Towards a comparison of pan-European open building data

Marco Minghini, Sara Thabit Gonzalez, Lorenzo Gabrielli, Patrizia Sulis

European Commission, Joint Research Centre (JRC), Ispra, Italy

Building data together: Building-related open microdata – OECD, Paris, 24 January 2025

Introduction

Context – Technology

- Building footprints (from now on buildings) are key geospatial datasets for several **use cases**
  - city planning, demographic analyses, modelling energy production/consumption, disaster preparedness/response, digital twins
- Buildings are traditionally produced, curated & updated by **governmental organisations**
  - National Mapping/Cadastral Agencies responsible for Spatial Data Infrastructures (SDIs)
- Recent technological trends have seen **other players** become producers of building datasets
  - private sector, research/academia, citizen-generated data initiatives

Context – Policy & Scientific interest

- JRC scientific support to EU policies on **data sharing** from two perspectives
  - **technical** enablers – data sources, standards, infrastructures, technologies
  - **organisational** enablers – business models, incentives, agreements, governance schemes
- (non sector-specific) EU policies relevant to building data
  - **INSPIRE Directive** (2007)
    - sets the framework to create a pan-European SDI
  - **Open Data Directive** (2019) & **Implementing Act on high-value datasets** (2023)
    - defines public sector data to be provided for free, under open licenses & through APIs
  - **Common European data spaces**, envisioned in the **European strategy for data** (2020)
    - sectorial, sovereign, interoperable, secure, federated data sharing

Scope, methods & results

Objective

- **Identify** existing building datasets
  - from governmental and **non-governmental organisations**
  - available (at least) at the **continental scale**, with a focus on the EU
  - accessible under an **open license**
- **Analyse** and **compare** the building datasets
  - characterisation based on a self-developed assessment framework
  - comparison through quantitative analyses
- Derive insights on **implications** for policy-relevant use cases
  - which dataset(s) can better address which policy need

Open building footprints

- From **public sector-led** initiatives
  - EU INSPIRE & Open Data Directives
- From **community-led** initiatives
  - OpenStreetMap (OSM)
- From **industry-led** initiatives
  - Microsoft Global ML Building Footprints (MS)
  - Google Open Buildings
  - Overture Maps
- From **research-led** initiatives
  - EUBUCCO
  - Digital Building Stock Model (DBSM)

MCC Mercator Research Institute on Global Commons and Climate Change

Joint Research Centre JRC

European Commission

Building characterisation – Methodology

- **Assessment framework**
  - inspired by & adapted from previous work
  - composed of **3 dimensions** and **13 attributes**

Building characterisation – Methodology

- **Assessment framework**
  - inspired by & adapted from previous work
  - composed of **3 dimensions** and **13 attributes**

1. DATA QUALITY
Attributes
Geographic coverage and completeness
Granularity, shape and positional accuracy
Timeliness
Semantic content
Quality assurance mechanisms and indicators

Building characterisation – Methodology

- **Assessment framework**
  - inspired by & adapted from previous work
  - composed of **3 dimensions** and **13 attributes**

1. DATA QUALITY	2. DATA USABILITY
Attributes	***Attributes***
Geographic coverage and completeness	Data findability
Granularity, shape and positional accuracy	Data accessibility
Timeliness	Data interoperability and manageability
Semantic content	Licence and reusability
Quality assurance mechanisms and indicators	

Building characterisation – Methodology

- **Assessment framework**
  - inspired by & adapted from previous work
  - composed of **3 dimensions** and **13 attributes**

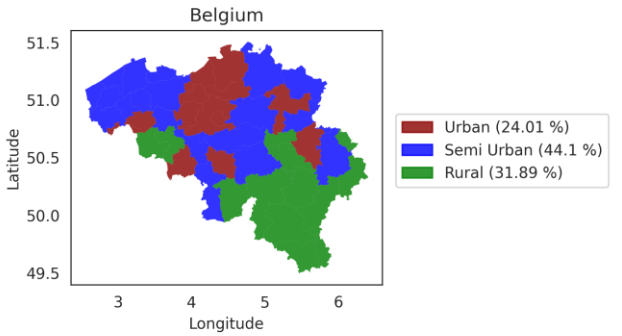
1. DATA QUALITY	2. DATA USABILITY	3. GOVERNANCE
Attributes	***Attributes***	***Attributes***
Geographic coverage and completeness	Data findability	Organisational structure
Granularity, shape and positional accuracy	Data accessibility	Business model and sustainability
Timeliness	Data interoperability and manageability	Openness, transparency and reproducibility
Semantic content	Licence and reusability	Community engagement
Quality assurance mechanisms and indicators		

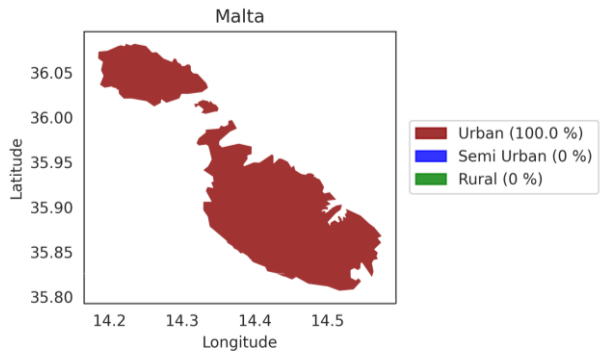
Building characterisation – Results

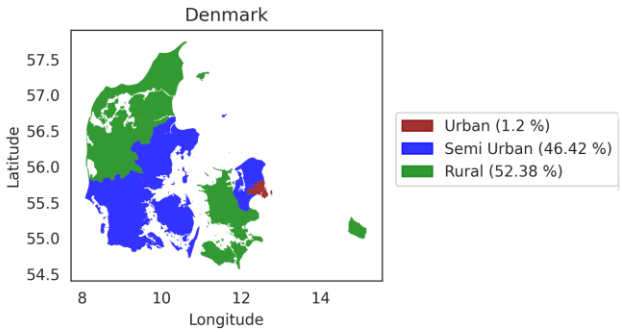
		PUBLIC SECTOR-LED	COMMUNITY-LED	INDUSTRY-LED			RESEARCH-LED	
		INSPIRE & Open Data Directive	OpenStreetMap (OSM)	Microsoft GlobalML Building Footprints	Google Open Buildings	Overture Maps	JRC Digital Building Stock Model (DBSM)	EUBUCCO
Data quality	*Geographic coverage and completeness*	Typically high	Influenced by demographic, socioeconomic, and contributor factors	Influenced by image resolution or distortions.	Influenced by image resolution or distortions.	Relatively higher than non-conflated datasets	Relatively high completeness compared with cadastral data	High variations per country, influenced by the dataset used
	Granularity, shape and positional accuracy	Tipically high	Influenced by demographic, socioeconomic, and contributor factors	Influenced by the specific LM algorythm and buildings characteristics	Influenced by the specific LM algorythm and buildings characteristics	*Insufficient existing comparative analysis*	Relatively high completeness compared with cadastral data	High variations per country, influenced by the dataset used
	Timeliness	Periodic releases by country	Constant/Real-time updates	Monthly/Bimonthly releases	Monthly/Bimonthly releases	Monthly/Bimonthly releases	Single release (Oct/2023)	Single release (Nov/2022)
	Semantic content	Different list of attributes per country, comprehensive	High number of attributes, partial coverage	Only building height, partial coverage	No semantic content	Multiple, imported from secondary sources	No semantic content	Height, use, and building age. Different coverage depending on the country/dataset
	Accuracy indicators and quality assurance mechanisms	Standardised data collection protocols and audits	Community of peers & academic analysis	AI-developed "Confidence score" per building	AI-developed "Confidence score" per building	No	No	No
Data usability	*Data Findability*	Different access tools depending on national portals	Highest variety of pre-visualization and search tools and platforms	No pre-visualization, only API search	Pre-visualization and search tools	Pre-visualization, but search tools only with account	No pre-visualization or search tools	No pre-visualization, only search tools
	Data Accessibility		Highly accessible by different user profiles	Limited to skilled users	Accessible by average user	Accessible by average user	Limited to high computational power	Accessible by average user
	Data Interoperability and Manageability	No cross-border harmonization; standard GIS formats	Harmonized data; highest variety of data formats	Harmonized data. Export only as .csv.gz	Harmonized data. Exports as .csv and .geojson	Harmonized data. Exports only as GeoParquet	Harmonized data. Exports as RDF/JSON, RDF/XML, Turtle formats.	Harmonized data. Exports in .gpkg and .csv
	Licence and reusability	CC0, CC BY-4.0 or equivalent or less restrictive license	ODbL	ODbL	CC BY-4.0 and ODbL	CDLA except from data derived from OSM (ODbL)	ODbL	ODbL except of 2 databases
Governance	*Organisational structure*	Hybrid (Top-down & bottom-up)	Bottom-up (open)	Top-down (closed)	Top-down (closed)	Top-down (open via annual fee)	Top-down (closed)	Top-down (closed)
	Business model and sustainability	Public sector funded	Foundation funding & voluntary mapping	Corporate funding	Corporate funding	Foundation funding	Project funding (limited timing)	Project funding (limited timing)
	Openness, transparency and reproducibility	Transparent and reproducible production process	Transparent and reproducible production process	Insufficient information for reproduction or process analysis	Insufficient information for reproduction or process analysis	Insufficient information for reproduction or process analysis	Insufficient information for reproduction or process analysis	Transparent and reproducible production process
	Community engagement	EU-led community of public sector collaborators	Very large open community	Not specified	Not specified	Payment-based membership tiers	Limited community of researchers	Limited community of researchers

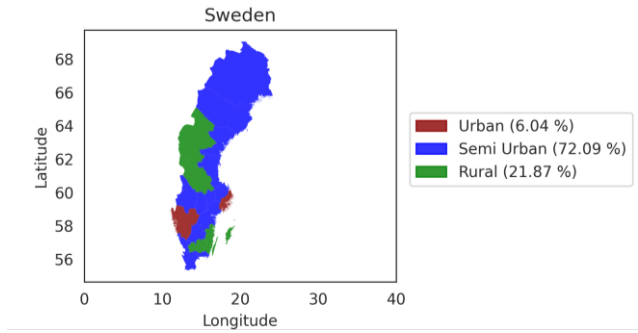
Quantitative comparison – Methodology

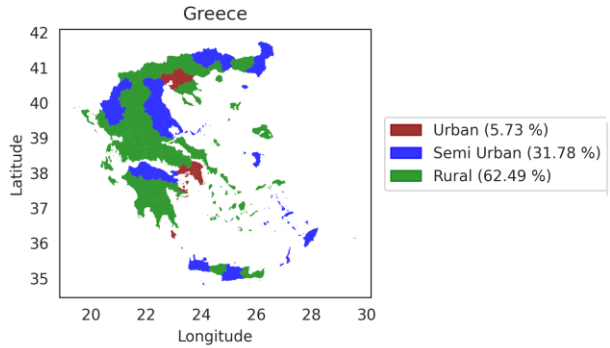
- Focused on the **geometry** of buildings
  - taking into account the **degree of urbanisation** (***urban***, ***semi-urban***, ***rural*** based on NUTS3)
  - limited to **4 building datasets**: OSM, EUBUCCO, MS, DBSM
  - limited to **5 EU countries**: Belgium, Denmark, Greece, Malta, Sweden







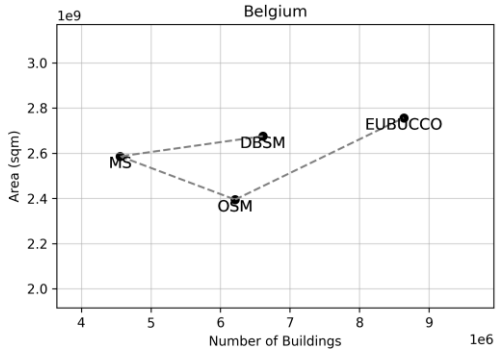


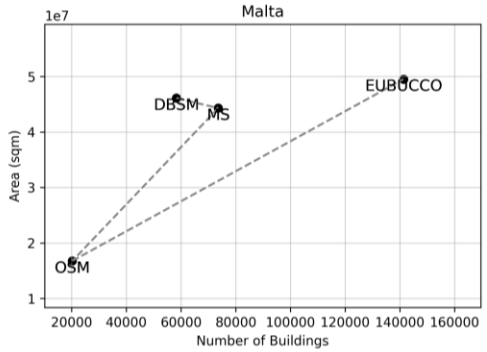


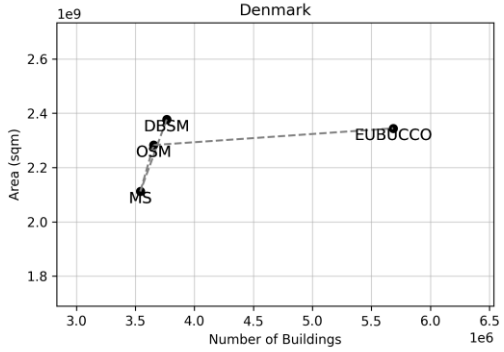
European Commission

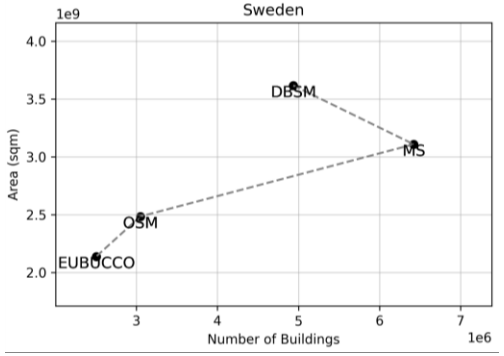
Total number & total area of buildings

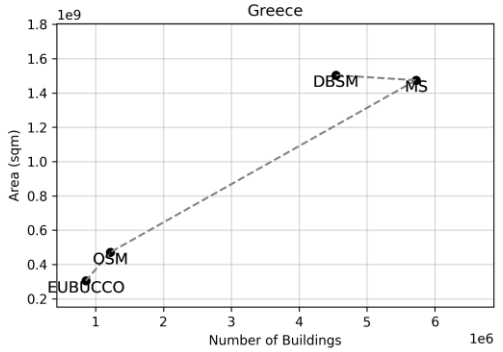
Dataset	Country	Number of buildings	Area of buildings [10^8 m^2]
EUBUCCO	Belgium	8,636,114	27.56
	Denmark	5,684,734	23.44
	Greece	856,140	3.04
	Malta	141,329	0.49
	Sweden	2,504,961	21.38
OSM	Belgium	6,211,451	23.94
	Denmark	3,654,875	22.82
	Greece	1,217,547	4.71
	Malta	20,225	0.16
	Sweden	3,050,667	24.84
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







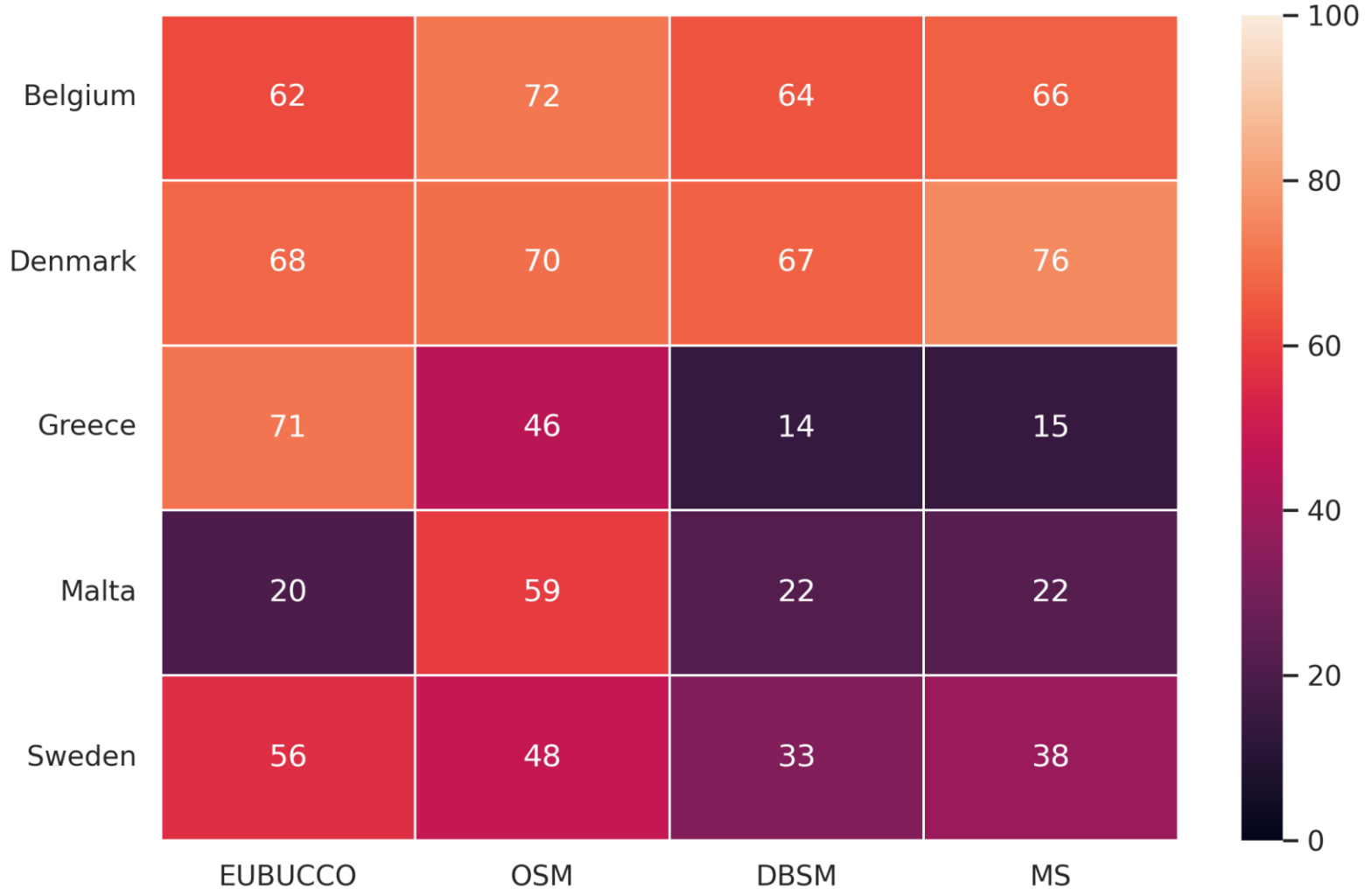




European Commission

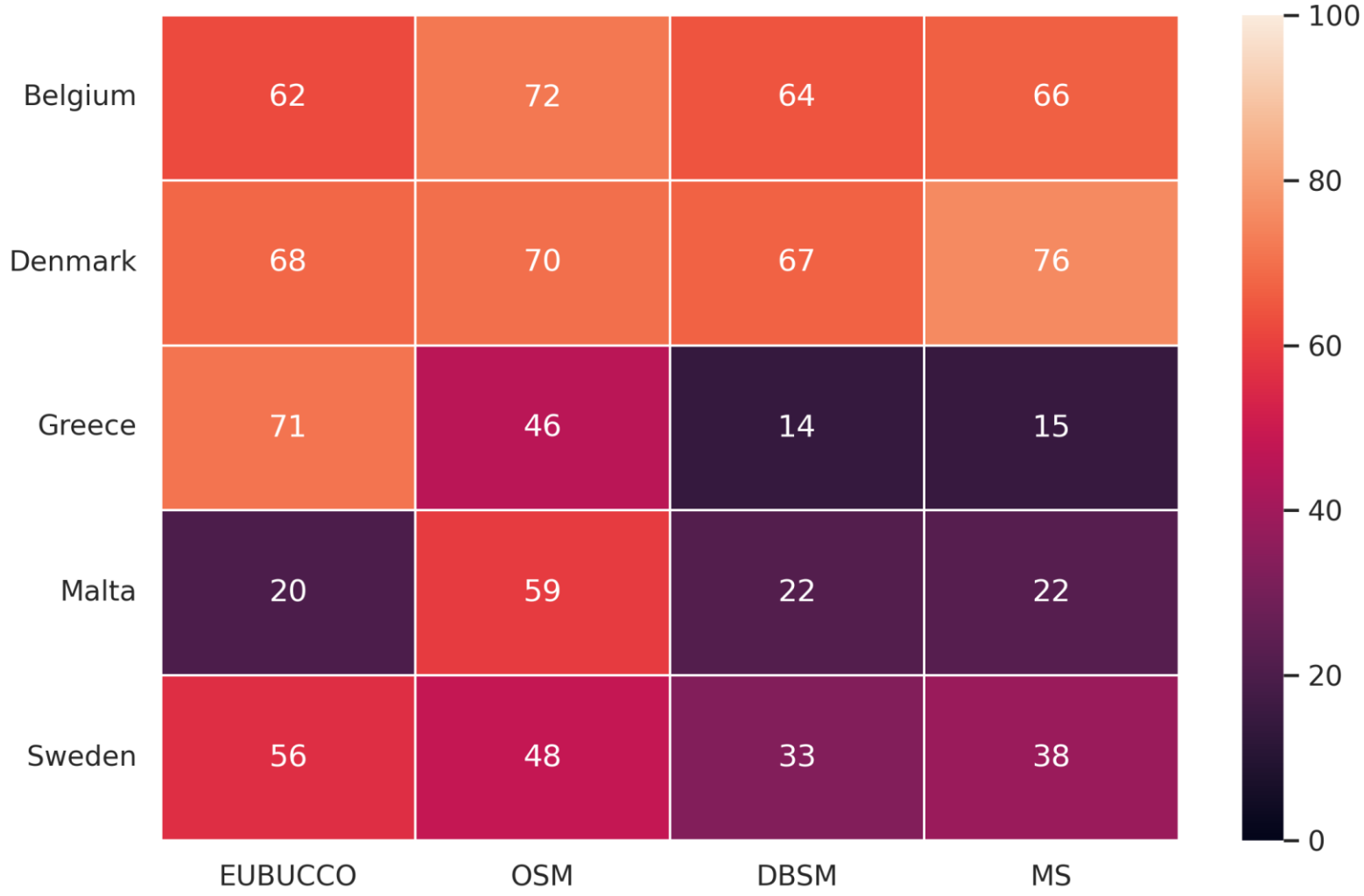
Similarity of building datasets

- Intersection between all the 4 datasets
  - % of the area of each dataset represented by the area of intersection between the 4 datasets



Similarity of building datasets

- Intersection between all the 4 datasets
  - % of the area of each dataset represented by the area of intersection between the 4 datasets
  - % **lower in rural areas** (minimum 7%) and **higher in urban areas** (maximum 79%)



Similarity of building datasets

- Intersection between each couple of datasets
  - % of the area of the dataset in the row, represented by the area of intersection between the dataset in the row and the dataset in the column

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

Similarity of building datasets

- Intersection between each couple of datasets
  - % of the area of the dataset in the row, represented by the area of intersection between the dataset in the row and the dataset in the column
  - **similarity higher in urban areas & lower in rural areas** (OSM vs MS: maximum 84-82%, minimum 68-75% in Denmark)

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

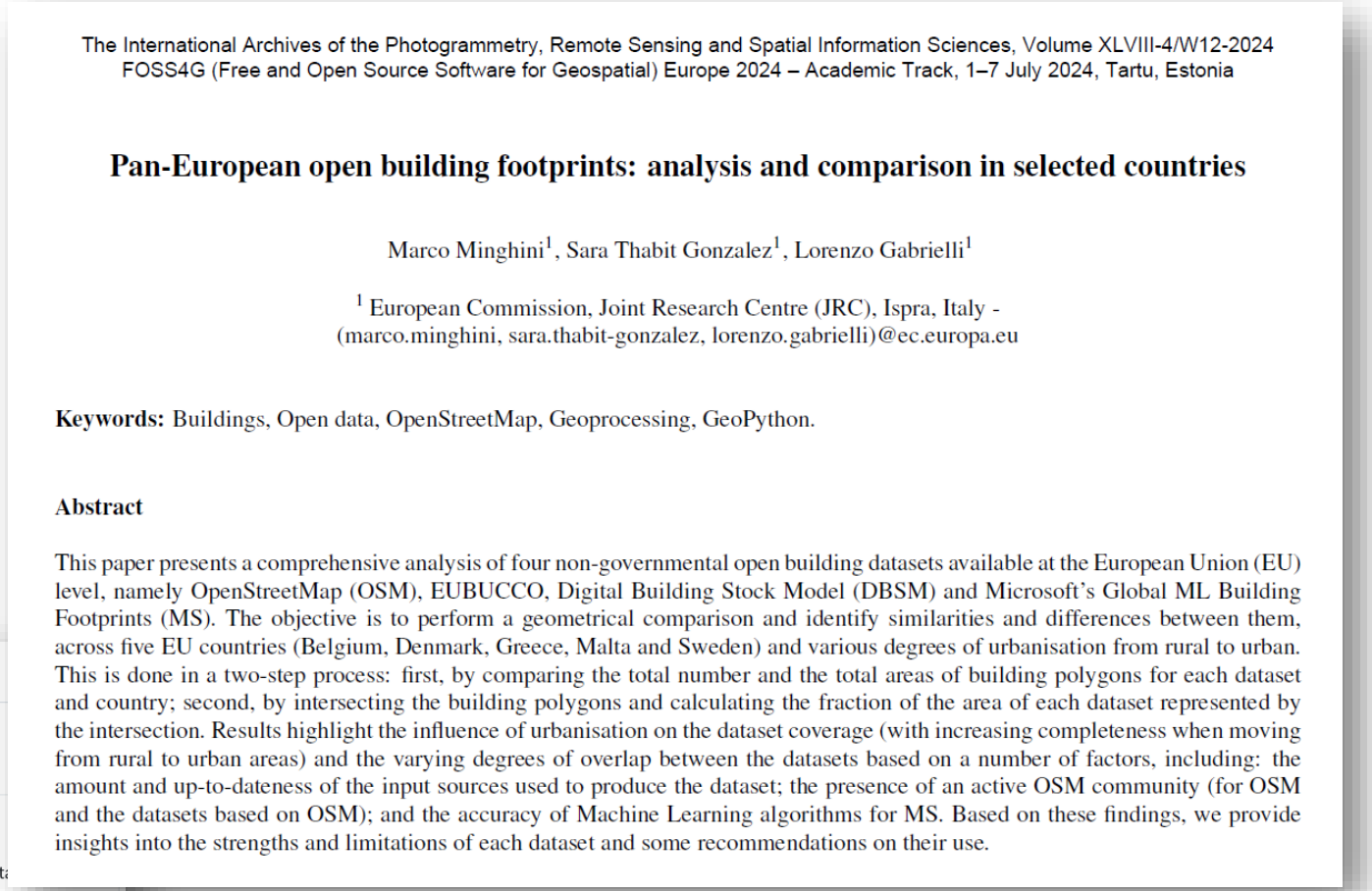
Greece	EUBUCCO	DBSM	MS	OSM
EUBUCCO	100	99	68	99
DBSM	20	100	88	31
MS	14	89	100	24
OSM	64	98	74	100

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

References

- Paper: https://isprs-archives.copernicus.org/articles/XLVIII-4-W12-2024/97/2024



The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-4/W12-2024
FOSS4G (Free and Open Source Software for Geospatial) Europe 2024 – Academic Track, 1–7 July 2024, Tartu, Estonia

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

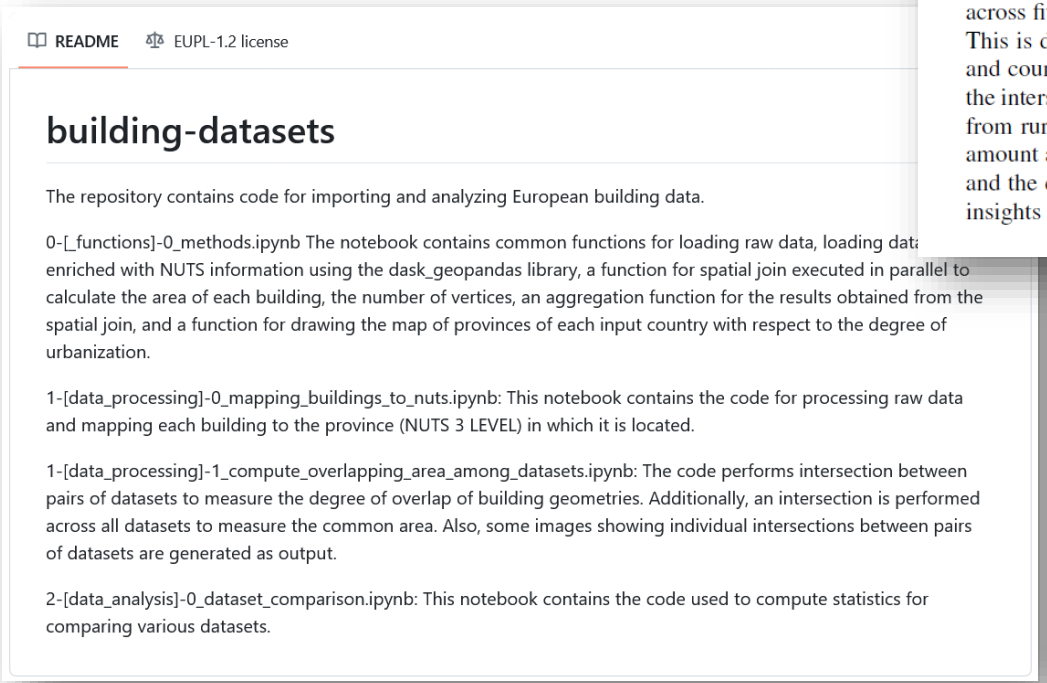
Marco Minghini[1], Sara Thabit Gonzalez[1], Lorenzo Gabrielli[1]

[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.



README EUPL-1.2 license

building-datasets

The repository contains code for importing and analyzing European building data.

0-[_functions]-0_methods.ipynb The notebook contains common functions for loading raw data, loading data enriched with NUTS information using the dask_geopandas library, a function for spatial join executed in parallel to calculate the area of each building, the number of vertices, an aggregation function for the results obtained from the spatial join, and a function for drawing the map of provinces of each input country with respect to the degree of urbanization.

1-[data_processing]-0_mapping_buildings_to_nuts.ipynb: This notebook contains the code for processing raw data and mapping each building to the province (NUTS 3 LEVEL) in which it is located.

1-[data_processing]-1_compute_overlapping_area_among_datasets.ipynb: The code performs intersection between pairs of datasets to measure the degree of overlap of building geometries. Additionally, an intersection is performed across all datasets to measure the common area. Also, some images showing individual intersections between pairs of datasets are generated as output.

2-[data_analysis]-0_dataset_comparison.ipynb: This notebook contains the code used to compute statistics for comparing various datasets.

- Code: https://github.com/eurogeoss/building-datasets

European Commission

Quantitative comparison – Methodology

- Focused on the **attributes** of buildings [**in progress**]
  - extended to **all building datasets**: OSM, EUBUCCO, MS, Overture, DBSM, GHSL
  - extended to **all EU countries**
  - taking into account the **degree of urbanisation**
- Methodology at a glance
  - map/**harmonise attributes** across datasets
  - for each dataset/country/attribute, **calculate the fraction of buildings with values**
  - for each dataset/country/attribute , **derive statistical distributions of values**
  - assess the **similarity of distributions** by country and dataset
  - assess **correlation with the degree of urbanisation**

Conclusions

Discussion

- First comparison of (some) non-governmental open building datasets
- Relative **comparison** (not quality assessment) of datasets is the way to go
  - **different sources**: governments, citizens, private companies, research institutions
  - **different production/update approaches**: digitalisation, machine learning, conflation
- There is **no best dataset** in general – the choice depends on the specific area & use case/application
- (Regular) **conflation** of multiple datasets is the most promising approach
- **Governance** (transparency, inclusivity & sustainability) considerations
- How will **production of open data** look like in the future?
  - non-governmental initiatives hold potential to **enhance public sector data**
  - which **role** for which **actor**? any chance for **collaboration**?

Thank you!

marco.minghini@ec.europa.eu

EU Science Hub
joint-research-centre.ec.europa.eu

@EU_ScienceHub
EU Science Hub – Joint Research Centre
EU Science, Research and Innovation
EU Science Hub
@eu_science