GeoStat

GSGF Europe:

Geospatial statistics according to data collection method

BIG DATA

GEOSTAT 4

eurostat

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

Author: GEOSTAT 4

Project: Eurostat ESSnet grant project GEOSTAT 4

Grant agreement number: 945503 - 2019-FI-GEOSTAT4

Summary

There are numerous current challenges to be tackled and overcome for incorporating Big Data into official statistics and their statistical production process, including accessibility, reliability, legal constraints, technical issues, ethical and organisational aspects. Nevertheless, what it seems increasingly unavoidable in the future of official is a combination of existing and traditional statistical data sources and processes with Big Data, including Geospatial Big Data, strengthening geospatial statistics and all statistical-geospatial outputs.

After discussing key characteristics of Big Data regarding their use in official statistics, two different use cases are considered through the prism of the GSGF Europe: Big Data as reference data on geography, and Big Data as location data of a phenomenon. Based on this analysis, additional Recommendations are added to the core set of GSGF, which are primarily designed to address Big Data issues, but may apply more generally to other types of data sources. These new recommendations aim to improve further consistency and quality of geocoding results, to monitor the coverage and temporal changes of geospatial data, and to promote the use of innovative geospatial dissemination platforms.

Content

1. Introduction

Over the past years, and especially since the Bucharest memorandum on Official Statistics in a Datafied Society (2018), Big Data has become a strong focus of interest, attracting more and more attention from European statistical organizations as a means of modernising the production of statistics, apart from classical sources for statistical data. The rapid and growing influx of Big Data, coming from many different types of sensors, has already considerably increased the data sources that include spatial components. Inside the various forms of Big Data, "Geospatial Big Data" designates these new data sources which are captured natively with geospatial data features.

However, there are major challenges in using Geospatial Big Data for official statistics, namely on changing and setting up classical data collection and production methods on the national level that may be time and money consuming. On the one hand, Big Data has a great potential to improve statistical accuracy and NSI ability to produce information at a very local scale; on the other hand, Big Data's benefits to statistical production are usually limited by issues such as private ownership, confidentiality and information reliability.

In this document, we describe key characteristics of Big Data that are important to consider as a collection method for statistical purposes. Technical and institutional topics to be handled are then discussed. Two different use cases are considered when it comes to geospatial statistics through the prism of the GSGF Europe: I) Big Data as reference data on geography; ii) and Big Data as location data of a phenomenon. Finally, additional suggestions are added to the core set of GSGF: Europe Requirements and/or Recommendations, considering topics like geocoding, quality management and data exploration with Geospatial Big Data.

2. Data collection method description

Big Data is usually defined as structured and unstructured datasets with massive data volumes that cannot be easily captured, stored, manipulated, analysed, managed and presented by traditional hardware, software and database technologies. Big Data is described with three dimensions that characterise the challenges and opportunities of large data: **Volume**, **Velocity** and **Variety** (3Vs). The additional fourth dimension of Veracity is sometimes added to describe data integrity and quality. Even if there is an infinite diversity of Big Data sources, most Big Data generated comes from three primary sources: social data, machine data and transactional data.

Social data comes from the Mails, Comments, Likes, Tweets & Retweets, Video Uploads and Views, and general media that are uploaded and shared via the world's social media platforms (including CV job offers, etc.) - this kind of data provides invaluable insights into citizens behaviour and sentiment. Using social data for official statistical faces numerous issues regarding data gaps, including availability, accessibility and discoverability, as large parts of social data are currently inaccessible mainly due to private providers. In this regard, most social data are often still collected for internal use from the social media platforms, and can be accessed by external users only according to payment.

Machine data is defined as information which is generated by technical equipment and sensors that are installed in devices. This type of data is expected to grow exponentially as the Internet of Things (IoT) grows ever more pervasive. It includes data from sensors such as mobile phones, GPS, smart meters, road cameras, satellites, health sensors, etc.

Transactional data is generated from all the daily transactions that take place. Invoices, payment orders, storage records, delivery receipts – all are characterized as transactional data, mostly useful to highlight economic trends. Potentially, the three primary sources of big data can include location information and be used with a geospatial perspective; however, machine data are the one with the greatest geospatial potential, as they usually natively included geolocated information (like x, y coordinates of the sensor).

In comparison with traditional types of data and conventional data collection methods, Big Data present several challenges regarding the production of statistics. Unlike survey data (see **GSGF Europe : Geospatial statistics according to data collection method – Survey Data**), they are not primarily designed to be used in statistical production, but in operational and business contexts depending on their actual provenance, which implies that concepts, definitions and classifications may differ from the ones used in the statistical community. Therefore, Big Data is usually outside of the scope of national or regional statistical authorities and present a more commercial purpose. Furthermore, Big Data might not be stable over time as operational needs can lead to immediate change in data pipelines, neither representative of the target population of interest, with discrimination and exclusion issues due to biased data. Part of Big Data are structured, similarly to administrative data, with named variables from which statistical materials can be directly derived, but parts are unstructured or loosely structured, for instance, satellite imagery, raw text from web scraping, etc.

Producing geospatial statistics based on Big Data enabled pipelines leads to even greater challenges, with both technical and institutional topics to be handled.

On the technical side, the increasing volume and varying formats of collected Big Data add new challenges, in comparison to the major requirements already specified in GSGF: Europe on topics such as storing, managing, processing, analysing, visualising and verifying the quality of data in a scope of producing geospatial statistics. Furthermore, the size, variety and update rate of datasets exceed the capacity of commonly used contemporary geospatial computing systems, and spatial database management technologies to learn, manage, process, and visualise the data with reasonable effort. Big Data paradigm led to a fast, continuous and pervasive data streams that exceed geotechnological capabilities of contemporary computing systems to process, analyse and display in real time. This is particularly relevant since there is a growing need from the users for quicker access to datasets in which up to date data is increasingly crucial to make decisions emphasising the data requirement of timeliness. High-frequency spatial data may even lead to a complete overhaul of the geospatial statistics process by rethinking the fundamental concepts related to geographical units of measurement – as the latter could be defined differently when the flow of data interweaves spatial and temporal elements at the same time.

On the institutional side, many Big Data sources are produced and owned by private sector companies, like mobile phone operators or web companies, and this aspect could become prominent with the increasing production of data from smart sensors and Internet Of Things (IoT) – especially regarding Geospatial Big Data, as IoT is usually linked with geotagging of data (for instance, with smart watches or health trackers, etc.). The fundamental question relies if the access to Big Data sources depends on some form of cooperation with the data owner. In most of the cases, such a cooperation is necessary, there are only some exceptional cases such as freely available data and information on the web.

Already in the past, NSI gained experiences in cooperating with data owners outside the statistical system, e.g., administrative data community. The decisive difference to the access to new data sources is the very progressive building of legal entitlement at national and international levels for accessing privately held data – meaning there are still many use cases without sufficient mandatory references to empower statistical offices regarding data access. In this regard, there is a gap within the legal framework that operates in the European statistical-geospatial framework, concerning official and binding mandates for the provision of statistical and geospatial data to the NSI and other relevant authorities on the national level, particularly addressed to private data owners. Furthermore, cooperation with data owners does not always include full access to the raw data. In fact, the focus of the statistical system should be on the extraction of the desired data for statistical purposes. One could differentiate between the case, where the NSI has to develop the technological processes applied to the raw data and the case where these technological processes were developed by the data owner or a third party - which also means a loss of control by the NSI over the way "data abstracts" are produced and their reliability.

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

In the 2019 version of the GSGF: Europe Implementation Guide (GEOSTAT 3, 2019), it was stated *"one of the key areas of the European Statistical System (ESS) Vision 2020 is to harness new data sources comprising Big Data, administrative data and geospatial data. Using data from a range of sources, for multiple purposes, requires their integration into a common reference system of harmonised concepts, but also common location and temporal framework*". Despite this inaugural statement, Big Data were so far only mentioned in a secondary way in the GSGF: Europe, as they were only seen as a kind of alternative source. Only two Requirements explicitly stated Big Data, either as input or output of statistical processes, but none of them really address specific challenges of setting up Geospatial Big Data Pipelines:

- Requirement 4.2,"Enable data integration through consistent semantics and concepts across domains" states *"the general ambition in the ESS is to increase the use of administrative data sources and alternative data sources such as big data. In doing so, semantic interoperability between different domains poses even greater challenges (...)*".
- Requirement 4.5,"Explore the potential of Linked Open Data for increased interoperability" emphasizes *"joining up of data has previously been done through methods such as data linkage but increasingly the diversity and complexity of administrative and big data sets compared to traditional surveys and census data requires new thinking to allow these datasets to be queried (...)*".

Nevertheless, general Recommendations of the GSGF: Europe about data collection activities fully apply to Big Data based on statistical processes: for Principles 1 and 2, using geospatial reference data and services from NSDIs when it comes to geocode data, building an effective and secure data management environment, ensuring consistency and quality of geocoding. However, Big Data may lead to special provisions within these Recommendations – and even complementary Requirements.

Regarding the possible use of Big Data for the production of geospatial statistics, two facets should be distinguished within the framework of the GSGF: i) the use of big data as a source to feed statistical processes with a spatial component; ii) and the use of big data to enrich reference geospatial infrastructures.

Here a distinction between two types of big data is presented :

- BD location data of a phenomenon/event: these data are processed by the NSDI to produce statistics;
- BD reference data on geography (like Earth Observation Data). They can be considered as infrastructure data and therefore likely to integrate the NSDI.

Regarding the use of Big Data as a source to feed statistical processes with a spatial component, related use cases can be described using GSBPM phases, with a special focus on Processing sub-processes, as there are particular issues regarding data integration and data cleaning.

A specific Big Data architecture framework in the ESS, based on GSBPM and EARF, has been designed during Big Data 2 ESSNet: the BREAL (Big Data Reference Architecture and Layers). The BREAL Application Architecture consists of a set of generic application services, proposed with the purpose of showing how the identified business functions could be implemented, with particular attention on data wrangling function, which manages the process of transforming and manipulating raw data acquired in the acquisition and recording phase into formats and sizes which are easy to handle by the following services. The BREAL Information Architecture consists of three layers, namely a raw data layer, a convergence layer and a statistical layer. In addition to the data concepts, some metadata concepts are introduced for each of the three layers. In particular, one specific category of metadata has been selected as very specific for big data, that is provenance metadata.

The use of Big Data as part of an NSDI is a more problematic topic whenever private sources are considered. Although the purpose is to examine Big Data whatever its origin, it is clear that, most often, this data is produced by private actors, in particular companies in the field of the digital economy. However, the place of private data was given little focus in the 2019 version of the GSGF Framework, in particular with regard to the notion of NSDIs.

Principle 1 of the GSGF Europe aims to highlight the notion of NSDIs, pressing public actors in the field of geospatial statistics (NSI, Mapping agencies and other national or regional public authorities relevant on producing statistical, geospatial and administrative data) to use a common base of localizing information (Requirement 1.1 - *use data from NDSI*), favour point-based geocoding (Requirement 1.2- *use point-based location data for geocoding*) and formal partnerships (Requirement 1.3 - *build formal working relationships on institutional agreements*). Overall, this principle does not address private sources as far as spatial data infrastructures are concerned. Private sources are certainly mentioned in Recommendation 1.2.1 (as stated in the 2019 version of the GSGF Europe), but only in the sense that the geocoding method and infrastructure used by public actors must be applied in the same way to public as to private data: "*Countries should agree on one single uniform national infrastructure for geocoding of all public and potentially private data*").

The question that can then arise is the following: can NSDI be built from a set of elements shared between public AND private actors, while considering that certain private data (and private actors) can be constitutive elements of NSDI? Because NSDI should remain a public capability, GSGF: Europe relies on the hypothesis that *private* Big Data sources are not intended to incorporate the NDSI and make part of reference data. Still they can be processed using NSDI services so as to enrich geospatially enabled statistics and produce more dissemination products. Furthermore, *public* Big Data sources are still to be considered as a valuable component to be included as part of NSDI.

In order to prepare statistical infrastructures for the new technical challenges related to Big Data, a new recommendation is thus added in the 2022 version of the GSGF: Europe, in order to build capacities suitable for processing large volumes of data, in particular streaming data.

Recommendation 1.1.5 : *"Though being in an early stage of development, countries should be open to and consider the need to include geospatial Big Data in their NSDIs and its potential. A common infrastructure for Big Data needs to be adapted concerning its features such as volume and velocity. If the spatial features and smart captors (e.g., sensor data and mobile/streaming data) are themselves elements that move over time, this may mean that the geographic reference frames must be scalable over suitable temporal frequencies, which could go as far as continuous updates (real-time basis). Earth Observation Data should be considered as a valuable Big Data source through its analytical potential on spatial resolution and time series encouraging its integration in official statistics (e.g., land use and land cover monitoring statistics) and management with authoritative and more traditional data sources and processes."*

4. Specific Requirements and Recommendations

Despite the limitations encountered in terms of accessibility and reliability, Big Data are an opportunity to strengthen statistical coverage on spatial phenomena at very different scales. In addition to the discussion on the strengths and weaknesses of these data sources and their possible place within an NSDI, the following Recommendations are presented as an additional support to the GSGF:Europe and may be useful to produce and disseminate geospatial statistics using Big Data.

Managing processes with pre-geocoded data In Big Data sources like IoT, geocoding is usually embedded at the very starting point of the process and cannot be managed within the statistical pipeline. This geocoding is natively taken into account and dependant on technical parameterisation of smart sensors. Hence, some particular issues about measuring the quality of geospatial information in the case of Big Data, both the quality of geocoding (change in technical parameterization for instance) and more generally the quality of all the data transmitted (missing observations for example). General GSGF: Europe Recommendations about geocoding workflow are to be extended to geospatial data stemming from smart sensors, in order to ensure a consistent and conform result whether they are managed by public or private actors. Technical parameters and methodological choices that underly the geocoding process embedded in smart sensors have to be documented and shared between actors.

Hence, actual Requirement 2.3 ("Ensure consistency and quality of geocoding results ") related to Principle 2 is completed in 2022 version of GSGF:Europe with a complementary and transversal Recommendation 2.3.4: "*Use of geo-enabled data that originates from private data providers implies little or no means of influencing the underlying technical infrastructure in data capture. It is highly recommended that technical parameters and methodological choices involved in the data capture process embedded in smart sensors should be documented and shared between stakeholders.*"

Ensure quality of Geospatial Big Data

Geospatial Big Data are also facing major challenges on the quality side, because of their Velocity and Variability characteristics. Velocity raises the question of the intertemporal consistency of measurements, and in particular of maintaining a constant and consistent field of observation, which is a particular difficulty in real-time processes. The following mobile phone use case is illustrative of this statement. Assuming that a Big Data statistical pipeline is based on a continuous stream of geocoded data allowing the production of very local estimation of population given time of the day. A one-off failure of an antenna, over a given time slot, can bias the statistical estimate with data gaps, such as missing values. Furthermore, the collection metadata does not provide information on punctual technical failures. Variability of geospatial big data may also indicate inconsistency, stemming from both the variety between different data channels as well as the unreliability of any of these sources. For example, sensor measures may vary with the location in the environment, depending on the transmission channel or even on the time of day: this dependency on location and time is non-linear and difficult to predict and model. Hence the need of an ad-hoc quality framework that can handle the 3Vs of Big Data and make it possible to estimate the relevance of computed statistics.

Overall, same approaches and principles regarding sampling data and its representativeness should be applied to Big Data through new Recommendations included in the 2022 version of GSGF: Europe Implementation Guide.

Requirement 2.5 "Define common data quality frameworks taking into account spatial consistency through time and the type of collection process"

- *Recommendation 2.5.1 " The coverage of collected data in relation to the population of interest must be identified, with particular attention to measuring the representativeness at the local scales. When data is collected on principles that do not guarantee representativeness (big data, or even administrative data, by comparison with sampled survey data and weighted observations), a measure of "non-coverage" must be established (relative importance of missing values) and correction/imputation methods must be defined by statistical institutes which use these data.*
- *Recommendation 2.5.2 " Changes over time in data coverage should be documented by data providers, whether the data is collected on an ad hoc basis (e.g. administrative data from one year to the next) or continuously (e.g. big data stream). This provides geo-historical data to support data treatment and analysis, and preferable reduce the time gap/stamp between the time of the event and data collection".*

Help exploring Geospatial Big Data through dissemination tools

Among the various mechanisms for disseminating statistical data, visualisation services are playing an increasing role in users' needs and expectations. These services make it possible to have a first overview of the data, or even to provide a first level of analysis, without having to download all datasets or having skills to manage geospatial data. Techniques for studying data based on graphical representations are usually grouped under the term visual analysis.

The purpose of visual analysis is to recognise and understand phenomena represented through (digital) graphical representations of data. It is also defined as the process for reaching a judgment about reliable or consistent effects by visually examining data. This analytical and graphical reasoning is supported by static or interactive visual techniques.

In this sense, the visualisation function takes on a particular importance for statistical-geospatial data since the spatial aspect of data can be directly shown in adequate maps, in a much more understandable and relatable way than long tables with associated geographic codes. In this regard, with geospatial big data statistical pipelines, visual (spatial) analysis plays a crucial function. It provides a much more malleable dissemination engine than file downloads, especially if it can be linked with an on-the-fly calculation system, so that the final user can directly explore the content of data deposits with geospatial visualisations.

The design of a statistical production process aimed at establishing geospatial statistics from Big Data may need to include a particular place in the Visual Analytics phase, taking into account its importance both for data cleaning, exploration and dissemination. As a priority, the GSGF should first deal with the contributions of visual analysis as a dissemination channel of geospatial data. This aspect is linked to Principle 5 (Accessible and usable geospatially enabled statistics). We suggest expanding the scope of Requirement 5.2 as stated in 2019 version of GSGF: Europe "*Use service-oriented data portals supporting dynamic integration of data*". This extension results in the addition of a new recommendation 5.2.5 in 2022 version of GSGF:Europe: *"Countries should explore innovative dissemination platforms by offering data visualization tools, highlighting the geographic dimension of statistics. When possible, these data visualization tools should help internally organisations and end users in their data exploration and to customize the production and dissemination of geospatial statistics"*

5. References

ESSNet Big Data 2 (2020). BREAL - Big data Reference Architecture Layers.

GEOSTAT 3 (2019). GSGF Europe - Implementation guide for the Global Statistical Geospatial Framework in Europe.

Lee, J., & Kang, M. (2015). Geospatial Big Data: Challenges and Opportunities. Big Data Res., 2, 74-81.

Li, S., Dragićević, S., Anton, F., Sester, M., Winter, S., Çöltekin, A., Pettit, C.J., Jiang, B., Haworth, J., Stein, A., & Cheng, T. (2015). Geospatial Big Data Handling Theory and Methods: A Review and Research Challenges. ArXiv, abs/1511.03010.

Percivall, G. (2017). Big Geospatial Data – an OGC White Paper.

Robinson, A.C., Demšar, U., Moore, A.B., Buckley, A., Jiang, B., Field, K., Kraak, M.J., Camboim, S.P., & Sluter, C.R. (2017). Geospatial big data and cartography: research challenges and opportunities for making maps that matter. International Journal of Cartography, 3, 32 - 60.

Shekhar, Shashi & Gunturi, Viswanath & Evans, Michael & Yang, Kwangsoo (2012). Spatial big-data challenges intersecting mobility and cloud computing.