Pan-European open building footprints: analysis and comparison in selected countries

Marco Minghini[1], Sara Thabit Gonzalez[1], Lorenzo Gabrielli[1]

[1] European Commission, Joint Research Centre (JRC), Ispra, Italy -
(marco.minghini, sara.thabit-gonzalez, lorenzo.gabrielli)@ec.europa.eu

Keywords: Buildings, Open data, OpenStreetMap, Geoprocessing, GeoPython.

Abstract

This paper presents a comprehensive analysis of four non-governmental open building datasets available at the European Union (EU) level, namely OpenStreetMap (OSM), EUBUCCO, Digital Building Stock Model (DBSM) and Microsoft's Global ML Building Footprints (MS). The objective is to perform a geometrical comparison and identify similarities and differences between them, across five EU countries (Belgium, Denmark, Greece, Malta and Sweden) and various degrees of urbanisation from rural to urban. This is done in a two-step process: first, by comparing the total number and the total areas of building polygons for each dataset and country; second, by intersecting the building polygons and calculating the fraction of the area of each dataset represented by the intersection. Results highlight the influence of urbanisation on the dataset coverage (with increasing completeness when moving from rural to urban areas) and the varying degrees of overlap between the datasets based on a number of factors, including: the amount and up-to-dateness of the input sources used to produce the dataset; the presence of an active OSM community (for OSM and the datasets based on OSM); and the accuracy of Machine Learning algorithms for MS. Based on these findings, we provide insights into the strengths and limitations of each dataset and some recommendations on their use.

1. Introduction

The current data economy is characterised by a multitude of actors involved in the production of data. Compared to the past, when the public sector was the main societal player responsible for collecting, maintaining, updating and disseminating datasets, today's landscape is much more heterogeneous. Leveraging new technologies such as Artificial Intelligence (AI), Internet of Things (IoT) and cloud, private, research and citizen-led initiatives have become relevant producers of valuable geospatial data for several applications and use cases (Kotsev et al., 2021). In the European Union (EU), the European strategy for data (European Commission, 2020) emphasised the need to make sense of the huge amount of data produced by all societal actors through both technological and organisational measures, supported by legal interventions, with the ultimate goal to create a fair, trustworthy and interoperable data sharing environment known as the common European data space, which is currently in the making (Farrell et al., 2023).

With this broad and complex context in mind, this work addresses the geospatial dimension of data sharing, which is horizontal across several societal domains, and zooms into specific geospatial datasets — building footprints (hereinafter simply referred to as buildings). These are fundamental datasets for several applications, including city planning, demographic analyses, modelling energy production and consumption, disaster preparedness and response, and digital twins. As key resources of Spatial Data Infrastructures (SDIs), buildings have been historically produced by governmental organisations — National Mapping or Cadastral Agencies — as part of their cartographic databases, with coverage ranging from local to national and licensing conditions being heterogeneous and not always open. This has typically made it challenging to derive open building datasets with a continental or global scale.

Over the last decade, however, the unparalleled developments in the resolution of satellite imagery, AI techniques and citizen engagement in geospatial data collection have enabled the production of several building datasets available at least at the continental scale under open licenses. A crucial role in this process was played by OpenStreetMap (OSM, `https://www.openstreetmap.org`), the most popular geospatial crowdsourcing project started in 2004 with the goal to create and maintain a database of the whole world licensed under the Open Data Commons Open Database License (ODbL, `https://opendatacommons.org/licenses/odbl`). With more than 2 million contributors active so far (OpenStreetMap Wiki, 2015), OSM has become the largest, most complete and most up-to-date geospatial database currently existing and its usage spans across multiple use cases and applications (Mooney and Minghini, 2017). OSM buildings are typically mapped through the digitisation of high-resolution satellite imagery, while in some cases they derive from the import of third-party datasets (e.g. released from governments) having an ODbL-compatible license. The quality of OSM buildings has been heavily studied in literature. While being heterogeneous across countries, regions and cities, OSM quality (mainly measured as positional accuracy and completeness) usually increases when moving from rural to urban areas, where it can equal or even outperform the quality of authoritative datasets (Hecht et al., 2013; Fan et al., 2014; Fram et al., 2015; Brovelli et al., 2016).

Nevertheless, concerns about OSM quality have stimulated the birth of multiple initiatives, led by either private or research actors, to create open building datasets using OSM as the foundation. Among the most promising initiatives led by the private sector is Overture Maps (`https://overturemaps.org`), founded in 2022 by four companies (Amazon, Meta, Microsoft and TomTom) with the mission to provide global, high-quality and interoperable open datasets from the combination of several input sources. Instead, open building datasets produced by research-led initiatives include EUBUCCO (Milojevic-Dupont et al., 2023; `https://eubucco.com`) and the Digital Building Stock Model (Florio et al., 2023; `https://europa.eu/!W9YJqy`), which combine OSM buildings with other sources to create

more complete and reliable products. In addition, today's AI and machine learning capabilities in the remote sensing domain make it possible to automatically produce buildings from high-resolution satellite imagery, as done e.g. in Microsoft's Global ML Building Footprints (https://github.com/microsoft/GlobalMLBuildingFootprints/) and Google's Open Buildings (https://sites.research.google/open-buildings).

In this paper we analyse four non-governmental open building datasets available (at least) at the EU level: OSM, EUBUCCO, Microsoft's Global ML Building Footprints, and DBSM. The objective of the work is to compare the four datasets, which derive from different approaches based on specific processing steps and governance rules, in terms of their geometry (i.e. attributes are out of scope) in order to draw conclusions on their similarities and differences. The comparison is performed on five EU countries and takes into account the degree of urbanisation to assess whether and how this influences the results. The work is developed as follows. Section 2 describes the building datasets used as well as the database providing information on the degree of urbanisation. The methodology for comparing the datasets is then illustrated in Section 3, while Section 4 presents the results. Based on these, Section 5 closes the paper by providing a critical perspective on the building datasets analysed and proposing potential avenues for future research.

2. Datasets

As mentioned in Section 1, in this work we analyse and compare four open building datasets produced by either private or research-led initiatives. The first building dataset is extracted from the OSM database (see Section 1) using the predefined extracts offered by Geofabrik (http://download.geofabrik.de) and corresponding to the OSM objects tagged with the *building* key (https://wiki.openstreetmap.org/wiki/Key:building). The second dataset, named EUBUCCO (Milojevic-Dupont et al., 2023) (https://eubucco.com/data/), was produced by a research team at the Mercator Research Institute of Global Commons and Climate Change and the Technical University Berlin and released in 2022. It includes buildings for the 27 EU Member States and Switzerland with three main attributes: building type, height and construction year. EUBUCCO is produced by merging multiple input sources: governmental building datasets for countries where these are available under an open license; and OSM otherwise. EUBUCCO is mostly licensed under the ODbL, with only exceptions for two regions in Italy and Czech Republic licensed under the Creative Commons Attribution-ShareAlike 2.0 Generic (CC BY-SA, https://creativecommons.org/licenses/by-sa/2.0) and the Creative Commons Attribution-NonCommercial 2.0 Generic (CC BY-NC 2.0 Deed, https://creativecommons.org/licenses/by-nc/2.0), respectively.

The third dataset is Microsoft's Global ML Building Footprints, hereafter simply called MS (https://github.com/microsoft/GlobalMLBuildingFootprints). It is generated through the application of machine learning technology on Bing Maps high-resolution satellite imagery acquired between 2014 and 2023, is available at the global scale under the ODbL and is regularly updated. The fourth dataset, called Digital Building Stock Model (DBSM) was released in 2023 by the Joint Research Centre (JRC) of the European Commission to support policies on energy-related purposes. It is an ODbL-licensed, pan-European dataset produced from the hierarchical conflation of three input datasets: OSM, MS and the European Settlement Map (Florio et al., 2023; https://europa.eu/!W9YJqy). The four open building datasets analysed in this work (OSM, EUBUCCO, MS and DBSM) were downloaded in January 2024.

As mentioned in Section 1, the quality of OSM buildings (that are also reused by EUBUCCO and DBSM) usually depends on the degree of urbanisation. Additional studies further highlight the variance of open building coverage between urban and rural areas (Gonzales, 2016; Ullah et al., 2023). For these reasons, we disaggregate the analysis described in Section 3 across urbanisation levels, using the EU Nomenclature of Territorial Units for Statistics (NUTS) as the reference framework to allow comparisons between regions with different urbanisation degrees. The smallest administrative areas (NUTS3), usually corresponding to municipalities or counties with a population between 150,000 and 800,0000 inhabitants, was chosen as the reference unit for the analysis. The geospatial dataset representing the EU NUTS3 boundaries, including their degree of urbanisation (classified as urban, semi-urban or rural), is produced by the Geographical information system of the Commission (GISCO) and downloaded from https://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/administrative-units-statistical-units/countries.

The comparison between the four building datasets was performed on five EU countries: Belgium, Denmark, Greece, Malta and Sweden. The choice was motivated by the needs to: i) select countries of different size and geographically distant from each other, which ensures that their national OSM communities are substantially different; ii) select countries having different portions of urban, semi-urban and rural areas; and iii) select two sets of countries for which the input source for EUBUCCO buildings was a governmental dataset (Belgium, Denmark, Malta) and OSM (Greece, Sweden) to detect possibly different behaviours. Figure 1 shows the maps of the five countries chosen, classified according to the three degrees of urbanisation of their NUTS3 areas, as well as the fractions of those areas on the total surface of each country.

3. Methodology

Two main processes were undertaken for the comparison of the four open building datasets (OSM, EUBUCCO, MS and DBSM). First, we developed an overview of the building coverage for each dataset across all the five countries and the three degrees of urbanisation. To achieve that, we calculated the total number of building polygons, on the one hand, and the total aggregated area of those polygons, on the other. For each country and degree of urbanisation, we plotted the total number and the total area of buildings on a bi-dimensional plane in order to visually detect potential patterns and trends. Second, in order to assess the geometrical similarity between the four datasets, we calculated the intersection of their building areas, and the corresponding statistics, in two scenarios: i) for each given couple of building datasets; and ii) for all the four building datasets, taken together. As a result, the quantitative analysis carried out for each country could provide a measure not only of the extent to which the geometries of different datasets overlap when representing a given building, but also of the overall similarity between the four datasets (which may be valid even beyond the study area considered in this paper). To perform the geometrical comparison between the four building datasets, we re-projected all of them to the same projection (WGS 1984, EPSG 4326).

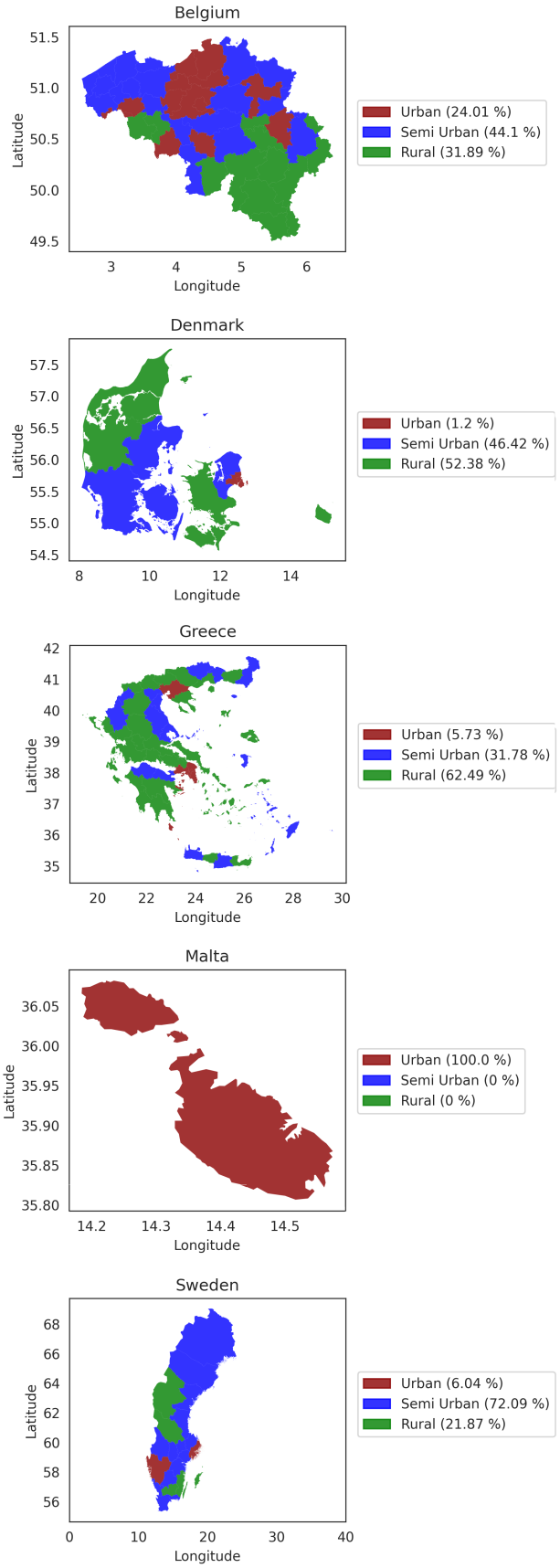

Figure 1. Maps of the five countries analysed in the study, classified according to the degree of urbanisation of their NUTS3 areas and corresponding fractions of their total area.

The whole procedure was written in Python, parallelised to increase efficiency and based on libraries such as Pandas, Geopandas, Dask-GeoPandas, Shapely, Plotly and Seaborn. The code is available under the open source European Union Public License (EUPL) v1.2 at `https://github.com/eurogeoss/building-datasets`.

4. Results

4.1 Total number and total areas of buildings

Table 1 summarises the total number and the total area of buildings for each of the four datasets in each of the five countries, which could be interpreted as a measure of their completeness. These results show that the four datasets perform highly differently in each of the five countries. While they also show variances across different degrees of urbanisation, the overall trends in terms of number and area of buildings were relatively homogeneous within each country and for this reason, the corresponding figures were not included in this section.

Dataset	Country	Number of buildings	Area of buildings [10^8 m^2]
EUBUCCO	Belgium	8,636,114	27.56
	Denmark	5,684,734	23.44
	Greece	856,140	3.04
	Malta	141,329	0.49
	Sweden	2,504,961	21.38
OSM	Belgium	6,211,451	23.94
	Denmark	3,654,875	22.82
	Greece	1,217,547	4.71
	Malta	20,225	0.16
	Sweden	3,050,667	24.84
DBSM	Belgium	6,610,034	26.75
	Denmark	3,765,255	23.76
	Greece	4,540,228	15.03
	Malta	58,247	0.46
	Sweden	4,936,573	36.16
MS	Belgium	4,557,403	25.86
	Denmark	3,541,845	21.11
	Greece	5,722,750	14.74
	Malta	73,579	0.44
	Sweden	6,422,594	31.07

Table 1. Total number and total area of buildings for each of the four datasets and in each of the five countries.

For the countries where the input sources of EUBUCCO correspond to open governmental building datasets (region- and city-level datasets in Belgium, and country-level datasets in Denmark and Malta), EUBUCCO reports the largest number of buildings in all cases. It also shows the largest total area of buildings in Belgium and Malta, and the second largest in Denmark, slightly surpassed by DBSM. In contrast, for the countries where the input source of EUBUCCO is OSM (Greece and Sweden), it reports the lowest values of number and total area of buildings. This can be explained by the fact that: i) EUBUCCO was released in 2022 and hence lacks all the subsequent updates included in OSM, which was extracted in January 2024; ii) DBSM, in addition to using OSM data from 2023, is further enriched by the use of two additional datasets (MS and the European Settlement Map); and iii) in MS, buildings are automatically identified by machine learning algorithms that typic-

ally result into larger numbers and areas of buildings than those available in OSM.

The MS dataset shows a very heterogeneous performance when assessed in comparison with the other datasets. In some cases, MS reports a relatively higher building coverage in rural areas, which even surpasses the total area reported by EUBUCCO's governmental dataset in the case of Belgium. This trend is also visible in the countries where EUBUCCO is not based on open governmental datasets (Greece and Sweden), where MS reports the highest number of buildings across all the four datasets, and the second highest total area after DBSM. However, exceptions were also found in our sample. For example, in Denmark MS reports by far the lowest coverage in semi-urban and rural areas.

The OSM dataset also shows different behaviours across the five countries. When compared with governmental data from EUBUCCO, it consistently reports lower numbers and total areas of buildings. However, when compared with MS the results vary. In Malta, Sweden and Greece OSM building coverage is consistently lower than MS. This difference is especially noticeable in semi-urban and rural areas, where the total area and number of buildings in OSM is significantly less than in MS. In Belgium, OSM reports a lower total area but a higher number of buildings than MS in urban and semi-urban areas, and a similar number of buildings in rural areas. Finally, Denmark stands out as an exception in this case, too. For this country, OSM shows significantly higher total areas than MS in semi-urban and rural areas, which comprise the majority of the country area (approximately 99%). This can be caused by multiple factors, e.g. an active OSM community in the country (which is confirmed by Neis, 2024) and/or a not so accurate performance of MS machine learning algorithms compared to other countries.

Finally, the DBSM dataset reports the largest total areas of buildings for all countries except two (Belgium and Malta), where EUBUCCO is based on governmental datasets and thus shows the highest values. This result is not surprising, since the methodology used to produce DBSM was based on a conflation process of multiple datasets including OSM and MS (Florio et al., 2023), which increase its completeness. Nonetheless, small variations were found. In Greece, Malta and Sweden, MS shows a higher number of buildings than DBSM, although its total area is lower. In contrast, in Belgium MS shows a slightly lower total area than DBSM, but a significantly (approximately 30%) lower number of buildings. Further research would be needed to understand whether these variations are either the result of MS updates after the release of DBSM, or are caused by the way MS footprints are actually used in the DBSM conflation process.

In Figure 2, the number of buildings and the total area of buildings for each of the four datasets in each of the five countries, already reported in Table 1, are plotted one against the other for each of the five countries, making it easier to identify the patterns described above.

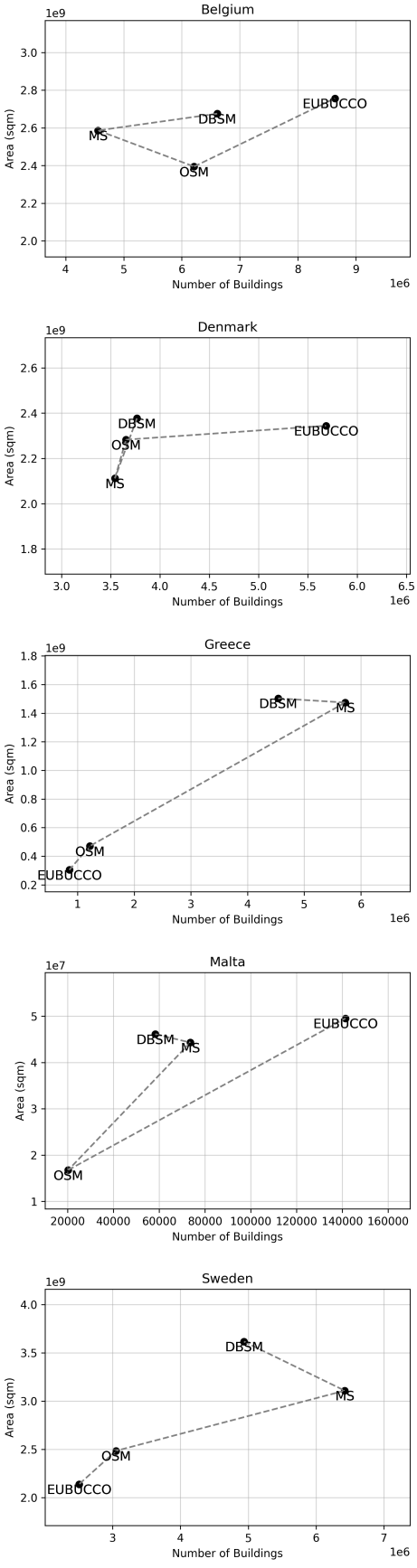

Figure 2. Number and total area of buildings for the four datasets in the five countries analysed.

4.2 Intersection of building datasets

To understand the extent to which the four datasets represent the same building footprints (i.e. their geometrical similarity), we performed the intersection of the building polygons of the four datasets and calculated the area of these intersections. We carried out this process in two steps.

First, we computed the area of intersection between the building polygons of all the four datasets. Figure 3 shows, for each

of the five countries and each of the four datasets, the fraction (expressed as a percentage) of the area of each dataset represented by the area of intersection between all the four dataset. In other words, a low percentage means that the geometries of building polygons of a given dataset are significantly different (or poorly intersecting) the geometries of building polygons of the other three datasets. In contrast, a high percentage means that the geometries of building polygons of a given dataset are similar (or largely overlapping) to the geometries of building polygons of the other three datasets.

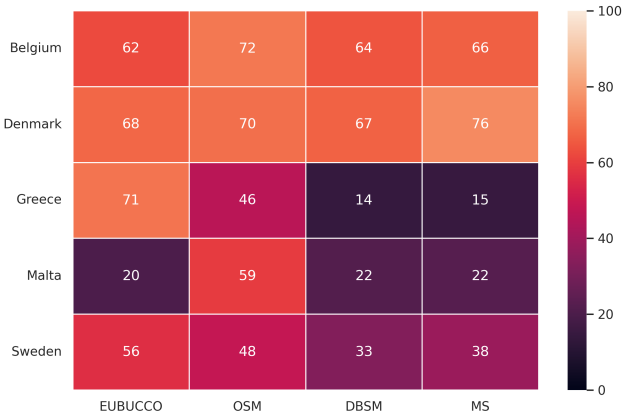

Figure 3. Percentage of the area of each dataset represented by the area of intersection between all the four datasets, for the five countries analysed.

The results of Figure 3 show that the areas of intersection between the building polygons of the four datasets highly vary across the five countries. For example, for Belgium and Denmark all the four datasets have a relatively high percentage of shared area (between 62% and 76%), with no significant differences across datasets. In Sweden the four percentages feature slightly lower values, ranging from 38% to 56%. In Malta, the shared area is around 20% for all datasets except OSM, for which the value raises to 59%. Finally, Greece shows the lowest values (only 14% and 15% for DBSM and MS, respectively). Analysing the same results from a dataset perspective, MS and DBSM (which also incorporates MS) are those featuring the lowest values for three out of the five countries (Greece, Malta and Sweden), which — with the only exception of Malta — are the countries where EUBUCCO is not based on governmental datasets. This might result from the very different production process of MS, which does not make use of OSM (in contrast to EUBUCCO and DBSM) but relies on machine learning algorithms. In Malta OSM building completeness is probably poor, which is the reason for the relatively high value for OSM (59%) and the low values of EUBUCCO, here based on the national governmental dataset (20%). The same exercise applied only to buildings in urban, semi-urban and rural NUTS3 areas shows that the varying building coverage across those areas consistently affects the percentages of overlap between the datasets. Accordingly, the fractions of the area of each dataset corresponding to the area of intersection between all the four dataset was found to be lower in rural areas (with minimum values equal to 7%) than in urban areas (with maximum values equal to 79%).

As a second step, to get a more detailed understanding of the similarity between each couple of datasets (taken separately from the others), we repeated the same process by analysing the intersections of the building polygons between each couple of datasets. For each of the five countries, each cell of Table 2 includes the fraction (expressed as a percentage) of the area of the dataset in the row represented by the area of intersection between the dataset on the row and the dataset on the column. Table 2 shows that OSM reports the highest values — ranging from 81% to 96% — when compared to EUBUCCO for the countries where the latter is based on governmental datasets (Belgium, Denmark and Malta). This shows that OSM completeness is very high in Belgium and Denmark, while it has room for improvement in Malta as also outlined above. The DBSM dataset shows the second-highest values when compared to EUBUCCO's governmental data with values between 74% and 93%, while the MS dataset ranges between 73% and 79%. In countries where EUBUCCO is not based on governmental datasets but solely on OSM (Greece and Sweden), EUBUCCO shows an overlap with OSM of 97% and 99%, respectively, which is lower than 100% due to the updates of OSM happened after EUBUCCO's release in 2022.

Belgium	**EUBUCCO**	**DBSM**	**MS**	**OSM**
EUBUCCO	100	86	71	81
DBSM	88	100	71	89
MS	76	74	100	69
OSM	94	99	74	100

Denmark	**EUBUCCO**	**DBSM**	**MS**	**OSM**
EUBUCCO	100	94	71	93
DBSM	93	100	70	96
MS	79	79	100	77
OSM	96	100	71	100

Greece	**EUBUCCO**	**DBSM**	**MS**	**OSM**
EUBUCCO	100	99	68	99
DBSM	20	100	88	31
MS	14	89	100	24
OSM	64	98	74	100

Malta	**EUBUCCO**	**DBSM**	**MS**	**OSM**
EUBUCCO	100	69	65	27
DBSM	74	100	86	36
MS	73	90	100	25
OSM	81	98	67	100

Sweden	**EUBUCCO**	**DBSM**	**MS**	**OSM**
EUBUCCO	100	98	57	97
DBSM	58	100	67	70
MS	39	78	100	45
OSM	84	99	56	100

Table 2. Percentage of the area of each dataset (in the row) represented by the intersection between each couple of datasets (in the row and in the column), for the five countries analysed.

The comparison between OSM and MS is particularly relevant, as these are the two primary sources that served to produce the other datasets. In this regard, we found that their intersection levels vary across countries and urbanisation levels. Across all five countries, the average area of intersection between the two datasets corresponds to 68% of the total area of OSM, but only to 48% of the total area of MS. This may be explained by the fact that MS shows a higher number of buildings and total areas

in most of the countries (see Table 1), also as a result of its generation process based on machine learning. Therefore, some of the buildings from MS (which might or might not correspond to buildings in the real world) are not available in OSM. This is particularly evident for Greece and Malta, where the number and total area of buildings in MS are significantly higher than in OSM (see Table 1). In these cases, the shared area between the two datasets represents 74% and 67% of OSM coverage, respectively, and just 24% and 25% of MS total area. Another interesting result is found for Sweden, where the values of shared areas represent only 56% of OSM coverage and 45% of MS coverage. This highlights that almost half of the building areas of each dataset (OSM and MS) do not overlap at all with the building areas of the other dataset.

Finally, once again the degree of urbanisation influences the results, in particular the fractions of the areas of OSM and MS represented by the area of their intersection. In most countries and as expected from literature, their similarity is higher in urban areas. The largest values for NUTS3 urban areas are observed in Denmark, where the area of intersection reaches 84% of the total area of OSM and 82% of the total area of MS. In the rural areas of this country, however, these values drop to 68% and 75%, respectively.

5. Discussion and conclusions

To the authors' knowledge, this work represents the first study focused on comparing some of the recently-emerged non-governmental open building datasets. Despite being limited to a few countries only, all located in the EU, the results of the analysis already shed some light on the pros and cons of these datasets and allow us to draw some preliminary conclusions on the opportunities and challenges of using them in real-world applications. A first important consideration is that assessing the quality of each building dataset, and hence determining which of them is the 'best' one, was not only out of scope for this work, but also hard if not possible at all. The four building datasets analysed in this work (OSM, EUBUCCO, DBSM and MS) basically depend on three underlying data sources: governmental data (when available and open), OSM (crowdsourced data) and MS (machine learning-generated data). Governmental data is generally considered to be a reference source and therefore of high quality, but — compared to alternative sources (e.g. OSM and MS) — it may not equal their extremely high level of detail and frequency of update (Antoniou and Skopeliti, 2015). In contrast, OSM traditionally suffers from quality biases depending on the actual presence and activity of national or regional communities (Hecht et al., 2013; Fan et al., 2014; Fram et al., 2015; Brovelli et al., 2016). Finally, MS relies on machine learning algorithms, whose accuracy may vary depending on factors such as the resolution of the image and the peculiar characteristics of the area. As an example of the broad variability of results when comparing OSM and MS, even within the same country, Figure 4 shows two examples from Malta. In the first, although the OSM and MS building polygons correspond to the same real-world buildings, they were most probably derived using different satellite imagery as input source, with the result that the intersection between the datasets only represents a small portion of the area of each dataset when taken alone. In the second example, a densely-built area is mapped very roughly (with a single polygon outlining the whole area) in OSM, while each specific building is available in MS. The result in this case is that the intersection between the two datasets corresponds almost exactly to the area of MS buildings. Hence, it is clear that the 'absolute' quality of each building dataset is strongly dependent on the specific area where it is measured and can hardly, if not at all be generalised for the whole dataset.

Figure 4. Extracts of building polygons from OSM and MS, and their intersection, for two regions in Malta.

The discussion above explains why a 'relative' comparison between the datasets aimed to identify common trends and patterns, is probably the most effective approach to undertake and the one we recommend for future studies. It also shows that analysing five countries only, while on the one hand representing an intrinsic limitation of the study, could on the other hand already lead to relevant conclusions. Based on the results described in Section 4, we provide some final reflections on the building datasets analysed and related recommendations for users in need of such datasets for their real-world applications. First, governmental building datasets such as those used in EUBUCCO usually represent a reliable source, but they can suffer from licensing issues (preventing their reuse for e.g. commercial applications) and/or outdated content. The presence of an active OSM community, which may vary even at the regional or local level, is a prerequisite for confidently relying on OSM buildings, with the recommendation to always make use of the latest version that captures all updates. Similarly, an OSM building database enriched with the import of a governmental dataset would typically represent a good choice. Finally, the results confirm that the completeness of OSM is higher when moving from rural to urban areas. Different considerations can be made about MS buildings, which show heterogeneous behaviours across the analysed countries. While MS appears as a rather complete dataset (i.e. it seems to overall include large portions of the real-world buildings), the sometimes poor comparison against the other building datasets analysed in the study questions its positional accuracy. Further work will be needed to explicitly analyse this dimension. Finally, combining OSM and MS, DBSM overcomes the potential issues of both these datasets and maximises the chance that real-world buildings are actually captured. This makes it fit-for-purpose for applic-

ations requiring high degrees of completeness such as disaster response. More in general, combining multiple open datasets into new (and better) products is a practice that will most probably become commonplace at all levels, from the global (see e.g. Overture Maps) to the national and regional (see e.g. Sarretta et al., 2023).

Future research could explore a number of directions. First, the methodology developed in the study could be extended to the continental scale (e.g. to the whole EU) in order to validate the results achieved. Similarly, the analysis could be extended to other open building datasets available at the continental or global scale, including Google's Open Buildings and Overture Maps. Stemming from the results of this study, future work would also be needed to better understand the reasons leading to the deviation in the geometrical comparison between the building datasets (see section 4). On the one hand, studying the geometric overlap based on each individual building would provide insights into how different data collection methods (from crowdsourcing to machine learning, or governmental data collection) respond to granularity and accuracy in building footprints. On the other hand, further analysis would be needed to understand the extent to which each dataset represents different building entities. This aspect could help provide recommendations around the complementing logic of various building datasets, the relevance of new building datasets built upon diverse input sources (as in the case of DBSM) as well as the underlying data integration/conflation methods.

Disclaimer

The views expressed are purely those of the authors and may not in any circumstances be regarded as stating an official position of the European Commission.

References

Antoniou, V., Skopeliti, A., 2015. Measures and indicators of VGI quality: An overview. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, II-3/W5, 345–351. https://doi.org/10.5194/isprsannals-II-3-W5-345-2015.

Brovelli, M. A., Minghini, M., Molinari, M. E., Zamboni, G., 2016. Positional accuracy assessment of the OpenStreetMap buildings layer through automatic homologous pairs detection: The method and a case study. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLI-B2, 615–620. https://doi.org/10.5194/isprsarchives-XLI-B2-615-2016.

European Commission, 2020. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: A European strategy for data. COM(2020) 66 final. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020DC0066.

Fan, H., Zipf, A., Fu, Q., Neis, P., 2014. Quality assessment for building footprints data on OpenStreetMap. *International Journal of Geographical Information Science*, 28(4), 700–719. http://doi.org/10.1080/13658816.2013.867495.

Farrell, E., Minghini, M., Kotsev, A., Soler Garrido, J., Tapsall, B., Micheli, M., Posada Sanchez, M., Signorelli, S., Tartaro, A., Bernal Cereceda, J., Vespe, M., Di Leo, M., Carballa Smichowski, B., Smith, R., Schade, S., Pogorzelska, K., Gabrielli, L., De Marchi, D., 2023. European Data Spaces - Scientific Insights into Data Sharing and Utilisation at Scale.

Florio, P., Giovando, C., Goch, K., Pesaresi, M., Politis, P., Martinez, A., 2023. Towards a pan-EU building footprint map based on the hierarchical conflation of open datasets: the Digital Building Stock Model-DBSM. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-4/W7-2023, 47–52. https://doi.org/10.5194/isprs-archives-XLVIII-4-W7-2023-47-2023.

Fram, C., Chistopoulou, K., Ellul, C., 2015. Assessing the quality of OpenStreetMap building data and searching for a proxy variable to estimate OSM building data completeness. *Proceedings of the GIS Research UK (GISRUK)*, 195–205.

Gonzales, J. J., 2016. Building-Level Comparison of Microsoft and Google Open Building Footprints Datasets. *Proceedings of the 12th International Conference on Geographic Information Science (GIScience 2023)*, 35:1–35:6. https://doi.org/10.4230/LIPIcs.GIScience.2023.35.

Hecht, R., Kunze, C., Hahmann, S., 2013. Measuring completeness of building footprints in OpenStreetMap over space and time. *ISPRS International Journal of Geo-Information*, 2(4), 1066–1091. https://doi.org/10.3390/ijgi2041066.

Kotsev, A., Minghini, M., Cetl, V., Penninga, F., Robbrecht, J., Lutz, M., 2021. INSPIRE: A Public Sector Contribution to the European Green Deal Data Space. EUR 30832 EN, Publications Office of the European Union, Luxembourg. https://doi.org/10.2760/8563.

Milojevic-Dupont, N., Wagner, F., Nachtigall, F., Hu, J., Brüser, G. B., Zumwald, M., Biljecki, F., Heeren, N., Kaack, L. H., Pichler, P.-P., Creutzig, F., 2023. EUBUCCO v0.1: European Building Stock Characteristics in a Common and Open Database for 200+ Million Individual Buildings. *Scientific Data*. https://doi.org/10.1038/s41597-023-02040-2.

Mooney, P., Minghini, M., 2017. A review of OpenStreetMap data. G. Foody, L. See, S. Fritz, P. Mooney, A.-M. Olteanu-Raimond, C. C. Fonte, V. Antoniou (eds), *Mapping and the Citizen Sensor*, Ubiquity Press, 37–59. https://doi.org/10.5334/bbf.c.

Neis, P., 2024. OSM Stats - Countries. http://www.osmstats.neis-one.org/?item=countries (15 May 2024).

OpenStreetMap Wiki, 2015. Stats. https://wiki.openstreetmap.org/wiki/Stats (26 April 2024).

Sarretta, A., Napolitano, M., Minghini, M., 2023. OpenStreetMap as an input source for producing governmental datasets: The case of the Italian Military Geographic Institute. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-4/W7-2023, 193–200. https://doi.org/10.5194/isprs-archives-XLVIII-4-W7-2023-193-2023.

Ullah, T., Lautenbach, S., Herfort, B., Reinmuth, M., Schorlemmer, D., 2023. Assessing completeness of OpenStreetMap building footprints using MapSwipe. *ISPRS International Journal of Geo-Information*, 12(4), 143. https://doi.org/10.3390/ijgi12040143.