Behavioural Analysis of Segmentation Accuracy Metrics on Synthetic Urban Objects

Emin Ozgur Avsar [1], Ozgun Akcay [1], Ahmet Batuhan Polat [1]

[1] COMU, Faculty of Engineering, Department of Geomatics Engineering, Çanakkale, Türkiye
ozguravsar@comu.edu.tr, akcay@comu.edu.tr, abpolat@comu.edu.tr

Keywords: Object-based image analysis, Segmentation accuracy, Synthetic dataset, Area-based metrics, Location-based metrics.

Abstract

This study investigates the behaviour of segmentation accuracy metrics in object-based image analysis (OBIA) using synthetic urban objects. Monitoring land cover through high-resolution imagery relies heavily on accurate segmentation, which directly influences classification performance. A synthetic dataset was created with a fixed-size reference square and varying-sized square segments positioned systematically to analyse spatial and geometric relationships. Six widely used accuracy metrics were evaluated: Area Fit Index (AFI), Match (M), Quality Rate (QR), Over-Segmentation (OS), Under-Segmentation (US), and Quality of Object Location (qLoc), representing both area-based and location-based criteria.
Results reveal that area-based metrics generally show consistent trends and similar sensitivity to changes in segment size and geometry, while location-based metrics exhibit independent patterns emphasizing spatial positioning and locational accuracy. This divergence highlights the limitations of relying solely on either metric type, advocating for an integrated evaluation framework combining both area and location criteria to achieve a more comprehensive assessment of segmentation quality. The study suggests that future research should incorporate more complex and irregular urban object shapes and explore additional metrics, such as boundary-based or context-aware measures. Furthermore, the identification of optimal segmentation configurations guided by these metrics could enhance training data quality for deep learning applications in urban object classification.

1. Introduction

Monitoring land cover is a fundamental requirement for the sustainable management of natural resources, environmental conservation, and the assessment of climate change impacts. Remote sensing techniques provide a robust framework for detecting and analysing land cover dynamics. The literature highlights that object-based image analysis (OBIA) significantly enhances classification accuracy for high-resolution imagery by addressing the inherent limitations of pixel-based techniques. OBIA interprets groups of neighbouring pixels -called segments- as visually meaningful objects, rather than analysing individual pixels. The segmentation process, which divides imagery into homogeneous regions, is a crucial pre-processing step in OBIA. The accuracy of subsequent classification depends on both the segmentation method and the selection of optimal parameters, which are influenced by data quality.

Segmentation methods are typically classified as edge-based, area-based, or threshold-based. Tools such as the Estimation of Scale Parameter (ESP) toolbox (Dragut et al., 2010; 2014) assist in identifying appropriate scale parameters for multi-resolution segmentation, but do not directly evaluate segmentation accuracy, which remains an ongoing challenge in OBIA workflows. Segmentation accuracy can be assessed using qualitative (visual interpretation) or quantitative (metric-based) approaches (Kotaridis and Lazaridou, 2021). Quantitative evaluations involve comparing classification accuracies or measuring geometric mismatches between segments and reference objects (Zhang et al., 2015). These mismatches are typically assessed using area-based or location-based metrics, focusing on overlap or spatial proximity (Clinton et al., 2010).

Urban features, especially buildings, are among the most frequently studied object classes due to their regular geometries and relevance in land use classification (Zhang et al., 2023; Vasavi et al., 2023). Moreover, their clearly defined shapes make them particularly suitable for evaluating segmentation accuracy (Akcay et al., 2022; Xu et al., 2023). For example, Jozdani and Chen (2020) first analysed eight regularly shaped buildings of varying sizes to identify patterns in segmentation metrics and then validated their findings on a randomly selected sample of 100 buildings. Their study revealed discrepancies among commonly used evaluation criteria, highlighting the need for further investigation. Simões et al. (2023), on the other hand, developed an R package called *segmetric* to enable the analysis of the metrics proposed and applied in segmentation accuracy studies.

The presented study constructs a synthetic dataset to represent the building class and systematically evaluates segmentation accuracy using six widely cited metrics: Area Fit Index (AFI), Match (M), Quality Rate (QR), Over-Segmentation (OS), Under-Segmentation (US), and Quality of Object Location (qLoc). By comparing the areal and spatial relationships between segments (represented as squares of varying sizes) and a fixed-size reference object, the study aims to clarify the behaviour of area-based and location-based metrics under controlled conditions. The results show that area-based metrics generally produce consistent values, while location-based ones display independent patterns, offering new insights into metric selection in OBIA validation.

2. Synthetic Dataset Design for Segmentation Metric Evaluation

The synthetic dataset consists of a fixed reference sample represented by a square with an edge length of fifty units. The segments to be compared are also squares, but with varying edge lengths ranging from ten to one hundred units, increasing in increments of ten units, and positioned differently relative to the sample. The dataset was generated within a two-dimensional Cartesian coordinate system, where the reference sample square was fixed at the origin for consistency. Segments were positioned systematically by defining their centroid coordinates relative to the sample to ensure precise control over spatial relationships. To investigate the impact of area differences between the sample and segments on the evaluation criteria, each segment set contains squares of uniform edge length distinct from other sets. Within

each segment set, the segment was initially positioned to create a fixed intersection area of one hundred square units with the sample. This intersection area was computed analytically using geometric formulas for the overlap of two axis-aligned squares. Following this initial positioning, the segment was incrementally shifted in ten-unit steps upward and to the right (along the positive y- and x-axes), creating multiple spatial configurations. This systematic repositioning was designed to reveal how changes in intersection and adjacency areas, as well as centroid displacements, influence the segmentation accuracy metrics.

For segment sets with edge lengths varying from ten to one hundred units, the number of positional shifts ranged from five to fourteen, depending on the segment size and the ten-unit increment steps. This resulted in a total of 985 sample-segment intersection configurations, calculated as the sum of all movement steps across segment sizes. The intersection pattern observed for 50-unit segments is also seen in the 10-, 30-, 70-, and 90-unit segments, while the pattern for 60-unit segments resembles those of the 20-, 40-, 80-, and 100-unit segments.

Given the symmetrical properties of the evaluation criteria employed in this study, specifically those related to area overlap and spatial proximity, redundant segment positions were identified and removed to optimize the dataset. For instance, rows and columns that were symmetrical with respect to the dataset matrix exhibited identical metric values and were therefore excluded. Figure 1 illustrates this data reduction approach, where intersections highlighted in blue and green represent redundant entries that were removed. Furthermore, diagonal symmetry allowed for additional pruning of the dataset, leaving only the unique intersection configurations within the magenta region necessary for evaluation, while excluding those in the red region.

The entire dataset, including segment coordinates and computed intersection areas, was stored in matrix form to facilitate efficient computational processing. It is important to note that while the use of square shapes simplifies geometric calculations and allows precise control of spatial parameters, it also introduces limitations. Real urban objects, such as buildings, often exhibit irregular shapes and orientations, and environmental factors affecting segmentation accuracy were not modelled in this synthetic dataset.

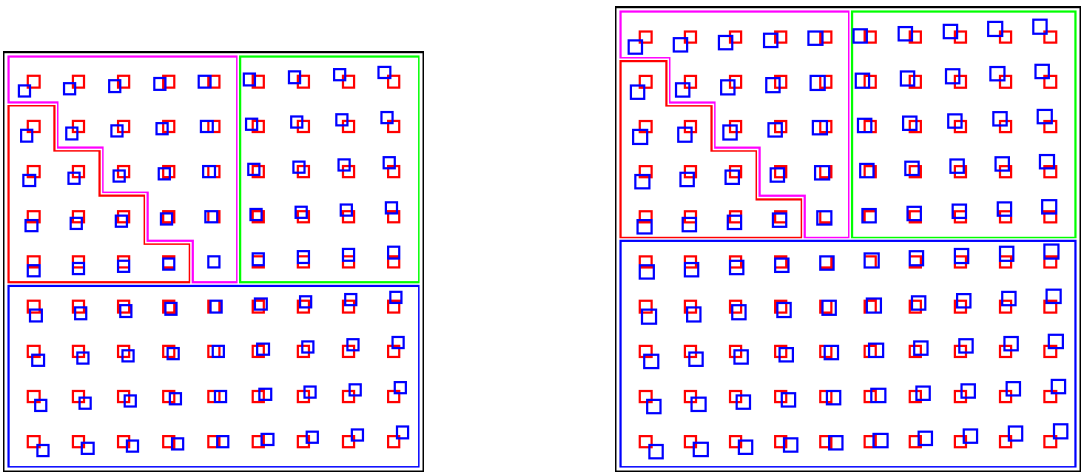

Figure 1. Visualization of sample–segment intersections in the synthetic dataset and reduction approach (left: segments with edge lengths of 50 units, right: segments with edge lengths of 60 units).

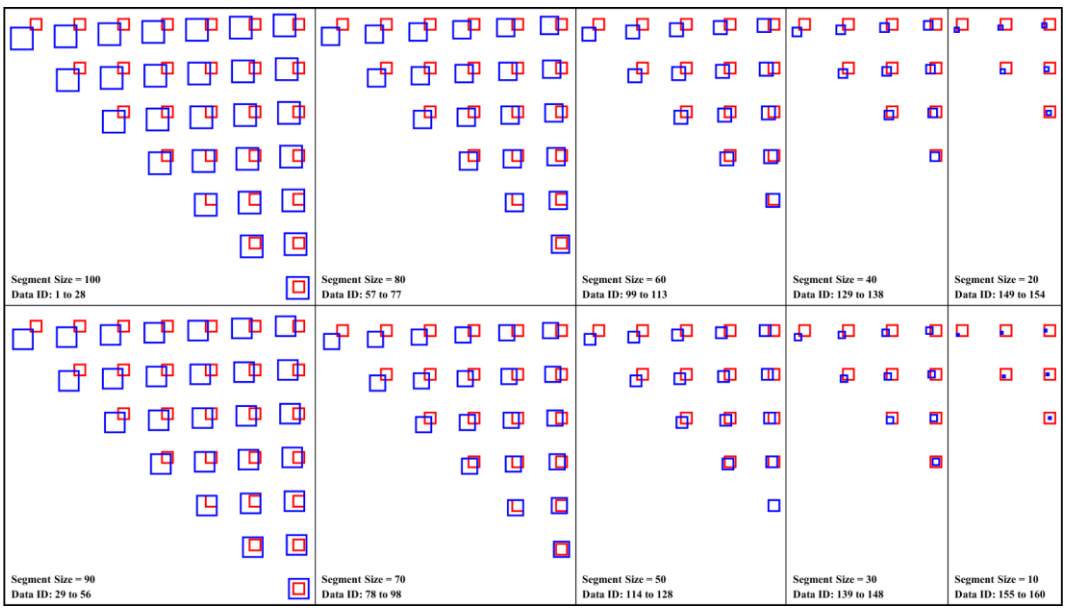

Figure 2. Synthetic segment sets categorized by size; segment IDs progress from left to right and top to bottom within each panel, starting at Ss = 100 (ID 1 - 28) and continuing sequentially down to Ss = 10 (ID 155 - 160).

As a result, in the synthetic dataset, segments with edge lengths of 100 and 90 units have 28 intersection positions each; those with edge lengths of 80 and 70 units have 21 intersections; segments of 60 and 50 units have 15 intersections; 40 and 30 unit segments have 10 intersections; and segments of 20 and 10 units have 6 intersections each. Figure 2 illustrates the total 160 segment sets organized by segment size, with their corresponding Data IDs start from the top-left corner of a segment set with 100-unit edge length and progresses in the horizontal direction (left to right). Upon reaching the end of each row, the numbering continues from the leftmost cell of the next row below. Once all segments in a given set are numbered, the same spatial pattern is maintained, and ID assignment continues from where it left off in the subsequent segment set.

3. Segmentation Accuracy Metrics

Research on segmentation accuracy criteria began in the late 1990s, leading to the proposal of various evaluation metrics by numerous researchers. Costa et al. (2018) provide a comprehensive review of these criteria, categorizing them into area-based and location-based metrics. Their review also includes combined geometric metrics. Another significant contribution is by Clinton et al. (2010), who analysed segmentation accuracy criteria by applying them to an RGB aerial image segmented with different parameter combinations of scale, shape, and compactness.

In the present study, several widely used area-based criteria - Area Fit Index (AFI), Match (M), Quality Rate (QR), Over Segmentation (OS), and Under Segmentation (US) - along with the location-based criterion, quality of the object's location (qLoc), are examined. For describing the criteria metrics; x_i denotes the sample where $X = \{x_i : i = 1 : n\}$ is the set of n training objects, while y_j denotes the segment where $Y = \{y_j : j = 1 : m\}$ is the set of m segments intersects with sample x_i.

The AFI (1) criterion, introduced by Lucieer and Stein (2002), measures the percentage of the intersection between a segment and the reference sample with which it overlaps the most. In the AFI metric, the optimal value, representing a perfect overlap between the sample and segment areas, is zero. The maximum possible value depends on the relative sizes of the sample and segment areas and can be as high as one, while the minimum value can approach negative infinity. A significant drawback of the AFI criterion is that it assigns the same value to segments of identical size regardless of the actual overlap with the sample, potentially overlooking differences in intersection quality.

$$AFI_{ij} = \frac{area(x_i) - area(y_j)}{area(x_i)} \tag{1}$$

Janssen and Molenaar (1995) proposed the M (2) criterion, which considers the relationship among the intersection area, sample area, and segment area. M is calculated by dividing the square of the intersecting area by the product of the sample and segment areas. A value of 1 indicates the best possible match between segment and sample, whereas 0 represents no overlap, corresponding to the worst-case scenario (Feitosa et al., 2010). Unlike other criteria, since the optimal value for M is 1, it was transformed to 1 – M for standardization during evaluation. Although a larger intersection area generally improves the M value, a smaller segment area can also lead to equal or better M values.

$$M_{ij} = 1 - \sqrt{\frac{area(x_i \cap y_j)^2}{area(x_i) * area(y_j)}} \tag{2}$$

The QR (3) criterion, suggested by Weidner (2008), is the ratio of the intersection area to the combined area of the sample and segment (including their junction). QR's advantage over other area-based criteria lies in accounting for both correctly intersecting areas and total junction areas (Clinton et al., 2010). Its values range from 0 to 1, with 0 being optimal. Since the junction area is smaller when segment areas are small, QR may yield better results even if the intersection area is reduced.

$$QR_{ij} = 1 - \frac{area(x_i \cap y_j)}{area(x_i \cup y_j)} \tag{3}$$

Persello and Bruzzone (2010) introduced the OS (4) and US (5) criteria, which represent the relations between the intersection area and the sample and segment areas, respectively. Under segmentation occurs when neighbouring pixels belonging to different classes are incorrectly grouped into a single object, while over segmentation happens when pixels that should belong to one object are split into multiple objects (Costa et al., 2018). These criteria are graded similarly to QR. However, when considered separately, OS and US may overlook the respective areas of segment and sample, potentially resulting in contradictory values; for example, a low US coupled with a high OS. Consequently, Weidner (2008) and Levine and Nazif (1982) recommended combining these measures, and their combined metric, Over-Under Segmentation (OUS) (6), is discussed in this study.

$$OS_{ij} = 1 - \frac{area(x_i \cap y_j)}{area(x_i)} \tag{4}$$

$$US_{ij} = 1 - \frac{area(x_i \cap y_j)}{area(y_j)} \tag{5}$$

$$OUS_{ij} = \sqrt{\frac{(OS_{ij})^2 + (US_{ij})^2}{2}} \tag{6}$$

Although the criteria discussed so far evaluate sample and/or segment areas or intersection and/or junction areas, they do not account for positional differences between the sample and segment. To address this, the location-based qLoc (7) criterion proposed by Zhan et al. (2005) was also evaluated. The qLoc criterion calculates the Euclidean distance between the centroids of the sample and segment. As the distance between centroids increases, the qLoc value increases, negatively impacting segmentation accuracy. The optimal qLoc value is 0 (indicating perfect centroid overlap), while the maximum value is unbounded, approaching infinity. This wide range necessitates normalization of qLoc values. Accordingly, the Relative Position (RP) (8) normalization method recommended by Möller et al. (2007) was applied. RP is computed by dividing the centroid distance of a sample and an intersecting segment by the maximum qLoc value among all segments intersecting that sample. Given that segment sizes are generally uniform; normalization was applied within each segment set. However, since qLoc is solely distance-based, it may not fully capture the fit quality between sample and segment.

$$qLoc_{ij} = dist[centroid(x_i), centroid(y_j)] \tag{7}$$

$$RP_{ij} = \frac{qLoc_{ij}}{dist_{max}} \quad (8)$$

4. Evaluation of Segmentation Accuracy Metrics on Synthetic Datasets

The synthetic datasets were evaluated using the criteria whose metrics and limitations were previously defined, and the corresponding numerical results are summarized in Figure 3. The evaluations of the intersections between the sample objects and the 50- and 60-unit segments are presented in tabular form as examples in Appendix A. The analyses yield the following key insights:

1. The Area Fit Index (AFI) criterion exhibits invariant values within each synthetic dataset.

2. Irrespective of the intersection (or junction) areas, discrepancies between sample and segment area sizes result in deviations of the AFI criterion from its optimal value.

3. When the intersection (or junction) areas of the sample and segment are equivalent:
   a. The metrics M, QR, and OUS display identical values within their respective categories.
   b. In contrast, the RP metric demonstrates variability.

4. As the intersection area between sample and segment increases (concomitant with a reduction in junction areas):
   a. M, QR, and OUS criteria values uniformly decrease, maintaining a consistent rank order.
   b. When the segment area is smaller than the sample area, RP values similarly decrease and conform to the order established by the other criteria.
   c. Conversely, if the segment area exceeds the sample area, RP values may remain constant or increase, thereby diverging in rank order relative to the other criteria. Specifically, within the following sample-segment intersections, despite increased intersection areas (which correspond to lower area-based criterion values), RP values are elevated:
      i. Ss100 dataset:
         1. $RP_9 > RP_5, RP_6, RP_7$
         2. $RP_{15} = RP_{12} > RP_{13}$
         3. $RP_{19} > RP_{17}, RP_{18}$
      ii. Ss90 dataset:
         1. $RP_{37} > RP_{33}, RP_{34}, RP_{35}$
         2. $RP_{47} > RP_{45}, RP_{46}$
      iii. Ss80 dataset:
         1. $RP_{64} > RP_{62}$

5. In cases where the sample is entirely enclosed within the segment (intersection area equals sample area and junction area equals segment area), or reciprocally, where the segment is fully enclosed within the sample:
   a. The M, QR, and OUS metrics yield consistent values within their groups.
   b. The RP metric may present variable values.

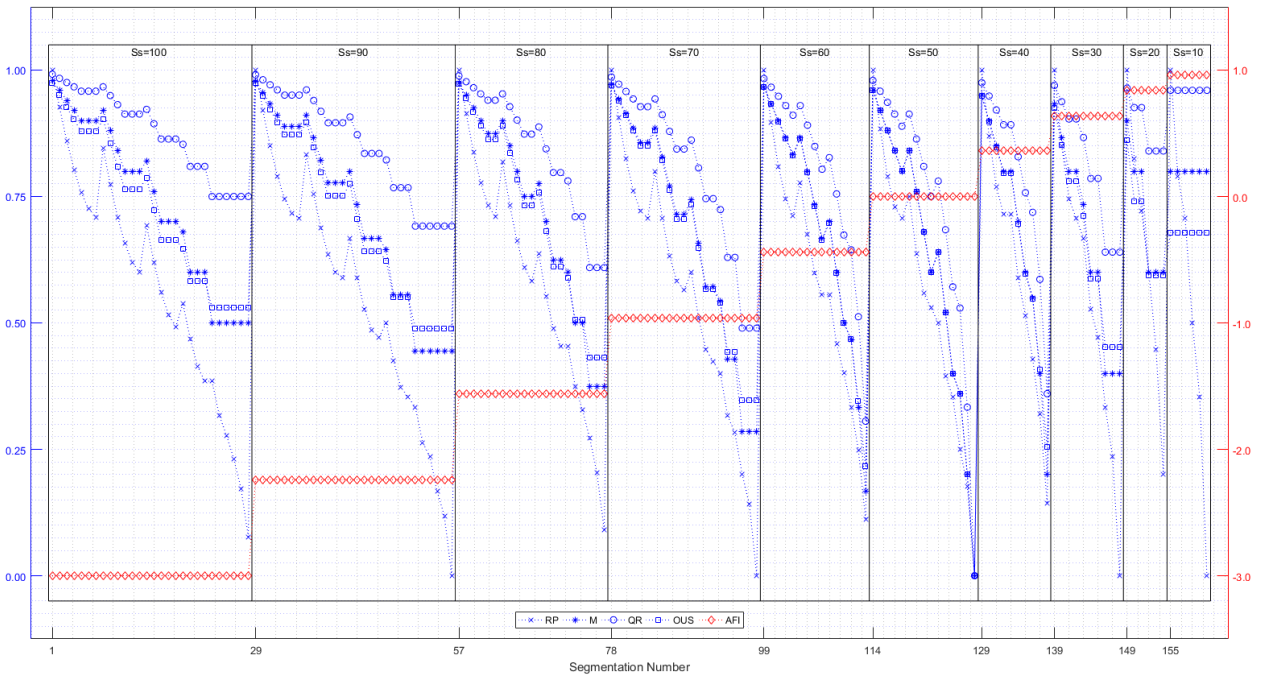

Figure 3. Segmentation accuracy criteria values of the synthetic datasets for a fixed sample size (50 units) and varying segment sizes (Ss = 100–10), with RP, QR, M, and OUS plotted on the left Y-axis and AFI on the right Y-axis.

The evaluations outlined above are exemplified in Figure 4. In cases where the intersection areas of the sample and segment are equivalent, whether the segment area is larger (ID: 44, 45, 46) or smaller (ID: 144, 145) than the sample, the area-based criteria remain constant within their group, while RP values vary slightly due to spatial alignment differences. This variation, however, has only a limited effect on the overall representativeness when the intersection ratio is fixed. A similar behaviour is observed when the segment is fully contained within the sample (ID: 146, 147, 148), where RP decreases gradually as the segment shifts within the sample, reflecting positional sensitivity but without substantially altering the representativeness relationship. By contrast, variations in RP become more pronounced when the sample is entirely enclosed within the segment (ID: 75, 76, 77), as the segment's dominance amplifies the influence of alignment on representativeness. Moreover, as the intersection between the sample and segment increases in cases where the segment area is smaller than the sample (ID: 142, 143, 144), RP values follow the

decreasing trend of the area-based criteria. Conversely, when the segment area exceeds the sample area (ID: 46, 47), RP values may increase despite the decline observed in the area-based criteria, highlighting the need to consider which type of metric should be prioritized when evaluating representativeness.

The area-based metrics—M, QR, and OUS—exhibit a consistent decreasing trend as intersection areas increase, maintaining their relative ranking and offering a stable characterization of geometric overlap. In contrast, the location-based RP metric occasionally diverges from this trend, showing constant or even elevated values under certain configurations. This behaviour reflects RP's sensitivity to spatial alignment, which complements the strictly area-focused nature of the other metrics. Notably, segments with identical area-based values can show varying RP scores, with some achieving the best alignment within their set. Conversely, segments sharing the same RP may differ in their area-based scores. This underscores the complementary roles of these metrics in evaluating segmentation performance. These patterns indicate that while area-based metrics reliably quantify the extent of overlap, RP captures differences in spatial configuration and object alignment—providing complementary insight into segmentation accuracy. To illustrate these behaviors, Appendix A presents a representative subset of 44 intersections selected to highlight extreme, median, and divergent scenarios across both area-based and location-based metrics.

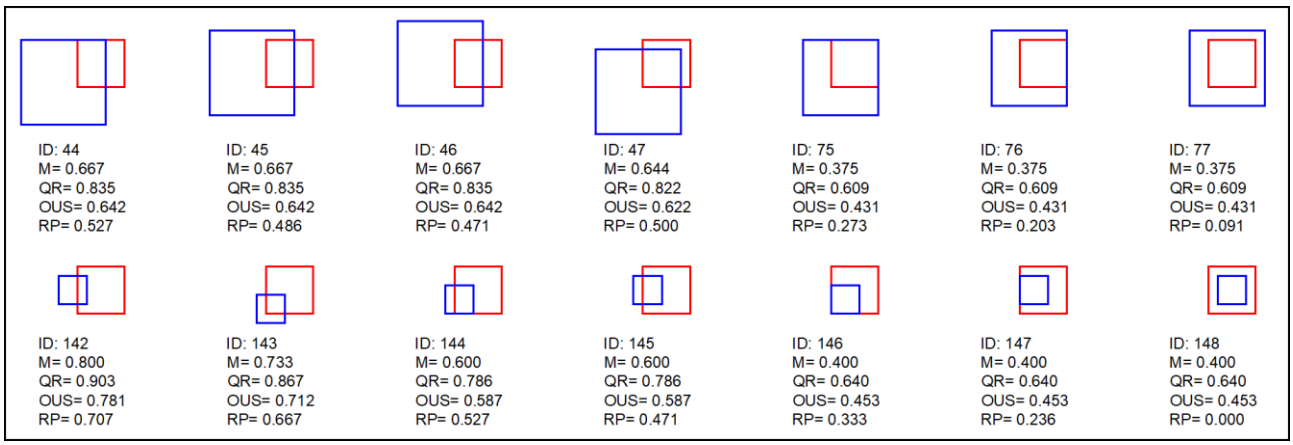

Figure 4. Comparison of M, QR, OUS, and RP metrics for varying intersection and segment sizes.

5. Conclusion and Future Work

The results of the segmentation accuracy assessment indicate that area-based metrics exhibit largely consistent trends across the synthetic datasets, demonstrating similar sensitivity to variations in segment geometry and scale. In contrast, location-based metrics display an independent behavioural pattern, highlighting a different dimension of segmentation quality—particularly spatial positioning and locational consistency. This divergence reveals the inherent limitations of relying solely on either metric type, suggesting that individual use may provide an incomplete assessment of segmentation quality. Specifically, metrics such as the Area Fit Index (AFI), despite their utility, can be insensitive to the degree of spatial congruence between segment and sample boundaries, potentially limiting their discriminative power in certain scenarios.

Therefore, it is critical to integrate both area-based and location-based criteria within a combined evaluation framework. While area-based metrics quantitatively capture segment–sample overlap, location-based metrics complement this by detecting positional deviations between objects. Such an integrated approach enables a multidimensional and more holistic analysis of segmentation performance, supporting more robust and reliable decision-making for optimal parameter selection and subsequent classification tasks.

This study employed a synthetic dataset with regular geometries to evaluate the behavioural differences among widely used segmentation accuracy metrics. While this controlled setup allows for isolating specific spatial and geometric factors, future studies can benefit from introducing more diverse and complex object forms. For example, datasets including rotated, L-shaped, or irregularly contoured building geometries would better reflect the variability of real urban environments. Such additions would enable further analysis of how different segmentation metrics respond to geometric complexity, orientation changes, or shape irregularities.

Another promising direction involves the inclusion of additional segmentation accuracy criteria. Beyond the six metrics evaluated here, future work could explore boundary-based, spectral–spatial, or context-aware measures that capture different dimensions of segmentation quality. For instance, metrics evaluating edge alignment or object topology may provide complementary insights, especially in high-resolution imagery where geometric details are critical. Incorporating such measures would enable a more holistic and multi-dimensional evaluation framework, potentially revealing metric sensitivities that are not apparent in area- or location-based analyses alone.

Furthermore, once optimal segmentation configurations are identified using appropriate accuracy metrics, the resulting high-quality segments, particularly for urban objects, can be utilized to generate training labels for deep learning applications. This metric-guided approach allows the creation of spatially consistent and semantically reliable labeled datasets with minimal manual effort. It also provides a methodological bridge between object-based image analysis and data-driven learning frameworks, improving training data quality and supporting more accurate classification outcomes.

References

Akcay, O., Kinaci, A. C., Avsar, E. O., Aydar, U., 2022: Semantic segmentation of high-resolution airborne images with dual-stream DeepLabV3+. *ISPRS International Journal of Geo-*

Information. 2022 Jan; 11(1):23. https://doi.org/10.3390/ijgi11010023.

Clinton, N., Holt, A., Scarborough, J., Yan, L. I., Gong, P., 2010: Accuracy assessment measures for object-based image segmentation goodness. *Photogrammetric Engineering & Remote Sensing*, 76, 289–299. https://doi.org/10.14358/PERS.76.3.289.

Costa, H., Foody, G. M., Boyd, D. S., 2018: Supervised methods of image segmentation accuracy assessment in land cover mapping. *Remote Sensing of Environment*, 205, 338-351. https://doi.org/10.1016/j.rse.2017.11.024.

Drăguț, L., Csillik, O., Eisank, C., Tiede, D., 2014: Automated parameterisation for multi-scale image segmentation on multiple layers. *ISPRS Journal of photogrammetry and Remote Sensing*, 88, 119-127. https://doi.org/10.1016/j.isprsjprs.2013.11.018.

Drăguț, L., Tiede, D., Levick, S. R., 2010: ESP: a tool to estimate scale parameter for multiresolution image segmentation of remotely sensed data. *International Journal of Geographical Information Science*, 24(6), 859-871. https://doi.org/10.1080/13658810903174803.

Feitosa, R. Q., Ferreira, R. S., Almeida, C. M., Camargo, F. F., Costa, G. A. O. P., 2010. Similarity metrics for genetic adaptation of segmentation parameters. In Geographic Object-Based Image Analysis (GEOBIA) 2010 conference held (Vol. 29).

Janssen, L.L., Molenaar, M., 1995: Terrain objects, their dynamics and their monitoring by the integration of GIS and remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 33(3), 749-758. https://doi.org/10.1109/36.387590.

Jozdani, S., Chen, D., 2020: On the versatility of popular and recently proposed supervised evaluation metrics for segmentation quality of remotely sensed images: An experimental case study of building extraction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 160, 275-290. https://doi.org/10.1016/j.isprsjprs.2020.01.002.

Kotaridis, I., Lazaridou, M., 2021: Remote sensing image segmentation advances: A meta-analysis. *ISPRS Journal of Photogrammetry and Remote Sensing*, 173, 309-322. https://doi.org/10.1016/j.isprsjprs.2021.01.020.

Levine, M. D., Nazif, A. M., 1982. An experimental rule based system for testing low level segmentation strategies, Multicomputers and Image Processing: Algorithms and Programs (K. Preston and L. Uhr, editors), New York: Academic Press, pp. 149–160.

Lucieer, A., Stein, A., 2002: Existential uncertainty of spatial objects segmented from satellite sensor imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 40(11), 2518-2521. https://doi.org/10.1109/TGRS.2002.805072.

Möller, M., L. Lymburner, M. Volk, 2007: The comparison index: A tool for assessing the accuracy of image segmentation, *International Journal of Applied Earth Observation and Geoinformation*, (9):311–321. https://doi.org/10.1016/j.jag.2006.10.002.

Persello, C., Bruzzone, L., 2010: A novel protocol for accuracy assessment in classification of very high resolution images. *IEEE Transactions on Geoscience and Remote Sensing*, 48, 1232–1244. https://doi.org/10.1109/IGARSS.2008.4778978.

Simões, M., Ferreira, M. P., & Pereira, M. N. (2023). The segmetric package: Metrics for assessing segmentation accuracy for geospatial data. R package. https://doi.org/10.32614/CRAN.package.segmetric.

Vasavi, S., Somagani, H. S., Sai, Y., 2023: Classification of buildings from VHR satellite images using ensemble of U-Net and ResNet. *The Egyptian Journal of Remote Sensing and Space Sciences*, 26(4), 937-953. https://doi.org/10.1016/j.ejrs.2023.11.008.

Weidner, U., 2008. Contribution to the assessment of segmentation quality for remote sensing applications. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 37, 479–484.

Xu, X., Zhang, H., Ran, Y., Tan, Z., 2023: High-Precision Segmentation of Buildings with Small Sample Sizes Based on Transfer Learning and Multi-Scale Fusion. *Remote Sensing*, 15(9), 2436. https://doi.org/10.3390/rs15092436.

Zhan, Q., Molenaar, M., Tempfli, K., Shi, W., 2005: Quality assessment for geo-spatial objects derived from remotely sensed data. *International Journal of Remote Sensing*, 26, 2953–2974. https://doi.org/10.1080/01431160500057764.

Zhang, H., Xu, C., Fan, Z., Li, W., Sun, K., Li, D., 2023: Detection and Classification of Buildings by Height from Single Urban High-Resolution Remote Sensing Images. *Applied Sciences*, 13(19), 10729. https://doi.org/10.3390/app131910729.

Zhang, X., Feng, X., Xiao, P., He, G., Zhu, L., 2015: Segmentation quality evaluation using region-based precision and recall measures for remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 102, 73-84. https://doi.org/10.1016/j.isprsjprs.2015.01.009.

Appendix A

ID	Ss	Jun. Area	Int. Area	AFI	M	QR	OUS	RP	Feature within the set
1	100	12400	100	-3.00	0.980	0.992	0.975	1.000	max RP
7	100	12000	500	-3.00	0.900	0.958	0.878	0.709	same RP, diff area metrics
10	100	11700	800	-3.00	0.840	0.932	0.809	0.709	
15	100	11300	1200	-3.00	0.760	0.894	0.723	0.620	area-based median-like
23	100	10000	2500	-3.00	0.500	0.750	0.530	0.385	equal the best in area-based metrics
28	100	10000	2500	-3.00	0.500	0.750	0.530	0.077	best
29	90	10500	100	-2.24	0.978	0.990	0.974	1.000	max RP
41	90	9600	1000	-2.24	0.778	0.896	0.751	0.589	same RP, diff area metrics
43	90	9400	1200	-2.24	0.733	0.872	0.706	0.589	
51	90	8100	2500	-2.24	0.444	0.691	0.489	0.333	equal the best in area-based metrics
56	90	8100	2500	-2.24	0.444	0.691	0.489	0.000	best, coincident centres
57	80	8800	100	-1.56	0.975	0.989	0.972	1.000	max RP
61	80	8400	500	-1.56	0.875	0.940	0.863	0.733	same RP, diff area metrics
64	80	8300	600	-1.56	0.850	0.928	0.836	0.733	
66	80	7900	1000	-1.56	0.750	0.873	0.732	0.610	area-based median-like
75	80	6400	2500	-1.56	0.375	0.609	0.431	0.273	equal the best in area-based metrics
77	80	6400	2500	-1.56	0.375	0.609	0.431	0.091	best
78	70	7300	100	-0.96	0.971	0.986	0.970	1.000	max RP
83	70	6900	500	-0.96	0.857	0.928	0.850	0.707	same RP, diff area metrics
85	70	6800	600	-0.96	0.829	0.912	0.821	0.707	
96	70	4900	2500	-0.96	0.286	0.490	0.346	0.200	equal the best in area-based metrics
98	70	4900	2500	-0.96	0.286	0.490	0.346	0.000	best, coincident centres
99	60	6000	100	-0.44	0.967	0.983	0.966	1.000	max RP
106	60	5300	800	-0.44	0.733	0.849	0.731	0.598	area-based median-like
107	60	5100	1000	-0.44	0.667	0.804	0.664	0.556	same RP, diff area metrics
108	60	5200	900	-0.44	0.700	0.827	0.697	0.556	
113	60	3600	2500	-0.44	0.167	0.306	0.216	0.111	best
114	50	4900	100	0.00	0.960	0.980	0.960	1.000	max RP
117	50	4600	400	0.00	0.840	0.913	0.840	0.729	same area metrics, diff RP
119	50	4600	400	0.00	0.840	0.913	0.840	0.750	
121	50	4200	800	0.00	0.680	0.810	0.680	0.559	median-like
128	50	2500	2500	0.00	0.000	0.000	0.000	0.000	perfectly coincident
129	40	4000	100	0.36	0.950	0.975	0.949	1.000	max RP
134	40	3500	600	0.36	0.700	0.829	0.696	0.589	median-like
138	40	2500	1600	0.36	0.200	0.360	0.255	0.143	best
139	30	3300	100	0.64	0.933	0.970	0.925	1.000	max RP
143	30	3000	400	0.64	0.733	0.867	0.712	0.667	median-like
146	30	2500	900	0.64	0.400	0.640	0.453	0.333	equal the best in area-based metrics
148	30	2500	900	0.64	0.400	0.640	0.453	0.000	best, coincident centres
149	20	2800	100	0.84	0.900	0.964	0.861	1.000	max RP
153	20	2500	400	0.84	0.600	0.840	0.594	0.447	equal the best in area-based metrics
154	20	2500	400	0.84	0.600	0.840	0.594	0.200	best
155	10	2500	100	0.96	0.800	0.960	0.679	1.000	max RP, equal the best in area-based metrics
160	10	2500	100	0.96	0.800	0.960	0.679	0.000	best, coincident centres

Table A1. Representative subset of segmentation evaluation results across all synthetic segment sets. For each set, rows were selected to capture extremes of RP and area-based metrics (M, QR, OUS), median-like cases, and instances where identical RP values correspond to different area-based outcomes.