AGILE: GIScience Series, 7, 29, 2026. https://doi.org/10.5194/agile-giss-7-29-2026
Proceedings of the 29th AGILE Conference on Geographic Information Science, 16–19 June 2026.
Eds.: Evelyn Uuemaa, Alexander Kmoch, Holger Virro, and Simon Scheider
This contribution underwent peer review based on a full paper submission.

Enhancing OpenStreetMap Building Footprints through nDSM-Based Geometric Segmentation for AI Training Data

Paul Kuper [1], Ruiqi Liu[1], Hanwen Deng[1] and Martin Breunig [1]

[1] Geodetic Institute, Karlsruhe Institute of Technology, Karlsruhe, Germany

Correspondence: Paul Kuper (kuper@kit.edu)

Abstract. High-quality building footprint labels are a critical prerequisite for training AI-based segmentation models, yet reliable ground truth data is rarely available at scale. On the one hand, vegetation often prevents the reliable determination of buildings when only using imagery data and on the other hand, community-driven open data sources such as OpenStreetMap (OSM) frequently exhibit spatial inconsistencies and incompleteness. This study brings both data sources together: it investigates the potential of utilizing airborne LiDAR-derived Normalized Digital Surface Models (nDSM) to improve building extraction and refine OSM labels. Two automated strategies are implemented and compared: 1) a rule-based region growing algorithm and 2) a Density-Based Spatial Clustering (DBSCAN) pipeline leveraging a multi-dimensional feature space that incorporates nDSM heights and local roughness. As a result, more reliable building footprint labels are generated to be used as training data for AI-based building segmentation. The two methods are evaluated on orthophoto-based ground truth data in Karlsruhe, Germany. Quantitative results demonstrate that the nDSM-based DBSCAN approach yields the most robust performance, achieving an F1-score of 0.94 and an Intersection-over-Union (IoU) of 0.89. This method systematically improves upon the raw OSM baseline by effectively filtering vegetation and correcting geometric misalignments through multi-source constraints, specifically the Normalized Difference Vegetation Index (NDVI) including OSM map data overlap. Finally, conclusions are drawn and the outlook indicates the way to AI-based building segmentation, trained on such labels, to be used in scenarios where high-quality ground truth is unavailable.

Submission Type. Model, case study, analysis, algorithm, dataset.

BoK Concepts. AM(AM8), GD(GD4), IP, DM.

Keywords. OpenStreetMap, nDSM, DBSCAN, Region Growing, Label Enhancement.

1 Introduction

1.1 Background and Motivation

Reliable building footprints are essential for tasks such as urban planning, urban heating analysis, and city modelling. While OpenStreetMap (OSM) provides a globally accessible database, its quality is often inconsistent due to its crowdsourced nature. Common issues include incompleteness, geometric misalignments, and semantic errors, i.e. vegetation or land use structures are incorrectly mapped or do not fit for purpose (Schott et al., 2024). Consequently, raw OSM data in general, and building footprints in particular, are often insufficient for applications that require high geometric accuracy (Senaratne et al., 2017; Brovelli and Zamboni, 2018; Fan et al., 2014). To address these quality gaps, CNN-based deep learning models have become a popular approach (Buyukdemircioglu et al., 2022; He et al., 2022). However, these methods rely heavily on large amounts of high-quality labeled data, which is often unavailable in real-world projects. Airborne LiDAR offers a robust alternative that is less dependent on ground truth data. By deriving a Normalized Digital Surface Model (nDSM), building roofs can be effectively distinguished from bare ground and vegetation based on their distinct elevation and planar geometry.

1.2 Problem Statement and Research Objectives

The core technical challenge addressed in this study is the discrepancy between the raw geometric information provided by nDSM and the semantic richness of OSM data. While the nDSM provides precise elevation data, it lacks semantic labels, making it difficult to distinguish buildings from other elevated objects like dense vegetation without additional processing. Conversely, OSM provides semantic building labels but suffers from spatial inconsistencies and incompleteness. Therefore, an automated framework is needed that can leverage the geometric accuracy of the nDSM to validate and refine geometrically inconsistent OSM footprints.

To address this problem, this study investigates two extraction strategies:

1. Rule-based segmentation: A region growing algorithm that delineates buildings based on local surface homogeneity constraints.
2. Unsupervised clustering: A DBSCAN pipeline that utilizes a custom multidimensional feature space to separate buildings from non-building objects.

It targets the following specific objectives:

1. Feature space analysis: evaluate how the design of the feature space, specifically the integration of nDSM height (h) and local roughness (σ), influences the ability of density-based clustering to separate building roofs from vegetation.
2. Benchmarking extraction strategies: To quantitatively compare the performance of the proposed region growing and DBSCAN pipelines against the OSM baseline, using manually corrected ground truth blocks to evaluate the overall extraction accuracy and reliability.

1.3 Related Work

Research on automated building extraction relies heavily on the geometric analysis of LiDAR-derived normalized digital surface models. To segment these models, traditional approaches such as region growing exploit local surface homogeneity to delineate roof planes (Awrangjeb et al., 2014). While effective, these methods can be rigid compared to density-based clustering (DBSCAN), which handles arbitrary shapes more flexibly. Wang et al. improved the robustness of DBSCAN by developing an automated parameter estimation method based on point density distributions (Wang et al., 2019). Furthermore, Du et al. demonstrated that integrating higher-order geometric features significantly improves the distinction between man-made structures and irregular vegetation (Du et al., 2017). However, height information alone is often insufficient to fully separate buildings from dense tree canopies. To resolve this, spectral information is frequently integrated. Meng et al. utilized indices like NDVI to adaptively mask vegetation and refine building boundaries using image consistency constraints (Meng et al., 2022). Zhuo et al. (2018) demonstrated that remote sensing data can be used to correct and align OSM footprints with physical structures. Microsoft® (2022) released a worldwide building footprints dataset including heights, under the same Open Database Licence as OSM, with polygons produced by machine learning algorithms having been trained with satellite imagery. Unlike prior work focusing on standalone building extraction, this study combines nDSM-derived geometric features with OSM semantic information to correct and enhance building footprint labels for use as training data.

2 Data and Methodology

2.1 Study Area and Data

Due to the high density of OSM data, the city of Karlsruhe, Germany, has been selected as study area, which is characterized by a diverse landscape, ranging from dense inner-city residential blocks to industrial zones and suburban areas characterized by substantial vegetation cover (Fig. 1). Due to the lack of pixel-wise building masks for the study area, an independent reference dataset was created to enable quantitative evaluation. Ten representative 100 × 100 m blocks were selected across different urban and suburban settings. For each block, building polygons were manually inspected and corrected in QGIS using orthophotos and the nDSM as ground truth.

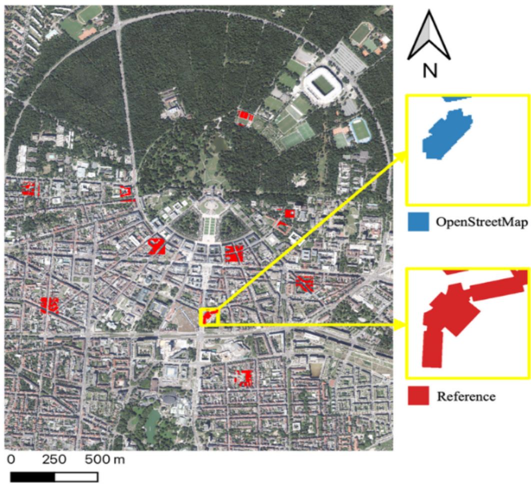

Figure 1. Study area in Karlsruhe, Germany, with locations of the selected 100 × 100 m evaluation blocks.

Although OSM building footprints in Karlsruhe exhibit high overall completeness, they still suffer from several issues, including spatial misalignment of building footprints, inaccurate boundaries, and missing buildings.

For the Karlsruhe region, Digital Elevation Model (DEM) datasets derived from LiDAR (Light Detection and Ranging) were acquired in the year 2022 from the Open Data Portal of the State Office for Geoinformation and Land Development (LGL), Baden-Württemberg, Germany. The datasets consist of a 1 m resolution Digital Surface Model (DSM), which includes the elevations of all surface features such as buildings and vegetation, and a 0.25 m resolution Digital Terrain Model (DTM), which represents the bare ground surface after removing above-ground objects. To eliminate the influence of terrain variations on building height information, a normalized Digital Surface Model (nDSM) was further derived in this paper, which is computed as follows:

$$nDSM(x,y) = DSM(x,y) - DTM(x,y) \quad (1)$$

Therefore, nDSM datasets serve as reliable geometric reference data for the correction and validation of OSM building footprints, as the example shown in Fig. 2.

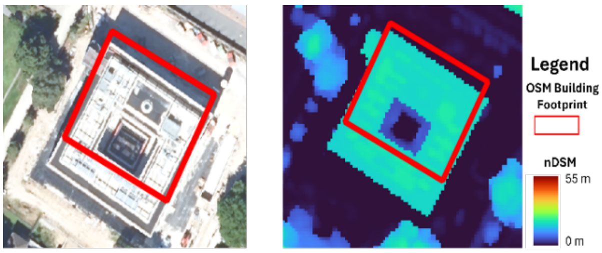

Figure 2. Example of OSM building footprint (in Karlsruhe, Germany) overlaid on orthophoto (left) and nDSM (right).

To distinguish buildings from vegetation, the Normalized Difference Vegetation Index (NDVI) was computed from the red and near-infrared bands of high-resolution orthophotos (Eq. 2).

$$NDVI = \frac{NIR - Red}{NIR + Red} \quad (2)$$

By incorporating this constraint, interference from tall vegetation can be effectively suppressed.

2.2 Building Label Enhancement Framework

The workflow of the proposed label enhancement framework is presented in Fig. 3.

In the first phase, multi-source data are pre-processed and transformed into complementary feature representations, including geometric characteristics extracted from the nDSM (i.e., terrain roughness and slope), vegetation masks, and semantic priors generated by rasterizing OSM building footprints. The second phase serves as the core processing stage and performs geometry-driven unsupervised segmentation to produce initial building candidates, using region growing and DBSCAN algorithms described in detail in Sections 2.2.1 and 2.2.2. In the final phase, a multi-source refinement stage is applied to enhance the quality of the segmentation outputs. NDVI masks are first used to suppress vegetation-related false positives. Subsequently, spatial consistency with OSM building footprints is enforced to remove geometrically inconsistent segments. Finally, morphological post-processing, including fragment elimination and topological smoothing, are performed to improve boundary continuity and spatial completeness.

2.2.1 Region Growing Algorithm

Region growing has been adopted as a geometry-driven segmentation strategy operating on the nDSM to extract building footprints. The segmentation is initialized from a set of seed pixels and iteratively expands under height and local homogeneity constraints to ensure structural consistency. To obtain reliable building cores, a height-based watershed segmentation is first applied to the inverted nDSM and constrained by the OSM building footprint mask, resulting in a set of compact roof core regions. For each core, a single representative seed is selected using a hierarchical quality criterion that prioritizes pixels with low local height roughness (3×3 standard deviation $\sigma \leq 0.4$ m) and sufficient elevation ($h \geq 2$ m). If no pixel satisfies the strict roughness criterion, the threshold is relaxed to $\sigma \leq 0.6$ m to tolerate DSM noise and slight roof slope variations while preserving geometric consistency of roof surfaces. Among the candidate pixels, the one whose height is closest to the median height of the roof core is selected as the seed.

Starting from these seeds, the region growing process is implemented in three successive passes. In the first pass, regions are expanded under strict homogeneity constraints to generate compact and conservative roof segments that preserve planar roof structures while suppressing irregular surfaces. In the second pass, relaxed thresholds are applied to allow controlled expansion into neighboring pixels, which facilitates boundary recovery and incorporation of adjacent building components without introducing background noise. In the final pass, small isolated fragments are merged into nearby regions to further reduce noise and improve the spatial completeness of the extracted roof segments.

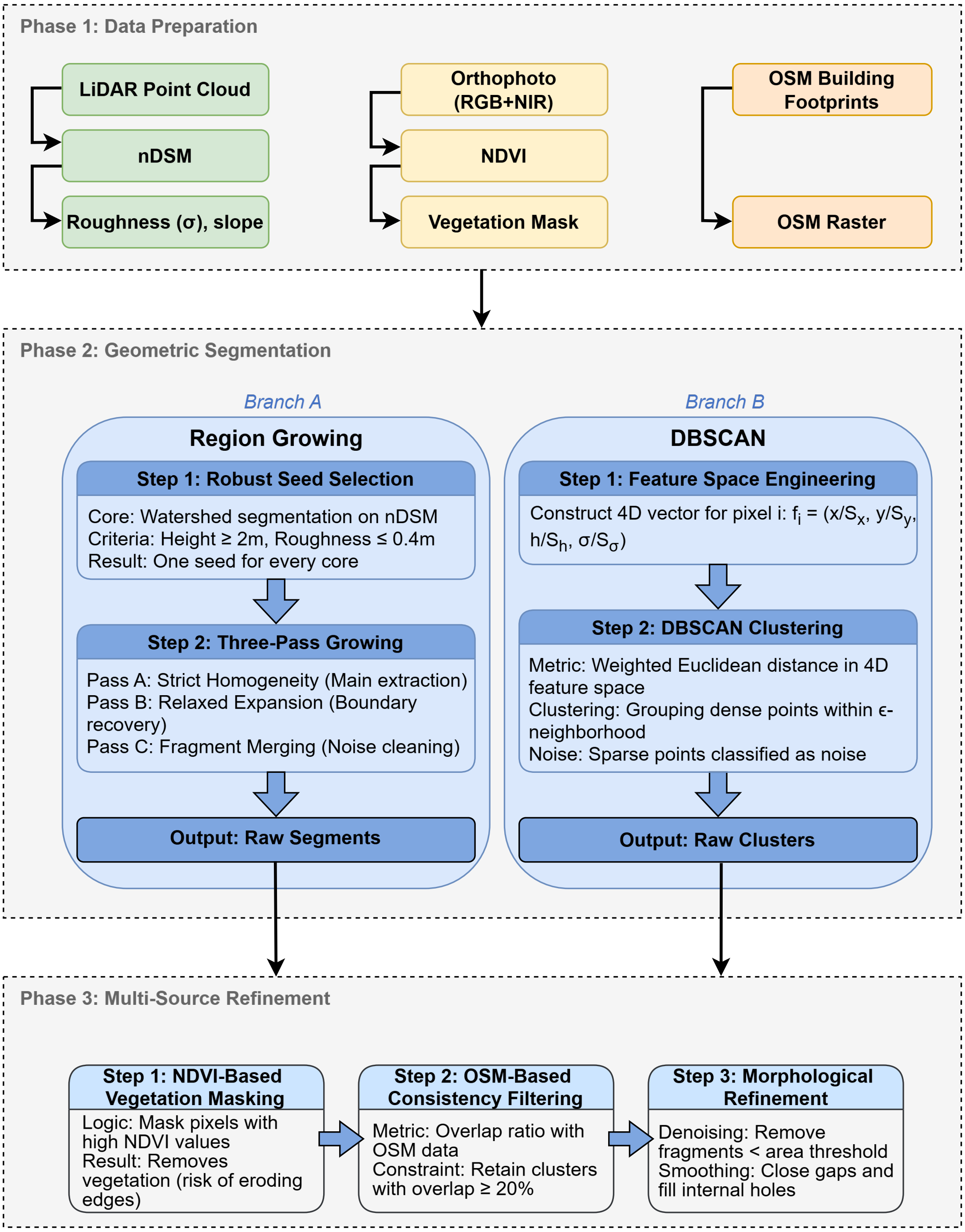

Figure 3. Workflow of the proposed label enhancement framework: The process moves from data preparation (Phase 1) to geometric segmentation (Phase 2) and multi-source refinement (Phase 3).

2.2.2 DBSCAN Algorithm

The Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm is governed by two hyperparameters: the neighborhood radius ε and the minimum number of points (MinPts) required to define a dense region. A point is classified as a core point if at least MinPts neighbors fall within a radius ε. Clusters are subsequently expanded by iteratively aggregating density-reachable points, whereas isolated points are labeled as noise. Conventionally, DBSCAN operates on spatial coordinates only. For each pixel i, a feature vector f_i is represented using spatial coordinates:

$$f_i = (\frac{x_i}{S_x}, \frac{y_i}{S_y}) \tag{3}$$

where x_i and y_i denote the pixel coordinates, and S_x and S_y are scaling factors used to normalize spatial units.

However, due to the absence of geometric information, spatially adjacent pixels belonging to physically different structures tend to be merged into the same cluster, such as roof pixels and neighbouring vegetation canopies. To enhance geometric discrimination, an augmented feature representation is introduced in Eq. (4) by extending the spatial feature vector with height and surface roughness information:

$$f_i = \left(\frac{x_i}{s_x}, \frac{y_i}{s_y}, \frac{h_i}{s_h}, \frac{\sigma_i}{s_\sigma}\right) \quad (4)$$

where h_i corresponds to the nDSM height and σ_i represents local surface roughness computed as the standard deviation within a 3×3 window. By incorporating nDSM height and roughness, pixels exhibiting spatial proximity but distinct geometric characteristics (e.g., a roof edge versus adjacent vegetation) become separable in the feature space. Similarity between pixels is quantified using the Euclidean distance:

$$d(i,j) = \|f_i - f_j\|_2 \quad (5)$$

A point j is considered a neighbor of point i if $d(i,j) \leq \varepsilon$. The scaling factors and DBSCAN hyperparameters were empirically selected on a validation set to balance spatial and geometric contributions and improve the separation between building and non-building clusters. The resulting building-label raster is then polygonised and passed to Phase 3 for refinement.

2.3 Data and Software Availability

The code used to reproduce the experiments in this paper is freely available in a GitHub repository: https://github.com/hanwendeng/label-enhancement-dbscan-region-growing. Datasets used in this paper: https://doi.org/10.5281/zenodo.18368398. The OpenStreetMap building footprints were obtained from Geofabrik: https://download.geofabrik.de/. DEMs and orthophotos were obtained from the State Office for Geoinformation and Land Development Baden-Württemberg (LGL): https://opengeodata.lgl-bw.de.

3. Results and Discussion

3.1 Quantitative Evaluation

As shown in Table 1, the post-processed building segmentation performance of DBSCAN, region growing, and the OSM baseline is evaluated over ten 100 × 100 m manually corrected validation regions.

Table 1. Mean building extraction performance over ten manually corrected 100 × 100 m validation blocks. The OSM baseline is included for comparison (Liu et al., 2026).

Method	IoU	Precision	Recall	F1
OSM baseline	0.844	0.934	0.900	0.905
Region Growing	0.881	0.931	**0.944**	0.936
DBSCAN	**0.894**	**0.956**	0.933	**0.943**

Overall, the DBSCAN-based method attains the best numerical performance, with an average IoU of 0.89 and F1-score of 0.94, and the highest precision (0.96) at still high recall (0.93). Region growing is a close second: it reaches an IoU of 0.88 and an F1-score of 0.94, with slightly lower precision but slightly higher recall than DBSCAN, indicating that it tends to avoid missed detections at the cost of a few more false positives, whereas DBSCAN behaves more conservatively.

3.2 Qualitative Evaluation

Fig. 4 compares both methods on a representative area in central Karlsruhe. The columns show, from left to right, manually created ground truth, OSM footprints, region growing and DBSCAN results.

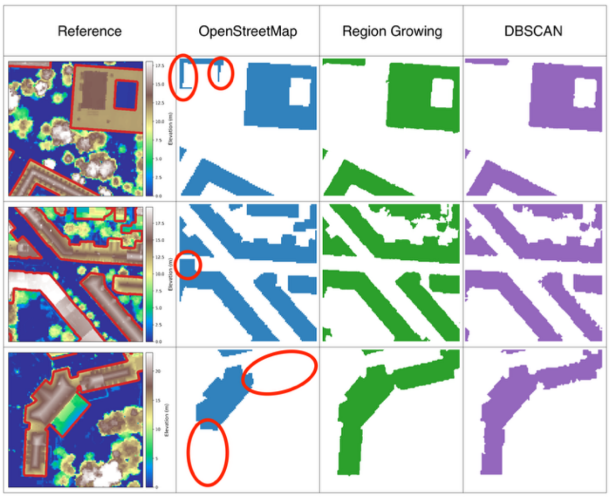

Figure 4. Comparison of our ground truth data (building footprints in red and nDSM as background), building footprints from OSM, region growing and DBSCAN.

Overall, the DBSCAN results seem to contain fewer unwanted small structures compared to the region growing method, while the structures itself appear fuzzier.

3.3 Limitations and Discussion

Although the proposed framework achieves competitive performance, several limitations should be considered. First, the building segmentation process is inherently tied to the specific characteristics and quality of the input nDSM. Systematic physical defects pose a significant challenge: dense tree canopies frequently occlude building roofs, causing local height and roughness signals to reflect the vegetation rather than the building. Furthermore, specific materials such as glass or lightweight roofing are poorly captured by airborne LiDAR, resulting in weak or missing elevation returns. In these scenarios, the geometric information is physically absent, rendering recovery impossible for both rule-based and learning-based strategies.

A further constraint arises from temporal and spatial mismatches between datasets. The nDSM, orthophotos, and OSM data originate from different acquisition times and processing chains. Despite careful co-registration, residual shifts and temporal changes (e.g., new construction or demolition) can result in discrepancies where buildings appear in the imagery or nDSM but are absent from the OSM data, or vice versa.

4 Conclusion and Future Work

In this paper we presented and compared two methods for generating accurate building footprints by combining OSM with nDSM data. Our results demonstrate that building footprints can be improved even in areas with high-quality OSM data, such as the city center of Karlsruhe, Germany. Furthermore, we showed that the DBSCAN-based method outperforms the region-growing approach used in this study, although both methods have their specific advantages.

Future research should address both data quality and methodological design. On the data side, improved separation of tree canopies from roofs would enhance the nDSM signal, benefiting all extraction methods. On the algorithmic side, hand-tuned parameters in the clustering pipelines could be replaced by adaptive, data-driven estimations. The next step is to use the results of this study, specifically the improved building footprints, as label data to train a U-net-based neural network.

Declaration of Generative AI in writing

AI tools were utilized for language editing, improving grammar, and sentence structure, but not for generating scientific content, research data, or substantive conclusions. All intellectual and creative work, including the analysis and interpretation of data, is original and has been conducted by the authors without AI assistance.

Acknowledgements

We sincerely thank the LGL (Landesamt für Geoinformation und Landentwicklung, Karlsruhe) for providing the DTM dataset. Map data by OpenStreetMap contributors, https://www.openstreetmap.org/copyright.

References

Awrangjeb, M. and Fraser, C. S.: Automatic Segmentation of Raw LIDAR Data for Extraction of Building Roofs, Remote Sensing, 6(5), 3716–3751, https://doi.org/10.3390/rs6053716, 2014.

Brovelli, M. A., & Zamboni, G. (2018). A New Method for the Assessment of Spatial Accuracy and Completeness of OpenStreetMap Building Footprints. ISPRS International Journal of Geo-Information, 7(8), 289. https://doi.org/10.3390/ijgi7080289

Buyukdemircioglu, M., Can, R., Kocaman, S., and Kada, M.: Deep Learning Based Building Footprint Extraction from Very High Resolution True Orthophotos and nDSM, ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci., V-2-2022, 211–218, https://doi.org/10.5194/isprs-annals-V-2-2022-211-2022, 2022.

Du, S., Zhang, Y., Zou, Z., Xu, S., He, X., and Chen, S.: Automatic Building Extraction from LiDAR Data Fusion of Point and Grid-Based Features, ISPRS Journal of Photogrammetry and Remote Sensing, 130, 294–307, https://doi.org/10.1016/j.isprsjprs.2017.06.005, 2017.

Fan, H., Zipf, A., Fu, Q., Neis, P.: Quality assessment for building footprints data on OpenStreetMap, International Journal of Geographical Information Science 28 (4), 700-719, 2014.

He, H., Gao, K., Tan, W., Wang, L., Fatholahi, S. N., Chen, N., Chapman, M. A., and Li, J.: Impact of Deep Learning-Based Super-Resolution on Building Footprint Extraction, Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XLIII-B1-2022, 31–37, https://doi.org/10.5194/isprs-archives-XLIII-B1-2022-31-2022, 2022.

Liu, R., Krikau, S., Kuper, P.V., Benz, S., Deng, H. and Breunig, M.: Spatio-Temporal Data Management and Enhanced Processing for Urban Heat Island Analysis supported by Digital Elevation Models, FIG Congress 2026 in Cape Town, South Africa, ISSN 2308-3441.

Meng, Y., Chen, S., Liu, Y., Li, L., Zhang, Z., Ke, T., and Hu, X.: Unsupervised Building Extraction from Multimodal Aerial Data Based on Accurate Vegetation Removal and Image Feature Consistency Constraint, Remote Sensing, 14(8), 1912, https://doi.org/10.3390/rs14081912, 2022.

Microsoft, Global ML footprints, https://wiki.openstreetmap.org/wiki/Microsoft_Building_Footprint_Data, 2022.

Senaratne, H., Mobasheri, A., Ali, A. L., Capineri, C., & Haklay, M. (Muki). (2017). A review of volunteered geographic information quality assessment methods. *International Journal of Geographical Information Science*, *31*(1), 139–167. https://doi.org/10.1080/13658816.2016.1189556

Schott, M., Zell, A., Lautenbach, S., Sümbül, G., Schultz, M., Zipf, A., Demir, B.: Analyzing and improving the quality and fitness for purpose of OpenStreetMap as labels in remote sensing applications. In: Burghardt, D., Demidova, E., Keim, D. (Eds.): Volunteered geographic information. (2024), 21-42 https://doi.org/10.1007/978-3-031-35374-1_2

Wang, C., Ji, M., Wang, J., Wen, W., Li, T., and Sun, Y.: An Improved DBSCAN Method for LiDAR Data Segmentation with Automatic Eps Estimation, Sensors, 19(1), 172, https://doi.org/10.3390/s19010172, 2019.

Zhuo, X., Fraundorfer, F., Kurz, F., and Reinartz, P.: Optimization of OpenStreetMap Building Footprints Based on Semantic Information of Oblique UAV Images, Remote Sensing, 10(4), 624, https://doi.org/10.3390/rs10040624, 2018.