Estimating the age of buildings from satellite and morphological features to create a pan-EU digital building stock model

Jeremias Wenzel, Ana M. Martinez, Pietro Florio, Katarzyna Goch
European Commission, Joint Research Centre
j.h.m.wenzel@student.utwente.nl, ana.martinez@ec.europa.eu
pietro.florio@ec.europa.eu, kgoch@twarda.pan.pl

Abstract

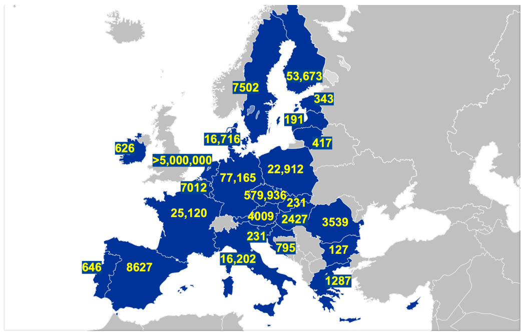
The acceleration in the effects of global warming and the recent turbulences in the energy market are further highlighting the need to act quicker and smarter in terms of decisions to transition to greener energy and reduce our overall energy consumption. With buildings accounting for about 40% of the energy consumption in Europe, it is crucial to have a comprehensive understanding of the building stock and their energy-related characteristics, including their age, in order to make informed decisions for energy savings. This study introduces a novel way to approach the age estimation of buildings at scale, using a machine learning method that integrates satellite-based imagery with morphological features of buildings. The findings demonstrate the benefits of combining these data sources and underscore the importance of incorporating local data to enable accurate prediction across different cities.

1 Introduction

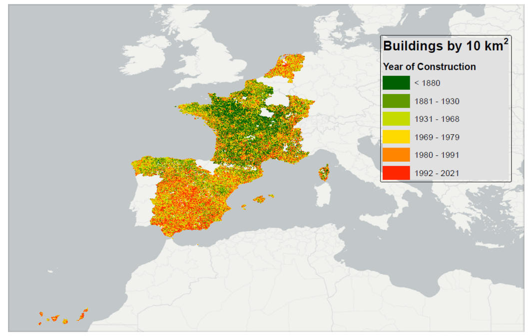
Buildings are significant contributors of the carbon emissions at global scale and particularly in Europe (Economidou et al., 2011). Supporting policy makers in developing legislation, propose concrete actions and monitor results to decarbonise the building stock, access to high-resolution building data is essential. Furthermore, it is estimated that 40 million Europeans were unable to heat their homes on 2022 (Eurostat, 2023), underscoring the critical importance of directing investments to where they are most needed. Therefore, transitioning from broad measures that benefit the majority to more informed decisions based on detailed data is imperative.

Access to building and energy-related data necessary for effective policy design at the European level is often limited or inconsistent. Information about buildings in the EU is typically available as aggregated statistics, such as the Building Stock Observatory (BSO) (Commission, 2016-2023) and Hotmaps (Pezzutto et al., 2018), or as coarse raster maps from the Global Human Settlement Layer (GHSL) (Pesaresi, 2023). However, there is a need to identify buildings at the individual unit level. Recent efforts in this direction include the publicly available EUBUCCO (Milojevic-Dupont et al., 2023) , DBSM (Florio et al., 2023), and Overture (Foundation, 2023) databases. The construction period of a building is a crucial indicator for gaining a better understanding of its energy-related characteristics and estimate its energy performance.

Three primary open and public data sources currently provide information on the construction age of individual buildings. The first source is OpenStreetMap (OSM), which offers sparse and unevenly distributed coverage of the EU with crowdsourced information (Contributors, 2022), as can be seen in Figure 1a. The second source is the combination of cadastral sources by GISCO from EUSTAT, which is currently restricted to large parts of Spain, France, and The Netherlands (Gisco, 2023), as shown in Figure 1b. Lastly, EUBUCCO covers 200 million buildings in EU countries, with 24% including the building age. This data encompasses $\sim$ 50M individual buildings, primarily distributed across five countries: France ($\sim$ 21.5M buildings), Spain ($\sim$ 16M buildings), the Netherlands ($\sim$ 10M buildings), Italy ($\sim$ 1.5M buildings), and Finland ($\sim$ 50K buildings). The data was obtained from national cadastral and municipal datasets.



(a) Number of buildings in Europe with information on age of construction from OSM



(b) Year of Construction of buildings from GISCO (EUSTAT)

In the literature, various methodologies have been proposed for deriving the age of buildings. Li et al. (2018) and Zeppelzauer et al. (2018) utilized Deep Learning models to estimate building age from non-openly available images sourced from Google Street View images, real estate evaluation reports, and web platforms. Garbasevschi et al. (2021) adopted a different approach, training a Random Forest classification model based on visual features of buildings, including geometric attributes and location-contextual information for selected cities in Germany. The most recent methodology, introduced by Nachtigall et al. (2023), demonstrated the feasibility of large-scale building age prediction using morphology features. The objective of our study is to derive an initial estimate of building construction age at the EU scale using openly available datasets. To achieve this, we aim to utilize non-commercial data and methodologies capable of scaling up, particularly those reliant on Earth Observation data. In this paper, we employ a traditional machine learning algorithm to integrate multi-temporal built-up surface grid rom the GHSL Global Human Settlement Layer and morphological features extracted from OSM. Our study demonstrates that this combined approach yields superior results compared to using either data source in isolation.

2 Methodology

A novel solution has been developed, integrating data on the age of building construction obtained from satellite imagery from the GHSL and morphological features derived from OSM. These two streams of information are combined using a supervised machine learning model to estimate the age of building construction.

2.1 Age of construction from GSHL multi-temporal built-up surface grid

The GHSL Global Human Settlement Layer project (European Commission, 2023) offers comprehensive global spatial data on human presence on the planet. Their products encompass built-up maps, population density maps, and settlement maps, among others. One of their key products is the multi-temporal built-up surface grid GHS-BUILT-S(Pesaresi, 2023), which is based on a series of temporally spaced-out Landsat and Sentinel-2 imagery that tracks the development of settlement growth over time. By comparing the relative difference in built-up areas between observations, insights into the age of added built-up areas can be gained, allowing for the identification of buildings constructed between 1975 and 2020. The resolution is harmonized to 100m, which restricts the ability to identify individual buildings or small additions. However, this approach aims to balance accuracy and biases over the entire age range. The GHS-BUILT-S provides consistent distribution of built-up surfaces in 5-year intervals. The process involves first identifying built-up areas in the Sentinel-2 image composite for year 2018 (Corbane et al., 2020) and then checking

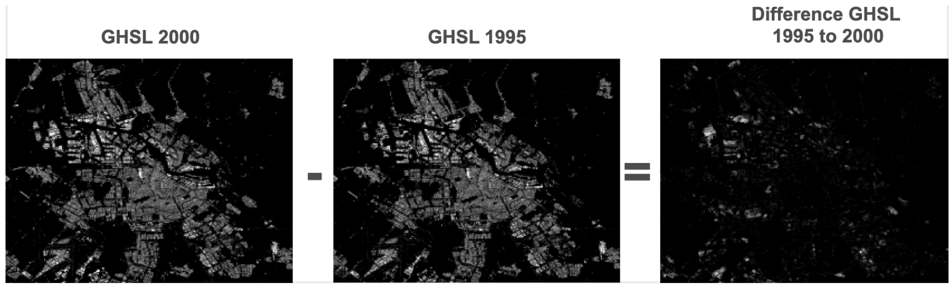


Figure 2: Built-up surfaces in period 1995-2000 derived from the multi-temporal built-up surface grid GHS-BUILT-S.

the blocks of identified built-up areas in subsequent steps. However, this approach has limitations, as it can only capture increases in built-up areas and cannot detect reductions, and it is unable to identify building areas that are re-built or renovated. Nevertheless, the advantage of this method is that it provides coverage at a global scale.

The subtraction of consecutive 5-year images from GHSL was performed to extract the most probable age of construction of a building, as illustrated in Figure 2.

2.2 Morphological features from OpenStreetMap

The morphological features extracted to drive the prediction of building age are shown in Table 1, based on the most significant ones identified by Nachtigall et al. (2023):

Building features	Neighbourhood features
Shape Complexity	Distance to closest neighbour
House element count	Number of adjoining buildings
House area	Distances to closest street and intersection
	Building count in 20m, 100m and 500 range

Table 1: Morphological predictive features derived from OSM.

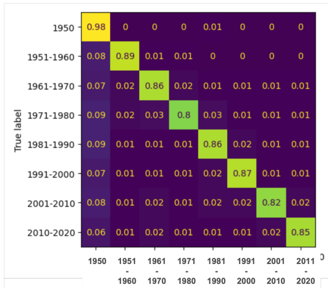
Building age labels for model training are also extracted from OSM, which are considered to be the closest to ground truth in this case (Figure 1a). The Netherlands stands out with the highest number of buildings with this information. Additionally, we observed that the samples are not representative in countries with a small number of labels, such as Italy, as most of them correspond to churches or other historic buildings. As a result, we have chosen to focus on two cities in The Netherlands for our feasibility study, specifically Amsterdam and Rotterdam. Amsterdam boasts a total of 249,546 buildings, and the building age distribution is presented in Table 2.

Table 2: Number of buildings per Age Category

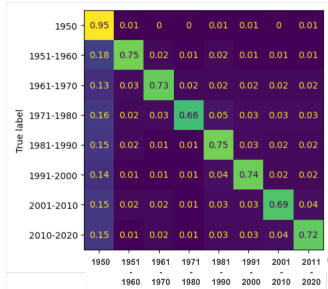
Pre 1950	1951-1960	1961-1970	1971-1980	1981-1990	1991-2000	2001-2010	2011-2020
95,693	29,073	18,871	13,992	29,821	27,872	17,301	16,923

3 Experimental Results

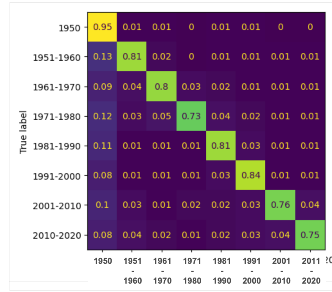
We trained a Random Forest model on 80% of the available data for Amsterdam. The confusion matrix in Figure 3a illustrates the model's performance on the test set from the remaining 20% of the data when using all predictive attributes, including information from GHS-BUILT-S multi-temporal built-up surface grid and morphological features derived from OSM. Upon removing GHS-BUILT-S features, performance deteriorates across all classes, as demonstrated in Figure 3b. Additionally, Figure 3c shows that when only GHS-BUILT-S features are utilized, the performance is not as robust as when all features are considered, but it is also not as poor as when only morphological attributes are considered.



(a) Includes GHSL and morphological features

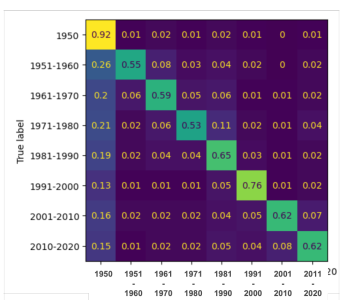


(b) Includes only morphological features

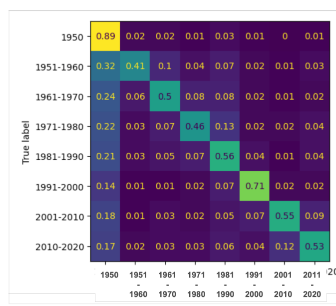


(c) Includes only GHSL features

We are particularly interested in exploring the model's ability to generalize to other geographic locations. To this end, we have chosen Rotterdam as an additional test case. After training a separate model for Rotterdam using a balanced dataset, we observed that without considering any labels or training data from Amsterdam, the model's performance in predicting the age of buildings in Amsterdam was notably poor. However, when a small amount of local data from Amsterdam (10% of available data) was used to re-train the model, the prediction capability increased, with an accuracy of 73.8%. Notably, when only 5% of the available data from Amsterdam was included in Rotterdam's model, the accuracy decreased to 67.3%. This finding is significant, as it reflects the challenge of limited data availability for building age across most geographical areas.



(a) Amsterdam predicted using 10% of Amsterdam data on the Rotterdam model



(b) Amsterdam predicted using 5% of Amsterdam data on the Rotterdam model

4 Conclusion

This study has demonstrated the effectiveness of combining satellite and morphological features for predicting the age of buildings using Machine Learning. Our feasibility study integrated the multi-temporal built-up surface grid from GHS-BUILT-S and various morphological features from OSM in a Random Forest model. As expected, the model exhibited strong local predictability, showing sensitivity to the use of local data, particularly from different cities within the same country. Notably, we observed that incorporating a small amount of local information could enhance prediction accuracy for a different location.

We view this work as an initial exploration into the combination of satellite and morphological features, aiming to provide a scalable solution for EU coverage. Further investigation is needed to better understand the over-prediction of pre-1950 buildings, test the model's generalization capability to other countries, and explore the use of other ML algorithms, such as graph neural networks, which consider spatial relations of the data. In future models, reference data could also be extracted from EUBUCCO and EUSTAT sources.

Moreover, for energy-related purposes, it is crucial to have information on the age of renovation, if any. Hence, this data should be combined with information from energy performance certificates in the future. We recognize that data availability in terms of geographical coverage in OSM is not uniform, even for extracting morphological features. However, we acknowledge ongoing initiatives such as DBSM, EUBUCCO, and Overture, which aim to address this limitation and enhance data availability.

5 Acknowledgement

Mr. Jeremias Wenzel, currently affiliated with Department of Philosophy, University of Twente (The Netherlands), developed this analysis during his traineeship at the Joint Research Centre of the European Commission, under the supervision of Dr. Ana Martinez and the support of Dr. Pietro Florio and Ms. Katarzyna Goch (the latter also affiliated with Institute of Geography and Spatial Organization Polish Academy of Sciences, Warsaw, Poland).

References

European Commission. Building stock observatory, 2016-2023. URL `http://energy.ec.europa.eu/topics/energy-efficiency/energy-efficient-buildings/eu-building-stock-observatory_en`. Accessed on February, 2024.

OpenStreetMap Contributors. Building data extracted from openstreetmap, 2022. URL `https://www.openstreetmap.org/`. Accessed on February, 2024.

C. Corbane, P. Politis, P. Kempeneers, D. Simonetti, P. Soille, A. Burger, M. Pesaresi, F. Sabo, V. Syrris, and T. Kemper. A global cloud free pixel- based image composite from sentinel-2 data. *Data in Brief*, 31:105737, 2020. ISSN 2352-3409. doi: https://doi.org/10.1016/j.dib.2020.105737. URL `https://www.sciencedirect.com/science/article/pii/S2352340920306314`.

Marina Economidou, Bogdan Atanasiu, Dan Staniaszek, Joana Maio, Ingeborg Nolte, Oliver Rapf, Jens Laustsen, Paul Ruyssevelt, David Strong, and Silvia Zinetti. *Europe's buildings under the microscope. A country-by-country review of the energy performance of buildings*. BPIE, 10 2011. ISBN 9789491143014.

Joint Research Centre European Commission. GHSL data package 2023, publications office of the european union, 2023. URL `https://data.europa.eu/doi/10.2760/098587`. Accessed on February, 2024.

Eurostat. EU-SILC statistics on income and living conditions, 2023. URL `http://ec.europa.eu/eurostat/web/income-and-living-conditions`. Accessed on February, 2024.

P. Florio, C. Giovando, K. Goch, M. Pesaresi, P. Politis, and A. Martinez. 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, 2023. doi: 10.5194/isprs-archives-XLVIII-4-W7-2023-47-2023. URL `https://isprs-archives.copernicus.org/articles/XLVIII-4-W7-2023/47/2023/`.

Overturemaps Foundation. Overturemaps foundation - world-wide open map dataset, 2023. URL `overturemaps.org`. Accessed on February, 2024.

Oana M. Garbasevschi, Jacob Estevam Schmiedt, Trivik Verma, Iulia Lefter, Willem K. Korthals Altes, Ariane Droin, Björn Schiricke, and Michael Wurm. Spatial factors influencing building age prediction and implications for urban residential energy modelling. *Computers, Environment and Urban Systems*, 88:101637, 2021. ISSN 0198-9715. doi: https://doi.org/10.1016/j.compenvurbsys.2021.101637. URL `https://www.sciencedirect.com/science/article/pii/S0198971521000442`.

EUSTAT Gisco. Building data from authoritative sources, 2023. URL `gisco-services.ec.europa.eu/pub/Inspire/ANNEX-1/Buildings/`. Accessed on February, 2024.

Yan Li, Yiqun Chen, Abbas Rajabifard, Kourosh Khoshelham, and Mitko Aleksandrov. Estimating Building Age from Google Street View Images Using Deep Learning. In Stephan Winter, Amy Griffin, and Monika Sester (eds.), *10th International Conference on Geographic Information Science (GIScience 2018)*, volume 114 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 40:1–40:7, Dagstuhl, Germany, 2018. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-083-5. doi: 10.4230/LIPIcs.GISCIENCE.2018.40. URL `https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.GISCIENCE.2018.40`.

Nikola Milojevic-Dupont, Felix Wagner, Florian Nachtigall, Jiawei Hu, Geza Boi Brüser, Marius Zumwald, Filip Biljecki, Niko Heeren, Lynn H. Kaack, Peter-Paul Pichler, and Felix Creutzig. EUBUCCO v0.1: European building stock characteristics in a common and open database for 200+ million individual buildings. *Scientific Data*, 10(1):147, 2023. ISSN 2052-4463. doi: 10.3929/ethz-b-000605346.

Florian Nachtigall, Nikola Milojevic-Dupont, Felix Wagner, and Felix Creutzig. Predicting building age from urban form at large scale. *Computers, Environment and Urban Systems*, 105:102010, 2023. ISSN 0198-9715. doi: https://doi.org/10.1016/j.compenvurbsys.2023.102010. URL `https://www.sciencedirect.com/science/article/pii/S019897152300073X`.

Martino Pesaresi. GHS-BUILT-S R2023A - GHS built-up surface grid, derived from sentinel2 composite and landsat, multitemporal (1975-2030), 2023. URL `http://data.europa.eu/89h/9f06f36f-4b11-47ec-abb0-4f8b7b1d72ea`.

Simon Pezzutto, Stefano Zambotti, Silvia Croce, Pietro Zambelli, Giulia Garegnani, Chiara Scaramuzzino, Ramón Pascual Pascuas, Alyona Zubaryeva, Franziska Haas, Dagmar Exner (eurac), Andreas Müller (e-think), Michael Hartner (tuw), Tobias Fleiter, Anna-Lena Klingler, Matthias Kühnbach, Pia Manz, Simon Marwitz, Matthias Rehfeldt, and Jan Steinbach. Hotmaps project, d2.3 wp2 report – open data set for the eu28. 2018. URL `www.hotmaps-project.eu`.

Matthias Zeppelzauer, Miroslav Despotovic, Muntaha Sakeena, David Koch, and Mario Döller. Automatic prediction of building age from photographs. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, ICMR '18, pp. 126–134, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450350464. doi: 10.1145/3206025.3206060. URL `https://doi.org/10.1145/3206025.3206060`.