Guidelines to Compare Semantic Segmentation Maps at Different Resolutions

Christian Ayala, Carlos Aranda, Mikel Galar, *Member, IEEE,*

***Abstract*—Choosing the proper ground sample distance (GSD) is a vital decision in remote sensing, which can determine the success or failure of a project. Higher resolutions may be more suitable for accurately detecting objects, but they also come with higher costs and require more computing power. Semantic segmentation is a common task in remote sensing where GSD plays a crucial role. In semantic segmentation, each pixel of an image is classified into a predefined set of classes, resulting in a semantic segmentation map. However, comparing the results of semantic segmentation at different GSDs is not straightforward. Unlike scene classification and object detection tasks, which are evaluated at scene and object level, respectively, semantic segmentation is typically evaluated at pixel level. This makes it difficult to match elements across different GSDs, resulting in a range of methods for computing metrics, some of which may not be adequate. For this reason, the purpose of this work is to set out a clear set of guidelines for fairly comparing semantic segmentation results obtained at various spatial resolutions. Additionally, we propose to complement the commonly used scene-based pixel-wise metrics with region-based pixel-wise metrics, allowing for a more detailed analysis of the model performance. The set of guidelines together with the proposed region-based metrics are illustrated with building and swimming pool detection problems. The experimental study demonstrates that by following the proposed guidelines and the proposed region-based pixel-wise metrics, it is possible to fairly compare segmentation maps at different spatial resolutions and gain a better understanding of the model's performance. To promote the usage of these guidelines and ease the computation of the new region-based metrics, we create the seg-eval Python library and make it publicly available at https://github.com/itracasa/seg-eval.**

***Index Terms*—Remote Sensing, Semantic Segmentation, Quality Assessment, Error Metrics.**

I. Introduction

Earth Observation (EO) has become a hot topic due to its significant impact on society. Open data policies adopted by agencies such as NASA [1] or ESA [2] have unleashed the power of remote sensing yielding meaningful insights about our world. As a result of the vast amount of data available today, traditional techniques to process EO data have become impractical. In the last decade, Deep Learning-based (DL) models have shown outstanding success in almost every application domain ranging from computer vision [3] to natural language processing [4]. When it comes to EO applications, DL models can accurately classify [5] and segment [6] images, detect objects [7], track changes over time [8], or answer text-based questions about images [9].

EO data pose new challenges compared to natural images (i.e., ImageNet) [10]. In remote sensing images, there is a high intra-class diversity and a small inter-class dissimilarity. That is, objects within the same semantic class may differ in shape and color (i.e., buildings) making it extremely difficult to distinguish them from objects in similar classes (i.e., industrial vs. residential). In addition, depending on the sensor, the data measured, the viewing angle, the scale, and the illumination may differ, making it almost impossible to build a single DL model that performs well in every scenario even for the same task.

When building a DL model for EO data, delicate decisions must be made about the type of sensor (i.e., radar vs. multi-spectral) and the ground sampling distance (GSD) (i.e., sub-meter, meter, or kilometer). While the former is relatively easy to make based on domain expertise, the latter is not since it heavily depends on the use case [11]. Accordingly, a slight variation in the GSD can be the difference between seeing an object or not. Hence, for accurately detecting any object, higher resolutions may seem to be the ideal solution. However, from an operational point of view, very high resolution imagery is costly and difficult to access. Furthermore, an increase in the spatial resolution results in larger images, which in turn require an increase in computing and storage resources [12].

Therefore, choosing an appropriate GSD is crucial in any remote sensing project. To make this decision objectively, it is necessary to have metrics that enable the comparison of different GSDs. For tasks such as image classification [13], [14] and object detection [15], [16], it is quite simple, since metrics are computed at the scene and object level, respectively. In these cases, only the input changes (the GSD), but given that the outputs are in the same unit and scale, it is easy to compare different input GSDs. However, for semantic segmentation tasks, the output changes with the size of the input, so different GSDs produce segmentation maps with different dimensions. As a result, deciding based on the metrics extracted from differently shaped outputs (at different scales) is not straightforward.

In this respect, there is no consensus in the literature on how these maps should be compared [17]. Most of the works agree on the fact that all segmentation maps must be rescaled to match the ground truth masks resolution [18]–[21], but there

C. Ayala was partially supported by the Government of Navarre under the industrial PhD program 2020 reference 0011-1408-2020-000008. M. Galar was supported by the Spanish Ministry of Science and Innovation under projects PID2019-108392GB-I00 and PID2022-136627NB-I00 (MCIN/AEI/10.13039/501100011033) and by the Public University of Navarre under project PJUPNA25-2022.

C. Ayala and C. Aranda are with Tracasa Instrumental, Calle Cabárceno 6, 31621 Sarriguren, Spain (cayala@itracasa.es; caranda@itracasa.es).

M. Galar is with the Institute of Smart Cities (ISC), Public University of Navarre (UPNA), Arrosadia Campus, 31006 Pamplona, Spain (mikel.galar@unavarra.es).

are others using the metrics computed at different resolutions (with varying scale ground truths) to make the comparison [22], or even works not providing any details on how the comparison is done [23], [24]. Therefore, it is necessary to study and propose a set of clear guidelines that can help researchers and practitioners deciding the most appropriate GSD for their semantic segmentation problem.

In addition, in [25] and [26], the authors suggest considering the usage of a tolerance buffer when computing the metrics to account for errors that may occur at object boundaries due to the resolution used. However, the tolerance buffer does not solve the aforementioned problems, since the metrics are still incomparable across different resolutions, requiring a clear guideline to set the buffer width according to its objective.

Applying these guidelines to find the most adequate GSD for a use case, several models are trained at different GSDs, looking for the best trade-off between accuracy and computational requirements. To reduce the number of experiments and their associated cost, it would be helpful to know in advance the upper bound of the evaluation metrics that can be achieved at each GSD. As a result of our research, we devised a method to compute these upper bounds without training any model.

In summary, this paper proposes a set of guidelines for fairly comparing segmentation maps with varying GSDs addressing the following issues:

- How to make segmentation maps at different GSDs comparable.
- Why the commonly used pixel-wise metrics can lead to a misleading view of model's performance especially in imbalanced scenarios.
- How to set the tolerance buffer to only account for the uncertainty in the edges due to the resolution.
- How to calculate upper bounds for semantic segmentation performance metrics at different GSDs.

Following these guidelines, commonly used pixel-wise semantic segmentation metrics will allow us to get a fairer overview of how our model performs. Nevertheless, it will still be difficult to identify specific cases where the model may be performing poorly. To address this issue, we propose a novel methodology for computing object level semantic segmentation metrics that can help in the interpretation of the results obtained by different models. The proposed region-based pixel-wise metrics calculate pixel-wise semantic segmentation metrics within a pre-defined set of evaluation regions (e.g., a region around each building). These metrics provide new insights into the effects of DL modeling choices such as architecture, loss function, or patch size on model's performance. Furthermore, region-based pixel-wise metrics can be used to identify the weaknesses of the model (e.g., difficulties detecting small buildings) and thus guide future labelling efforts.

In summary, the two main contributions of this paper are:

1) A clear set of guidelines for fairly evaluate remote sensing semantic segmentation maps across different spatial resolutions (presented in Section IV).
2) A methodology for computing region-based pixel-wise semantic segmentation metrics in remote sensing (presented in Section V).

Despite their differences, the proposed guidelines and the region-based pixel-wise metrics are both useful and complementary for assessing the quality of semantic segmentation maps. Accordingly, the former should be applied when computing region-based pixel-wise metrics, while the latter complements the traditional analysis performed by using standard pixel-wise metrics.

Two different use cases have been considered to illustrate the necessity of the proposed guidelines and the usefulness of region-based pixel-wise metrics. On the one hand, the Massachusetts Building Aerial dataset [25] has been chosen due to the high intra-class diversity of building footprints. In addition, we have added high-resolution satellite imagery to the dataset to simulate a super-resolution semantic segmentation (SRSS) use case where segmentation maps at multiple resolutions must be compared, which is a very common use case where our guidelines should be applied. Finally, the BH-Pools dataset [27] has been used to extract swimming pools from submeter aerial imagery, which has low intra-class diversity and much more imbalance than building segmentation. Moreover, annotations are not provided in vector format allowing us to show how the proposed guidelines may be applied in these cases.

The experiments conducted demonstrate that the proposed guidelines can be useful not only for comparing segmentation maps with different spatial resolutions, but also for better understanding the performance of the model. Finally, the code developed in this work has been organized into a Python library named seg-eval, which has been made freely available at GitHub[1]. We believe that these guidelines will facilitate the comparison of research results within the remote sensing community.

The remainder of this article is organized as follows. Section II summarizes the related works with regard to the accuracy assessment procedure commonly followed for comparing semantic segmentation maps at different spatial resolutions. Then, Section III describes the materials and methods used in the experiments carried out. Thereafter, Section IV presents the proposed guidelines for comparing segmentation maps at multiple spatial resolutions. Furthermore, Section V presents a novel approach for computing region-based semantic segmentation pixel-wise metrics. Finally, Section VI summarizes the lessons learned and presents some future research.

II. Related works

Accurately evaluating the performance of a DL model is essential for the advancement of research. The sampling unit serves as the foundation for accuracy assessment and ranges from single pixels and polygons to the whole scene [28]. Accordingly, a prediction and a reference ground truth are compared on a sampling unit scale. For example, image classification and object detection tasks are evaluated on scene and polygon scales, respectively, while semantic segmentation tasks are commonly evaluated at a pixel level.

If a pixel is chosen as the sampling unit, metrics must be computed pixel-wise by mapping each pixel from the predicted

[1] https://github.com/itracasa/seg-eval

segmentation map to the ground truth map [29]. Due to this sampling design, segmentation maps at different GSDs cannot be directly compared because the number of pixels that cover a specific area varies depending on the spatial resolution. However, this is essential, for example, to assess the effect of super-resolution on subsequent semantic segmentation tasks or to evaluate the suitability of different GSDs for a given task.

The SRSS problem is a clear example where segmentation maps at different GSDs should be compared. To obtain segmentation maps with a higher spatial resolution than the one provided at the input, either stage-by-stage or end-to-end approaches can be adopted [18]. While the former often super-resolves the low-resolution input prior to the segmentation network [23], the latter integrates the super-resolution process into the segmentation task [22]. As a result, it is necessary to compare the super-resolved segmentation maps to those derived from the native GSD in order to determine whether these approaches outperform traditional segmentation techniques (i.e., without super-resolving).

Both stage-by-stage and end-to-end approaches have been fairly compared to other state-of-the-art SRSS techniques in previous works [18], [19], using the same high-resolution ground truth. To evaluate the individual contribution of the super-resolution and semantic segmentation modules, it is common practice to rescale the resulting low-resolution semantic segmentation maps to match the GSD of the ground truth [20], [21], [30], [31]. As we will show in Section IV, interpolating the segmentation maps is not the most accurate procedure, since class probabilities before the final segmentation can be used for a better estimation.

Pereira et al. [32] proposed an end-to-end SRSS framework, but conducted the experimentation from a different point of view. Rather than resampling the generated segmentation maps to match the GSD of the ground truth, inputs were downscaled using the bicubic interpolation prior to being fed to the network. As a result, all the output segmentation maps have the same dimensions and can be compared to the same high-resolution ground truth. However, this method has the drawback of losing critical information during the downscaling process.

Otherwise, in [22] a novel DL architecture that learns how to super-resolve an image internally to produce fine-grained semantic segmentation maps was proposed. In the experimental study, each predicted segmentation map was compared to its corresponding ground truth mask (previously rasterized to the target GSD). It must be noted that this differs from previous works since ground truth masks at multiple spatial resolutions are available. Anyway, this methodology cannot be considered fair, since a fair comparison should be done in the same unit scale. Finally, there are also other works whose accuracy assessment procedure is unclear [23], [24].

Based on these previous works, we believe that it is necessary to establish a consistent set of guidelines for comparing the quality of semantic segmentation maps across different GSDs that can be easily adopted by the remote sensing community.

III. Materials and methods

This section introduces the materials (dataset, DL model, training strategy, evaluation metrics, and seg-eval Python package) used in the experiments carried out to illustrate the proposed accuracy assessment guidelines and the region-based pixel-wise metrics.

A. Datasets

To illustrate how the proposed guidelines can be used in practice, the Massachusetts Building Aerial [25] and BH-Pools [27] datasets have been considered. Table I summarizes both datasets.

TABLE I
SUMMARIZED DESCRIPTION OF THE DATASETS CONSIDERED TO PUT INTO PRACTICE THE PROPOSED GUIDELINES.

	Massachusetts Building Aerial [25]	BH-Pools [27]
Abbreviation(s)	MB-Aerial, MB-Sat	BH-Pools
Color spectrum	RGB	RGB
Native GSD (m)	1	0.15
Simulated GSD (m)	MB-Aerial: 2, 3, 4, 5, 10 / MB-Sat: 10	0.3, 0.6, 1.2
Total Coverage (km^2)	340	248
Target object	Buildings	Swimming pools
Avg. object size (m^2)	236.57	467.51
Vector annotations	Yes	No
Intra-class diversity	High	Low
Class imbalance (% positive class)	Medium (40.48%)	High (0.83%)

1) Massachusetts Building Aerial dataset: The Massachusetts Building Aerial dataset [25] (MB-Aerial) contains 151 aerial images of 1500×1500 pixels covering urban and suburban areas of Boston. Approximately 2.25 km^2 are covered by each aerial image at a GSD of 1 m. These aerial images have been labeled using annotations from the OpenStreetMap project [33]. These comprise buildings of diverse sizes, including factories, individual houses, and garages. It must be noted that ground truth vector annotations are also provided within the dataset. This allows one to compute the metrics more accurately, since the boundaries of the objects are better preserved. Images were randomly split into training, validation, and test sets following Mnih's indications [25], resulting in 137, 4, and 10 images, respectively. A sample of the dataset is shown in Figure 1.

To evaluate if the proposed guidelines can help compare the accuracy of semantic segmentation maps at different spatial resolutions, images have been degraded to 2, 3, 4, 5, and 10 m using the bicubic interpolation. To train the corresponding semantic segmentation models, the ground truth annotations in vector format have been rasterized to the different spatial resolutions. It is important to mention that, during the rasterization process the *all_touched* flag has been set. In other words, all pixels that intersect with a geometry are considered positive class pixels. Figure 1 shows degraded images and corresponding ground truth masks for a representative region.

In order to assess the validity of the conclusions drawn in a simulated SRSS use case, real satellite imagery with a GSD of 10 m has been included as a supplement to the dataset. This is a common scenario in which segmentation maps at different spatial resolutions are compared. For this purpose, freely available Sentinel-2 imagery, using only the RGB bands, has been considered. In the experimental study,

a SRSS approach is utilized with satellite imagery to enhance the spatial resolution of semantic segmentation maps up to 5 m and 2.5 m. To train the SRSS models, ground truth annotations in vector format need to be rasterized to 5 m and 2.5 m. This dataset will be referred to as MB-Sat.

2) BH-Pools dataset: We have also considered the BH-Pools dataset [27], since it represents the imbalance problem better than the MB-Aerial dataset. The BH-Pools dataset consists of 200 images covering 8 different neighborhoods in the city of Belo Horizonte, Minas Gerais, Brazil. The images were exported from Google Earth Engine and have an eye altitude of 330 meters with a resolution of 3840×2160 pixels. Nevertheless, since Google Earth combines different images from different observations to build up its high-resolution images, it is not possible to determine the exact spatial resolution. However, given the previous information and through a visual analysis the GSD was estimated to be 0.15 m [34]. Images were manually annotated resulting in 3980 ground truth polygons. In this case, only the final segmentation masks are provided in raster format, which reduces the precision of the accuracy assessment. To fit and evaluate the models, images were randomly split into training, validation, and test sets, resulting in 571, 122, and 122 images, respectively. Finally, we have degraded the images to 0.15, 0.3, 0.6, and 1.2 m using the bicubic interpolation to test if the proposed guidelines can help compare the accuracy of semantic segmentation maps at different spatial resolutions. Since this time no vector annotations have been provided, ground truth masks have been generated by rescaling the one given at the maximum GSD using the nearest neighbor interpolation. A sample of the dataset is shown in Figure 2.

B. Model

The model consists of a U-Net architecture [35] with a ResNet-34 [36] encoder. To generate segmentation maps with a higher spatial resolution than the one provided at the input (MB-Sat dataset), a super-resolution module can be added before the encoder [22]. Specifically, it consists on a nearest neighbor upsample layer prior to the feature extractor. It must be noted that we are not looking for the state-of-the-art performance, so we focus on a simple yet powerful and reproducible semantic segmentation architecture that can be used to put the proposed guidelines into practice.

C. Training strategy

All the experiments were conducted under the same conditions. Specifically, models were trained for 100 epochs, taking batches of 16 samples. As mentioned earlier, to isolate the effects of spatial resolution on final performance, it is crucial that all patches cover the same area (in each dataset). However, covering the same area with different GSDs results in patches with varying numbers of pixels. To ensure that the input patch size for the network is consistent across all resolutions, we have chosen to resize them to 128×128 for the MB-Aerial and MB-Sat datasets, and 256×256 pixels for the BH-Pools dataset. The patch size was set based on expert knowledge and has determined the maximum possible degraded resolution for each dataset as described in Section III-A. Additionally, in the case of the BH-Pools dataset, due to the heavy class imbalance, for training, we only considered samples with at least 10% of pixels corresponding to the positive class (pool) to prevent the training process from being dominated by samples without positive pixels. Nevertheless, to fairly assess the performance of the models, we consider all testing samples, including those without any pixels belonging to the positive class. Finally, the best model based on validation loss was selected to prevent overfitting.

As in other works [22], [37], [38], we have used the well-established Combo Loss [39] as a loss function, which combines the Binary Cross-Entropy [40] and Dice [41] losses. Adam [42] has been used as optimizer with a fixed learning rate of $1e^{-3}$ and a weight decay of $1e^{-2}$.

To increase the generalization capability of the models, affine and photometric data augmentation techniques have been used. In particular, images have been augmented using combinations of horizontal and vertical flips, as well as 90-degree rotations, which is known as Dihedral data augmentation [43]. Moreover, photometric transformations [44] such as randomly changing the brightness, saturation, and contrast of the image have also been applied. Augmentations at testing time are applied to enhance the quality of the resulting segmentation maps. Accordingly, the final predicted segmentation map is computed by aggregating predictions across transformed versions of a testing sample using dihedral transformation (resulting in 8 different predictions). It should be emphasized that all the conclusions drawn from the experimental study (Sections IV and V) would remain unchanged even if test time augmentation had not been applied.

The experiments have been run on a computing node with a 2 $\times$ Intel Xeon E5-2609 v4 @ 1.70 GHz processor with 128 GB of RAM and a NVIDIA RTX2080Ti GPU (11 GB of RAM).

D. Evaluation metrics

There are various evaluation metrics in the semantic segmentation literature that have been used to assess the quality of the resulting segmentation maps [18], [24], [45]. While most of them are based on the number of true positive (TP), false positive (FP), true negative (TN), and false negative (FN) derived from the confusion matrix [46], a variety of indices have emerged to better assess different types of geometric errors [47], [48]. Despite of this, the choice of the evaluation metric depends on the use case. Therefore, we implemented the full set of metrics in the open-source Python package provided with this work, but we consider the most commonly used ones for the analysis in this paper.

The considered metrics are Precision (Prec.), Recall (Rec.), F1-score (F1), and Intersection over Union (IoU). For a binary semantic segmentation problem, Precision [49] (Eq. (1)) is defined as the proportion of true positive predictions among all positive predictions made by the model, while Recall [49] (Eq. (2)) is the proportion of true positive predictions among all pixels that should have been classified as belonging to the positive class. The F1-score [50] (Eq. (3)) is the harmonic

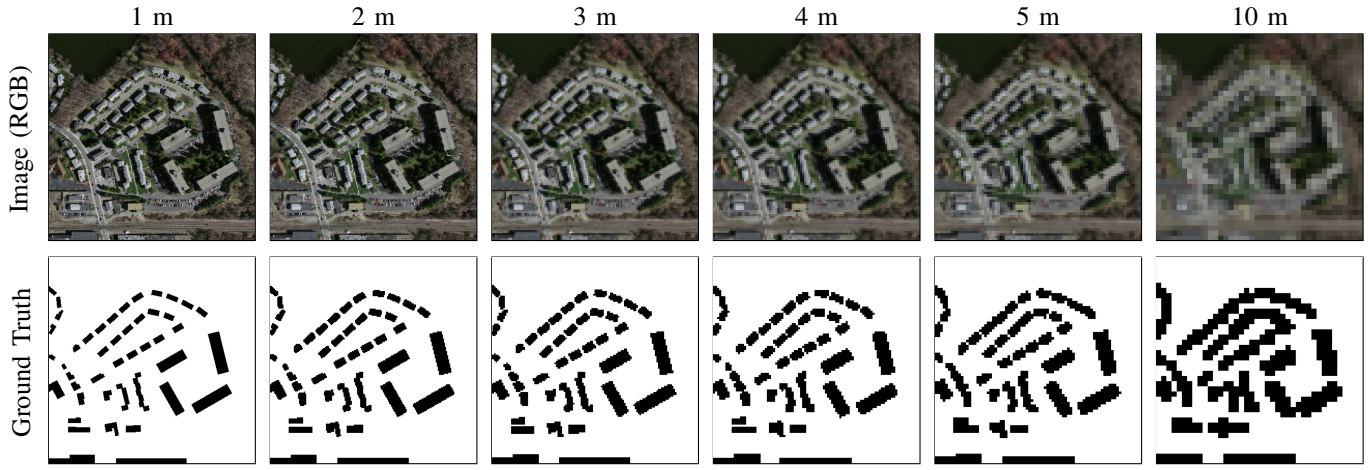


Fig. 1. Sample images from the MB-Aerial dataset, including real imagery (1 m) and degraded versions (2-10 m), along with their corresponding ground truth masks.

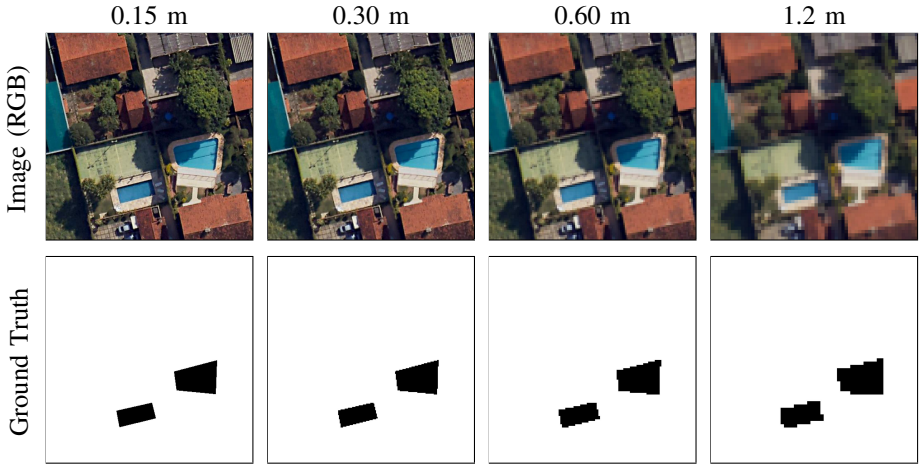


Fig. 2. Sample images from the BH-Pools dataset, including real imagery (0.15 m) and degraded versions (0.30-1.2 m), along with their corresponding ground truth masks.

mean of Precision and Recall, and has a similar definition to the IoU [51] (Eq. (4)), although the latter penalizes false positives and false negatives more.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F1\text{-}score = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{3}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{4}$$

These metrics are frequently applied in multi-class problems by computing each metric individually per class and then averaging the results in different ways (micro, macro, or weighted) [52]. It must be noted that the macro-average is often used, since it equally weights the contribution of all the classes, making it robust against imbalanced datasets [53].

E. Seg-eval Python Package

As a result of this paper, we have developed the seg-eval Python package available at GitHub[2]. This library provides useful tools for evaluating semantic segmentation models, including commonly used pixel-wise semantic segmentation metrics to quantitatively assess the quality of a segmentation map. In addition to the metrics described in Section III-D, the library also includes methods to compute the following metrics: Overall Accuracy, Detection Probability or Hit Rate), Specificity (also known as True Negative Rate), False Negative Rate (also known as Fall-out), False Positive Rate (also known as Miss Rate), Area Under the Curve, Cohen's Kappa, Matthews correlation coefficient, and Hausdorff distance. It is important to emphasize that the novelty of the seg-eval Python package lies not in the implementation of the metrics themselves, but in their adaptation to be computed over shapes. This represents a significant departure from the conventional practice of computing metrics over rasters in remote sensing.

Furthermore, the library allows relaxing these metrics by applying a tolerance buffer and it also includes functionality

[2]https://github.com/itracasa/seg-eval

for generating a set of evaluation regions for computing the region-based pixel-wise semantic segmentation metrics as proposed in Section V. Finally, it provides utilities for analyzing these metrics through visual IoU maps (e.g., Figure 3) or bi-variate probability density functions (e.g., Figure 12).

Overall, the use of this library by the remote sensing community can contribute to a more fair evaluation of the performance of semantic segmentation models and the standardisation of the style of the results and figures presented in future research works.

IV. Guidelines to compare semantic segmentation maps at the pixel level

This section presents common misconceptions that arise when comparing semantic segmentation maps at multiple spatial resolutions and provides guidelines for addressing them. Each proposed guideline can be applied independently, although we suggest using them all together. The order in which the guidelines are presented is not relevant, but we believe this order is adequate for readers to understand them.

The proposed guidelines are tested using the MB-Aerial, MB-Sat, and BH-Pools datasets described in Section III-A. To summarize, after processing each dataset as explained in Section III-A and training the corresponding models, the following information is available for each testing set:

- *Ground truth annotations* in both raster and vector formats at all spatial resolutions. If the original dataset does not include the vector format, it is generated by vectorizing the raster ground truth at the highest spatial resolution.
- *Predicted segmentation maps* in raster and vector formats. Vector formats are derived from raster predictions. The raster predictions are obtained for the original input resolution and are also rescaled to the highest spatial resolution.

The quality of the resulting segmentation maps has been evaluated using the metrics described in Section III-D. For the sake of accuracy, all the metrics presented in the experimental study are computed using the vector format. To complement the results presented in this paper, a full report containing all the metrics for the experiments conducted can be found in the supplementary material. Additionally, we have conducted additional experiments using the DeepLabV3+ architecture [54] to demonstrate that all the conclusions drawn within the experimental study remain the same and thus, the proposed guidelines are useful regardless of the segmentation model considered. Accordingly, the DeepLabV3+ metric report have been also included in the supplementary material.

To get a better idea of the spatial resolutions that are being used and their derived segmentation maps, Figure 3 shows a randomly taken test sample from each dataset (MB-Aerial, MB-Sat, and BH-Pools) together with the predicted segmentation maps using IoU visualization (where the true positives, true negatives, false positives, and false negatives are depicted with different colors).

In the following sections, we will follow a consistent structure: we will identify a problem, provide a detailed description of the issue, propose a guideline for addressing the problem, and experimentally validate the proposed guideline.

A. Choosing a reference map

Problem. Metrics computed at different spatial resolutions cannot be directly compared.

Description. When ground truth masks at multiple spatial resolutions are available, one may mistakenly think that the comparison should be performed at the predicted segmentation map's GSD. However, this is not the most suitable approach because the target population varies, resulting in incomparable results and potentially misleading conclusions.

Guideline. To accurately compare semantic segmentation maps at different GSDs, the same reference ground truth mask must be used, carefully considering the method used to rescale the predicted segmentation maps. In this regard, it is suggested to interpolate the class probabilities before thematic classification.

Illustrative Example. The results comparing the predicted segmentation maps with the ground truth masks at the same GSD and with the ground truth mask at the greatest GSD (rescaling class probabilities) are shown in Table II for the MB-Aerial, MB-Sat, and BH-Pools datasets. These results are presented in terms of the IoU, F1, Prec., and Rec. metrics. The best results achieved for each dataset and performance metric are presented in **boldface**.

TABLE II
Results obtained for the MB-Aerial, MB-Sat, and BH-Pools datasets comparing each predicted segmentation map with the ground truth mask at its corresponding GSD and with the ground truth at greatest GSD after rescaling class probabilities.

		w/o rescaling				w/ rescaling			
	GSD (m)	IoU	F1	Prec.	Rec.	IoU	F1	Prec.	Rec.
MB-Aerial	1	0.5794	0.7331	**0.8286**	0.6596	**0.5840**	**0.7366**	**0.7631**	0.7190
	2	0.6065	0.7548	0.8023	0.7171	0.5443	0.7044	0.6526	0.7859
	3	0.5952	0.7455	0.7766	0.7181	0.4941	0.6601	0.5720	0.7943
	4	0.5888	0.7406	0.7706	0.7158	0.4502	0.6182	0.5213	0.7935
	5	0.6022	0.7508	0.7430	0.7615	0.4058	0.5734	0.4489	0.8264
	10	**0.6142**	**0.7589**	0.6942	**0.8450**	0.2841	0.4339	0.3005	**0.8596**
MB-Sat	2.5	0.3925	0.5624	0.6200	0.5182	0.3353	0.4990	**0.4711**	0.5442
	5	0.4781	0.6447	**0.6891**	0.6119	**0.3417**	**0.5057**	0.4218	0.6635
	10	**0.5855**	**0.7348**	0.6838	**0.7972**	0.2784	0.4286	0.2982	**0.8220**
BH-Pools	0.15	0.7163	0.8319	0.8796	0.7975	0.7163	0.8319	0.8796	0.7975
	0.3	**0.7372**	**0.8464**	**0.9034**	0.8028	**0.7363**	**0.8460**	**0.8809**	0.8204
	0.6	0.7254	0.8376	0.8436	**0.8411**	0.6817	0.8075	0.7612	**0.8695**
	1.2	0.6743	0.8017	0.8706	0.7496	0.6158	0.7591	0.7097	0.8227

We can observe how the results change completely depending on the approach taken. When the predicted segmentation map and ground truth mask are compared at the same GSD (without rescaling), the best overall result is generally achieved at the lowest resolution (MB-Aerial and MB-Sat). However, the performance does not follow exactly the same trend in the BH-Pools dataset due to small variations in the spatial resolution and class imbalance. This result may seem counterintuitive, as Figure 3 shows that the highest resolution segmentation maps in the MB-Aerial and MB-Sat datasets produce a large number of false positives. This issue arises because metrics calculated over different target populations are being compared.

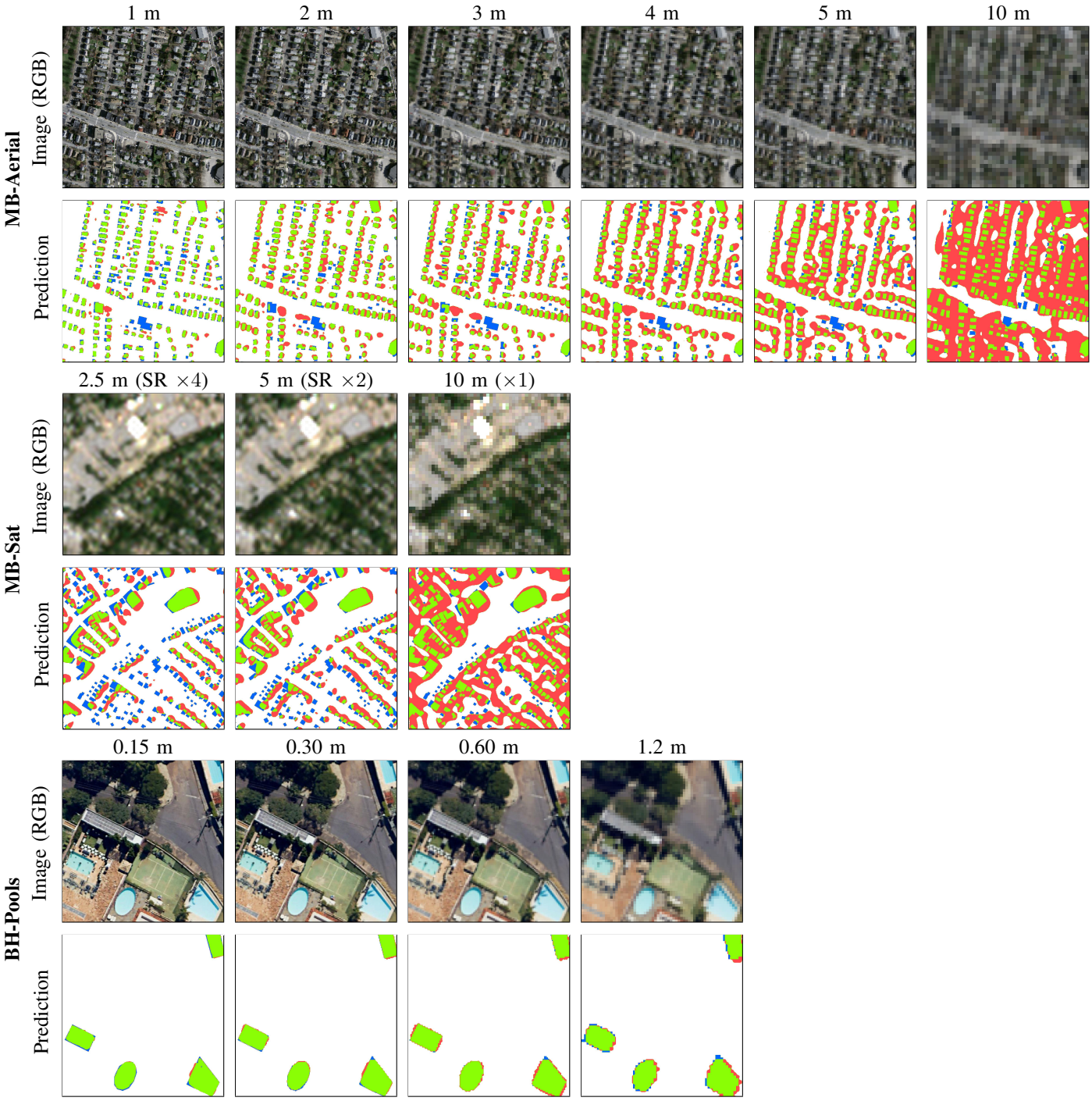


Fig. 3. Image and corresponding predicted segmentation maps obtained at multiple GSDs for the MB-Aerial, MB-Sat, and BH-Pools datasets. Predicted segmentation maps are presented in terms of TP in green, FN in blue, FP in red, and TN in white. It should be noted that randomly taken testing samples are shown in this figure, and thus, the quantitative results presented in the following tables can slightly differ.

If we fix the target population at the maximum GSD as suggested and rescale the outputs (w/ rescaling columns), we can observe that the performance metrics can be fairly compared regardless of spatial resolution and better reflects the qualitative results. This is because pixels from different segmentation maps can be properly matched and compared. In most cases [18], [19], [32], [55], the predicted segmentation map is directly rescaled to the maximum GSD, although it may be more accurate to interpolate the class probabilities before thematic classification, as this preserves more details (see Figure 4).

Following this approach, the overall IoU and F1 results accurately reflect what is shown in the figures, being the best models the 1 m for MB-Aerial, the 5 m for MB-Sat (note that in this case the 2.5 m does not perform as well as the 5 m due to the effect of super-resolution), and the 0.3 m for BH-Pools. In fact, the differences between higher and lower resolutions become more pronounced when fixing the target population,

which is more in line with what we observe qualitatively.

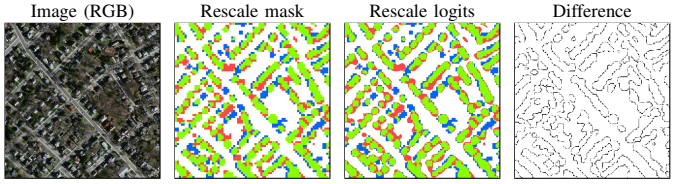


Fig. 4. Image and corresponding predicted segmentation map obtained at 5 m GSD for a representative zone (*22828930_15*) from the MB-Aerial dataset. Predicted segmentation maps are presented in terms of TP in green, FN in blue, FP in red, and TN in white. Both the segmentation mask rescaled from the final prediction (with nearest neighbor) and from the class probabilities (bicubic interpolation) are presented with a map showing where they differ.

B. Choosing a fair metric

Problem. Commonly used single-class semantic segmentation metrics focus only on the foreground class, leading to misleading conclusions in scenarios where the background class is equally important.

Description. We can use a variety of metrics to evaluate the performance of a DL model. Looking at Table II, rescaling the class probabilities to match the greatest GSD, we find that in some metrics such as recall, lower spatial resolutions perform better than higher ones, due to the coarse predictions. This may lead to misleading conclusions if they are analyzed on their own without looking at other metrics. For this reason, the F1-score is usually considered since it takes into account both the precision and recall metrics. A major drawback of the F1-score is that it is computed only for the foreground class without taking the number of TNs into account [56]. As a result, the F1-score is not suitable for scenarios where both classes are relevant such as the building and swimming pool semantic segmentation [57]. Therefore, metrics that consider both foreground and background classes are preferable, since they provide a more accurate picture of the model's performance.

Guideline. Binary semantic segmentation problems should be evaluated as multi-class problems, macro-averaging the performance metrics obtained for both the foreground and background classes.

Illustrative Example. Table III shows the performance in terms of the IoU, F1, Prec., and Rec. metrics for both the foreground and background classes, as well as their macro-average, on the MB-Aerial, MB-Sat, and BH-Pools datasets. Overall, the results are more consistent with what we observe qualitatively. While foreground metrics provide insight into the model's performance regarding under-segmentation errors, background metrics account for over-segmentation errors. To gain a comprehensive understanding of the model's performance, metrics must account for both under- and over-segmentation errors. In this regard, it is a common practice to compute the arithmetic average of the partial accuracies of each class [58]. As a result, macro-averaging both the foreground and background metrics gives a fairer view of the model's performance [59], which is more in line with the qualitative results. This strategy can be easily applied to any binary semantic segmentation metric. However, notice that metrics such as the Matthews Correlation Coefficient (MCC) [56], which takes into account the four components of the confusion matrix could be used without macro-averaging. In this case, the MCC is symmetric, and thus the same results are obtained for both the foreground and background classes (this can be observed in the supplementary material).

C. Addressing the lack of detail due to the limited spatial resolution

Problem. There is no agreement on how the tolerance buffer commonly used for relaxing semantic segmentation metrics in the context of remote sensing should be set.

Description. In remote sensing, it is common to relax metrics to deal with the limited spatial resolution of the imagery used. This is done by ignoring pixels within a tolerance buffer, α, of the object boundaries during metric computation [25], [26]. For instance, Mnih et al. [25], relaxed the metrics by using a tolerance buffer of three times the GSD. On the other hand, Feng et al. [26], chose a tolerance buffer of five times the GSD. We believe these buffers are too optimistic for the imagery used since they ignore many more pixels than necessary.

Guideline. Set the tolerance buffer to the diagonal of the pixel at the corresponding GSD (Eq. (5)).

Illustrative Example. To determine the proper tolerance buffer for a given GSD, we conducted qualitative and quantitative analyses. Our approach is to define the tolerance buffer by comparing the ground truth mask at a given GSD with the ground truth mask at the highest GSD. This allows us to set a buffer that only accounts for the errors caused by the limited spatial resolution, such as those that may occur due to the rasterization process at the edges of objects. As a result, the proposed tolerance buffer is equal to the diagonal of the pixel at the corresponding GSD:

$$\alpha = \sqrt{2 \times GSD^2} \tag{5}$$

Figure 5 visually compares the proposed tolerance buffer with those suggested by Minh et al., and Feng et al., for a sample randomly taken from the MB-Aerial dataset. To illustrate this, we use the ground truth at both 1 and 2 m, where the 2 m ground truth mask is considered the best possible prediction that could be achieved at this resolution. If we want to use the buffer to account for errors in boundaries due to resolution, the buffer should cover the difference between the highest resolution ground truth (1 m) and the ground truth mask at the corresponding resolution (2 m in this case).

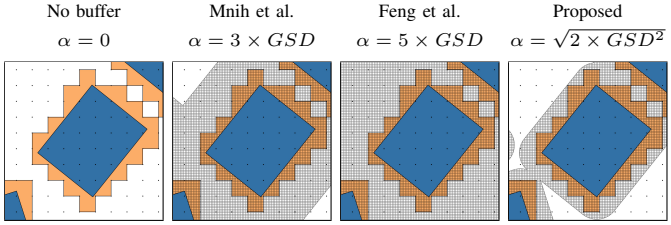


Fig. 5. Visual comparison of different tolerance buffers. The ground truth mask is shown in blue while the predicted segmentation mask is shown in orange. The tolerance buffer is illustrated with a hatched pattern. Additionally, a point grid at 2m GSD has been included for ease of visualization.

TABLE III
RESULTS OBTAINED FOR THE MB-AERIAL, MB-SAT AND BH-POOLS DATASETS CONSIDERING BOTH THE FOREGROUND AND BACKGROUND CLASSES, AS WELL AS THEIR MACRO-AVERAGE.

		Foreground metrics				Background metrics				Macro-averaged metrics			
	GSD (m)	IoU	F1	Prec.	Rec.	IoU	F1	Prec.	Rec.	IoU	F1	Prec.	Rec.
MB-Aerial	1	**0.5840**	**0.7366**	**0.7631**	0.7190	**0.8917**	**0.9421**	0.9317	**0.9537**	**0.7378**	**0.8394**	**0.8474**	0.8364
	2	0.5443	0.7044	0.6526	0.7859	0.8618	0.9250	0.9414	0.9109	0.7031	0.8147	0.7970	**0.8484**
	3	0.4941	0.6601	0.5720	0.7943	0.8264	0.9037	0.9455	0.8664	0.6603	0.7819	0.7587	0.8303
	4	0.4502	0.6182	0.5213	0.7935	0.7946	0.8839	0.9420	0.8354	0.6224	0.7510	0.7316	0.8145
	5	0.4058	0.5734	0.4489	0.8264	0.7341	0.8430	**0.9495**	0.7624	0.5699	0.7082	0.6992	0.7944
	10	0.2841	0.4339	0.3005	**0.8596**	0.4975	0.6391	0.9466	0.5117	0.3908	0.5365	0.6235	0.6856
MB-Sat	2.5	0.3353	0.4990	**0.4711**	0.5442	**0.7788**	**0.8734**	0.8898	**0.8584**	**0.5571**	**0.6862**	**0.6804**	0.7013
	5	**0.3417**	**0.5057**	0.4218	0.6635	0.7327	0.8422	0.9110	0.7860	0.5372	0.6740	0.6664	**0.7247**
	10	0.2784	0.4286	0.2982	**0.8220**	0.5257	0.6720	**0.9338**	0.5455	0.4021	0.5503	0.6160	0.6838
BH-Pools	0.15	0.7163	0.8319	0.8796	0.7975	0.9430	0.9646	0.9971	0.9454	**0.8297**	**0.8983**	0.9383	0.8715
	0.3	**0.7363**	**0.8460**	**0.8809**	0.8204	0.9170	0.9471	0.9974	0.9190	0.8267	0.8965	**0.9392**	0.8697
	0.6	0.6817	0.8075	0.7612	**0.8695**	**0.9490**	**0.9688**	**0.9982**	**0.9504**	0.8153	0.8881	0.8797	**0.9100**
	1.2	0.6158	0.7591	0.7097	0.8227	0.9406	0.9645	0.9975	0.9426	0.7782	0.8618	0.8536	0.8826

As shown in the Figure 5, the tolerance buffers proposed by Minh et al., and Feng et al. occupy much more space than the space corresponding to the lack of resolution. In contrast, the proposed tolerance buffer only considers pixels at the edges of the object, addressing errors caused by the limited spatial resolution without being overly optimistic. This will allow us to know if at a given resolution we can achieve the same metrics as at a higher resolution ignoring the boundary areas where we cannot do as well due to the lack of resolution.

Table IV shows the results of using different tolerance buffers for the MB-Aerial, MB-Sat and BH-Pools datasets in terms of macro-averaged IoU and F1. The tolerance buffer is applied to both the foreground and background classes before the results are macro-averaged. It is important to note that the *all_touched* flag was set during the rasterization process (as described in Section III-A1), and thus the negative buffer is equal to the positive one.

TABLE IV
COMPARISON IN TERMS OF MACRO-AVERAGED IOU AND F1-SCORE BETWEEN DIFFERENT TOLERANCE BUFFERS FOR THE MB-AERIAL, MB-SAT, AND BH-POOLS DATASETS.

		Without buffer		Mnih et al.		Feng et al.		Proposed	
	GSD (m)	IoU	F1	IoU	F1	IoU	F1	IoU	F1
MB-Aerial	1	**0.7378**	**0.8394**	0.7787	0.8692	0.7793	0.8702	0.7685	0.8617
	2	0.7031	0.8147	**0.8121**	**0.8923**	**0.8088**	**0.8909**	**0.7937**	**0.8798**
	3	0.6603	0.7819	0.8061	0.8890	0.7997	0.8858	0.7847	0.8740
	4	0.6224	0.7510	0.7949	0.8823	0.7776	0.8704	0.7746	0.8677
	5	0.5699	0.7082	0.8060	0.8896	0.7896	0.8769	0.7698	0.8648
	10	0.3908	0.5365	0.8105	0.8886	0.7802	0.8564	0.7361	0.8409
MB-Sat	2.5	**0.5571**	**0.6862**	0.6425	0.7676	0.6358	0.7660	0.6116	0.7381
	5	0.5372	0.6740	0.6953	0.8129	0.6635	0.7849	**0.6752**	0.7951
	10	0.4021	0.5503	**0.7563**	**0.8507**	**0.7262**	**0.8168**	0.7092	**0.8213**
BH-Pools	0.15	**0.8297**	**0.8983**	0.8345	0.9014	0.8365	0.9027	0.8321	0.8999
	0.3	0.8267	0.8965	0.8424	0.9066	0.8476	0.9098	0.8355	0.9022
	0.6	0.8153	0.8881	**0.8580**	**0.9165**	**0.8725**	**0.9255**	**0.8384**	**0.9038**
	1.2	0.7782	0.8618	0.8418	0.9063	0.8592	0.9175	0.8130	0.8870

These results support the conclusions drawn from examining Figure 5. Both the Minh et al., and Feng et al., approaches are overly optimistic as they ignore a large number of false positives, reducing the number of evaluated pixels and increasing the impact of true positives on the final metric. The MB-Sat dataset clearly shows how this can lead to undesirable results, such as lower GSD outperforming higher ones. Additionally, it may seem counterintuitive that Minh et al.'s approach outperforms Feng et al.'s, even though the latter evaluated fewer pixels. This can be explained by the fact that, we are looking at the macro-averaged results. When considering only foreground metrics, a larger tolerance buffer outperforms a smaller one. For further details, please refer to the complete metric report provided as a supplementary material.

Overall, the proposed tolerance buffer results in a more accurate way of assessing the model's performance, as it only ignores pixels that are subject to uncertainty due to a lack of resolution.

D. Establishing an upper bound for the pixel level metrics

Problem. There is no well-established strategy to estimate the upper bound that can be achieved with a GSD for a given semantic segmentation performance metric.

Description. Determining the maximum possible value of a performance metric at a specific spatial resolution is useful because it allows for the efficient use of resources by eliminating unnecessary experiments. This can save time and money by allowing researchers and practitioners to discard GSDs that are known to be ineffective for the given use case. Finding the upper and lower bounds for performance metrics is relevant and has also been studied in other domains such as object detection [60]. Despite attempts being made in SRSS to define the upper and lower boundaries for performance metrics [18], to the best of our knowledge, there is no work describing a methodology to extrapolate this to multiple spatial resolutions. Hence, we propose a simple but effective method for defining the upper bound for any performance metric at a given GSD, without having to train a model, just using ground truth annotations.

Guideline. Rasterize ground truth vector annotations to any spatial resolution. To determine the upper bounds, compute performance metrics between the rasterized ground truth and the ground truth at the greatest spatial resolution.

Illustrative Example. Degraded ground truth masks are the result of rasterizing ground truth vector annotations to many GSDs. These ground truth masks can be seen as the most accurate predictions that a semantic segmentation model can make at those resolutions. Therefore, the upper bound can be easily defined by computing performance metrics between the degraded ground truth masks and the one at the highest spatial resolution.

Table V presents the upper bounds in terms of macro-averaged IoU and F1-score for the MB-Aerial, MB-Sat, and BH-Pools datasets. Furthermore, the performance metrics derived from using real predicted segmentation maps (after training a model) are included in the comparison with and without applying a tolerance buffer. These results are also shown in Figure 6.

TABLE V
COMPARISON BETWEEN THE ESTIMATED UPPER BOUNDS AND REAL VALUES (OBTAINED AFTER TRAINING A MODEL) IN TERMS OF MACRO-AVERAGED IOU AND F1-SCORE FOR THE MB-AERIAL, MB-SAT AND BH-POOLS DATASETS.

		Upper bound		Real w/o buffer		Real w/ buffer	
	GSD (m)	IoU	F1	IoU	F1	IoU	F1
MB-Aerial	1	**0.9093**	**0.9510**	**0.7378**	**0.8394**	0.7685	0.8617
	2	0.8420	0.9105	0.7031	0.8147	**0.7937**	**0.8798**
	3	0.7858	0.8737	0.6603	0.7819	0.7847	0.8740
	4	0.7389	0.8411	0.6224	0.7510	0.7746	0.8677
	5	0.6984	0.8116	0.5699	0.7082	0.7698	0.8648
	10	0.5529	0.6954	0.3908	0.5365	0.7361	0.8409
MB-Sat	2.5	**0.8130**	**0.8918**	**0.5571**	**0.6862**	0.6116	0.7381
	5	0.6984	0.8118	0.5372	0.6740	**0.6752**	0.7951
	10	0.5529	0.6961	0.4021	0.5503	0.7092	**0.8213**
BH-Pools	0.15	**1.0000**	**1.0000**	**0.8297**	**0.8983**	0.8321	0.8999
	0.3	0.9763	0.9878	0.8267	0.8965	0.8355	0.9022
	0.6	0.9342	0.9648	0.8153	0.8881	**0.8384**	**0.9038**
	1.2	0.8664	0.9233	0.7782	0.8618	0.8130	0.8870

Closely looking at Figure 6 it becomes evident that the estimated upper bounds are reasonable, as no real without buffer results surpass them. For all the datasets, there is a consistent shift between the estimated and real measurements. When using a tolerance buffer, the real data outperforms the estimated upper bounds for some GSDs. This is because the relaxed metrics are directly compared to upper bounds derived from the ground truth at their respective GSDs, rather than their relaxed counterparts. In this regard the relaxed upper bounds are all perfect performances since spatial resolution-related errors are ignored. It is worth noting that the impact of the tolerance buffer on the BH-Pools dataset is minimal due to the slight variations in the GSD.

The estimated upper bounds may help to directly discard working with a GSD for a specific problem. For instance, for the building detection task, a 10 m GSD would lead to a maximum possible macro-averaged IoU of 0.5529 which may be considered not enough. If we want to ensure a higher detection accuracy we may discard GSDs whose estimated macro-averaged IoU upper bound is below 0.75. Under these circumstances, we would end up choosing at least a 3 m GSD which leads to a maximum possible macro-averaged IoU of 0.7858. However, depending on the use case, if errors at the building boundaries are acceptable (e.g., for demographic analyses) a 5 m GSD would result in a buffered macro-averaged IoU of 0.7698, which is close to the 3 m GSD (0.7847), while reducing computing requirements by 64%.

V. REGION-BASED PIXEL-WISE METRICS

In the field of semantic segmentation, performance is often measured using pixel level metrics. These metrics provide insight into the model's ability to make detailed predictions about individual pixels in an image. However, it is also useful to compute metrics at the object level, which provides information about the model's ability to classify and identify objects within an image. By combining metrics at both the pixel and object levels, it is possible to gain a more comprehensive understanding of the performance of a semantic segmentation model.

To assess the performance of a semantic segmentation model at the object level, it is necessary to map objects from the reference ground truth to the predicted segmentation map. This is commonly done in biomedical image segmentation, where object association criteria are applied to map segmented nucleus with their corresponding ground truth annotations [61]. The simplest method for doing this is to label connected regions, which are groups of adjacent pixels that can be reached from one another within a certain number of orthogonal hops. However, at low spatial resolutions, this method can lead to many connected regions being mapped to the same reference object, as shown in Figure 7. This can make it challenging to accurately evaluate the model's performance at the object level.

Unfortunately, there is no well-established strategy to compute metrics pixel-wise at an object level regardless of the spatial resolution. For this reason, this section presents a methodology to define evaluation regions where pixel-wise metrics can be computed to assess the performance of a semantic segmentation model at an object level. It is important to recognize that the proposed methodology is specifically intended to work with polygonal geometries. To deal with more complex geometries, such as line strings (e.g., roads), it may be necessary to reconsider how evaluation regions are defined.

A. Defining evaluation regions

In order to evaluate the quality of a semantic segmentation map with respect to individual objects, we must first define evaluation regions for each object. Pixels within these regions will be used to compute the confusion matrix for the corresponding object. It is important that these regions cover the entire scene area and are resolution-agnostic, in order to allow for a fair comparison between different resolutions.

To generate evaluation regions, we propose using Voronoi diagrams, which have been used in many contexts including semantic segmentation [62], [63]. Given a set of n points on a plane, the Voronoi diagram of those points subdivides the plane into n cells, each enclosing the portion of the plane that

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2024.3369310

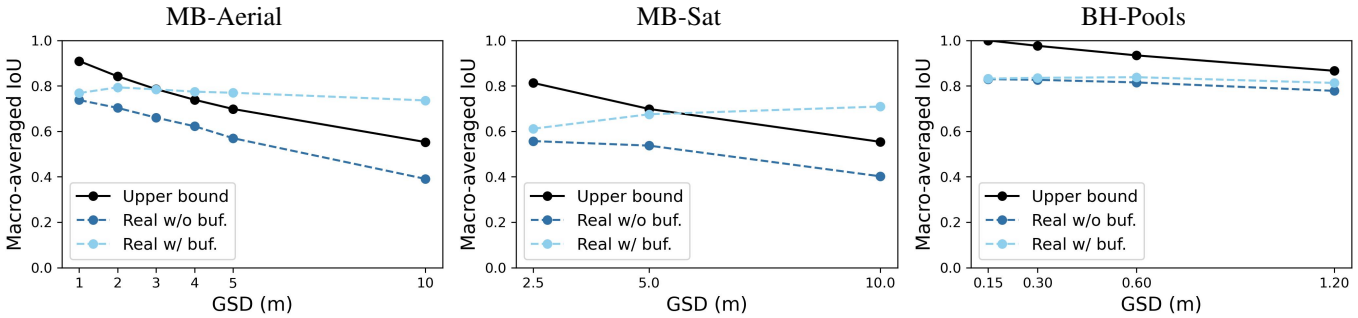


Fig. 6. Visual comparison between the estimated upper bounds and real values (obtained after training a model) in terms of macro-averaged IoU for the MB-Aerial, MB-Sat and BH-Pools datasets.

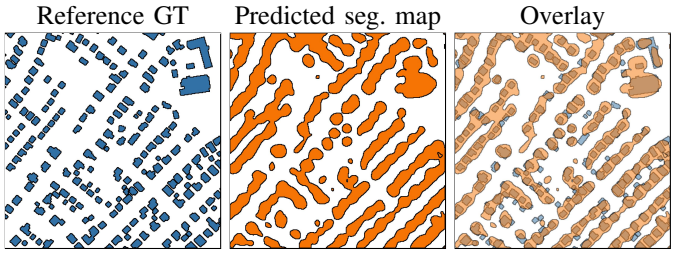


Fig. 7. Visual comparison between reference ground truth objects and connected regions extracted from the predicted segmentation map at 5 m for the MB-Aerial dataset.

is closest to a given point. This approach can be extended to polygons by sampling points from their boundaries and merging the resulting Voronoi cells based on their source polygons. The more points we sample, the more accurate the resulting Voronoi cells will be. It should be noted that the definition of cell boundaries is heavily influenced by the distance metric used. In our experiments, we used the Euclidean distance metric. Figure 8 illustrates the process of generating Voronoi cells.

B. Computing region-based pixel-wise metrics

Once the evaluation regions have been defined, metrics can be calculated pixel-wise for each region. This allows us to gain new insights into the performance of the model, such as how it is performing with respect to the object size. By looking at the metrics for individual regions, it is possible to determine which objects are consistently being misclassified or are otherwise challenging for the model to predict. This information can then be used to improve the model or to focus future research efforts on specific areas of difficulty. Additionally, comparing the performance of the model on different types of objects (e.g., large vs. small objects, objects with complex shapes, etc.) can provide valuable insight into the strengths and weaknesses of the model. It is worth noting that the results are fairly comparable across different spatial resolutions as the same evaluation regions are used.

Since we are working at pixel level in each region we can put into practice the guidelines described in the previous section. In this regard, ground truth segmentation masks can be used to estimate the upper bound of any performance metric within the evaluation region. By analyzing the relationship between performance metrics and features such as object area, we can identify patterns or trends in the data that may not be apparent from looking at the variables individually.

Figure 9 shows the relationship between the macro-averaged IoU and the object area through a bi-variate probability density function for the MB-Aerial, MB-Sat, and BH-Pools datasets. Although we have focused on the object area, other features may be of interest depending on the use case. For the purpose of comparison, both upper bound estimation and actual predicted values are shown.

In the case of the MB-Aerial dataset, the macro-averaged IoU decreases gradually from 1 to 5 m. However, there is a significant drop in performance at the lowest resolution of 10 m. The results of Figure 9 suggest that this decline in performance is primarily due to difficulties in detecting small and medium-sized objects, rather than larger objects which can be easily identified regardless of the GSD. Same conclusions can be drawn for the MB-Sat dataset. For the BH-Pools dataset, Figure 9 shows only minor differences in performance across spatial resolutions. However, it is still possible to observe the gradual loss of delineation of small swimming pools as spatial resolution decreases.

C. Exploring model performance with region-based pixel-wise metrics

This novel approach provides new insights into how the model is performing and can be used to compare models and identify their strengths and weaknesses. As an example, we have compared segmentation results derived from super-resolved satellite imagery (MB-Sat) to those obtained from downgraded aerial imagery (MB-Aerial).

Figure 10 compares the relationship between the macro-averaged IoU and the object area through a bi-variate probability density function for the MB-Aerial and MB-Sat datasets. It should be noted that since there is no 2.5 m degraded aerial imagery, the super-resolved satellite imagery at 2.5 m has been compared to both 2 and 3 m degraded aerial imagery. Upon closer inspection of the results, it can be concluded that up to 5 m, both degraded aerial imagery and super-resolved satellite imagery perform similarly (note that satellite uses 10 m images as input). Furthermore, it has been found that as the spatial resolution increases, the performance gap between the super-resolved satellite imagery and the degraded aerial imagery grows.

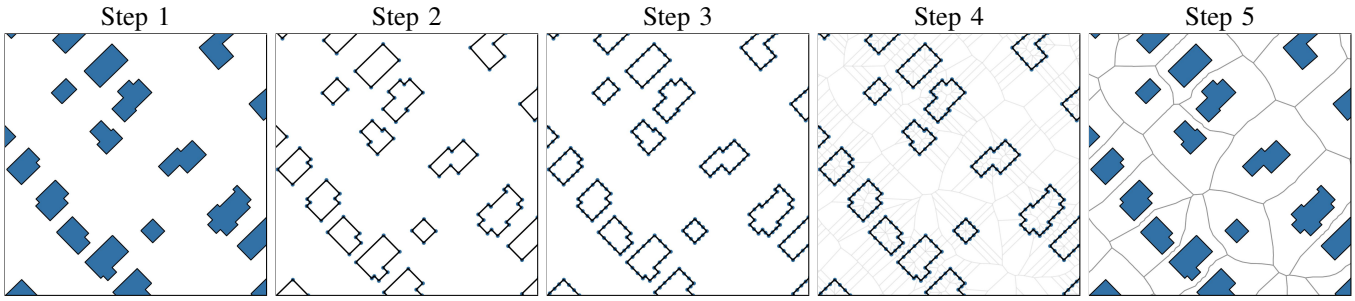


Fig. 8. Visual example of the procedure to generate evaluation regions for computing region-based pixel-wise performance metrics. After sampling points from the boundary of objects (step 3), Voronoi cells are computed (step 4) and merged based on the source polygon (step 5).

Additionally, this novel approach can be particularly useful in situations where commonly used pixel-wise metrics are similar, making it difficult to determine which model configuration is superior. For example, at low resolutions, it is common to modify the patch size to increase the context provided to the model. However, pixel-wise metrics for different patch sizes often only vary slightly, making it difficult to determine whether one patch size is superior to another based solely on these metrics.

To illustrate this, we have compared several patch sizes at 1 m GSD for the MB-Aerial dataset. Figure 11 shows the predicted segmentation maps for two testing samples randomly taken from the MB-Aerial dataset. It becomes evident that larger patch sizes result in less noisy segmentation maps, and better detection of medium and large-sized buildings.

Table VI compares the commonly used pixel-wise metrics and the proposed region-based macro-averaged metrics for each patch size configuration. Pixel-wise metrics provide information on the overall segmentation performance for each patch size configuration. In this regard, larger patch sizes outperform lower ones, although their differences are more significant quantitatively than qualitatively. This is because pixel-wise metrics do not give information about the detection capability of the models. However, region-based macro-averaged metrics provide a more accurate picture of the performance, reflecting the detection accuracy of the models. Therefore, when we compare just the detection capabilities of the models, we observe that there are slight differences between patch size configurations, meaning that all of them have a similar detection capability which is in line with what we observe qualitatively.

TABLE VI
COMPARISON BETWEEN DIFFERENT PATCH SIZE CONFIGURATIONS IN TERMS OF PIXEL-WISE AND REGION-BASED MACRO-AVERAGED IOU AND F1-SCORE FOR THE MB-AERIAL DATASET.

	Pixel-wise				Region-based			
Patch size	IoU	F1	Prec.	Rec.	IoU	F1	Prec.	Rec.
128×128	0.7378	0.8394	0.8474	0.8364	0.6866	0.7801	0.7935	0.7915
256×256	0.7639	0.8584	0.8600	0.8592	0.6966	0.7848	0.7919	0.7998
512×512	0.7819	0.8710	0.8801	0.8643	0.7082	0.7928	0.8040	0.7993
1024×1024	**0.7952**	**0.8802**	**0.8819**	**0.8803**	**0.7208**	**0.8056**	**0.8120**	**0.8159**

Figure 12 compares all patch size combinations in pairs, in terms of the bi-variate probability density function of the macro-averaged IoU and object area. This allows one to have a deeper look at which specific scenarios each patch size configuration is outperforming the others. Overall, we see that increasing the patch size improves detection accuracy for large buildings. In addition, patch sizes of 512×512 and 1024×1024 perform similarly for large buildings, but there is a slight loss of precision for small buildings compared to lower patch size configurations. Depending on the specific use case and the target building size, different patch sizes may be more appropriate. For example, if accuracy in small buildings is a priority, a patch size of 256×256 should be chosen. On the other hand, if the focus is on medium-sized and large buildings, using a larger patch size such as 512×512 or 1024×1024 may provide better accuracy.

VI. CONCLUSIONS AND FUTURE WORK

In this work, we present a clear set of guidelines for fairly comparing segmentation maps obtained at different spatial resolutions in the context of remote sensing. To begin, a reference ground truth must be selected at the highest resolution available. Then, the class probabilities should be rescaled to match the ground truth resolution. Thereafter, pixel-wise performance metrics can be computed and fairly compared. However, due to the imbalance between background and foreground classes, we suggest treating the problem as a multi-class semantic segmentation problem and macro-averaging the performance metrics. Additionally, to address the lack of detail in coarse spatial resolutions, some works have relaxed the performance metrics. We propose adjusting the tolerance buffer to account for resolution issues without being overly optimistic or pessimistic. Furthermore, we demonstrate how upper bounds for any performance metric can be estimated without training a model using only ground truth masks. Based on these bounds, the optimal GSD can be selected for a given use case to maximize the trade-off between computing resources required and accuracy achieved. Finally, we propose a methodology for computing region-based pixel-wise metrics by defining disjoint evaluation regions for each object. By following these guidelines, it is possible to compare segmentation maps at different spatial resolutions and gain a better understanding of how a model performs. We believe that the open-source Python library developed in this project will be useful for facilitating the comparison of research results in the remote sensing community.

In addition to the novelties proposed in this paper, the application of these guidelines to multiclass problems, such

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2024.3369310

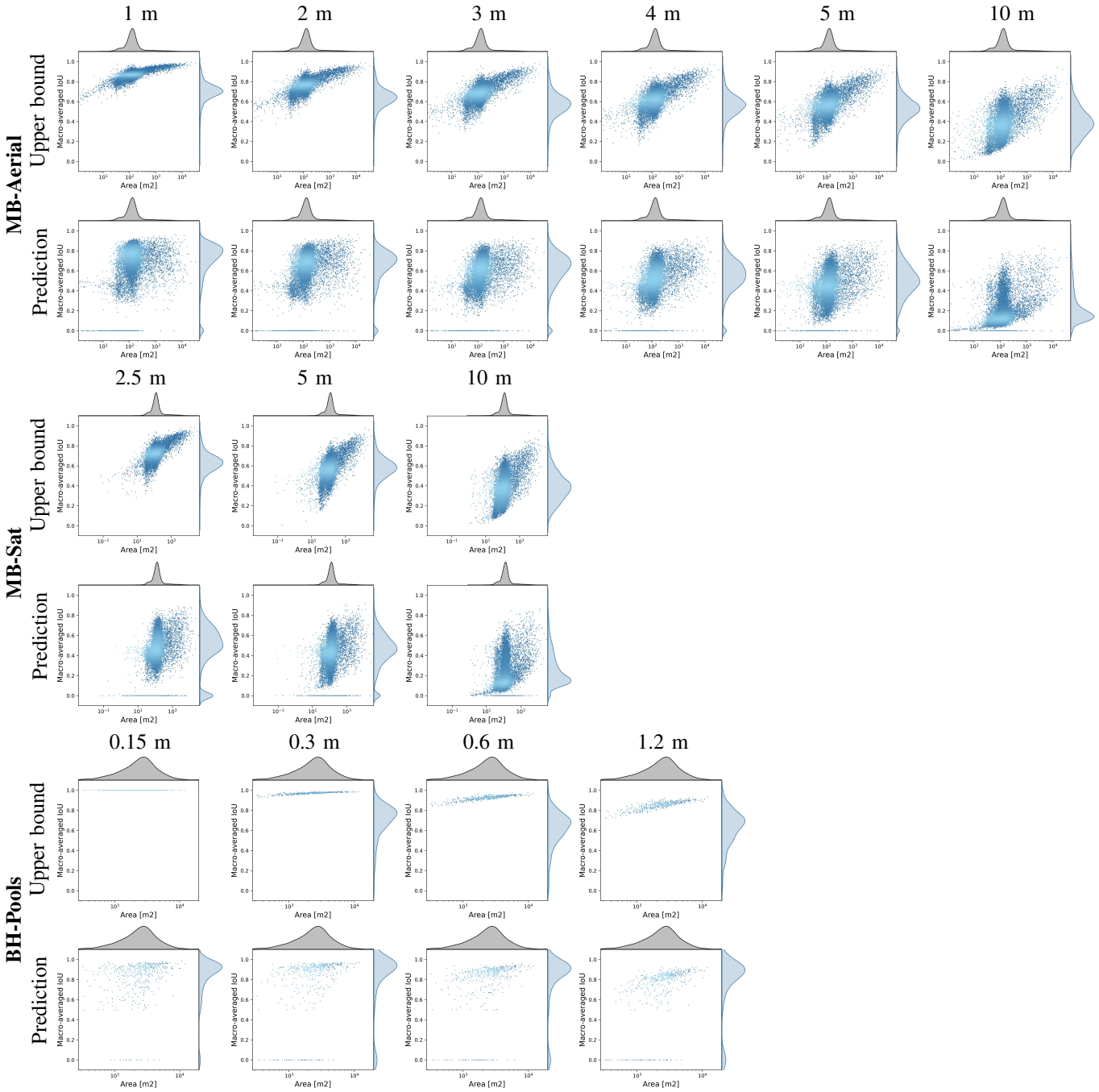


Fig. 9. Bi-variate probability density function of the macro-averaged IoU and the object area in m^2 for the MB-Aerial, MB-Sat, and BH-Pools datasets.

as land use and land cover classification, requires further investigation. Additionally, there is a need for a more in-depth study on how to handle more complex geometries, such as roads, in the process of defining evaluation regions for computing region-based pixel-wise metrics. Moreover, the inclusion of other datasets and sensors is necessary to enhance the reliability of the conclusions drawn. Finally, it would be interesting to further investigate whether the performance at a given spatial resolution can be predicted by considering a few upper bounds and actual performances to feed a machine learning model.

REFERENCES

[1] National Aeronautics and Space Administration, "Landsat programme," urlhttps://earthexplorer.usgs.gov/, Accessed 16.11.2022.

[2] European Spatial Agency, "Copernicus programme," urlhttps://www.copernicus.eu, Accessed 16.11.2022.

[3] K. Han, Y. Wang, H. Chen *et al.*, "A survey on vision transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2022.

[4] D. Khurana, A. Koli, K. Khatter *et al.*, "Natural language processing: state of the art, current trends and challenges," *Multimedia Tools and Applications*, Jul 2022.

[5] H. Ouchra and A. Belangour, "Satellite image classification methods and techniques: A survey," in *2021 IEEE International Conference on Imaging Systems and Techniques (IST)*, 2021, pp. 1–6.

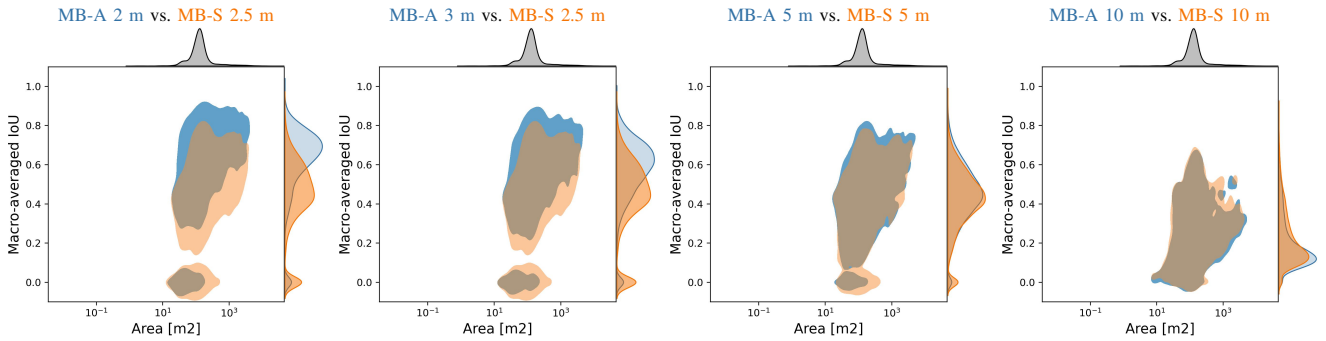


Fig. 10. Bi-variate probability density function of the macro-averaged IoU and the object area in m^2 for the MB-Aerial (MB-A) in blue and MB-Sat (MB-S) in orange datasets.

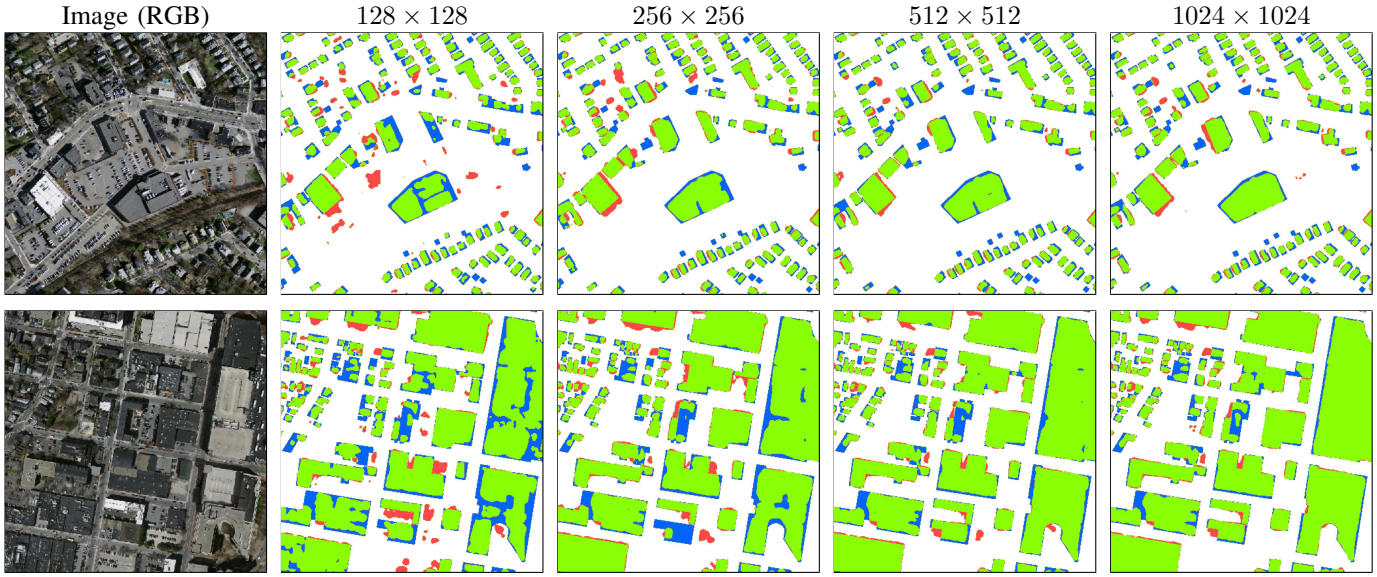


Fig. 11. Visual comparison between predicted segmentation maps for each patch size configuration. Predicted segmentation maps are presented in terms of TP in green, FN in blue, FP in red, and TN in white.

[6] P. Anilkumar and P. Venugopal, "A survey on semantic segmentation of aerial images using deep learning techniques," in *2021 Innovations in Power and Advanced Computing Technologies (i-PACT)*, 2021, pp. 1–7.

[7] Z. Zheng, L. Lei, H. Sun *et al.*, "A review of remote sensing image object detection algorithms based on deep learning," in *2020 IEEE 5th International Conference on Image, Vision and Computing (ICIVC)*, 2020, pp. 34–43.

[8] D. Wen, X. Huang, F. Bovolo *et al.*, "Change detection from very-high-spatial-resolution optical remote sensing images: Methods, applications, and future directions," *IEEE Geoscience and Remote Sensing Magazine*, vol. 9, no. 4, pp. 68–101, 2021.

[9] S. Lobry, D. Marcos, J. Murray *et al.*, "Rsvqa: Visual question answering for remote sensing data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 12, pp. 8555–8566, 2020.

[10] G. Cheng, X. Xie, J. Han *et al.*, "Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 3735–3756, 2020.

[11] S. Liang and J. Wang, "Chapter 1 - a systematic view of remote sensing," in *Advanced Remote Sensing (Second Edition)*. Academic Press, 2020, pp. 1–57.

[12] R. Schwartz, J. Dodge, N. A. Smith *et al.*, "Green ai," 2019.

[13] F. M. Senalp and M. Ceylan, "Effects of the deep learning-based super-resolution method on thermal image classification applications," *Multimedia Tools and Applications*, vol. 81, no. 7, pp. 9313–9330, Mar. 2022.

[14] L. Zhou, G. Chen, M. Feng, and A. Knoll, "Improving low-resolution image classification by super-resolution with enhancing high-frequency content," in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 1972–1978.

[15] A. V. Etten, "Satellite imagery multiscale rapid detection with windowed networks," *CoRR*, vol. abs/1809.09978, 2018.

[16] J. Shermeyer and A. Van Etten, "The Effects of Super-Resolution on Object Detection Performance in Satellite Imagery," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Long Beach, CA, USA: IEEE, Jun. 2019, pp. 1432–1441.

[17] L. Zhang, R. Dong, S. Yuan, and H. Fu, "Srbuildingseg-E ^{2} : An Integrated Model for End-to-End Higher-Resolution Building Extraction," in *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*. Kuala Lumpur, Malaysia: IEEE, Jul. 2022, pp. 1356–1359.

[18] P. Xu, H. Tang, J. Ge *et al.*, "Espc_nasunet: An end-to-end super-resolution semantic segmentation network for mapping buildings from remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 5421–5435, 2021.

[19] L. Wang, D. Li, Y. Zhu *et al.*, "Dual super-resolution learning for semantic segmentation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3773–3782.

[20] S. Hao, W. Wang, Y. Ye, E. Li, and L. Bruzzone, "A deep network architecture for super-resolution-aided hyperspectral image classification with classwise loss," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 8, pp. 4650–4663, 2018.

[21] J. Xie, L. Fang, B. Zhang, J. Chanussot, and S. Li, "Super resolution guided deep network for land cover classification from remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2022.

[22] C. Ayala, R. Sesma, C. Aranda, and M. Galar, "A deep learning approach to an enhanced building footprint and road detection in high-resolution

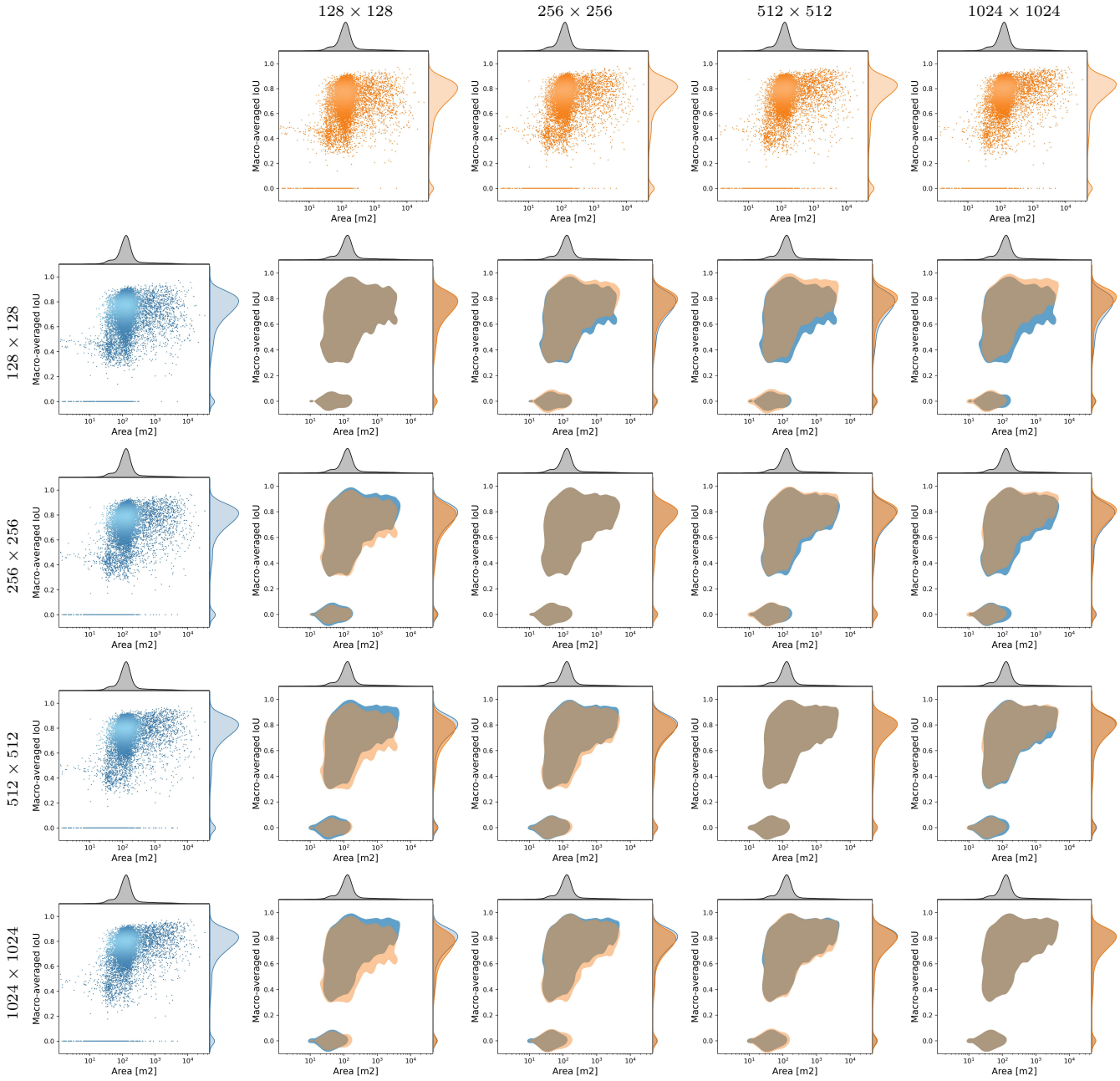


Fig. 12. Visual comparison between bi-variate probability density function of the macro-averaged IoU and the object area for multiple patch sizes configurations for the MB-Aerial dataset.

satellite imagery," *Remote Sensing*, vol. 13, no. 16, 2021.

[23] Z. Guo, G. Wu, X. Song *et al.*, "Super-resolution integrated building semantic segmentation for multi-source remote sensing imagery," *IEEE Access*, vol. 7, pp. 99 381–99 397, 2019.

[24] S. Lei, Z. Shi, X. Wu *et al.*, "Simultaneous super-resolution and segmentation for remote sensing images," in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, 2019, pp. 3121–3124.

[25] V. Mnih, "Machine learning for aerial image labeling," Ph.D. dissertation, University of Toronto, 2013.

[26] Y. Feng, C. Yang, and M. Sester, "Multi-scale building maps from aerial imagery," *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2020.

[27] E. Fernandes, P. Wildemberg, and J. dos Santos, "Water tanks and swimming pools detection in satellite images: Exploiting shallow and deep-based strategies," in *Anais do XVI Workshop de Visão Computacional*. Porto Alegre, RS, Brasil: SBC, 2020, pp. 117–122.

[28] S. V. Stehman and R. L. Czaplewski, "Design and analysis for thematic map accuracy assessment: Fundamental principles," *Remote Sensing of Environment*, vol. 64, no. 3, pp. 331–344, 1998.

[29] S. V. Stehman and J. D. Wickham, "Pixels, blocks of pixels, and polygons: Choosing a spatial unit for thematic accuracy assessment," *Remote Sensing of Environment*, vol. 115, no. 12, pp. 3044–3055, 2011.

[30] C. Ayala, C. Aranda, and M. Galar, "Pushing the limits of sentinel-2 for building footprint extraction," in *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, 2022, pp. 322–325.

[31] ——, "Sub-pixel width road network extraction using sentinel-2 imagery," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, 2021, pp. 2174–2177.

[32] M. B. Pereira and J. A. dos Santos, "How effective is super-resolution to improve dense labelling of coarse resolution imagery?" in *2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2019, pp. 202–209.

[33] M. Haklay and P. Weber, "Openstreetmap: User-generated street maps," *IEEE Pervasive Computing*, 2008.

[34] H. S. Cunha, B. S. Sclauser, P. F. Wildemberg *et al.*, "Water tank and swimming pool detection based on remote sensing and deep learning: Relationship with socioeconomic level and applications in dengue control," *PLoS One*, vol. 16, no. 12, p. e0258681, Dec. 2021.

[35] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2015.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016.

[37] F. H. Wagner *et al.*, "U-net-id, an instance segmentation model for building extraction from satellite images—case study in the joanópolis city, brazil," *Remote Sensing*, vol. 12, no. 10, 2020.

[38] Y. Lin, D. Xu, N. Wang, Z. Shi, and Q. Chen, "Road extraction from very-high-resolution remote sensing images via a nested se-deeplab model," *Remote Sensing*, vol. 12, no. 18, 2020.

[39] Taghanaki *et al.*, "Combo loss: Handling input and output imbalance in multi-organ segmentation," *Computerized Medical Imaging and Graphics*, vol. 75, 05 2019.

[40] Ma Yi-de, Liu Qing, and Qian Zhi-bai, "Automated image segmentation using improved pcnn model based on cross-entropy," in *Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, 2004.*, 2004, pp. 743–746.

[41] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2017.

[42] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.

[43] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS Journal of Photogrammetry and Remote Sensing*, 2020.

[44] R. Wu, S. Yan, Y. Shan, Q. Dang, and G. Sun, "Deep image: Scaling up image recognition," 01 2015.

[45] M. B. Pereira and J. A. d. Santos, "An end-to-end framework for low-resolution remote sensing semantic segmentation," in *2020 IEEE Latin American GRSS ISPRS Remote Sensing Conference (LAGIRS)*, 2020, pp. 6–11.

[46] A. E. Maxwell, T. A. Warner, and L. A. Guillén, "Accuracy assessment in convolutional neural network-based deep learning remote sensing studies—part 1: Literature review," *Remote Sensing*, vol. 13, no. 13, 2021.

[47] A. A. Taha and A. Hanbury, "Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool," *BMC Medical Imaging*, vol. 15, no. 1, p. 29, Aug 2015.

[48] L. Bruzzone and C. Persello, "A novel protocol for accuracy assessment in classification of very high resolution multispectral and sar images," in *IGARSS 2008 - 2008 IEEE International Geoscience and Remote Sensing Symposium*, vol. 2, 2008, pp. II–265–II–268.

[49] D. L. Olson and D. Delen, *Advanced Data Mining Techniques*, 1st ed. Springer Publishing Company, Incorporated, 2008.

[50] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[51] P. Jaccard, "The distribution of the flora in the alpine zone.1," *New Phytologist*, vol. 11, no. 2, pp. 37–50, 1912.

[52] C. Ferri, J. Hernández-Orallo, and R. Modroiu, "An experimental comparison of performance measures for classification," *Pattern Recognition Letters*, vol. 30, no. 1, pp. 27–38, 2009.

[53] D. M. W. Powers, "Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.

[54] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation," in *Computer Vision – ECCV 2018*. Cham: Springer International Publishing, 2018, vol. 11211, pp. 833–851.

[55] S. Abadal, L. Salgueiro, J. Marcello *et al.*, "A dual network for super-resolution and semantic segmentation of sentinel-2 imagery," *Remote Sensing*, vol. 13, no. 22, 2021.

[56] D. Chicco and G. Jurman, "The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, p. 6, Jan 2020.

[57] D. M. W. Powers, "What the f-measure doesn't measure: Features, flaws, fallacies and fixes," 2015.

[58] N. Japkowicz, "Assessment Metrics for Imbalanced Learning," in *Imbalanced Learning*. Hoboken, NJ, USA: John Wiley & Sons, Inc., Jun. 2013, pp. 187–206.

[59] W. Zheng and M. Jin, "The Effects of Class Imbalance and Training Data Size on Classifier Learning: An Empirical Study," *SN Computer Science*, vol. 1, no. 2, p. 71, Mar. 2020.

[60] A. Borji and S. M. Iranmanesh, "Empirical Upper Bound in Object Detection and More," Dec. 2019, arXiv:1911.12451 [cs].

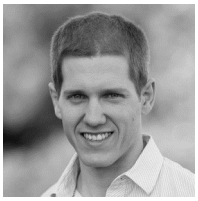
[61] N. Kumar, R. Verma, S. Sharma, S. Bhargava, A. Vahadane, and A. Sethi, "A dataset and a technique for generalized nuclear segmentation for computational pathology," *IEEE Transactions on Medical Imaging*, vol. 36, no. 7, pp. 1550–1560, 2017.

[62] D. Dan, "Image segmentation using Voronoi diagram," in *Eighth International Conference on Digital Image Processing (ICDIP 2016)*, vol. 10033, International Society for Optics and Photonics. SPIE, 2016, p. 100331P.

[63] P. Arbelaez and L. Cohen, "Constrained image segmentation from hierarchical boundaries," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

Christian Ayala received the BSc degree in computer science from the Public University of Navarre (UPNA), Navarre, Spain, in 2018, and the MSc degree in general computer science from the Public University of Navarre (UPNA), Navarre, Spain, in 2020. He is currently pursuing the Ph.D. degree in artificial intelligence with Tracasa Instrumental S.L., Navarre, Spain, and the Public University of Navarre, Navarre, Spain. Currently, he is involved in deep learning and process automation projects at Tracasa Instrumental's R+D+i department.

Carlos Aranda received the BSc degree in electronic engineering from the Zaragoza University, Spain, in 1997 and the MSc degree in technology and computer sciences in the University Oberta de Catalunya, Spain in 2007. He also has a MSc degree in Finance from the IEB Stock Market Studies Institute, Madrid, Spain in 2017. He has worked during his career in several consulting companies, covering positions of analyst and head of technical office and since 2018 he has held the position of Director of R+D+i (CINO) in Tracasa, Spain.



Mikel Galar (M14) received the MSc and PhD degrees in Computer Science in 2009 and 2012, both from the Public University of Navarre, Pamplona, Spain. He is currently an associate professor at the Department of Statistics, Computer Science and Mathematics at the Public University of Navarre. He is the author of 50 published original articles in international journals and more than 90 contributions to conferences. He is also reviewer of more than 35 international journals. He is a co-author of a book on imbalanced datasets and he has acted as guest editor for special issues in journals such as Cognitive Computation. His research interests are machine learning, deep learning, ensemble learning, evolutionary algorithms, fuzzy systems and big data. He is involved in the applications of these techniques to the industry, healthcare and remote sensing. He is an external scientific advisor at Tracasa Instrumental, S.L. and Co-founder of Neuraptic AI. He is a member of the IEEE, the European Society for Fuzzy Logic and Technology (EUSFLAT) and the Spanish Association of Artificial Intelligence (AEPIA). He has received the extraordinary prize for his PhD thesis from the Public University of Navarre and the 2013 IEEE Transactions on Fuzzy System Outstanding Paper Award for the paper "A New Approach to Interval-Valued Choquet Integrals and the Problem of Ordering in Interval-Valued Fuzzy Set Applications" (bestowed in 2016).