EXPLORING THE USE OF CLASSIFICATION UNCERTAINTY TO IMPROVE CLASSIFICATION ACCURACY

D. Moraes [1,2]*, P. Benevides[1], F. D. Moreira [1], H. Costa [1,2], M. Caetano [1,2]

[1] Direção-Geral do Território, Rua da Artilharia Um, 107, 1099-052 Lisboa, Portugal (pbenevides, hcosta, mario.caetano)@dgterritorio.pt, franciscodmoreira@gmail.com
[2] NOVA Information Management School (NOVA IMS), Universidade Nova Lisboa, Campus de Campolide, 1070-312 Lisbon, Portugal – moraesd90@gmail.com

Commission III, WG III/1

KEY WORDS: Supervised Classification, Classification Uncertainty, Remote Sensing, Land Cover, Accuracy Assessment.

ABSTRACT:

Supervised classification of remotely sensed images has been widely used to map land cover and land use. Since the performance of supervised methods depends on the quality of the training data, it is essential to develop methods to generate an enhanced training dataset. Active learning represents an alternative for such purpose as it proposes to create a dataset of optimized samples, normally collected based on classification uncertainty. However, it is heavily dependent on human interaction, since the user has to label selected samples over a number of iterations. In this paper, we explore the use of uncertainty to improve classification accuracy through a single iteration. We conducted experiments in a region of Portugal (Trás-os-Montes), using multi-temporal Sentinel-2 images. The proposed approach consisted in computing the classification uncertainty of a Random Forest to collect additional training data from areas of high uncertainty and perform a new classification. An accuracy assessment was performed to compare the overall accuracy of the initial and new classifications. The results exhibited an increase in accuracy, though considered not statistically significant. Obstacles related to labelling additional sampling units resulted in a lack of additional training data for various classes, which might have limited the accuracy improvement. Additionally, an uneven proportion of additional training sampling units per class and the collection of new sample data from a limited number of uncertainty regions might also have prevented a higher increase in accuracy. Nevertheless, visual inspection of the maps revealed that the new classification reduced the confusion between some classes.

1. INTRODUCTION

The recent availability of new data and developments in computing processing and classification algorithms has contributed to map and monitor land cover and land use (LCLU) efficiently (Wulder *et al.*, 2018). Supervised classification of remotely sensed images has been widely used to map LCLU, as a variety of studies suggest that these methods generally produce higher accuracy compared to unsupervised methods (Maxwell *et al.*, 2018). The success of supervised methods depends on the quality of the training dataset, which should preferably contain balanced and representative training samples (Belgiu and Drăguţ, 2016).

Different approaches for the creation of enhanced training datasets have been proposed, consequently contributing to an improvement in classification accuracy. Such is the case of active learning, which proposes creating a small training dataset containing optimized sample units, collected based on a query criterion, whose performance can be similar to larger training datasets composed by randomly collected samples (Li *et al.*, 2013). The process is based on the interaction between analyst and model, in which the model provides the analyst with unlabelled sampling units which yield maximal information. Then, the analyst is responsible for labelling such units, which are incorporated to the prior training dataset for a new classification. This cycle is repeated until a satisfactory result is achieved (Tuia *et al.*, 2011).

Among the various query criteria employed to select new training sampling units mentioned in the literature, e.g. uncertainty, representativeness, inconsistency, variance and error (Ahmad *et al.*, 2019), classification uncertainty is the most common. Samples with high classification uncertainty normally represent difficult examples, and their inclusion in training can contribute to improve the model's predictive capabilities. The uncertainty can be determined based on a range of approaches (Tuia *et al.*, 2011). Breaking Ties (BT) is a popular approach, suitable to be employed with classifiers that output posterior probabilities (Crawford *et al.*, 2013), as is the case of Random Forest (RF). BT consists in computing the uncertainty as the difference between the two highest class membership probabilities.

Despite the encouraging results, active learning is strongly dependent on human interaction, as the analyst has to label additional training sampling units throughout multiple iterations. In this context, the study conducted by Mack *et al.* (2017) proposed to apply the active learning principles in a single iteration. An initial classification with training samples derived automatically from an existing reference dataset was conducted, following the computation of classification uncertainty to determine areas of high uncertainty in the map, from which unlabelled sampling units were collected, labelled and incorporated into the initial training dataset to produce the final classification. Although the potential of the methodology

* Corresponding author

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLIII-B3-2021-81-2021 | © Author(s) 2021. CC BY 4.0 License.

was demonstrated, the impact of introducing additional samples was not assessed.

In this paper experiments are performed in order to further investigate the use of classification uncertainty to improve classification accuracy. It is proposed to assess whether the introduction of additional training samples collected from areas of high classification uncertainty can improve an existing training dataset and, consequently, increase land cover classification accuracy. The paper is presented according to the following structure: study area and data, methods, results and conclusions.

2. STUDY AREA AND DATA

2.1 Study area

The chosen study area is the region of Trás-os-Montes, located in the North of Portugal (Figure 1). It is characterized by mountainous land, with rocks, forest and bushes, besides agricultural areas in the lower lands. The pronounced land cover diversity found in this area poses a good scenario for uncertainty classification experiments.

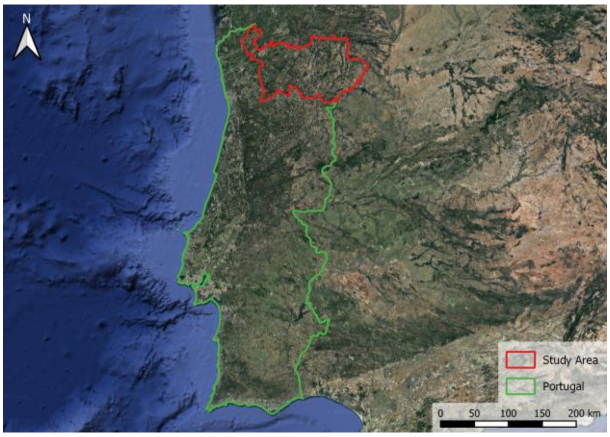

Figure 1. Location of the study area.

2.2 Data

The data used in this study can be separated into remotely sensed data and auxiliary data. The remotely sensed data consisted of 457 Sentinel-2 images with less than 50% cloud cover from the agricultural year of 2018 in Portugal (October 2017 to September 2018) downloaded from the Theia Land Data Centre. In addition, an orthophotomap with 25cm spatial resolution from 2018 was used for the purpose of labelling by photointerpretation sampling units for training and validation.

The auxiliary data consist of multiple datasets used either as reference data to automatically extract training samples or as filtering data to refine the quality of the training samples. The national land use and land cover cartography (COS), the Portuguese Land Parcel Identification System (LPIS) and roads network from OpenStreetMap were used as reference data. In terms of filter data, the national cartography of burned areas, the Copernicus Land Monitoring Service's High Resolution Layers (HRL) products from 2015 and a mask of NDVI for Forest change detection in 2015-2018 (Costa *et al*., 2020) were used.

3. METHODS

Sentinel-2 monthly composites were calculated from the median value of single image to remove pixels contaminated by clouds and their shadows. For each composite, 10 bands (B2, B3, B4, B5, B6, B7, B8, B8A, B11 and B12) were obtained, from which 5 spectral indices were computed. In addition, 7 spectro-temporal metrics were computed for each band and index. The final Sentinel-2 dataset consisted of 285 bands: 10 bands and 5 indices for each month and 7 metrics for each band and index.

Reference data were used to delineate polygons from which training samples were collected automatically. Filtering data were employed to refine this process and prevent mislabelling; e.g. removing areas from the reference dataset that are more heterogeneous or not related to a specific land cover type, and preserving those prone to follow a condition expected for a specific land cover class. Some classes in particular, however, needed training data collected manually as preliminary results indicated that some classes have low accuracies when sampled automatically. Manual collection of training samples was based on delineation of polygons through visual interpretation of the 2018 orthophoto map. The automatic and manual training samples were extracted from the corresponding filtered or manual data sets, but subject to spatial constraints. A negative buffer of 40 and 10 m was applied to the automatic and manual training polygons, respectively, and automatic areas smaller than 1000 m² were eliminated before sample extraction. Our approach proposes to use different nomenclatures for the training and final map. The final map nomenclature results from the aggregation of training classes. A total of 22 and 10 LCLU classes were used for the training and final map, respectively. Such nomenclatures can be seen in Table 1.

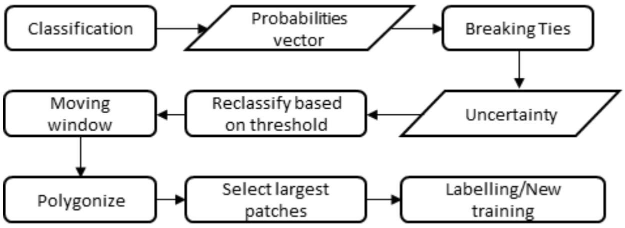

Figure 2. Proposed workflow.

The proposed workflow (Figure 2) consists in conducting an initial classification using Random Forest and up to 6000 training sampling units per class, depending on sample availability. Besides predicting classes, the scikit-learn RF implementation (Pedregosa *et al*., 2011) can also predict class probabilities vectors, which were used to compute the classification uncertainty according to the BT approach. Uncertainty is given by the difference between the highest and second-highest class probabilities, with values ranging from 0 (high uncertainty) to 1 (low uncertainty).

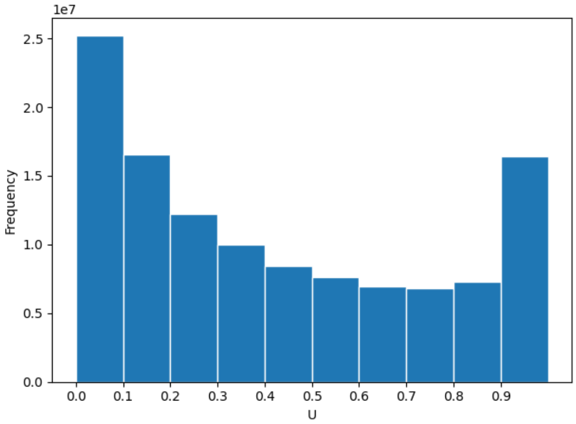

Figure 3. Histogram of classification uncertainty values (U).

Next, a reclassification based on a threshold of uncertainty was conducted to generate a binary map, which was smoothed using a 5x5 pixel moving window to reduce the salt and pepper effect and create contiguous groups of pixels of high uncertainty. The value of 0.1 was adopted as the uncertainty threshold, since the distribution of the uncertainty in the map of the initial classification revealed that a sufficient portion of the pixels had an uncertainty (U) ≤ 0.1 (Figure 3). Then, the raster was converted to vector and the 20 largest patches of high uncertainty per class were selected. Such processes are illustrated in Figure 4.

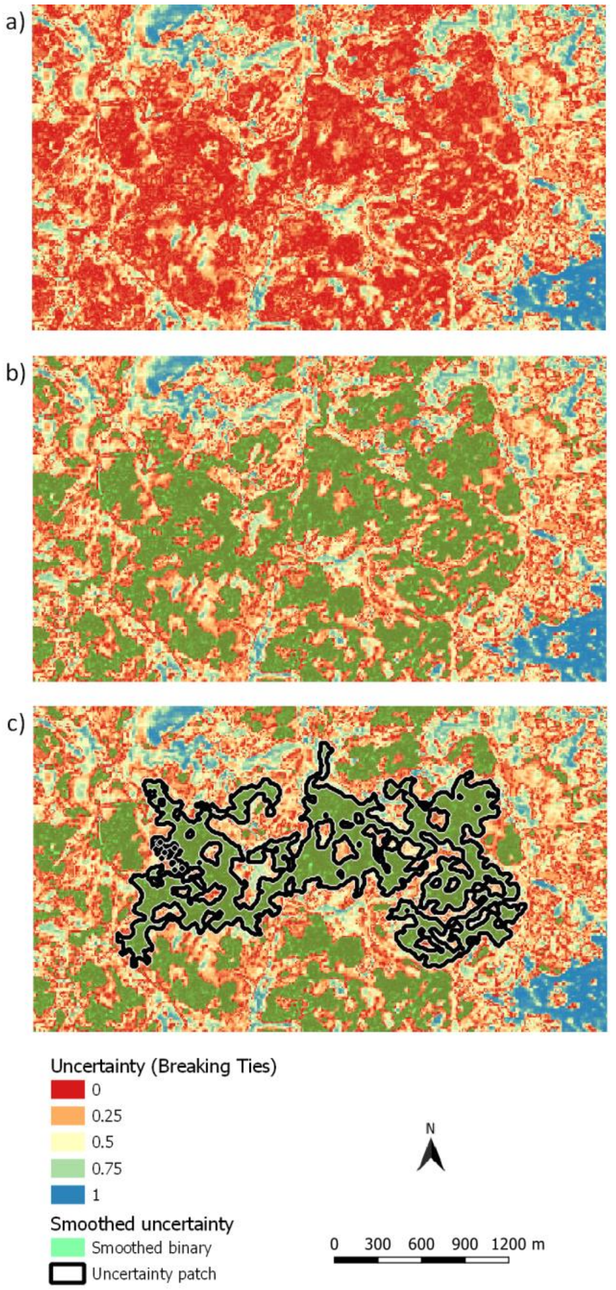

Figure 4. Delineation of an uncertainty patch: a) uncertainty distribution; b) result of the application of threshold followed by smoothing (moving window) in green; c) delineation of a contiguous uncertainty patch.

Polygons were manually delimited and labelled within these patches, assisted by the 25cm orthophotomap from 2018, to collect additional training samples units to be introduced to the initial training dataset. Since it is very unlikely to correctly identify crop types by visual interpretation of the orthophoto, patches or parts of patches located on top of cropland areas were ignored, which means that no additional polygons were delimited for the classes of agricultural crops.

Map Class	Training Class	Labelled	
		Polygons	Sampling units
BUP	Built up	4	40
	Industrial	-	-
	Road Network	-	-
AGR	Oat	-	-
	Wheat	-	-
	Barley	-	-
	Ryegrass	-	-
	Triticale	-	-
	Rye	-	-
	Corn	-	-
	Sunflower	-	-
	Managed Grasslands	5	308
NGL	Agric. Nat. Grassland	23	3722
	Mount. Nat. Grassland	2	220
EUC	Eucalyptus Adult	6	579
OBL	Other Broadleaf	18	1906
MTP	Maritime Pine	9	769
OCF	Other Coniferous	20	1418
SBL	Dense Shrubland	47	4195
NVS	Baresoil	27	4782
	Bare Rock	1	48
WTR	Water	18	1812

Table 1. Class nomenclature and additional sample units extracted from areas of high classification uncertainty. BUP: Built up, AGR: Agriculture, NGL: Natural Grasslands, EUC: Eucalyptus, OBL: Other Broadleaf, MTP: Maritime Pine, OCF: Other Coniferous, SBL: Shrubland, NVS: Non-vegetated Surfaces, WTR: Water.

Since the size of the uncertainty patches might vary, training classes may have a different number of additional training sampling units available. Therefore, it is important to mind class balance when adding new sampling units. Moreover, it is preferable to incorporate the additional sampling units into an initial sample of compatible size in order to ensure that the additional units can have an influence over the representativeness of the aggregated sample. With this in mind, and considering the availability of additional sampling units (Table 1), we decided to add up to 500 sampling units per class to a subset of the initial sample. The subset of the initial sample consisted of 500 units per class. Therefore, our new training dataset had up to 1000 sampling units per class.

Finally, a new classification was performed using the new training dataset. An accuracy assessment was conducted with an independent stratified random validation dataset with 535 sampling units labelled through visual interpretation of the previous orthophotomap. The results of the initial and new classification were compared in order to evaluate whether the additional training samples increased classification accuracy.

4. RESULTS

The overall accuracies of the initial and new classifications are shown in Table 2. The new classification, performed with additional sampling units in the training dataset, exhibited an overall accuracy of 69.72%, with only a small increase in accuracy compared to the initial classification. According to the confidence intervals (3.9%), the increase in accuracy was considered not statistically significant.

Classification	Overall accuracy (%)
Initial	68.78 ± 3.9
New	69.72 ± 3.9

Table 2. Accuracy assessment of the classifications.

The analysis of the accuracy metrics per class (Table 3) indicates that the new classification was advantageous to seven out of the 10 classes, although the degree of improvement varied. The addition of new sampling units was most beneficial for natural grasslands and eucalyptus, which had an increase of 13% and 19.3% in F1-score, respectively. The other five classes had only a small increase in F1-score. It is noticeable, however, that despite the increase in F1-score, most of these classes exhibited a tradeoff between reduction and growth in commission and omission errors. On the other hand, agriculture, shrubland and non-vegetated surfaces exhibited a decrease in accuracy, with the latter having a reduction of 16.5% in terms of F1-score.

Class	Precision (%)		Recall (%)		F1-score (%)	
	Initial	New	Initial	New	Initial	New
BUP	50.0	56.5	93.8	86.7	65.2	68.4
AGR	55.0	63.6	91.7	72.9	68.8	68.0
NGL	80.0	81.6	36.4	51.3	50.0	63.0
EUC	83.3	64.7	22.7	47.8	35.7	55.0
OBL	97.4	95.7	57.8	64.3	72.6	76.9
MTP	76.0	81.6	67.9	65.1	71.7	72.5
OCF	27.0	24.6	35.7	51.9	30.8	33.3
SBL	68.9	67.4	85.9	84.8	76.5	75.1
NVS	46.2	29.4	85.7	83.3	60.0	43.5
WTR	100	100	98.0	100	99.0	100

Table 3. Precision, recall and F1-score of both classifications. BUP: Built up, AGR: Agriculture, NGL: Natural Grasslands, EUC: Eucalyptus, OBL: Other Broadleaf, MTP: Maritime Pine, OCF: Other Coniferous, SBL: Shrubland, NVS: Non-vegetated Surfaces, WTR: Water.

A few problems were detected regarding the delimitation and labelling of polygons in the selected uncertainty patches. Besides not being possible to label the type of crops based on visual interpretation of the orthophoto, no uncertainty patches corresponded to classes industrial or road network. As a result, all these classes did not have additional training sampling units. Moreover, the amount of training sampling units available after labelling varied depending on the class. These conditions resulted in an unbalanced training dataset, which might have contributed to prevent a higher increase in classification accuracy. For instance, training classes built up and bare rock had only 40 and 48 additional sampling units, respectively.

Furthermore, collecting additional sample units from a small number of polygons (e.g. built up, mountain natural grassland, eucalyptus adult and bare rock) might have resulted in acquiring redundant new samples, which could have contributed to limit the spectral diversity, thus potentially reducing the impact on classification accuracy.

Besides the accuracy assessment, a visual inspection was conducted to evaluate the effects of the new classification on the map. The analysis of the maps revealed that the additional samples may have been responsible for certain improvements, as illustrated in Figure 5. In this example, the new classification mapped more accurately an area identified as broadleaf forest according to COS, reducing not only the confusion between other broadleaf and other species but also between other broadleaf and agriculture.

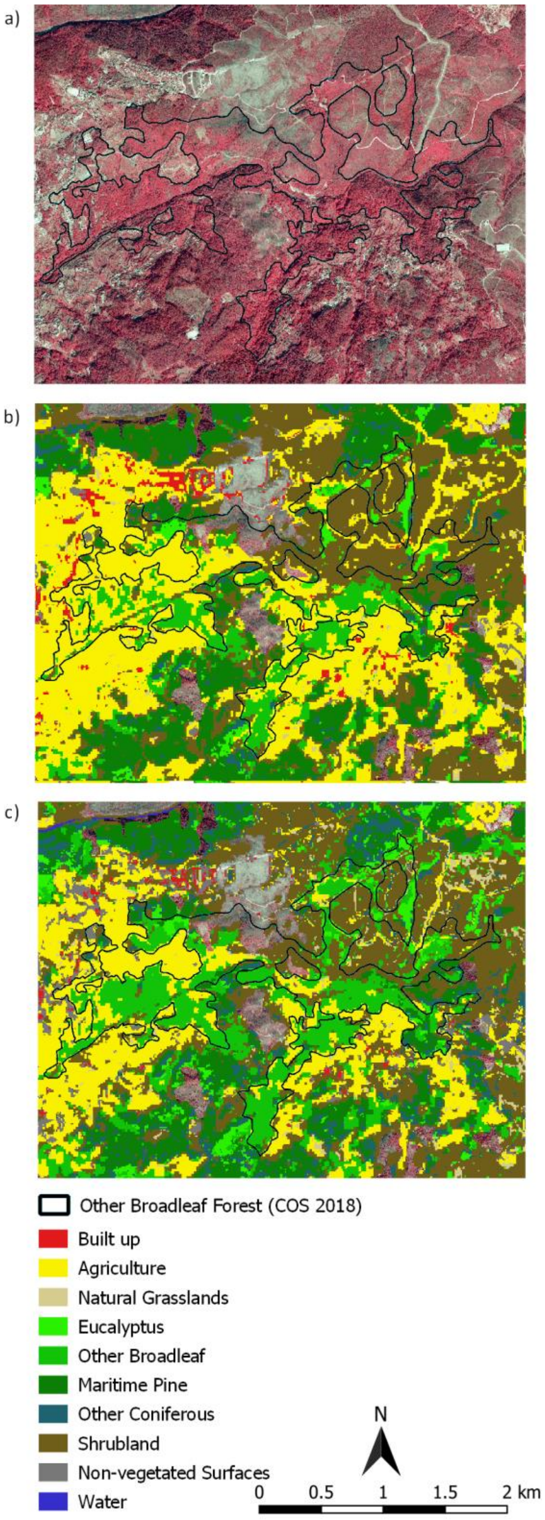

Figure 5. Highlight of the classification of other broadleaf – a) false color orthophoto; b) reference classification; c) new classification.

Another example of improvements is exhibited in Figure 6, where areas incorrectly mapped as built up in the initial classification were correctly mapped as non-vegetated surfaces in the new classification. Additionally, the new classification also exhibited less confusion between agriculture and shrubland.

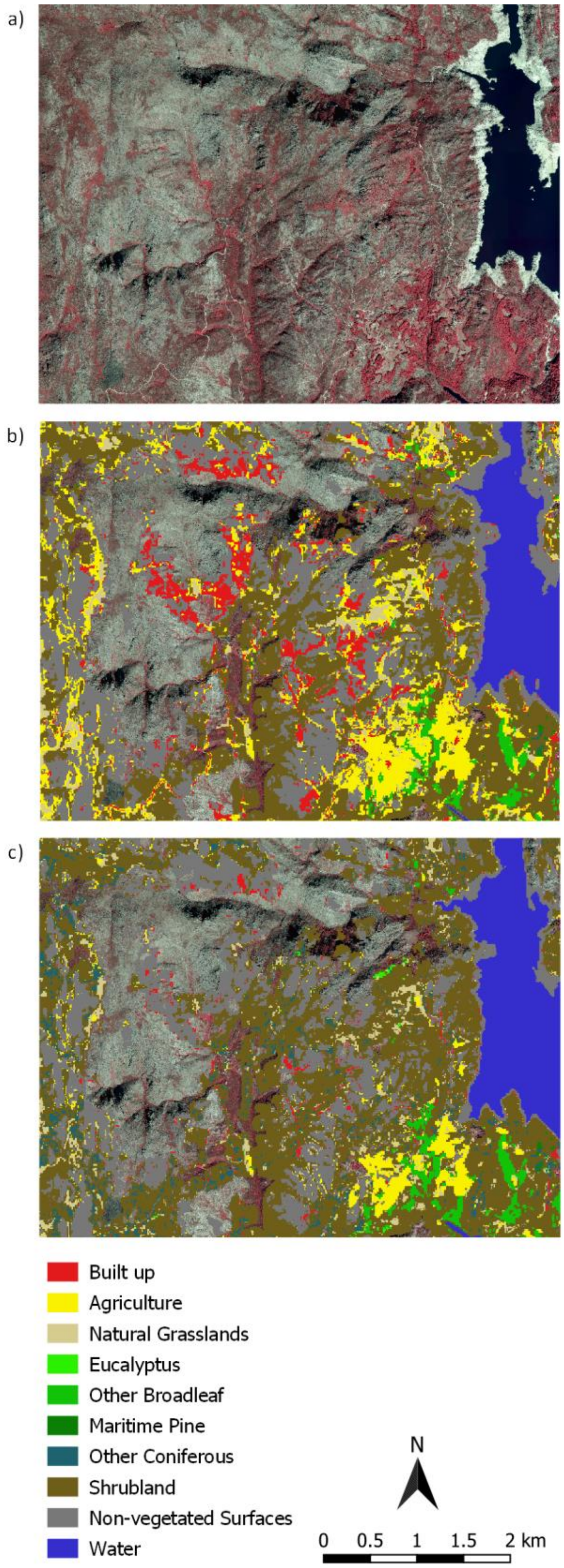

Figure 6. Reduction of misclassifications after introducing additional samples – a) false colour orthophoto of a mountainous area; b) reference classification; c) new classification.

5. CONCLUSIONS

This work proposed to explore the use of classification uncertainty to extract additional training data within areas of high classification uncertainty, as an attempt to improve classification accuracy. The accuracy assessment exhibited only a small increase in performance, which was considered not statistically significant. Such result might have been caused by limitations in the methodology, especially related to the impossibility of labelling additional sampling units corresponding to the agricultural crops. Moreover, the unequal distribution of additional training sampling units among classes may also have limited the improvements. Another factor which can be considered detrimental to the results was the collection of new training sampling units from a limited number of polygons, which may cause the additional sample data to be redundant.

Besides the accuracy assessment, a visual inspection of the classification maps was conducted, revealing important improvements in some classes through reducing the confusion between classes with similar spectral patterns, as is the case of built up and non-vegetated surfaces.

Future studies can further explore the potential of uncertainty to improve classification performance, especially addressing the issues involving the impossibility of labelling new sampling units as well as proposing alternatives to ensure that a sufficient amount of training data per class is collected. This could be achieved by modifying the uncertainty threshold or by increasing the number of uncertainty patches. Furthermore, special attention can be dedicated towards including a spatial criterion in the query strategy in order to select new samples from spatially disperse polygons.

ACKNOWLEDGEMENTS

This work has been supported by project foRESTER (PCIF/SSI/0102/2017), SCAPEFIRE (PCIF/MOS/0046/ 2017),and by Centro de Investigação em Gestão de Informação (MagIC), all funded by the Portuguese Foundation for Science and Technology (FCT). Value-added data processed by CNES for the Theia data centre www.theia-land.fr using Copernicus products. The processing uses algorithms developed by Theia's Scientific Expertise Centres.

REFERENCES

Ahmad, M., Khan, A., Khan, A. M., Mazzara, M., Distefano, S., Sohaib, A., Nibouche, O., 2019: Spatial prior fuzziness pool-based interactive classification of hyperspectral images. *Remote Sensing*, 11:9, 1136.

Belgiu, M., 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.

Costa, H., Benevides, P., Marcelino, F., Caetano, M., 2020: Introducing automatic satellite image processing into land cover mapping by photo-interpretation of airborne data. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci*. XLII-3/W11, 29–34.

Crawford, M. M., Tuia, D., Yang, H. L., 2013: Active learning: Any value for classification of remotely sensed data?. *Proceedings of the IEEE*, 101(3), 593-608.

Li, J., Bioucas-Dias, J. M., Plaza, A., 2013: Spectral–spatial classification of hyperspectral data using loopy belief propagation and active learning. *IEEE Transactions on Geoscience and Remote Sensing*, 51:2, 844-856.

Mack, B., Leinenkugel, P., Kuenzer, C., Dech, S., 2017: A semi-automated approach for the generation of a new land use and land cover product for Germany based on Landsat time-series and Lucas in-situ data. *Remote Sensing Letters*, 8(3), 244-253.

Maxwell, A.E., Warner, T.A., Fang, F., 2018: Implementation of machine-learning classification in remote sensing: an applied review. *International Journal of Remote Sensing*, 39(9), 2784-2817.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011: Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.

Tuia, D., Volpi, M., Copa, L., Kanevski, M., Munoz-Mari, J., 2011: A survey of active learning algorithms for supervised remote sensing image classification. *IEEE Journal of Selected Topics in Signal Processing*, 5(3), 606-617.

Wulder, M. A., Coops, N. C., Roy, D. P., White, J. C., Hermosilla, T., 2018: Land cover 2.0. *International Journal of Remote Sensing*, 39:12, 4254-4284.

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLIII-B3-2021-81-2021 | © Author(s) 2021. CC BY 4.0 License.