DOI: 10.15201/hungeobull.74.4.1 Hungarian Geographical Bulletin **74** 2025 (4) 339–357.

Lithological mapping with pseudo-labelling: Promise or overestimation in data-scarce settings?

Szilárd SZABÓ[1], Abdelmajeed A. ELRASHEED[2], Lilla KOVÁCS[3], Imre J. HOLB[4], Szilárd B. LIKÓ[5] and Dávid ABRIHA[6]

Abstract

Reference data are the most crucial points in model building. In geoscience, a scarcity of sufficient reference data is common. Pseudo-labelling (PL), i.e. incorporating high-probability data in the model-building process, offers a potential solution. We aimed to reveal the efficiency of PL in lithological mapping in a vegetation-free arid region of Sudan. Multiple Adaptive Regression Splines (MARS) and Random Forest (RF) were used to classify a Landsat 9 image. Reference data were collected during fieldwork and through visual interpretation. Image processing yielded classified maps with associated probability layers, from which 1000 additional traditional samples (PL data) were extracted at a 95 percent probability. A detailed accuracy assessment was conducted, and accuracy measures were evaluated using statistical analysis and visual inspection. MARS was found to be an ambiguous classifier because the probability was too optimistic related to the overall accuracy (OA) (81% of samples had above 99% probability, OA = 98.2%) compared to RF (21% above 99%, OA = 98.1%); that is, despite the high probability, the accuracy improvement was only 0.1 percent. At the class level, the correlation between probability and the F1-score was low (0.21%). The original and PL-based models resulted in different maps with improved accuracy, although the new model version showed lower probability values for both the classifiers. Visual inspection proved essential for better insights into the spatial patterns: expert knowledge is crucial for controlling the occurrence of rock types and identifying false classifications. The main finding is that probability should be handled carefully, as it does not guarantee high model performance in classification, although the PL approach can lead to more reliable maps.

Keywords: Multiple Adaptive Regression Splines (MARS), Random Forest (RF), self-training, probability, data augmentation

Received June 2025, accepted November 2025.

Introduction

Lithological mapping is the process of identifying different lithological units (rocks) within a given area and represents a fundamental step in most geological investigations, including mineral exploration, groundwater assessment, petroleum studies, natural resource

[1] Department of Physical Geography and Geoinformatics, Faculty of Sciences and Technology, University of Debrecen. National Laboratory for Water Science and Water Safety, University of Debrecen. Egyetem tér 1, H-4032 Debrecen, Hungary. Corresponding author's e-mail: szabo.szilard@science.unideb.hu

[2] Department of Physical Geography and Geoinformatics, Faculty of Sciences and Technology, University of Debrecen. Egyetem tér 1, H-4032 Debrecen, Hungary. Department of Geology, University of Khartoum. 1115 Khartoum, Sudan. E-mails: aaelrasheed@uofk.edu, ali.abdelmajeed.adam.elrasheed@science.unideb.hu

[3] Doctoral School of Geosciences, Faculty of Sciences and Technology, University of Debrecen. Egyetem tér 1, H-4032 Debrecen, Hungary. E-mail: kovacs.lilla@science.unideb.hu

[4] Institute of Horticulture, University of Debrecen. Böszörményi út 138, H-4032 Debrecen, Hungary. Plant Protection Institute, Centre for Agricultural Research, HUN-REN. Herman Ottó út 15, H-1022 Budapest, Hungary. E-mails: holb.imre@agr.unideb.hu, holbimre@gmail.com

[5] Freelancer. E-mail: liko.szilard.balazs@gmail.com

[6] Department of Physical Geography and Geoinformatics, Faculty of Sciences and Technology, University of Debrecen. Egyetem tér 1, H-4032 Debrecen, Hungary. E-mail: abriha.david@science.unideb.hu

evaluation, and environmental management (Amusuk, D.J. *et al.* 2016; Abrams, M. and Yamaguchi, Y. 2019). Consequently, obtaining an accurate lithological map is crucial for mineral exploration. Traditionally, lithological mapping is executed mainly through interpretation of aerial photographs, followed by extensive field-based investigations in which geologists collect samples, register observations, measure geological structures (such as dip and strike), identify mineral composition and textures, or interpret field relationships (Szaniawska, L. 2018). However, this traditional mapping approach can be challenging because it requires skilled field geologists, is time-consuming, and requires a suitable budget to cover costs, especially in harsh environments and inaccessible or isolated places. These difficulties have recently decreased through the integration of cutting-edge technologies, such as remote sensing and machine learning, in the lithological mapping process (Bachri, I. *et al.* 2019; Abdelkareem, M. *et al.* 2021; Shirmard, H. *et al.* 2022).

Geological features can be identified using extensive datasets of remote sensing technology (El-Omairi, M.A. and Garouani, A.E. 2023). Valuable insights can be drawn from the geographic data obtained by sensors, based on the distinctive properties of the local area (Nair, P. *et al.* 2023). Machine learning offers an efficient way of processing remote sensing data. After the training phase, algorithms can identify rock types, faults, or mineral deposits if the spectral resolution of the images makes it possible. These algorithms can handle large datasets, allowing the discovery of plausible patterns in complex geological datasets (Han, W. *et al.* 2023). However, machine learning requires high-quality training data to build a reliable model. Additionally, independent testing data (known as reference data) are essential for validating a model's performance (James, G. *et al.* 2013).

Reference data is the most crucial constituent of all models. Training subsets are used for model building, while testing subsets are used to assess accuracy. In remote sensing, we may have millions of pixels (i.e. data points), which may give the impression that the delineation of reference data is easy, however, this is not true for all tasks. For example, in land cover mapping, the traditional approach is to classify surface objects into simple classes such as forests, grasslands, and water bodies. These classes can include thousands of pixels as reference data because simple visual interpretations of the images can provide sufficient information. However, when the aim is a more specific problem, such as identifying plant or tree species, detecting plant diseases, or classifying roof types, reference data collection requires field observations and/or ground measurements, which makes this step labour-intensive and time-consuming. Lithological mapping faces the same problem: Field observations are essential for a reliable reference dataset. If the number of labelled instances is insufficient, the model cannot be adequately trained because of the insufficient size of the reference dataset.

The question of how much reference data is needed depends on the algorithms and is widely discussed in the literature. While there are basic rules, generally, the more the reference data, the better. Studies have shown that accurate, balanced, and large training datasets are often more important than the choice of algorithms in terms of the output (Li, C. *et al.* 2014; Maxwell, A.E. *et al.* 2018; Collins, L. *et al*. 2020). There is no universally accepted minimum number of training data points. However, Foody, G.M. (2009) defined a method for calculating the minimum amount of testing data based on the desired overall accuracy, acceptable standard error, and classification error. For example, for four categories with an 85 percent target accuracy and a standard error of 0.02 (2%), at least 1275 testing data samples are required, which is approximately 320 per class. Class imbalance also can be a crucial point when the target class or some classes are underrepresented causing false accuracy metrics (Luque, A. *et al.* 2019), which is common when the target features are limited owing to unique characteristics (e.g. rare species, specific roofing types, or uncommon rock types or plant species), collecting a sufficient number of data

points may be infeasible, while easy to collect non-target classes. Even with fewer reference data than the optimal, accurate outcomes are still possible depending on the data distribution and representativeness of the data which introduces a higher risk of uncertain results.

An alternative method, known as self-training or pseudo-labelling (PL), involves selecting high-probability data from the predictions of an initial model conducted with fewer data points, and the model training is repeated with data from a classified map with the highest probability. The effectiveness of this method has been proven to improve the classification results, including soil class mapping (Zhang, L. *et al.* 2021), detection of geochemical anomalies (Chen, Y. *et al.* 2023), prediction of invasive species distribution (Cruz, C. *et al.* 2023; Kim, E. *et al.* 2024), and identification of sporadically distributed species (Likó, Sz.B. *et al.* 2024). However, PL does not always improve classification accuracy or even result in worse predictions. Although PL appears to be a good solution for overcoming the issue of limited reference data, users cannot assume that the outcome will be better than that of the model output based on the original data. Geological mapping presents a unique challenge, as both natural and anthropogenic processes (e.g. physical and chemical weathering and mining) can alter rock characteristics. However, the spatial pattern can easily be validated by visual interpretation (e.g. the given rock type can occur at a given location).

We aimed to determine the effectiveness of PL in geological mapping. The selected study site was in an arid environment, where the lack of vegetation made it possible to observe rock types in satellite imagery. We had the following questions related to the probabilities and PL: (i) what was the probability of the related rock type and how did it change with the increased training data; (ii) what was the standard deviation (SD) of classes in the changed areas, and how large areas were influenced; and (iii) what was the direction of the changes in terms of probability? Accordingly, we formulated the following hypotheses: (i) high probability values ensure high model performance and (ii) pseudo-labelling improves map quality by providing additional training data for the modelling process.

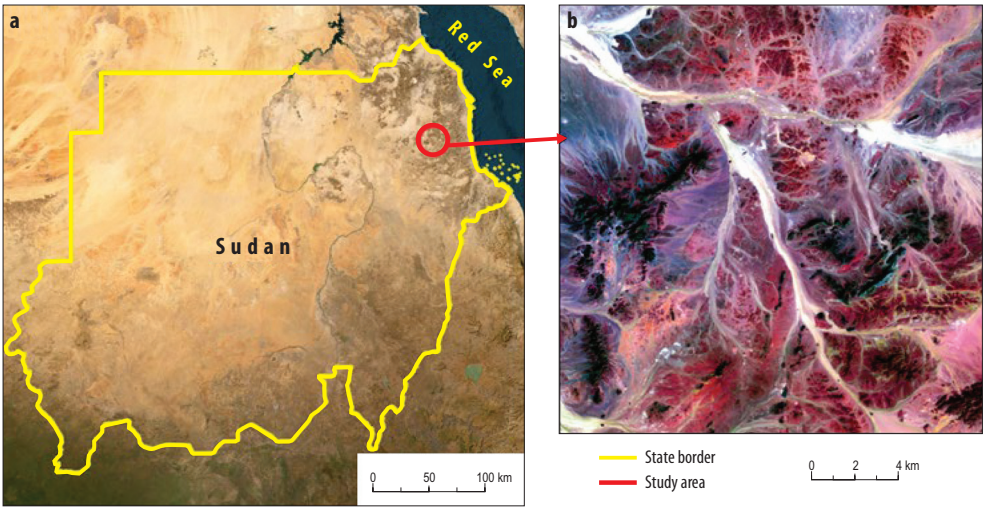

Fig. 1. Location of the study area: Sudan and study area location (a), and Landsat 8 colour composite of the study area in 7, 5, and 2 bands (b). *Source*: Authors' own elaboration.

Data and methodology

Study area

The study area was located in the Red Sea State, northeast Sudan, 90 km from Port Sudan *(Figure 1).* Geologically, the area is part of the Red Sea Hills (RSHs), the Sudanese section of the Arabian Nubian Shield, which is a juvenile continental crust formed between 900 Ma and 550 Ma (Abdelsalam, M.G. *et al.* 2000; Hamimi, Z. *et al.* 2021; Abdelrahman, S. *et al.* 2024). These rocks are highly sheared and deformed, and their geology is complex (Abdelsalam, M.G. and Stern, R.J. 1993). The study area is part of the Gebeit terrane, and the major lithological units include highly sheared low-grade meta-volcanics (meta-basaltic andesite, meta-andesitic basalt, and meta-dacite), meta-sediments (marble), sheared granitoids, and superficial deposits. Accordingly, we identified and mapped the following six rock types: artisanal (art), granite (gra), marble (marb), meta-volcanic rocks (mtvo), ophiolite (ophi), and wadi deposits (WDi).

Applied data

A Landsat 9 image (LC09_L2SP_172046_20231014_20231015_02_T1, US Geological Survey) was used in the analysis on a cloud-free date. Through a thorough review of current geological maps, close visual inspection, and analysis of processed Landsat data, as well as high spatial resolution imagery from Google Earth, we carefully produced reference data. In addition, extensive fieldwork was conducted along regularly planned traverses.

Image processing

Machine learning

Multivariate Adaptive Regression Spline (MARS) and Random Forest (RF) and algorithms were tested in this study because both provide a probability layer that indicates the reliability of the classification pixel-by-pixel regarding the classes.

MARS is a nonparametric, robust algorithm that efficiently manages a large amount of input data and nonlinear relationships between the target and explanatory variables with no assumptions (Friedman, J.H. 1991). The feature space is split into regions based on knots, which define the boundaries of the piecewise linear basis functions that contribute to the overall prediction. Each basis function votes on a class for the data instances, and finally, the function with the majority votes is selected as the prediction. Probability calculation for classification tasks are similar to the procedure of logistic regression (logistic transformation to convert the continuous values to 0 and 1 probabilities), but the main difference is that the MARS, instead of using linear terms, calculates the weighted sum of the basic functions (Friedman, J.H. 1991; Boehmke, B. and Greenwell, B.M. 2019, 2020). In the R implementation, the earth package (Milborrow, S. *et al.* 2024), the 'degree' should be specified, which is the degree of interaction among the input variables, 'nprune' refers to the number of basic functions. The performance of MARS models is high; however, a grid search of hyper-parameters is important.

RF is a widely used and robust algorithm (Breiman, L. 2001), and its efficiency has been proven in several fields such as land cover mapping, soil science, and geology (Belgiu, M. and Drăguț, L. 2016; Shahare, Y.R. *et al.* 2024; Simarmata, N. *et al.* 2025). RF does not assume normal distribution, homoscedasticity, or multi-collinearity due to its calculation method, and classification is based on hundreds of decision trees (DTs) using randomly chosen data (36.8% subset of the training dataset) and variables. The number of trees (ntree) parameter is usually set between 100 and 500, and we used the default setting of the 'caret' package (Kuhn, M. 2022), which is 500. The number of variables (mtry) is chosen at each split of a single DT, the default is the square root of the number of all variables, and hyper-parameter tuning testing can be tested between 1:20.

Models were developed and conducted in R 4.4.2 (R Core Team, 2024) with the 'caret' package (Kuhn, M. 2022). Both models resulted in two outputs: prediction of the classes (classified map) and calculation of the related probabilities (probability map). Probabilities were calculated by classes, but in the analysis, we used only the maximum values, i.e. instead of having six values (for the six classes) with the related probabilities in six raster layers, only the highest values were extracted to a single raster layer. It was important, because the maximums determined the resulting classified output, and we were able to pair and evaluate the classes with the probabilities by pixels: for six classes 16.67 percent (100/6 = 16.66) of maximum probability was enough to classify a pixel, whereas the maximum was 100 percent.

Accuracy assessment

Model testing is a crucial step in all predictions (Bui, D.H. and Mucsi, L. 2022). We used our previously developed module, programmed for an automated accuracy assessment, the Classification Assessment Tool (Szabó, Sz. *et al.* 2024), which calculates accuracies using advanced solutions by taking random subsamples from the entire testing dataset based on a predefined ratio obtained from the testing data (0–1) and the number of repetitions of random sampling. We applied 0.6 for the fraction (60% of data were used at a time with stratified random sampling), and for repetitions, we applied 10. Boxplot diagrams were used to visualise the differences among the classes. The following class-level metrics were calculated: Precision (or User's Accuracy, UA), Sensitivity (Producer's Accuracy, PA) (Congalton, R.G. 1991; Barsi, Á. *et al.* 2018), Specificity (True Negative Rate), F1-scores (or Dice Similarity Coefficient), Jaccard Index (or Intersection over Union, IOU) (Willem, 2017; Grandini, M. *et al.* 2020), and Matthews correlation coefficient (MCC) (Cao, C. *et al.* 2020; Chicco, D. and Jurman, G. 2020).

Pseudo-labelling and statistical evaluation

For pseudo-labelling, users must choose a probability map based on an accuracy assessment (e.g. the highest accuracy). A threshold value should be set to define pixels with a minimum probability to delineate the area where the newly labelled classes will be collected. Although the threshold can be between 1 and 100 percent, only higher values (e.g. 0.9 [> 90%]) are useful; accordingly, we chose 95 percent. This step assigned the relating area to the probability maps. We selected 1000 spatially random data points from the assigned area as PL-data (abbreviated as PL1000), involved in the new training phase, merged with the original training data; possible duplicates were removed. Next, the classification step was repeated with an accuracy assessment.

Furthermore, we compared the class-level accuracy metrics (independent variables: Precision, Sensitivity, Specificity, MCC, F1, IOU) of the original maps with the map where the model was trained with the largest number (1000) of pseudo-labelled, resampled data (PL1000) (i.e. input datasets as dependent variables). A multivariate method, Hotelling's T-squared test, was used to test H_0 (group means for all independent variables were equal). The analysis was conducted with R 4.4.2 with the Hotelling package (Curran, J. and Hersh, T. 2012). Besides the p-values, effect sizes of partial η^2 and Mahalanobis Distance Squared (D^2) were also determined to express the magnitude of the differences between the two values of the dependent variable. For partial η^2, 0.01–0.06 is considered as small, 0.06–0.14 as medium, > 0.14 as large effect; for D^2, 0.25–0.50 is considered as small, 0.50–1.00 as medium, and > 1.00 as large effect (Cohen, J. 2013; Matcharashvili, T. *et al.* 2019; Shaker, A. 2023).

Maps produced with the original training and PL1000 data were compared using cross-tabulation and quantified by cross-entropy and visual analysis. Cross-entropy is a robust index for identifying hotspot areas of change (the higher the value, the larger the

change) (Shim, J.W. 2024). Values below the upper quartile were blanked to enhance the relevant differences between the two maps, and the differences were quantified in comparison (agreement) tables. The two maps were also compared using the Interspersion and Juxtaposition Index (IJI), which showed the isolation of the intermixing of patches (i.e. rock types) (Mead, R.A. *et al.* 1981). Cross-entropy was determined using the 'spatialEco' package (Evans, J.S. *et al.* 2023), and IJI with the 'landscapemetrics' package (Hesselbarth, M.H.K. *et al.* 2025).

We also performed a difference analysis between the original and PL1000 maps, focusing directly on changes and probabilities. We determined the probabilities of the classifications and the relationship between the accuracy metrics and probabilities at the class level by rock type. We focused on the areas where the classification was different in the two approaches, and investigated the probability of the pairs (e.g. in the original approach, a pixel was "art", and in the PL1000 it was "WDi"). Finally, we compared the F1-scores and mean probabilities by rock type using Spearman correlation.

Results

Geological maps with original training data

The MARS and RF algorithms provided seemingly similar maps of rock types, but there were also significant differences. The main difference was in the case of WDi: MARS considerably overestimated WDi and underestimated all the other types *(Table 1)*. Considering the possible occurrence of WDi, the RF model was more reliable, as MARS indicated the deposits at irrelevant locations as well (NW and NE corners of the area, *Figure 2*).

Table 1. Estimated area of rock types by two classification models: Multiple Adaptive Regression Spline (MARS) and Random Forest (RF)

Rock type	Area based on	
	MARS model, km²	RF model, km²
art (astisanal)	54.09	58.98
gra (granit)	133.08	150.57
marb (marble)	47.56	60.64
mtvo (meta-volcanic)	54.93	64.81
ophi (ophiolite	29.60	22.88
WDi (wadi deposits)	139.59	100.96

Source. Compiled by the authors.

However, the class-level accuracy metrics showed different results: WDi had the best performance with MARS *(Figure 3)*. In the case of other rock types, the accuracies were similar (or at least slightly, 2–3%, better with the RF), and the RF had a narrower range based on the repetitions of the testing procedure, that is, it acted more reliably with the existing testing data.

Classification accuracies and probabilities

Regarding the probabilities, maps showed that MARS was supposed to be more accurate as 371.56 km² of the area had > 99 percent probability (considering the maximum probabilities of all classes by pixels), whilst in the case of RF it was only 94.85 km² *(Figure 4)*. Although the near-100 percent pixels dominated the MARS model, it did not perform better; only the probabilities were overoptimistic, and the accuracy measures were only slightly better than RF (*Figure 5*, a). The median (derived from the accuracy assessment data) differences between the two models in the case of the robust F1 and MCC did not demonstrate the superiority of any of the models: for art, marb, and ophi, the RF performed better than the MARS with 4.6–2.2–6.6 percent (MCC) and 4.0–1.8–5.9 percent (F1), whereas the MARS was shown to be a better model for gra, mtvo, and WDi with 2.9–1.4–12.2 percent (MCC) and 2.7–1.0–10.9 percent (F1).

Class level accuracy and the F1-score had no connection with the two classifiers; the Spearman correlation coefficient was 0.21 ($p = 0.51$, i.e. not significant). In half of the rock types, the RF performed better, and in the other half, the MARS performed better based on the F1-scores (*Figure 5*, b). Although the values theoretically followed a linear relationship,

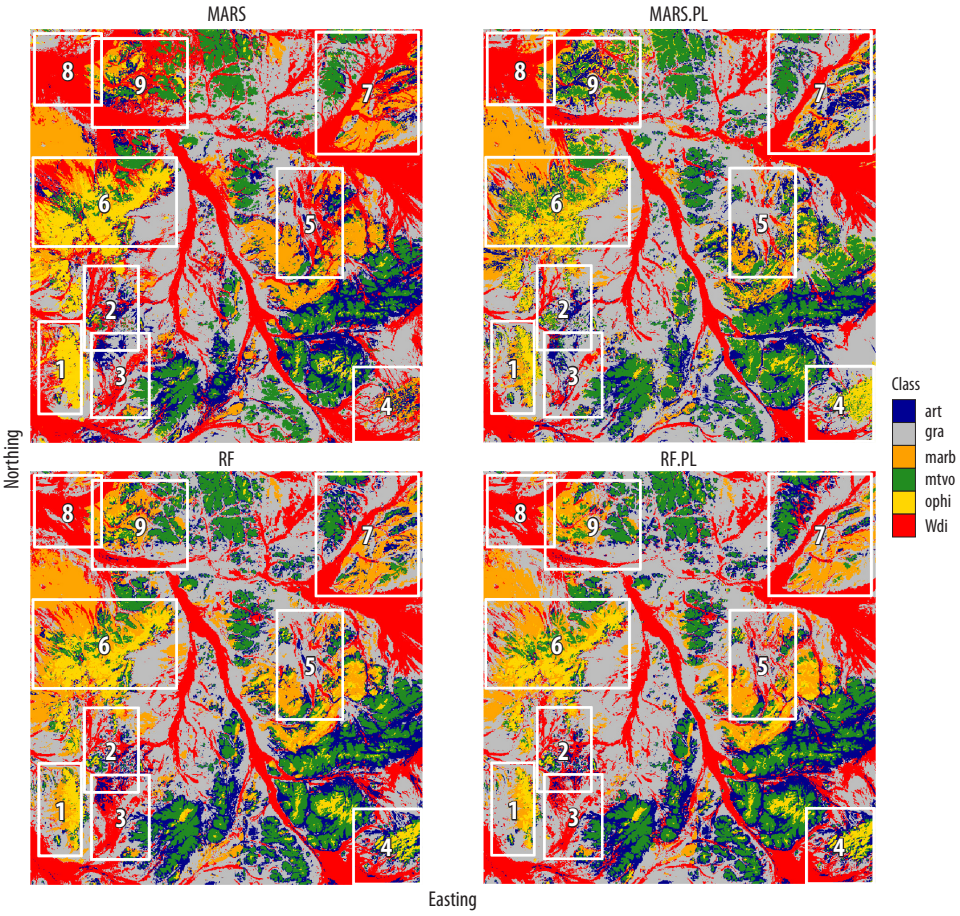

Fig. 2. Geological maps of the Multiple Adaptive Regression Spline (MARS), and Random Forest (RF) models and their pseudo-labelled (PL) versions. Rock types: art = artisanal; gra = granite; marb = marble; mtvo = metavolcanic; ophi = ophiolite; WDi = wadi deposits. (PL = 1000 pseudo-label data with 95% probability; white rectangles and numbers: evaluated areas). *Source*: Authors' own elaboration.

there were outliers in both models: in the case of $MARS_{art}$, the mean accuracy (F1) was one of the lowest (0.81), while the probability was 92 percent, and RF_{marb} had a large F1-score (0.98) with the lowest probability (58%).

Models trained with original and pseudo-labelling

We compared the maps produced with the original and PL1000 (95% probability) training datasets and found that the original training set performed similarly to the case with an additional 1000 data points, at least on the level of accuracy metrics. Comparison matrices showed smaller agreements for MARS (art, marb, and ophi had < 50%, WDi 58%) and slight differences for the RF (all rock types had > 80% agreement, except marb, having only 65%).

Visual analysis brought controversial observations: although both the RF and MARS maps differed by 25 percent from the PL versions regarding the hot-spot changing areas (in addition to the simple expres-

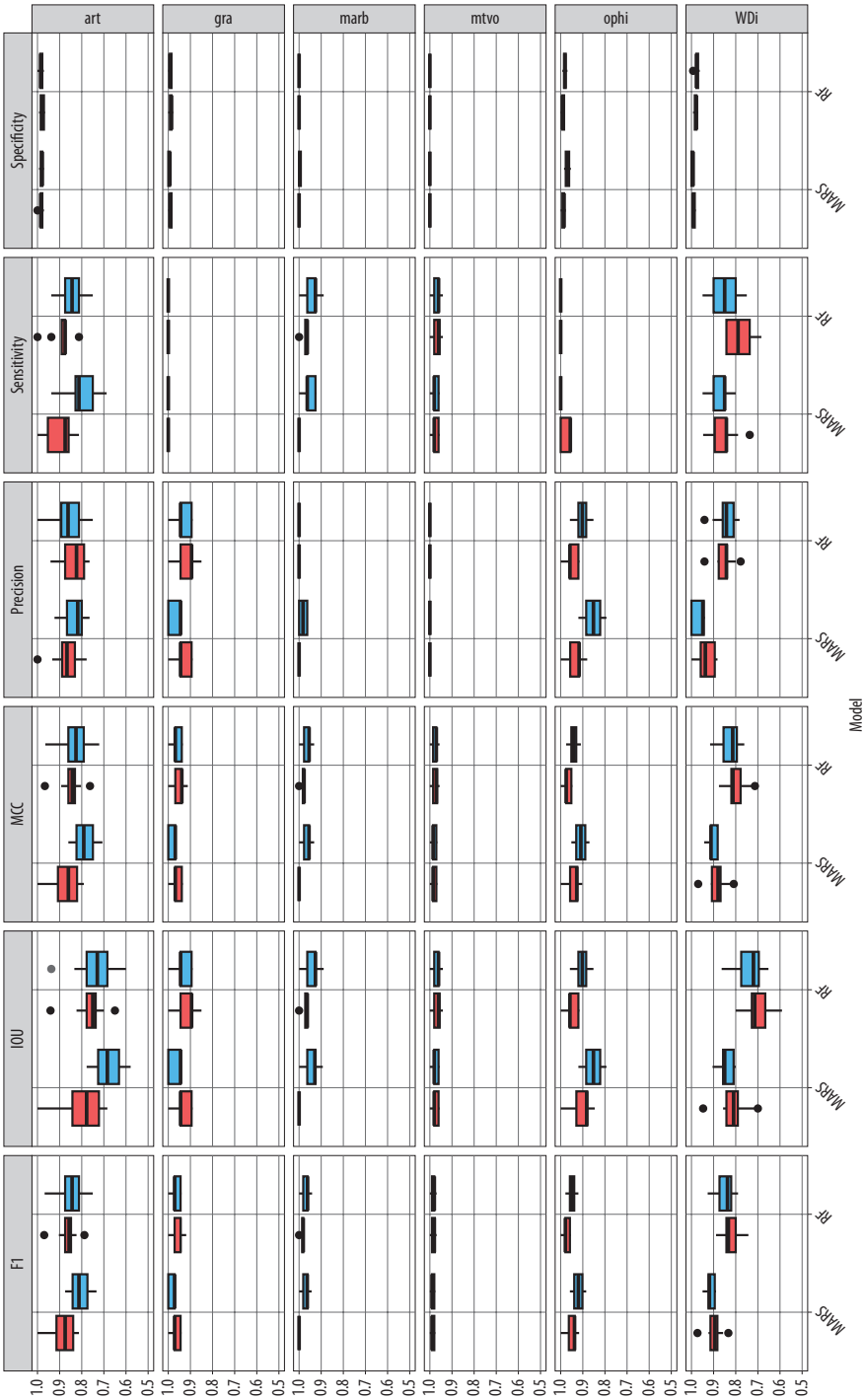

Fig. 3. Accuracy metrics of the MARS and RF classifications. Original = Models trained with the original training dataset; PL1000 = Models trained with an additional 1000 pseudo-labelled data sampled from 95% probability pixels; IOU = Intersection over Union; MCC = Matthews correlation coefficient. For rock types see *Fig. 2*. *Source:* Authors' own elaboration.

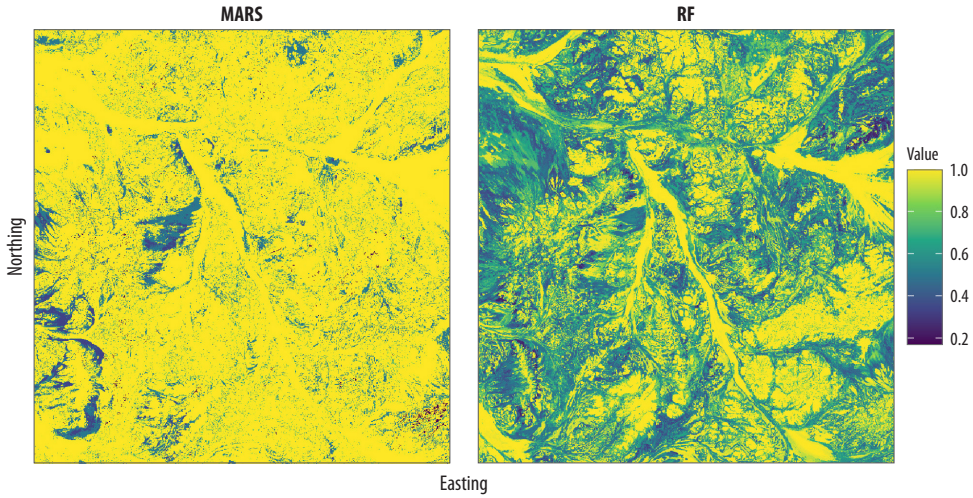

Fig. 4. Probability layers of classification models of the MARS and RF models, visualising the maximum probabilities *Source*: Authors' own elaboration.

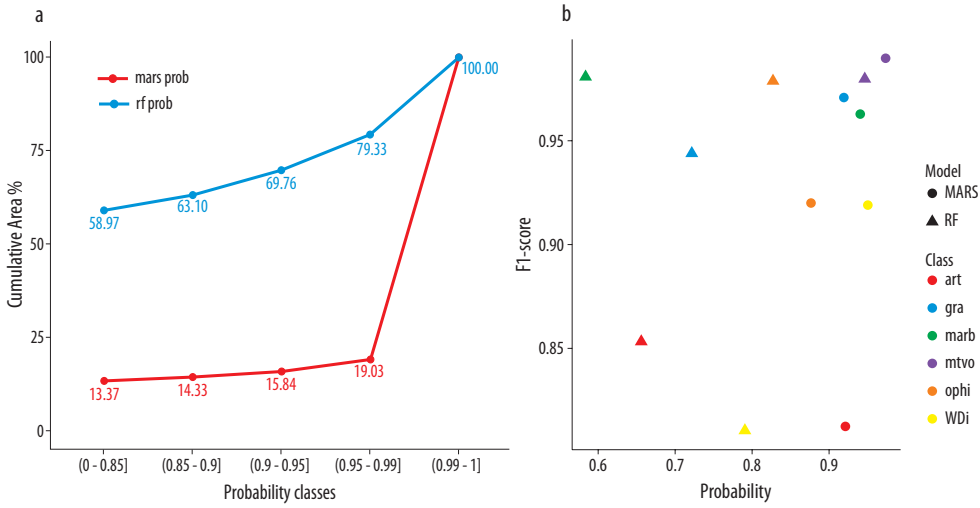

Fig. 5. Cumulative probabilities and accuracies of the MARS and RF models. Proportions of pixels by probability thresholds (a); Probabilities and F1-scores by rock type (b). For rock types see *Fig. 2. Source*: Authors' own elaboration.

sion of the area, the cross-entropy levels differed significantly), the changes caused by the additional data improved the maps *(Table 2).* In area 1, the PL versions outperformed the original classifications for both RF and MARS; in area 2, RF.PL outperformed all other versions, and MARS.PL outperformed the original MARS. For area 3, MARS.PL provided the best outcome, but RF.PL was also better than the original version. In areas 4

*Table 2. Summary of best models based in visual inspection of spatial patterns**

Area	Characteristic rock type	Rock type	Best model
1	ophi, WDi	generally	RF.PL and MARS.PL
2	gra, WDi	generally	RF.PL
3	gra, mtvo, WDi	generally	MARS.PL and RF.PL
4	art, ophi, WDi	generally	RF.PL
5	ophi, marb, gra	generally	RF.PL
6	marb, mtvo, ophi	ophi	RF.PL
6	marb, mtvo, ophi	marb	MARS.PL
7	art, marb, mtvo, WDi	mtvo	MARS and MARS.PL
7	art, marb, mtvo, WDi	WDi	RF.PL
8	gra, WDi	generally	RF.PL
9	art, marb, mtwo, WDi	art	RF.PL
9	art, marb, mtwo, WDi	WDi, mtvo	MARS.PL

*Area codes depicted in *Figure 2*. RF = Random Forest; MARS = Multiple Adaptive Regression Spline; PL = pseudo labelled; art = artisanal, gra = granite, marb = marble, mtvo = meta-volcanic, ophi = ophiolite, WDi = wadi deposits. *Source*. Compiled by the authors.

and 5, the RF.PL was the most reliable solution. For area 6, MARS was the least accurate, and RF.PL provided the best solution for ophi, and MARS.PL for marb. MARS and MARS.PL mapped mtvo better than the other models, and RF.PL mapped WDi the best in the case of area 7. Generally, RF.PL was the best for area 8. In case area 9, RF.PL had the best classification for art and MARS.PL for mtvo and WDi. Accordingly, the PL model versions performed well, and based on the visual analysis, the spatial patterns were determined in several cases as the best outcomes.

Although cross-entropy showed that the hot-spot areas of differences were 25 percent for both model pairs (original vs. PL), in the case of RF, the values were higher, indicating that there was a difference in sparse pixels; that is, the level of mixing of standalone pixels had changed *(Figure 6)*. IJI confirmed that

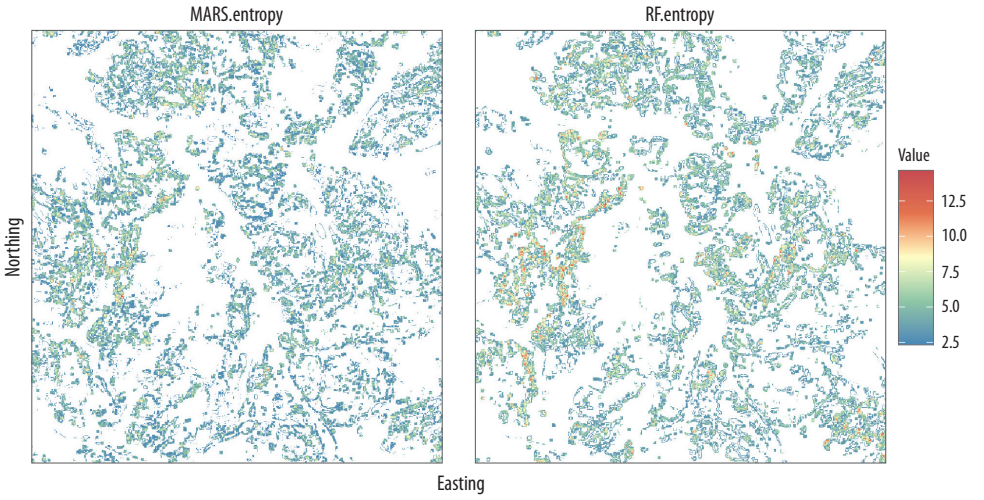

Fig. 6. Differences of the pseudo-labelled model (PL1000) to the original models using the cross-entropy of MARS and RF (values < 2.5, i.e. upper quartile, were blanked). *Source*: Authors' own elaboration.

interspersion became more uneven with the RF.PL with 2.0 percent (77.2) related to the RF (79.2), while in the case of MARS it was 86.2, and for MARS.PL it was 89.5.

The class-level evaluation of the probability levels in the two classification approaches showed that the classifiers reacted differently. In the case of MARS, the probabilities were initially high in the changing pixels of the rock-type pairs, and usually, the larger probabilities had lower SDs. A small decrease in the mean probability levels caused an increase in SD (*Figure 7*, a). The relationship between the mean probability and SDs was almost perfect (e.g. as a second-order polynomial), but the marb-WDi pair was an influential data point with a low mean and SD. Although the changes in probabilities were not significant according to the Wilcoxon test, they exceeded 10 percent in 14 of 30 cases (*Figure 7*, b). In 12 instances, the probability increased. The results were different for RF; typically, lower mean probabilities had lower SDs and followed a linear relationship (*Figure 8*, a). The probabilities of the original approach were lower than those of MARS, and the probabilities of the PL versions were even lower, with significant differences (*Figure 8*, b). The number of cases in which the change was > 10 percent was nine, and the number increased by only five. The magnitudes of the changes also differed; the maximum increase was 62 percent in the case of MARS, and 0.09 for RF (additionally, the largest change regardless of the direction of changes was only 0.27).

Multivariate comparison of the traditional and pseudo-labelling methods

Multivariate comparison of the MARS models

The MARS models performed better with PL1000 than with the original training data, except for three out of 36 cases (without specific rock types). The original data provided more accurate results, and among the three exceptions, the mean differences were below 2 percent (mostly < 0.5%). The decreases in F1, IOU, and MCC were up to 18.8 percent, with a maximum increase of 1.8 percent. For RF models, PL1000 was not very useful; nevertheless, it provided better metrics in 17 out of 36 cases, particularly for gra and WDi rock types *(Figure 9)*.

The difference between the models based on the original and PL1000 training data was significant according to the Hotelling test ($T^2 = 447.77$, $F[6, 593] = 74.63$, $p < 0.001$). $\eta^2 = 0.4302$ indicated a very large effect size and accounted for 43.02 percent of the multivariate variance by group differences in the independent variable. The large effect size was also justified by $D^2 = 0.746$. Both indicated a strong association between the grouping variables and the set of dependent variables. Effect sizes suggested that there were substantial differences in how the models performed across different datasets, considering all performance metrics simultaneously. Accordingly, the groups were well-separated in the multivariate space defined by the dependent variables.

Multivariate comparison of the RF models

For the RF models, the accuracy metrics showed varied results related to the MARS, and the PL1000 training dataset provided better accuracy measures than the original in 15 out of 36 cases. The increase was 5.5 percent and the maximum decrease was 4.8 percent (see *Figure 9*). The Hotelling test revealed a significant difference ($T^2 = 35.31$, $F[6, 593] = 5.89$, $p < 0.001$), but the effect sizes were not as large as in the case of MARS, indicating less pronounced differences between the groups. $\eta^2 = 0.056$ was close to the threshold for a medium effect (0.06), suggesting moderate significance of the group difference. $D^2 = 0.0588$ indicated a relatively small separation between the groups; accordingly, the difference was statistically significant, but the effect was not large.

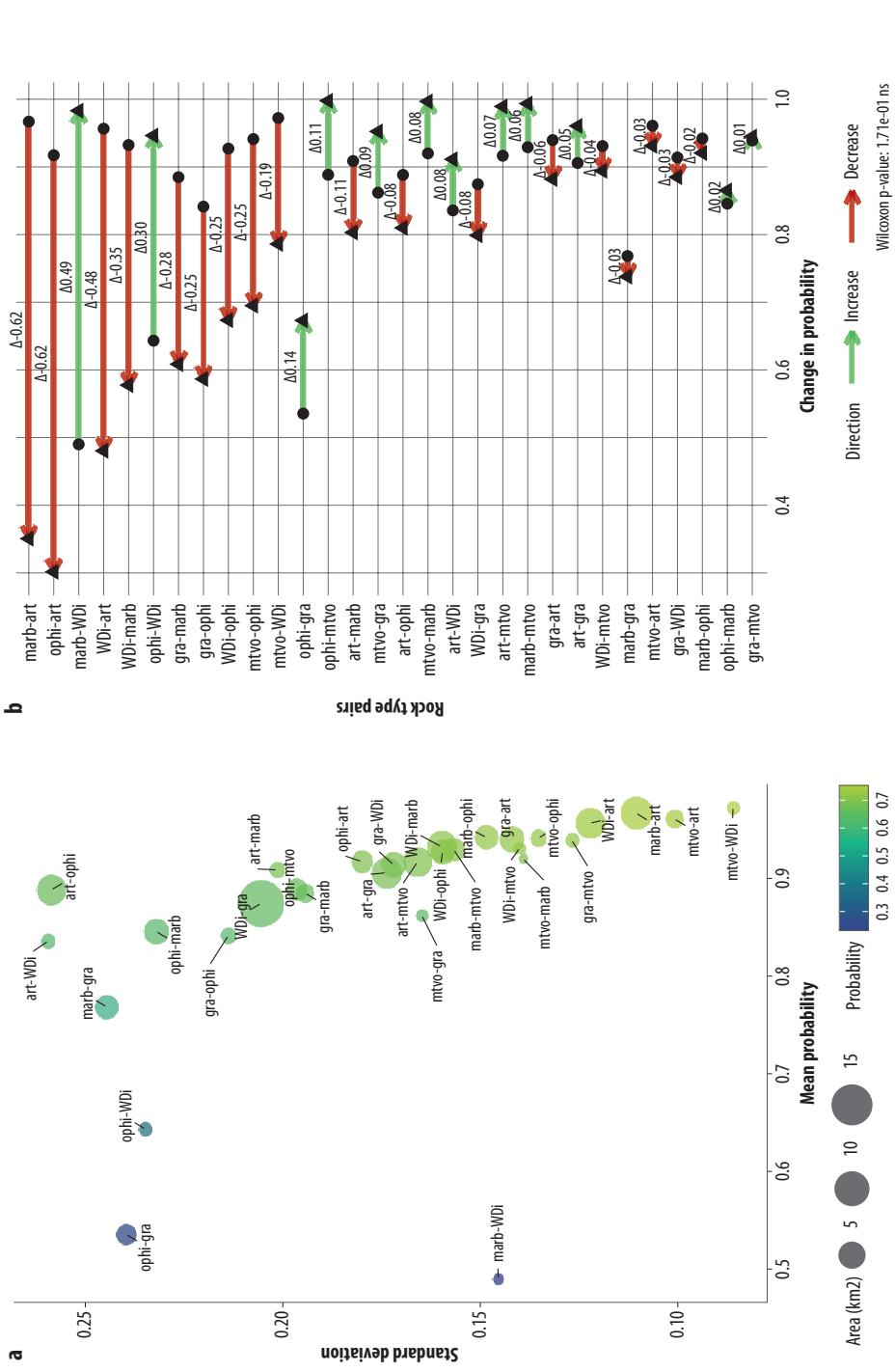

Fig. 7. Changes of rock type classifications in light of probabilities in the case of MARS classifier. Original probabilities by changing rock types (a); Magnitudes of probabilities of rock type pairs (b). For rock types see *Fig. 2. Source*: Authors' own elaboration.

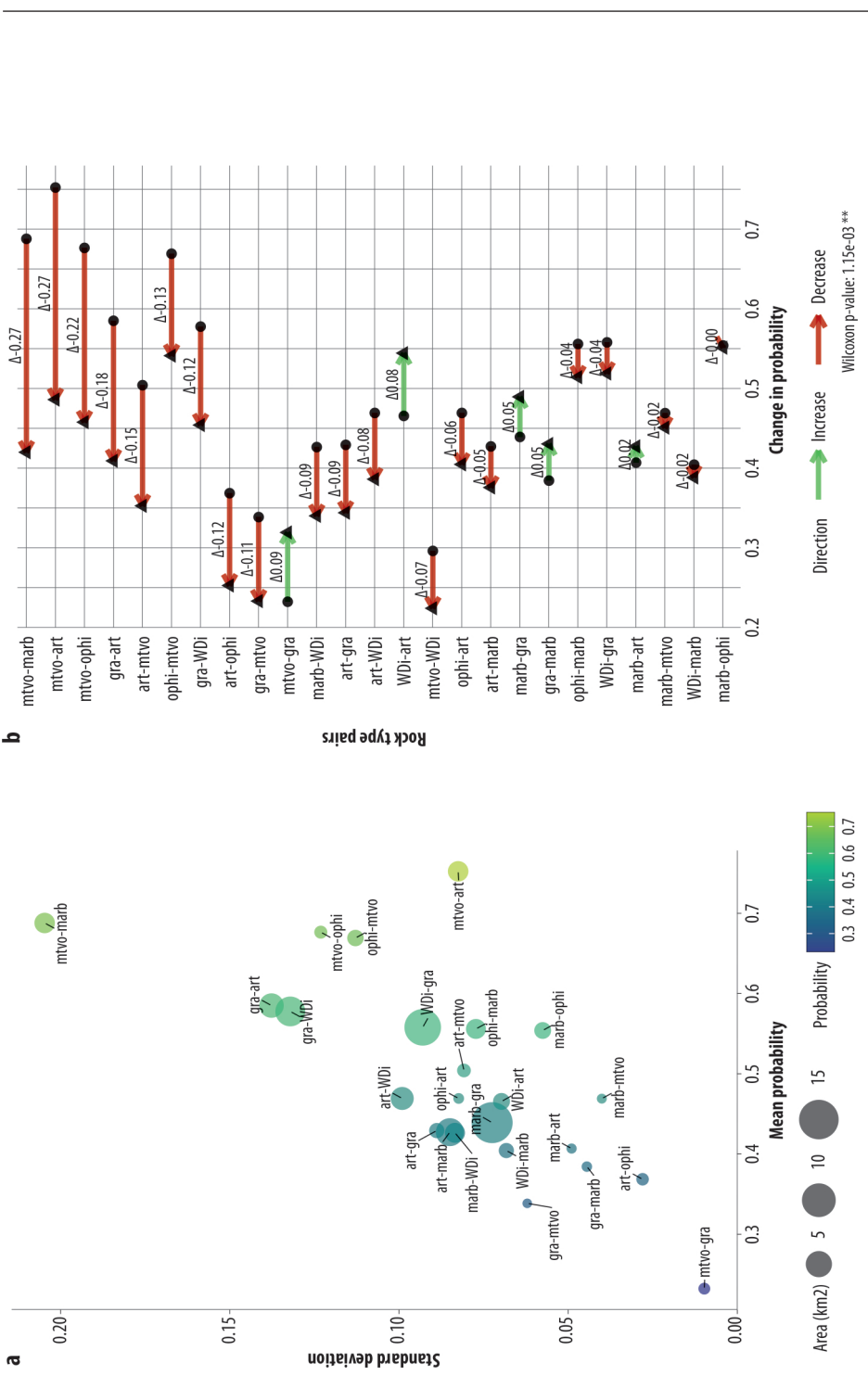

Fig. 8. Changes of rock type classifications in light of probabilities in case of RF classifier. Original probabilities by the changing rock types (a); Magnitudes of probabilities of the rock type pairs (b). For rock types see *Fig. 2. Source:* Authors' own elaboration.

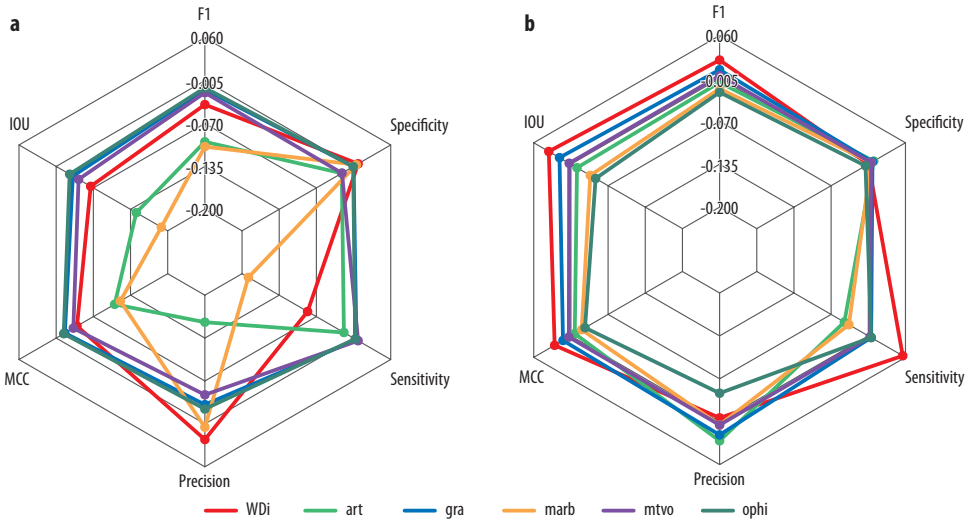

Fig. 9. Difference in model accuracies (Original – PL1000 training data) by rock type and accuracy metrics MARS (a), and RF (b). IOU = Intersection over Union; MCC = Matthews correlation coefficient. For rock types see *Fig. 2.* *Source*: Authors' own elaboration.

Discussion

Insights on classification algorithms

Studying the probability maps alone could lead to misjudging the real performance of the algorithms, as MARS provided unreliably high values, and one could state that RF has poor accuracy. Although probabilities provided new insights, MARS indicated almost perfect reliability (i.e. > 99%) for 81 percent of the study area; however, this was not validated by the accuracy metrics, and RF was not worse than MARS. The OAs only slightly differed between the MARS and RF models; at the class level, there were differences in the misclassification. The reason for the high probabilities of MARS can amplify small spectral differences through its model calculation procedure, whereas RF naturally moderates its probabilities through vote averaging. The key point of this result is that, although MARS seems to be a better classifier based on the probabilities, RF was not worse when considering the accuracy metrics. Even with class-level comparisons regarding the medians, MARS performed only 1–2 percent better than RF in half of the cases. The general issue with confirming the results is that MARS, unlike RF, SVM, XGB, or KNN, is not frequently used. Quirós, E. *et al.* (2009) found that MARS was the best classifier related to Maximum Likelihood and Parallelepiped methods in 13 of 17 test zones in Spain. Nagpal, A. and Singh, V. (2019) used MARS and RF and found that RF was better in six out of 10 datasets. Based on the study by Rotigliano, E. *et al.* (2018), MARS outperformed binary logistic regression for landslide identification, which is consistent with our finding. Harvey, A.S. and Fotopoulos, G. (2016) found that RF was the best classifier; however, in their study, MARS was not applied. Thus, although direct comparison is not a common practice, because MARS is not as popular as RF, we can conclude that both classifiers can generally perform well.

Common misclassifications were caused due to the complexities of geological features, high inter-class similarity, and difficulties presented by remote sensing data. Misclassifications are frequently encountered in lithological mapping, especially when employing this type of data (El-Omairi, M.A. and Garouani, A.E. 2023). In our study, the rock type classification focused on marble, meta-volcanic, ophiolite, altered rock, and wadi deposit units. Discriminating these lithologies is inherently challenging because of their spectral similarities (Othman, A.A. and Gloaguen, R. 2014). Misclassifications occurred primarily among granitoids, wadi deposits, and altered rocks owing to their heterogeneous nature and overlapping spectral characteristics. Specifically, wadi deposits are largely influenced by the weathering of granitoids, whereas altered rocks, often the result of traditional mining activities, are typically associated with these deposits. In contrast, marble, metavolcanic, and ophiolite units predominantly comprise iron (Fe) and magnesium (Mg) minerals, which are less prone to weathering and, thus, less likely to be misclassified. These observations are consistent with Liu, H. *et al.* (2021), who reported common misclassifications between similar lithologies such as granitoids and alluvial sediments owing to their spectral similarities.

Evaluation of pseudo-labelling

PL is generally regarded as a promising method for model-based predictions and focuses on geosciences (e.g. well-log classification: Dunham, M.W. *et al.* 2020; seismic facies classification: Asghar, S. *et al.* 2020). In remote sensing, PL also ensured better models, and several authors have proven its relevance (Aydav, P.S.S. and Minz, S. 2018; Li, J. *et al.* 2022). However, in our study, the outcomes were not always straightforward in terms of usefulness, as reflected by the F1 and MCC values (see *Figure 9*). We evaluated the results both quantitatively (based on accuracy metrics) and qualitatively (based on visual inspection), and the final judgement was not obvious.

A comparison of the original and PL approaches showed that 25 percent of the study area was affected by the changed training data, and the accuracy metrics were affected by the differences between the two approaches, although the improvement was not consistent across all rock types. Focusing on the areas where rock types were classified differently using the two approaches, we found that the consequences differed for the MARS and RF maps. The reason is the initially high probability of the MARS model, which was the opposite of the results experienced in case of RF: the areas of differently classified pixels had high probability with low SD in case of MARS where a small decrease of probability caused the increase of SD; while in case of RF followed a more common trend, having a linear relation with low SDs for low mean probabilities, and high SDs for high probabilities by rock pairs. The main observation was that the probabilities did not improve in the case of MARS and had significantly lower values with RF. Although the MARS was dominated by > 99 percent probabilities, in the areas of change, these values were lower, and the PL approach ensured a higher probability. In the case of RF, the probabilities were lower and could result in lower values with PL, but this did not mean that the performance would have been lower.

Multivariate analyses showed that using 1000 additional data with 95 percent probability changed the accuracy metrics according to the Hotelling test. Furthermore, visual analysis justified these differences. However, PL1000 resulted in conflicting results with several rock types; in these circumstances, PL1000 was useless, and training the model using the original data was more adequate. A summary of the visual analysis results showed that RF.PL was more useful than MARS. The PL was not as successful as the metrics suggested. This seems to be a contradiction, but the explanation can be simple: the amount of testing data was not sufficient to reveal the accuracy in detail. A geologically complex area, such as the study area, would require a higher amount of testing data, but testing only

validated spots is recommended to avoid false accuracy metrics (Meyer, H. and Pebesma, E. 2022; Gao, M. *et al.* 2023). Accordingly, we did not use pseudo-labelled data for testing, but only for training; therefore, the testing points did not cover the entire area. Furthermore, spatial patterns cannot be captured using points, and visual inspection with specific geological knowledge is more important. The main question is, based on analogies, whether there are reasonable (i.e. geologically possible) occurrences of rock types. The artisanal small-scale mining exploits certain rock types (meta-volcanics), and occurrences can be excluded when they are identified on marble, ophiolite, and granite. Wadi deposits (WDi) also have typical areas where wadis can be found, corresponding to the topography.

Limitations

Training data is always the main question of all models, and in our case, we provided a possible approach by augmenting the available training data. Testing is another important point, and we had only a limited amount of data, 256 records that were collected during extensive fieldwork, and rock samples that were investigated and registered. This was sufficient to conduct an accuracy assessment; however, visual inspection of the spatial patterns was useful. Accordingly, further testing is needed with more testing data and other datasets where probabilities can be tested as well.

Conclusions

Our aim was to study the efficacy of pseudo-labelling in the geological mapping of an African area. We applied the RF and MARS classifiers, which provided classified maps and probability maps, and evaluated the results using an accuracy assessment and visual analysis. RF provided more reliable probability levels, whereas the MARS probability map was too optimistic; 85.7 percent of the classified pixels were above 90 percent probability and 81 percent above 99 percent probability, which did not correspond to the accuracy assessment. MARS performed only slightly better than RF, and as the PL data were obtained within the 95 percent range of probability, PL was useful for MARS (with 1000 PL data of 95% probability). For RF, PL helped to obtain better accuracies, but its relevance was smaller owing to its robustness. Visual analysis enhanced the relevance of specific knowledge of the area by confirming or excluding the outcomes of the best and impossible occurrences of rock types. We revealed the relevance of PL in geological mapping for both RF and MARS, and the additional data helped to gain better maps. Based on class-level accuracy metrics, PL provided better maps in the case of MARS (33 out of 36) and fewer cases with RF (17 out of 36), considering 1000 additional samples of 95 percent probability.

Acknowledgements: The project was funded by the NKFI K138079, the framework of the Széchenyi Plan Plus program with the support of the RRF 2.3.1 21 2022 00008 project and the TKP2021-NKTA-32 has been implemented with the support provided by the National Research, Development and Innovation Fund of Hungary.

REFERENCES

Abdelkareem, M., Hamimi, Z., El-Bialy, M.Z., Khamis, H. and Abdel Wahed, S.A. 2021. Integration of remote-sensing data for mapping lithological and structural features in the Esh El-Mallaha area, west Gulf of Suez, Egypt. *Arabian Journal of Geosciences* 14. (6): 497. https://doi.org/10.1007/s12517-021-06791-3

Abdelrahman, S., Ibrahim, M.A.E., Li, H., Abdelrahman, E.M. and Faisal, M. 2024. Geochemical characteristics of Neoproterozoic meta-volcanic rocks of Ariab Auriferous Volcanogenic Massive Sulfide deposit, Red Sea Hills, North-East Sudan. *Journal of African Earth Sciences* 216. 105305. https://doi.org/10.1016/j.jafrearsci.2024.105305

Abdelsalam, M.G. and Stern, R.J. 1993. Tectonic evolution of the Nakasib suture, Red Sea Hills, Sudan: Evidence for a late Precambrian Wilson Cycle. *Journal of the Geological Society* 150. (2): 393–404. https://doi.org/10.1144/gsjgs.150.2.0393

Abdelsalam, M.G., Stern, R.J. and Berhane, W.G. 2000. Mapping gossans in arid regions with Landsat TM

and SIR-C images: The Beddaho Alteration Zone in northern Eritrea. *Journal of African Earth Sciences* 30. (4): 903–916. https://doi.org/10.1016/S0899-5362(00)00059-2

Abrams, M. and Yamaguchi, Y. 2019. Twenty years of ASTER contributions to lithological mapping and mineral exploration. *Remote Sensing* 11. (11): 1394. https://doi.org/10.3390/rs11111394

Amusuk, D.J., Hashim, M., Pour, A.B. and Musa, S.I. 2016. Utilization of Landsat-8 data for lithological mapping of basement rocks of plateau state North Central Nigeria. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLII-4-W1*, 335–337. TC IV
 International Conference on Geomatic and Geospatial Technology (GGT) 2016. Vol. XLII-4/W1. 3–5 October 2016, Kuala Lumpur, Malaysia. https://doi.org/10.5194/isprs-archives-XLII-4-W1-335-2016

Asghar, S., Choi, J., Yoon, D. and Byun, J. 2020. Spatial pseudo-labelling for semi-supervised facies classification. *Journal of Petroleum Science and Engineering* 195. 107834. https://doi.org/10.1016/j.petrol.2020.107834

Aydav, P.S.S. and Minz, S. 2018. Classification of hyperspectral images using self-training and a pseudo-validation set. *Remote Sensing Letters* 9. (11): 1109–1117. https://doi.org/10.1080/2150704X.2018.1511932

Bachri, I., Hakdaoui, M., Raji, M., Teodoro, A.C. and Benbouziane, A. 2019. Machine learning algorithms for automatic lithological mapping using remote sensing data: A case study from Souk Arbaa Sahel, Sidi Ifni Inlier, Western Anti-Atlas, Morocco. *ISPRS International Journal of Geo-Information* 8. (6): 1–20. https://doi.org/10.3390/ijgi8060248

Barsi, Á., Kugler, Zs., László, I., Szabó, Gy. and Abdulmutalib, H.M. 2018. Accuracy dimensions in remote sensing. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLII–3*. 61–67. https://doi.org/10.5194/isprs-archives-XLII-3-61-2018

Belgiu, M. and Drăguţ, L. 2016. Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing* 114. 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

Boehmke, B. and Greenwell, B.M. 2019. *Hands-On Machine Learning with R*. London, Chapman and Hall/CRC. https://doi.org/10.1201/9780367816377

Boemhke, B. and Greenwell, B.M. 2020. *Hands-On Machine Learning with R*. London, Chapman and Hall/CRC. https://bradleyboehmke.github.io/HOML/

Breiman, L. 2001. Random forests. *Machine Learning* 45. (1): 5–32. https://doi.org/10.1023/A:1010933404324

Bui, D.H. and Mucsi, L. 2022. Predicting the future land-use change and evaluating the change in landscape pattern in Binh Duong province, Vietnam. *Hungarian Geographical Bulletin* 71. (4): 349–364. https://doi.org/10.15201/hungeobull.71.4.3

Cao, C., Chicco, D. and Hoffman, M.M. 2020. *The MCC-F1 Curve: A Performance Evaluation Technique for Binary Classification.* arXiv:2006.11278, arXiv. https://doi.org/10.48550/arXiv.2006.11278

Chen, Y., Sui, Y. and Shayilan, A. 2023. Constructing a high-performance self-training model based on support vector classifiers to detect gold mineralization-related geochemical anomalies for gold exploration targeting. *Ore Geology Reviews* 153. 105265. https://doi.org/10.1016/j.oregeorev.2022.105265

Chicco, D. and Jurman, G. 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* 21. (6): https://doi.org/10.1186/s12864-019-6413-7

Cohen, J. 2013. *Statistical Power Analysis for the Behavioural Sciences.* 2nd edition. London, Routledge. https://doi.org/10.4324/9780203771587

Collins, L., McCarthy, G., Mellor, A., Newell, G. and Smith, L. 2020. Training data requirements for fire severity mapping using Landsat imagery and random forest. *Remote Sensing of Environment* 245. 111839. https://doi.org/10.1016/j.rse.2020.111839

Congalton, R.G. 1991. A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment* 37. (1): 35–46. https://doi.org/10.1016/0034-4257(91)90048-B

Cruz, C., McGuinness, K., Perrin, P.M., O'Connell, J., Martin, J.R. and Connolly, J. 2023. Improving the mapping of coastal invasive species using UAV imagery and deep learning. *International Journal of Remote Sensing* 44. (18): 5713–5735. https://doi.org/10.1080/01431161.2023.2251186

Curran, J. and Hersh, T. 2012. *Hotelling: Hotelling's T^2 Test and Variants.* p. 1.0-8 (Dataset). https://doi.org/10.32614/CRAN.package.Hotelling

Dunham, M.W., Malcolm, A. and Kim Welford, J. 2020. Improved well-log classification using semi-supervised label propagation and self-training, with comparisons to popular supervised algorithms. *Geophysics* 85. (1): 1–15. https://doi.org/10.1190/geo2019-0238.1

El-Omairi, M.A. and Garouani, A.E. 2023. A review on advancements in lithological mapping utilizing machine learning algorithms and remote sensing data. *Heliyon* 9. (9): https://doi.org/10.1016/j.heliyon.2023.e20168

Evans, J.S., Murphy, M.A. and Ram, K. 2023. *spatialEco: Spatial Analysis and Modelling Utilities.* Version 2.0-2. (Computer software). https://cran.r-project.org/web/packages/spatialEco/index.html

Foody, G.M. 2009. Sample size determination for image classification accuracy assessment and comparison. *International Journal of Remote Sensing* 30. (20): 5273–5291. https://doi.org/10.1080/01431160903130937

Friedman, J.H. 1991. Multivariate adaptive regression splines. *The Annals of Statistics* 19. (1): 1–67. https://doi.org/10.1214/aos/1176347963

Gao, M., Wang, G., Xu, Y., Mou, N., Huang, L., Zuo, L. and Wu, R. 2023. 3D mineral exploration Cu-Zn targeting with multi-source geoscience datasets in the Weilasituo-bairendaba district, Inner Mongolia, China. *Frontiers in Earth Science* 11. https://doi.org/10.3389/feart.2023.1102640

Grandini, M., Bagli, E. and Visani, G. 2020. *Metrics for Multi-Class Classification: An Overview*. arXiv:2008.05756, arXiv. https://doi.org/10.48550/arXiv.2008.05756

Hamimi, Z., Fowler, A.-R., Collins, A., Liégeois, J.-P., Abdelsalam, M. and Abd El-Wahed, M. 2021. *The Geology of the Arabian-Nubian Shield*. Regional Geology Reviews. Cham, Springer Nature. https://doi.org/10.1007/978-3-030-72995-0

Han, W., Zhang, X., Wang, Y., Wang, L., Huang, X., Li, J., Wang, S., Chen, W., Li, X., Feng, R., Fan, R., Zhang, X. and Wang, Y. 2023. A survey of machine learning and deep learning in remote sensing of geological environment: Challenges, advances, and opportunities. *ISPRS Journal of Photogrammetry and Remote Sensing* 202. 87–113. https://doi.org/10.1016/j.isprsjprs.2023.05.032

Harvey, A.S. and Fotopoulos, G. 2016. Geological mapping using machine learning algorithms. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLI-B8. 423–430. XXIII ISPRS Congress, Commission VIII, Volume XLI-B8. 19 July 2016, Prague, Czech Republic. https://doi.org/10.5194/isprsarchives-XLI-B8-423-2016

Hesselbarth, M.H.K., Sciaini, M., Nowosad, J. and Hans, S. 2025. *Landscape Metrics for Categorical Map Patterns*. p. 2.2.1 (Dataset). https://doi.org/10.32614/CRAN.package.landscapemetrics

James, G., Witten, D., Hastie, T. and Tibshirani, R. 2013. *An Introduction to Statistical Learning: With Applications in R*. 1st edition. Corr. 7th printing 2017 edition. Cham, Springer.

Kim, E., Moon, J., Shim, J. and Hwang, E. 2024. Predicting invasive species distributions using incremental ensemble-based pseudo-labelling. *Ecological Informatics* 79. 102407. https://doi.org/10.1016/j.ecoinf.2023.102407

Kuhn, M. 2022. *caret: Classification and Regression Training*. https://CRAN.R-project.org/package=caret

Li, C., Wang, J., Wang, L., Hu, L. and Gong, P. 2014. Comparison of classification algorithms and training sample sizes in urban land classification with Landsat thematic mapper imagery. *Remote Sensing* 6. (2): 964–983. https://doi.org/10.3390/rs6020964

Li, J., Sun, B., Li, S. and Kang, X. 2022. Semi-supervised semantic segmentation of remote sensing images with consistency self-training. *IEEE Transactions on Geoscience and Remote Sensing* 60. 1–11. https://doi.org/10.1109/TGRS.2021.3134277

Likó, Sz.B., Holb, I.J., Oláh, V., Burai, P. and Szabó, Sz. 2024. Deep learning-based training data augmentation combined with post-classification improves the classification accuracy for dominant and scattered invasive forest tree species. *Remote Sensing in Ecology and Conservation* 10. (2): 203–219. https://doi.org/10.1002/rse2.365

Liu, H., Wu, K., Xu, H. and Xu, Y. 2021. Lithology classification using TASI thermal infrared hyperspectral data with convolutional neural networks. *Remote Sensing* 13. (16): 3117. https://doi.org/10.3390/rs13163117

Luque, A., Carrasco, A., Martín, A. and De las Heras, A. 2019. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition* 91. 216–231. https://doi.org/10.1016/j.patcog.2019.02.023

Matcharashvili, T., Czechowski, Z. and Zhukova, N. 2019. Mahalanobis distance-based recognition of changes in the dynamics of a seismic process. *Nonlinear Processes in Geophysics* 26. (3): 291–305. https://doi.org/10.5194/npg-26-291-2019

Maxwell, A.E., Warner, T.A. and Fang, F. 2018. Implementation of machine-learning classification in remote sensing: An applied review. *International Journal of Remote Sensing*. 39. (9): 2784–2817. https://doi.org/10.1080/01431161.2018.1433343

Mead, R.A., Sharik, T.L., Prisley, S.P. and Heinen, J.T. 1981. A computerized spatial analysis system for assessing wildlife habitat from vegetation maps. *Canadian Journal of Remote Sensing* 7. (1): 34–40. https://doi.org/10.1080/07038992.1981.10855007

Meyer, H. and Pebesma, E. 2022. Machine learning-based global maps of ecological variables and the challenge of assessing them. *Nature Communications* 13. (1): 2208. https://doi.org/10.1038/s41467-022-29838-9

Milborrow, S., Hastie, T. and Tibshirani, R. 2024. *Package 'earth': Multivariate Adaptive Regression Splines*. CRAN. https://doi.org/10.32614/CRAN.package.earth

Nagpal, A. and Singh, V. 2019. Coupling multivariate adaptive regression spline (MARS) and random forest (RF): A hybrid feature selection method in action. *International Journal of Healthcare Information Systems and Informatics* 14. (1): 1–18. https://doi.org/10.4018/IJHISI.2019010101

Nair, P., Srivastava, D.K. and Bhatnagar, R. 2023. Application of machine learning in mineral mapping using remote sensing. In *IOT with Smart Systems*. Eds.: Choudrie, J., Mahalle, P., Perumal, T. and Joshi, A., Cham, Springer Nature, 27–35. https://doi.org/10.1007/978-981-19-3575-6_4

Othman, A.A. and Gloaguen, R. 2014. Improving lithological mapping by SVM classification of spectral and morphological features: The discovery of a new chromite body in the Mawat Ophiolite Complex (Kurdistan, NE Iraq). *Remote Sensing 6*. (8): 30. https://doi.org/10.3390/rs6086867

Quirós, E., Felicísimo, Á.M. and Cuartero, A. 2009. Testing multivariate adaptive regression splines (MARS) as a method of land cover classification of TERRA-ASTER satellite images. *Sensors* 9. (11): 9011–9028. https://doi.org/10.3390/s91109011

R Core Team, 2024. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rotigliano, E., Martinello, C., Agnesi, V. and Conoscenti, C. 2018. Evaluation of debris flow susceptibility in El Salvador (CA): A comparison between multivariate adaptive regression splines

(MARS) and binary logistic regression (BLR). *Hungarian Geographical Bulletin* 67. (4): 361–373. https://doi.org/10.15201/hungeobull.67.4.5

Shahare, Y.R., Singh, M.P., Singh, S.P., Singh, P. and Diwakar, M. 2024. ASUR: Agriculture soil fertility assessment using random forest classifier and regressor. *Procedia Computer Science* 235. 1732–1741. https://doi.org/10.1016/j.procs.2024.04.164

Shaker, A. 2023. *Chapter 6 Effect sizes | STM1001 Topic 7: One-way ANOVA*. https://bookdown.org/a_shaker/STM1001_Topic_7/6-effect-sizes.html

Shim, J.W. 2024. Enhancing cross entropy with a linearly adaptive loss function for optimized classification performance. *Scientific Reports* 14. (1): 27405. https://doi.org/10.1038/s41598-024-78858-6

Shirmard, H., Farahbakhsh, E., Heidari, E., Beiranvand Pour, A., Pradhan, B., Müller, D. and Chandra, R. 2022. A comparative study of convolutional neural networks and conventional machine learning models for lithological mapping using remote sensing data. *Remote Sensing* 14. (4): 819. https://doi.org/10.3390/rs14040819

Simarmata, N., Wikantika, K., Tarigan, T.A., Aldyansyah, M., Tohir, R.K., Fauzi, A.I. and Fauzia, A.R. 2025. Comparison of random forest, gradient tree boosting, and classification and regression trees for mangrove cover change monitoring using Landsat imagery. *The Egyptian Journal of Remote Sensing and Space Sciences* 28. (1): 138–150. https://doi.org/10.1016/j.ejrs.2025.02.002

Szabó, Sz., Holb, I.J., Abriha-Molnár, V.É., Szatmári, G., Singh, S.K. and Abriha, D. 2024. Classification Assessment Tool: A program to measure the uncertainty of classification models in terms of class-level metrics. *Applied Soft Computing* 155. 111468. https://doi.org/10.1016/j.asoc.2024.111468

Szaniawska, L. 2018. Lithological maps visualizing the achievements of geological sciences in the first half of the 19th century. *Polish Cartographical Review* 50. (2): 87–109. https://doi.org/10.2478/pcr-2018-0006

Willem, 2017 – https://stats.stackexchange.com/users/159052/willem, 2017, *F1/Dice-Score vs IoU*. https://stats.stackexchange.com/q/276144

Zhang, L., Yang, L., Ma, T., Shen, F., Cai, Y. and Zhou, C. 2021. A self-training semi-supervised machine learning method for predictive mapping of soil classes with limited sample data. *Geoderma* 384. 114809. https://doi.org/10.1016/j.geoderma.2020.114809