Energy and AI 17 (2024) 100408

Contents lists available at ScienceDirect

Energy and AI

journal homepage: www.sciencedirect.com/journal/energy-and-ai

Leveraging machine learning to generate a unified and complete building height dataset for Germany

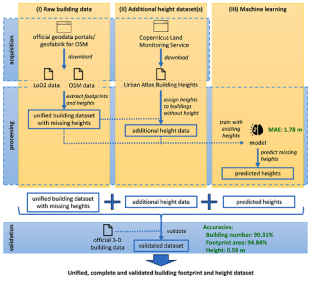
Kristina Dabrock [a,b,*], Noah Pflugradt [a], Jann Michael Weinand [a], Detlef Stolten [a,b]

[a] Forschungszentrum Jülich GmbH, Institute of Climate and Energy Systems, Jülich Systems Analysis, 52425 Jülich, Germany
[b] RWTH Aachen University, Chair for Fuel Cells, Faculty of Mechanical Engineering, 52062 Aachen, Germany

HIGHLIGHTS

- XGBoost machine learning model suitable for imputing missing height data.
- 3-D building data and OpenStreetMap serve as data base.
- Complete nation-wide building footprint and height dataset for Germany.
- Validation shows high overall accuracy with region-dependent error.

GRAPHICAL ABSTRACT



ARTICLE INFO

Original content: ETHOS.BUILDA: Building Footprint and Height Dataset Germany (Original data)

Keywords:
Machine learning
XGBoost
Building height
Building footprint
3-D building data
Geodata
Spatial analysis

ABSTRACT

Building geometry data is crucial for detailed, spatially-explicit analyses of the building stock in energy systems analysis and beyond. Despite the existence of diverse datasets and methods, a standardized and validated approach for creating a nation-wide unified and complete dataset of German building heights is not yet available. This study develops and validates such a methodology, combining different data sources for building footprints and heights and filling gaps in height data using an XGBoost machine learning algorithm. The XGBoost model achieves a mean absolute error of 1.78 m at the national level and between 1.52 m and 3.47 m at the federal state level. The goal is proving the applicability of the methodology at a large scale and creating a useful dataset. The resulting dataset is thoroughly evaluated on a building-by-building level and spatially resolved statistics on the quality of the dataset are reported. This detailed validation found that the building number and footprint area of German building stock is 90.31 % and 94.84 % correct, respectively, and the building height accuracy is 0.59 m at the national level. However, errors are not homogeneous across Germany and further research is needed into the impact of including additional datasets, especially for regions and building types with lower accuracies. This study proves that the chosen methodology is useful for generating a building height dataset and the workflow, with some modifications for regional data availability, can be transferred to other countries. The generated building dataset for Germany constitutes a valuable data basis for the research community in fields such as energy research, urban planning and building decarbonization policy development.

* Corresponding author.
E-mail address: k.dabrock@fz-juelich.de (K. Dabrock).

https://doi.org/10.1016/j.egyai.2024.100408

Available online 31 July 2024

1. Introduction

Buildings contribute significantly to energy consumption and greenhouse gas emissions in the European Union, with shares of 40 % and 36 %, respectively [1]. In order to analyze the feasibility and efficacy of decarbonization measures, detailed knowledge of the building stock is required. When aiming for bottom-up analyses, especially spatially-explicit ones, data on individual building levels is required. However, although there has been a surge in open data initiatives and datasets on building attributes have been published, this data is often fragmented and no standardized methodology for combining it has been established.

Building footprint and height are the key attributes for describing a building's geometry and are crucial for estimating its heating energy demand. Building geometry directly affects, e.g., the wall areas and thereby heat transmission losses of the building envelope. The geometry is also a valuable source of information for deriving other building characteristics. Energy consumption, for example, is often given per square meter of living area of a building, which can be approximated when knowing the footprint area and height. Other use cases include the scaling of building archetypes, e.g., from the Tabula typology [2], to better match the actual buildings characteristics, or estimating the number of apartments per building. Information on building geometry can also be a valuable feature in machine learning tasks, e.g., for classifying building types [3]. Height information therefore is of utmost importance when estimating the energy demand of buildings on a per-building level as it allows, for example, to derive the heated volume of a building, the area of its thermal envelope, and the conditioned living area. These characteristics could not be derived from a building's footprint alone.

Therefore, the existence of complete and validated building datasets on the (inter-)national scales is crucial for energy related analyses and the lack thereof has been noted, e.g., by Bandam et al. [3]. Building height data is available from different sources, in different formats and qualities and can be combined and harmonized. However, the resulting datasets still contain missing data, an issue that has to be addressed by developing novel approaches. Whereas various approaches exist and have been tested for some regions, consolidating, adapting and applying these approaches on larger scales and validating the resulting dataset has not yet been done. The objective of this paper is to develop and evaluate a methodology inspired by previous work in the field of machine learning for creating and, most importantly, validating a unified and complete nation-wide building height dataset in order to fill that gap and facilitate further research in energy system analysis and beyond. Out study therefore lies at the intersection of artificial intelligence and energy analyses, with the former used as a methodological basis and the latter being the main beneficiary of the generated dataset.

The German building stock serves as a case study for this paper as it has the highest final energy consumption of the EU countries, according to the EU building stock observatory [4], and is therefore of particular importance for the energy transition. However, this study is also relevant for other countries as it presents a generalized workflow. While the presented implementation makes full use of the height data available in Germany and may therefore not be directly transferable to regions with pronounced data scarcity, it can be adapted to other regions using OpenStreetMap data and additional global and local data sets. The overall goal of our study is not the comparison of various algorithms but proving the applicability of the proposed methodology and creating a useful dataset for Germany.

The article is organized as follows: Section 2 outlines the previous research in the field of building height assignment and presents the applied methodologies and results of existing studies. Section 3 describes the methodology used in this study, including the basic datasets, processing, machine learning, and validation steps. The results of the machine learning model training and of the nation-wide validation are presented in Section 4. These results and the benefits and limitations of the applied methodology are discussed in Section 5. Finally, Section 6 concludes the study and provides an outlook for further research.

2. Previous research

Milojevic-Dupont et al. [5] present a unified dataset of building footprints for Germany. However, the height data they provide is incomplete and a clear description of how building heights are derived is missing. Furthermore, the dataset of building footprints and building heights provided by Milojevic-Dupont et al. [5] only includes the heights reported in open governmental datasets or OpenStreetMap, resulting in a coverage of 66 % of buildings with height information in Germany. However, they do not conduct any further steps for assigning heights to buildings.

Different approaches for measuring and deriving building height information can be found in the existing literature (see Table 1). Where available, remote sensing data has been used to calculate building heights. Light Detection and Ranging (LiDAR) data is especially useful for this objective, because of the high accuracy of the generated spatial coordinates of surface points and it is available in many European areas [6]. Saraf et al. [7] assess the accuracy of 3-D building extraction from LiDAR data and conclude that the determined heights are within a tolerance of 1 m. Wu et al. [8] and Bonczak et al. [9] use LiDAR data to assign heights to buildings in the United Kingdom and New York City, United States, respectively. Meanwhile, Teo [10] constructs LoD1 buildings, i.e., 3-D buildings in a simplified block shape with flat roofs, from LiDAR data using a fully convolutional network and extract both building footprints and heights from the point clouds. Park and Guldmann [11] address the issue of identifying LiDAR points belonging to rooftops and propose a random forest machine learning algorithm for the classification. As an alternative to LiDAR data, satellite data available at a larger geographical scale can be used. The European Environment Agency [12] provides a 10 m resolution dataset of building height for major urban areas in the European Union, based on a combination of satellite and LiDAR data. In Germany, data is available for 97 cities. Frantz et al. [13] generate a raster dataset of building heights with a resolution of 10 m for all of Germany using satellite imagery from Sentinel-1A/B and Sentinel-2A/B time series. Although they limit their dataset to residential areas by using the European Settlement Map, they do not provide a building dataset assigned with the respective heights. Similarly, Li et al. [14] use Sentinel-1 data for estimating building heights in the United States within the 500 m grid, achieving a root mean square (RMSE) error of 1.5 m. Che et al. [15] use a combination of satellite data and other datasets to train a random forest model for predicting height at the building level and report an RMSE of 3.35 m. Cao et al. [16] use multi-spectral and multi-view images of three satellites to train a deep learning network for cities in China and achieve an RMSE error of 6.3 m. Instead of satellite data, Liu et al. [17] use single aerial images in combination with a convolutional–deconvolutional deep neural network architecture to generate a digital surface model and Pang et al. [18] use street view images with the Differentiable Volumetric Renderer deep learning architecture. The Microsoft GlobalMLBuildingFootprint dataset [19] includes heights derived by using a neural network trained on imagery and height measurements, however, it has some gaps, e.g., the city of Magdeburg is currently not covered.

In an effort to address the lack of elevation measurements in some areas, machine learning without the use of elevation data has been the subject of multiple studies. Biljecki et al. [20] provide a review on the methods applied and show that the three main categories for approximating building heights are the number of floors, local building height regulations, and the analysis of shadows in the imagery cast by buildings. They did not find metrics for describing the quality of height predictions using the number of floors, but report accuracies of 3–13 m for studies that estimate heights based on shadows in imagery. The authors also propose their own approach for determining building heights using a random forest algorithm. Support vector machines and

Table 1
Related studies in the field of building height estimation with the main data source, region and accuracy. The "open result dataset" column indicates whether the result dataset is openly available (only datasets explicitly referenced from the article are considered).

Study	Main data source/ approach	Region	Accuracy	Open result dataset
Saraf et al. [7]	LiDAR data	Malaysia (selected buildings)	>96 %	No
Wu et al. [8],	LiDAR data + OS MasterMap	United Kingdom	MAE: 0.3 floors	No
Teo [10]	LiDAR data + fully convolutional network	Taipei City, Taiwan	NA	No
Park and Guldmann [11]	LiDAR data + random forest model	City of Columbus and Franklin Country, Ohio, United States	MAE: 0.48 m; RMSE: 1.35 m	No
European Environment Agency [12]	LiDAR data, satellite imagery	Europe (selected cities and urban centers)	3 m	Yes, 10 m grid
Bonczak et al. [9]	LiDAR	New York City, United States	Volume median percentage error: 0.6 % (LoD1), -7.2 % (LoD2)	No
Frantz et al. [13]	Satellite imagery + regression model	Germany	Frequency-weighted RMSE: 2.92–3.55 m; RMSE: 3.83 – 8.14 m	Yes, 10 m grid
Li et al. [14]	Satellite imagery	United States (seven cities)	RMSE: 1.5 m at 500 m grid	No
Microsoft GlobalMLBuilding-Footprint dataset [19]	Satellite imagery + neural network	Mainly United States, Europe, Australia	NA	Yes, building-level with gaps
Cao et al. [16]	Satellite imagery + deep learning M^3Net	China (42 cities)	RMSE: 6.3 m	No
Pang et al. [18]	Street view imagery	Helsinki, Finland	Volume estimation: -9.198 % - -36.167 %	No
Wu et al. [23]	Satellite imagery, building features and proximity features + regression model	Shenzhen City, China	RMSE: 7.4 m	
Che et al. [15]	Satellite imagery, geometry features + random forest model	United States	RMSE: 3.35 m	No
Biljecki et al. [20]	2-D building data (cadaster) + random forest model	Rotterdam and Leeuwarden, Netherlands	MAE: 0.8 m	No
Bernard et al. [21]	2-D building data (OpenStreetMap) + random forest model	France (14 communes)	RMSE: 2.05 -2.2 m	No
Milojevic-Dupont et al. [22]	2-D building data (OpenStreetMap/ cadaster) + XGBoost	France, Italy, Netherlands, Germany (selected regions)	MAE: 1.47 m	No

multiple linear regressions were excluded due to worse performance. 2-D building data and additional attributes from the cadaster, e.g., population density, average household size, and average income were included as features. A model combining the number of floors, age, and net internal area performed best, with an accuracy of 0.8 m for the city of Rotterdam. Using a combination of the three geometric features of footprint area, normalized perimeter index and number of neighbors, a mean absolute error of 1.8 m was achieved. The model was also shown to be applicable to other cities in the Netherlands. Bernard et al. [21] use a random forest model for predicting building heights in France and report a RMSE of 2.05 - 2.2 m. Milojevic-Dupont et al. [22] use a gradient boosting algorithm to predict building heights based on a variety of urban form features such as building geometry, street networks, and blocks. They found that random forest and linear regression had a lower performance. They focus on cities in France, Italy, the Netherlands, and Germany and analyze the transferability of the trained model by performing cross-country validation. They achieved accuracies of between 0.91 and 1.65 m in their study areas in Italy, France and the Netherlands and show the potential for cross-country generalization. Wu et al. [23] compare the performance of extreme gradient boosting, random forest and artificial neural network regression models and find that extreme gradient boosting outperforms the other approaches and achieves an RMSE of 7.4 m. This is in line with Grinsztajn et al. [24], who demonstrate that XGBoost outperforms deep learning models, such as Resnet and the transformer models FT Transformer and SAINT on regression tasks for large datasets with numerical features. They conclude that tree-based models are state-of-the-art and mention the advantage of a higher training speed.

In summary, there have been successful attempts using both elevation measurements and machine learning to generate building footprint and height datasets. However, as shown in Table 1, the existing approaches have either not been applied to all of Germany or the resulting datasets are incomplete, not validated or not available at the individual-building level. Therefore, developing an approach for combining existing datasets and filling gaps using machine learning to generate a unified, easily usable, complete, and, importantly, validated nation-wide dataset containing building footprints and building heights is, to our knowledge, still missing. Describing this methodology and validating the generated data is therefore the goal of this study.

3. Methodology and data

The methodology for generating a unified building footprint and height dataset is based on three pillars: the raw building data as a primary input source, additional height datasets as secondary inputs, and lastly machine learning to fill remaining gaps in the dataset (see Fig. 1). The following sections describe the acquisition of the basic building data (Section 3.1) and the extraction of footprint and height information from it and from additional datasets (Section 3.2), and the methodology for training and using a machine learning model to impute missing data (Section 3.3). In a last step, the dataset is validated against official 3-D building data.

3.1. Basic building data

In a similar approach to that of Milojevic-Dupont et al. [5], basic building data for Germany was extracted where possible from open 3-D building governmental datasets downloaded from the respective open geodata portals of the federal states, and from OpenStreetMap for all other states. Open governmental data is available and was downloaded for the German federal states Bavaria (2022) [25], Berlin (2013) [26], Brandenburg (2022) [27], Hamburg (2021) [28], Hesse (2021) [29], Lower Saxony (2021) [30], North Rhine–Westphalia (2021) [31], Saxony (2022) [32], Saxony–Anhalt (2021) [33], and Thuringia (2021) [34]. The year in parenthesis specifies the reference year of the data. All datasets were downloaded between April 2022 and January 2023. The early reference year in the case of Berlin is due to its discontinuation of publishing building-level 3-D data. The included 3-D building datasets are in Level of Detail 2 (LoD2), i.e., buildings are represented by simplified blocks, including roof shapes, but with smaller scale geometries such as chimneys or windows neglected; the format follows the citygml standard [35]. Buildings are described through a set of

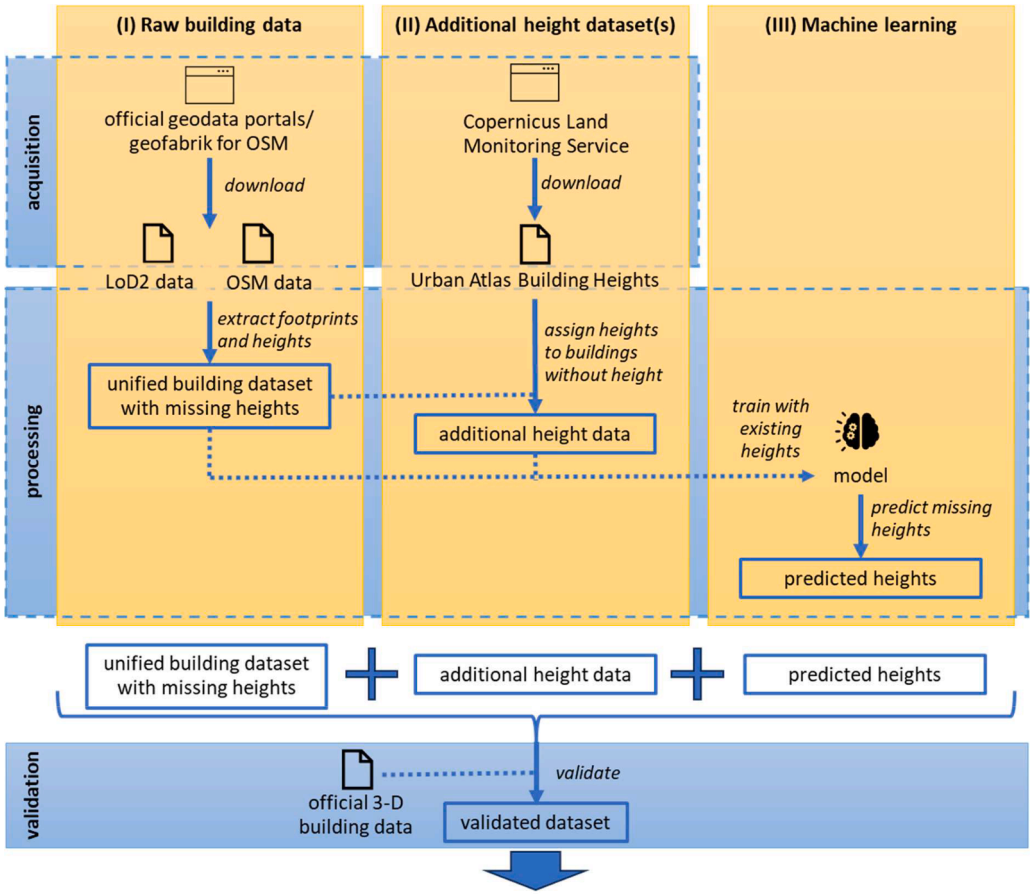


Fig. 1. Illustration of the three pillars of the data processing workflow (OSM=OpenStreetMap, LoD2=3-D building data in level of detail 2, i.e., including standardized roof shapes).

attributes, including, e.g., ID, function type and height, and through the coordinates of their surfaces. One building can be represented as either a single block or multiple sections that make up the building. For Baden–Wuerttemberg, Bremen, Mecklenburg–Western Pomerania, Rhineland–Palatinate, the Saarland, and Schleswig–Holstein, OpenStreetMap [36] data version 230315 in the osm.pbf format was downloaded from the geofabrik download server [37]. OpenStreetMap data for buildings combines the 2D geometric data with additional data in the form of key value pairs, providing information on the building type or its height. Data from both sources was then processed to extract building footprints and height data, as described in the following sections.

3.2. Building footprint and height extraction

The open governmental datasets are unpacked, and all ground surfaces, i.e., the boundary between the building and the ground, of each 3-D building, including its building parts, are extracted. They are merged into one 2-D building polygon using the *unary_union* function provided by shapely, which returns the union of the polygons [38]. Heights are extracted from the "*measuredHeight*" attribute, which provides the height difference between the ground and the ridge line of the building or building part. In the case of multiple building parts with this information, heights are combined by calculating the area-weighted height $h_{building}$ of all building part heights $h_{building\_part}$ according to Equation 1, with $A_{building\_part}$ being the footprint area of the building part. Building parts with a height below 2 m are not considered when calculating the weighted average height as they are assumed to be invalid.

$$h_{building} = \sum h_{building_{part}} * \frac{A_{building_{part}}}{\sum A_{building_{part}}} \quad (1)$$

Using the OpenStreetMap data, all buildings were filtered using osmosis [39] to reduce the file size and resulting memory consumption. Subsequently, buildings were extracted for further processing using the *get_buildings* function from Pyrosm [40]. The geometry polygons were used as building footprints without modifications. The building height, which in accordance with the description in the OpenStreetMapWiki should provide the maximum height of the building from the ground surface to the top of the roof [41], was provided for some buildings in the "*height*" tag. Building heights provided in the height tag were cleaned by removing invalid characters and values that do not conform to the floating-point number format. All buildings, both from open governmental data and from OpenStreetMap data with footprint areas below 1 m^2 or a height below 2 m, were deemed invalid and removed from the dataset.

As most of the buildings from OpenStreetMap do not contain height information, additional data sources were included for the states for which no open governmental 3-D building data is available. Due to a relatively high spatial resolution of 10 m and a coverage of the major cities in Germany, Urban Atlas Building Height [12] was deemed a suitable dataset for increasing the share of building heights in our data. This step is included in order to derive as many buildings' heights as possible from the elevation measurements, assuming that they are more accurate and reliable than machine learning algorithms without elevation data. All datasets for German cities from [12] were downloaded. Heights were then assigned to all buildings that could not be assigned a height in the previous step. This was undertaken by calculating the mean value of all raster cells of [12], touching the building footprint polygons extracted from OpenStreetMap using the rasterstats package [42]. Height values above 368 m, which is the height of the highest building in Germany (the TV tower in Berlin), were disregarded from all input

datasets.

3.3. Machine learning approach

Following the two previous steps, 16 % of buildings still do not contain building heights (cf. Section 4.1). To fill the missing values, a machine learning model was trained. As mentioned in Section 2, Biljecki et al. [20] find that random forests outperform support vector machines and multiple linear regression. Milojevic-Dupont et al. [22] and Wu et al. [23] show that gradient boosting performed as least as well as random forest or better. Wu et al. [23] also confirm the superiority of extreme gradient boosting over a deep learning model as found by Grinsztajn et al. [24].

In order to examine the validity of these findings for our dataset, we performed experiments on a subset of the full training dataset, containing a random 10 % of the data, with an 80:20 train-test split ratio. We trained the tree-based models *RandomForestRegressor* and *XGBoostRegressor*, the nearest neighbor model *KNeighborsRegressor*, the linear model *LinearRegression*, the support vector machine *LinearSVR*, and the neural network *MLPRegressor* from the sklearn package [43]. A randomized grid-search hyperparameter optimization was conducted, testing a maximum of 10 random combinations from the hyperparameter search spaces of the respective models presented in Table 2, including a 3-fold cross validation. We then compared the performance metrics and fit times of the models with the best hyperparameter combinations.

Due to its good performance in regression tasks, and a high computational efficiency for training on large datasets, found in literature and confirmed in our experiments, as will be described in more detail in Section 4.1, we chose extreme gradient boosting for our study. The implementation available as "*XGBRegressor*" in the sklearn package [43] was used.

A share of 80 % of all buildings assigned heights using the information in the basic building data or that provided by the additional dataset served as training data. The remaining 20 % were used in equal shares for the evaluation of early stopping in hyperparameter optimization and for testing of the final model (see Fig. 2). Data was split in a stratified fashion to ensure that buildings from all federal states (NUTS-1 regions) were included in all three partitions.

The feature set was based on the geometric features employed by Biljecki et al. [20], who use the footprint area, normalized perimeter index, and the number of neighboring buildings in a 30 m radius. We explicitly included the perimeter of the building footprint, the area-to-perimeter ratio, the number of touching buildings and the length of shared walls with neighboring buildings as well as additional ranges around a building to be considered for counting the neighboring buildings. The features used for training were as follows:

- Area of the building footprint A
- Perimeter of the building footprint P
- Perimeter of the equal area circle $P_{circ} = 2\sqrt{A\,\pi}$
- Area-to-perimeter ratio A/P
- Normalized perimeter index $NPI = P_{circ}/P$
- Number of touching buildings
- Length of shared walls with touching buildings
- Number of neighboring buildings in a [30,50, 100, 500] m radius

As Biljecki et al. [20] point out, the advantage of using these features is their easy availability and non-reliance on additional data sources, as they can be calculated directly from the dataset itself, while still being able to reach an accuracy of 1.8 m.

First, hyperparameter tuning using randomized grid search with five-fold cross-validation was performed to determine the optimal parameters for training. The tuned hyperparameters and potential values are presented in Table 3. 25 parameter combinations were randomly selected and evaluated. The objective of the learning task was set to regression with a squared loss (*reg:squarederror*). For evaluating the cross-validated model during hyperparameter optimization using the negative mean squared error, the scoring parameter was set to *neg_mean_squared_error*. The early stopping rounds parameter was set to 20, and evaluated by the RMSE.

Based on the optimal parameters, an *XGBRegressor* model was trained. This model was then evaluated using the test set and performance metrics for both the entire test set, as well as for subsets for the different federal states. The importance of each feature was determined using the 'gain' approach, which measures how much a split in the decision tree using the respective feature contributes to the performance of the model. The final model was then used to assign heights to all buildings still lacking height information. The minimum and maximum valid height of 2 m and 368 m, respectively, were enforced in a post-processing step by setting all values exceeding these limits to the boundary values.

Table 2
Hyperparameter search space for experiments on a subset of the data.

Model	Hyperparameter search space
RandomForestRegressor	n_estimators: [50, 100, 500], max_depth: [5, 10, 15], min_samples_split: [2, 3, 4], min_samples_leaf: [1, 5, 10], max_features: [0.8, 1.0]
XGBoostRegressor	n_estimators: [500, 1000], max_depth: [5, 10, 15], eta: [0.05, 0.1, 0.3], subsample: [0.5, 1], min_child_weight: [1,10,50,100], gamma: [0, 1, 5]
LinearSVR	C: [0.1, 1, 10], epsilon: [0, 0.5, 1]
LinearRegression	*No hyperparameter optimization*
KNeighborsRegressor	n_neighbors: [1, 3, 5, 8], leaf_size: [10, 20, 30, 40], p: [1, 2]
MLPRegressor	hidden_layer_sizes: [(50,), (100,), (200,)], activation: ["identity", "logistic", "tanh", "relu"], alpha: [0.001, 0.0001, 0.00001], learning_rate: ["constant", "invscaling", "adaptive"], learning_rate_init: [0.01, 0.001, 0.0001]

3.4. Validation

Building footprints and heights were validated against a building footprint polygon dataset and a full 3-D building dataset for Germany, respectively, which is not openly available. Instead, the dataset was provided by the Federal Agency for Cartography and Geodesy (BKG) [44] and made available only internally for research purposes. The building footprint polygon dataset dates from April 2020, whereas the 3-D building dataset is the first all-German version and dates from 2020 (month not specified). According to the NUTS classification, Germany can be divided into multiple statistical regions at different levels. For this study, the relevant levels range from the national level 0 down to the 401 districts of level 3. Validations were carried out at all three levels, from the national (NUTS-0) to the district (NUTS-3) levels.

In a first step, the completeness of the generated dataset regarding the building footprints was evaluated. As a direct building-by-building comparison between object IDs was not possible due to the different data sources, alternative measures were required. First, the number of buildings from the validation dataset were compared to that of the generated dataset on a per-region basis. Second, the total footprint areas per region were compared.

In a second step, the building heights were validated against the height information provided in the validation 3-D building dataset. As this dataset was in the citygml format and so it was not possible to easily

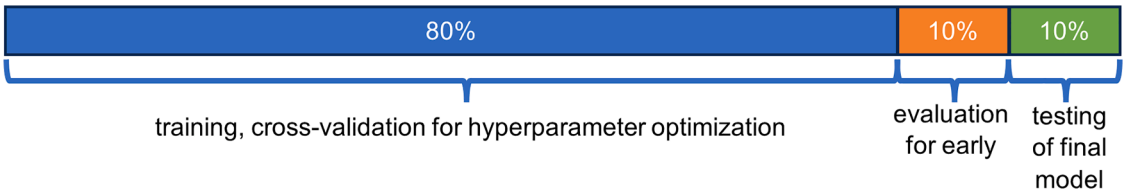


Fig. 2. Splitting of data for the machine learning process.

Table 3
Possible hyperparameter values used in the randomized grid search for hyperparameter optimization.

Parameter	Values	Description [43]
n_estimators	500, 1000	Number of trees
min_child_weight	1, 10, 50, 100	Minimum sample size in the node
gamma	0, 1, 5	Loss reduction required for further splitting of the node
subsample	0.5, 1	Ratio of training data sub-sampled for each boosting iteration
max_depth	5, 10, 15	Maximum depth of tree
eta	0.05, 0.10,3	Learning rate

read height data for all buildings, the methodology used to extract building footprints and heights follows that described in Section 3.1 for the open governmental 3-D data. First, the generated height dataset was validated by comparing the distribution of heights in the generated and validation datasets. Then, the heights of individual buildings in the generated building dataset were compared to the height of the building with the most similar footprint area in a 5 m radius in the validation dataset based on the buildings' centroids. The mean absolute error (MAE) was calculated at various aggregation levels according to Equation 2, with h_g being the height of the building in the generated dataset, h_v the height of the nearest neighboring building in the validation dataset, and n the number of buildings.

$$MAE = \frac{\sum_{i=1}^{n} |h_g - h_v|}{n} \tag{2}$$

Similarly, the mean relative error (MRE) was calculated according to Equation 3:

$$MRE = \frac{\sum_{i=1}^{n} \frac{|h_g - h_v|}{h_v}}{n} \tag{3}$$

Then, the distribution of individual building errors depending on the height and footprint area of the buildings in the validation dataset were analyzed for selected NUTS-1 regions. Furthermore, the distribution of average errors at the NUTS-3 level depending on the degree of urbanization according to the Global Human Settlement Layer's degree of urbanization classification level 2 [45] was assessed. This classification differentiates between seven urbanization levels, from mostly uninhabited area to cities. Finally, the sum over all building volumes in each NUTS-1, NUTS-3 and LAU region in the generated dataset was compared to that in the validation dataset. A building's volume was approximated by multiplying its footprint area with its height.

4. Results

The machine learning results, including selected hyperparameters, performance metrics and feature importance of the trained model, are presented in Section 4.1. Furthermore, the share each of the data sources contributes to the completeness of the final height dataset is given. The results show that the accuracy of the machine learning model for all of Germany is 1.78 m and varies between the federal states. The share of buildings for which the model was used to predict the highest in Saarland with 92 %. In Section 4.2, the validation results of the generated footprint and height dataset are shown. They show that the number of buildings and the total footprint area have an accuracy of -9.69 % and –5.16 % over all of Germany, respectively. The height accuracy is 0.59 m over all of Germany. All accuracies vary depending on the federal state.

4.1. Data extraction and imputation

The building footprint dataset contains a total of 50,815,696 buildings, combining open governmental 3-D data for the German federal states where available and OpenStreetMap building data for the others. The share of buildings that can be assigned a height directly from the raw data range between 0.04 % and 100 % for the different federal states, and 81.4 % for all of Germany. Adding the Urban Atlas Building Height data [12] increased the share of buildings with a height by between 0 and 86 percent points (see Fig. 5).

In total, 84% of German buildings can be assigned a height attribute. These buildings can then be used for training and evaluating the machine learning model. The available building data are unevenly distributed across the federal states, with shares between 8 % and 100 % of buildings in each federal state that can be used for machine learning (see Fig. 5).

Experiments on a subset containing 10 % of the total training dataset yielded the results presented in Table 4. The errors are lowest across all metrics for the *XGBoostRegressor*, followed closely by the *RandomForestRegressor*. The *MLPRegressor* performs similarly, with an MAE that is 0.07 m above that of the *XGBoostRegressor*. The *KNeighborsRegressor, LinearRegression* and *LinearSVR* perform worse with an MAE almost twice as high for the worst performing *LinearSVR* than for the *XGBoostRegressor*. Fit times are lowest for *LinearRegression* and the *KNeighborsRegressor*. Training the *XGBoostRegressor* is slower than the *KNeighborsRegressor* by a factor of 8.8, but it is faster than the *MLPRegressor* by a factor of 2.5 and faster than the *RandomForestRegressor* by a factor of 4. This shows that the *XGBRegressor* does not only have the highest accuracy, but it also beats models with a similar performance in terms of fit times. Therefore, this model was deemed the best choice for training on the full dataset.

The hyperparameter optimization using the *XGBoostRegressor* on the full training dataset yielded the parameters presented in Table 5.

Validating the model trained with the hyperparameters on the test dataset resulted in a mean absolute error (MAE) of 1.78 m, an RMSE of 2.8 m, a mean absolute percentage error (MAPE) of 33.29 %, and a mean squared error (MSE) of 7.38 m^2. However, the validation results on test data varied significantly between the federal states (see Fig. 3). Values above an MAE of 2.5 m, i.e., a common ceiling height of a floor, were noted for Berlin, Bremen, Hamburg, Baden-Württemberg, Schleswig-Holstein, Mecklenburg-Western Pomerania, Brandenburg and Saarland.

The feature importance of the trained model based on the 'gain' approach is depicted in Fig. 4. According to this metric, the area-to-perimeter ratio is of particularly high importance, followed by the normalized perimeter index. When considering the number of neighboring buildings as an indication for the neighborhood characteristics, a range of 50 m appears to have the highest predictive power.

The share of each source contributing data for assigning the height attribute in all German federal states is shown in Fig. 5. In most federal states where open governmental 3-D building data is available, the share

Table 4
Results of the machine learning experiments for a subset of the total training dataset. (MAE=mean absolute error, MAPE=mean absolute percentage error, RMSE=root mean squared error, MSE=mean squared error).

Model	Hyperparameters	MAE [m]	MAPE [%]	RMSE [m]	MSE [m²]	Fit time [s]
RandomForest-Regressor	n_estimators: 100, max_depth: 15, min_samples_split: 3, min_samples_leaf: 5, max_features: 0.8	1.81	33.84	2.82	7.96	1010.07
XGBoostRegressor	n_estimators: 1000, max_depth: 10, eta: 0.05, subsample: 1, min_child_weight: 10, gamma: 5	1.80	33.64	2.82	7.95	254.91
LinearSVR	C: 1, epsilon: 0.5	3.54	88.45	4.87	23.73	1196.73
LinearRegression		2.37	47.06	3.59	12.89	2.07
KNeighbors-Regressor	n_neighbors: 8, leaf_size: 40, p: 1	1.98	37.11	3.11	9.68	28.79
MLPRegressor	hidden_layer_sizes: (100,), activation: “relu”, alpha: 0.00001, learning_rate: "adaptive", learning_rate_init: 0.0001	1.87	34.52	2.97	8.84	632.40

Table 5
Best parameter values according to the hyperparameter optimization. For a description of parameters, see Table 3.

Parameter	Values
n_estimators	1000
min_child_weight	100
gamma	1
subsample	1
max_depth	15
eta	0.1

of height data taken from this source equals 100 %. Berlin constitutes an exception; the raw data does not contain a “*measuredHeight*” tag in some cases. Where OpenStreetMap provides the data basis, the source distribution is more diverse. The share of buildings that can be assigned a height attribute from OpenStreetMap is relatively low. It generally remains well below 1 % and only reaches a maximum of 2.4 % in Mecklenburg–Western Pomerania. In some federal states, Urban Atlas building heights contribute significantly to the number of buildings assigned a given height. This is especially the case for Bremen which features a percentage of about 86 %. In other federal states where Urban Atlas building heights significantly contribute, the share lies in the range of 7–17 %. Buildings that could not be assigned a height using the raw data sources were assigned one via the machine learning model. Therefore, the federal states for which OpenStreetMap data is used as a basis except for Bremen rely most heavily on machine learning, i.e., Baden–Württemberg, Mecklenburg–Western Pomerania, Rhineland-Palatinate, Saarland, and Schleswig–Holstein. With a share of 92 % of building heights estimated with the machine learning model, Saarland is at the top.

4.2. Validation

Comparing the number of buildings in the generated and validation dataset yields a deviation of -9.69 % for all of Germany and varies between 0.27 % for Brandenburg (BB) and -46.9 % for Schleswig–Holstein (SH), as depicted in Fig. 6. Deviations of 25 % or more can be observed in most federal states with OpenStreetMap as a data basis, whereas the deviation remains below 10 % for all federal states in which open governmental 3-D building data could be used, with the exception of Berlin, with a relative error of 18.6 %.

As is shown in Fig. 7, the relative error of the footprint area for Germany is -5.16 %, and therefore lower than the relative error for the number of buildings. At the NUTS-1 level, the error lies within the much smaller range of between -0.52 % for Mecklenburg-Western Pomerania and -10.2 % for Hamburg. No dependency on the source of the raw data was apparent, i.e., whether official 3-D data or OpenStreetMap data was used.

The results of the building heights validation are depicted in Fig. 8, Fig. 9, and Fig. 10. Fig. 8 shows the distribution of building heights in the generated dataset compared to the building heights in the validation dataset for all of Germany. Overall, the distributions display very high agreement. However, in the validation dataset, some buildings taller than the maximum expected height of 368 m are present. These heights were deemed invalid in the methodology described in this study and therefore do not appear in the generated dataset. As is shown in Fig. 9, the MAE for all of Germany is 0.59 m, which is the same as the median absolute error. For the various federal states, it ranges between 0.11 m for Hesse and 3.27 m (median absolute error of 2.96 m) for Bremen. Errors are not homogeneous across regions, and even within the NUTS-1 states, there is significant variance. Amongst the NUTS-3 regions, Stuttgart and Heidelberg, both located in Baden–Württemberg, exhibited the lowest accuracies, with an MAE of above 5 m. Five other NUTS-3 regions featured an MAE above 4 m; 13 others had an MAE above 3 m. On the other side, 289 NUTS-3 regions had accuracies in the sub-meter

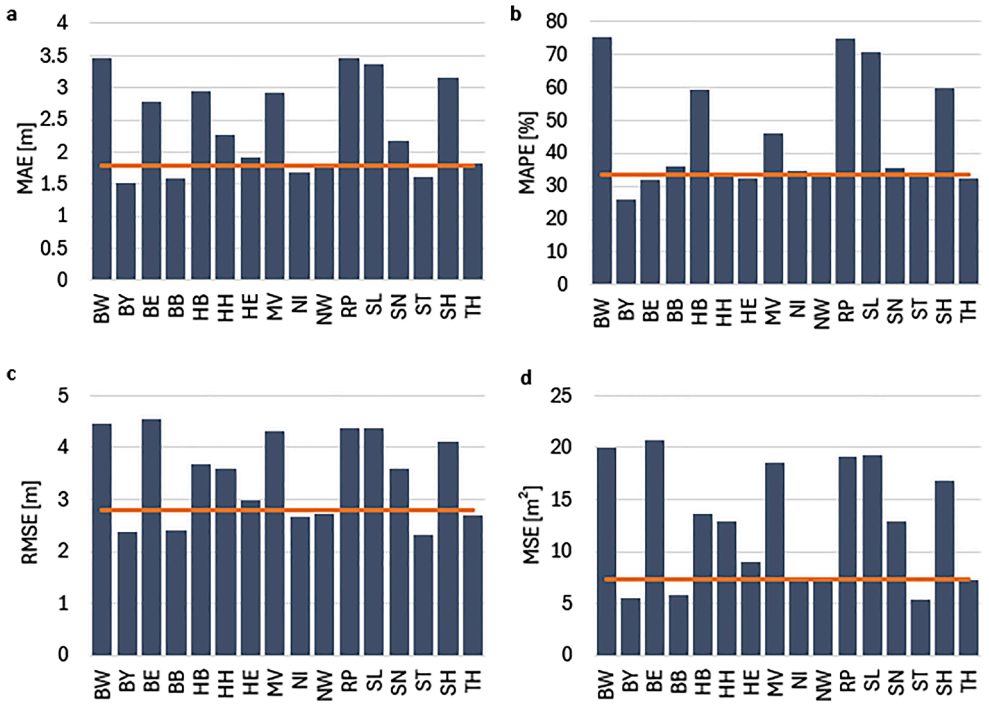


Fig. 3. (a) Mean absolute error (MAE), (b) mean absolute percentage error (MAPE), (c) root mean squared error (RMSE), and (d) mean squared error (MSE) in building heights of the XGBoost machine learning model for the German federal states (BW=Baden–Württemberg, BY=Bavaria, BE=Berlin, BB=Brandenburg, HB=Bremen, HH=Hamburg, HE=Hesse, MV=Mecklenburg–Western Pomerania, NI=Lower Saxony, NW=North Rhine–Westphalia, RP=Rhineland–Palatinate, SL=Saarland, SN=Saxony, ST=Saxony–Anhalt, SH=Schleswig–Holstein, TH=Thuringia). Orange line shows mean over all federal states.

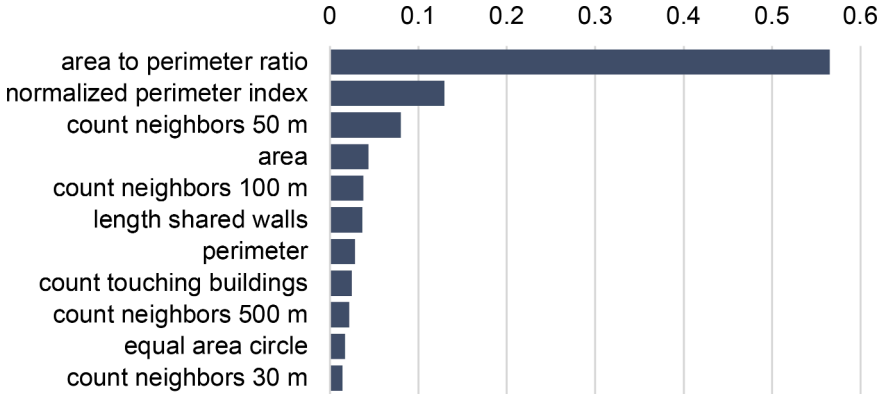


Fig. 4. Feature importance ('gain') of the features used for training the XGBoost machine learning model.

range.

The MRE of all four aggregation levels is shown in Fig. 10. For Germany as a whole, the MRE is 9.12 %. The values vary between federal states and range between 1.67 % for Hesse and 41.12 % for Schleswig–Holstein. MREs at NUTS-3 level are below 10 % for 288 regions. Only three regions, the city of Stuttgart, Freiburg, and Ulm, located in Baden-Württemberg, exhibit an MRE above 50 %.

Fig. 13 shows the distribution of the MAE at the NUTS-3 level grouped by degree of urbanization. The median is almost the same for all urbanization levels, apart from the "suburban or peri-urban area" category, which contains only two regions. 75 % of the NUTS-3 regions across all urbanization categories have an error below 2 m. From dispersed rural area to semi-dense town, the error for all regions remains below 2.5 m. Higher errors occur only in dense towns and cities, with errors up to 3.5 m and 5.39 m, respectively.

Fig. 14 shows the relative error of the total building volume aggregated by NUTS-3, NUTS-1 and LAU level. At the NUTS-1 level, the relative error in 12 of the 16 federal states is below 10 %. The relative error is above 10 % in Bremen (-28.49 %), Mecklenburg–Western Pomerania (18.45 %), Berlin (13.43 %) and Bavaria (10.39 %). At the NUTS-3 and LAU level, 71.32 % and 63.01 % of the regions, respectively, have an accuracy above 90 %. In 3.74 % of the NUTS-3 regions the accuracy is below 75 % and in 2.09 % of LAU regions, the accuracy is below 50 %.

5. Discussion

This study shows the feasibility of the presented methodology and quantifies the accuracy of the results on different aggregation levels. Using this approach, it is possible to generate a complete building dataset of acceptable quality, combining different data sources. The following section discusses the benefits and limitations of the chosen approach, compares it with existing studies and presents ideas for further improvement and suggestions for future investigations.

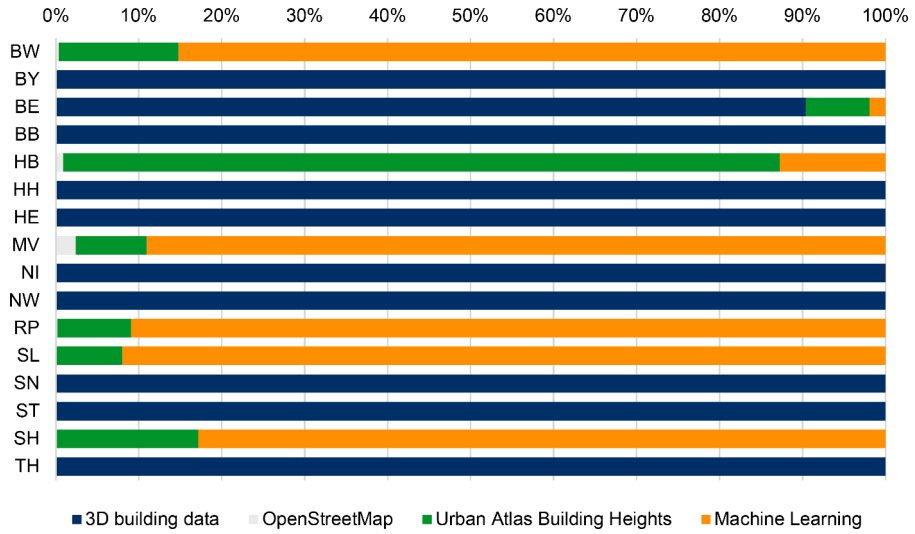


Fig. 5. Share of buildings in each German federal state, derived from the various source datasets and approaches for assigning a height value (BW=Baden–Württemberg, BY=Bavaria, BE=Berlin, BB=Brandenburg, HB=Bremen, HH=Hamburg, HE=Hesse, MV= Mecklenburg–Western Pomerania, NI=Lower Saxony, NW=North Rhine–Westphalia, RP=Rhineland–Palatinate, SL=Saarland, SN=Saxony, ST=Saxony–Anhalt, SH=Schleswig–Holstein, TH=Thuringia).

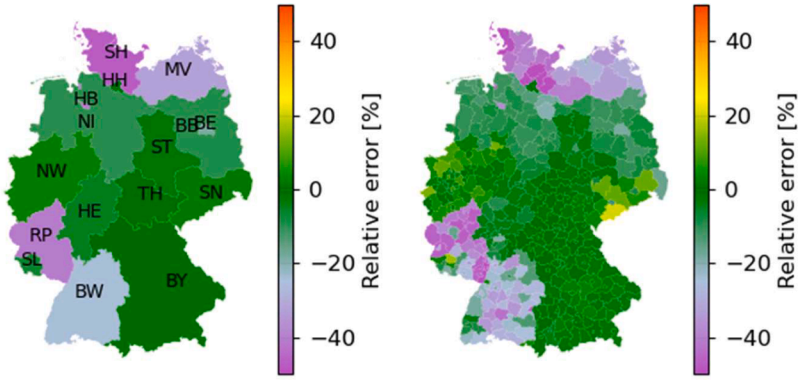


Fig. 6. Relative error of the number of buildings in the generated dataset compared to the validation data at the NUTS-1 and NUTS-3 levels. For negative values, the number of buildings was lower in the generated dataset than in the validation one.

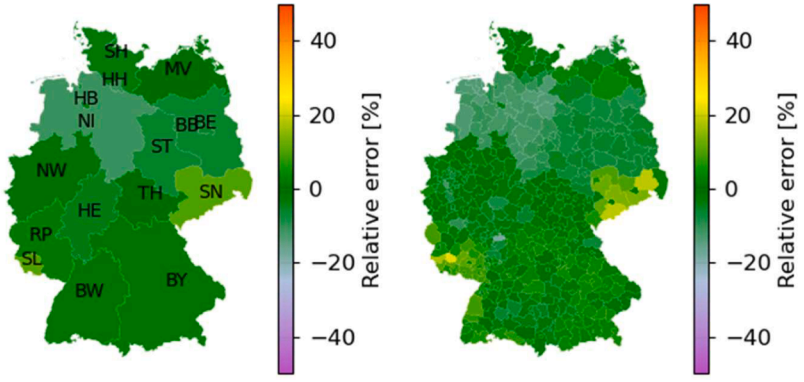


Fig. 7. Relative error of the footprint area of buildings in the generated dataset compared to the validation data at the NUTS-1 and NUTS-3 levels. For negative values, the footprint area of buildings was lower in the generated dataset than in the validation dataset.

The validation of building footprints shows that while there is a good match for some federal states, in others there are large discrepancies in the number of buildings in the generated and validation datasets for federal states relying on OpenStreetMap data. A visual examination of the validation data revealed that this is caused by the validation dataset being much more detailed and including smaller buildings and parts thereof. As the total number of polygons is counted and there is no trivial way of combining building parts with buildings, this leads to a high number of buildings in the validation data. For this reason, validation based on footprint area is deemed more meaningful. Considering the footprint area, a very high degree of similarity between the generated data and validation data can be observed for all federal states,

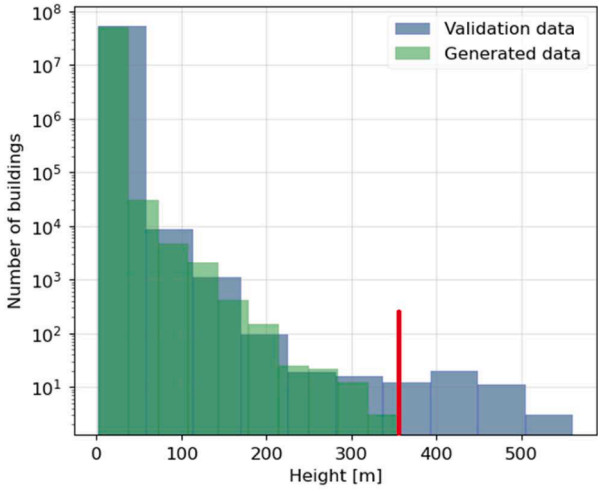


Fig. 8. Distribution of building heights in generated and validation datasets for all of Germany (y-axis in logarithmic scale). The red line indicates the height of the highest building in Germany; heights above it are inapplicable.

independent of the source of the building data. This permits the conclusion that using the approach of combining official 3-D building data with volunteered geographical information from OpenStreetMap data is a viable approach for generating a complete building footprint dataset for Germany.

The average error when comparing the generated dataset to the validation one for Germany is 0.59 m. This is well below common ceiling heights and should therefore not have a significant effect when using building data on an aggregated level. However, the observed errors strongly depend on the region and source of the height data. Although errors are negligible in regions based on open governmental 3-D data, errors are higher, as expected, in the other regions. However, even within the NUTS-1 regions, errors are not homogeneous, and some NUTS-3 regions were found to have significantly higher accuracies than others. Amongst the NUTS-1 regions, Bremen proves to be most problematic due to a high reliance on the machine learning model combined with a poorer performance of the model for this state. At the NUTS-3 level, the model shows low accuracies for several cities in Baden–Württemberg. Analyzing accuracies depending on the type of region showed that higher errors occur in cities and dense towns. This is most likely due to the more complex and heterogeneous structure of cities compared to villages and dispersed rural areas. This finding is in line with the higher error for, e.g. Paris, in the study by Bernard et al.

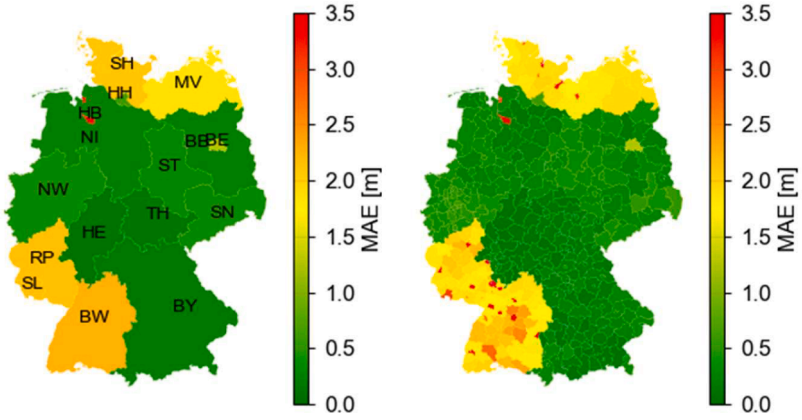


Fig. 9. Mean absolute error (MAE) of building heights in the generated dataset compared to the validation data at the NUTS-1 and NUTS-3 levels.

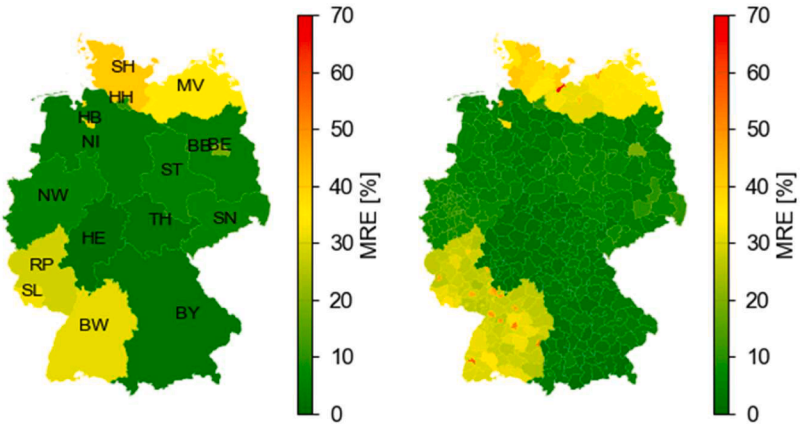


Fig. 10. Mean relative error (MRE) of building heights in the generated dataset compared to the validation data at the NUTS-1 and NUTS-3 levels. Figs. 11 and 12 show the distribution of absolute building height errors in the generated data over real height and footprint area of the buildings for those federal states that have the lowest height accuracies (MRE >30 %), as described above. The error does not exhibit any clear dependence on the height of buildings. A tendency for errors to be higher for buildings with heights above 15 m can be observed across most regions though; it is most pronounced in Baden–Württemberg, Rhineland–Palatinate, and Saarland. However, there are only very few buildings that fall into that height category. The median error for buildings in the range between 3 and 10 m, which is the most frequent height range for all regions depicted in Fig. 11, is below 2 m for all regions apart from Bremen. The majority of errors is even lower, whereas there is a tail of higher errors. As Fig. 12 shows, building height errors appear to be mostly independent of the footprint area of buildings, with only a slight increase of errors for buildings with footprints of more than 200 m^2.

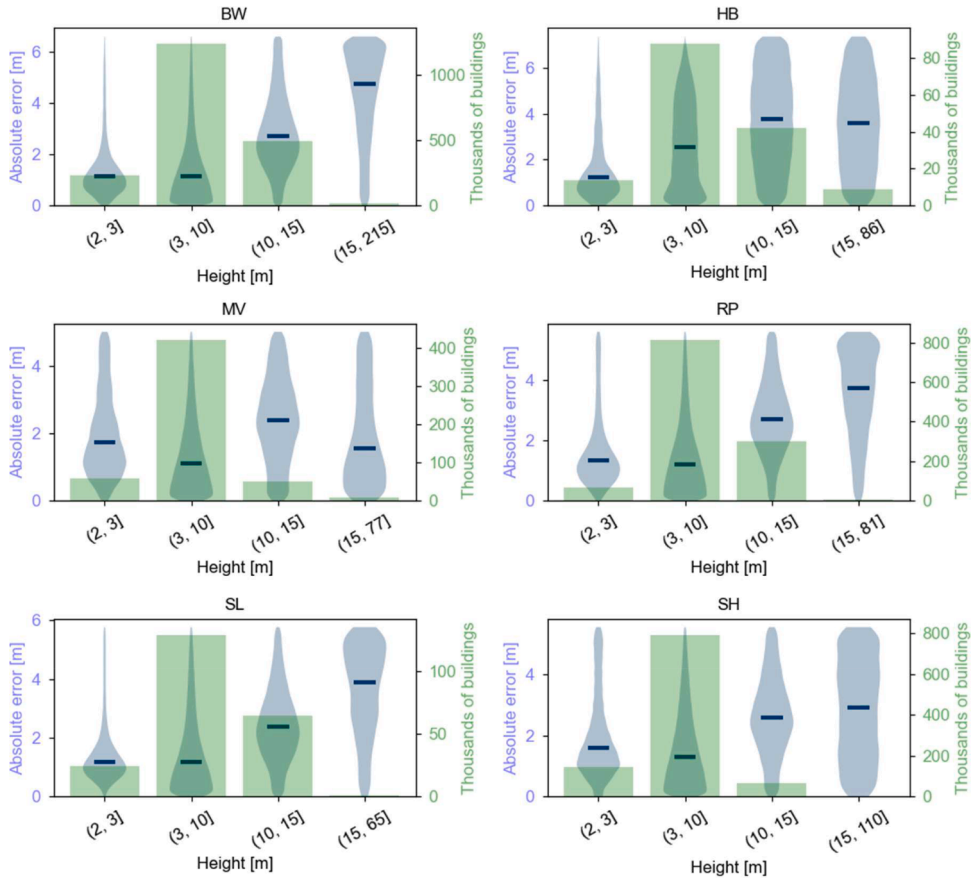


Fig. 11. Distribution of the absolute building height errors (upper 5 % removed for display reasons) and number of buildings for height ranges in selected NUTS-1 regions in Germany (BW=Baden–Württemberg, HB=Bremen, MV=Mecklenburg–Western Pomerania, RP=Rhineland–Palatinate, SL=Saarland, SH=Schleswig–Holstein).

[21]. Using the generated dataset for these regions could therefore lead to errors in further analyses. Looking at the relative error instead of absolute errors reveals a similar picture. The main difference is that less densely populated states with overall lower building heights show a higher relative error, indicating that the absolute deviance between heights in the generated and validation dataset is similar, independent of the height of the building.

It must be pointed out that the errors in the building height validation are not independent of the building footprint validation. As discussed above, it was found that the similarity between generated and validation dataset is considerably higher when comparing footprint areas than when comparing building numbers, indicating that the 3-D building data is more detailed than the OpenStreetMap data. This could lead to the effect that the height of a building in the generated dataset is compared to that of a building in the validation dataset, to which it is not actually comparable, even though the buildings are in close proximity. For example, the height of a large building complex might be compared to that of a garage, despite the limitation of a 5 m radius and selecting the building with the most similar footprint area for comparison. As an additional validation that eliminates this error source, we compared the total building volume per region at the NUTS-1, NUTS-3 and LAU level and found a high agreement between the generated and the validation dataset for most regions. Due to the averaging effect, errors are lowest at the highest aggregation level, i.e., at the federal state level and only a small percentage of NUTS-3 exhibits accuracies below 75 %, making the dataset particularly suitable for large-scale analyses at the aggregated level. However, the higher errors in some of the NUTS-3 and, particularly, LAU regions need to be considered and corrected for when using the data.

Whether the presence of additional building height data from the Urban Atlas building heights increases the overall accuracy of the generated data compared to using machine learning models only is not evident and requires further investigation in subsequent studies. The fact that the Urban Atlas dates to 2012 could significantly increase average errors by including buildings that no longer exist in the generated and validation dataset and vice versa. Other methods of assigning heights to buildings from the Urban Atlas raster data could be tested in the future, e.g., only considering the pixels overlapping the centroid of the building polygon or taking the maximum instead of the average value of the touching pixels.

Additionally, it must be kept in mind that validating the generated data against the 3-D building dataset assumes that this dataset represents reality. However, as mentioned previously, even the validation dataset contains a few unrealistic building heights. It is expected that for the regions where open governmental 3-D data is available, the accuracy compared to this dataset is very high, as it is based on the same methodology and data basis. Minor discrepancies for those states in the validation can be attributed to the source data for generating the dataset itself being, in most cases, more recent than the validation data. In the case of Berlin this is reversed, and the validation data is more recent by seven years, explaining in part the lower accuracy in the city state. Furthermore, the pre-processing steps taken to extract building footprints and height information from the raw citygml files was conducted for both generating and validating the dataset. The validation dataset was therefore modified to be able to carry out the validation, which presents a potential source of uncertainty.

The distribution of errors shows that for the most frequently occurring height range of 3–10 m, the majority of errors as well as the mean

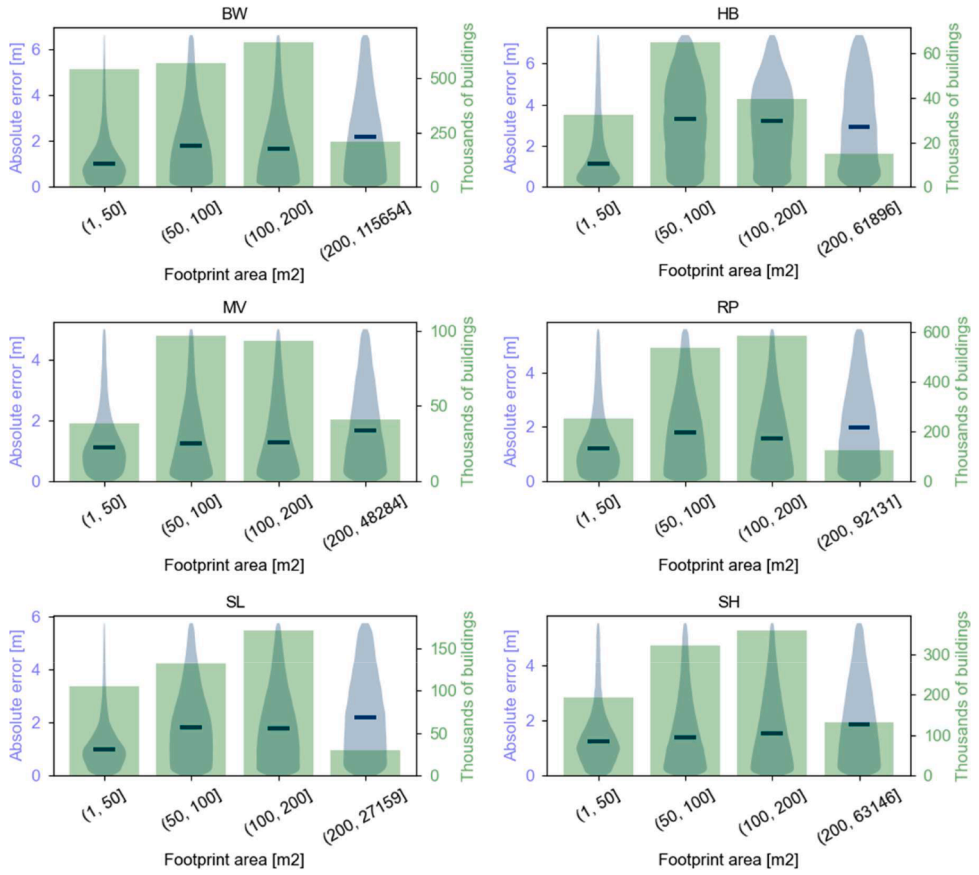


Fig. 12. Distribution of the absolute building height errors (upper 5 % removed for display reasons) and number of buildings for footprint area ranges in selected NUTS-1 regions in Germany (BW=Baden–Württemberg, HB=Bremen, MV=Mecklenburg–Western Pomerania, RP=Rhineland–Palatinate, SL=Saarland, SH=Schleswig–Holstein).

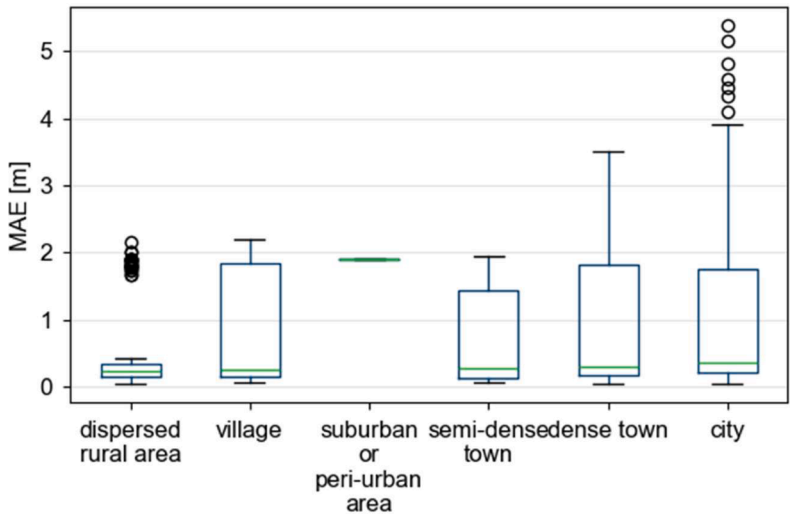


Fig. 13. Distribution of mean absolute error (MAE) in NUTS-3 regions grouped by degree of urbanization.

error, is in a range below the average ceiling heights. However, some errors are significantly higher across all height and footprint ranges. These buildings will require further investigation in future research. Including datasets containing height data specifically for high-rise buildings would be an interesting addition. As the outliers with high errors amongst a few buildings are likely to have a significant impact on the average error in federal states, reducing the error of these buildings should also decrease the overall error.

Furthermore, the accuracy of the machine learning model (MAE=1.78 m), which slightly outperforms the model based solely on geometric properties by Biljecki et al. [20] (MAE=1.8 m), but as expected performs worse than their more complex models with more features (MAE=0.8 m), could potentially be increased by including more features. Similarly, Milojevic-Dupont et al. [22] use more features and, despite including only small samples from the respective areas, achieve slightly better MAEs when applying their model to their study area

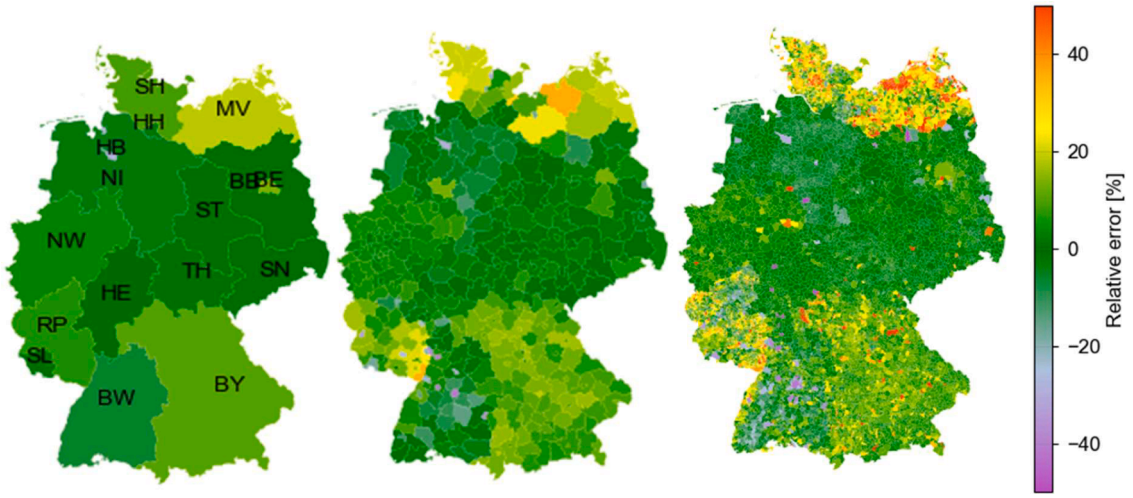


Fig. 14. Relative error of building volume in the generated dataset compared to the validation data at the NUTS-1, NUTS-3 and LAU level.

(MAE=1.47 m) and to Berlin (MAE=2.06 m) and Brandenburg (MAE=1.47 m). As they do not validate their model for other German regions, a more detailed comparison was not possible. The model by Bernard et al. [21] performs slightly better (RMSE=2.02–2.2 m) than our model (RMSE=2.8 m), which is not surprising considering that their study is limited to a small number of French communes. Compared to the nation-wide random forest model for the United States by Che et al. [15] (RMSE=3.35 m) and the deep learning model for Chinese cities by Cao et al. [16] (RMSE=6.3 m), our model shows a higher accuracy. Adding a feature that contains information about the location, e.g., a NUTS code or coordinates, of the building would be especially interesting, as it could allow the model to take regional differences into account and might help balance the model's performance in the various federal states. Furthermore, features that provide information on the environmental and socio-economic context, such as the proximity to public infrastructure or the heritage status of a building could improve the model's performance. However, the model should not be made too complex in order to keep computation times at a manageable level and avoid overly strong dependencies on the availability of regional datasets, which would limit transferability to other regions.

Our experiments on a subset of the full training data confirmed the findings in literature that *XGBoostRegressor* is the best suitable model for the regression task at hand. Not only did it show the highest performance in terms of accuracy metrics, but it also had lower fit times than other similarly performing models. In order to further increase the accuracy of the XGBRegressor model, a more detailed hyperparameter optimization could be carried out in future studies.

The advantage of this progressive approach of integrating different data sources lies in the ability to incorporate the best data available while guaranteeing a complete dataset without missing data. It provides a minimal workflow that is transferable to other regions while allowing for extension and adaptation to local data availability. For Germany, a large height dataset is available, which was valuable for training and, especially, validating the selected approach. This is not necessarily the case in other regions. Within this study, we attempted to train the best possible model for Germany using as much data as possible. This does not allow us to make statements on the requirements for data availability and potentially limits the generalizability and transferability of the machine learning to regions with severe data scarcity. However, our proposed approach makes it possible, in the most basic case, to transfer the approach to other regions, where OpenStreetMap can serve as a minimal data basis and the machine learning model can be used to impute missing heights. As pointed out by Milojevic-Dupont et al. [22], adding only a small sample from the target region can already significantly improve model performance. However, while the completeness of building footprints in OpenStreetMap is 71 % in Europe, according to Herfort et al. [46], only 2.8 % of all buildings in OpenStreetMap worldwide contain a height key [47]. Moreover, whereas 13.41 % of buildings in the United States have a height key [48], the coverage in European countries is often much lower (Germany: 0.42 % [49], France: 0.28 % [50], Hungary: 0.23 % [51]). Therefore, relying solely on OpenStreetMap data may not be sufficient. Ideally, 3D building datasets or other height data sources are available for at least some areas within the region of interest. Considering the increasing coverage of building heights in the Microsoft GlobalMLBuildingFootprint dataset [19], it is likely that many regions can use this dataset as one height data input in the future. However, formats and semantics may vary between sources, requiring tailored data processing steps. As Milojevic-Dupont et al. [22] show, applying the machine learning model to other regions, ideally with additional local data, leads to acceptable results. However, while they consider a scenario with 2 % of additional height data available, even this exceeds what is available in OpenStreetMap for many regions. Therefore, whereas model improvement is an important aspect, the acquisition of high-quality input data on an individual building level is crucial to fully leverage the presented approach and enhance the dataset's accuracy. Whereas the focus of this study lies in using openly available datasets and features that can easily be derived from building geometries to increase transferability, it might be worth analyzing the effect of including additional datasets. It would, for example, be interesting to evaluate the availability of LiDAR data for buildings that have been assigned a height using machine learning in our approach and to test whether its inclusion increases the accuracies of the height assignment. In future studies, though making the approach more region-specific, including other datasets could increase accuracy in regions with higher errors both directly and through increases in the accuracy of the machine learning model by providing more spatially-balanced training data.

6. Conclusions and outlook

The presented methodology, including footprint and height data extraction from basic 3-D and OpenStreetMap data, enriching it with additional height datasets and imputing missing height data using machine learning, is suitable for creating a unified and complete dataset of building footprints and heights for all of Germany. Apart from a few regions that require further attention in future studies due to lower accuracies of building heights, the accuracy of both building footprints and height data is high. We find a high correspondence between the generated and reference datasets for the building number and per-building height validation in those states where open governmental data is

available, the difference is larger in areas where OpenStreetMap is the basis for building data. For the footprint area and building volume validation, the accuracy is less dependent on the federal state's main data source. The nation-wide building number and footprint area accuracy is 90.31 % and 94.84 %, respectively. The XGBoost machine learning model for height prediction has an overall accuracy of 1.78 m, the nation-wide height accuracy based on a per-building validation is 0.59 m and varies between 0.11 m and 3.27 m depending on the federal state. It remains under the common ceiling height of 2.5 m for all states apart from Bremen. The total building volume accuracy at the federal state level lies between 71.5 % and 99.91 % and is above 90 % for 12 of the 16 federal states. At the NUTS-3 and LAU level, 71.32 % and 63.01 % of the regions have an accuracy above 90 %, respectively. Based on the results of the validation it can be stated that the dataset is suitable for individual-building and aggregated analyses.

In future studies, more features could be included in the machine learning model. Additionally, if possible, more datasets could be incorporated, especially for regions with higher errors. Furthermore, the transferability of our model to regions outside of Germany, especially those with marked data scarcity, is an interesting question to evaluate.

To summarize, the methodology described in this study is useful for creating a nation-wide, high-quality building height dataset. The dataset itself is of high value to the scientific community and can easily be used for further analyses, e.g., in energy system modeling.

CRediT authorship contribution statement

Kristina Dabrock: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Noah Pflugradt:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Jann Michael Weinand:** Writing – review & editing, Supervision. **Detlef Stolten:** Supervision, Resources, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The result dataset is available to download from https://doi.org/10.5281/zenodo.11845992.

Acknowledgements

This work was supported by the Helmholtz Association under the program "Energy System Design".

References

[1] European Commission. In focus: Energy efficiency in buildings. European Commission - European Commission; 2022. Accessed: Apr. 04,[Online]Available, https://ec.europa.eu/info/news/focus-energy-efficiency-buildings-2020-lut-17_en.
[2] Loga T, Stein B, Diefenbach N. Tabula building typologies in 20 European countries-making energy-related features of residential building stocks comparable. Energy Build 2016;132:4–12. https://doi.org/10.1016/j.enbuild.2016.06.094. Nov.
[3] Bandam A, Busari E, Syranidou C, Linssen J, Stolten D. Classification of building types in Germany: a data-driven modeling approach. Data (Basel) 2022;7(4). https://doi.org/10.3390/data7040045. Artno. 4Apr.
[4] "EU Building Stock Observatory - Database - Energy Consumption." Accessed: Nov. 21, 2023. [Online]. Available: https://building-stock-observatory.energy.ec.europa.eu/database/.
[5] Milojevic-Dupont N, et al. Eubucco v0.1: European building stock characteristics in a common and open database for 200+ million individual buildings. Sci Data 2023;10(1). https://doi.org/10.1038/s41597-023-02040-2. Art. no. 1, Mar.
[6] European Commission. Joint Research Centre.. Non-commercial light detection and ranging (LiDAR) data in Europe. LU: Publications Office; 2021. Accessed: Feb. 11, 2022. [Online]Available, https://data.europa.eu/doi/10.2760/212427.
[7] Saraf NM, Hamid JRA, Halim MA, Rasam ARA, Lin S. Accuracy assessment of 3-dimensional LiDAR building extraction. 2018 Ieee 14th International colloquium on signal processing & its applications (cspa2018). New York: Ieee; 2018. p. 261–6. Accessed: Feb. 23, 2022[Online]. Available, https://www.webofscience.com/wos/woscc/full-record/WOS:000435278600049.
[8] Wu Y, Blunden LS, Bahaj AS. City-wide building height determination using light detection and ranging data. Environ Plan B Urban Anal City Sci 2019;46(9):1741–55. https://doi.org/10.1177/2399808318774336. Nov.
[9] Bonczak B, Kontokosta CE. Large-scale parameterization of 3D building morphology in complex urban landscapes using aerial LiDAR and city administrative data. Comput Environ Urban Syst 2019;73:126–42. https://doi.org/10.1016/j.compenvurbsys.2018.09.004. Jan.
[10] Teo T-A. Deep-learning for lod1 building reconstruction from airborne lidar data. In: IGARSS 2019 - 2019 IEEE International geoscience and remote sensing symposium; 2019. p. 86–9. https://doi.org/10.1109/IGARSS.2019.8897810. Jul.
[11] Park Y, Guldmann J-M. Creating 3D city models with building footprints and LIDAR point cloud classification: A machine learning approach. Comput Environ Urban Syst 2019;75:76–89. https://doi.org/10.1016/j.compenvurbsys.2019.01.004.
[12] European Environment Agency (EEA). Building height 2012 — copernicus land monitoring service. 2023. Accessed: Mar. 14,[Online]. Available, https://land.copernicus.eu/local/urban-atlas/building-height-2012.
[13] Frantz D, et al. National-scale mapping of building height using sentinel-1 and sentinel-2 time series. Remote Sens Environ 2021;252. https://doi.org/10.1016/j.rse.2020.112128.
[14] Li X, Zhou Y, Gong P, Seto KC, Clinton N. Developing a method to estimate building height from Sentinel-1 data. Remote Sens Environ 2020;240:111705. https://doi.org/10.1016/j.rse.2020.111705. Apr.
[15] Che Y, Li X, Liu X, Zhang X. Characterizing the 3-D structure of each building in the conterminous United States. Sustain Cities Soc 2024;105:105318. https://doi.org/10.1016/j.scs.2024.105318. Jun.
[16] Cao Y, Huang X. A deep learning method for building height estimation using high-resolution multi-view imagery over urban areas: A case study of 42 Chinese cities. Remote Sens Environ 2021;264:112590. https://doi.org/10.1016/j.rse.2021.112590. Oct.
[17] Liu C-J, Krylov VA, Kane P, Kavanagh G, Dahyot R. IM2ELEVATION: Building height estimation from single-view aerial imagery. Remote Sens (Basel) 2020;12(17). https://doi.org/10.3390/rs12172719. Art. no. 17Jan.
[18] Pang HE, Biljecki F. 3D building reconstruction from single street view images using deep learning. Int J Appl Earth Obs Geoinf 2022;112:102859. https://doi.org/10.1016/j.jag.2022.102859. Aug.
[19] microsoft/GlobalMLBuildingFootprints. Microsoft. 2023. Aug. 09,Accessed: Aug. 09, 2023. [Online]. Available, https://github.com/microsoft/GlobalMLBuildingFootprints.
[20] Biljecki F, Ledoux H, Stoter J. Generating 3D city models without elevation data. Comput Environ Urban Syst 2017;64:1–18. https://doi.org/10.1016/j.compenvurbsys.2017.01.001. Jul.
[21] Bernard J, Bocher E, Le Saux Wiederhold E, Leconte F, Masson V. Estimation of missing building height in openstreetmap data: a French case study using Geoclimate 0.0.1. Geosci Model Dev 2022;15(19):7505–32. https://doi.org/10.5194/gmd-15-7505-2022. Oct.
[22] Milojevic-Dupont N, et al. Learning from urban form to predict building heights. PLoS One 2020;15. https://doi.org/10.1371/journal.pone.0242010.
[23] Wu X, Ou J, Wen Y, Liu X, He J, Zhang J. Developing a data-fusing method for mapping fine-scale urban three-dimensional building structure. Sustain Cities Soc 2022;80:103716. https://doi.org/10.1016/j.scs.2022.103716. May.
[24] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on tabular data?" arXiv, Jul. 18, 2022. doi: 10.48550/arXiv.2207.08815.
[25] Bayerische Vermessungsverwaltung, "3D-Gebäudemodelle (LoD2)." in open data - kostenfreie geodaten der bayerischen vermessungsverwaltung. Accessed: Jan. 31, 2023. [Online]. Available: https://geodaten.bayern.de/opengeodata/OpenDataDetail.html?pn=lod2.
[26] Berlin Partner für Wirtschaft und Technologie GmbH. Berlin 3D - Downloadportal des Business Location Centers. 2023. AccessedJan. 31,[Online]. Available, https://www.businesslocationcenter.de/downloadportal.
[27] GeoBasis-DE /LGB. 3D-Gebäudemodelle LoD2 (Level of Detail 2) Brandenburg. 2022. Accessed: Apr. 26,[Online]. Available, https://data.geobasis-bb.de/geobasis/daten/3d_gebaeude/lod2_gml/.
[28] Freie und Hansestadt Hamburg, Landesbetrieb geoinformation und vermessung (LGV), "3D-gebäudemodell LoD2-DE hamburg - metaver." Accessed: Jul. 19, 2022. [Online]. Available: h.t.tps://metaver.de/trefferanzeige?docuuid=2C1F2EEC-CF9F-4D8B-ACAC-79D8C1334D5E#detail_links.
[29] Hessische Verwaltung für Bodenmanagement und Geoinformation. Geodaten online - Downloadcenter. 2022. Accessed: Jul. 19,[Online]. Available, https://gds.hessen.de/INTERSHOP/web/WFS/HLBG-Geodaten-Site/de_DE/-/EUR/ViewDownloadcenter-Start.
[30] Landesamt für Geoinformation und Landesvermessung Niedersachsen. 3D-Gebäudemodell (LoD2). OpenGeoData.NI; 2022. Accessed: Apr. 25,[Online]. Availableh.t.t.ps://opengeodata.lgln.niedersachsen.de/#lod2.
[31] Geobasis NRW. 3D-Gebäudemodell LoD2 - Paketierung: Einzelkacheln." GDI-NW. 2023. Accessed: Jan. 05,[Online]. Available, https://www.opengeodata.nrw.de/produkte/geobasis/3dg/lod2_gml/lod2_gml/.

[32] Landesamt für Geobasisinformation Sachsen (GeoSN). Offene Geodaten. 2022. Accessed: Apr. 26,[Online]. Available, http://www.geodaten.sachsen.de/batch-download-4719.html.
[33] Landesamt für Vermessung und Geoinformation Sachsen-Anhalt (LVermGeo). Kostenfreies 3D-Gebäudemodell - LoD2 - landesweit. Landesportal Sachsen-Anhalt/Geodatenportal Sachsen-Anhalt; 2022. Accessed: May 24,[Online]. Available, https://www.lvermgeo.sachsen-anhalt.de/de/3d-gebaeude-lod2-landesweit.html.
[34] Kompetenzzentrum Geodateninfrastruktur Thüringen (GDI-Th). 3D Gebäudedaten LoD1 & LoD2. Geoportal Thüringen; 2022. Accessed: Apr. 28,[Online]. Availableh.t.t.ps://geoportal.geoportal-th.de/gaialight-th/_apps/atomfeedexplorer/?#feed=h.t.t.ps%3A%2F%2Fgeoportal.geoportal-th.de%2Fdienste%2Fatom_th_gebaeude%3Ftype%3Ddataset%26id%3D97d152b8-9e00-49f3-9ae4-8bbb30873562.
[35] CityGML. Open Geospatial Consortium. 2023. Accessed: Jun. 27,[Online] Available, https://www.ogc.org/standard/citygml/.
[36] OpenStreetMap contributors. OpenStreetMap. 2015 [Online]. Available, https://www.openstreetmap.org.
[37] "Geofabrik download server." Accessed: Dec. 19, 2022. [Online]. Available: https://download.geofabrik.de/europe/germany.html.
[38] "The Shapely User Manual — Shapely 2.0.1 documentation." Accessed: Jun. 27, 2023. [Online]. Available: https://shapely.readthedocs.io/en/stable/manual.html.
[39] Osmosis. OpenStreetMap on GitHub. 2023. Jun. 19,Accessed: Jun. 27, 2023. [Online]. Available, https://github.com/openstreetmap/osmosis.
[40] Henrikki Tenkanen and pyrosm contributors. Pyrosm. 2023. Accessed: Apr. 13, [Online]. Available, https://pyrosm.readthedocs.io/en/latest/.
[41] "Key:height – OpenStreetMap Wiki." Accessed: Jul. 04, 2023. [Online]. Available: h.t.t.ps://wiki.openstreetmap.org/wiki/Key:height#Height_of_buildings.
[42] M. Perry, "Rasterstats: summarize geospatial raster datasets based on vector geometries." Jul. 21, 2022. Accessed: Aug. 26, 2022. [OS Independent]. Available: https://github.com/perrygeo/python-raster-stats.
[43] xgboost developers, "XGBoost documentation." Accessed: Aug. 15, 2023. [Online]. Available: https://xgboost.readthedocs.io/en/stable/python/python_api.html.
[44] © GeoBasis-DE /BKG. 3D Gebäudemodell LoD2 Deutschland. 2021 [Online]. Available, https://gdz.bkg.bund.de/index.php/default/3d-gebaudemodelle-lod2-deutschland-lod2-de.html.
[45] European Commission. Joint Research Centre.. GHSL data package 2023. LU: Publications Office; 2023. Accessed: Mar. 21, 2024. [Online]. Available, https://data.europa.eu/doi/10.2760/098587.
[46] Herfort B, Lautenbach S, Porto de Albuquerque J, Anderson J, Zipf A. A spatio-temporal analysis investigating completeness and inequalities of global urban building data in open street map. Nat Commun 2023;14(1). https://doi.org/10.1038/s41467-023-39698-6. Art. no. 1, Jul.
[47] "height | Keys | OpenStreetMap Taginfo." Accessed: Nov. 22, 2023. [Online]. Available: h.t.t.ps://taginfo.openstreetmap.org/keys/height#combinations.
[48] "height | Keys | OpenStreetMap Taginfo United States of America." Accessed: Nov. 22, 2023. [Online]. Available: https://taginfo.geofabrik.de/north-america:us/keys/height#combinations.
[49] "height | Keys | OpenStreetMap Taginfo Germany, Austria, Switzerland." Accessed: Nov. 22, 2023. [Online]. Available: https://taginfo.geofabrik.de/europe:dach/keys/height#combinations".
[50] "height | Keys | Taginfo - France métropolitaine." Accessed: Nov. 22, 2023. [Online]. Available: https://taginfo.openstreetmap.fr/keys/height#combinations.
[51] "height | Keys | OpenStreetMap Hungary Taginfo." Accessed: Nov. 22, 2023. [Online]. Available: https://taginfo.openstreetmap.hu/keys/height#combinations.