Automatic upgrade of 3D building models to LoD3 using 3D Point Clouds and Grounding DINO

Anass Yarroudh, Abderrazzaq Kharroubi, Imane Jeddoub, Roland Billen

GeoScITY, Spheres Research Unit, University of Liège, 4000 Liège, Belgium (ayarroudh, akharroubi, i.jeddoub, rbillen)@uliege.be

Keywords: City Model, CityJSON, Building, Level of Detail, Mobile Mapping System, Point Cloud.

Abstract

The advancement of urban digital twins depends on the accurate representation of 3D city models, particularly Level of Detail 3 (LoD3) models, which incorporate detailed façade features essential for urban planning applications. However, generating LoD3 models is challenging due to the complexities of semantic segmentation in 3D point cloud data and the high resource demands of traditional methods. This paper presents an automated methodology for upgrading existing Level of Detail 2.2 (LoD2.2) building models to LoD3 using mobile mapping point cloud data and the Grounding DINO model. The approach begins with extracting façade surfaces from LoD2.2 models while maintaining geometric integrity. Point cloud data is then transformed into a 2D image format to facilitate the application of Grounding DINO, which accurately detects and segments façade elements such as windows and doors. The identified features are re-integrated into the 3D model, resulting in an enhanced LoD3 representation. This methodology demonstrates effectiveness and scalability, providing a practical solution for improving urban digital twins with detailed and reliable building models.

1. Introduction

Urban digital twins as advanced digital representations of cities enable comprehensive urban planning, management, and analysis. At the core of these digital twins are 3D city models, which includes buildings, roads, terrain, and other elements of the built environment. Buildings as dominant element of the urban landscape play an important role for several applications (Biljecki et al., 2015), such as solar potential estimation (Matsuoka et al., 2024), Energy demand estimation (Kaden & Kolbe, n.d.), 3D cadastre (Dursun et al., 2022). The geometric accuracy and semantics complexity of building representations are managed into Levels of Details (LoDs), which range from simple geometric shape to highly detailed structures (Biljecki et al., 2016). Higher level of details, such as LoD 3, include façade elements such as windows, doors, and architectural elements.

Two main approaches are commonly used to create LoD3: either create LoD3 models from scratch or update the facade details of models at a lower level of detail. Scan2LoD3, developed by Wysocki et al., (2023), employs ray casting and Bayesian networks to generate LoD3 models. The method processes three inputs: point cloud, an existing building model, and façade texture. It creates three probability maps: a point cloud probability map via a modified Point Transformer network, a conflicts probability map from laser scanner visibility analysis with the building model, and a texture probability map using Mask-RCNN. These maps are combined using a Bayesian network to produce a target probability map, which guides the 3D reconstruction into a detailed, CityGML-compliant LoD3 building model. Alternatively, enhancing existing LoD2 models with façade details offers a practical approach to upgrading city models to LoD3. This method leverages the existing structural information and augments it with additional data from mobile mapping data, ensuring a more efficient and scalable process. By focusing on the enhancement of specific elements, such as windows and doors, this approach can produce detailed and accurate models without the need for a complete overhaul.

The main challenges with existing methods include the need for semantic segmentation of 3D point cloud data, which is complex and time-consuming. Additionally, there is a significant shortcoming in the availability of labelled datasets, with only one existing dataset that limits the training and validation of models (Wysocki et al., 2023). Furthermore, current methods do not leverage the advances in image foundation models, such as Grounding DINO (Liu et al., 2023), Segment Anything (Kirillov et al., 2023), and Grounded Segment Anything (Grounded-SAM) (Ren et al., 2024). These state-of-the-art models offer powerful tools for image-based object detection and segmentation, which can significantly enhance the accuracy and efficiency of detecting façade details.

This paper explores the latter approach, demonstrating how mobile mapping point cloud data can be used to refine LoD2 models, resulting in comprehensive LoD3 city models suitable for various urban applications. Our methodology involves processing the point cloud data into images and applying Grounding DINO to extract façade details and integrating these details into the existing LoD2.2 models. The contributions of this paper include a novel method for model refinement, a detailed evaluation of the approach. We start with a CityJSON LoD2.2 building model. The first step is to separate the façade surfaces from each building. Next, we identify the points from mobile mapping data that cover these façade surfaces. In the third step, we project the point cloud data of the façade into an image format and use Grounding DINO to detect and delineate windows and doors. The fourth step involves projecting these detected openings back onto the building model. Finally, we conducted a performance evaluation of the opening detection and an assessment of the reconstruction process.

The structure of this paper is as follows: Section 2 reviews related work, Section 3 describes the methodology, Section 4 presents the results and discusses the findings, Section 5 concludes the paper and outlines future work.

2. Related Work

Given the high demand of 3D city models under the Urban Digital Twin umbrella, many semi-automated and fully automated approaches are developed to generate LoD 2 city models (Peters et al., 2022). While these LoD2 city models can fulfill the requirements of a variety of urban applications, the growing need to produce LoD3 models with the lightest possible implementation is necessary (Biljecki et al., 2015). Generating LoD 3 models is still tedious and a non-straightforward process considering the complexity of detecting the facade elements (e.g., windows, doors, to name a few) and their shape diverity. Furthermore, few scholars attempt to bridge the gap in the literature by improving the manual modeling processes by semantically refining the existing LoD 2 models with the facade details. Thanks to the availability of MLS data, it provides accurate, high-density, and street-level point clouds and images. They represent a solid data source for the semantic facade enrichment of city models. Earlier studies focused mainly on facade detection without going further to explain the 3D reconstruction. For instance, the authors in Hoegner & Gleixner (2022) present a method for automatically extracting 3D models of facades and windows from MLS point clouds using voxel segmentation and visibility analysis. Furthermore, a study introduces a semi-automated method to extract building windows from MLS point clouds using voxel-based segmentation, statistical noise filtering, and conditional Euclidean clustering to identify building facades (Zhou et al., 2018). Another related work focused on developing various approaches to generate from scratch LoD3 models from multiple data sources, namely airborne/mobile LiDAR data and oblique, aerial, and UAV images, to reconstruct a CityGML standard-compliant LoD3 model (Gruen et al., 2020). In addition, Tan et al. (2024) propose a method that fuses geometric features with deep learning networks to improve facade-level classification of point clouds. The method addresses the challenge of capturing local geometric information, making it a valuable approach for high-Level of Detail (LoD3) 3D building reconstructions. In this paper, the focus is on the enrichment of an existing LoD 2 semantic building model, also known as refinement strategy (Wysocki et al., 2022, 2024). (Wysocki, Grilli, et al., 2022; Wysocki et al., 2023, 2024), provide a Scan2LoD3 approach based on the ray-casting multimodal fusion method for 3D reconstruction of the LoD3 building model. The Scan2LoD3 method uses (MLS) point clouds and semantic 3D building models prior to probabilistically detecting model conflicts, integrating these with multimodal data through a Bayesian network. This approach allows an automatic LoD3 reconstruction of models with facade elements like windows and doors compliant with the CityGML standard. The multimodal probabilistic fusion shows its potential to guide the semantic LoD3 facade element reconstruction. Furthermore, a study conducted by Froech et al. (2024) proposes a method for reconstructing 3D facade details using MLS point clouds combined with a predefined 3D model library, leveraging an enhanced Bag of Words (BoW) approach using semi-global features, allowing to address the assumption of rectangularity and the use of bounding boxes. On the other hand, advances in machine learning rely heavily on 2D image-based approaches to detect the facade elements in images. (Hensel et al., 2019) propose a workflow for enhancing LoD2 CityGML models through object detection using 2D images and deep learning, with a focus on improving facade modeling. The authors deployed a modified Faster R-CNN to effectively segment and predict depth for more detailed facade features like windows and balconies. However, the lack of annotated datasets for depth estimation remains a challenge, requiring assumptions or further research to improve neural network performance. Moreover, Pang & Biljecki (2022) present a deep learning approach to reconstructing 3D building models from single street view images, using image-to-mesh reconstruction for outdoor scenes across three scenarios: standalone reconstruction, footprint-aided reconstruction, and refinement of existing models. Results indicate that the latter two scenarios can accurately reconstruct building geometry, while standalone reconstruction estimates building mass. This method facilitates 3D modeling in areas lacking data, enabling new geospatial analyses. Fan et al. (2021) propose a VGI3D platform, a novel web-based tool designed for fast and cost-effective 3D building modeling using convolutional neural networks applied to volunteered geographic information (VGI) and images. It simplifies the process by automatically extracting facade elements like windows and doors. (Murtiyoso et al., 2021) present a novel method for semantic segmentation of building facades using a deep learning (DL) approach combined with transfer learning. The method first segments 2D orthophoto images of buildings and then back-projects the segmented data into 3D point clouds. The study highlights the importance of image quality and suggests further improvements like enhancing orthophotos and experimenting with different DL models for better accuracy in facade classification. A recent work by Salehitangrizi et al. (2024) proposes a new method that leverages multi-view images, Faster R-CNN, and Segment Anything (SAM) deep learning models to detect and accurately project 2D façade elements into 3D space. This paper addresses challenges in creating detailed 3D building façade models (LoD 3), particularly the issues of occlusions and spatial uncertainties in 3D locations of façade elements. While the current state of the art relies on combined approaches (3D point clouds and 2D images) for facade element detection and extraction, as well as upgrading the LoD2 to LoD 3 models, we are currently leveraging advances in image-based approaches, namely Grounded Segment Anything (Grounded-SAM) (Ren et al., 2024).

3. Methodology

Our methodology for upgrading LoD2.2 building models to LoD3 using mobile mapping point cloud data consists of several systematic steps to integrate façade details into existing models. as summarized in Figure 1 and explained in subsections below.

The preprocessing step merges wall surfaces within the LoD2.2 models based on topological and coplanarity criteria, simplifying the geometry for subsequent processing (**1.1**). Next, façade points are extracted from the mobile laser scanning data, focusing on points that correspond to the building's wall surfaces (**1.2**). These points are then transformed into a 2D image format by aligning the coordinate system with the façade's normal vector (**1.3**) to facilitate the use of the Grounded Segment Anything model for detecting façade openings such as windows and doors (**1.4**). The detected features are re-projected onto the original 3D model to integrate these detailed elements into the LoD3 model (**1.5**). Finally, postprocessing refines the model, resulting in an upgraded representation suitable for urban applications.

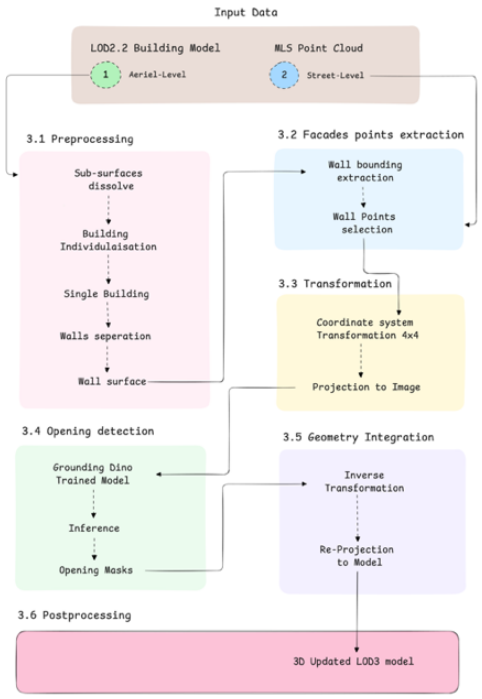

Figure 1. Overview of the full workflow.

1.1 Preprocessing

As shown in Figure 2, the preprocessing of LoD2.2 models involves merging wall surfaces that represent the same facade based on two criteria: topology and coplanarity. First, a topological check is performed to identify adjacent surfaces that share at least one edge. Secondly, the coplanarity of these adjacent surfaces is evaluated by comparing their normal vectors. Surfaces with parallel normal vectors are considered coplanar. If both the topological and coplanarity criteria are satisfied, the wall surfaces are merged to form a single, larger wall surface. This ensures accurate segmentation of the 3D point cloud into sections corresponding to each facade, thereby avoiding missing or incomplete detections of openings.

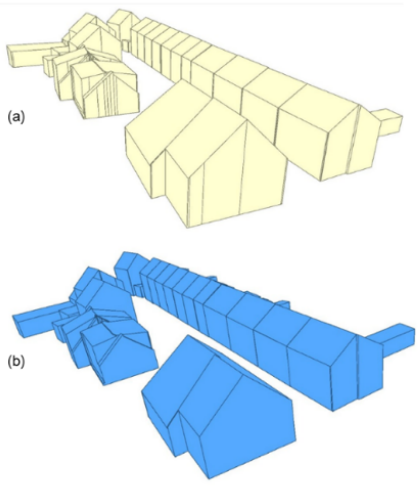

Figure 2. Preprocessing: (a) Input model (b) Merged surfaces.

1.2 Facades points extraction

To process each facade individually, we proceed to a segmentation step to separate different facades. For each 3D building model, we iterate through the wall surfaces in the building's geometry to extract corresponding points from the MLS data. Initially, we filter points within the building's spatial extent. Then, inliers that align with each façade's surface plane are identified. These points are identified based on a perpendicular distance threshold relative to the plane.

1.3 Transformation

As our approach involves using Grounding DINO for inference, 3D points of each facade are projected into a 2D image format by aligning the global coordinates system of the points with the façade's plane as shown in Figure 3.

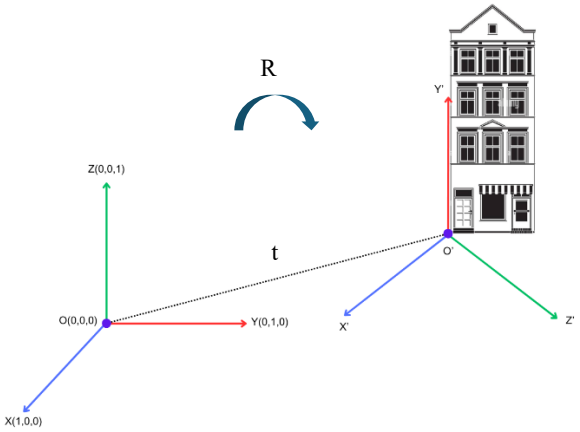

Figure 3. Transformation of original points to a new local coordinates system defined on the facade.

The new local coordinates system has its z-axis aligned with the normal vector of the wall surface. The transformation to this local system is performed as follows:

First, we construct the local coordinate system axes:

- The z-axis of the new coordinate system is aligned with the normal vector n.
- The x-axis is computed as the cross product of the normal vector n and the global z-axis:

$$X' = \frac{n \times Z}{|n \times Z|}$$

  Where X' is the x-axis of the new coordinate system and Z is the global coordinate system's z-axis.
- The y-axis is determined by the cross product of the x-axis and the z-axis.

The new coordinates system's origin O' is at a point on the plane which could be any vertex of the wall surface.

The translation vector t is the difference between the new origin O' and the origin of the global coordinate system O:

$$t = O' - O$$

The rotation matrix R maps the global coordinates system basis vectors to the new local basis vectors. It is calculated as the product of individual rotations around the global coordinates system:

$$R = R_x(\alpha) \cdot R_y(\beta) \cdot R_z(\gamma)$$

Where α, β and γ are rotation angles around the x-axis, y-axis and z-axis respectively.

The full transformation matrix M is then given by:

$$M = \begin{bmatrix} R & t \\ 0 & 1 \end{bmatrix}$$

Finally, the transformation of a point from the global to the local coordinate system is performed as follows:

$$p' = M \cdot p$$

Where $p = (x, y, z, 1)$ are the homogeneous coordinates of the point in the global coordinate system, and $p' = (x', y', z', 1)$ are the homogenous local coordinates.

Once the points are transformed into the local coordinate system, they are then projected onto a 2D plane by mapping their x and y coordinates to pixel coordinates. The resolution σ of this mapping is defined to control granularity of the image. The pixel coordinates (u, v) are calculated as follows:

$$u = \frac{x' - \min(x')}{\sigma}$$

$$v = \frac{\max(y') - y'}{\sigma}$$

As shown in Figure 4, the result is a 2D image array where each pixel corresponds to a point, or a set of points, in the wall surface's plane, with its color determined by the original point cloud data. In case two points have the same pixel coordinates, the pixel is assigned the color value of the farthest point, i.e. the one with the highest z' coordinate.

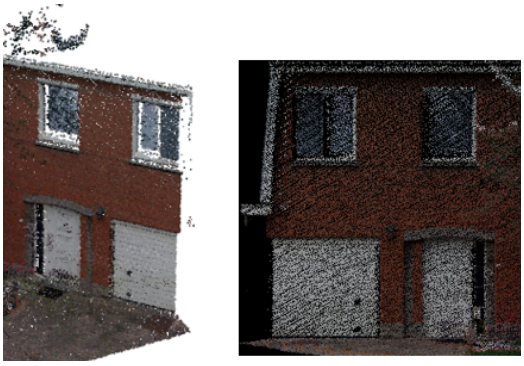

Figure 4. Example of generated image (right) from 3D point cloud of facade (left).

1.4 Openings detection

In the opening detection step, the façade points, now projected into a 2D image format, are analysed using the Grounding DINO model to identify doors and windows. Grounding DINO[1], an open-set object detector, is designed to detect arbitrary objects based on human language inputs, such as category names or referring expressions (see example in Figure 5**).** This model combines the Transformer-based detector DINO with grounded pre-training, enabling it to generalize to open-set concepts by efficiently fusing language and vision modalities.

The architecture of Grounding DINO includes an image backbone for feature extraction, a text backbone for processing language inputs, and a feature enhancer for integrating these modalities. The model employs a language-guided query selection to refine the queries based on the cross-modality features, followed by a cross-modality decoder that retrieves relevant features from both image and text inputs. This approach allows Grounding DINO to detect and label façade openings, such as windows and doors, by aligning textual descriptions with visual data.

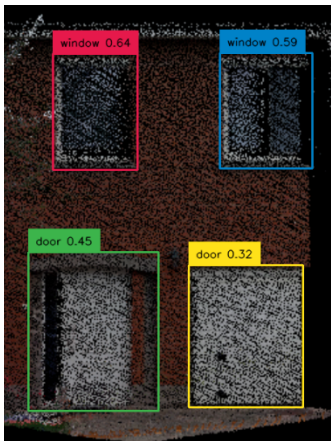

Figure 5. Example of prompt to detect and label windows and doors.

1.5 Geometry integration and post-processing

In the fifth step, the detected features are re-projected onto the original 3D model. The global coordinates of the detected façade elements are then recalculated using the inverse transformation matrix M^{-1}. The geometry of each facade is modified by creating openings in the wall surface polygon, then the corresponding geometries of the detected elements are added to the model. The orientation of each added geometry is aligned with the façade's plane normal vector.

Steps 2 to 5 are then repeated for each building in the dataset. Finally, a postprocessing step refines the model, we employed two filtering criteria: elevation and area. Elevation helps distinguish between windows and doors, as doors generally are at lower elevation close to the ground floor level, while the area criterion filters openings based on their size, filtering some false detections.

[1] https://github.com/IDEA-Research/GroundingDINO

4. Experiments and Results

We assessed the performance of Grounding DINO in detecting façade openings. The model showcased a relatively good capability to accurately identify bounding boxes of different elements, based on simple text prompts. However, the detection quality depends on the chosen box and text thresholds. Depending on the completeness of the 3D point cloud of the façade, these values should be adjusted to control the correctness of the prediction.

Our test dataset, which is an MLS colorized point cloud data from Belgium as shown in Figure 6. The dataset comprised 65 windows. Grounding DINO successfully detected 40, resulting in a detection rate of 61.53%. This performance indicates that the model can recognize the façade features.

Figure 6. Testing dataset: buildings' footprints (left), corresponding MLS data (right).

Additionally, the detected features were correctly mapped and added to the 3D models as illustrated in Figure 7. The results are written to CityJSON format and inspected in Ninja Viewer. The correct orientation of windows is ensured as the surface geometry of each element is aligned with the wall surface normal vector.

However, we also noted some limitations in the method. The model occasionally produced false positives, which were primarily attributed to occlusions from nearby objects, such as cars and trees. These occlusions hindered the model's ability to accurately discern façade openings. This issue was particularly evident in dense areas, where a high concentration of surrounding objects increased the likelihood of occlusions in the 3D point cloud. Furthermore, we found that the density of the point cloud had a significant impact on detection accuracy and completeness. Lower point cloud densities resulted in lower-resolution façade images, which, in turn, led to less accurate and less complete object detection outcomes. In particular, the lower resolution limited the model's ability to detect smaller façade features, resulting in some missed openings or imprecise boundary definitions.

Besides the nearby objects, the angle and distance of the data capture also played a role in the detection quality and the presence of occlusions. Point clouds generated from oblique angles or greater distances tended to yield fewer points per façade, further complicating accurate detection. This was noted especially for higher buildings where the façade's points density of the higher floors is less than This sensitivity to capture conditions highlights the need for careful planning in data acquisition, as certain angles or positions might exacerbate occlusions or reduce point density in critical areas.

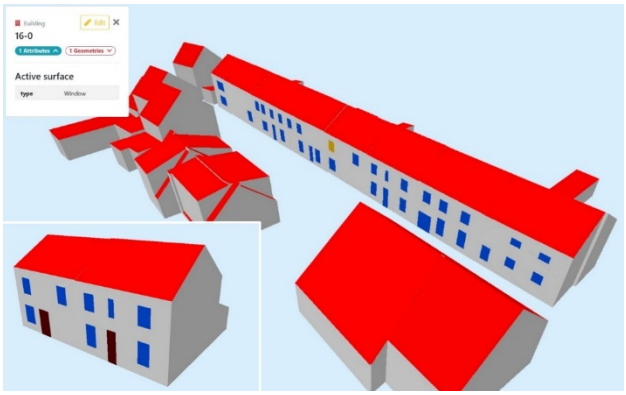

Figure 7. LoD3 Building Models.

5. Conclusions

In this paper, we presented a novel methodology for upgrading LoD2.2 building models to LoD3 using mobile mapping 3D point clouds and advanced image-based object detection techniques, specifically the Grounding DINO model. Our approach systematically integrates façade details into existing models, enriching urban digital representations with greater spatial detail and accuracy.

The automated detection of façade elements facilitated by Grounding DINO significantly reduces the time and effort required to annotate and train context-specific detectors. By using colorized 3D point clouds, the model enables direct detection of elements without the preliminary step of selecting the most representative image of the façade, as is typically required in image-based methods. This methodology is also designed to be scalable, making it suitable for large urban areas where traditional modeling techniques may be impractical due to time and resource constraints.

However, there are some limitations associated with this approach. One significant limitation is its dependency on data quality; the effectiveness of the approach heavily relies on the completeness and density of the input point cloud, as low-density data can lead to inaccuracies and omissions in the detected façade elements. Our results demonstrated that occlusions from nearby objects such as cars, trees, and other urban elements, particularly in dense areas, can create false positives or obscure façade features, affecting the accuracy of detection. Additionally, the angle and distance of data capture play a crucial role in detection quality. For example, oblique or distant point clouds yield fewer points per façade, especially on higher floors, making detection more challenging.

Furthermore, the lack of extensive annotated 3D point cloud datasets for training and validating 3D detection models remains a challenge. While Grounding DINO operates as a zero-shot detector in this methodology, it requires a transformation step to convert façade points to image format for processing, which can add computational overhead and impact performance in large-scale implementations.

Overall, despite these limitations, the presented approach holds significant promise for enhancing digital urban models through efficient, automated detection of architectural features, paving the way for more detailed and accurate urban representations.

Acknowledgements

This research was conducted as part of the Belgian TrackGen project and supported by Logistics in Wallonia and SPW Recherche. We sincerely appreciate their financial support.

We would like also to thank our partners GIM Wallonie and Sirris for their contribution to this work, and Cyclomedia for the provided testing data.

References

Chan, K.L., Qin K., 2017: Biomass burning related pollution and their contributions to the local air quality in Hong Kong. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci*., XLII-2/W7, 29-36. doi.org/10.5194/isprs-archives-XLII-2-W7-29-2017.

Dubayah, R.O., Swatantran, A., Huang, W., Duncanson, L., Tang, H.,Johnson, K., Dunne, J.O., Hurtt, G.C., 2017. CMS: LiDAR-derived Biomass, Canopy Height and Cover, Sonoma County, California, 2013. ORNL DAAC, Oak Ridge, Tennessee, USA. doi.org/10.3334/ORNLDAAC/1523.

Förstner, W., Wrobel, B., 2016: *Photogrammetric Computer Vision*. Springer Nature, Cham.

Gago-Silva, A., 2016. GRASS GIS in Grid Environment. doi.org/10.6084/m9.figshare.3188950.

GRASS Development Team, 2015. Geographic Resources Analysis Support System (GRASS) Software, Version 6.4. Open Source Geospatial Foundation. grass.osgeo.org (1 June 2017).

GRASS Development Team, 2017. Geographic Resources Analysis Support System (GRASS) Software. Open Source Geospatial Foundation. grass.osgeo.org (20 September 2017).

Lennert, M., GRASS Development Team, 2017. Addon i.segment.stats. Geographic Resources Analysis Support System (GRASS) Software, Version 7.2, Open Source Geospatial Foundation. grass.osgeo.org/grass7/manuals/addons/i.segment.stats (1 June 2017).

Maas, A., Rottensteiner, F., Heipke, C., 2017. Classification under label noise using outdated maps. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, IV-1/W1, 215-222. doi.org/10.5194/isprs-annals-IV-1-W1-215-2017.

Michalis, P., Dowman, I., 2008: A Generic Model for Along-Track Stereo Sensors Using Rigorous Orbit Mechanics. *Photogrammetric Engineering & Remote Sensing* 74(3), 303-309.

Smith, J., 1987a. Close range photogrammetry for analyzing distressed trees. *Photogrammetria*, 42(1), 47-56.

Smith, J., 1987b. Economic printing of color orthophotos. Report KRL-01234, Kennedy Research Laboratories, Arlington, VA, USA.

Smith, J., 2000. Remote sensing to predict volcano outbursts. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XXVII-B1, 456-469.