Automated Pipelines for Large-Scale Height-Based Vegetation Segmentation

Oleksandr Tsaryniuk[1], *, Andriy Hlybovets[2], Oleksiy Oletsky[3]

[1,2] *National University of "Kyiv-Mohyla Academy",[2] Skovorody St., Kyiv, 04655, Ukraine*

Abstract

Height-based vector vegetation segmentation is one of the critical aspects of spatial analysis. This segmented data is used in radio propagation modeling, environmental monitoring, and vegetation mapping. Many studies on vector vegetation segmentation focus on delineating individual tree crowns, allowing detailed data sets to be obtained. However, the high level of detail results in a substantial data volume, making it impractical to use these datasets over large areas, such as an entire country. Segmentation of large vector data sets remains a significant challenge in geospatial data creation. In our study, we developed three different segmentation pipelines: hexagon segmentation, convolution segmentation, and random points. A test data fragment was processed to compare the proposed methods and accuracy and volume metrics were calculated.

Keywords[1]

Vegetation segmentation, spatial analysis, hexagonal grid, random points, convolution filters

1. Introduction

Integrating diverse datasets is a pivotal challenge in geospatial data production, particularly in vegetation analysis, where combining vector-based vegetation cover with Canopy Height Models (CHM) is essential for depth-enhanced segmentation. This study tackles such integration, aiming to segment vegetation based on height — a crucial step for comprehensive environmental and geographical analyses. Through the lens of satellite and aerial imagery, vegetation segmentation unlocks insights into vegetation distribution, health, and variety across vast areas. We introduce and assess three segmentation approaches: Hexagon Segmentation, Convolution Segmentation, and Random Points prioritizing their applicability to large-scale datasets, potentially encompassing entire countries. This comparative evaluation showcases the method's precision and practicality and advances our methodological toolkit for environmental studies.

2. Literature review

Image segmentation is one of the most challenging tasks in image processing. Currently, there are numerous approaches and methods for image segmentation, such as the hexagon segmentation method Hofmann & Tiede [1] and the Point Initialization Approach Mueller & Corcoran [2]. Most of the research in vegetation segmentation has focused on identifying individual tree crowns. This direction has been instrumental in detailed studies of forest ecosystems, as exemplified by the works of Douss et al. [3], Li et al. [4], Lindberg et al. [5], and Jakubowski et al. [6]. These studies have significantly advanced our understanding of individual tree characteristics, forest structure, and biomass distribution.

In contrast to the detailed focus on individual tree crowns, our research aims to develop methods for generalized segmentation that represent large arrays of vegetation with similar (or nearly identical) heights. These approaches are well-suited for segmenting vegetation over vast areas, such as entire

14th International Scientific and Practical Conference from Programming UkrPROG'2024, May 14-15, 2024, Kyiv, Ukraine

* Corresponding author.

† These authors contributed equally.

o.tsaryniuk@ukma.edu.ua (O. Tsaryniuk); a.glybovets@ukma.edu.ua (A. Hlybovets); oletsky@ukma.edu.ua (O. Oletsky)

0000-0003-1394-2040 (O. Tsaryniuk); 0000-0003-4282-481X (A. Hlybovets); 0000-0002-0553-5915 (O. Oletsky)

countries, addressing the need for macro-level vegetation analysis. Such analysis is essential for regional and national environmental assessments, land use planning, and large-scale conservation efforts.

Our study on vegetation segmentation will leverage CHM data with a 10-meter resolution, as developed by Liu et al. [7]. This CHM data is crucial for our methodology as it provides a detailed representation of vegetation height across large areas. Using a 10-meter resolution matrix allows for a fine-grained analysis of vegetation structure, making it manageable for large-scale applications like country-wide segmentation.

3. Methodology

We developed three distinct automated pipelines to address the challenge of segmenting vegetation based on height. We aimed to understand the complexity of accurately determining vegetation at different altitudes on large datasets. The methods described in the article were developed using FME (Feature Manipulation Engine). FME is a data integration platform developed by Safe Software. It is widely used for transforming, integrating, and automating spatial data workflows. A series of specific metrics were selected to assess the effectiveness and appropriateness of these approaches. These metrics serve as a foundation for evaluating each method's performance, ensuring a balanced analysis between the innovative aspects of our methodologies and their practical outcomes.

The following metrics were used for comparison:

Accuracy (1). This is the ratio of correctly identified pixels, TruePixels (2) to the total number of pixels. It is a straightforward measure of how accurately a model classifies or segments pixels.

$$Accuracy = \frac{TruePixels}{TotalNumberofPixels} \quad (1)$$

Where: *Total Number of Pixels* is the sum of all pixels within all vegetation segments.

$$TruePixels = \sum_{i=0}^{n} \left(\left| h_{input}(p) - h_{output}(p) \right| \leq 3 \right) \quad (2)$$

Where: $h_{input}(p)$ is the height associated with pixel p in the input data, $h_{output}(p)$ is the height associated with pixel p in the output data, as determined by the segmentation process.

Volume. This metric is expressed in the number of vertices after segmentation. It reflects the segmentation's complexity and detail. A more significant number of vertices usually implies a more detailed segmentation but negatively affects the display speed and processing.

3.1. Hexagon segmentation

Hexagonal grids offer several advantages over square grids, primarily due to their low perimeter-to-area ratio, which reduces sampling bias related to edge effects. Unlike circles with the lowest ratio but cannot tessellate, hexagons can form a continuous grid while being the most circular-shaped polygon. This allows hexagonal grids to more naturally represent curves in data patterns compared to square grids. Additionally, points within hexagons are closer to the centroid than points within equal-area squares or triangles, making hexagons ideal for analyses involving connectivity or movement paths. Hexagons also reduce orientation bias and distortion over large areas, and finding neighbors is simpler due to the equidistant centroids of adjacent hexagons.

Figure 1: Hexagon segmentation workflow.

The Hexagon segmentation method (Figure 1) involves creating a hexagonal grid with uniform hexagons (each side is 100 meters long) and generalizing the height matrix to a 3-meter interval. The vegetation vector is clipped according to the hexagon grid to form segments. Heights from the height matrix are then assigned to each segment, with the most frequent height value in the segment being selected (using the MODE function). Adjacent segments with the same height are merged.

We used Pierre's Gauthier algorithm [8] to generate a hexagonal grid. This algorithm's core involves generating a grid of points that will serve as the centers of the hexagons. The primary parameter is *SIDE_LENGTH*, the length of a hexagon's side.

Two point grids are generated with the following parameters: the first grid is defined by hoffset(1) and *voffset*(2):

$$\text{offset} = \text{SIDE\_LENGTH} * 3 \quad (1)$$
$$\text{voffset} = \cos 30° * \text{SIDE\_LENGTH} * 2 \quad (2)$$

The second grid is a copy of the first grid with shifts applied to the x and y coordinates:

$$X_{\text{shift}} = \text{hoffset}/2$$
$$Y_{\text{shift}} = \text{voffset}/2$$

The last step of the algorithm involves creating circles at the generated points with a radius of SIDE_LENGTH and then simplifying these circles into six-sided polygons. The result is a grid of regular hexagons with a side length of SIDE_LENGTH.

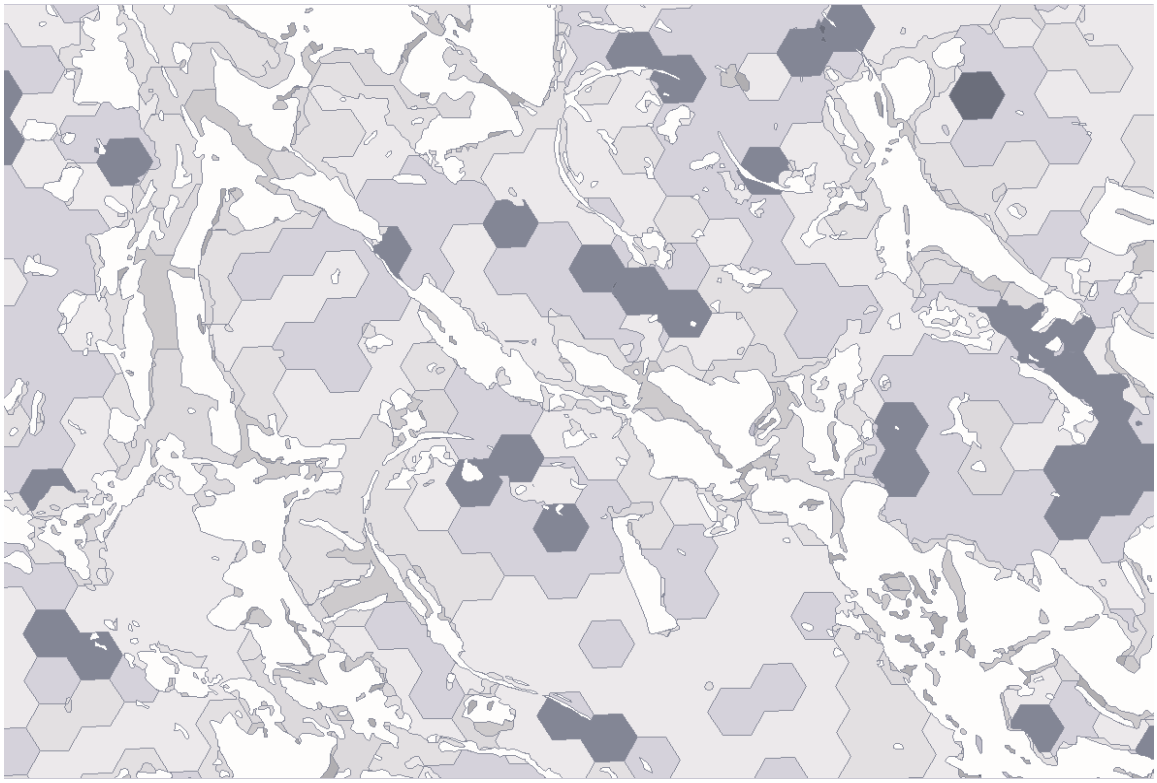

Figure 2: Result of the hexagon segmentation workflow.

3.2. Convolution Segmentation

The convolution function filters the pixel values in an image, which can be used for sharpening, blurring, edge detection, or other kernel-based enhancements. Filters enhance raster image quality by removing spurious data or highlighting features. These convolution filters are applied with a moving, overlapping kernel (window or neighborhood). They calculate pixel values based on the weights of neighboring pixels. In our approach, we used several iterations of convolutional filters to obtain areas with the same height.

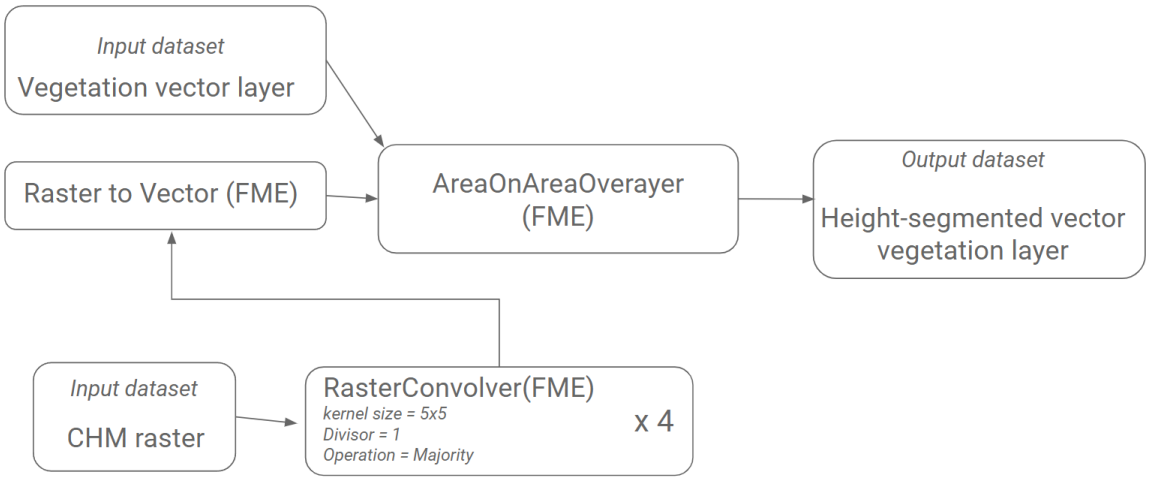

Figure 3: Convolution segmentation workflow.

Like the first method, the Convolution Segmentation method also generalizes the height matrix to a 3-meter interval. The matrix is then generalized using a convolutional filter (kernel = 5x5). Several iterations are conducted using the "Majority" operation (4 iterations), selecting the most frequently occurring value, as in the first method. The next stage is converting the raster to a vector. RasterToPolygonCoercer(FME) and AreaGapAndOverlapCleaner(FME) are used. To make a better shape of polygons after conversion, we used a combination of generalization and smoothing: Douglas–Peucker(Generalize 7 meters)[9] → NURBfit(Smooth) [10] → Douglas–Peucker(Generalize

2 meters). Such a combination of generalization and smoothing allows for eliminating pixel steps and obtaining an acceptable density of polygon vertices. The final stage combines the resulting polygons and vector vegetation layer by AreaOnAreaOverlayer(FME).

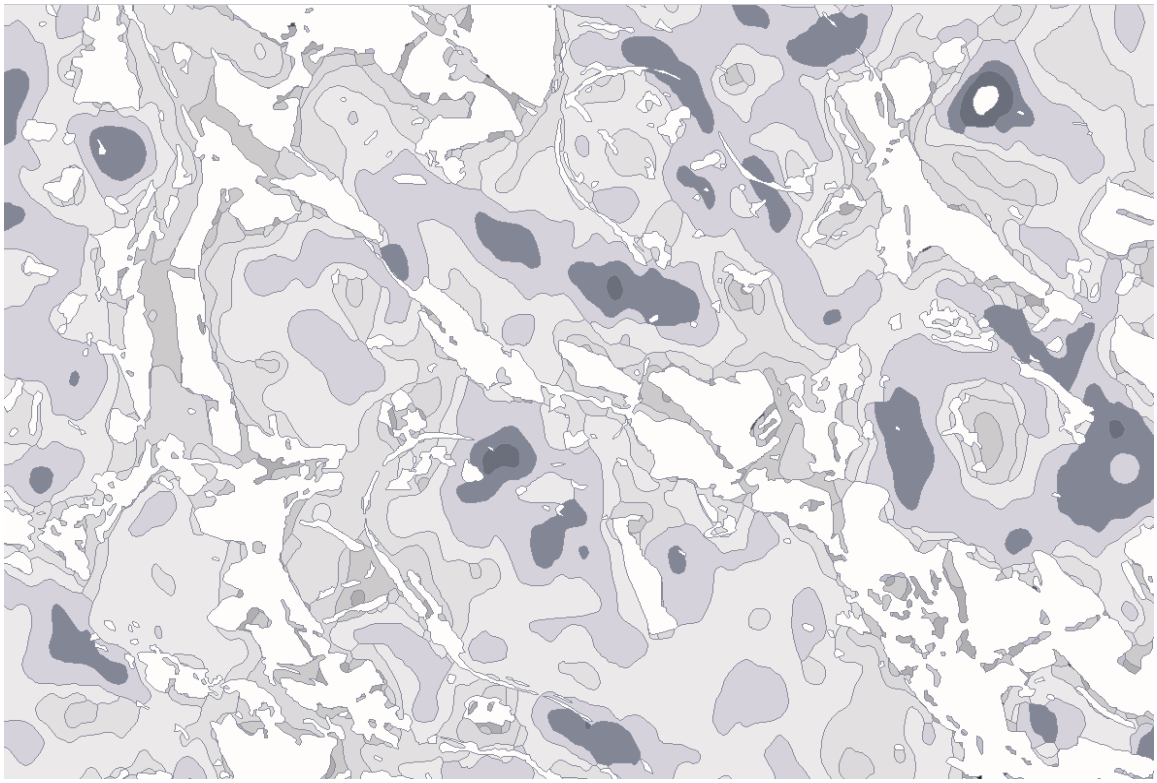

Figure 4: Result of the convolution segmentation workflow.

3.3. Random Points

The Random Points method is based on creating random points within a vegetation polygon using several steps: 1) Generation of random points across the polygon's bounding box. Two sets of coordinates are generated (X, Y). The number of random coordinates depends on the polygon`s area: larger area – more coordinates; 2) Generating points along the central line of the polygon created by CenterLineReplacer(FME); 3) Extracting the central point of the polygon: CenterPointExtractor(FME). 4) Snapper(FME) and DuplicateFilter(FME) were applied to the resulting points to avoid duplicates or near-duplicate points. Different approaches are used to generate random points due to the different shapes of the input polygons. This combination of point-generation methods allows us to get uniform point distribution over all polygons. The next step involves using the ArcGIS procedure 'Generate Subset Polygons'[11] activated by Python script. This function creates a subset of polygon features from input points without gaps and overlaps. The goal is to divide the points into compact, nonoverlapping subsets, and create polygon regions around each subset of points. The minimum and maximum number of points in each subset can be controlled.

Generate Subset Polygons' function based on Thiessen polygons also known as Voronoi diagram or Voronoi polygons. [12, 13].

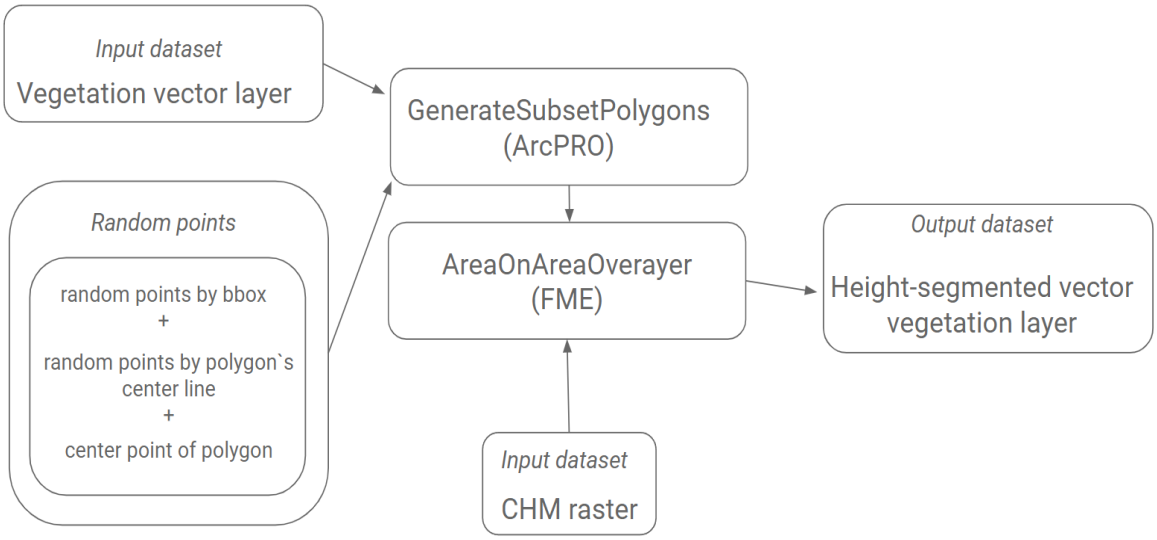

Figure 5: Random points segmentation workflow.

The methodology for assigning elevations to segments follows the approach established in previous methods. Each segment intersects with a generalized elevation matrix up to 3 meters. The elevation assigned to each segment is determined by the most frequently occurring pixel values within that intersection. This technique ensures consistency in elevation assignment across different segments, leveraging the established practices from prior methodologies for effective elevation mapping.

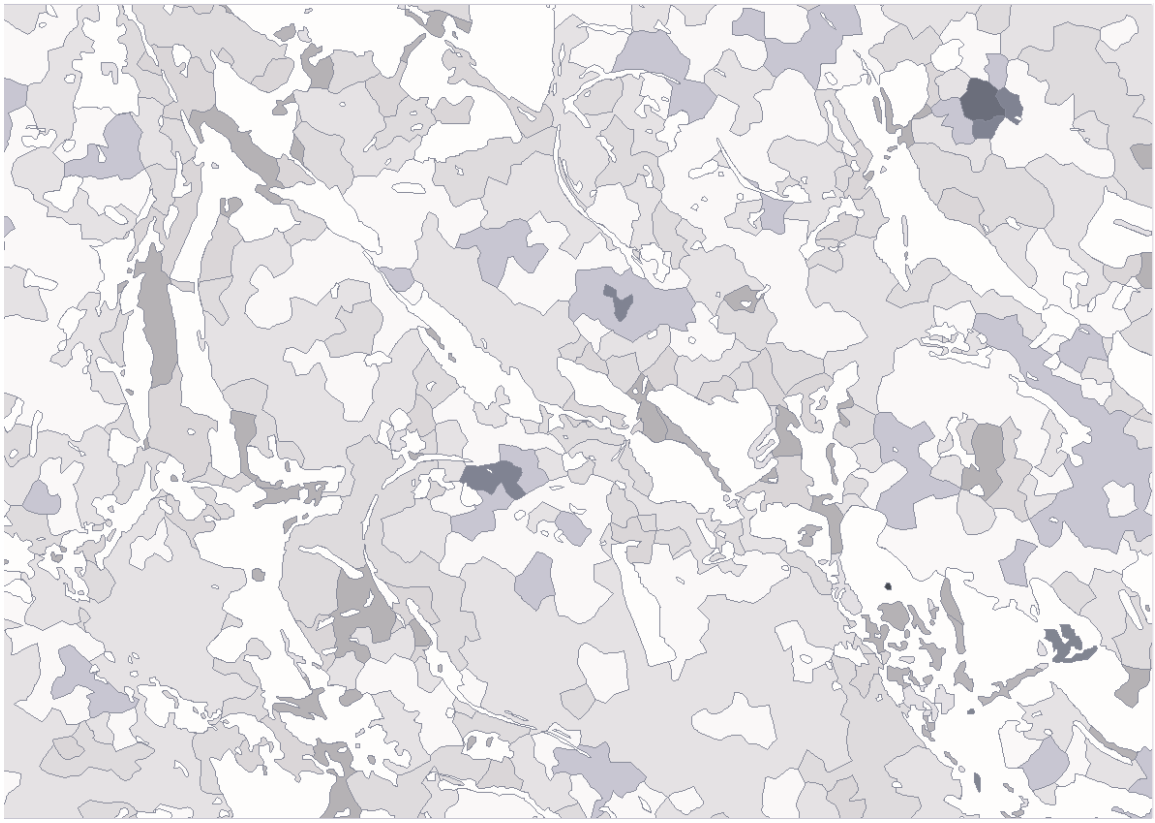

Figure 6: Result of the convolution segmentation workflow.

4. Evaluation of the quality of the proposed approaches

For this study, a test site covering an area of 430 square kilometers in the western Czech Republic was selected as the primary focus (Figure 7). The data concerning vegetation heights was sourced

from a detailed 10-meter CHM. The vegetation data itself was derived from a comprehensive vector dataset. This dataset was generated through machine learning techniques to automatically analyze high-resolution satellite imagery, a process meticulously carried out by the Visicom company.

Figure 7: Research area location.

The methods discussed in this article, as well as the analysis of the results, were implemented on PC using the Feature Manipulation Engine (FME). The obtained Accuracy and Volume results are shown in Tables 1,2,3.

Table 1

Hexagon method statistics

Vegetation Height	Accuracy %	Total pixels in CHM	Volume
0	66.22	980	558338
3	58.92	1020	
6	87.51	4485	
9	94.05	28114	
12	93.19	80631	
15	89.95	145203	
18	82.01	219782	
21	80.75	343390	
24	82.07	512259	
27	85.62	749204	
30	88.73	905916	
33	90.17	517650	
36	90.29	94701	
39	84.83	3723	

Table 2

Convolution method statistics

Vegetation Height	Accuracy %	Total pixels in CHM	Volume
0	76.29	949	752412
3	58.6	1256	
6	69.77	8657	
9	79.43	44705	
12	87.72	98143	
15	92.41	156412	
18	95.2	238215	
21	96.38	360859	
24	97.26	534555	
27	98.17	741339	
30	98.92	836759	
33	99.39	485668	
36	99.65	94605	
39	99.73	5176	

Table 3

Random point method statistics

Vegetation Height	Accuracy %	Total pixels in CHM	Volume
0	65.57	909	737853
3	58.26	1567	
6	83.38	6361	
9	88.01	41188	
12	88.55	87758	
15	82.67	141607	
18	82.01	213139	
21	80.75	360787	
24	82.07	542611	
27	85.62	780707	
30	88.73	905794	
33	90.17	516782	
36	90.29	93784	
39	84.83	4905	
42	86.36	374	

To evaluate the segmentation's accuracy, 3-meter height ranges were selected. After testing various height range options (1m, 3m, and 5m), the 3-meter range was chosen as optimal. This selection was based on its ability to accurately reflect the vegetation's true height while minimizing the amount of "noise" from individual pixels with varying heights. This compromise ensures a balance between precision and the reduction of outliers, providing a more reliable assessment of segmentation performance.

It is worth noting that in some methods, the 42-meter height category is not represented on the histogram. This is due to the very small number of pixels in this category. The most representative heights are those between 12 and 36 meters, with a sufficient number of pixels.

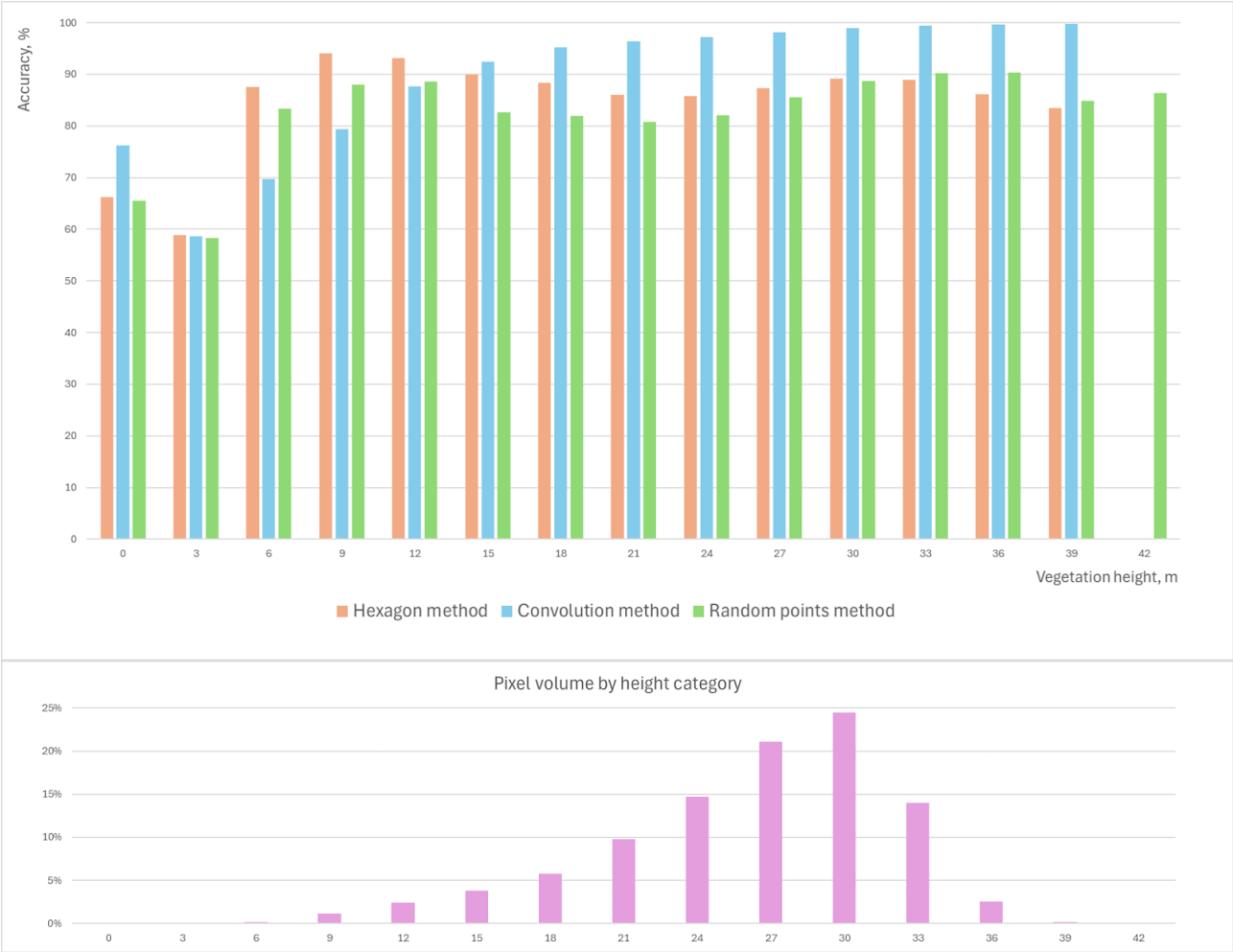

Figure 8: Comparative assessment of accuracy by height categories.

We did not consider the performance evaluation of the segmentation methods within the scope of this study. This decision was based on the understanding that performance assessments conducted on a limited test dataset would not yield representative results.

5. Conclusion

The comparative analysis reveals that each method has its merits in terms of accuracy and volume of the final segmented vector. The choice of method may depend on specific research needs, available computational resources, and the scale of the analysis. Although the hexagon method has the lowest accuracy, it differs from the simplicity of the other presented methods and can be successfully applied to large data sets. The convolutional method has the highest accuracy in representing heights but has a "bottleneck" at the raster-to-vector conversion stage. This stage requires significant computing resources and can become an obstacle in processing extensive data.

Future work should focus on refining these methodologies, exploring their application in different ecological contexts, and integrating additional data sources to enhance the accuracy and utility of vegetation segmentation for environmental monitoring and management.

Considering the rapid development and high efficiency of machine learning methods, future development of this research aims to incorporate AI-based approaches alongside the methods already compared. This expansion will comprehensively evaluate traditional segmentation techniques against AI-powered models, potentially setting a new benchmark in vegetation segmentation methodologies.

Additionally, plans are underway to apply the described segmentation methods to large countrywide datasets. In this context, it would be prudent to analyze each method's performance speed and calculate the computational resources required for its implementation. This comprehensive evaluation will ensure the methods' scalability and efficiency when applied to extensive data sets.

6. Authorship Contribution Statement

A. Hlybovets, O. Oletsky: Selection of metrics and assessment of the complexity of the proposed algorithms.
O. Tsaryniuk: Development and implementation of segmentation pipelines.

7. References

[1] P. Hofmann, D. Tiede,Image segmentation based on hexagonal sampling grids, South□Eastern European Journal of Earth Observation and Geomatics 3, 2014 pp. 173-177.
[2] J.N. Mueller, J.N. Corcoran, A Random Point Initialization Approach to Image Segmentation with Variational Level-sets. 2021, http://arxiv.org/abs/2112.12355.
[3] R. Douss, I.R Farah, Extraction of individual trees based on Canopy Height Model to monitor the state of the forest. Trees, Forests and People 8, 2022, doi: 10.1016/j.tfp.2022.100257.
[4] W. Li, Z. Niu, S. Gao, N. Huang, H. Chen, Correlating the horizontal and vertical distribution of LiDAR point clouds with components of biomass in a Picea crassifolia forest. Forests 5(8), 2014, pp. 1910–1930. doi: 10.3390/f5081910.
[5] E. Lindberg, J. Holmgren, H. Olsson, Classification of tree species classes in a hemi-boreal forest from multispectral airborne laser scanning data using a mini raster cell method. International Journal of Applied Earth Observation and Geoinformation 100, 2021, doi: 10.1016/j.jag.2021.102334.
[6] M.K. Jakubowski, W. Li, Q. Guo, M. Kelly, Delineating individual trees from lidar data: A comparison of vector- and raster-based segmentation approaches. Remote Sensing 5(9), 2013, pp. 4163–4186. doi: 10.3390/rs5094163.
[7] S. Liu, et al, The overlooked contribution of trees outside forests to tree cover and woody biomass across Europe. Science Advances 9(37), 2023, doi: 10.1126/sciadv.adh4097.
[8] FME Hub, HexagonSampler, https://hub.safe.com/publishers/larry/transformers/hexagonsampler
[9] D. Douglas, T. Peucker, Algorithms for the reduction of the number of points required to represent a digitized line or its caricature, Volume 10 Issue 2, 1973, pp. 112-122, doi: 10.3138/FM57-6770-U75U-7727
[10] K. Versprille, Computer-aided design applications of the rational b-spline approximation form. Ph.D. Dissertation, 1975, Syracuse University, USA.
[11] ArcGIS Tool Reference, Generate Subset Polygons (Geostatistical Analyst), https://pro.arcgis.com/en/pro-app/latest/tool-reference/geostatistical-analyst/generate-subset-polygons.htm
[12] Voronoï, Georges (1908a). "Nouvelles applications des paramètres continus à la théorie des formes quadratiques. Premier mémoire. Sur quelques propriétés des formes quadratiques positives parfaites" (PDF). Journal für die Reine und Angewandte Mathematik. 1908 (133): 97–178. doi:10.1515/crll.1908.133.97. S2CID 116775758.
[13] Voronoï, Georges (1908b). "Nouvelles applications des paramètres continus à la théorie des formes quadratiques. Deuxième mémoire. Recherches sur les parallélloèdres primitifs" (PDF). Journal für die Reine und Angewandte Mathematik. 1908 (134): 198–287. doi:10.1515/crll.1908.134.198. S2CID 118441072.