control (SDC) method is then a three-step process that consists in identifying the units at risk, processing them according to a given method and evaluating the reduction of data utility in exchange of more protection.

As an example from New Zealand's Statistics Act of 1975, Section 37(4)(b) states:

> (4) All statistical information published by the Statistician shall be arranged in such a manner as to prevent any particulars published from being identifiable by any person (other than the person by whom those particulars were supplied) as particulars relating to any particular person or undertaking, unless—
> (b) Their publication in that manner could not reasonably have been foreseen by the Statistician or any employee of the department.

Currently, this legislation is under review, in part due to advancements in technology over the past 35 years, but still raises a valid line of inquiry for other countries to examine their own national contexts: *Do other countries have this as a bottom line? What are the common threads throughout legislation? Do other countries' legislation put the onus of protection on the user rather than the producer of the data?*

Figure 2 New Zealand's Statistics Act, 1975

Nevertheless, traditionally, SDC methods do not take spatial features and spatial correlations or patterns into account. As such they may lead to a very distorted spatial pattern after processing.

The challenge of confidentiality in managing geospatially enabled statistical data

The disclosure risk is higher when considering geospatially enabled data

Using geospatially enabled statistical data can increase the risk of disclosure, because it is an even more strongly identifying information when it is collected in different variables for different purposes or places (for example, place of birth, place of residence, of study, of work), with time being another important covariate.

Economic geographers and social studies often highlight the strong spatial autocorrelation of this phenomenon. Perhaps this is best encapsulated by Tobler's First Law of Geography "Everything is related to everything else (Tobler 1970). But near things are more related than distant things", spatial autocorrelation then refers to the pattern in which observations from nearby locations are more likely to have similar features than that from distant locations. For areas with a low number of observations, that are often those where the density is weak, the risk of attribute disclosure is then higher. Tobler's

law is the critical contribution to statistical (table-based) confidentiality methods that we use to describe, measure, and represent geospatial confidentiality methods.

The disclosure risk also increases for geospatially enabled statistical data because of geographic differencing issues, that occur when the same data is disseminated in different non-nested geographies. In some cases, attributes can be deduced for several statistical units, below the threshold, by subtracting the counting of an area from the counting of another enclosing area. Therefore, anyone proficient with geographic information systems (and subtraction) could potentially uncover the underlying statistical data, inadvertently resulting in unintended disclosure [8] (Figure 3).

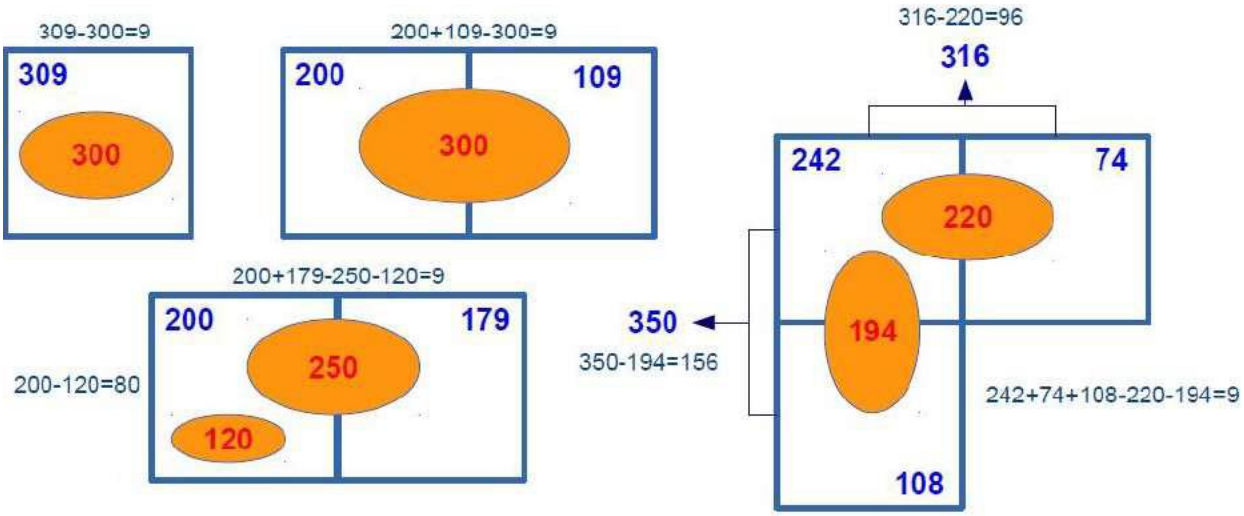


Figure 3 An exemplar of the possible cases of geographic differencing

Finally, trajectory data is a specific kind of geospatially enabled statistical data that can lead to specific breach of confidentiality. Indeed, beside Tobler's Law, a user may then use the uniqueness and regularity of human mobility to recover data without any prior knowledge.

A growing demand cascades to a growing risk

The theoretical risks identified in the previous section are even more pertinent due to the growing need for geospatially enabled statistical data. The integration of statistical and geospatial information relies on specific processes, as summarized by the GSGF, that require data sources such as address and cadastral parcels registers. These capacities, whether they already exist or are yet to be built, will make identity disclosure all the easier, if these datasets are released as open data or made available by the institutional that is nationally responsible for their production and dissemination. Other globally

[8] Figure 3 presents examples of possible cases of geographical differencing. Overlapping zones between the circles (A) and rectangles (B) are highlighted in orange. In the first case, zoning B encompasses zoning A. A user could then reconstruct information about B-A by subtraction and this may lead to revealing data concerning a small number of individuals. In the second box, the user can combine two areas of zoning B to perform the operation (B1 + B2) - A and thus obtain information concerning a non-released area. The last two cases show that differencing can occur with a combination of any number of the two zonings. With the frequency counts included in these examples, if information cannot be disseminated if it concerns less than 10 individuals, then there is a breach of confidentiality by geographical differentiation in each of these four examples

available mapping tools such as Google Earth or OpenStreetMap can also help provide greater context in the production and dissemination of statistical units, and for personal data should support the obfuscation of data.

Citizens are often most affected by decisions which influence their immediate neighborhood and this has resulted in governments/local authorities/political opponents increasingly seeking information at a very precise level of detail so they may analyse and illustrate the impact of various programs and policies. As a result, policymakers and analysts are looking for detailed information across a broad range of spatial dimensions, such as cities and/or rural areas, local administrative units and/or 1 km^2 grid cells, or below. The smaller size of the areas increases the risk of attribute disclosure, while the growing need of tailor made, or non-nested geographies increases the risk of geographic differencing issues.

With the prevalence of mobile devices, human mobility data are ubiquitously collected through cellular networks and mobile applications and publicly released for academic research, commercial purposes or official statistics leading to higher and new risk of breach of confidentiality.

A more difficult management of confidentiality with geospatially enabled statistical data?

A territorial classification used for the dissemination of data is nothing else than a categorical variable like any other. It is therefore possible, through the application of standarised methods, to deal with disclosure risk without any geographical consideration, simply by considering the geography as a variable with hundreds or even thousands of modalities (Finland, Germany).

Yet, a geographically intelligent management of disclosure issues will preserve underlying spatial phenomena, but no specific methodology has yet been developed yet. In practical terms, dealing with geospatially enabled statistical data adds a layer of complexity in the disclosure control process because it requires implementing specific methods that need great computing power. On micro-data, some SDC methods involve specifying the neighborhood structure with matrices, whose size can easily become unmanageable without using big-data techniques. On tabular data as well, detecting the risky observations by differencing sometimes requires the combination of many dimensions.

Dealing with the confidentiality aspects of geospatially enabled statistical data

Traditionally, in disclosure control literature, a distinction is made between post-tabular methods applied to tables (hypercubes) and pre-tabular methods applied on micro-data. Another way of categorizing protection methods is to classify them as either information reductive/non-perturbative or perturbative.

Census data, in practical terms, most countries adopt classical statistical non-perturbative/post-tabular methods, for example aggregating cells until sufficient thresholds are reached or suppressing cells. However, for gridded-data, data suppression is not an option; there must be a numeric value in each cell. These types of methods avoid issues of disclosure risks from geographic differencing and must be applied several times. Thus, this becomes very cumbersome when different geographies are used or when consistency is required between different linked tables. Moreover, post-tabular methods or removing valuable data within dataset can distort relationships between variables (Kamlet, Klepper and Frank 1985) and spatial correlations.

Perturbative methods appear to be a very attractive alternative solution. Firstly, pre-tabular methods of this class only have to be applied once, because if microdata is safe, so all possible aggregations from them will be safe too, and consistency is preserved. Secondly, they are more customisable and they allow a great flexibility of statistical products, both with grid data or hypercubes (they also permit tailored data for users). Another advantage of pre-tabular methods (like record swapping) can be unbiased, whereas most post-tabular methods involve suppressing cells and then introducing bias in the estimation of parameters or making some parameters impossible to estimable. Further, some post-tabular pertubative methods (like addition of random noise based on cell keys) are reductive.

In practice, the development of appropriate pre-tabular perturbative methods is challenging: a single perturbed microtable table file from which every table could be safely extracted, is not realistic, because for a given level of risk, the SDC expert will have to alter too many records (Young et al. 2009), which is not reasonable for an NSI. Moreover, pre-tabular methods can lead the users to believe that nothing is done to ensure confidentiality (Longhurst, et al. 2007), (N. Shlomo 2007), because applied alone, they can lead to releasing small cells for sensitive variables.

A classic compromise is to implement basic protection in the micro-data file, and then to add protection to tables when needed (Massell, Zayata and Funk 2006), (Hettiarachchi 2013). Post-tabular information reduction (e.g. suppression) methods are indeed applied under some conditions for output products, to reduce remaining disclosure risks assessed by threshold rules (such as minimum frequency, thresholds, the n;k rule, p% rule, etc.). For example, after perturbation on micro-data, cells that are still unique regarding a given variable will be suppressed. An alternative strategy could be a combination of protection afforded by slightly perturbed microdata with post-tabular noise addition as suggested in the context of the European Census 2021 (Giessing and Schulte-Nordholt, 2017). However, in cases where users have direct access to the micro-data file itself, there might be some risk of confidentiality breach of the data due to the absence of advanced protection on the micro data.

Acknowledging the specific issues of geospatially enabled statistical data

Before identifying and processing the data at risk because of their spatial features, a preliminary condition is to be aware that there are specific issues for this type of data. Different SDC initiatives can be undertaken for different target audiences.

In its Quality Assurance Framework, Eurostat directly refers to *addresses* as identifiers. The framework currently recommends that these *identifiers* are deleted from data as soon as possible, but this is being debated in refinements to the framework. In its work to develop a *statistical and geographical information confidentiality management policy*, Mexico also refers to address as identifiers but to the broader concept of geolocation as well. The latter encompasses other kinds of spatial information such as cadastral identifiers. The Australian Bureau of Statistics releases specific Guidance Material to protect privacy for Geospatially Enabled Statistics, recommends methods for de-identification of data and explores the specific aspects of geographic differencing.

Besides national statistical or privacy laws, data release policies, nationally agreed guidelines, national, regional or global quality assurance frameworks, or just acknowledged practices (not written anywhere),

the specific issues of geospatially enabled statistical data for the management of confidentiality have attracted a lot of interest among the scholars community or official bodies over last years (Brown 2003), (Curtis, Mills and Leitner 2006), (Domingo-Ferrer and Trujillo Rasua 2011), (Hundepool, et al. 2012), (Markkula 1999), (de Montjoye, et al. 2013), (Nagy 2015), (VanWey, et al. 2005), (Xu, et al. 2017). Nevertheless, despite this profusion, no reference handbook summarizing the studies carried out so far is available.

Apart from highlighting new risks and tackling some methodological issues, many of the latter studies mention the technologies challenges of producing geospatially enabled statistical data. They mainly point out that there is no standardised tools that would help implement the various methods, while their implementation requires very specific skills in many fields, such as statistics, geography, algorithms and/or coding optimization.

Identifying datasets, groups of units and units at risk

Adapting existing methods

Several risk metrics have been developed and discussed to evaluate the disclosure risk of an entire dataset. For example, a dataset will satisfy k-anonymity if, for each combination of quasi-identifiers, there are at least k observations. L-diversity is a broader approach that encompasses the latter and allows the consideration of intra-group diversity for sensitive variables. Both approaches may be used with geospatially enabled statistical data, since it is considered as a quasi-identifier or a sensitive variable.

Whether the data is spatial or non-spatial, one approach is to build the tabular data just as if it were disseminated without any constraint, and to flag risky cells as cells that do not satisfy the dissemination constraints. Risky units are then all the units inside risky cells. For grid data or small mesh data, risky areas can be flagged with these same rules, considering the mesh or the square as a dimension like any tabular dimension.

Another approach is to work directly on the micro-data. Each observation is has a probability of being re-identified by a data user. The underlying idea is that an observation is risky if it is not surrounded by similar observations. Conditionally to a list of quasi-identifiers, a score evaluates, for each record, how likely it is to find someone else sharing the same characteristics in the neighborhood. An individual alone in an empty area will always be considered as risky, but an elderly man located in an area with mainly young people will be risky as well. Ideally, such a score requires choosing a definition of distance or neighborhood between two records, and to build a huge matrix crossing all the units of the exhaustive data. For populous areas, this computation quickly encounters computing power issues. To solve this, an alternative is to base the risk measure on frequency counts of sensitive variables (Elliot, et al. 2005) . Another solution is to adopt a simpler definition of the neighborhood: belonging to a same area at a superior hierarchical level. That supposes to have a nested system of geographical levels (Nagy 2015)

Exploring new paths

Identifying units at risk with geospatially enabled statistical data has attracted the attention of scholars and researchers in recent years. The prevalence of human mobility data has led to raise the question of confidentiality. (de Montjoye, et al. 2013), proved that in a dataset where the location of an individual is

specified hourly and with a spatial resolution equal to that given by the carrier's antennas, four spatio-temporal points are enough to uniquely identify 95% of the individuals[9]. Using only already published aggregated mobility data, (Xu, et al. 2017) revealed an attack system able to recover users' trajectories with a 73%~91% accuracy at the scale of tens of thousands to hundreds of thousands users, which indicates severe privacy leakage in such datasets.

Another recent line of research aims at identifying the units at risk because of geodifferencing issues, when disseminating the same data according to two non-nested geographies. In algorithmic research (Costemalle 2019) develops an algorithm that efficiently allows to find such units even for non-nested territorial classifications having a huge number of items. The implementation of this method with the 35,000 French municipalities and more than 2 million grid cells have led to identifying 10,000 households at risk among the 30 million geolocated households, but adding a new dwelling to a statistical area can introduce a temporal differencing issue that should be considered.

Finally, recent research involving the confidentiality of locations when publishing smoothed density maps, shows that it is possible to retrieve the underlying location of the statistical units whenever the used parameters are published (Lee, Chun and Griffith 2019), (Wang, et al. 2019), (Hut, et al. 2020).

Processing data at risk

Adapting existing methods

Regarding questions of statistical confidentiality, geospatial data is not a fundamentally a new type of data, but rather aggregated statistical data that, in principle, can be examined and processed in accordance with the already existing post-tabular methods and procedures. Using standardized methods of analysis may nevertheless lead to the distortion of data as its aggregation will be undertaken without accounting for the distance between aggregated areas, in contradiction with the underlying spatial pattern as stated by Tobler's law.

To circumvent this problem, one possible way, when the number of risky areas is low, is to aggregate the areas on a case-by-case basis (Finland). When this option is not available, for grid data for example, an intelligent aggregation or disaggregation of cells is possible while preserving spatial correlation. The idea is to benefit from an existing or purpose-built hierarchy of geographies, to aggregate or impute cells at risk with other cells belonging to the same geographical unit at the previous level. The French and Finnish Statistical Offices (Costemalle 2019), (Markkula 1999) use such a strategy to release their data, along with other classical SDC methods. The Leibniz Insitute of Ecological Urban and Regional Development used a similar approach for Building stock visualization in Germany.

Swapping in general consists in exchanging the attributes of two observations. Targeted swapping targets the riskiest records of the data for exchanging attributes. Targeted record swapping (TRS) is a pre-tabular method that has been tested for some synthetic data from census data in Great Britain. In Japan, (Ito and Hoshino 2014) has also tested targeted swapping for the 2005 Census micro-data release. The main addition of TRS is to concentrate the swapping on the observations with the greatest risk of reidentification, defined at the level of a given geography, and to coerce swapped records not to

[9] https://www.nature.com/articles/srep01376

be too distant geographically. First versions of TRS were developed for hierarchical geographies (Brown 2003), or for grid-based data in Hungary (Nagy 2015). In these initiatives, two individuals cannot be swapped if they do not belong to the same area at a superior hierarchical level.

Exploring new paths

The growing need for spatial information often concerns non-statistical geographies, such as user defined geographies defined by x- and y-coordinates increasing the risk of disclosure due to geographic differencing. Statistics New-Zealand recently explored ways to provide enough protection when disseminating data, sometimes on-the-fly, for these areas. Among the three options studied, one is recommended since it better matches some prior constraints that include easy to use and capable of being automated within output tools.

Last and not least, computer scientists have developed the concept of "differential privacy", initially as a rigorous privacy or risk measure, along with differentially private (noisy) output mechanisms that are engineered to manifestly guarantee a given differential privacy level. It's adaptability to official statistics is expected to be a major topic in the future. Differential privacy is not a single tool, but rather a criterion, which many tools for analyzing sensitive personal information have been devised to satisfy. Differential privacy essentially ensures that using an individual's data will not reveal any personally identifiable information that is specific to them. "Specific" refers to information that cannot be inferred unless the individual's information is used in the analysis. There are several national examples of how countries use differential privacy, including:

- In the United States, the US Census Bureau is engaged in a major project that uses differential privacy. When preparing for the 2020 US census, the US Census Bureau performed verifications using 2010 census data, and investigated the potential of differential privacy as a way of maintaining data accuracy, while ensuring data security for statistical tables containing nationwide data such as gender, race, age, and relation to head-of household;
- In Japan, the possibility of adapting differential privacy for detailed geographical data from the Japanese census has been examined along with the potential of this set of methods as an anonymisation method for all statistical data.

Recommendations

- Increase the level of awareness of the unique and specific issues that come with managing the confidentiality of geospatially enabled statistical data at the global, regional and national levels;
- Acknowledge these specific issues in national statistical or privacy laws, data release policies, nationally agreed guidelines, national, regional or global quality assurance frameworks;
- Foster the collaboration with the scholar community and other official bodies in order to explore new paths and collaborate with academics and official bodies to develop geospatial-statistical disclosure control tools and techniques;
- Develop a handbooks of:
    - Statistical disclosure control methods for geospatially enabled statistical data, that will identify pros and cons of each approach and help countries set up their own policy; and,
    - Spatial-statistical disclosure control techniques that are applicable regardless of the national context.

- Increasing the security level in processing and disseminating geospatial data and fostering the integration of the geographic dimension within the methodologies and techniques of existing software for the management of confidentiality;
- The geospatial data dissemination policy must pay attention to differences in authorization between the government and general users in meeting statistical needs -- to ensure governmental bodies have access to certain geospatial information for the management of confidentiality in geospatially enabled statistical data; and,
- Develop the capacity and capabilities of NSOs in the domain of statistical disclosure control methods for geospatially enabled statistical data. This would include highlighting resources for agencies that produce geospatial information, capacity development organisations and other relevant stakeholders ensure the inclusion of the geospatial dimension in existing workshops and other capacity development materials on management of confidentiality within NSOs.

Works Cited

Brown, D. 2003. "Different approaches to disclosure control problems associated with geography." *Proceeding of the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality.*

Costemalle, V. 2019. "Detecting geographical differencing problems in the context of spatial data dissemination." *Statistical Journal of the IAOS* 4 (35): 559-568.

Curtis, A J, J W Mills, and M Leitner. 2006. "Spatial confidentiality and GIS: re-engineering mortality locations from published maps about Hurricane Katrina." *International Journal of Health Geographics* 1 (12): 1-12.

de Montjoye, Y A, J Quoidbach, F Robic, and A S Pentland. 2013. "Predicting personality using novel mobile phone-based metrics." *International conference on social computing, behavioral-cultural modeling, and prediction.* Berlin, Heidelberg: Springer. 48-55.

Domingo-Ferrer, J, J M J., Mateo-Sanz, and V Torra. 2001. "Comparing SDC methods for microdata on the basis of information loss and disclosure risk." 807-826.

Domingo-Ferrer, Josep , and Rolando Trujillo Rasua. 2011. "Anonymization of trajectory data." *Advanced Research in Data Privacy.*

Elliot, M, Manning A, K Mayes, J Gurd, and M Bane. 2005. "SUDA: A program for detecting special uniques." *Proceedings of the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality.* 353-362.

Hettiarachchi, Raja. 2013. "Data Confidentiality, Residual Disclosure and Risk Mitigation: Challenges in Managing the Demand for full Disclosure and the Need to Safeguard Restricted Information." Edited by UNECE and European Commission. Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality: Conference of European Statisticians/Eurostat. 1-7. https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2013/Topic_4_IMF.pdf.

Hundepool, Anco, Josep Domingo-Ferrer, Luisa Franconi, Sarah Giessing, Eric Schulte Nordholt, Keith Spicer, and Peter-Paul De Wolf. 2012. *Statistical disclosure control.* Chichester: Wiley Series in Survey Mthodology.

Hut, Douwe, Jasper Goseling, Marie-Colette van Lieshout, Peter-Paul de Wolf, and Edwin de Jonge. 2020. "Statistical disclosure control when publishing on thematic maps." *International Conference on Privacy in Statistical Databases.* Springer, Cham. 195-205.

Ito, S, and N Hoshino. 2014. "Data swapping as a more efficient tool to create anonymized census microdata in Japan." *In Privacy in Statistical Databases.* Eivissa: Springer. 1-14.

Kamlet, M S, S Klepper, and R G Frank. 1985. “Mixing micro and macro data: Statistical issues and implication for data collection and reporting.” *Proceedings of the 1985 Public Health Conference on Records and Statistics.*

Lee, M, Y Chun, and D A Griffith. 2019. “An evaluation of kernel smoothing to protect the confidentiality of individual locations.” *International Journal of Urban Sciences* 3 (23): 335-351.

Longhurst, J, C Young, Tromans N, and C Miller. 2007. “Statistical disclosure control for the 2011 UK census.” *Joint UNECE/Eurostat conference on Statistical Disclosure Control.* Manchester: UNECE/Eurostat.

Markkula, J. 1999. “Statistical disclosure control of small area statistics using local restricted imputation.” *Bulletin of the International Statistical Institute*, 52nd Session ed.: 50.

Massell, Paul, Laura Zayata, and Jeremy Funk. 2006. “Protecting the Confidentiality of Survey Tabular Data by Adding Noise to the Underlying Microdata: Application to the Commodity Flow Survey.” Edited by J Domingo-Ferrer and L Franconi. *Privacy in Statistical Databases. PSD 2006.* Heidelberg: Heidelberg. 304-317. doi:10.1007/11930242_26.

Nagy, Beata. 2015. “Targeted record swapping on grid-based statistics in Hungary.” *Submission for the 2015 IOAS Prize for Young Statisticians.*

Shlomo, N. 2005. “Assessment of statistical disclosure control methods for the 2001 UK Census.” *Monographs of official statistics* 141.

Shlomo, N, and J Marés. 2013. “Comparison of Perturbation Approaches for Spatial Outliers in Microdata.” *Conference of European Statisticians.* Ottawa: UNECE and Eurostat . 1-12. https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2013/Topic_2_ShlomoMares.pdf.

Shlomo, N, C Tudor, and P Groom. 2010. “Data swapping for protecting census tables.” *Privacy in Statistical Databases.* Berlin, Heidelberg: Springer. 41-51. doi:https://doi.org/10.1007/978-3-642-15838-4_4.

Shlomo, Natalie. 2007. “Statistical Disclosure Control Methods for Census Frequency Tables.” *International Statistical Review* 199/217.

Tobler, Waldo. 1970. “A computer movie simulating urban growth in the Detroit region.” *Economic Geography* (46): 234-240.

VanWey, Leah K., Ronald R. Rindfuss, Myron P. Gutmann, Barbara Entwisle, and Deborah L. Balk. 2005. “Confidentiality and spatially explicit data: Concerns and challenges.” *Proceedings of the National Academy of Sciences* no. 43 (102): 15337–15342. https://www.pnas.org/content/pnas/102/43/15337.full.pdf.

Wang, Zengli, Liu Lin, Hanlin Zhou, and Minxuan Lan . 2019. “How Is the Confidentiality of Crime Locations Affected by Parameters in Kernel Density Estimation?” *ISPRS International Journal of Geo-Information* 8 (12): 544.

Xu, F, Z Tu, Y Li, P Zhang, X Fu, and D Jin. 2017. “Trajectory recovery from ash: User privacy is not preserved in aggregated mobility data.” *Proceedings of the 26th international conference on world wide web.* Perth: International World Wide Web Conferences Steering Committee. 1241-1250.

The Terminology of the Integration of Statistical and Geospatial Information

The purpose of these definitions is to propose an agreed initial set of definitions of key concepts, to help share knowledge of existing terminologies and practices and align the description of concepts in order to reach a common understanding among representatives of statistical and geospatial communities. The terms and their definitions identified are based on their usage from the reports of UN-GGIM, the GSGF and other prevailing resources used by both the statistical and geospatial communities. The EG-ISGI's WS on Interoperability undertook a review of terms, elaborating the below terminologies, building on previous work of the EG-ISGI[10]. These aim to be living definitions, updated and refined accordingly, when appropriate.

Index

[10] A proposal for a common statistical-geospatial terminology database, 2nd meeting of the EG-ISGI, Lisbon Portugal, May 2015: http://ggim.un.org/meetings/2015-2nd_Mtg_EG-ISGI-Portugal/documents/UN-GGIM%20EG%20Lisbon%20meeting%20session%204%20background%20paper%20terminology.pdf

Frameworks, Themes and Global Agendas

2030 Agenda for Sustainable Development

The 2030 Agenda for Sustainable Development[11] aims to be a plan of action for people, planet and prosperity. It also seeks to strengthen universal peace in larger freedom and recognises that eradicating poverty in all its forms and dimensions, including extreme poverty, is the greatest global challenge and an indispensable requirement for sustainable development. All countries and all stakeholders, acting in collaborative partnership, are called to implement the 2030 Agenda. The Agenda resolves to free the human race from the tyranny of poverty and want and to heal and secure our planet, calling for the bold and transformative steps which are urgently needed to shift the world onto a sustainable and resilient path. Supported by five pillars of People, Planet, Prosperity, Peace and Partnership the Agenda is anchored by a pledge that no one will be left behind.

Global Statistical Geospatial Framework

The Global Statistical Geospatial Framework (GSGF) facilitates the integration of statistical and geospatial information. A Framework for the world, the GSGF enables a range of data to be integrated from both statistical and geospatial communities. Through the application of its five Principles and supporting key elements, the GSGF permits the production of harmonised and standardised geospatially enabled statistical data. The resulting data can then be integrated with statistical, geospatial, and other information to inform and facilitate data-driven and evidence-based decision making to support local, sub-national, national, regional, and global development priorities and agendas, such as the 2020 Round of Population and Housing Censuses and the 2030 Agenda for Sustainable Development.

The GSGF was developed by the United Nations Expert Group on the Integration of Statistical and Geospatial Information (EG-ISGI), to inform and report to both the UN-GGIM and UNSC (as subsidiary bodies of the Economic and Social Council – ECOSOC), with the mandate to develop an international statistical geospatial framework due to consensus for an urgent need for a mechanism to facilitate consistent production and integration approaches for geo-statistical information. The GSGF is the culmination of this work, and now the EG-ISGI moves towards developing material that enables the promotion and awareness-raising of the GSGF to enable adoption at national, regional, and the global level.

Integrated Geospatial Information Framework

The Integrated Geospatial Information Framework (IGIF) provides a basis and guide for developing, integrating, strengthening and maximising geospatial information management and related resources in all countries. It will assist countries in bridging the geospatial digital divide, secure socio-economic prosperity, and support countries with the overarching goal of the 2030 Agenda for Sustainable Development, in enabling action to leave no one behind.

[11] A/RES/70/1 https://www.un.org/ga/search/view_doc.asp?symbol=A/RES/70/1&Lang=E

The IGIF comprises three parts as separate, but connected, documents: Part 1 is an Overarching Strategic Framework; Part 2 is an Implementation Guide; and, Part 3 is a Country-level Action Plan. The three parts comprise a comprehensive Integrated Geospatial Information Framework that serves a country's needs in addressing economic, social and environmental factors; which depend on location information in a continually changing world. The Implementation Guide communicates to the user what is needed to establish, implement, strengthen, improve, and maintain a national geospatial information management system and capability.

The IGIF focuses on location information that is integrated with any other meaningful data to solve societal and environmental problems, acts as a catalyst for economic growth and opportunity, and to understand and take benefit from national development priorities and the SDGs. In providing the fundamental geospatial infrastructure for a country, the IGIF provides countries with a framework that anchors a previous work to develop NSDIs and standards, technologies, policies, best practices, amongst other key elements to enable the provision of geospatial information with a country.

Global Fundamental Geospatial Data Themes

The 14 global fundamental geospatial data themes are a foundation to support global geospatial information management, notably used to support the integrated geospatial information framework, among other global initiatives to strengthen geospatial information. They are the fundamental data sets and minimum primary sets of data that cannot be derived from other data sets, and that are required to spatially represent phenomena, objects, or themes important for the realisation of economic, social, and environmental benefits consistently across local, national, sub-regional, regional and global levels.

Sustainable Development Goals

The 2030 Agenda is composed of 17 Sustainable Development Goals and 169 targets, with which to build on the progress achieved by the Millennium Development Goals and complete what was not achieved. They seek to realise the human rights of all and achieve gender equality and the empowerment of all women and girls. They are integrated and indivisible and balance the three dimensions of sustainable development: Economic, Social and Environmental.

Core Concepts

Common Geography

Related concept *Geospatially enabled statistical data*

A common geography are an agreed set of geographies for the display, storage, reporting, and analysis of social, economic and environmental comparisons across statistical datasets from different sources. They enable the production and dissemination of integrated statistics and geospatial information within a country to support informed decision-making.

Data Management Environment

A data management environment holistically encompasses the tools, storage, and environment for acquiring, validating, storing, protecting, and processing required data to ensure the accessibility, reliability, and timeliness of the data for its users.

Discrete Global Grid System (DGGS)

The Discrete Global Grid System (DGGS) represent the Earth as a hierarchy of equal area cells with progressively finer geographic resolution. Individual observations can be assigned to a cell corresponding to both the position and size (or uncertainty) of the phenomenon being observed. DGGSs provide significant benefits when encoding, scaling, threading, streaming, combining, and analysing geospatially enabled statistical data.

Geographic Feature

A geographic feature is the geometric representation of a feature; this could be a physical feature such as a unit record, a dwelling, or property or a functional area such as an administrative boundary or an economic area.

Geographic Location

Related concept: ***Place***

A geographic location describes the geographic features and their relationship to other features and associated statistical information; and can be presented in many forms and mediums including maps, satellite imagery, aerial photography, and even sophisticated, interactive and highly visual dashboards. Unlike its corollary and related concept **Place**, the concept of Geographic Location enables the establishment and measurement of accuracy and precision, whereas **Place** does not.

Geospatial Information

Synonym: geographic data, geospatial data

Geospatial Information provides the digital connection between a geographic place, location, its people and their activities, and is used to illustrate what is happening – where, how and why.

Geospatially enabled statistical data

Synonyms: location-based statistics, geospatially enabled statistics, geographically referenced statistics, location-enabled statistics, small area statistics, statistical geography, spatial statistics, geo-statistical
Related concept: Geospatial enabling

Geospatially enabling statistics provides more information and capacity to generate and use knowledge than just statistics alone. Geospatially enabled statistical data enables the geographic breakdown of an area of interest, on which statistics are collected or disseminated. An effective geospatially enabled statistical area is one which supports many uses and enables comparisons over time. They are often hierarchically nested to collect or disseminate geospatially enabled statistical data. The construction of geospatially enabled statistical data may be functional but also population or socio-economic driven.

Geographies defined by a set of rules or a methodology meant to represent a geographic concepts (e.g., metropolitan or core-based functional areas, labour market areas outside of metropolitan regions or areas, neighboorhoods, urban, rural, a rural to urban continuum). This type of geographic area is often termed statistical

Location and geographic extent are the main characteristics of geospatially enabled statistics. The geography used in geospatially enabled statistics should meet the users' perception of their area of interest, e.g. What is the situation within a neighbourhood or areas of interest, responsibility? Geospatially enabled statistics are used to answer questions from a geographic perspective, e.g. What is close? How many are within distance *x*? How many per surface area? Further, it is recommended that all statistical unit record data should be collected or associated with a location reference and that ideally, it should allow for geospatial coordinates with x- and y-values to be produced for each record.

Integrative Geographies

Geographies designed to integrate social, economic and environmental data without the requirements and limitations of administrative and statistical geographies (e.g., the grid-based approach proposed by the Discreet Global Grid System). The term integrative is used here to denote that this type of approach is not dependent on any legal or other framework.

Interoperability

Interoperability is the ability of a system to exchange and use information, enabled through the application of open standards.

Legal/Administrative Geographies

Geographies defined in law, regulations or constitution. This type of geographic area is often termed administrative.

Location information

Location information can include addresses, property or building identifiers, and other location descriptions, such as enumeration geographies and other standardised and non-standardised, e.g. village names or other geographic names.

National Spatial Data Infrastructure

Related and core conceptual framework: Integrated Geospatial Information Framework

A National Spatial Data Infrastructure (NSDI) identifies technology, policies, standards, good practices, and human resources necessary to acquire, process, store, disseminate, and analyse the use of geospatial information. The NSDI concept has been replaced by the IGIF as the overarching framework for strengthening geospatial information.

Place

Related concept: Geographic Location

Citizens, communities, business sectors, governments, and other stakeholders benefit daily, and often unknowingly, from the use of geospatial information and related location-based services. These groups understand their physical location as their **place,** which is often described through a geographical name or some other vague or fuzzy concept which lacks the capability for accuracy and precision to be measured.

Statistical Area

A unit of measurement used for the dissemination or collection of statistics.

Statistical Geography (Geo-Statistical)

***Related concept** Geospatially enabled statistical data*

Geographies are defined by a set of rules or a methodology meant to represent a geographic concept (e.g., metropolitan or core-based functional areas, labour market areas outside of metropolitan regions or areas, neighbourhoods, urban, rural, a rural to urban continuum).

Statistical Unit Records

Statistical Unit Records can include persons, households and living quarters, businesses, buildings or parcels and units of land.

Reproducibility

Reproducibility or reliability is the degree of stability of the data when the measurement is repeated under similar conditions.

Definition of Data Integration Practices

Integration of Statistical and Geospatial Information

Core conceptual framework: Global Statistical Geospatial Framework

Synonym: integrated geospatial information

The integration of statistical and geospatial information describes the use of geospatial information for the production and dissemination of statistical data, leading to geospatially enabled statistical data. Integration can occur at any stage of the statistical production process. The GGSF is a principles-based framework that guides countries with the production of geospatially enabled statistics, noting that many, if not all, statistical phenomena are connected to a geographic location.

Linking

Related term: Linked Data

Linking defines a process of connecting structured data sources using a system of unique identifiers. Linking builds upon standard Web technologies such as HTTP, RDF and URIs. While integration describes the process of combining data from different thematic communities from a conceptual viewpoint, linking refers to technically connecting data in a machine-to-machine environment irrespective of the subject.

Geographic Differencing

Geographic differencing is the process where the same data is obtained for two different, but overlapping geographic areas, where the data from the smaller of these regions is subtracted from the data for the larger region. By using this method, it is possible to obtain data for the area that is not common to both regions; however, obtaining data for small areas using this method may result in a risk to privacy or confidentiality.

Geocoding and Georeferencing

Geocode

Geocodes are, preferably, fine-scale geospatially referenced objects that are stored as a geometry data type, such as location coordinates, e.g. x-, y-, and z-coordinates, or small area geographies, e.g. mesh blocks, block faces or similar small building block geographies. Larger geographic units, such as enumeration geographies, can be used as geocodes where finer scale geospatial units are not available.

Geocoding

Geocoding is the method of linking a description of a location to the location's measurable position in space. Geocoding links unreferenced location information (e.g., an address, or other location description) associated with a statistical unit (e.g., housing unit or business) to a set of coordinates within a coordinate system (also referred to as a spatial reference systems). These resulting coordinates are the geocode. More formally stated, geocoding is generally defined as the process of geospatially enabling statistical unit records or other nonspatial data (such as address lists or housing unit records) by creating x- and y- (and potentially z) coordinates and linking them to each record (x- and y-coordinates referring to a Latitude and Longitude or an Eastings and Northings, with the z- coordinate referring to elevation are the most commonly used, but other references are in use). Once geocoding is performed on individual statistical unit records, they (or the associated data) can be aggregated into larger geographic units (e.g., states, provinces, or municipalities) for statistical analysis. The records are ready for further applications such as methodologies to ensure confidentiality and avoid data disclosure.

Georeferencing

Georeferencing, in its broadest definition, is understood to be the process of linking geospatially enabled data to a common geospatial reference frame that allows geospatial presentation and analysis of those data, usually in Geographic Information System (GIS) software. Georeferencing requires linking coordinates to a defined geospatial reference frame (i.e. a geospatial datum, ellipsoid, coordinate system, and often a projection). Georeferencing may refer to the alignment of Orthoimagery or digital copies of paper maps with their inherent geographic coordinates (i.e., geocodes); or the transformation of geospatial data from a one defined geospatial reference frame to another.

Geospatial enabling

Synonym: location enabling

Geospatial enabling describes the process of taking location information such as an address or administrative area code and linking this information to a geospatial feature. The geocodes (e.g. location coordinates, address ids, or geographic areas codes), obtained from this process can be stored directly on the statistical unit record or linked in some way to the record. Unless geographical coordinates can be stored with the unit record, linking via key relationships is safer to avoid the changing geographies disrupt the time series.

Location Information

Location information can include addresses, property or building identifiers, as well as other location descriptions, such as enumeration geographies and other standardised (e.g. postal codes) and non-standardised (e.g. village names) textual descriptions of a location.

Aggregated statistical information

Aggregated statistical information is aggregated from geocoded unit record level data into the dissemination geography instead of disaggregated statistical information created using a spatial distribution model and larger statistical geographies as source data.

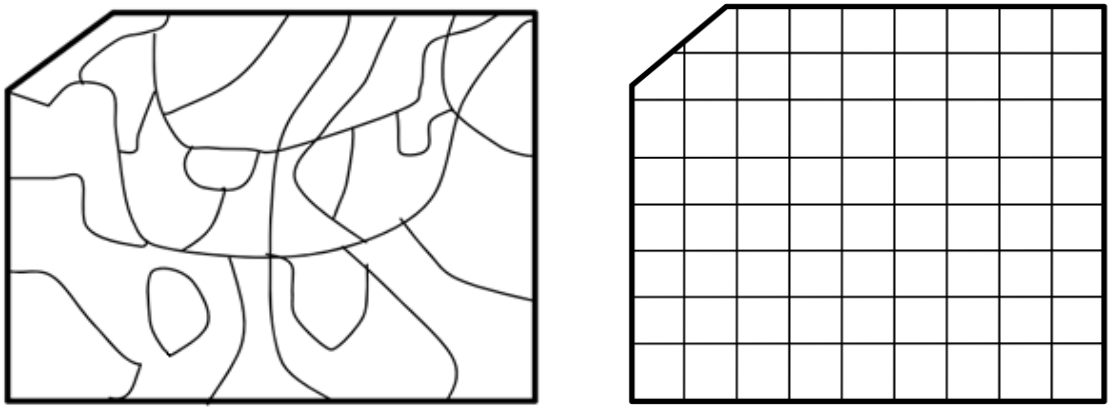
Geographical Classifications

Geographic Classification

Synonyms: Sub-national typologies, regional typologies, territorial typologies

Related concept: Gridded geographies

Geographical classifications are a method to group geographies according to objective criteria. The GSGF considers two main classifications, Administrative and Gridded Geographies respectively, from which other geographies are derived. The resulting geographies are characterised by how they are geographically represented. The GSGF, in its Annex B, considers, compares and contrasts administrative and gridded geographies.



Administrative Geographies

Gridded Geographies

Figure 4 Administrative and Gridded Geographies

Administrative Geographies

Synomym: enumeration geography, functional geographies, functional geography

Administrative geographies are primarily the geographic representation of the administrative boundaries of a country. The largest administrative subdivision of a country is called the "first-level administrative level", and often the smallest areas of measurement are enumeration areas. Enumeration geography is the division of a country into areas for census purposes. They represent the smallest area for which in most countries' population information is available. However, in certain countries, enumeration areas are further subdivided into blocks, e.g. bounded by physical features such as streets or rivers.

Functional geographies are defined by characteristics other than their surface area or administrative level. These include geographical characteristics such as mountain areas; social characteristics such as less-favoured areas, areas in need for development, areas by type of economic activity etc.

Gridded geographies

Related concept:

Administrative Geographies

Gridded geographies are of a consistent size, identified with a unique geocode and independent to the underlying geography.

Degree of Urbanisation

The Degree of Urbanisation (DEGURBA) is a classification of municipalities based on population densities and urban clusters. Based on the share of the local population living in urban clusters and urban centres, it classifies municipalities into three types of area: thinly populated area (rural area); intermediate density area (towns and suburbs/small urban area), and densely populated area (cities/large urban area).

Dissemination Geography

Synonym output system, output areas

System of often hierarchically nested geographies to be either particularly suitable for analysis (administrative geographies, gridded geographies).

Locality

A locality is a term used by different people to mean different areas, and assumptions should not be made about the term in any given usage. An increasingly important official use of the term is in connection with the census. A locality in this sense is a contiguous built-up area use for settlement reaching a minimum population threshold.

Spatial Analysis

Spatial Analysis

Synonym: location analytics

The process of examining the locations, attributes, and relationships of spatial features in spatial information through overlay, distances, spatial selection, intersection, aggregation and other analytical techniques to address a question or gain useful knowledge. Spatial Analysis extracts or creates new information from geospatial information.

Supporting Resources

Integrated Statistical and Geospatial Resources

- The European Forum for Geography and Statistics (EFGS): https://www.efgs.info/information-base/introduction/terminology/
  nb. The work of the EFGS is primarily based on work presented by Eurostat to the EG-ISGI's second meeting in Lisbon on 24 May 2015: http://ggim.un.org/meetings/2015-2nd_Mtg_EG-ISGI-Portugal/documents/UN-GGIM%20EG%20Lisbon%20meeting%20session%204%20background%20paper%20terminology.pdf
- The EG-ISGI's Wiki (for EG-ISGI members and Secretariat only): https://unstats.un.org/wiki/display/ISGI/Common+Statistical+and+Geospatial+Definitions

Geospatial Vocabularies

- ISO: https://unstats.un.org/wiki/download/attachments/36143402/Glosario%20de%20Terminos%20INEGI.pdf?version=1&modificationDate=1539014218647&api=v2
- OGC http://www.opengeospatial.org/ogc/glossary

Statistical Vocabularies:

- GSIM https://statswiki.unece.org/display/gsim/Generic+Statistical+Information+Model
- https://www.efgs.info/information-base/introduction/terminology/EUROSTAT http://ec.europa.eu/eurostat/statistics-explained/index.php?title=Thematic_glossaries#Special-topic_glossaries
- OECD https://stats.oecd.org/glossary/
- NQAF (National Quality Assurance Framework) https://unstats.un.org/unsd/dnss/docs-nqaf/NQAF%20GLOSSARY.pdf
- EFGS: https://www.efgs.info/information-base/introduction/terminology/

Other Resources

- The UNdata Glossary: http://data.un.org/Glossary.aspx
- International Statistical Institute: https://www.isi-web.org/index.php/publications/glossary-of-statistical-terms
- Bank of International Settlements https://www.bis.org/statistics/glossary.htm
- SDMX https://sdmx.org/wp-content/uploads/SDMX_Glossary_Version_1_0_February_2016.docx
- FMI http://www.imf.org/external/np/exr/glossary/showTerm.asp
- World Bank http://databank.bancomundial.org/data/metadataglossary/all/series
- European Central Bank https://www.ecb.europa.eu/home/glossary/html/index.en.html