GeoStat

GSGF Europe:

Managing Confidentiality in Geospatial Statistics

GEOSTAT 4

eurostat

Content

1. Guidelines for taking confidentiality issues into account when integrating statistical and geographical information

Major guidelines specified by the UN Expert Group on the Integration of Statistical and Geospatial Information (UN EG-ISGI), aligned with the Global Statistical Geographical Framework (GSGF) principles and recommendations (UNSC/UN-GGIM, 2019), can be summarized as follows:

- Increase the level of awareness of spatial data's specificity for the management of confidentiality at the global, regional and national levels.
- Acknowledge this specificity 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.
- Define a policy on which disclosure control methods has to be applied on spatial data, according both to the features of data and to the dissemination targets.
- Include the monitoring of actual disclosure control in Quality Assurance Framework.

- Foster the integration of the spatial dimension within existing softwares for management of confidentiality.

2. Why geospatial data raise new confidentiality challenges

Geospatial data are to be considered as personal data, since addresses and geospatial coordinates allow the identification of individuals just as well as their names do. The issue here is still to "secur[e] sensitive and confidential data while ensuring appropriate access to meet user needs for analysis and decision making" (UNSC/UN-GGIM, 2019).

Geospatial data actually raise new challenges in terms of personal data protection, so that the disclosure risk is a great matter of concern when you deal with geocoded data. This is because:

- socio-economic data are not spatially independent (see 2.1);
- the growing demand for spatial information leads to dissemination of more and more data, at lower levels of geographic nomenclatures which might not interlock, without checking their consistency one with another: this raises a specific disclosure risk known as "geographic differencing" (see 2.2);
- data visualization by the means of maps makes data easier to read for the human eye; at the same time, open data and APIs make it easier for everyone to compute data on their own, cross a huge amount of data and even hack some information that was not meant to be disseminated;
- statistical disclosure control (SDC) techniques require higher computer capacities than usual because of the huge amount of data involved;
- the spatial consistency of the data disseminated must be preserved throughout the SDC process.

2.1 Clusters and personal data protection

Spatial and socio-economical characteristics are correlated, as Tobler's first law of geography "everything is usually related to all else but those which are near to each other are more related when compared to those that are further away" (Dempsey C., 2014).

This observation made by Waldo Tobler in 1970 aimed to describe the urban growth system with a "universal gravitation" model. This idea underlies the concept of spatial autocorrelation.

Data are spatially autocorrelated when many similar values are located near each other (positive autocorrelation) or, on the contrary, when very different results are found near each other (negative autocorrelation). The most commonly used indicator for spatial autocorrelation is Moran's I (see chapter 3 by Loonis, V., 2018).

When data is disseminated on an area showing a strong positive spatial autocorrelation, there is a disclosure risk, since the aggregated data are actually the same as the individual data. Controlling the disclosure risk while dealing with spatial data then begins with detecting clusters.

2.2 Geographic differencing

The disclosure risk also increases for spatial data because of geographic differencing issues, which occurs when the same data is disseminated using different, not interlocked geographical nomenclatures like, for instance, grids and administrative geographies (see Figure 1). If the difference between two areas belonging to two different nomenclatures is small, it may contain very few units, raising a disclosure risk for these units.

Trajectory data are more complex to deal with, as the disclosure risk must be tackled for both departure and arrival points.

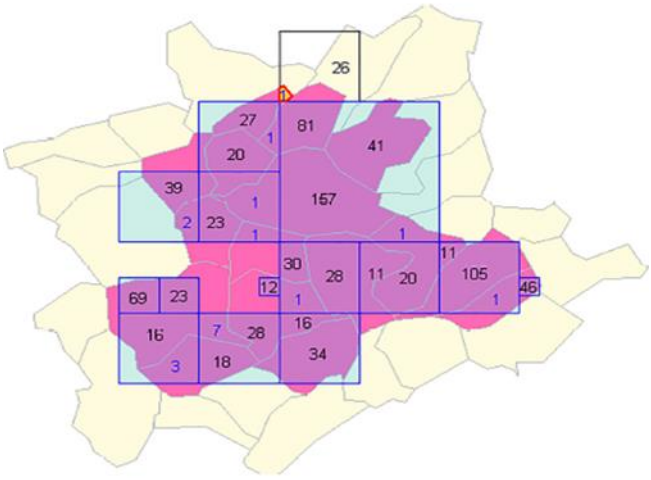

Figure 1. Differencing issue with 16 municipalities.

3. Enhancing personal data protection in the dissemination process of geospatial data

3.1 SDC techniques applied to geospatial data

Grid data and data on small areas generally present high disclosure risks and must be treated with specific methods, especially when other datasets have already been released on the same population with the same scale, or when the same dataset has already been released on different, overlapping areas (risk due to “geographic differencing”).

Technical recommendations to handle statistical disclosure control on geospatial data rely on a combination of two types of methods: pre-tabular methods applied on micro-data and post-tabular methods applied on tables.

For more details on SDC techniques applied to geospatial data, you can refer to chapter 14 in (Loonis, 2018).

3.1.1 Pre-tabular (perturbative) methods

Pre-tabular methods are based on changes on individual data so as to slightly alter individual features to prevent units from reverse identification while assuring that all aggregated indicators remain accurate and unbiased. Commonly used techniques are record swapping and blurring.

Record swapping consists in swapping the geographical positions of two units while keeping other variables unchanged. An iterative approach is required to optimize the consistency of the data.

Blurring consists in introducing a random noise in the values of each risky unit. It makes the data less accurate.

3.1.2 Post-tabular (suppressive) methods

Post-tabular methods are based on suppressing risky cells or aggregating them to larger cells, having enough units to ensure personal data protection.

The most common way to do this consists in building the tabular data (just as if it were disseminated without any constraint), then flagging risky cells (cells that do not satisfy the dissemination constraints) and finally treating them iteratively until no risky cell remains. The process is the same for spatial and non-spatial data, geography being here one particular dimension of the data to disseminate.

Another approach is to work directly on the micro-data and associate to each observation a probability of reverse identification, based on the distribution of the variables of interest across space: 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 also be risky.

3.2 Computing considerations

There are two types of computing considerations regarding geospatial data. The first one regards the daily work of the statistician: dealing with spatial data adds complexity in the SDC process because it requires implementation of specific methods which require great computing resources, far beyond the capacities of a standard machine. It is then necessary to use multi-thread programming and reduce the complexity of the algorithms (see 3.2.1).

The second type of computing considerations regards what "the world" may do with the huge amount of geospatial data disseminated and the numerous resources made available by the "open" trend. One possible way to address this topic could be to build highly performant and secure platforms called "virtual data enclaves" (see 3.2.2).

3.2.1 Reducing complexity when dealing with geographic differencing

The method detailed by Costemalle (2019) gives an example of the inherent complexity of detecting geographical differentiation problems.

In the case of grid data and municipalities, the order of magnitude of the set of combinations of cells and municipalities (and therefore of the possible differentiations to be tested) increases exponentially with the number of municipalities.

It is then necessary to reduce the number of dimensions of the problem, which can be achieved by modelling the differentiation problem with graphs.

Each vertex of the graph represents a municipality and two vertices are linked if they are intersected by the same cell. The graph makes it possible to target the cells which are directly involved in geographic differencing and thus strongly reduces the complexity of the algorithm.

3.2.2 Building Virtual Data Enclaves (VDE)

The growing demand for geospatial data dissemination reinforces risks of unit identification.

First, there is a global trend to open-source administrative data and core information belonging to the National Spatial Data Infrastructure, like address registers, cadastral parcels boundaries, geolayers used for building informative maps, etc. Making these data and tools accessible is aligned with GSGF Europe principles but makes it much easier for their users to identify statistical units.

Second, there is a growing focus on spatial impacts of public and private actions, so that policymakers, analysts and even citizens are looking for detailed information across a broad range of spatial dimensions at a very small scale, with may again raise disclosure of individual characteristics.

As an answer to these concerns, some countries have experimented sharing and archiving confidential geo-referenced micro-data using a "Virtual Data Enclave" (VDE) that is to say a virtual machine environment designed specifically for research needs when dealing with personal data. VDE lets researchers share, use, and analyse remotely hosted data on their desktop computers but doesn't allow them to download them.

VDEs, alongside SDC techniques, show a new possible way to protect data while ensuring appropriate access to individual data.

4. Confidentiality issues throughout the statistical business process

Statistical confidentiality and security are key quality dimensions of the statistical business process. Considering the generic statistical business process model GSBPM (UNECE, 2017), confidentiality issues have to be tackled in the following steps:

1.3 Establish output objectives

- To what extent have legal constraints regarding statistical outputs been considered, for example but not limited to ensuring confidentiality of data and preventing the disclosure of sensitive information?

1.5 Check data availability

- To what extent have legal constraints regarding data collection, acquisition and use been assessed and any necessary changes been proposed?

2.1 Design outputs

- Have the confidentiality rules and micro data access procedures been designed?

4.2 Set up collection

- Is there a risk of a breach while data is being transferred (survey and administrative data sources)?

6.4 Apply disclosure control

In sub-process 6.4 of GSBPM, specific methods and procedures are applied to assess and reduce disclosure risks in maps and other dissemination products dealing with locationally identifiable data, in order to securely and efficiently protect analytical outputs that anonymize data and prevent them from reverse identification of individuals. Those SDC methods can be applied to micro-data, before processing ("pre-tabular" methods, like swapping) or to processed data ("post-tabular" methods, divided in two main groups: perturbative vs. non-perturbative).

- To what extent is the business process using standard or well-known methods for avoidance of identification and protection of sensitive information?

- To what extent is the data protected from the risk of disclosure of sensitive information?
- To what extent is the data actually protected? What is the residual risk of disclosure?
- To what extent has the usability of the data been degraded? What is the loss in precision or level of detail?

7.3 Manage release of dissemination products

- Are researchers who have access to micro data legally bound to uphold confidentiality and security protocols of the NSI?
- Are research proposals submitted for approval by NSI analysts (analysts must approve the relevance of the analysis and the appropriateness of the methods)?
- Are there policies in place that ensure outputs are vetted prior to their dissemination?
- Are there confidentiality rules in place, such as a minimum number of units in a cell when doing cross-tabulations, and a maximum number of data requests per day with a maximum number of variables per request (to protect against penetration by an automated data mining process)?

5. References

Costemalle V., 2019. Detecting geographical differencing problems in the context of spatial data dissemination. Statistical Journal of the IAOS, vol. 35, no. 4, pp. 559-568.

Dempsey C., 2014. Tobler's First Law of Geography. Geography Realm information site. (https://www.geographyrealm.com/toblers-first-law-geography/)

Loonis, V., 2018. Handbook of spatial analysis, Theory and practical application with R, Insee Méthodes n°131, directed by Loonis V. (https://ec.europa.eu/eurostat/documents/3859598/9462709/INSEE-ESTAT-SPATIAL-ANA-18-EN.pdf/c4f87d5b-b508-4aff-ad7d-264da463077e?t=1545319662000)

UNECE, 2017. *Quality Indicators for the Generic Statistical Business Process Model (GSBPM) - For Statistics derived from Surveys and Administrative Data Sources*, United Nations Economic Commission for Europe (UNECE), Version 2.0, October 2017. (https://statswiki.unece.org/display/GSBPM/Quality+Indicators)

UNSC/UN-GGIM (2019). The Global Statistical Geospatial Framework. The GSGF. (https://ggim.un.org/meetings/GGIM-committee/9th-Session/documents/The_GSGF.pdf)

6. See also

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Naylor, J., Nordholt, E.S., Seri, G., De Wolf, P-P., 2010. Handbook on Statistical Disclosure Control. ESSNet-handbook SDC.
(https://ec.europa.eu/eurostat/cros/system/files/SDC_Handbook.pdf)

de Jonge, E., de Wolf P.-P., 2021. Introducing sdcSpatial - Privacy protected density maps. publication of R Package.
(https://edwindj.github.io/sdcSpatial/articles/privacy_protecting_maps.html)