COMPARATIVE EVALUATION OF VALIDATION APPROACHES FOR WATER BODY MAPPING USING PLANETSCOPE IMAGERY

Adrian Yordanov

University of Architecture, Civil Engineering and Geodesy
e-mail: a.yordanov_fgs@uacg.bg

Keywords: *PlanetScope; water-body mapping; spatial cross-validation; shoreline detection; model generalization*

Abstract: *This paper uses one PlanetScope dataset (14 four-band acquisitions from 2023 over one study area, with 31 water bodies digitised manually on the PlanetScope images and checked against the nearest-date high-resolution images in Google Earth) and a random-forest classifier with eight fold-selected features. With the classifier held fixed, 4 different validation strategies were compared – group-shuffle and leave-one-out validation, grouped by scene or by water body – on a shared set of sampled pixels. Scored on identical samples, holding out whole scenes and holding out single water bodies gave IoU values within one point of each other (94.1% and 95.0%). Reported accuracy changed far more with the averaging method and the choice of land samples. Under leave-one-waterbody-out, pooled IoU was 95.0%, but the per-object mean was 73.8% (median 85.7%). Per-object IoU rose with the number of water samples (Spearman ρ = 0.85), the three smallest water bodies scored below 5%, and one water body held 72% of the water samples. The same scene-held-out predictions labelled 3.9% of land pixels as water across whole scenes, but 19.6% near shorelines. This paper aims to show that reporting the grouping unit, the averaging method and the land-sampling scheme, and giving both pooled and per-object scores, provides important information about the models. All data come from one region, so transfer to new areas remains untested.*

ОЦЕНКА НА РАЗЛИЧНИ ПОДХОДИ ЗА ВАЛИДАЦИЯ ПРИ КАРТИРАНЕ НА ВОДНИ ТЕЛА ОТ ИЗОБРАЖЕНИЯ PLANETSCOPE

Адриан Йорданов

Университет по Архитектура, Строителство и Геодезия
e-mail: a.yordanov_fgs@uacg.bg

Ключови думи: *PlanetScope; water-body mapping; spatial cross-validation; shoreline detection; model generalization*

Резюме: *В настоящата статия е използван набор от 14 четириканални изображения PlanetScope от 2023 г. за район между Плевен и Севлиево с 31 водни тела, векторизирани ръчно върху изображенията PlanetScope и проверени спрямо най-близките по дата изображения с висока разделителна способност в Google Earth. Класификацията е извършена с алгоритъма „Случайна гора“ обучен с 8 признака избирани независимо за всяко обучение. При един и същ класификатор са сравнени 4 различни метода за валидация на модела – случаен групов набор и изключване на един елемент, с групиране по изображение и по водно тяло – върху общ набор от пикселни извадки. При оценяване върху идентични извадки, изключването на цели изображения и на единични водни тела довежда до промяна в стойностите на коефициента на Джакард под 1 процентен пункт (94.1% и 95.0%). Отчетената точност се променя значително в зависимост от метода на осредняване и от избора на извадки от сушата. При изключване на единично водно тяло, обединената стойност на коефициента на Джакард е 95.0%, а средната стойност, по обекти, е 73.8% (медианата е 85.7%). Отчетена е и висока степен на корелация между нарастване на стойностите на коефициента на Джакард пропорционално на нарастването на броя извадки за дадено водно тяло (ρ на Спиърман от 0.85). Трите най-малки водни тела имат стойности под 5%, а едно от водните тела заема 72% от общия брой извадки. При изключване на цели изображения, 3.9% от пикселните извадки за сушата са класифицирани като вода в рамките на цялото изображение, докато в районите около бреговата линия, процентът нараства на 19.6%. Настоящата статия цели да акцентира, че посочването на единицата за групиране, метода на осредняване и начина на формиране на извадките от сушата, заедно с представянето на обединени и на осреднени по обекти стойности, дава по-пълна представа за качеството на модела. Всички данни са от един район, поради което устойчивостта на модела към други територии не е проверена.*

1. Introduction

Reviews of remote-sensing accuracy assessment show that sampling designs and metric definitions are often reported incompletely [Morales-Barquero et al., 2019; Maxwell et al., 2021], yet these choices can change the reported accuracy of a classifier. Good practice for the reference sample, the response and the analysis is well documented [Stehman and Foody, 2019]. A single headline number is still often quoted without the details a reader needs to interpret it.

Neighboring pixels in a satellite image are spatially correlated. If pixels are split at random into training and test sets, neighbors end up on both sides and the measured accuracy is optimistic [Roberts et al., 2017; Ploton et al., 2020; Karasiak et al., 2022]. Grouped validation avoids this by holding out whole units. For water-body mapping the unit can be an image or a water body, and it is not obvious which is the better choice. Two more choices change the reported score: which samples are scored, land samples in particular, and how the scores of units of different size are averaged. Errors concentrate along shorelines, the zone that remote-sensing contour mapping of lakes targets [Yordanov and Atanasova-Evdenova, 2025]. Small water bodies are also routinely mapped from PlanetScope imagery [Mullen et al., 2023], so these choices are likely to matter for this kind of data.

The paper doesn't aim to rank validation designs. With the classifier and the samples held fixed, the question is how much the reported IoU depends on three things: the grouping unit, the land samples in the scored set, and the averaging method (pooled or per unit). On this dataset the grouping unit changes the pooled IoU by less than one point and the choice of scored land samples by about one point, whereas averaging per water body changes it by more than twenty. These results lead to a short list of items that should accompany any reported accuracy.

2. Data and study area

The study area lies between Pleven and Sevlievo, Bulgaria. It contains 31 water bodies: one reservoir and 30 ponds, of which 5 are large (above 50,000 m²), 6 medium (10,000–50,000 m²), 10 small (5,000–10,000 m²) and 9 very small (below 5,000 m²) (Fig. 1). The data are 14 four-band (blue, green, red, near-infrared) PlanetScope images from 2023, with a pixel size of 3 m, dated 2 and 19 January, 19 February, 13 March, 13 April, 7 May, 20 June, 21 July, 20 August, 12 September, 15 October, 14 November, 4 December and 28 December. All are surface reflectance products harmonized with Sentinel-2.

The reference data were vectorized manually on the PlanetScope images and checked against the nearest-date high-resolution images in Google Earth. Polygons are drawn separately for every scene, which accounts for changes in size between dates and for seasonal water bodies. Each water body keeps one persistent integer ID across scenes. Twenty-nine water bodies are present in all scenes; two are strongly seasonal and almost absent in some. An ID contributes to a scene only where it has sampled pixels, so some water bodies appear in only one or two scenes.



Fig. 1. Study area with the 31 water bodies

3. Methodology

3.1. Classifier

The classifier is a Random Forest with 500 trees, the square root of the number of features tried at each split, and the Gini criterion. It used eight features in every experiment, selected independently inside each training fold from the raw bands and from derived features (differences, normalized differences, ratios and proportions of band pairs), so no information from the held-out units reaches feature selection. In the 14 leave-one-scene-out folds seven features were chosen every time; the eighth was the green–blue difference in 10 folds and the green–red difference in 4. All features were selected based on ranking using average precision, F1-score and Fisher score and a correlation filtering between every possible feature and the already selected ones. The correlation threshold for choosing a new feature was set to 0.95. The classifier is the same in all designs. Only the way training and test data are separated changes.

3.2. Samples and classes

Reference polygons were rasterized and split into four classes with morphological operations. A 6 m buffer, which is 2 pixels at 3 m (converted with the raster transform), separates the classes. Interior water is the reference water shrunk by 2 pixels. Water boundary is the remaining 2-pixel ring of reference water. Land boundary, or shoreline land, is the 2-pixel ring of land found by dilating the reference water by 2 pixels and removing the two water classes. Land interior is all other land. Every scene supplied a fixed quota: 33,000 interior-water, 5,000 water-boundary, 5,000 land-boundary and 33,000 land-interior pixels. The 14 scenes give 1,064,000 samples with a nominal water-to-land ratio of 1:1. Pixels were drawn at random with a minimum spacing of 4 pixels, relaxed to 2 and then to 0 pixels when a quota could not be filled. All samples are kept in one table, one row per sample, with scene ID, water-body ID, and pixel and global coordinates. Land-interior samples carry water-body ID 0.

3.3. Validation designs

Four designs were compared on the same table of samples: leave-one-scene-out (14 folds, 988,000 training samples each), leave-one-water-body-out (31 folds), and group-shuffle by scene and by water body (10 random splits each, with 4 test scenes or 3 test water bodies per split). Training and test groups are always disjoint. The designs leak different information. With a scene held out, the same water bodies remain in training through the other dates. With a water body held out, the same dates remain in training through the other water bodies. The held-one-out designs use all datasets except one and loop until all individual datasets were used once as validation. For the image one-held-out design, 1 dataset is equal to 1 image while for the waterbody design, 1 dataset is 1 waterbody. For the group-shuffle design, a number of iterations is defined (in this study – 10) and a number of held-out datasets as well (for images – 3 and for waterbodies – 4). Each iteration consists of choosing the defined number of held-out datasets at random and using them for validation excluding them from training. Some of the held-out datasets can appear in more than 1 iteration.

The scored rows differ between designs. A leave-one-scene-out fold is scored on all 76,000 samples of the scene. A leave-one-water-body-out fold is scored on the samples of the held-out water body: its water samples and the nearby land-boundary samples attributed to it. Land-interior samples carry ID 0, are never held out and so are never scored in this design. To compare the designs on identical samples, leave-one-scene-out was also scored on these shared rows, that is, on all water samples and the shoreline land samples (532,000 water and 70,000 land samples over all folds).

3.4. Metrics

Water is the positive class. Accuracy is reported as the intersection over union (IoU, Jaccard index):

(1) $IoU = TP / (TP + FP + FN),$

where TP, FP and FN are the numbers of true positive, false positive and false negative samples. The pooled IoU sums TP, FP and FN over all folds before applying (1). The per-unit IoU applies (1) to the held-out samples of one scene or one water body, and is summarized by the unweighted mean, median and minimum over units. Per-unit scores exist for leave-one-scene-out (14 scenes) and leave-one-water-body-out (31 water bodies); the group-shuffle designs are reported as pooled values only. For water bodies, the Spearman correlation between per-object IoU and the number of water samples is also given. The false-water rate is the share of land samples predicted as water in the leave-one-scene-out predictions. It was calculated over all land samples, over land-interior samples and over shoreline (land-boundary) samples.

4. Results

4.1. Grouping unit and scored rows

On all 76,000 samples of each scene, leave-one-scene-out gave a pooled IoU of 93.0% (Table 1). On the shared rows it gave 94.1%, against 95.0% for leave-one-water-body-out, a gap of less than one point. Dropping the land-interior samples from the scored set moves the leave-one-scene-out score by 1.2 points, as much as switching the grouping unit does. Group-shuffle gave 92.8% for scenes with 1.05% standard deviation (SD) and 89.0% for water bodies with 2.94% SD. The six-point gap between group-shuffle by water body and leave-one-water-body-out has a likely cause. Each split holds out only three of the 31 water bodies, and the largest one, which holds most of the water samples and scores well, was not among them in this run. The pooled score then rests on smaller water bodies, which score lower.

Table 1. IoU (%) of the two leave-one-out designs. Per-unit values are on the shared rows.

Design	Units	Pooled, all rows	Pooled, shared rows	Mean per unit	Median per unit	Min per unit
Leave-one-scene-out	14 scenes	93.0	94.1	94.1	94.3	90.9
Leave-one-water-body-out	31 water bodies	n/a	95.0	73.8	85.7	0.0

4.2. Pooled and per-unit scores

For scenes, pooling and averaging agree: the pooled IoU is 94.1% and the mean per-scene IoU is also 94.1% (median 94.3%, range 90.9–96.7%). For water bodies they do not. The pooled IoU is 94.96%, the mean of the 31 per-object IoUs is 73.8% and the median is 85.7%. The reason is unit size. Every scene contributes the same quota of 38,000 water samples, while the water bodies range from 35 to about 381,000 water samples, and the largest holds 72% of them. The pooled value mostly reflects that one water body. Per-object IoU rises with the number of water samples (Spearman ρ = 0.85, Fig. 2), and the three smallest water bodies score below 5%. Averaging therefore matters only when units differ strongly in size, and the equal quota per scene hides that difference for the scene unit.

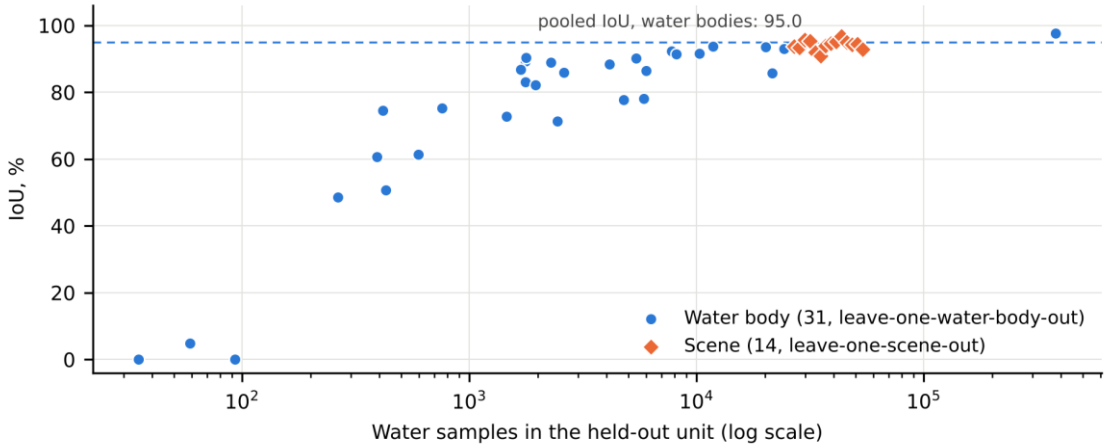


Fig. 2. IoU of each held-out unit (shared rows) against its number of water samples. Dashed line: pooled IoU of leave-one-water-body-out. Scene points are shifted horizontally for visibility.

4.3. Land samples

In the leave-one-scene-out predictions, 3.9% of all land samples were labeled as water (20,535 of 532,000). The rate was 1.5% for land-interior samples (6,789 of 462,000) and 19.6% for shoreline samples (13,746 of 70,000). Shoreline samples are 13% of the land samples but produce 67% of the false positives. Leave-one-water-body-out gives a similar shoreline rate of 19.3% (13,495 of 70,000), so this error does not depend on the grouping unit. A score computed mostly from distant land samples hides it. The scene of 13th of April shows how much: its IoU is 85.1% on all rows and 92.0% on the shared rows. The shoreline ring is 2 pixels (6 m) wide, about the size of the positional error of manual digitizing and of a mixed pixel, so part of the 19.6% may be label uncertainty and not model error. This was not tested.

5. Discussion

Three choices that are rarely reported changed the IoU by very different amounts here. The grouping unit moved it by less than one point (94.1% and 95.0%), the scored land samples by about one point (93.0% and 94.1%), and averaging over water bodies by more than twenty (94.96% and 73.8%). Neither the pooled nor the per-unit score is wrong; they answer different questions. The pooled score weights every sample equally and so reflects the largest water body. The per-object mean weights every water body equally and so reflects the small ones, where the model is least accurate. Readers cannot tell which was used unless the paper says so. Whether the two differ depends on how unequal the units are, and for scenes that is set by the sampling quota, not by the landscape. Details beyond the unit matter too: group-shuffle by water body differed from leave-one-water-body-out by six points although both hold out water bodies.

At least four things should be reported: the grouping unit and the number of folds or splits; the rows that were scored, with their class composition (here 532,000 water against 70,000 land samples); pooled and per-unit scores side by side, with the number and size range of the units; and separate scores for the shoreline zone. This study has clear limits: one area, one year, one classifier and manual reference polygons. The group-shuffle designs were not scored per unit. Transfer to other areas was not tested, and a test of that kind should first assess the area of applicability of the model [Meyer and Pebesma, 2021].

6. Conclusion

On PlanetScope imagery of one region, holding out scenes or single water bodies gave pooled IoU values within one point of each other (94.1% and 95.0%), and adding land-interior samples to the scored set changed the leave-one-scene-out score by a similar amount (93.0%). Averaging per water body lowered the reported accuracy from 94.96% to 73.8%, because the water bodies differ so much in size. Shoreline land had a false-water rate of 19.6%, against 1.5% for interior land. Reports of water-body mapping accuracy should state the grouping unit, the scored rows and the averaging method, and give both pooled and per-unit scores.

7. Acknowledgements

This research was financially supported by the Research and Design Centre at the University of Architecture, Civil Engineering and Geodesy (UACEG) under Contracts No. D-173/2026.

References:

1. Karasiak, N. et al. Spatial dependence between training and test sets: another pitfall of classification accuracy assessment in remote sensing. Machine Learning, 111, 2022, pp. 2715–2740.
2. Maxwell, A. E. et al. Accuracy assessment in convolutional neural network-based deep learning remote sensing studies – Part 1: Literature review. Remote Sensing, 13(13), 2021, 2450.
3. Meyer, H., E. Pebesma. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. Methods in Ecology and Evolution, 12(9), 2021, pp. 1620–1633.
4. Morales-Barquero, L. et al. Trends in remote sensing accuracy assessment approaches in the context of natural resources. Remote Sensing, 11(19), 2019, 2305.
5. Mullen, A. L. et al. Using high-resolution satellite imagery and deep learning to track dynamic seasonality in small water bodies. Geophysical Research Letters, 50, 2023, e2022GL102327.
6. Ploton, P. et al. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. Nature Communications, 11, 2020, 4540.
7. Roberts, D. R. et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography, 40(8), 2017, pp. 913–929.
8. Stehman, S. V., G. M. Foody. Key issues in rigorous accuracy assessment of land cover products. Remote Sensing of Environment, 231, 2019, 111199.
9. Yordanov, A., T. Atanasova-Evdenova. Remote sensing-based contour mapping of Burgas lake. София: УАСГ, 2025, pp. 843–864.