Geographical data quality for spatial analysis and geospatial statistics

Julien Gaffuri[1]

[1]Eurostat, Luxembourg, Luxembourg

Abstract

Geographical information science studies the representation of real-world phenomena into digital databases, as well as their analysis. Geographical datasets are, like statistical datasets, a representation of the real-world for specific purposes – their quality need to be described to assess how they meet the requirements or their users. This paper presents common practices, standards and frameworks in place to describe geographical data quality, illustrating the specificity of this type of data compared to statistics. The importance of controlling GIS data quality for the production of spatial analyses is then illustrated with two examples of geospatial statistics production performed at Eurostat. A first example concerns gridded statistics for building density analysis. The input GIS data represent the location and size of buildings and allows assessing the density of various building types on 100m and 1km resolution grids. A second example concerns the production of an accessibility indicator to healthcare services. The input GIS data represent the road transport network and the location of main healthcare services. GIS routing algorithms based on graph theory allow computing travel time from grid cells to the nearest healthcare service and assessing their accessibility on a 100m resolution grid. For both examples, different input GIS datasets of various quality are used (National topographic data, Tomtom MultiNet dataset and OpenStreetMap). The comparison of the outputs illustrates the importance of using reliable input GIS data, with properly controlled and documented quality.

Keywords: Geographical information, quality, GIS, spatial analysis, gridded statistics, CENSUS 2021.

1. Introduction

Together with official statistics, geographical information science (GIS) provides crucial inputs to inform citizens and policy makers on various thematic domains. The specificity of GIS analyses is of course to focus on the spatial dimension and provide advanced quantitative information on the spatial distribution of real-world phenomena and their potential correlations. The most significant progress for pan-European spatial analyses is the recent publication of the Eurostat Census 2021 population grid[1]. Like any computer-based analysis, the reliability of GIS analyses depends on various factors, such as the methodology used and the input data and their quality.

This paper presents common practices, standards and frameworks to describe GIS data quality. Two examples of geospatial statistics production performed at Eurostat are presented. They illustrate the impact of geographical data quality for spatial analyses.

[1] https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Population_and_housing_census_2021_-_population_grids&stable=1

2. Geographical data quality

Geographical data used in GIS software are stored as gridded data (or image, raster) and, more often, as vector data. Vector GIS databases represent and classify individual real-world entities as features with some properties (or attributes) and a geometrical representation in vector format. This geometrical representation describes explicitly the position, shape, and extent of the corresponding entity. A specificity of vector GIS databases lies in the high variety of different ways to define their content: There is an infinite number of ways to represent the real-world, and each GIS dataset represents only some aspects, with some approximations and generalisation, depending on its intended utilisation. The content of a geographical dataset is usually described in a specification document, which defines the concepts represented in the dataset and their representation modes. Data specifications allows creating and using a dataset – it relies on a 'universe of discourse', which is a view of the real world that includes the information of interest (and excludes the information not of interest). Quality, as the capacity to address user requirements, is of course a crucial aspect to address for geographical data producer. It encompasses two aspects:

- External quality, the degree to which dataset specifications meet user needs.
- Internal quality, the degree to which a dataset complies with its own specifications.

External quality is addressed at the preliminary stage of the dataset production. Considering the high cost to produce and maintain geographical datasets, geographical data producers define generic specifications intended to address a wide variety of user requirements. For a specific utilisation, customisations are necessary to comply with implicit user specifications: Unnecessary data is filtered out, data is transformed (sometimes generalised), and complementary information can be integrated and combined from other data sources.

Internal quality is easier to be formalised. *ISO 19157:2013 – Geographical information – Data quality* proposes a standard framework including a standard terminology for geographical data quality elements, procedures for quality evaluation and reporting. Some of the most common data quality elements are the following:

- **Completeness** is the degree to which a dataset representing features has values for all expected entities and their attributes. Omission occurs when data is unexpectedly absent (ex: a building is missing) – and commission on the contrary when data is present but should not (ex: a building that no longer exist in the real world is still represented). Data specifications usually precisely describe selection criteria for the entities: some entities may be omitted on purpose depending on their type (ex: represent road section which can be driven by car, exclude foot tracks), their size (ex: Buildings smaller than 10m² should not be represented), their importance in their spatial context (ex: An isolated building, even if too small, should be kept).

- **Conceptual consistency** is the degree to which the entities are represented according to the specified data structure.
- **Domain consistency** refers to the degree to which values provided for attributes are within the set of values defined in the specifications. For example, the attribute on the height of a building should be a positive number - "high" text would not be a valid value.
- **Topological consistency** relates to the degree to which topological relations of the geometries are respected. For example, for a dataset representing a network, initial and final vertices of network sections should be located exactly at the same position as the initial or final vertices of the other line sections they are connected to. A dataset representing a tessellation of surfaces, such as administrative units, should not contain overlaps nor gaps between touching surfaces.
- **Absolute positional accuracy** refers to the closeness of the feature geometry vertices positions to their true position in the real world. For example, the position of a punctual feature in a 1m resolution dataset should be represented with this positional accuracy. **Relative position accuracy** refers to the relative position of the features. Entities that are nearby, or compose some special spatial patterns as a group, should be represented accordingly.
- **Thematic accuracy** refers to the correctness of classification of the entities. For example, a motorway road section should not be misclassified as a foot track.
- **Temporal validity** refers to the validity of the data according to time, that is how the dataset represent the real world at the specified reference date.
- **Temporal accuracy** refers to the closeness of temporal measures to the expected values. For example, an attribute on the date of construction of a building should be within the specified accuracy interval. **Temporal consistency** refers to the chronological order of events. For example, the construction date of a building should be before its renovation date.
- Finally, **metaquality** refers to the quality of the quality description. Evaluating and documenting quality is crucial: The quality of geographical dataset may not be evaluated and documented in a satisfying manner, and sometimes not at all.

Data quality evaluation is performed by measuring quality elements defined in the data specifications. These evaluations can be either performed on the entire dataset (for example to measure domain consistency) or on a sample (for example for completeness). The result of this quality evaluation is usually included in the dataset specifications, as metadata elements. Examples of such quality reports are available for the BD TOPO dataset[2].

Geographical datasets should be selected depending on their quality and the user needs. In the next section, we present two example of GIS-based spatial analyses and the impact of quality.

3. Controlling geographical data quality on spatial analyses

Geographical data quality has an impact on the result of the spatial analyses it is used for. For example, deriving a simple statistical indicator based on object counts will result in different levels of reliability depending on the completeness of the input geographical dataset. The relative importance of the quality elements depends of course on the type of spatial analysis and the target accuracy level. We present below two examples of such GIS-based spatial analyses which were tested using different input geographical datasets having different quality.

1.1 Building density analysis

Eurostat aims at measuring the evolution of European building stocks and their characteristics with geographical data. For this objective, topographic and cadastral geographical datasets are used to derive statistical indicators at grid cell level (100m and 1km resolution) such as the number, the total ground area, and the total floor area of buildings. These three indicators are derived for specific types of buildings such as residential buildings, economic activity buildings (industrial, commercial, etc.) and buildings with a cultural heritage value (churches, castles, fortresses, etc.). Deriving such indicators requires detailed geographic datasets describing the geometry of buildings and their characteristics such as their height or number of floors (to assess floor areas), their usage (residential, industrial, etc.) and their nature. Other characteristics such as date of construction, energy performance, etc. could be used for other indicators.

These indicators could then be used in combination with gridded population data such as the Eurostat Census 2021 to assess population pressure and detect touristic regions with a predominance of secondary residences.

For this spatial analysis, two candidate input geographic datasets on buildings have been selected:

[2] https://geoservices.ign.fr/documentation/donnees/vecteur/bdtopo/rapport-controle-qualite

- The **BD TOPO®** dataset[3] produced and maintained by the French mapping agency IGN-France, with reference date 31/03/2024. The full documentation including quality control indicators are available[4].
- The **OSM** OpenStreetMap[5] dataset which is a popular community-based geographic dataset. The dataset was downloaded in March 2024. The documentation is available and does not include quality control indicators[6]. The reference date is unknown.

Computing the specified statistical indicators with both datasets and comparing the results reveals differences and limitations of the input datasets in terms of quality. Figure 1 illustrates this comparison for some selected indicators, at 1km resolution. The maps based on both geographical datasets show comparable distributions over space. The map showing the differences between both (right column) shows however significant differences. An investigation on the areas with significant differences reveal the impact of the limited and heterogeneous quality of OpenStreetMap compared to the BD TOPO® dataset. It concerns mainly the following quality elements:

- **Completeness**: Some buildings are missing in OpenStreetMap while they meet the selection criteria. Buildings are segmented differently in both datasets, which result in different counts.
- **Thematic accuracy**: Some buildings in OpenStreetMap have incorrect and often missing description on their type, usage, and height. Missing information on building height explains the difference for floor area indicators. Some buildings are misclassified as light constructions.
- **Temporal validity**: OpenStreetMap reference date is expected to be the present. Some differences result in a lack of updating.
- **Absolute positional accuracy**: The geometric representations of both datasets differ significantly due to capture at different scales. Several meters differences can be found.

1.2 Healthcare services accessibility analysis

Eurostat produces datasets on the localisation of basic services, such as main healthcare services, over Europe. This geographical dataset is used to assess the accessibility to these services by road transport network over Europe, allowing to find population clusters “left behind”

[3] https://geoservices.ign.fr/bdtopo

[4] https://geoservices.ign.fr/documentation/donnees/vecteur/bdtopo

[5] https://www.openstreetmap.org

[6] https://wiki.openstreetmap.org/

with low accessibility. This information is important to guide for example the EU regional policy. This GIS-based accessibility analysis relies on GIS data representing the road transport network. GIS routing algorithms based on graph theory allow computing travel time from populated cells (from Eurostat Census 2021 population grid) to healthcare services and assessing their accessibility at 100m and 1km resolution. Three candidate datasets were selected for the road network:

- The **OME2** Open Map for Europe 2[7] prototype dataset on transport networks which is a high-scale geographic dataset relying on several national topographic datasets, such as the BD TOPO® dataset presented in the previous section.
- The Multinet **Tomtom** road transport dataset[8], which is a high-scale geographic dataset designed for embedded car navigators.
- The **OSM** OpenStreetMap dataset, described in the previous section.

Accessibility indicators on fastest travel time to the nearest healthcare service were computed using the three road transport network datasets. Figure 2 maps based on the three geographical datasets show comparable distributions over space. The maps showing the differences between them (figure 3) shows however significant differences. OME2 and Tomtom comparison show little differences – the main differences are observed with OpenStreetMap dataset. An investigation on the areas with significant differences reveal the impact of the limited and heterogeneous quality of OpenStreetMap compared to the two other datasets. It concerns mainly the following quality elements:

- **Completeness** and **temporal validity**: Some road sections are represented in a dataset and not in the other, due to different selection criteria in their specifications, or different up-to-dateness.
- **Thematic accuracy**: Road sections are classified differently from one dataset to the other, and thematic classification is not always provided in OpenStreetMap. This results it different travel time estimations.
- **Topological consistency**: Some road sections are not properly connected to the network. This result in impossible routes through some road junctions and unexpected detours that affect the travel times.

[7] https://eurogeographics.org/open-maps-for-europe/ome2-progress/

[8] https://download.tomtom.com/open/banners/MultiNet-product-info-sheet.pdf

4. Conclusion

In this paper we presented some specificities of geographical data quality. This type of data is different from statistical data and the description and control of its quality is necessary when using it for specific spatial analyses. We illustrated this with two examples on gridded statistics.

Geographical information offers incredible opportunities for statistical analyses – using it is a pertinent solution to better describe how indicators vary across space, with a better spatial granularity than the traditional administrative unit levels. Using geographical data for official statistics purposes would require more attention on the specific quality elements of this type of data. Geographical data sources should be selected depending on their quality. A pertinent approach would be to further involve the geographical information science community in the development of new geospatial statistics products. Existing geographical datasets could also be improved in quality to better meet the needs of official statistics production.

Figure 1: Results of building density analysis and comparison

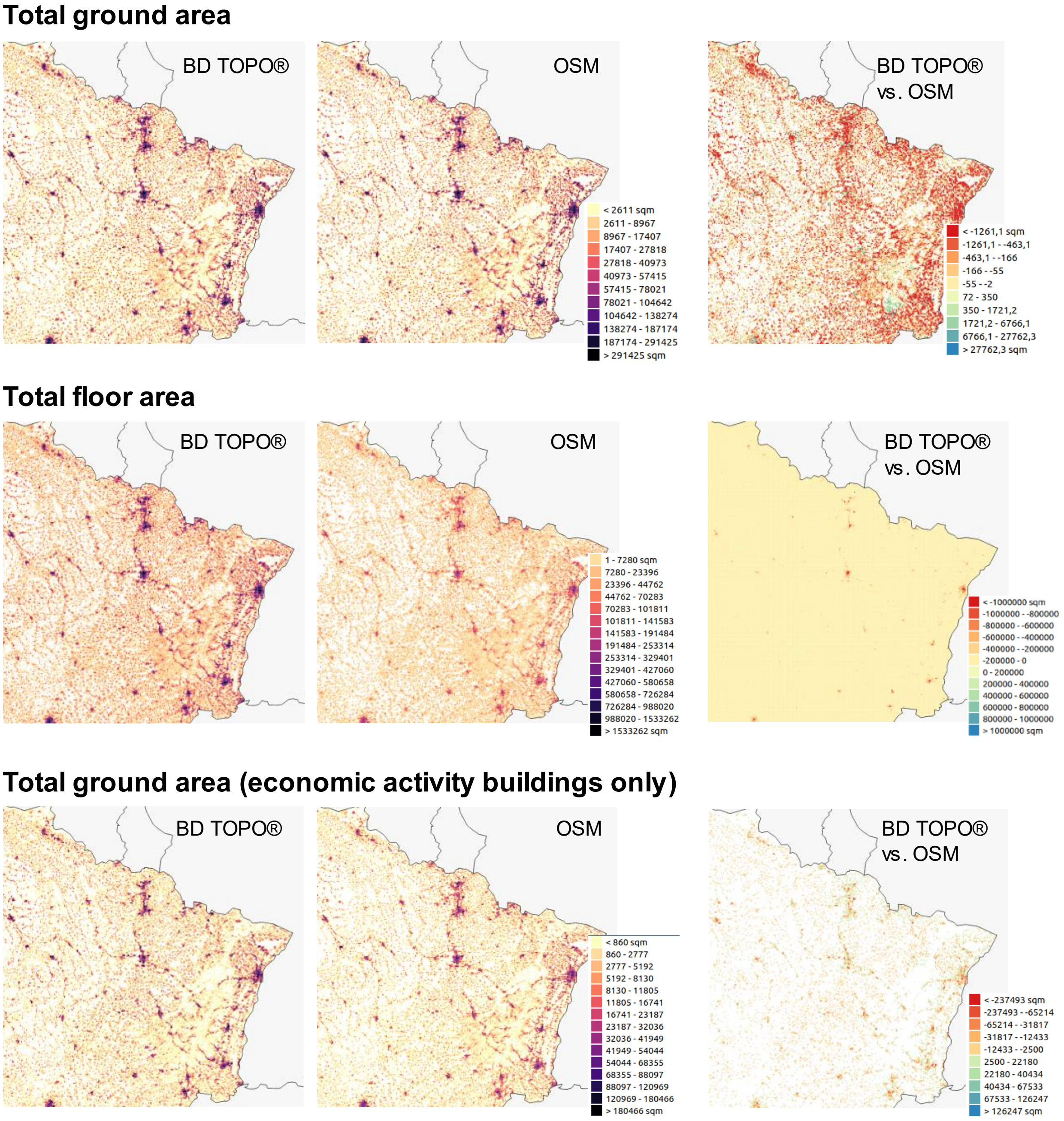


Figure 2: Results of accessibility analysis

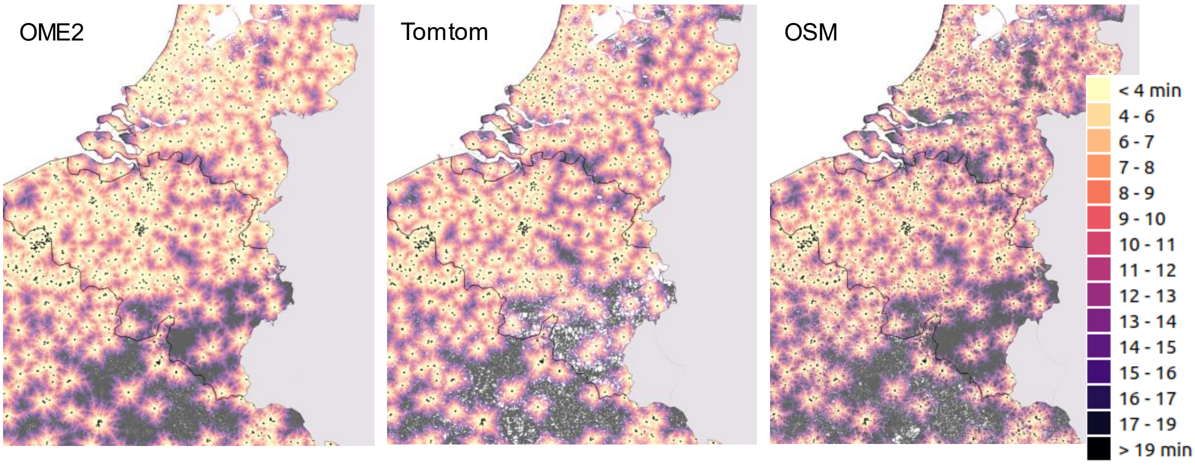
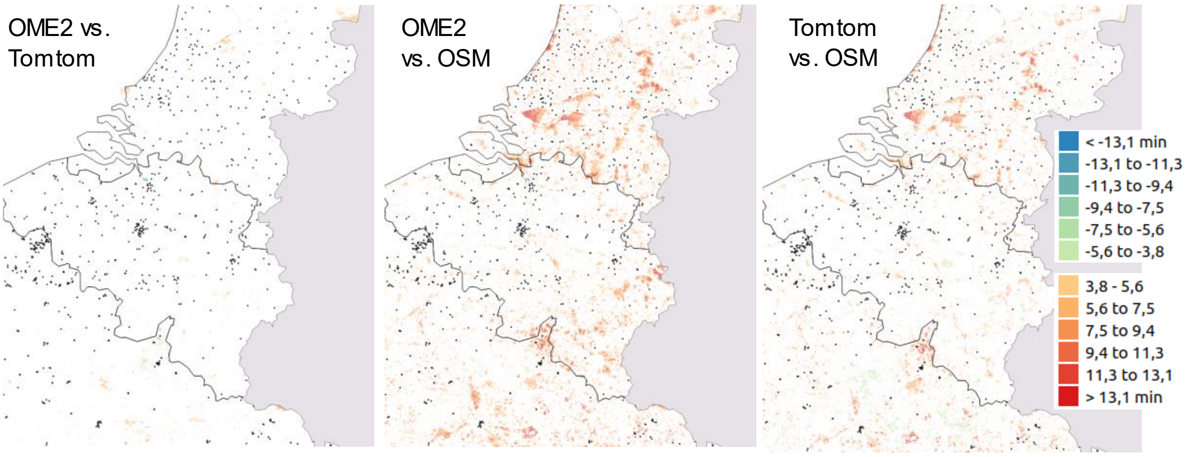


Figure 3: Comparison of accessibility analysis results


OME2 vs. Tomtom
OME2 vs. OSM
Tomtom vs. OSM
< -13,1 min
-13,1 to -11,3
-11,3 to -9,4
-9,4 to -7,5
-7,5 to -5,6
-5,6 to -3,8
3,8 - 5,6
5,6 to 7,5
7,5 to 9,4
9,4 to 11,3
11,3 to 13,1
> 13,1 min