A CONTRIBUTION TO LAND COVER AND LAND USE MAPPING
in Portugal with multi-temporal Sentinel-2 data and supervised classification

Daniel Moraes

Dissertation submitted in partial fulfilment of the requirements for the Degree of Mestre em Ciência e Sistemas de Informação Geográfica (Master in Geographical Information Systems and Science)

NOVA Information Management School

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

A CONTRIBUTION TO LAND COVER AND LAND USE MAPPING

in Portugal with multi-temporal Sentinel-2 data and supervised classification

By

Daniel Moraes

Dissertation submitted in partial fulfillment of the requirements for the Degree of Mestre em Ciência e Sistemas de Informação Geográfica (Master in Geographic Information Systems and Science)

Supervisor: Professor Ph.D. Mário Sílvio Rochinha de Andrade Caetano

Co-supervisor: Ph.D. Pedro Benevides

February 2021

DECLARATION OF ORIGINALITY

I declare that the work described in this document is my own and not from someone else. All the assistance I have received from other people is duly acknowledged and all the sources (published or not published) are referenced.

This work has not been previously evaluated or submitted to NOVA Information Management School or elsewhere.

Lisbon, February 2021

Daniel Moraes

ACKNOWLEDGEMENTS

I would like to thank my supervisor Prof. Dr. Mário Caetano for the guidance, patience and all the hard work. His contributions were essential to the development of this thesis. I would also like to thank Dr. Pedro Benevides, my co-supervisor, who provided great assistance during this process. In addition, I am grateful to all the people at Direção-Geral do Território (DGT), especially Hugo Costa and Francisco Moreira, who also collaborated in this work.

Also, I would like to thank my wife Letícia for the all the love and constant support. I love you.

Finally, I wish to thank my parents for making it all possible and for always teaching me how to be a better professional and human being.

This thesis was developed under the framework of the project SCAPEFIRE: A SUSTAINABLE LANDSCAPE PLANNING MODEL FOR RURAL FIRES PREVENTION - PCIF/MOS/0046/2017 funded by Fundação para a Ciência e Tecnologia.

A CONTRIBUTION TO LAND COVER AND LAND USE MAPPING

in Portugal with multi-temporal Sentinel-2 data and supervised classification

ABSTRACT

Remote sensing techniques have been widely employed to map and monitor land cover and land use, important elements for the description of the environment. The current land cover and land use mapping paradigm takes advantage of a variety of data options with proper spatial, spectral and temporal resolutions along with advances in technology. This enabled the creation of automated data processing workflows integrated with classification algorithms to accurately map large areas with multi-temporal data. In Portugal, the General Directorate for Territory (DGT) is developing an operational Land Cover Monitoring System (SMOS), which includes an annual land cover cartography product (COSsim) based on an automatic process using supervised classification of multi-temporal Sentinel-2 data. In this context, a range of experiments are being conducted to improve map accuracy and classification efficiency. This study provides a contribution to DGT's work. A classification of the biogeographic region of Trás-os-Montes in the North of Portugal was performed for the agricultural year of 2018 using Random Forest and an intra-annual multi-temporal Sentinel-2 dataset, with stratification of the study area and a combination of manually and automatically extracted training samples, with the latter being based on existing reference datasets. This classification was compared to a benchmark classification, conducted without stratification and with training data collected automatically only. In addition, an assessment of the influence of training sample size in classification accuracy was conducted. The main focus of this study was to investigate whether the use of

classification uncertainty to create an improved training dataset could increase classification accuracy. A process of extracting additional training samples from areas of high classification uncertainty was conducted, then a new classification was performed and the results were compared. Classification accuracy assessment for all proposed experiments was conducted using the overall accuracy, precision, recall and F1-score. The use of stratification and combination of training strategies resulted in a classification accuracy of 66.7%, in contrast to 60.2% in the case of the benchmark classification. Despite the difference being considered not statistically significant, visual inspection of both maps indicated that stratification and introduction of manual training contributed to map land cover more accurately in some areas. Regarding the influence of sample size in classification accuracy, the results indicated a small difference, considered not statistically significant, in accuracy even after a reduction of over 90% in the sample size. This supports the findings of other studies which suggested that Random Forest has low sensitivity to variations in training sample size. However, the results might have been influenced by the training strategy employed, which uses spectral subclasses, thus creating spectral diversity in the samples independently of their size. With respect to the use of classification uncertainty to improve training sample, a slight increase of approximately 1% was observed, which was considered not statistically significant. This result could have been affected by limitations in the process of collecting additional sampling units for some classes, which resulted in a lack of additional training for some classes (eg. agriculture) and an overall imbalanced training dataset. Additionally, some classes had their additional training sampling units collected from a limited number of polygons, which could limit the spectral diversity of new samples. Nevertheless, visual inspection of the map suggested that the new training contributed to reduce confusion between some classes, improving map agreement with ground truth. Further investigation can be conducted to explore more deeply the potential of classification uncertainty, especially focusing on addressing problems related to the collection of the additional samples.

KEYWORDS

Sentinel-2

Land Cover Mapping

Random Forest

Intra-annual time series

Image Classification

Training Sample Size

Classification Uncertainty

Portugal

ACRONYMS

COS – Carta de Uso e Ocupação do Solo (Land use and land cover cartography)

CV – Coefficient of Variation

DGT – Direção-Geral do Território – General Directorate for Territorial Management

HRL – High-Resolution Layers

ICNF – Instituto da Conservação da Natureza e das Florestas – Institute for Nature Conservation and Forests

IFAP – Instituto de Financiamento da Agricultura e Pescas – Agriculture and Fisheries Financing Institute

LCLU – Land Cover Land Use

LPIS – Land Parcel Identification System

INDEX OF THE TEXT

INDEX OF TABLES

INDEX OF FIGURES

1. INTRODUCTION

1.1. Background

Land cover is considered an element of extreme relevance for the description and study of the environment (Herold *et al.*, 2006), therefore it is necessary to quantify and map land cover and its changes over time. Land cover can be defined as a descriptor of Earth's terrestrial surface, important to characterize anthropogenic activity, biogeographical and eco-climatic diversity and is often mapped in conjunction with land use, a descriptor of how humans use the land (Wulder *et al.*, 2018). Land cover and its derived products can benefit society in a variety of areas, such as disasters, climate, water, agriculture, among other areas listed by the Group on Earth Observations (GEO) (Wulder *et al.*, 2008). In terms of sustainable development, land cover and land use (LCLU) data are decisive to combat land degradation and promote sustainable land management (Anderson *et al.*, 2017). Therefore, it is essential to develop methods to gather such valuable information.

Remote sensing techniques have been widely adopted to map and monitor land cover (Cihlar, 2000) since they can provide data in a variety of spatial and temporal scales (Gómez et al., 2016). While early works, which date back to the 1970s, were limited due to the quality of the information (Townshend, 1992) and less advanced and costly technology (Cihlar, 2000), the current land cover mapping paradigm benefits from a variety of data options with adequate spatial, spectral and temporal resolutions and the developments in technology (Wulder *et al.*, 2018). Nowadays, automated data processing workflows integrated with state-of-the-art machine learning classification algorithms allow to accurately map land cover of large areas using multi-temporal imagery (Inglada *et al.*, 2017; Hermosilla *et al.*, 2018). Although these conditions contributed to the creation of some operational LCLU monitoring programs (Wulder *et al.*, 2008), most countries, including Portugal, still do not have any. Therefore, there is a demand to further develop LCLU mapping methodologies so that more countries can implement their monitoring programs.

1.2. Problem

In Portugal, the General Directorate for Territory (DGT), the National Reference Center for Land Cover of the European Environment Agency, is developing a new project: the Land Cover Monitoring System (SMOS). SMOS results from the integration of three products. The first product is the already consolidated domestic LCLU cartography, *Carta de Uso e Ocupação do Solo (COS)*, which was first released in 1995. The production of COS is based on visual

interpretation of orthophotos, a manual process with a high cost in terms of time and resources. Such cartography adopts a nomenclature of 83 classes and is available in vector format, with a minimum mapping unit (MMU) of 1 ha, having a periodicity of 3 years. The second product that integrates SMOS is a simplified COS called COSsim, which aims to map land cover exclusively in a yearly basis. COSsim's production relies on an automatic process based on supervised classification of Sentinel-2 images, though DGT has also been conducting research to evaluate whether combining data from Sentinel-1 and Sentinel-2 can increase mapping accuracy. COSsim provides data in raster format with a MMU of 100 m² (10 m x 10 m pixel) in a nomenclature of 13 classes and with annual periodicity, therefore having greater spatial detail and periodicity when compared to COS. The third SMOS product is the Vegetation State Intra-annual Map (MIAEV), which aims to monitor the conditions of the vegetation. Such product is also generated by an automatic process based on Sentinel-2 imagery, thus being available in raster format with 100 m² MMU. MIAEV provides monthly continuous values of the vegetation state.

The automatic methodology adopted by DGT to produce COSsim employs state-of-the-art machine learning supervised classification algorithms, which have been widely used to map LCLU. Recent studies indicate a preference for supervised algorithms in particular, as they tend to yield higher accuracies when compared to unsupervised methods (Maxwell *et al.*, 2018; Yu *et al.*, 2014). In spite of the classification being automatic, supervised classification requires collecting training samples, a traditionally human dependent activity, thus costly and time consuming. Since training samples can have a significant impact on classification accuracy, it is important to dedicate special attention to the process of sample collection. In this context, DGT is experimenting a process of automatic sample extraction from existing databases in the production of COSsim (Hernandez *et al.*, 2020).

In order to produce LCLU maps with adequate accuracy, DGT has been using time-series images. Multi-temporal intra-annual imagery can improve land cover classification accuracy (Townshend, 1992; Griffiths *et al.*, 2019), since the collection of seasonal variability data can help distinguish land cover classes related to vegetation, e.g. forest and crops (Gómez *et al.*, 2016). The characteristics of Landsat products and the availability of its extensive open access time series imagery (Wulder *et al.*, 2018; Gómez *et al.*, 2016) made Landsat the main source of data for land cover classification. However, the satellite's revisit time of 16 days generally provide an insufficient data availability for a variety of locations, especially in regions subjected

to constant cloud cover, what prevents conducting proper inter and intra-annual analysis using medium resolution data over large areas (Gómez *et al.*, 2016).

The launch of the European Space Agency (ESA) Sentinel-2 mission brought a systematic global coverage, with a 5 day revisit time, high spatial resolution (10 to 60m) and an appropriate spectral band range (Drusch *et al.*, 2012). The more frequent revisit time means an increase in the number of observations and thus a higher probability of acquiring cloud free images, what supports creating intra-annual time-series. Despite the opportunities brought by more observations, there are still some challenges to be met. Incorporating vast time-series into the classification results in more predictor variables. Although additional predictor variables might help separate distinct classes, they also increase the dimensionality and complexity of the feature space, which might result in a decrease in classification accuracy. This issue happens because the number of training sampling units is insufficient to describe the complexity of the feature space, which is called the Hughes phenomenon (Maxwell *et al.*, 2018). Therefore, multi-temporal land cover classification might require collecting a sufficiently large training sample.

Although the higher dimensionality of the feature space demands a larger training sample, it is unclear how changes in sample size can influence classification accuracy. Overall, literature lacks advice on the minimum sample size, however, there is a broad understanding that increasing sample size results in higher classification accuracy (Maxwell *et al.*, 2018). Whilst the sensitivity to sample size was evaluated for classifications with reduced predictor variables (Rodríguez-Galiano *et al.,* 2012; Thanh Noi and Kappas, 2018; Huang *et al.*, 2002), which indicated small sensitivity, further investigation is required for classifications with a large number of variables. Identifying a minimal or optimum number of training sampling units can be an aspect that facilitates sample collection works.

In addition, training sample quality, i.e. the representativeness of each class, and class balance can influence classification performance (Maxwell *et al.*, 2018). Thus, it is necessary to not only collect a proper amount of training data but also to observe its quality and class balance, while also considering the feasibility in terms of cost and resources dedicated to such task.

Multiple works use existing reference datasets to draw training samples automatically (Leinenkugel *et al.*, 2019; Griffiths *et al.*, 2019; Inglada *et al.*, 2017). However, despite the encouraging results, classification accuracy still needs to be improved, especially in the case of classes which carry errors or lack quality from the reference dataset, for instance, classes whose spectral response is heterogeneous, having contributions of various cover types. Poor training

sampling can also cause confusion in distinguishing some similar classes. In order to address these issues, DGT is experimenting the implementation of a stratification of the study area based on reference cartography to group regions that share the same land cover characteristics. Additionally, DGT is also investigating whether introducing manual training for the classes that normally exhibit low accuracies when automatically sampled can improve classification accuracy.

Since one of the major causes of classification inaccuracy is a certain degree of imperfection in the training dataset, it is convenient to develop means to improve it. Active Learning represents an alternative to enhance the performance of the classification through the evaluation of the classification uncertainty, which can help identify areas where the classification was the most uncertain, so that additional training samples could be collected manually in these areas (Tuia *et al.*, 2011). Training samples collected from areas of high uncertainty are considered difficult examples, therefore having potential to improve the model. The additional training data is then included in the training dataset to produce a reinforced model. Active Learning approaches begin with a small number of training sampling units, which gradually increases after adding new well-chosen sampling units in each cycle. Then, one can achieve good classification performance using a much smaller training dataset. In spite of the good results, Active Learning approaches are based on recurrent interactions between the model and the analyst, with the latter being responsible for manually collecting and labeling new sampling units. In contrast to such demanding human intervention, classification uncertainty can assist refining classifiers trained on a large training dataset simply by providing a single set of additional training sampling units collected in areas of high uncertainty (Mack *et al.*, 2017).

This thesis was developed in collaboration with DGT within the context of experiments with the COSsim methodology, currently in implementation in other regions. Besides assessing whether the methodology can be suitable to the specificities of other type of landscape, characterized mainly by mountainous land occupied with rocks, forest and bushes, the purpose of this thesis is to contribute to improve the methodology of COSsim production in terms of efficiency and accuracy, suitable for the production in subsequent years. The approach adopted in this work innovates by introducing manual training samples, stratification of the area to be mapped and most importantly the improvement of the training dataset using classification uncertainty. Furthermore, a series of experiments were conducted in order to assess the impact of training sample size in classification accuracy.

1.3. Research Question

Mapping LCLU by performing supervised classification of multi-temporal imagery requires adequate training sample size and quality. Most studies have not investigated the effect of variations in training sample size on classifications with a large number of predictor variables. Additionally, further investigation can be conducted regarding the use of classification uncertainty to produce a new training dataset of enhanced quality. This study proposes to classify LCLU using multi-temporal Sentinel-2 data and to assess the sensitivity of classifications with a large number of predictor variables to variations in training sample size by comparing classification performance in different sample size scenarios. Moreover, the study proposes to evaluate whether using classification uncertainty to generate a new training sample can increase performance. The new training sample is created by incorporating additional sampling units extracted from areas of high classification uncertainty.

Therefore, the following research questions are proposed:

- Can the stratification of the mapping area and the introduction of manual training samples for classes in which the automatic training performs poorly improve classification accuracy?
- Can variations in the sample size affect the performance of classification of high dimensionality?
- Can the introduction of new training samples collected from areas of high classification uncertainty improve classification performance?

1.4. Thesis structure

This thesis is organized according to the following sections:

- Section 2 - Literature Review: a review and discussion of relevant works focused on land cover mapping, temporal compositing, machine learning classification, random forest classifier, effects of training sample size on classification accuracy, use of reference data to extract training samples and classification uncertainty in land cover mapping.
- Section 3 – Study Area and Data: a description of the study area and datasets utilized.
- Section 4 – Methods: a comprehensive description of the methods employed.
- Section 5 – Results and Discussion: presents the results based on classification accuracy metrics and other statistics and analyzes the results in the light of the literature.
- Section 6 – Conclusion: a summary of the research, with the main findings, limitations and recommendations for future works.

2. LITERATURE REVIEW

2.1. Land cover mapping

Land cover refers to the biophysical characteristics of the Earth's terrestrial surface, including vegetation, water, bare soil and anthropogenic structures (Gómez *et al.*, 2016). Land cover can outline the existing functional relationship between the terrain, climate and soils, thus providing an overview of the environment and the factors that induce changes. It is also relevant to describe anthropogenic activity and biogeographical and eco-climatic diversity (Wulder *et al.*, 2018), hence being considered the most significant element for the description and study of the environment (Herold *et al.*, 2006) and an indispensable climate variable (GCOS, 2003). Changes in land cover have a considerable influence in climate change processes, especially in the case of deforestation, which is a major anthropogenic source of carbon dioxide (Anderson *et al.*, 2017). In addition, Hermosilla *et al.* (2018) pointed out that changes in land cover also heavily impacts on hydrology and global biophysical and biogeochemical cycles of the terrestrial surface. They consider land cover and land cover change to be crucial information in the process of monitoring Earth ecosystems and capable of providing insights about their status and tendencies. Furthermore, land cover products can contribute to the Earth observation societal benefits presented by GEO in nine different areas: disasters, health, energy, climate, water, weather, ecosystems, agriculture and biodiversity (Wulder *et al.*, 2008). It is also acknowledged that land cover data is decisive to achieve the target of neutral land degradation and to promote sustainable land management (Anderson *et al.*, 2017).

Remote sensing techniques emerged as an opportune alternative to map land cover, mainly since the availability of Landsat 1 data, which prompted the use of satellite data for numerous studies involving mapping land cover (Cihlar, 2000). However, the early researches, which were developed in a time when Landsat-TM and SPOT-HRV were the main source of data, found some limitations regarding the quality of the information extracted by the sensors in terms of spectral, spatial and temporal resolutions, affecting the ability to distinguish the different cover types of interest (Townshend, 1992). In terms of spectral limitations, Townshend (1992) considered the TM and SPOT-HRV broad spectral bands sensors to hinder the ability to separate multiple cover types spectrally. In addition, the SPOT-HRV did not have the valuable short wave infrared band, important for vegetation characterization. Concerning the spatial resolution, the author views the relatively fine resolution as insufficient to capture the details in the context of urban applications. In spite of such criticism, he highlights the importance of the detailed view provided by the aforementioned sensor to map and monitor land cover. With respect to the

temporal resolution, he affirms that the use of multi-temporal images can improve performance, although the recurring presence of clouds makes it difficult to acquire multiple usable images. The author mentions the use of data provided by short observation intervals from sensors such as the National Oceanic and Atmospheric Administration (NOAA) Advanced Very High Resolution Radiometer (AVHRR) as an attempt to increase the number of cloud-free observations, however, these sensors' coarse spatial resolution fails to capture important spatial detail. Other limitations at the time were related to the lack of technology and cost of data storage, which constrained most of the studies with fine resolution data to limited areas (Cihlar, 2000).

Currently, the variety of data options in terms of spatial, spectral and temporal resolutions combined with the advances in data storage, computing processing and classification algorithms contributed to create a new land cover mapping paradigm (Wulder *et al.*, 2018). Open access and analysis-ready satellite imagery also play an important role in this new paradigm. It is possible to automate data processing and use advanced classification algorithms, typically based on signature-extension methods, to produce accurate land cover maps of large areas using multi-temporal images (Hermosilla *et al.*, 2018). These encouraging conditions favored the development of some operational land cover mapping programs. In spite of the good efforts of the aforementioned initiatives, they represent an exception, since most countries do not have remote sensing based land cover mapping operational programs. In addition, some programs, e.g. the Portuguese COS and European CORINE, rely on mapping land cover through computer aid photointerpretation (DGT, 2018; Bossard *et al.*, 2000), which is a costly and time consuming approach. Therefore, it is necessary to expand land cover mapping operational programs as well as to develop automated mapping approaches.

Landsat has been the main source of data for land cover classification due to its convenient spatial detail (30m), multi decade image archive, radiometric calibration, open access and capacity of covering large areas (185 x 185 km) (Wulder *et al.*, 2018; Gómez *et al.*, 2016). In spite of that, the availability of Landsat data for a range of locations, especially in constantly cloudy areas, is often considered insufficient and inadequate for both inter and intra-annual analysis (Gómez *et al.*, 2016). The last authors outline that even though data availability can be increased by adopting compositing strategies, large areas remain discontinuous to some extent.

The ESA Sentinel-2 mission, operating with two identical satellites (Sentinel-2A and Sentinel-2B, launched in 2015 and 2017, respectively), provides an unprecedented combination of

systematic global coverage, frequent revisit time of 5 days (considering both satellites), high spatial resolution (10, 20 or 60m depending on the spectral band) and 13 spectral bands including visible, near infrared and short wave infrared (Drusch *et al.*, 2012). The shorter revisit time translates into a higher number of observations, which boosts the likelihood of acquiring cloud free data, hence favoring the formation of multi-temporal intra-annual composites. Additionally, the data can be downloaded with no cost. The aforementioned characteristics make Sentinel-2 a valuable and adequate source of data for mapping and monitoring land cover. Multiple studies successfully mapped land cover and land use utilizing Sentinel-2 imagery (Weigand *et al.*, 2020; Paris *et al.*, 2019). Nevertheless, some of them do not take advantage of the full potential of the constellation (Sentinel-2A and 2B) revisit time of 5 days (Close *et al.*, 2018; Griffiths *et al.*, 2019; Vuolo *et al.*, 2018), as data from Sentinel-2B was not available at the time. As Griffiths *et al.* (2019) outlines, within the agricultural domain temporal information is crucial to distinguish different crop types. The authors claim that intra-annual observations are required to record the differences in seasonal growing characteristics of a determined crop. In this context, the work by Vuolo *et al.* (2018), despite using only Sentinel-2A observations, demonstrates that multi-temporal Sentinel-2 data can improve crop type classification.

In spite of Sentinel-2's increased revisit time being an opportunity, it also represents a challenge, as more images are provided. The availability of vast multi-temporal data can be translated into a growing number of predictor variables, e.g. in the form of statistical metrics (Gómez *et al.*, 2016). Whilst adding more predictor variables can enhance the classification ability of separating classes, it also expands the dimensionality and complexity of the feature space, which might result in a decreased classification accuracy. This occurs because the amount of training data becomes insufficient to describe the overly complex and high dimensional feature space, which is known as the Hughes phenomenon. Such situation is more critical in supervised classifications with a small number of training sampling units (Maxwell *et al.*, 2018). Therefore, the additional predictor variables brought by using multiple time-series might require the collection of a large number of training sampling units.

2.2. Temporal compositing

Sentinel-2 temporal resolution facilitates the use of multi-temporal data. According to Wulder *et al.* (2018), most of the current studies in land cover classification derive spectral band and indices from image time series. As Townshend (1992) and Griffiths *et al.* (2019) observed, acquiring remotely sensed data in an intra-annual frequency could be important to improve

classification performance. However, the latter mention that achieving such frequency depends strongly on atmospheric conditions, i.e. clear sky, since cloud cover can mask relevant phases of crop development. Furthermore, the authors and Hermosilla *et al.* (2018) pointed out that cloud shadow also prevent obtaining consistent pixel values. Besides cloud-related obstacles, Griffiths *et al.* (2013) also cite discontinuity in image archives and data or sensor related errors as other issues that might affect data availability. Therefore, it is necessary to implement strategies to minimize such data losses.

Image compositing within a regular time window or period represents an alternative to address the data availability problem. Within a pixel-based approach (pixel-based compositing), analysis are not limited to few images with satisfying cloud cover. Instead, information availability grows since clear pixels that belong to a cloudy image can be computed (Griffiths *et al.*, 2013).

Griffiths *et al.* (2019) explored the concept of temporal compositing as an option to increase data availability. The authors mention various existing approaches, most of them based on best-pixel selection, also known as best available pixel. The selection criteria are numerous, for instance, selecting the pixels based on the maximum Normalized Difference Vegetation Index (NDVI), selecting the median of a single band or index and a selection based on parametric scoring. According to Hermosilla *et al.* (2015), the selection aims to not only exclude pixels affected by cloud and cloud shadow, haze or sensor related problems but also include pixels that meet users' particular requirements, e.g. closeness to a target day-of-year.

Although pixel compositing might improve data availability, there could still be pixels that do not meet all the requirements, hence forming data gaps. In this context, Hermosilla *et al.* (2015) propose using proxy composites. This approach consists in filling data gaps according to the complete spectral information of the pixel series, which allows deriving artificial pixel values. This method is similar to what is presented in the work conducted by Inglada *et al.* (2017). Their procedure also consists in characterizing pixels by a time series of image features. In order to fill data gaps, they perform a linear interpolation using the prior and following cloud-free dates (Inglada *et al.*, 2015). This gap filling and resampling process yields a set of virtual acquisition dates, which allows choosing common dates for all pixels when proceeding to feature extraction.

2.3. Use of reference data to extract training samples

As previously discussed, supervised classification depends on collecting training samples. However, such task might be costly in terms of time and resources, especially considering that a sufficiently large, representative and balanced training dataset should be sought. Moreover,

the increasing availability of multi-temporal data contributes to expand the dimensionality of the feature space, which demands an even larger training sample. Therefore, multiple studies have adopted the strategy of using existing reference data to collect training samples automatically (Inglada *et al.*, 2017; Leinenkugel *et al.*, 2019; Pflugmacher *et al.*, 2019; Griffiths *et al.*, 2019).

Whilst some works use a single reference dataset to extract training samples (Pflugmacher *et al.*, 2019; Hermosilla *et al.*, 2018), others use a combination of different reference dataset (Leinenkugel *et al.*, 2019; Inglada *et al.,* 2017; Griffiths *et al.*, 2019). The latter approach can be justified due to the availability of more reliable datasets covering specifics classes, such as agricultural and urban.

However, Foody *et al.* (2016) remarked that a range of problems may arise due to using reference data from different sources. They outlined that problems might be caused by the different acquisition methods of each source. For instance, sampling efforts may differ, what might result in imbalanced sampling across regions and thus impact class balance.

In addition, the authors also discussed the normal errors associated with reference datasets. They mentioned errors made due to mislabeling, which they believe can be caused either by usual typographical and transcription errors or by an ambiguity in class membership. Such errors, they remind, can influence the training phase of the classification, hence affecting the classification accuracy.

Foody *et al.* (2016) further studied the influence of mislabeling in classification accuracy. They found mislabeled training data typically degrade the classification accuracy, especially when the incorrect labels involve similar classes. The authors emphasize that such issue might be particularly relevant when training sampling units are drawn from border locations. In this perspective, a potential approach to mitigate such issue could be what Hermosilla *et al.* (2018) adopted, which is simply to avoid border pixels by excluding areas within a certain distance from the border.

2.4. Feature selection

Feature selection consists in determining which features of the sample dataset should be employed as predictor variables in the classification process. Most studies use satellites' native bands and common spectral indices, such as NDVI (Normalized Difference Vegetation Index), NDWI (Normalized Difference Water Index), NDBI (Normalized Difference Built-Up Index) and

Normalized Burn Ratio (NBR), as classification features. Inglada *et al.* (2017) use six Landsat 8 spectral bands and 3 radiometric indices: NDVI, NDWI and brightness. The authors used data from 22 dates, what amounts to 198 features. Pflugmacher *et al.* (2019) use the blue, green, red, NIR, SWIR1 and SWIR2 bands from Landsat 8. The authors also use the NDVI, NBR, Modified Soil-adjusted Vegetation Index (MSAVI2), Tasseled Cap Brightness (TCB), Tasseled Cap Greenness (TCG) and Tasseled Cap Wetness (TCW). Furthermore, they use spectral-temporal metrics, which are statistical metrics that characterize the behavior of a spectral band or index within a time period. The statistical metrics might be, for instance, minimum, maximum, mean, standard deviation, quantiles, percentiles and other common descriptive statistics. In one of their classification models, the authors calculated 9 metrics for 6 bands and 6 indices, what sums up to 108 features.

The works based on Sentinel-2 data employ a similar feature selection. Paris *et al.* (2019), Vuolo *et al.* (2018) and Close *et al.* (2018) used only 10 spectral bands. The first authors composed a time series of four images, hence totaling 40 features. Griffiths *et al.* (2019) utilized all 13 Sentinel-2 spectral bands, with the number of up to 324, depending on the analyzed time series. On the other hand, Weigland *et al.* (2020) used a combination of bands, indices (NDVI, NDWI and NDBI), metrics and auxiliary imperviousness information, adding up to 229 features.

Despite the broad range of features used in the studies cited above, there is a lack of agreement about which are the most advantageous features to be used in remote sensing classification. In this context, studies regarding the selection of an optimal combination of features, such as the one conducted by Feng *et al.* (2019) which is based on methods of feature importance, can provide a significant contribution in this topic.

2.5. Classification

Machine learning classification has been largely utilized in remote sensing studies. Its algorithms can model complex class signatures and accept a range of predictor variables as inputs. In addition, such algorithms are non-parametric, which means that they do not make assumptions about the data distribution. A variety of studies indicated that these methods generally produce higher accuracy compared to traditional parametric classifiers (Maxwell *et al.*, 2018).

Supervised learning can be considered the most relevant paradigm of machine learning applied to remote sensing. Boutaba *et al.* (2018) describe supervised learning as a method that requires a training dataset labeled with the corresponding ground truth. Then, the model learns to

identify patterns in the data so that it can distinguish classes of a set of new input data of an unknown class label.

In terms of machine learning workflow, specifically the supervised learning paradigm, Boutaba *et al.* (2018) describe the process according to the following steps: data collection, feature engineering, model learning and model validation. Data collection is simply the process of gathering training data, what requires representative data as suggested by the authors. Maxwell *et al.* (2018) highlight that training sample size and quality are crucial aspects to be considered in the process of planning a classification. Although the last authors acknowledge the absence of advice by the literature towards an appropriate minimum number of sample size, they suggest that a large and accurate training dataset is preferable since researches indicate that increasing training sample size results in an enhanced classification accuracy. Feature engineering can be understood as a preprocessing step that aims to remove noise and clean the data. Furthermore, it also encompasses the process of feature selection and extraction. Model learning is the actual process of machine learning, in which the algorithm recognizes patterns in the training data in order to create a signature for each class. Finally, the trained model is used to predict the class of new input data whose class is unknown. Then, the model performance is validated through an evaluation of a variety of metrics, such as the overall accuracy (Boutaba *et al.*, 2018).

There is a variety of machine learning classifiers employed in mapping LCLU. In the detailed review by Maxwell *et al.* (2018), some of the most common classifiers presented are Random Forest (RF), Support Vector Machines (SVM), Artificial Neural Networks (ANN) and k-nearest neighbors (k-NN). Among them, special attention has been dedicated towards RF in exercises involving mapping LCLU.

The Random Forest (RF) classifier is an ensemble classifier that applies the aggregation of multiple classification and regression trees. The growth of each tree in the ensemble can be determined by random vectors, which can be generated through strategies such as bagging, also known as bootstrapping, in which trees are grown based on a random selection of a subset of the training sample. Then, after having a large number of trees, the classification is performed by computing their votes for the most popular class (Breiman, 2001).

Belgiu and Drăguţ (2016) highlighted the recent interest of the remote sensing community in ensemble classifiers, especially RF. They mention a variety of studies that were successful in using RF to map land cover classes, urban buildings, insect defoliation levels, tree canopy cover

and others based on images from different satellites. Moreover, Maxwell *et al.* (2018) showed that ensemble methods have yielded better classification accuracy when compared to simple single classifiers such as Decision Tree.

The study by Lawrence and Moran (2015) presented a systematical comparison of machine learning classification algorithms using 30 different datasets. The results concluded that RF had the highest average classification accuracy. However, RF was the most accurate in only 18 of the 30 classifications. Belgiu and Drăguţ (2016) also compared RF to other machine learning classifiers, finding that RF achieved better classification results when using multi-dimensional or multi-source data. They too concluded that RF is faster than high performance classifiers such as Support Vector Machine (SVM) and AdaBoost, besides being simpler in terms of parameters to be configured. Furthermore, Maxwell *et al.* (2018) found RF to be less sensitive to the dimension of the training sample and training mislabeling.

Random Forest can be considered a classifier of easy optimization, since it only requires two user-defined parameters: the number of trees and the number of random variables used to determine the best split when growing the trees (Maxwell *et al.*, 2018). Regarding the number of trees, multiple studies concluded that such parameter does not have a major influence on the classification results, nevertheless, a value of 500 trees is recommended (Belgiu and Drăguţ, 2016; Maxwell *et al.*, 2018). As to the number of random variables available at each split, Maxwell *et al.* (2018) outlined that such parameter could have a moderate impact on the classification accuracy, thus suggesting it should be optimized.

With respect to RF limitations, Belgiu and Drăguţ (2016) pointed out that the classifier is sensitive to imbalanced training data, which results in favoring the most represented class. Therefore, the authors recommend training samples to be balanced and representative of each class.

2.6. Effects of training sample size on classification accuracy

As stated by Maxwell *et al.* (2018), the literature lacks a recommendation regarding the minimum training sample size needed for machine learning classification. The authors, however, support the broad understanding that increasing the number of training sampling units results in higher classification accuracy. Notwithstanding, they also acknowledge that the sensitivity to training data size varies depending on the classifier.

Huang *et al.* (2002) suggest that a satisfactory sample size might vary according to the classifier, number of classification variables and size and spatial variability of the study area. In terms of classifiers, Rodríguez-Galiano *et al.* (2012) verified that RF was significantly less sensitive to a reduction in training sample size when compared to single decision tree classification. Their findings are in accordance with results obtained by Maxwell *et al.* (2018), which demonstrated that RF has a superior performance compared to single decision trees when the number of training sampling units is smaller. Furthermore, the experiments conducted by Thanh Noi and Kappas (2018) ratify RF's low sensitivity to the number of training sampling units. The authors tested fourteen sample size scenarios and two sampling strategies: balanced and imbalanced. Then, they computed the overall accuracy with an unchanged validation dataset to compare the scenarios. The results showed that a reduction of 95% in the training sample size resulted in a decrease of less than 5% in the RF classification accuracy, regardless of class balance. When taking into account the computed confidence intervals, the decrease might be less than 2%.

Since the works by Rodríguez-Galiano *et al.* (2012) and Thanh Noi and Kappas (2018) use only 9 and 10 classification variables, respectively, further investigation needs to be conducted to evaluate the impact of training sample size on classifications with a large number of predictor variables. This could be particularly relevant in the case of RF, which is considered a classifier that has good performance in classifications with high dimensionality.

Besides samples size, class imbalance is another common issue discussed within the training sample topic. Although Maxwell *et al.* (2018) suggested using balanced training data, the results presented by Thanh Noi and Kappas (2018) indicated that neither balanced nor imbalanced strategy is predominant in terms of classification performance specifically using RF and Sentinel-2 images.

2.7. Classification uncertainty and land cover mapping

One of the main aspects of this thesis is using classification uncertainty to generate a new and improved training dataset. Therefore, this section is dedicated to presenting and discussing methodologies related to how uncertainty can be derived from classification and how it can contribute to create new training datasets.

There is a need to improve classification accuracy, which, in the case of supervised classification, implies optimizing the training dataset. In this context, active learning approaches emerge as an alternative to build a more informative and representative training set. As Tuia *et al.* (2011) explained, active learning's main idea consists in creating a small training dataset of optimized

samples, whose performance can be as good as a large training dataset composed by randomly chosen samples. The process is focused on the interaction between user and model, so that the model provides the user with pixels whose classification is the most uncertain. Then, the user is responsible for manually labelling such pixels, which will be incorporated into the prior training set in order to reinforce the model. This process occurs for a number of iterations until a satisfactory result is achieved. The authors explained that including difficult samples contributes to maximize the model optimization for generalization capabilities. Li *et al.* (2013) considered that active learning methods select new training samples that provide maximum information, resulting in higher classification accuracy when compared to a training set of the same size built collecting random selected sampling units.

Active learning selects unlabeled sampling units based on a query strategy, which can adopt measures such as uncertainty, representativeness, inconsistency, variance and error (Ahmad *et al.*, 2019). Usually, the query strategy relies on selecting sampling units with highest classification uncertainty. Therefore, it is important to discuss the various uncertainty criteria, i.e. forms of quantifying uncertainty. Tuia *et al.* (2011) conducted an extensive review of strategies employed to determine uncertainty, dividing them into 3 families: committee-based, large margin-based and posterior probabilities-based heuristics. The strategies have distinct advantages and degrees of suitability to the different classifiers. For instance, the authors claim that when using SVM the best choice is the large margin-based family. A potential method to be used in conjunction with a Random Forest classifier is the posterior probability-based Breaking Ties (BT). Posterior probability-based heuristics use the estimates of posterior probabilities of class membership to select the best candidates. The BT approach computes the difference between the two highest class membership probabilities and considers that when they have similar values, i.e. are close to a tie, the classifier confidence is the lowest (Tuia *et al.*, 2011). As Crawford *et al.* (2013) suggested, the BT strategy is suitable to be applied to models that output posterior probabilities.

Since Random Forest classification can output for each pixel a class probability distribution, various studies used such RF output to determine the uncertainty (Mack *et al.*, 2017; Liu *et al.*, 2018). Although not implementing an active learning methodology, Loosvelt *et al.* (2012), Loosvelt *et al.* (2014), Thonfeld *et al.* (2020) and Roodposhti *et al.* (2019) used RF class probabilities to compute uncertainty, which they used to simply map classification uncertainty. Mapping areas where the classification was the most uncertain allowed the authors to perform

a spatial analysis to identify patterns and elaborate on possible causes of high uncertainty in specific areas and which classes are the most affected. Although important information could be inferred, such use of uncertainty does not interfere in the final output of the classification. Additionally, most of these works use the Shannon entropy (H) as the uncertainty measure (Shannon, 1948). In contrast to the BT heuristics, the entropy H evaluates the disagreement existing in the whole class probabilities vector, with a high entropy indicating a higher disagreement, therefore a high uncertainty. According to Tuia *et al.* (2011), entropy is largely employed in committee-based active learning.

In terms of performance, Tuia *et al.* (2011) concluded that large margin-based heuristics performed better than the other families, although in some of their tests the BT approach yielded better results. The authors acknowledge that the large margin-based family had a better performance due to its enhanced compatibility with SVM, the classifier employed in their study. Nevertheless, all families of heuristics performed better than selecting new sampling units randomly. As to RF classifications, the BT approach seems the most convenient in terms of easiness of implementation (Mack *et al.*, 2017) and fair performance (Liu *et al.*, 2018).

While usual active learning approaches rely on using uncertainty alone, new researches propose combining uncertainty with other criteria. Crawford *et al.* (2013) outline the importance of a diversity criterion in active learning as a strategy to avoid selecting redundant sampling units. They argue that introducing a diversity measure contribute to select pixels that are most dissimilar among the highly ranked by the uncertainty query. Furthermore, the authors discuss the incorporation of spatial information in the active learning process, what can reduce spatial redundancy and contribute to further differentiate sampling units. Spatial information in active learning is also addressed by Lu *et al.* (2017) and Li *et al.* (2013), whose experiments indicate that utilizing spatial features can improve performance when compared to approaches based exclusively on spectral features. The work by Ahmad *et al.* (2019) proposes a method that, besides taking into account the spatial domain, utilizes fuzziness and diversity criteria. The results showed that their method outperformed all other sample selection methods employed. Despite the good results, the aforementioned approaches tend to have a more complex implementation.

Although active learning has presented encouraging results, it is heavily dependent on human interaction. Therefore, other works try to take advantage of using uncertainty to improve classification accuracy. The study by Gonçalves *et al.* (2009) presents a hybrid approach, in which

they propose to classify landscape units using multispectral images by combining a standard probabilistic classification and its associated classification uncertainty. The final class is determined based on a set of decision rules, which consider the type of surface element and an uncertainty value. The results indicated an increase in overall accuracy when compared to a classification without the incorporation of uncertainty. However, their work does not focus on collecting new samples to improve the training dataset.

On the other hand, Mack *et al.* (2017) proposes implementing an initial RF classification, from which they derive the classification uncertainty. Then, they determine an uncertainty threshold, conduct a segmentation and identify the largest connected patches of high uncertainty, from which they extract new training sampling units. Finally, they incorporate the additional sampling units to the initial training set and perform a final classification. Although their method yielded an overall accuracy of 87%, it is unclear whether adding new samples improved the classification, as the study lacks a comparison between initial and final classifications and therefore the cost benefit of adding new training based on the uncertainty analysis. Moreover, there was insufficient detail regarding the definition of the uncertainty threshold.

This thesis attempts to take a step further on the use of uncertainty to improve the training dataset of supervised algorithms, namely by proposing and evaluating a methodology to aggregate into the training data additional sampling units collected from areas where the initial classification was the most uncertain.

3. STUDY AREA AND DATA

This section presents a description of the study area and the various data sets used in this research.

3.1. Study area

The study area is located in the North of Portugal and corresponds to the landscape unit of Trás-os-Montes (Figure 1), comprising an area of 11,778 km² characterized by mountainous land occupied with rocks, forest and bushes, in addition to agriculture in the lower lands. Landscape units are areas of similar biogeographic aspects, in which DGT is experimenting specific methodologies. Working with landscape units rather than image tiles is preferable, since the first gathers areas of similar landscape characteristics, hence contributing to enhance spectral distinction between LCLU classes.

Figure 1: Location of study area.

3.2. Data

The data utilized in this study can be divided into two groups: ancillary and remotely sensed data. The ancillary data comprise various datasets employed in the process of automatic training sample collection, whereas the remotely sensed data correspond to the Sentinel-2 imagery used for classification.

3.2.1. Ancillary data

The ancillary data were used to delineate regions from where training data will be collected. The data can be divided into reference and filter data. Reference data aim to provide the base polygons to delineate training sample areas, while filter data aim to refine the reference polygons to reinforcing label consistency.

The first dataset used as reference data was the Portuguese Land Use and Land Cover Cartography from 2010, 2015 and 2018 (COS 2010, COS 2015 and COS 2018). COS is a thematic

cartography that aims to map land cover and land use in continental Portugal with a high level of detail, having a minimum map unit of 1 hectare and a minimum distance between lines of 20 meters. The production of such cartography is based on visual interpretation of high resolution orthorectified aerial imagery and the final product is available in vector format. The latest version of COS (COS 2018) contains 83 LULC classes (DGT, 2019).

The second reference dataset is the national parcel registry from the Portuguese Land Parcel Identification System (LPIS) of the *Instituto de Financiamento da Agricultura e Pesca* (IFAP). The data used in this study corresponds to the agricultural year of 2018 and hereinafter is referred as IFAP 2018. Such dataset consists of land parcels reported by farmers who applied to agricultural subsidies provided by the European Union. A fraction of about 5% of these parcels is subjected to control in the form of visual interpretation on orthophotos and field validation of the type of crop grown.

Other dataset used in the study was the map of burned areas of 2018 and 2017 from the *Instituto da Conservação da Natureza e Florestas* (ICNF). The map is provided in vector format, containing polygons of burned areas larger than 5 Ha during the years of 2016, 2017 and 2018 (ICNF, 2018).

In addition, Copernicus Land Monitoring Service's High Resolution Layers products from 2015 (HRL 2015) were used as filters to refine the reference dataset (IFAP and COS 2018). HRL provide information on particular land cover characteristics. Within the Forests domain, two products were incorporated: Tree Cover Density (TCD) and Dominant Leaf Type (DLT). TCD refers to the degree of tree cover density in a range from 0 to 100% whilst DLT determines whether there is a majority of broadleaf or coniferous leaves. Considering the Imperviousness domain, the Imperviousness Density (IMD) product was used. This product aims to provide information about the imperviousness degree (0 to 100%), which contributes to identify built-up areas. The 2015 HRLs products are generated mainly based on Sentinel 1 and 2 satellite imagery through a combination of automatic processing and interactive rule based classification and provided in raster format with 20m spatial resolution (Langanke, 2016; Langanke, 2017).

Moreover, a mask of NDVI changes detected from 2015 to 2018 using Landsat 8 images was used to remove clear cuts areas (Costa *et al.*, 2020). Clear cuts are zones where trees were uniformly cut down as part of forest management cycle of forest plantations. However, the land use of such zones is still mapped as forest in reference datasets. Then, removing the clear cuts helps preventing training forests classes using pixels that do not correspond spectrally to forests.

Lastly, the OpenStreetMap (OSM) primary roads and motorways of Continental Portugal were used as reference data to collect training data for the road network class.

3.2.2. Remotely sensed data

The remote sensing data utilized are an orthophoto map of 2018 with 25 cm spatial resolution, used to assist the collection of manual training and validation, and a composite of Sentinel-2 images of the study area acquired from October 2017 to September 2018, corresponding to the agricultural year of 2018. The year of 2018 was selected not only due to the availability of the orthophoto to assist in the process of validation, but also due to the existence of official LCLU cartography for such year (COS 2018), allowing a comparison between COS and the map generated by our proposed methodology.

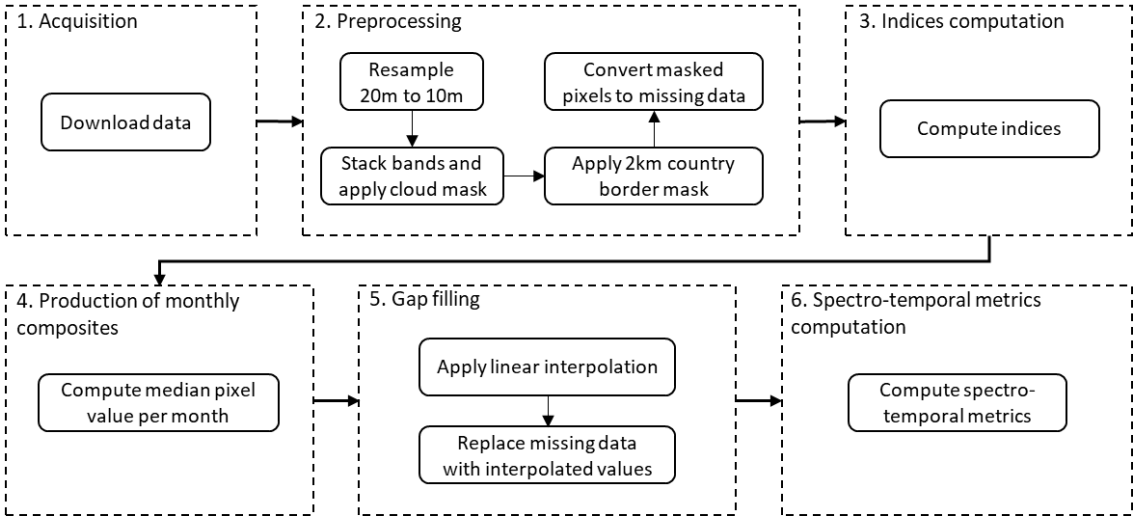

Figure 2: Workflow involved in the processing of generating Sentinel-2 dataset.

The images utilized in this study were produced by DGT according to the subsequent procedures, which follow DGT's technical specifications for the generation of intra-annual Sentinel-2 surface reflectance composites (DGT, 2020). The workflow comprises 6 main activities (Figure 2): acquisition, preprocessing, indices computation, production of monthly composites, gap filling and spectro-temporal metrics computation.

The acquisition consists in downloading the imagery from Theia for the agricultural year of 2018, with a filter of less than 50% of cloud cover. The images are provided as Level-2A processing products, with ortho-rectification, atmospheric correction to the bottom of atmosphere (BOA), water, snow, cloud and cloud shadow masks and slope effect correction. The study area is covered by 6 tiles of Sentinel-2 images: 29TNG, 29TPG, 29TQG, 29TNF, 29TPF and 29TQF (Figure

3). In total, 457 images with less than 50% cloud cover were acquired. The per-tile distribution is exhibited in Table 1.

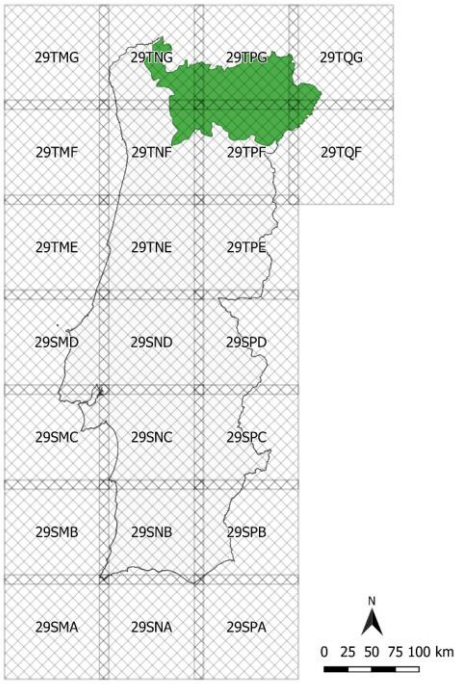

Figure 3: Sentinel-2 tiling for Continental Portugal.

Tile	Sentinel-2 images acquired
29TNG	72
29TPG	71
29TQG	81
29TNF	78
29TPF	78
29TQF	77
Total	**457**

Table 1: Number of images per tile.

In the preprocessing stage, Sentinel-2 spectral bands B2, B3, B4, B5, B6, B7, B8, B8A, B11 and B12 are selected and the bands with spatial resolution different from 10m are disaggregated to 10m. Bands B1 and B10 are used only for atmospheric correction. Next, the bands corresponding to a single acquisition are stacked and masked according to the cloud mask. Pixels contaminated with cloud or cloud shadow are reclassified to “missing data”. Additionally, pixels beyond 2km of the country border and coastline are also converted as missing data. Then, rasters with pixel size of 10m and 10 bands are generated.

The preprocessed rasters are then used to compute spectral indices, thus generating new rasters, each corresponding to a specific index. Such indices were selected based on a

bibliographic review of various papers about production and implementation of national land cover maps. The computed indices and their correspondent equations are described in Table 2.

Index	Equation	Sentinel-2 bands	Reference
NDVI	(NIR-R) / (NIR+R)	(B08 - B04) / (B08 + B04)	Rouse et al. (1974)
NBR	(NIR-MIR2) / (NIR+MIR2)	(B8A - B12) / (B8A + B12)	Hislop et al. (2018)
NDWI	(G-NIR) / (G+NIR)	(B03 - B08) / (B03 + B08)	McFeeters (1996)
NDBI	(MIR1-NIR) / (MIR1+NIR)	(B11 - B8A) / (B11 + B8A)	Zha et al. (2003)
NDMIR (or NBR2)	(MIR1-MIR2) / (MIR1+MIR2)	(B11 - B12) / (B11 + B12)	Roteta et al. (2019)

Table 2: Spectral indices computed for the Sentinel-2 dataset.

Having the bands and indices set, the next step is the production of monthly composites. This process consists in computing at the pixel level the median pixel value for each month. As such approach takes into account multiple instead of single acquisitions, it increases the probability of acquiring pixels not contaminated by clouds, hence reducing the occurrence of missing values. However, there still could be pixels covered by clouds during an entire month, what would lead to them being flagged as missing values, forming gaps. Therefore, it is necessary to perform an additional step to fill these gaps.

Metric	Description
q10	Quantile 10^{th}
q25	Quantile 25^{th}
q50	Quantile 50^{th}
q75	Quantile 75^{th}
q90	Quantile 90^{th}
q75-q25	Difference between 75^{th} and 25^{th} quantiles
q90-q10	Difference between 90^{th} and 10^{th} quantiles

Table 3: Spectro-temporal metrics computed for the Sentinel-2 dataset.

The process of gap filling consists in applying a simple linear interpolation based on time. For instance, if a given pixel has a missing value in a specific month, values from the previous and following months are used to interpolate the value for the missing month.

Name	Quantity
Spectral Bands	120
Spectral Indices	60
Spectral-temporal metrics	105
Total	**285**

Table 4: Band composition of the Sentinel-2 final dataset.

Lastly, the results from the previous step are used to compute spectro-temporal metrics. Such metrics are composed of quantiles and differences between quantiles, computed for each of the

10 bands and 5 indices considering the whole year of analysis. Table 3 presents the list of metrics employed.

The final composite consists of 285 bands: 10 bands and 5 indices for each month of the year and 7 metrics for each of the 10 bands and 5 indices (Table 4). A summary of the data described in the previous sections, with their brief description, origin, year and function is presented in Table 5.

Dataset	Description	Source	Year	Function
COS	Land use and land cover official cartography	DGT	2010, 2015, 2018	Reference data
IFAP	National parcel registry (crop types)	IFAP	2018	Reference data
OSM Roads	OpenStreetMap roads network	OpenStreetMap	2020	Reference data
Burned Areas	Wildfire burned areas	ICNF	2017, 2018	Filter data
OSM Roads	OpenStreetMap roads network	OpenStreetMap	2020	Reference data
HRL-TCD	Tree cover density	Copernicus Land Monitoring Service	2015	Filter data
HRL-DLT	Dominant leaf type	Copernicus Land Monitoring Service	2015	Filter data
HRL-IMD	Imperviousness degree	Copernicus Land Monitoring Service	2015	Filter data
NDVI Clear Cuts Mask	Forest cuts alerts	Costa *et al.* (2020)	2015-2018	Filter data
Orthophoto	Orthophotomap (25 cm)	DGT	2018	Remote sensing data
Sentinel-2 Imagery	Satellite imagery	Theia Land Data Centre	2017-2018	Remote sensing data

Table 5: Summary of the data used in the research.

4. METHODS

The proposed methodology is based on a combination of manual collection and automatic extraction of training sampling units from preprocessed reference datasets. The collected sampling units are used to retrieve the spectral information of a multi-temporal Sentinel-2 composite, which will form the feature space of a supervised learning classifier using Random Forest. Prior to the learning process, a stratification of the study area is conducted. The classification accuracy is assessed computing the overall accuracy, which is used to compare results. This base workflow is implemented to evaluate the impact of training sample size on classification accuracy and to assess whether training samples can be improved using classification uncertainty. Therefore, the methods can be divided into 5 sections: study area stratification, COSsim technical specifications, supervised learning and classification, impact of sample size and classification uncertainty and improved training.

4.1. Study area stratification

The study area was divided into 5 strata (Figure 4) with distinct land cover characteristics, based on COS 2018 and ICNF cartography: Cork and holm oak, Burned areas in 2017, Burned areas in 2016, Forest cuts from 2015 to 2018 and a complementary stratum (Table 6). Stratum 1 was originated from areas of cork and holm oak of COS 2018. Strata 2 and 3 were defined according to the ICNF burned areas of 2017 and 2016, respectively. Stratum 4 resulted from the intersection of areas of forest and shrubland of COS 2018 with the NDVI clear cuts mask, hence representing zones where vegetation cuts occurred. Lastly, the complementary stratum represents the remaining areas.

The stratification aims to provide a combination of land cover classes suitable for each stratum, thus considering the inherent specific spectral characteristics. An individual supervised learning was performed for each stratum. The stratification is used to determine the combination of classes, as the samples utilized in each learning belong to the whole study area, regardless of stratum. Additionally, the generation of the classification map is conducted by merging individual classifications of each stratum.

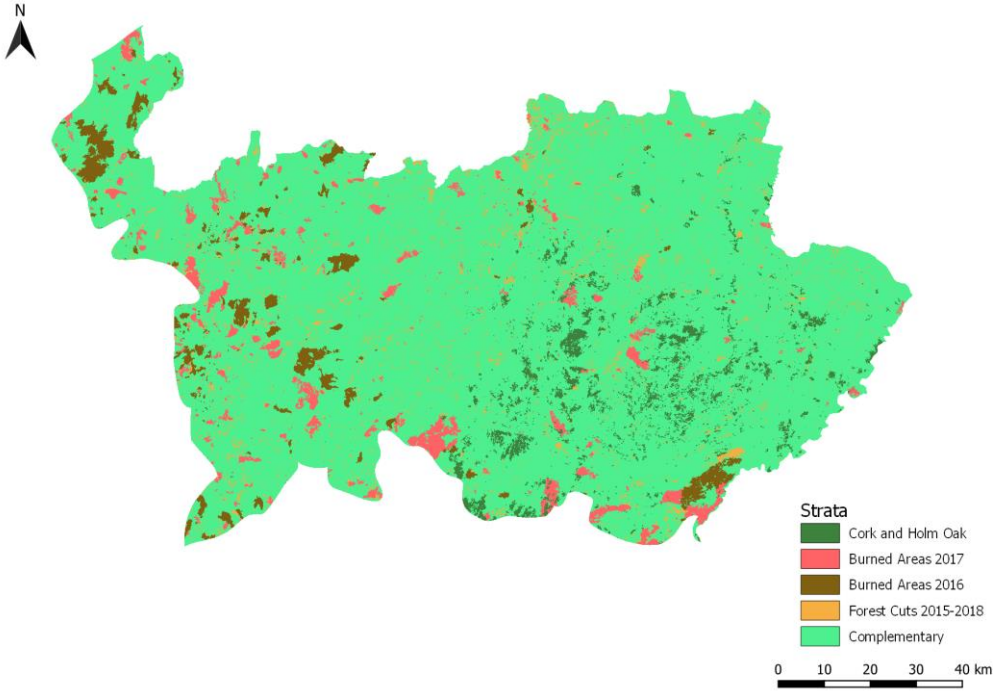

Figure 4: Map of the stratification of the region of study.

Number	Stratum	Area (ha)	%
1	Cork and Holm Oak	39113	3.3
2	Burned Areas 2017	42981	3.6
3	Burned Areas 2016	35418	3.0
4	Forest Cuts 2015-2018	45513	3.9
5	Complementary	1014777	86.2
	Total	1177802	100.0

Table 6: Stratification of the study area.

4.2. COSsim technical specifications

The following processes aim to produce a preprocessed reference dataset to be used to extract training samples according to DGT's COSsim technical specifications. The main purpose of the preprocessing is to create a reference dataset in accordance to the defined class nomenclature by applying rules to generate polygon features corresponding to each class. Such dataset is used as a reference from which training samples are automatically extracted. COSsim production adopts a strategy in which the training phase has a particular and more detailed class nomenclature, whilst the final product, COSsim Level 3, has a broader class nomenclature resulting from the aggregation of the training classes (Table 7). Using more training classes means separating COSsim Level 3 classes with known intrinsic variability into subclasses whose spectral characteristics are simpler to distinguish. This can be seen, for instance, in class natural grasslands (COSsim Level 3), which is divided into 3 classes with distinct spectral responses:

agricultural natural grasslands, mountain natural grasslands and natural grasslands BA2017 (specific of the Burned areas 2017 stratum).

Class COSsim Level 3	Training Class	Stratum				
		1	2	3	4	5
Built up	Built up	X	X	X	X	X
	Industrial	X	X	X	X	X
	Road Network	X	X	X	X	X
Agriculture	Oat	X	X	X	X	X
	Wheat	X	X	X	X	X
	Barley	X	X	X	X	X
	Ryegrass	X	X	X	X	X
	Triticale	X	X	X	X	X
	Rye	X	X	X	X	X
	Corn	X	X	X	X	X
	Sunflower	X	X	X	X	X
	Managed Grasslands	X	X	X	X	X
Natural Grasslands	Agricultural Natural Grassland	X	X	X	X	X
	Mountain Natural Grassland	X	X	X	X	X
	Natural Grasslands BA2017		X			
Cork and Holm Oak	Cork Oak				X	
	Holm Oak				X	
Eucalyptus	Eucalyptus Adult	X	X	X	X	X
	Eucalyptus BA2017		X			
	Eucalyptus 1 year cuts			X		
	Eucalyptus Young Cuts			X		
Other Broadleaf	Other Broadleaf	X	X	X	X	X
Maritime Pine	Maritime Pine	X	X	X	X	X
Stone Pine	Stone Pine	X	X	X	X	X
Other Coniferous	Other Coniferous	X	X	X	X	X
Shrubland	Dense Shrubland	X	X	X	X	X
	Shrubland BA2017		X			
Non-vegetated surfaces	Baresoil	X	X	X	X	X
	Bare Rock	X	X	X	X	X
Water	Water	X	X	X	X	X

Table 7: COSsim Level 3 class nomenclature and training class nomenclature according to each stratum.

This work adopts a hybrid process for collecting training samples: a combination of automatic and manual collection. The manual collection is the traditional approach used in supervised classification, in which training data are acquired through digitization of polygons by photointerpretation, which in this study was assisted by the 2018 ortophoto in conjunction with one Sentinel-2 image for each season. Manual training is needed because previous experiments indicated that some classes in the reference dataset data are considerably heterogeneous, having several contributions from very different land cover types. Table 8 reveals which classes are based on manual and automatic training.

The reference datasets that serve as a base for defining COSsim automatic training areas are COS and IFAP 2018. Whilst COSsim is a land cover cartography with 100 m² Minimum Mapping

Unit (MMU), COS maps land use and land cover with MMU of 1 ha. Therefore, to use COS as reference data for COSsim training it is important to apply filters to exclude mislabeled pixels.

Training Class	Method	Dataset of Origin	Filters				
			HRL IMD	HRL DLT	HRL TCD	NDVI	Clear Cuts
Built up	Automatic	COS2018	≥ 80%				
Industrial	Manual	-					
Road network	Automatic	OSM				max(NDVI) ≤ 0.3	
Wheat	Automatic	IFAP18, COS2018					
Rye	Automatic	IFAP18, COS2018					
Oat	Automatic	IFAP18, COS2018					
Ryegrass	Automatic	IFAP18, COS2018					
Triticale	Automatic	IFAP18, COS2018					
Corn	Automatic	IFAP18, COS2018					
Sunflower	Automatic	IFAP18, COS2018					
Barley	Automatic	IFAP18, COS2018					
Managed Grasslands	Automatic	IFAP18, COS2018					
Agricultural Natural Grassland	Manual	-					
Mountain Natural Grassland	Manual	-					
Natural Grasslands BA17*	Manual	-					
Cork Oak	Manual	-					
Oak Canopy	Manual	-					
Holm Oak	Manual	-					
Eucalyptus Young Cuts	Manual	-					
Eucalyptus Adult	Automatic	COS2015, COS2018		Broadleaf	≥ 90%	min(NDVI) ≥ 0.3	Outside
Eucalyptus BA17*	Manual	-					
Eucalyptus 1 Year Cuts**	Manual	-					
Other Broadleaf	Automatic	COS2018		Broadleaf	≥ 90%	min(NDVI) ≥ 0.3	Outside
Maritime Pine	Automatic	COS2018		Coniferous	≥ 90%	min(NDVI) ≥ 0.3	Outside
Stone Pine	Automatic	COS2018		Coniferous	≥ 70%	min(NDVI) ≥ 0.3	Outside
Other Coniferous	Automatic	COS2018		Coniferous	≥ 90%	min(NDVI) ≥ 0.3	Outside
Dense Shrubland	Manual	-					
Shrubland BA17*	Manual	-					
Baresoil	Automatic	COS2018				min(NDVI) > 0 & max(NDVI) < 0.3	
Bare Rock	Manual	-					
Water	Automatic	COS2018				max(NDVI) ≤ 0	

Table 8: Class nomenclature, their methods of collecting training samples, origin and filters applied. **Manually collected within burned areas in 2017; **manually collected within forest cuts 2015-2018.*

Regarding the automatic process, the first step consists in eliminating burned areas from COS 2018, based on a geometric difference between COS 2018 and ICNF burned areas of 2018, which prevents training vegetation classes with burned pixels. Next, the IFAP 2018 dataset is considered, selecting only the 10 most abundant crops in the country (wheat, barley, oat, ryegrass, triticale, rye, tomato, corn, sunflower and rice) and managed grasslands. However,

there are no rice and tomato polygons within the study region, therefore the subsequent processing is conducted with the 8 remaining classes. An additional filter is applied to the managed grasslands to exclude parcels that might eventually contain significant abundance of trees. Parcels with area of less than 1000m² are excluded from IFAP.

For each of the eight most abundant crops in the country, a geometric intersection with COS 2018 annual crops was carried out. Then, a negative buffer of 40m was applied, eroding polygons to eliminate boundaries and transition zones where there could be mixed pixels. Lastly, polygons generated by the intersection with area of less than 100m² were excluded. The processing of the class managed grasslands is similar, however, the intersection is performed with COS 2018 class grasslands.

The remaining of the automatic classes are generated directly from the correspondent class in the COS 2018 preprocessed dataset, also having a negative buffer of 40m applied to the polygons. The only exceptions are road network and eucalyptus adult. Road network reference data is acquired by extracting points along the main roads (primary and motorway) in the OpenStreetMap dataset, while the eucalyptus adult results from the intersection between eucalyptus in COS 2018 and 2015, what suggests that these eucalyptus might be older.

Besides the processes described above, additional filters are applied to specific classes. The filters intend to refine reference polygons by intersecting it with other datasets, such as HRL, in order to attempt delineating areas with more accurate spectral responses of a given class, hence preventing acquiring potential mislabeled pixels. Three HRL products were used as filters: Imperviousness Degree (IMD), Tree Cover Density (TCD) and Dominant Leaf Type (DLT). Whenever the TCD and DLT filter are used, a raster shrink function with size of 1 pixel (20m) is applied to further refine the filtering. Since HRL and COSsim have different reference periods, areas mapped by HRL as forest in 2015 might have been cut later. Therefore, additional filters are needed to prevent delineating areas where forest was cut, hence avoiding pixels whose spectral characteristics do not match forest's. For this purpose, an annual NDVI raster was employed as a filter, with different thresholds rules depending on the class, along with the NDVI clear cuts mask (Costa *et al.*, 2020). After generating the polygons a cleansing process was conducted, in which features with less than 1000m² of area were deleted. Table 8 presents a summary of the nomenclature adopted in COSsim, the source datasets and filter rules of each class.

Special attention was dedicated to the classification of cork and holm oaks. Within the cork and holm oak stratum, training samples were collected by visual interpretation of the orthophotomap. A total of 12 classes were used to represent a gradual level of abundance of the following elements: oak canopy, natural grasslands, shrubland and baresoil (Table 9). The abundance of each cover type in a pixel was defined according to the occurrence of oak. The vicinity or understory can have elements such as shrubs, grass and soil, which can generate distinct spectral responses. Therefore, the classes also considered a combination of different understory and neighboring cover types based on oak canopy percentage estimation.

Class	Description	% oak
1	Oak and shrubland	100
2	Oak and natural grasslands	100
3	Oak and baresoil	100
4	Oak and shrubland	80
5	Oak and natural grasslands	80
6	Oak and baresoil	80
7	Shrubland and oak	20
8	Natural grasslands and oak	20
9	Baresoil and oak	20
10	Shrubland	0
11	Baresoil	0
12	Natural grasslands	0

Table 9: Cork and holm oak canopy classification training classes.

4.2.1. Training database

The preprocessed polygons were used to automatically extract random training samples from each class. The sampling process was implemented within a GIS environment. It consisted in generating points inside the polygons, corresponding to the image composite pixel centroids, then performing a random selection and retrieving the composite values. The composite bands correspond to the predictor variables in the classification. Although previous experiments performed with variable selection to reduce the dimension of the feature space indicated an increase in classification efficiency, this study was conducted with all the predictor variables available, following what was being experimented by DGT. With respect to the sampling process, a total of 6000 sampling units per class were collected. In some classes the number of sampling units available was less than 6000, therefore their sample size was the largest possible. Besides the automatic training, the manually collected sampling units were also aggregated, thus forming the training database. Table 10 presents the number of polygons, descriptive statistics of their areas and number of sampling units per training class.

Although some works reviewed in sections 2.5 and 2.6 address the topic of imbalanced training samples, our approach did not consider balancing the samples, since our investigation focused on other topics.

Regarding the cork oak canopy classification, Table 11 presents the number of sampling units collected by photointerpretation. For this particular classification, only the 10 bands and 5 spectral indices from August 2018 were used as classification variables.

Training Class	**Nº of polygons**	**Area (ha)**						**Sample size**
		Min	Max	Mean	Median	Std	Total	
Built up	223	0.01	2.32	0.18	0.03	0.34	39.44	3943
Industrial	322	0.01	1.87	0.09	0.05	0.17	30.24	2793
Road Network	-	-	-	-	-	-	-	6000
Oat	1146	0.01	8.17	0.47	0.25	0.65	535.55	6000
Wheat	303	0.01	6.88	0.30	0.11	0.63	90.26	6000
Barley	22	0.01	4.59	0.41	0.10	0.94	9.06	910
Ryegrass	34	0.01	0.94	0.21	0.12	0.23	7.11	704
Triticale	66	0.01	3.70	0.27	0.11	0.52	17.91	1777
Rye	751	0.01	3.79	0.30	0.13	0.46	222.92	6000
Corn	460	0.01	3.58	0.26	0.10	0.42	118.30	6000
Sunflower	1	0.14	0.14	0.14	0.14	0.00	0.14	17
Managed Grasslands	840	0.01	5.90	0.39	0.18	0.60	327.76	6000
Agricultural Natural Grassland	100	0.07	5.53	0.74	0.52	0.83	74.33	6000
Mountain Natural Grassland	47	0.03	8.14	1.23	0.77	1.40	57.88	5801
Natural Grasslands AA2017	57	0.06	7.57	1.33	0.72	1.58	75.80	6000
Cork Oak	140	0.03	29.65	2.73	1.22	4.30	382.66	6000
Holm Oak	105	0.05	12.03	2.51	1.67	2.58	263.84	6000
Eucalyptus Adult	24	0.07	52.66	3.10	0.52	10.40	74.44	428
Eucalyptus AA2017	22	0.31	20.84	4.08	2.80	4.49	89.73	6000
Eucalyptus 1 year cuts	30	0.64	11.39	3.09	2.46	2.40	92.72	6000
Eucalyptus Young Cuts	16	0.02	0.92	0.27	0.08	0.30	4.28	6000
Other Broadleaf	211	0.01	4.90	0.20	0.03	0.57	42.48	4245
Maritime Pine	872	0.01	10.19	0.48	0.18	0.87	420.39	6000
Other Coniferous	140	0.02	3.76	0.34	0.16	0.47	48.03	4796
Dense Shrubland	255	0.07	88.55	3.09	1.46	6.43	787.23	6000
Shrubland AA2017	80	0.01	10.17	0.80	0.27	1.56	63.88	6000
Baresoil	453	0.01	15.32	0.58	0.12	1.38	264.08	6000
Bare Rock	953	0.01	1.72	0.06	0.03	0.11	58.64	5803
Water	492	0.01	593.20	4.53	0.08	38.76	2227.24	6000

Table 10: Number of training polygons resulted from the preprocessing, descriptive statistics of their areas and the training sampling units collected. Training for road network is derived from linear elements.

Class	1	2	3	4	5	6	7	8	9	10	11	12
Sample Size	100	150	100	100	180	100	100	100	100	230	100	300

Table 11: Training sample size for the cork and holm oak canopy classification.

4.2.2. Image classification

Image classification was performed using the Random Forest classifier. The classification was implemented in Python, using the Scikit-learn library (Pedregosa *et al.*, 2011). The parameters adopted were 500 trees, $\sqrt{n}$ as the number of features available at each node and entropy as

the criterion used to determine the quality of a split. The remaining parameters were left as default.

A series of classifications were performed. First, five models were trained according to the classes of each stratum (see Table 7 in section 4.2) to classify the correspondent stratum in the Sentinel-2 composite. In this case, the number of features is equal to 285 (n = 285). Next, the pixels were reclassified from training class to the COSsim Level 3 class nomenclature.

A further classification was conducted to map cork oak canopy. The same parameters were applied, though the number of features was different (n = 12). The classification was performed only in the Cork and Holm Oak stratum. Then, the classes with occurrence of canopy (class 1-9 in Table 11) were reclassified to oak canopy. The final map consisted in overlaying the original map of such stratum with the oak canopy pixels, what resulted in converting these pixels in the original map to oak canopy. Afterwards, the pixels were also converted to COSsim Level 3 class nomenclature. The final LCLU map consisted in merging all the products described above.

4.2.3. Accuracy assessment

An independent validation dataset composed by 600 sampling units drawn from stratified random sampling and manually labeled by visual interpretation of the orthophoto map of 2018 with 25 cm spatial resolution was used to assess the classification accuracy of the final product (COSsim Level 3). The labels were assigned considering a 3x3 pixel window, with the sampling unit being located in the central pixel. This approach aims to address possible spatial displacement of the Sentinel-2 composite. For each validation sampling unit one or more labels were allocated, when adequate (e.g. transition between two land cover patches). A sampling unit is considered correctly classified if the class predicted by the RF classifier matches one of the labels assigned to it. Figure 5 illustrates the validation window of the sampling unit represented by the yellow point. In this case, 2 labels were assigned to the sampling unit: built up and agriculture. The metrics utilized in the accuracy assessment are the overall accuracy, precision, recall and F1-score. The last three metrics are computed according to the following equations, where tp: true positives, tn: true negatives, fn: false negatives.

$$Precision = \frac{tp}{tp + fp}$$

$$Recall = \frac{tp}{tp + fn}$$

$$F1\text{-}score = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

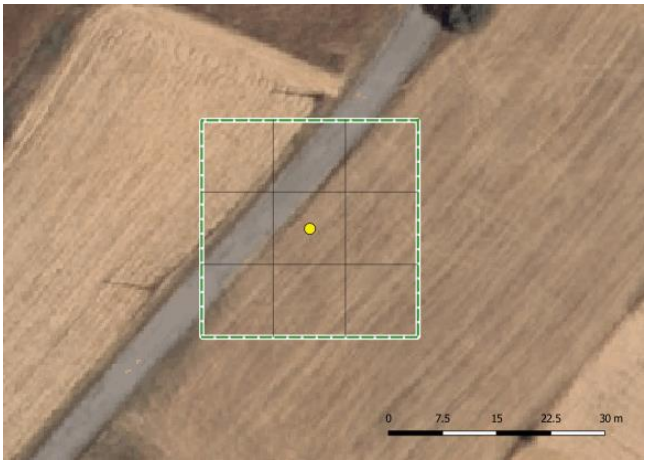

Figure 5: 3x3 pixel neighborhood of the validation sampling unit.

Confidence intervals for the overall accuracy were computed based on an error tolerance, as presented in Baraldi *et al.* (2006). In addition, confusion matrices are used to assess the confusion between predicted and reference classes. As the reference can have multiple labels, the disagreement between prediction and reference is computed considering the predominant label in the reference.

4.3. Assessment of the impact of stratification and manual training

In order to evaluate the impact of adopting stratification and manual training, a benchmark classification with training without stratification and using only automatically collected samples is conducted. In this context, some of the manual classes need to be eliminated from the training process, while others are replaced with their equivalent automatic class. Nevertheless, nomenclature for COSsim Level 3 remained the same, thus a comparison between the benchmark and the classification performed with stratification and combination of manual and automatic training can determine whether the latter strategies contribute to improve classification accuracy. In order to assure compatibility, a training sample size of up to 6000 sampling units per class was adopted for both classifications.

4.4. Assessment of the impact of sample size

The impact of sample size in the classification accuracy was evaluated testing different scenarios: 50, 500, 1000, 2000, 3000, 4000, 5000 and 6000 training sampling units per class collected randomly. For the classes in which the number of sampling units was less than the intended amount, the number of collected units was the largest available. These experiments were conducted in the complementary stratum only, as it represents the vast majority of the

study region (86.2%). In this case, the validation is performed excluding sampling units from other strata, resulting in a 535 sampling units validation dataset.

4.5. Classification uncertainty and improved training

The purpose of this process is to map areas where the classification was the most uncertain, then collect new training samples from within these areas. The process requires a reference classification, which in this case is the aggregation of the Sentinel-2 composite classifications of the 5 strata, with sample size of 6000 or the largest number of samples per class (see column sample size in Table 10). In this process, the output of the reference classification keeps the training nomenclature instead of converting to the map nomenclature.

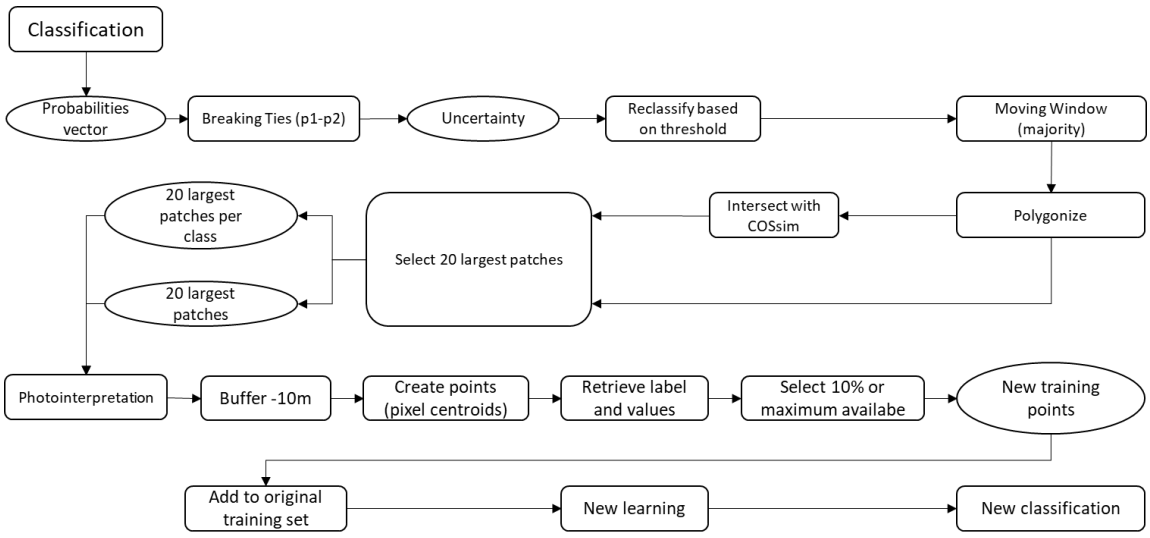

Figure 6: Workflow of the process of acquisition of new training samples from areas of high classification uncertainty and subsequent incorporation in initial training sample to perform a new classification.

The overview of the workflow is exhibited in Figure 6. The process consists in computing the probabilities vector of the reference classification. The Scikit-learn Random Forest implementation allows predicting the class probabilities of an input sample, which are defined in the documentation as the mean predicted class probabilities of the trees in the forest. From the probabilities vector, classification uncertainty can be derived using an uncertainty criterion. Breaking Ties (BT), a posterior-probability based approach, was used to compute uncertainty. The underlying assumption is that elements whose highest and second-highest class probabilities have similar values are considered as having high classification uncertainty. Then, BT calculates the uncertainty as the difference between the highest and second-highest class probabilities, with values ranging from 0 (high uncertainty) to 1 (low uncertainty). The result of

such computation is a raster where each pixel has its uncertainty value associated. The uncertainty U of a sampling unit i, calculated by the BT approach is:

$$U_i = max_{\omega \in N}\{p(\, y_i = \omega | x_i)\} - max_{\omega \in N \setminus \omega^+}\{p(\, y_i = \omega | x_i)\}$$

where $p(\, y_i = \omega | x_i)$ represents the probability of a sampling unit $x_i$belonging to the class ω, and ω^+ is the class of highest probability.

The next step consisted in reclassifying the raster based on a threshold, generating a binary map where pixels with values greater than the threshold are converted to no data. The threshold value was defined based on an analysis of the distribution of uncertainty values (U) computed for the whole map (Figure 7), which revealed that a sufficient amount of pixels have U ≤ 0.1, which was considered adequate to produce an adequate number of uncertainty patches. Then, 0.1 was adopted as the threshold. Next, a 5x5 majority moving window was applied in order to smooth the results, reducing a potential salt and pepper effect. The resulting raster is converted to vector and the 20 largest patches of uncertainty are collected. In addition, the resulting vectors are also intersected with the classification output, in order to select the 20 largest patches of uncertainty per class. These procedures are illustrated in Figure 8.

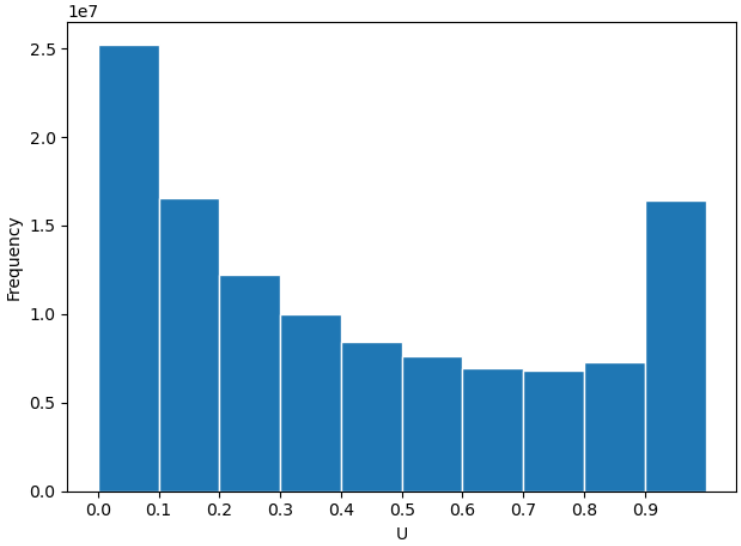

Figure 7: Histogram of classification uncertainty values.

The 20 largest patches regardless of class in conjunction with the 20 largest patches per class are photointerpreted in order to collect sampling units in these areas. Next, a negative buffer of 10m is applied to prevent capturing pixels in transition areas. Since it is not possible to identify the crop type by visual interpretation of the orthophoto, a patch or part of a patch located on

top of agricultural areas in the orthophoto is ignored. Figure 9 illustrates photointerpreted polygons on top of an uncertainty patch.

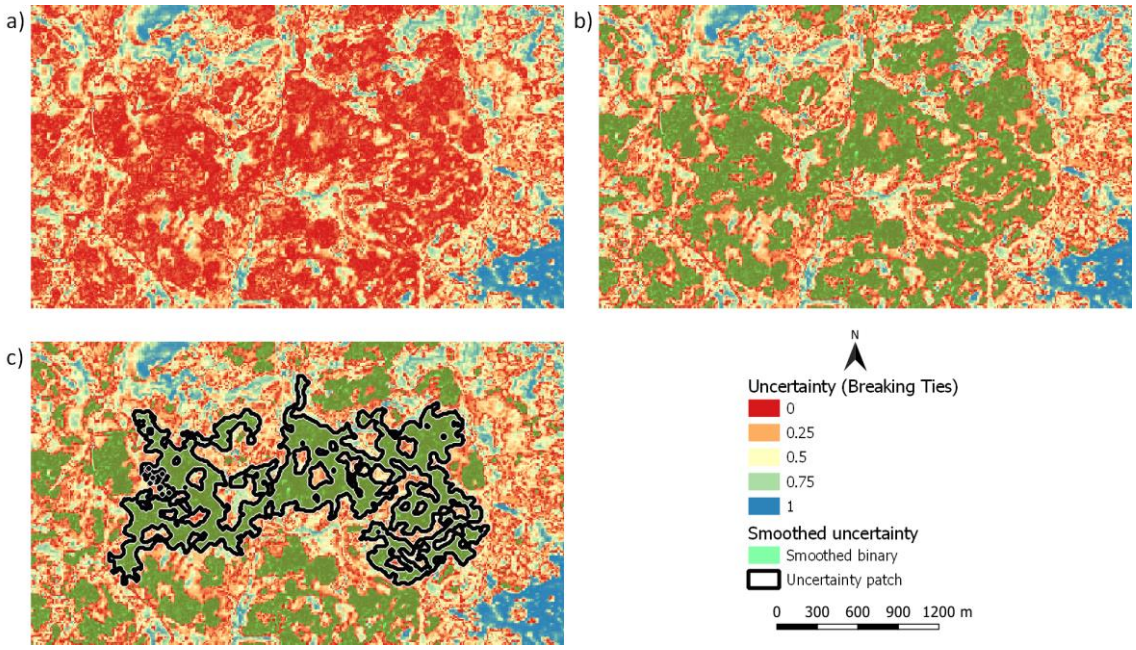

Figure 8: Delineation of an uncertainty patch: a) raw uncertainty distribution; b) result of the application of 0.1 threshold followed by smoothing (moving window) in green; c) delineation of a contiguous uncertainty patch.

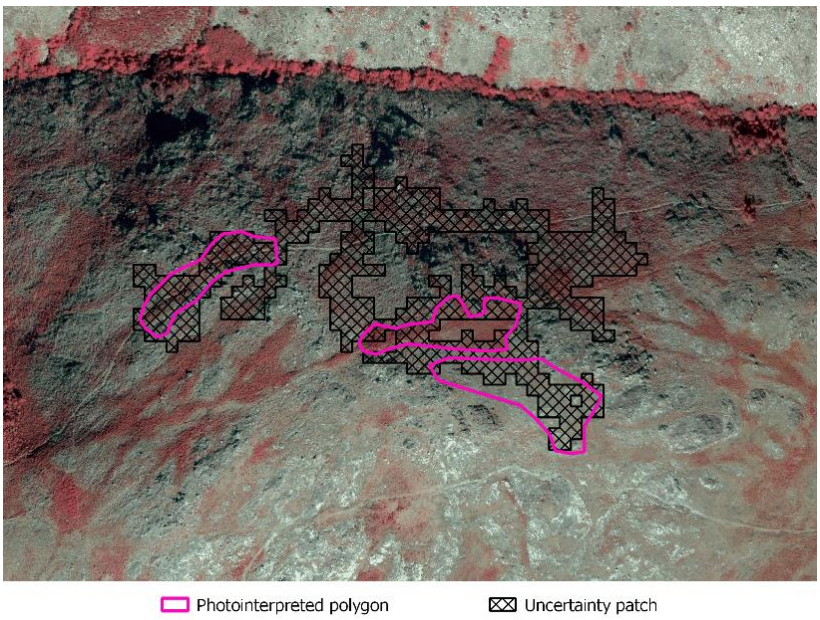

Figure 9: Photointerpretation of an uncertainty patch.

Finally, points corresponding to pixel centroids are generated within the polygons and their corresponding labels and composite values are retrieved. Regarding the addition of the new sampling units into the original sample, some aspects had to be addressed. Since the size of the

patches might vary, classes may have distinct number of sampling units. The new training should consider adding sufficient sampling units while minding class balance. With this in mind, it is preferable to incorporate additional sampling units into an initial sample of compatible size, so that the new units could have a growing influence in the representativeness of the aggregated sample. Thus, we adopted a strategy which consisted in adding up to 500 sampling units per class to an initial sample of 500 units per class. The initial sample was taken from the experiments with 500 sampling units of section 4.5. Having the new training set, a new classification is conducted only for the complementary stratum.

5. RESULTS AND DISCUSSION

This chapter presents the results of the innovative classification performed with stratification and introduction of manual sampling, a comparison with the benchmark classification, conducted without stratification and manual training, the assessment of the influence of sample size in classification accuracy and the evaluation of the improvement of training samples using classification uncertainty.

5.1. Stratification and introduction of manual training samples

The classification of the Sentinel-2 composite conducted with stratification and combination of manual and automatic training sample collection generated the COSsim Level 3 product seen in Figure 10. The map was produced by a classification performed with 6000 or the largest amount available of training sampling units per class. The overall accuracy of the classification, computed with the independent validation dataset, was 66.7%. The distribution of the validation sampling units is also shown in Figure 10.

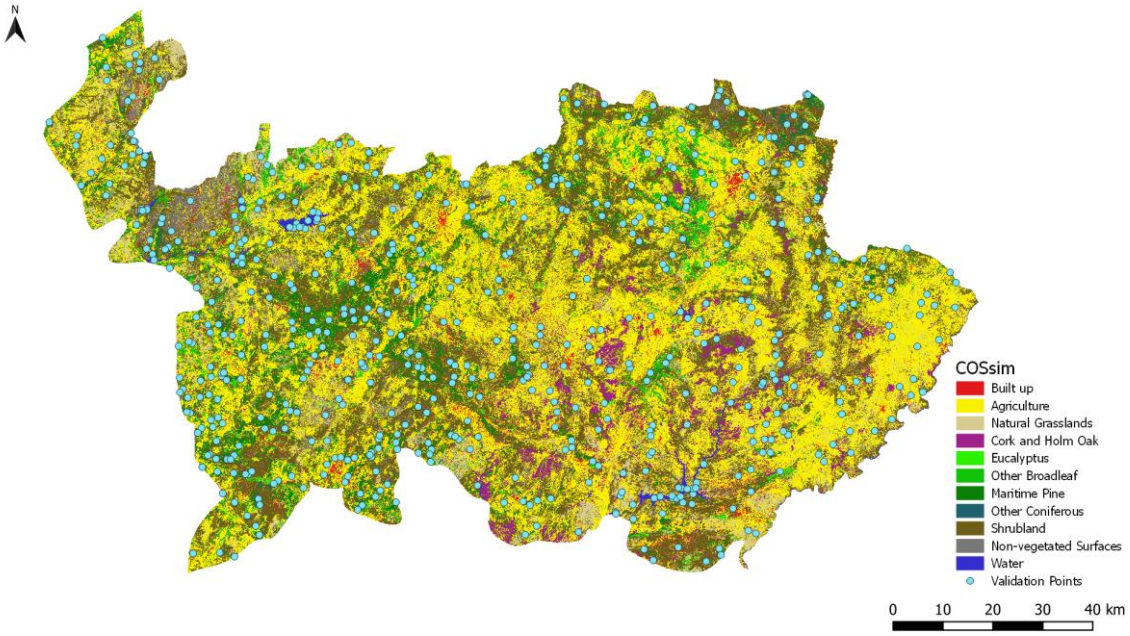

Figure 10: Classification map produced using stratification and combination of manual and automatic training. Points represent the distribution of the validation sample.

The map exhibits a predominance of agriculture throughout the study area, with cork and holm oak restricted to the eastern part of the region. Patches of eucalyptus occur especially in the southeast and southwest part of the map. The northwest portion of the map depicts a concentration of non-vegetated surfaces, corresponding to the rocks present in this mountainous area. In this particular location, it is possible to observe some errors caused by

classifying bare rock, i.e. non-vegetated surfaces, as built up. Maritime pine forest are condensed in the western part of the map, in agreement with COS. In addition, natural grasslands and shrublands are spread throughout the entire region.

		Reference										
	Class	BUP	AGR	NGL	CHO	EUC	OBL	MTP	OCF	SBL	NVS	WTR
Classification	BUP	16	2	14	1			1		1	2	
	AGR		56	31			15	1		7		
	NGL		1	33			2			3	1	
	CHO			1	5							
	EUC			1		6	1	2	1	3		
	OBL			1			38					
	MTP		1			7	3	84	13	1		
	OCF					5	1	20	11	1		
	SBL		1	5		6	8	16	8	96	1	
	NVS	1		4						4	7	1
	WTR											49

Table 12: Confusion matrix of the classification. BUP: Built up, AGR: Agriculture, NGL: Natural Grasslands, CHO: Cork and Holm Oak, EUC: Eucalyptus, OBL: Other Broadleaf, MTP: Maritime Pine, OCF: Other Coniferous, SBL: Shrubland, NVS: Non-vegetated Surfaces, WTR: Water.

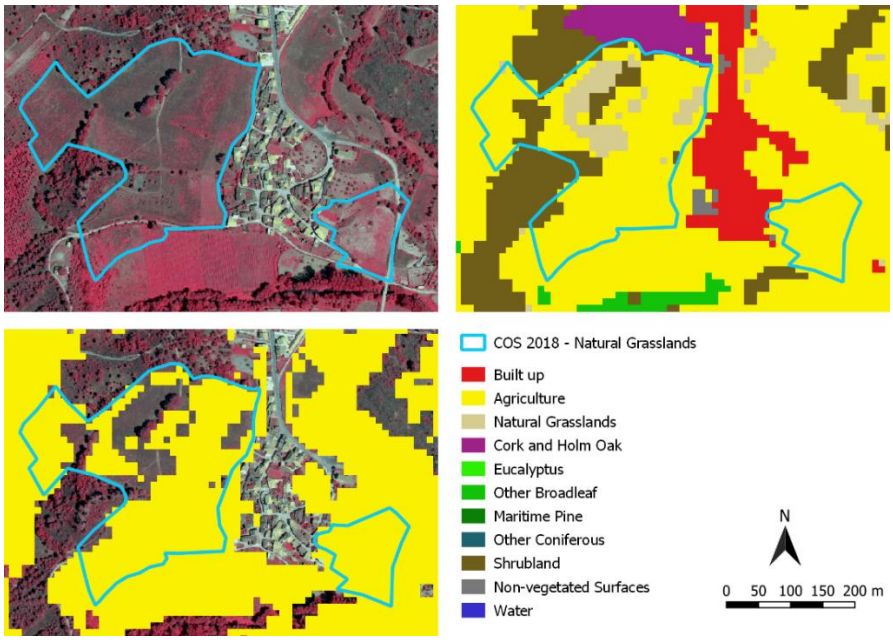

Figure 11: Example of confusion between agriculture and natural grasslands.

The confusion matrix of the classification is presented in Table 12. The data indicates a notable confusion between natural grasslands and built up, natural grasslands and agriculture, other broadleaf and agriculture, maritime pine and other coniferous and shrubland and maritime pine. Confusion among agriculture and natural grasslands is evidenced in Figure 11, where an area identified as natural grasslands in COS 2018 is classified as managed grasslands, i.e. agriculture according to COSsim Level 3. Overall, confusion was predominant between vegetated classes. Cork and holm oak appeared to be the class which benefited the most from the process of

stratification and manual training, showing small omission and commission errors. Agriculture and built up exhibited only a few omission errors, whilst the commission errors were significantly more abundant. Moreover, water was the class with smaller classification errors, as expected.

5.2. Comparison with benchmark classification

The benchmark classification, conducted without stratification and manual training, is depicted in the map of Figure 12. The classification overall accuracy was 60.2%. In comparison with the accuracy of the classification performed with stratification and combination of automatic and manual training (66.7%), this result indicates that the former classification has slightly higher accuracy. However, considering the confidence interval of 95% (approximately ±4% for both classifications), the difference in accuracy was not statistically significant.

The analysis of the map reveals a predominance of shrubland, natural grasslands and agriculture. In contrast with the classification map of section 5.1, there was a reduction in the abundance of agricultural areas, which transitioned to shrubland and natural grasslands in the benchmark map. Another dissimilarity involves cork and holm oak, which in the benchmark are no longer limited to the eastern part of the region. In addition, the rocks in the northwest are still present, despite an increase in misclassification of rocks as built up. Fewer eucalyptus were mapped, most of them being located in the southwest.

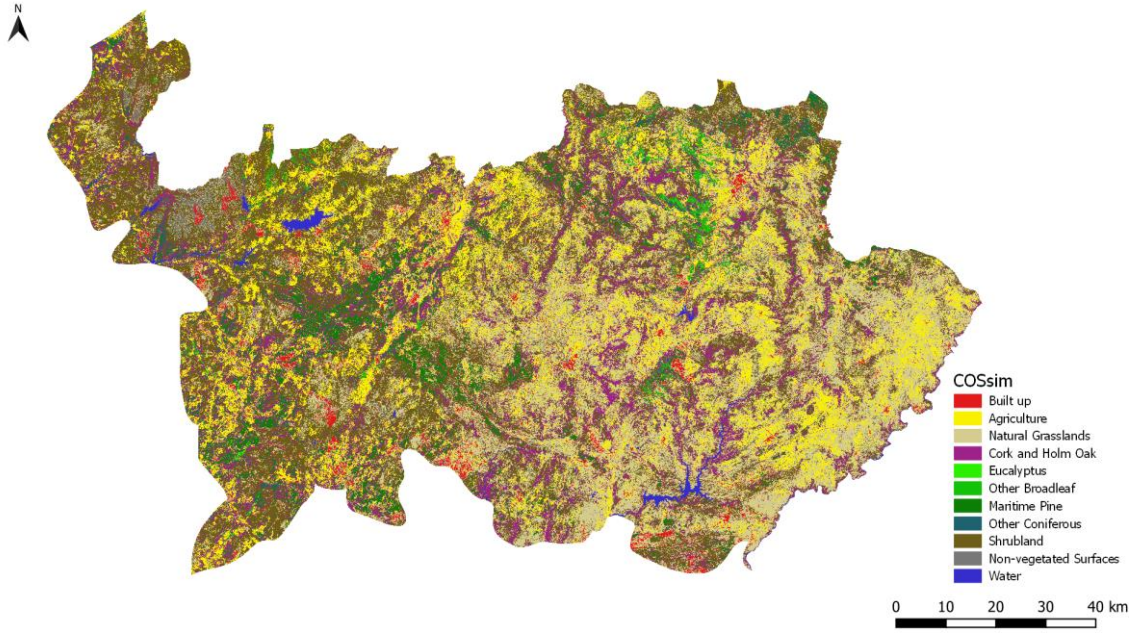

Figure 12: Benchmark classification map, produced without stratification and manual training sample.

		Reference										
	Class	BUP	AGR	NGL	CHO	EUC	OBL	MTP	OCF	SBL	NVS	WTR
Classification	BUP	16	1	12				1			4	
	AGR		27	16			4					
	NGL		9	79			6			6	1	
	CHO			1	6	14	27	18	3	15		
	EUC					2	1					
	OBL						21					
	MTP					7	2	77	15	2		
	OCF					4	3	24	11	2		
	SBL		3	12	1	1	7	9	4	70		
	NVS									2	4	1
	WTR											49

Table 13: Confusion matrix of the benchmark classification. BUP: Built up, AGR: Agriculture, NGL: Natural Grasslands, CHO: Cork and Holm Oak, EUC: Eucalyptus, OBL: Other Broadleaf, MTP: Maritime Pine, OCF: Other Coniferous, SBL: Shrubland, NVS: Non-vegetated Surfaces, WTR: Water.

Table 13 presents the confusion matrix of the benchmark classification and the accuracy metrics per class are shown in Table 14. In terms of F1-score, the introduction of stratification and manual training benefited all classes, except for non-vegetated surfaces, built up and natural grasslands. The last two showed a decrease in F1-score, even though they were expected to benefit from the manual training of their spectral subclasses. Cork and holm oak, eucalyptus and other broadleaf were the classes that benefited the most, with an increase in F1-score of 70.15%, 18.68% and 25.38%, respectively. In the case of cork and holm oak, stratification and manual training caused a substantial reduction in commission error, as seen in the precision and in the comparison between both confusion matrices. Despite not having manual training, some classes, e.g. other broadleaf, exhibited increases in F1-score. Regarding the precision and recall, only maritime pine and shrubland presented an increase in both metrics simultaneously. A reduction was observed in built up and the remaining classes had a tradeoff between increase and decrease in precision and recall.

Class	Precision (%)		Recall (%)		F1-score (%)	
	Benchmark	SMT	Benchmark	SMT	Benchmark	SMT
Built up	47.06	43.24	100.00	94.12	64.00	59.26
Agriculture	57.45	50.91	67.50	91.80	62.07	65.50
Natural Grasslands	78.22	82.50	65.83	36.67	71.49	50.77
Cork and Holm Oak	7.14	83.33	85.71	83.33	13.19	83.33
Eucalyptus	66.67	42.86	7.14	25.00	12.90	31.58
Other Broadleaf	100.00	97.44	29.58	55.88	45.65	71.03
Maritime Pine	74.76	77.06	59.69	67.74	66.38	72.10
Other Coniferous	25.00	28.95	33.33	33.33	28.57	30.99
Shrubland	65.42	68.09	72.16	82.76	68.63	74.71
Non-vegetated Surfaces	57.14	41.18	44.44	63.64	50.00	50.00
Water	100.00	100.00	98.00	98.00	98.99	98.99

Table 14: Benchmark classification accuracy assessment and comparison with classification performed with stratification and manual training (SMT).

In addition to the analysis of the accuracy metrics, a visual inspection was conducted to evaluate the differences among the maps. Figure 13 presents the classifications for an area affected by

forest fires in 2017 (stratum 2). The comparison between classifications reveals that stratification and manual training of burned natural grasslands and eucalyptus (Figure 13c) might have contributed to reduce the misclassification of built up within burned areas.

Another demonstration of how stratification and incorporation of manual training might have improved the classification is exhibited in Figure 14. In this example, an area encompassed by stratum 4 (forest cuts 2015-2018), identified as eucalyptus forest in COS 2018, was mapped mostly as shrubland by the benchmark classification (Figure 14b). The new classification, on the other hand, mapped correctly most of the eucalyptus. The benefit observed in this situation can be explained by the introduction of manual training classes eucalyptus young cuts and eucalyptus 1 year cuts.

Regarding the cork and holm oak stratum, limiting the areas where pixels can be classified as cork and holm oak caused a significant reduction in the dispersion of cork and holm oak throughout the region, as seen in Figure 15. In comparison with COS 2018 (Figure 15a), the benchmark classification was notably dissimilar, whereas the new classification exhibited a spatial distribution closer to COS's.

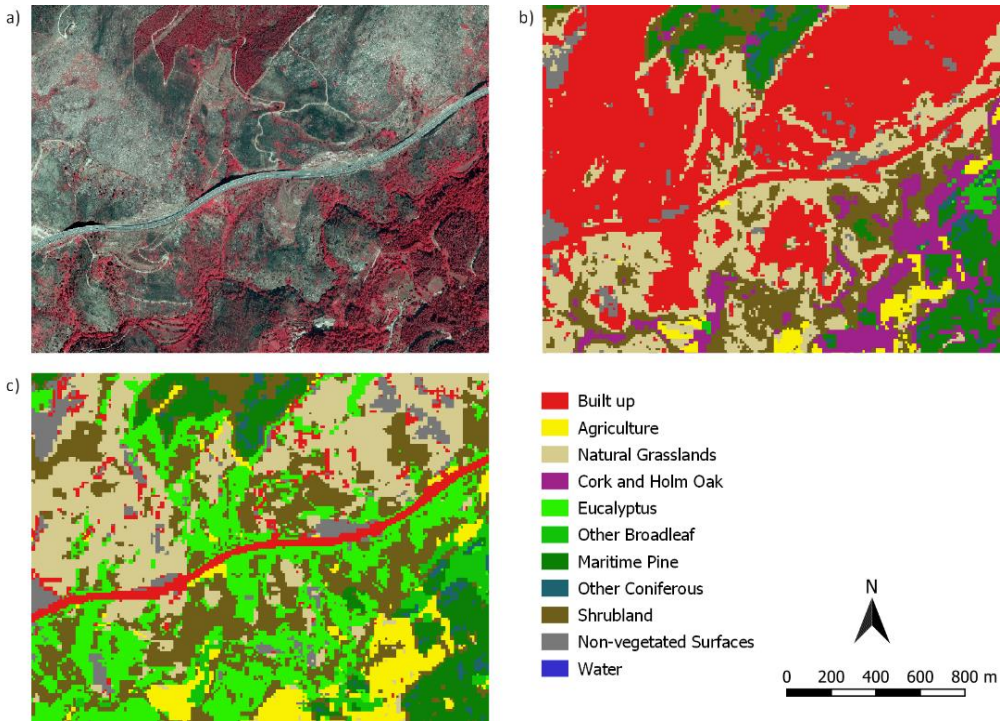

Figure 13: Benefits of stratification and manual training – a) orthophoto of an area affected by fires in 2017 (stratum 2); b) benchmark classification map; c) map produced with stratification and manual training.

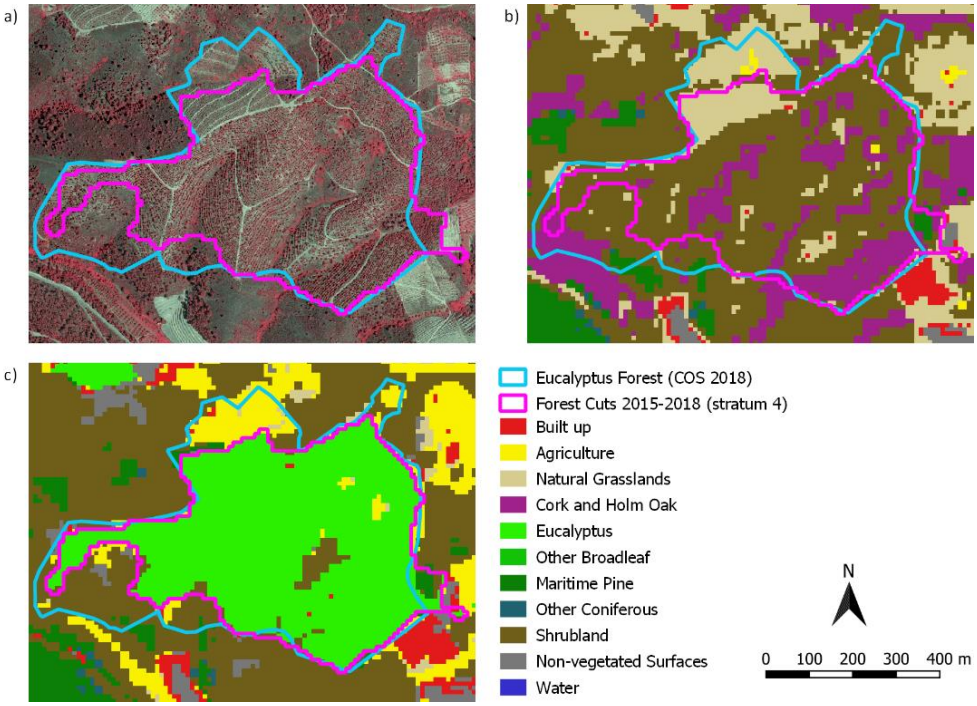

Figure 14: Benefits of stratification and manual training – a) orthophoto of an area where forest cuts occurred (stratum 4); b) benchmark classification map; c) map produced with stratification and manual training.

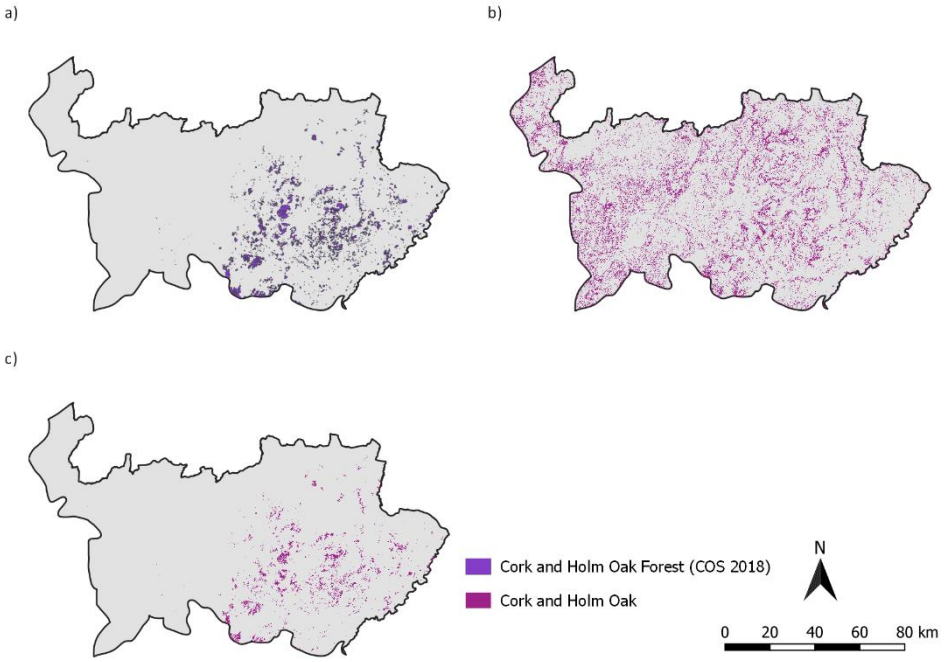

Figure 15: Spatial distribution of cork and holm oak according to a) COS 2018; b) the benchmark classification; c) the classification conducted with stratification and manual training

Validation and on map visual inspection provided different insights about the impact of employing stratification and manual training. The small improvement observed by the accuracy assessment of the validation sample was considered not statistically significant. This can be

explained by the predominance of the complementary stratum in the study region, as it accounts for 86.2% of the area, whereas the second largest stratum, Forest Cuts 2015-2018, corresponds to only 3.9%. In addition, 89.17% of the sampling units in the validation sample belong to the complementary stratum. Therefore, it was expected that the results of the accuracy assessment would be heavily influenced by the complementary stratum, which, despite having classes manually trained, encompass general spectral characteristics instead of being distinguished by particular land cover (e.g. burned areas), making it similar to the benchmark approach of classifying the study area regardless of stratum. However, visual inspection of both maps demonstrated that the innovative approach might have contributed to improve the map, although most improvements were observed outside of the complementary stratum, which represent a small fraction of the total area.

5.3. Influence of sample size

Eight classifications with variable training sample size were performed in the complementary stratum. The results computed with the validation dataset exhibited fairly similar classification accuracies, despite the significantly different sample sizes (Figure 16). The values remained relatively stable even after a reduction of more than 90% in the number of sampling units per class. The highest accuracy (69%) was yielded by the classifications with 2000 and 4000 sampling units per class, whilst the lowest accuracy (66.2%) was observed using 50 sampling units per class. The variation in accuracy was approximately 3% and it was not possible to identify a trend in accuracy as a function of the size of the training sample. The error tolerance of the accuracy estimates was approximately 4% and the confidence intervals overlap, which means that the differences among the classifications' accuracy are not statistically significant.

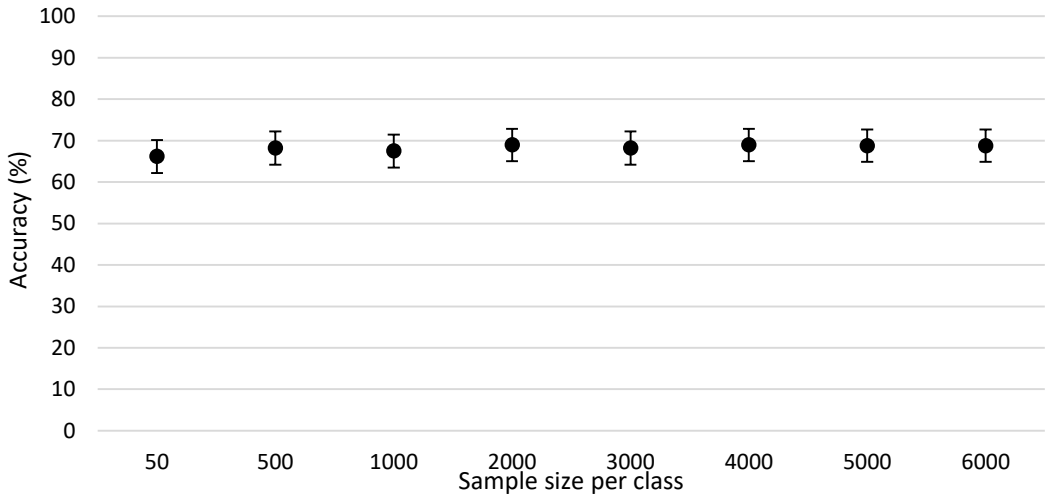

Figure 16: Accuracy estimates and confidence interval of classifications with various sample size.

These results are in accordance with the findings of Rodriguez-Galiano *et al.* (2012) and Thanh Noi and Kappas (2018), which suggested that RF has low sensitivity to variations in sample size. Furthermore, there is indication that smaller samples might be as capable as larger samples to distinguish land cover classes adequately. As shown in Table 10, sampling is not restricted to a small group of polygons, which could mean reduced class variability. A possible explanation for the similar accuracies is the training strategy, which was conducted with spectral subclasses instead of map classes. Spectral subclasses are used to distinguish the spectral diversity present within a map class, thus ensuring that different cover types are taken into account. As a result, such strategy incorporates by design a certain spectral diversity in the samples, regardless of their size.

Spectral diversity can be examined through the distribution of the coefficient of variation (CV) computed for all bands and training classes (Figure 17). The histogram illustrates that samples of 50 and 6000 sampling units per class have similar CV distribution. A closer investigation of the CVs of the near-infrared band for three distinct months (Table 15) also shows comparable values for both sample sizes.

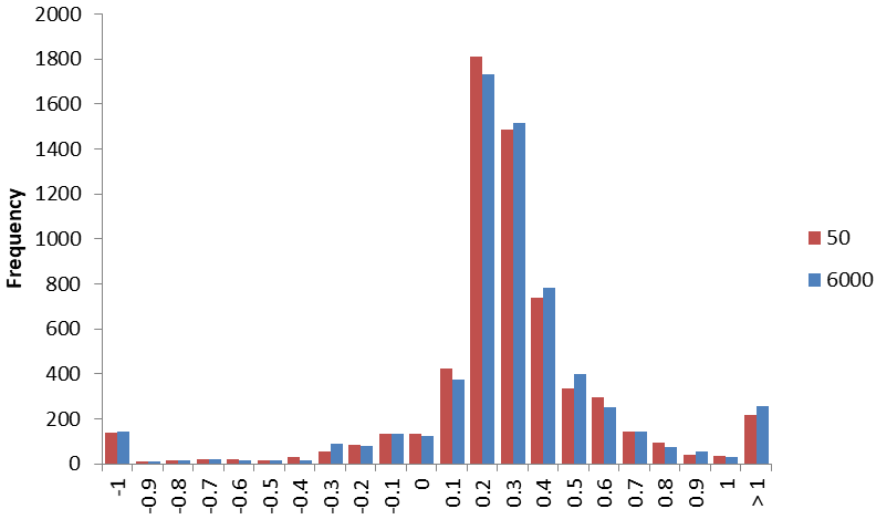

Figure 17: Coefficient of variation computed for all bands and training classes.

Class	Sampling	Oct		Feb		Jul	
		50	6000	50	6000	50	6000
Other Broadleaf	Automatic	0.22	0.23	0.20	0.22	0.10	0.10
Martitime Pine	Automatic	0.18	0.16	0.20	0.18	0.14	0.12
Other Coniferous	Automatic	0.13	0.16	0.14	0.17	0.13	0.17
Agricultural Natural Grasslands	Manual	0.22	0.20	0.17	0.19	0.15	0.16
Mountain Natural Grasslands	Manual	0.13	0.14	0.13	0.17	0.10	0.12
Dense Shrubland	Manual	0.18	0.17	0.26	0.23	0.17	0.16

Table 15: Coefficient of variation of the near-infrared band calculated for October, February and July.

Intra-class variability is exhibited in the scatterplots of Figure 18 and Figure 19, which depict the relationship between red and near-infrared bands of the samples with 50 and 6000 sampling units per class. In spite of having less units, the smaller sample seems to have fairly similar distribution in comparison with the larger. Furthermore, considering the sampling method, the analysis of Table 15 in conjunction with Figure 18 and Figure 19 reveals that variability is similar regardless of the sampling method being automatic of manual. This is particularly relevant since one could expect that the automatic sampling method, due to the application of filters, would result in more homogeneous samples, i.e. low spectral variability, when compared to the manual method.

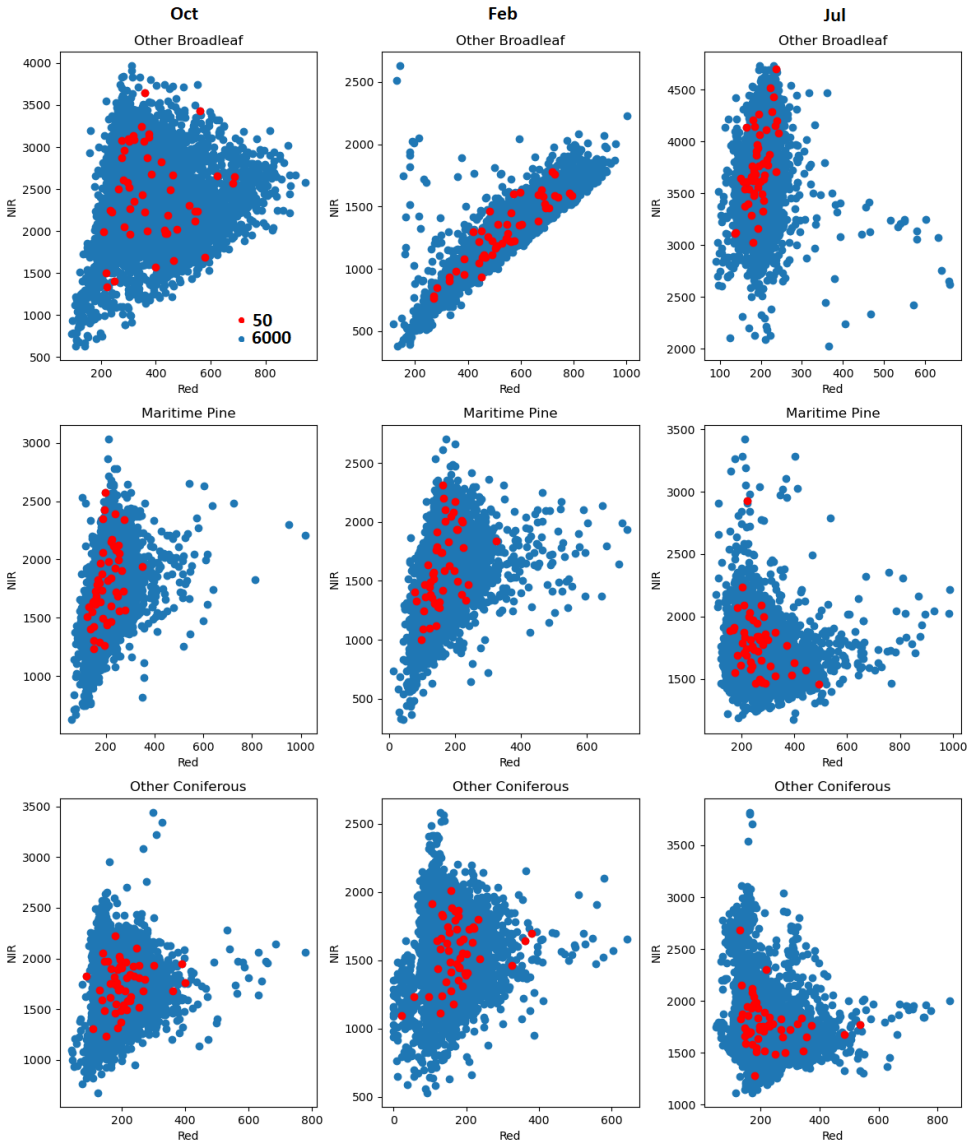

Figure 18: Scatterplots exhibiting the correlation between red (horizontal axis) and near-infrared (vertical axis) bands of samples with 50 (red) and 6000 (blue) sampling units per class for October, February and July. Only automatically sampled training classes considered.

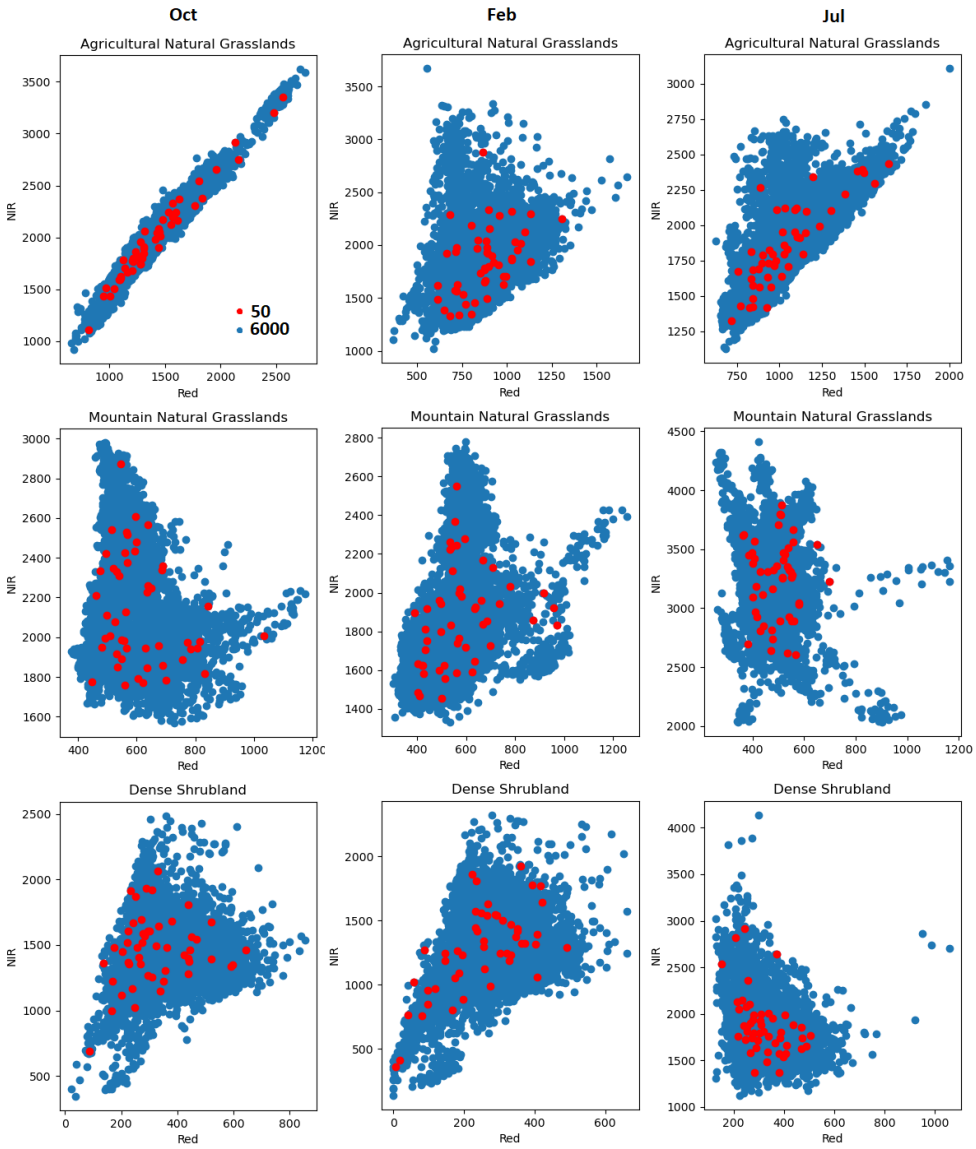

Figure 19: Scatterplots exhibiting the correlation between red (horizontal axis) and near-infrared (vertical axis) bands of samples with 50 (red) and 6000 (blue) sampling units per class for October, February and July. Only manually sampled training classes considered.

Therefore, although collecting large samples automatically might seem advantageous, it may not result in enhancing classification accuracy in the specific case of Random Forest trained with subclasses that contribute to increase spectral diversity. Moreover, RF's sensitivity to variations in sample size in classifications with a large number of predictor variables was found to be comparable to classifications with fewer predictor variables. Despite the sensitivity evaluation, our experiments did not look for the minimum number of training sampling units that could be used before accuracy drops dramatically.

5.4. Improvement of training sample using classification uncertainty

Computation of uncertainty (U) according to the Breaking Ties heuristics produced the map depicted in Figure 20, where values closer to 0 represent high uncertainty, whilst values close to 1 represent low uncertainty. In spite of the following analysis being focused on the complementary stratum, samples were collected regardless of stratum, therefore, the uncertainty was computed for the whole region. The map indicates that an expressive portion of the pixels have U ≤ 0.25. This suggests that those pixels might have spectral responses for which the classifier had difficulties to predict a class, increasing the chance of error. In fact, an analysis of the uncertainty of the validation dataset shows that 47.42% of the sampling units with U ≤ 0.1 were classified incorrectly. This could mean that training samples' spectral diversity did not encompass such pixels. Another hypothesis is that these could be mixed pixels, e.g. transition between classes, which the classifier finds difficult to distinguish. Thus acquiring new sampling units in these areas is expected to improve the classifier's predictive ability.

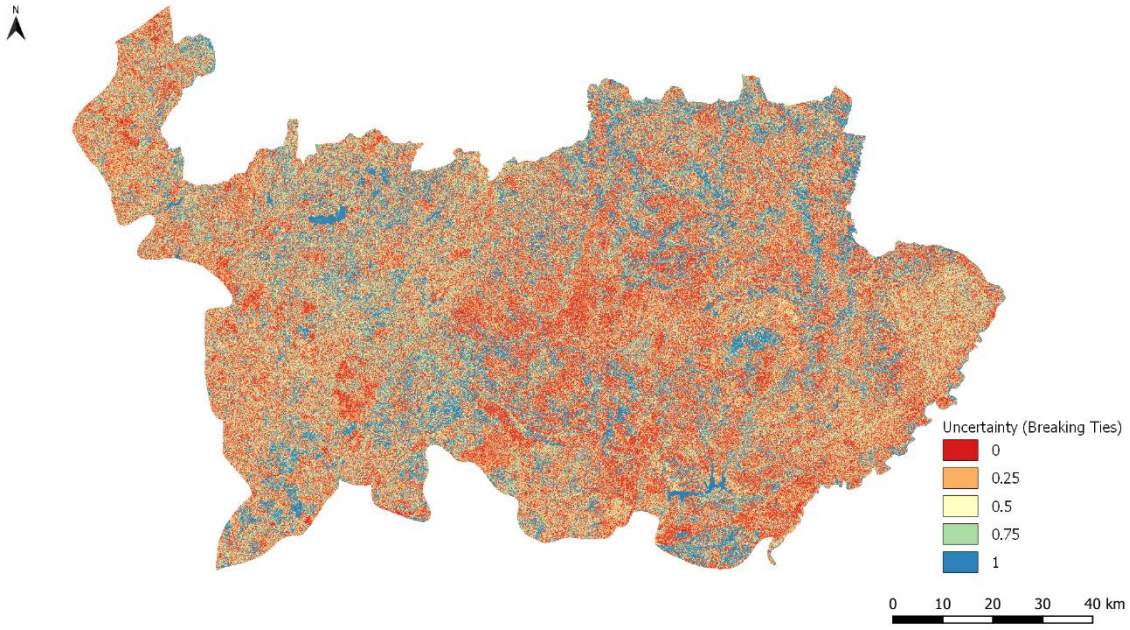

Figure 20: Map of the classification uncertainty, computed using Breaking Ties heuristics.

After the application of the 0.1 threshold, following the 5x5 moving window smoothing, uncertainty patches were delineated. The selected patches of high classification uncertainty comprised a total area of 5866.2 ha. A total of 180 polygons were digitized and labeled on top of the selected uncertainty patches, comprising a total of 19799 new sampling units. Patch delineation of classes industrial and sunflower yielded very small polygons, which were

discarded. In addition, no photointerpreted polygon was labeled as road network or industrial. Then, no sampling units were collected for such classes. As already mentioned in section 4.5, patches or parts of patches located on top of agricultural crops were not photointerpreted, since distinguishing crop type on the orthophoto is not possible. Table 16 presents a summary of the uncertainty patches and the new samples derived from the photointerpretation of polygons contained in the patches. Polygons digitized within the 20 largest patches overall were assigned to their correspondent class in the table.

A number of up to 500 sampling units were collected from the patches, which were incorporated to an initial training sample of up 500 units per class. The validation of the classification conducted with the improved training dataset showed a slight increase in accuracy (approximately 1%), however, according to the confidence intervals (± 3.9%), the difference can be considered not statistically significant (Table 17).

Patch/Training Class	Patch Area (Ha)	Photo-interpreted		Initial Sample Size	Additional Sample Size	Final Sample Size
		Polygons	Sampling units			
20 largest overall	4955.9	N/A	N/A	N/A	N/A	N/A
Built up	36.6	4	40	500	40	540
Industrial	-	-	-	500	-	500
Road Network	38.3	-	-	500	-	500
Oat	98.2	-	-	500	-	500
Wheat	33.3	-	-	500	-	500
Barley	5.1	-	-	500	-	500
Ryegrass	3.4	-	-	500	-	500
Triticale	5.1	-	-	500	-	500
Rye	254.8	-	-	500	-	500
Corn	78.9	-	-	500	-	500
Sunflower	-	-	-	17	-	17
Managed Grasslands	73.6	5	308	500	308	808
Agric. Natural Grassland	31.9	23	3722	500	500	1000
Mount. Natural Grassland	25.1	2	220	500	220	720
Eucalyptus Adult	24.8	6	579	428	500	928
Other Broadleaf	21.6	18	1906	500	500	1000
Maritime Pine	37.4	9	769	500	500	1000
Other Coniferous	22.5	20	1418	500	500	1000
Dense Shrubland	38.7	47	4195	500	500	1000
Baresoil	35.4	27	4782	500	500	1000
Bare Rock	44.1	1	48	500	48	548
Water	13	18	1812	500	500	1000
Total	**5866.2**	**180**	**19799**	**10445**	**4616**	**15061**

Table 16: A summary of the total area collected employing the uncertainty workflow: number of polygons and sampling units available after photointerpretation, number of additional sampling units and final sample size for new training.

Classification	Description	Accuracy (%)
Reference	Original classification, trained with up to 6000 sampling units per class	68.8 ± 3.9
Improved	New classification with additional sampling units collected from areas of high uncertainty	69.7 ± 3.9

Table 17: Comparison of the overall accuracy for the reference and improved classification.

The confusion matrices of both classifications are exhibited in Table 18 and Table 19, respectively. Accuracy metrics by class (precision, recall and F1-score) are shown in Table 20. Considering the F1-score, the improved classification benefited seven classes: built up, natural grasslands, eucalyptus, other broadleaf, maritime pine, other coniferous and water. The addition of new sampling units was most beneficial for natural grasslands and eucalyptus, which had an increase of 12.99% and 19.29% in F1-score, respectively. There was only a small increase in the case of the other five classes, with a tradeoff between reduction and growth in commission and omission errors, which overall contributed to an increase in classification accuracy. Natural grasslands was the only class in which both precision and recall metrics increased, meaning a reduction in commission and omission errors. On the other hand, agriculture, shrubland and non-vegetated surfaces exhibited a decrease in F1-score, thus contributing to a decrease in classification accuracy, having the last two classes seen a reduction in both precision and recall. Non-vegetated surfaces had the highest decrease (16.52%) in F1-score.

An important aspect to be discussed is the training class balance, which was highly affected by the unequal incorporation of additional training sampling units. This was particularly relevant in the case of the agricultural classes, but was also observed in built up, mountain natural grasslands and bare rock, besides all agricultural classes.

		Reference										
	Class	BUP	AGR	NGL	CHO	EUC	OBL	MTP	OCF	SBL	NVS	WTR
Classification	BUP	15	2	11	1						1	
	AGR		55	25			13	1		6		
	NGL		1	24			2			3		
	CHO											
	EUC					5		1				
	OBL			1			37					
	MTP		1			6	3	76	13	1		
	OCF					5	1	19	11	1		
	SBL		1	4		6	8	15	7	91		
	NVS	1		1						4	6	1
	WTR											49

Table 18: Confusion matrix of the reference classification of the complementary stratum for COSsim Level 3. BUP: Built up, AGR: Agriculture, NGL: Natural Grasslands, CHO: Cork and Holm Oak, EUC: Eucalyptus, OBL: Other Broadleaf, MTP: Maritime Pine, OCF: Other Coniferous, SBL: Shrubland, NVS: Non-vegetated Surfaces, WTR: Water.

		Reference										
	Class	BUP	AGR	NGL	CHO	EUC	OBL	MTP	OCF	SBL	NVS	WTR
Classification	BUP	13		9							1	
	AGR		35	14			4			2		
	NGL		3	40			1			5		
	CHO											
	EUC					11	1	3	1	1		
	OBL			1			45	1				
	MTP					3	2	71	10	1		
	OCF		3	2	1	5	2	25	15	5		
	SBL		4	7		4	15	9	4	89		
	NVS	2	3	5						2	5	
	WTR											50

Table 19: Confusion matrix of the improved classification of the complementary stratum for COSsim Level 3. BUP: Built up, AGR: Agriculture, NGL: Natural Grasslands, CHO: Cork and Holm Oak, EUC: Eucalyptus, OBL: Other Broadleaf, MTP: Maritime Pine, OCF: Other Coniferous, SBL: Shrubland, NVS: Non-vegetated Surfaces, WTR: Water.

Class	Precision (%)		Recall (%)		F1-score (%)	
	Reference	Improved	Reference	Improved	Reference	Improved
Built up	50.00	56.52	93.75	86.67	65.22	68.42
Agriculture	55.00	63.64	91.67	72.92	68.75	67.96
Natural Grasslands	80.00	81.63	36.36	51.28	50.00	62.99
Eucalyptus	83.33	64.71	22.73	47.83	35.71	55.00
Other Broadleaf	97.37	95.74	57.81	64.29	72.55	76.92
Maritime Pine	76.00	81.61	67.86	65.14	71.70	72.45
Other Coniferous	27.03	24.56	35.71	51.85	30.77	33.33
Shrubland	68.94	67.42	85.85	84.76	76.47	75.11
Non-vegetated Surfaces	46.15	29.41	85.71	83.33	60.00	43.48
Water	100.00	100.00	98.00	100.00	98.99	100.00

Table 20: Precision, recall and F1-score for both classifications.

Since no additional sampling units were incorporated in the agricultural classes (except for managed grasslands), such lack of improvement might have prevented a higher growth in accuracy. Furthermore, only 40 additional sampling units were collected for RF training class built up, accounting for all the new units of the COSsim Level 3 class built up, as no units were collected for industrial and road network. This might also have contributed to inhibit an increase in classification accuracy.

Furthermore, the analysis of Table 16 reveals that in the case of RF training classes built up, managed grasslands, mountain natural grassland, eucalyptus adult and bare rock only few polygons were delineated within the uncertainty patches. Mountain natural grassland, for instance, had only two polygons from which 220 sampling units were extracted. Assuming that polygons encompass relatively homogeneous areas, the 220 units could be considered redundant, i.e. lack spectral diversity. This could be extrapolated to some extent for the aforementioned classes, which means that RF training classes whose additional sampling units were extracted from a very limited number of polygons might have only a narrow improvement, if any, which may result in minimum impact on classification accuracy. However, while

eucalyptus had only 6 polygons, it was the second class that benefited the most in terms of improvement in F1-score. This could be explained due to eucalyptus being derived from few polygons in the initial training, which were subjected to strict filtering rules. Hence, even though there were few additional polygons derived from the uncertainty patches, they could have generated a gain in terms of spectral diversity. For future works, in order to increase the number of polygons, a larger number of uncertainty patches could be collected. This may involve a fine-tuning of the threshold parameter as well as the number of uncertainty patches collected.

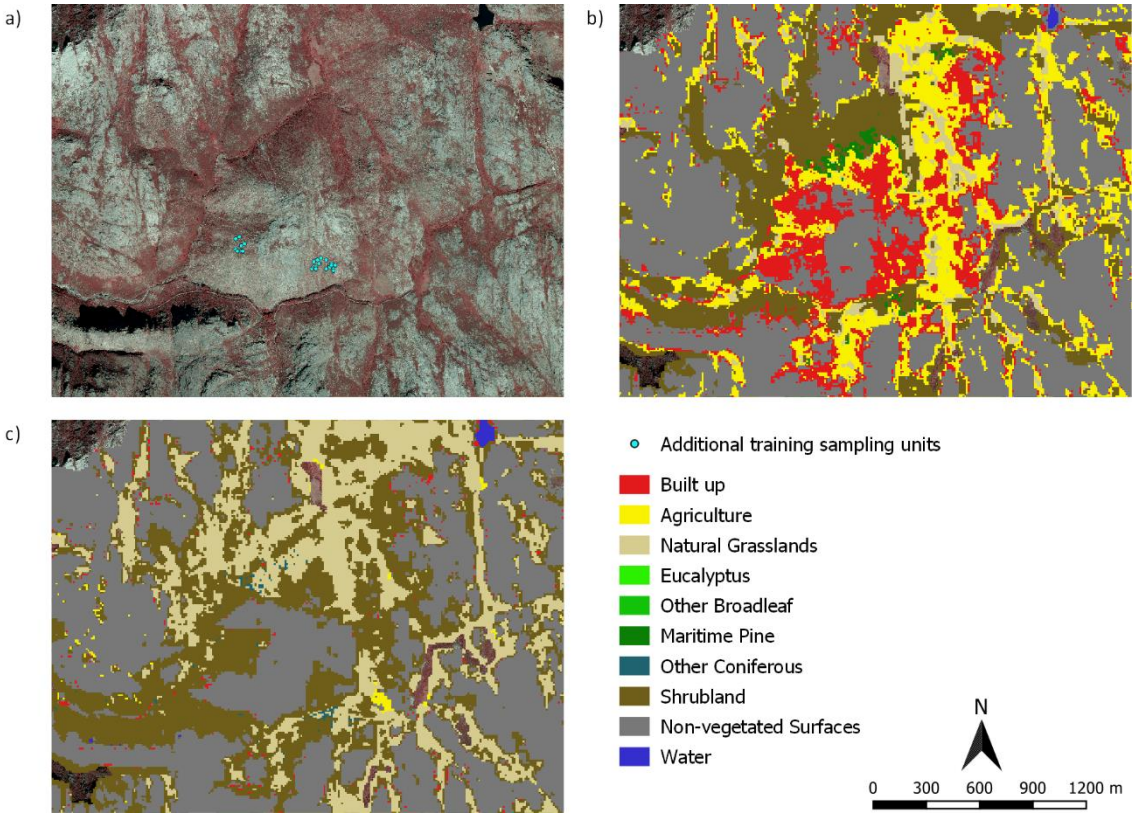

Figure 21: Reduction of misclassifications possibly caused by adding new training sampling units – a) orthophoto of a mountainous area and location of additional training sampling units derived from areas of high uncertainty; b) reference classification; c) improved classification. Pixels outside the complementary stratum were not classified.

In addition to the accuracy assessment, a visual inspection of the reference and improved classification maps was conducted. Overall, the classifications were relatively similar, as 74.11% of the pixels had identical values in both maps. Besides, the spatial distribution of specific cover types, e.g. eucalyptus and maritime pine, was similar among the classifications. However, it could be observed that numerous pixels classified as agriculture in the reference classification shifted to natural grasslands or non-vegetated surfaces. Although occurring throughout the whole region, this trend was most apparent in the mountains to the northwest, where it too evidenced less misclassifications involving built up, shrubland and non-vegetated surfaces

(Figure 21). This was more evidenced in regions close to additional sampling units, however the same effect could be observed in areas where no additional units were collected (Figure 22).

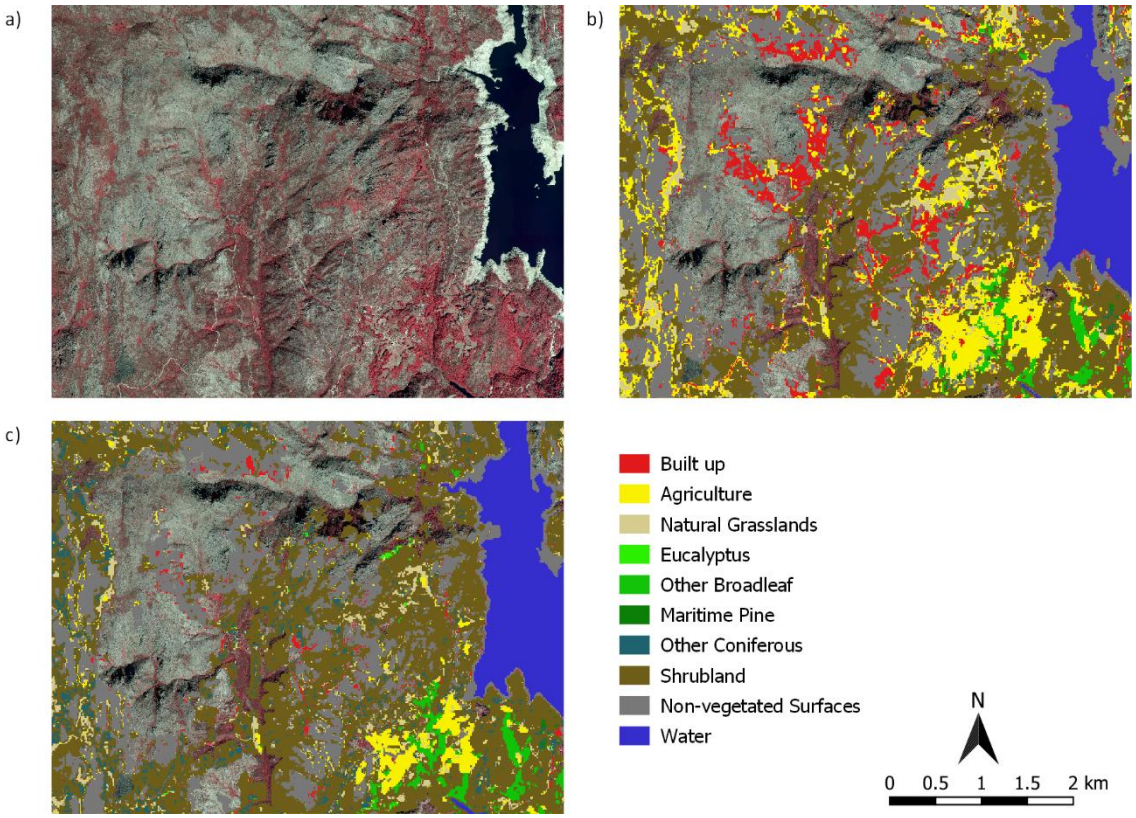

Figure 22: Reduction of misclassifications observed in areas where no additional sampling units were collected – a) orthophoto of a mountainous area; b) reference classification; c) improved classification. Pixels outside the complementary stratum were not classified.

The effect of adding new training sampling units extracted from patches of high uncertainty was also noticeable for other classes. Figure 23 illustrates how the improved classification mapped an area of eucalyptus forest more competently when compared to the reference classification. In this case, a reduction of pixels misclassified as maritime pine is observed. The impact of the additional sampling units is noticed in the whole vicinity, including the in the southeast of the map.

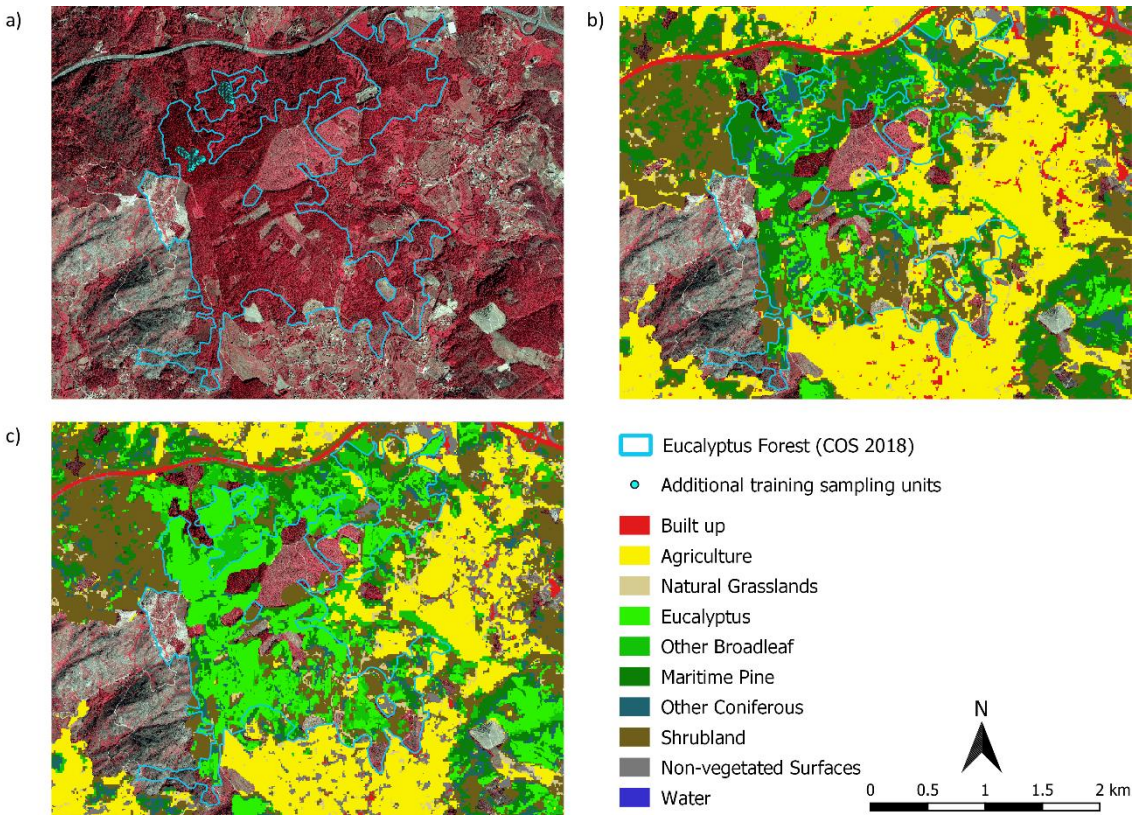

Figure 23: Highlight of the classification of eucalyptus forest – a) false color orthophoto and distribution of additional eucalyptus adult training sampling units; b) reference classification; c) improved classification. Pixels outside the complementary stratum were not classified.

Another example is exhibited in Figure 24, indicating that the additional sampling units might also have contributed to improve the classification in this area. The effects of the new training points are extended beyond their vicinity, for instance, in the south of the mapped region. Zones misclassified as eucalyptus and other coniferous in the reference classification switched to other broadleaf in the improved classification. Again, it is possible to see that the improvement in classification was evidenced not only in areas near the additional training sampling units, but also spread across the entire study region (Figure 25).

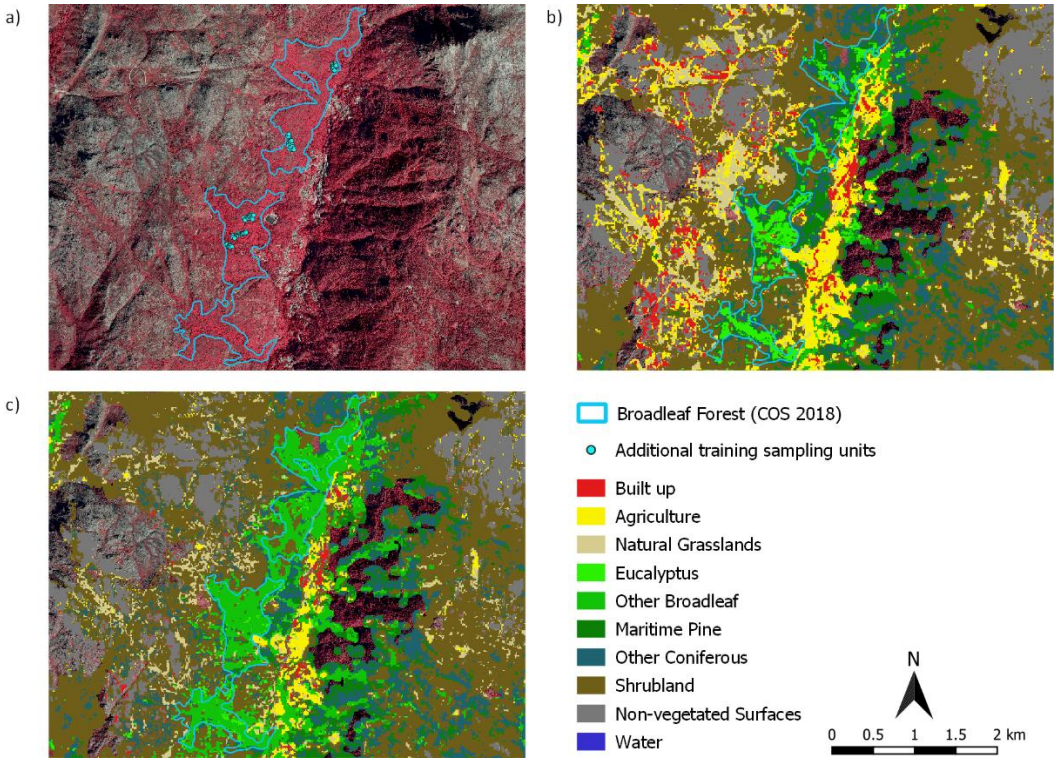

Figure 24: Highlight of the classification of other broadleaf – a) false color orthophoto and distribution of additional other broadleaf training sampling units; b) reference classification; c) improved classification. Pixels outside the complementary stratum were not classified.

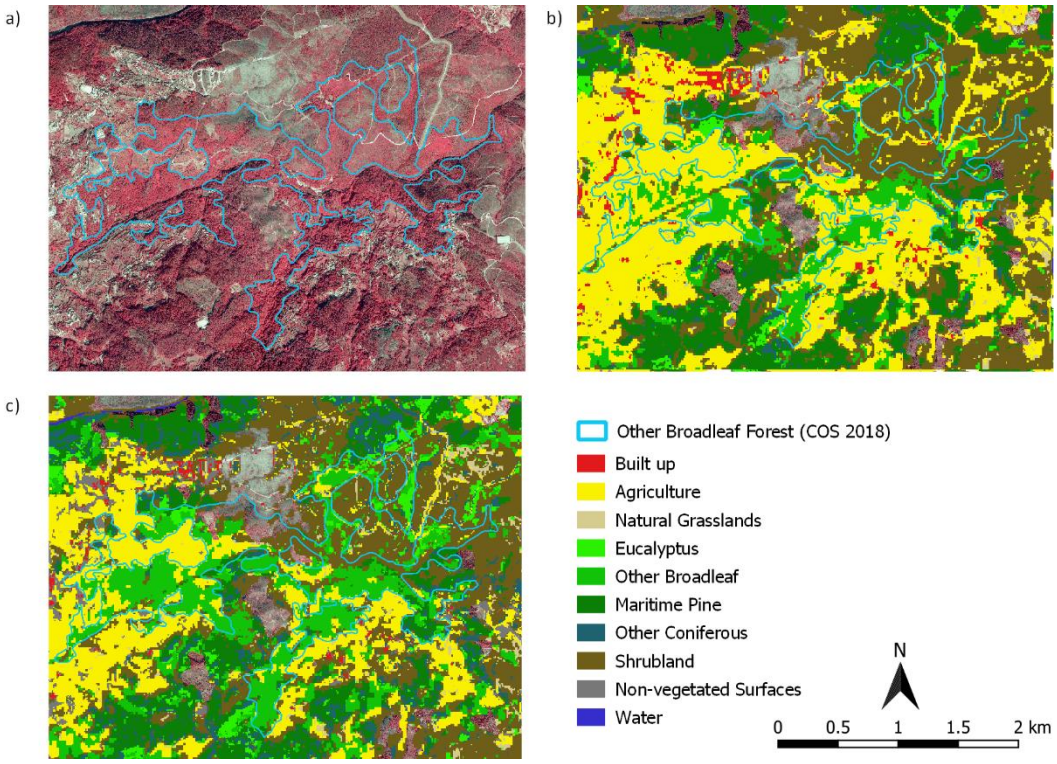

Figure 25: Highlight of the classification of other broadleaf in areas where no additional sampling units were collected – a) false color orthophoto; b) reference classification; c) improved classification. Pixels outside the complementary stratum were not classified.

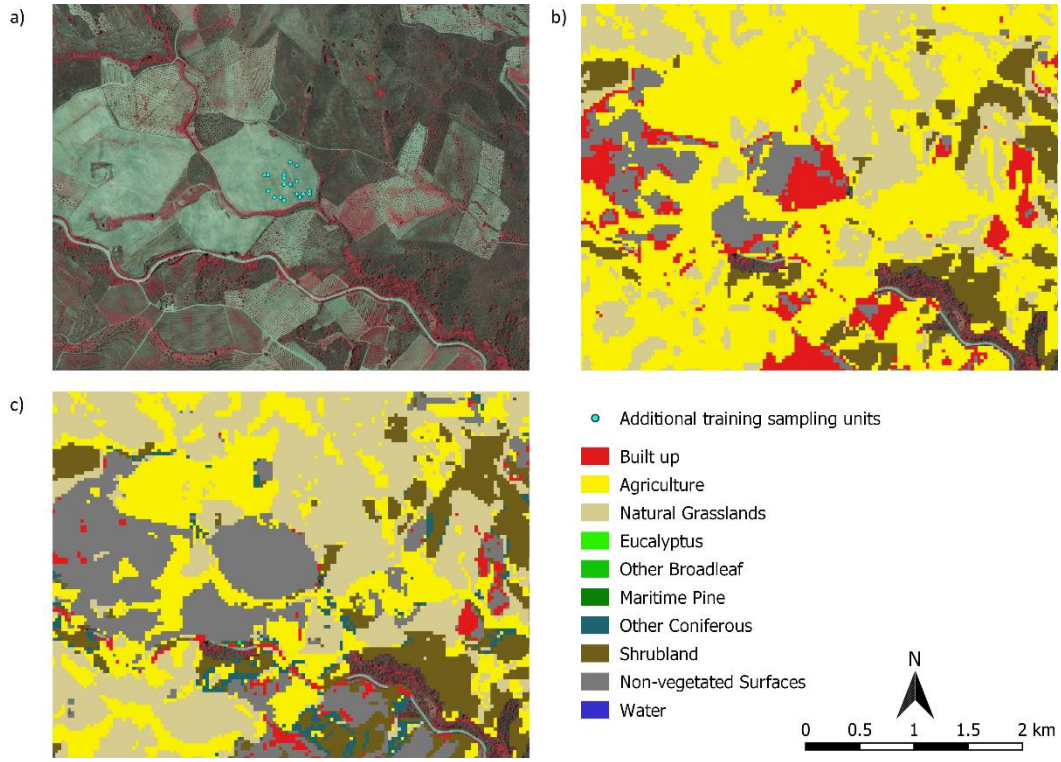

Figure 26: Highlight of the classification of baresoil – a) false color orthophoto and distribution of additional baresoil training sampling units; b) reference classification; c) improved classification. Pixels outside the complementary stratum were not classified.

Additional training might also have contributed to reduce the confusion between built up and baresoil, as illustrated in Figure 26. In this example, areas misclassified as built up in the reference classification were classified as non-vegetated surfaces after the introduction of new training.

The use of additional sampling units collected from areas of high classification uncertainty to improve the training dataset did not provoke a statistically significant impact on COSsim Level 3 classification accuracy. The results might have been hindered by the lack of new sampling units for some training classes as well as by extracting new units from a limited number of polygons. Despite the poor statistical results, visual inspection of the classification map suggested that the additional sampling units might have contributed to improve the classification in particular areas.

6. CONCLUSION

In this work, a Random Forest supervised classification of multi-temporal Sentinel-2 data adopting an innovative process of stratification and combination of automatic and manual training was conducted to map land cover in Trás-os-Montes, Portugal. Three main research objectives were proposed: to assess the impact of incorporating stratification and manual training in the process of classification, to assess the influence of variation in training sample size in classification accuracy and to evaluate whether incorporating new training sampling units extracted from areas of high classification uncertainty could improve classification accuracy.

The classification workflow consisted in mapping LCLU based on satellite imagery and using existing reference datasets to extract training samples automatically. A process of stratification of the study region and introduction of manual training samples was adopted to improve the classification. The implementation of the classification allowed mapping LCLU with an accuracy of 66.7%. Therefore, this method can be considered appropriate to be employed as an operational LCLU mapping strategy at the country level and serve as a model to other countries, provided the necessary reference cartographies. However, due to the complexity and dimensionality of the feature space, the classification workflow had a high computational cost. In this context, future studies within the topic of variable selection can contribute to reduce the feature space dimension and make the classification more efficient. Preliminary experiments have already suggested that a similar classification performance could be achieved using about 40 of the 285 features used in this study.

The assessment of the employment of stratification and manual training indicated that the difference in classification accuracy was not statistically significant when compared to the benchmark classification, conducted without stratification and using only automatic training sampling. The analysis of the accuracy metrics by class revealed that there were improvements for the majority of the classes. Yet, some classes whose training was partially manual exhibited a deterioration in accuracy. Since the complementary stratum accounts for over 86% of the study area and over 89% of the validation sample, the results were strongly influenced by such stratum, which comprises general spectral characteristics instead of covering areas with a particular landscape pattern. Therefore, the effects of stratification and manual training in classification accuracy might have been hindered by the size of the complementary stratum, which biased the comparison with the benchmark classification. In addition to the accuracy assessment, a visual inspection was conducted in order to compare both maps, which evidenced

potential improvements following the use of stratification and manual training. Most of the improvements were observed outside the complementary stratum, therefore representing a small fraction of the study area. However, experiments conducted by DGT in other study regions, where the distribution of the strata is less imbalanced, indicated that stratification and manual training resulted in better classifications, what encouraged DGT to adopt this approach in the production of COSsim. In terms of further investigations, it could be convenient to assess the impacts of stratification and manual training separately.

Regarding the variation in training sample size, the results converged with what the literature suggests, revealing that differences in the accuracy of Random Forest classifications of complex feature space were not statistically significant, even after a reduction of over 90% in the training sample size. The investigation of such results indicated that the approach employed in the training, which consists in using spectral subclasses of the land cover class, might have contributed to produce small differences, as it ensures that spectral diversity is introduced in the samples independently of their size. Furthermore, evaluation of the automatically and manually collected samples showed that there is minimum difference in spectral variability between the smaller and larger samples. The experiments ratified Random Forest's low sensitivity to alterations in sample size, which means that increasing training sample size does not necessarily produce higher accuracy for classifications with a large number of predictor variables. This could affect the design of operational mapping workflows that rely on automatic training sample collection, since in the case of Random Forest collecting large samples might not yield higher classification accuracy if using spectral subclasses to assure spectral diversity.

Concerning the incorporation of training sampling units extracted from areas of high classification uncertainty, the change in accuracy was considered not statistically significant. The analysis of the accuracy metrics by class revealed that built up, other broadleaf, maritime pine, other coniferous and water had a minor overall improvement in accuracy, whilst natural grasslands and eucalyptus exhibited a higher improvement. On the other hand, agriculture, shrubland and non-vegetated surfaces had an overall deterioration in accuracy. The lack of additional training sampling units collected from areas of high classification uncertainty in the case of road network, industrial and all agricultural classes except for managed grasslands might have prevented achieving higher classification accuracy. Furthermore, some classes had their additional sampling units extracted from a small number of polygons, which can limit the spectral diversity of the new sample, thus reducing the impact on classification accuracy. Despite

the poor performance regarding accuracy statistics, visual inspection of the classification map indicated that the additional samples collected from areas of high classification uncertainty might have contributed to enhance map accuracy, particularly by mitigating the confusion associated to specific classes. Nevertheless, further investigation can be conducted to better explore the potential of classification uncertainty, which involves collecting new sampling units for the classes that lacked and developing a strategy to increase the number of training polygons drawn within the uncertainty patches.

BIBLIOGRAPHIC REFERENCES

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

Anderson, K., Ryan, B., Sonntag, W., Kavvada, A. and Friedl, L., 2017, Earth observation in service of the 2030 Agenda for Sustainable Development. *Geo-spatial Information Science*, 20:2, 77-96.

Baraldi, A., Puzzolo, V., Blonda, P., Bruzzone, L. and Tarantino, C., 2006, Automatic spectral rule-based preliminary mapping of calibrated Landsat TM and ETM+ images. *IEEE Transactions on Geoscience and Remote Sensing*, 44:9, 2563-2586.

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.

Bossard, M., Feranec, J., & Otahel, J., 2000, CORINE land cover technical guide: Addendum 2000.

Boutaba, R., Salahuddin, M. A., Limam, N., Ayoubi, S., Shahriar, N., Estrada-Solano, F., and Caicedo, O. M., 2018, A comprehensive survey on machine learning for networking: evolution, applications and research opportunities. *Journal of Internet Services and Applications*, 9:1, 16.

Breiman, L., 2001, Random forests. *Machine learning*, 45:1, 5-32.

Cihlar, J., 2000, Land cover mapping of large areas from satellites: Status and research priorities. *International Journal of Remote Sensing*, 21:6-7, 1093-1114.

Close, O., Benjamin, B., Petit, S., Fripiat, X., and Hallot, E., 2018, Use of Sentinel-2 and LUCAS database for the inventory of land use, land use change, and forestry in Wallonia, Belgium. *Land*, 7:4, 154.

Costa, H., Benevides, P., Marcelino, F., and Caetano, M., 2020, Introducing automatic satellite image processing into land cover mapping by photo-interpretation of airborne data. *The International Archives of Photogrammetry*, *Remote Sensing and Spatial Information Sciences,* 42, 29-34.

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

Direção-Geral do Território, 2018, Especificações técnicas da Carta de uso e ocupação do solo de Portugal Continental para 1995, 2007, 2010 e 2015. Relatório Técnico. Direção-Geral do Território, Portugal.

Direção-Geral do Território, 2019, Especificações técnicas da Carta de uso e ocupação do solo de Portugal Continental para 2018. Relatório Técnico. Direção-Geral do Território, Portugal.

Direção-Geral do Território, 2020, Technical specification for intra-annual Sentinel-2 surface reflectance composites. Relatório Técnico. Direção-Geral do Território, Portugal.

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., Meygret, A., Spoto, F., Sy, O., Marchese, F.and Bargellini, P., 2012, Sentinel-2: ESA's optical high-resolution mission for GMES operational services. *Remote sensing of Environment*, 120, 25-36.

Feng, S., Zhao, J., Liu, T., Zhang, H., Zhang, Z. and Guo, X., 2019, Crop type identification and mapping using machine learning algorithms and sentinel-2 time series data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12:9, 3295-3306.

Foody, G. M., Pal, M., Rocchini, D., Garzon-Lopez, C. X., and Bastin, L., 2016, The sensitivity of mapping methods to reference data quality: Training supervised image classifications with imperfect reference data. *ISPRS International Journal of Geo-Information*, 5:11, 199.

GCOS, 2003, The second report on the adequacy of the Global Observing System for Climate Support of the UNFCC, GCOS-82. Secretariat of the World Meteorological Organization: Geneva, Switzerland, 74 pp.

Gómez, C., White, J. C., and Wulder, M. A., 2016, Optical remotely sensed time series data for land cover classification: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 116, 55-72.

Gonçalves, L. M., Fonte, C. C., Júlio, E. N., and Caetano, M., 2009, A method to incorporate uncertainty in the classification of remote sensing images. *International Journal of Remote Sensing*, 30:20, 5489-5503.

Griffiths, P., van der Linden, S., Kuemmerle, T., and Hostert, P., 2013, A pixel-based Landsat compositing algorithm for large area land cover mapping. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 6:5, 2088-2101.

Griffiths, P., Nendel, C., and Hostert, P., 2019, Intra-annual reflectance composites from Sentinel-2 and Landsat for national-scale crop and land cover mapping. *Remote sensing of environment*, 220, 135-151.

Hermosilla, T., Wulder, M. A., White, J. C., Coops, N. C., and Hobart, G. W., 2015, An integrated Landsat time series protocol for change detection and generation of annual gap-free surface reflectance composites. *Remote Sensing of Environment*, 158, 220-234.

Hermosilla, T., Wulder, M. A., White, J. C., Coops, N. C. and Hobart, G. W., 2018, Disturbance-Informed Annual Land Cover Classification Maps of Canada's Forested Ecosystems for a 29-Year Landsat Time Series. *Canadian Journal of Remote Sensing*, 44:1, 67-87.

Hernandez, I., Benevides, P., Costa, H. and Caetano, M., 2020, Exploring Sentinel-2 for land cover and crop mapping in Portugal. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43, pp. 83-89.

Herold, M., Latham, J. S., Di Gregorio, A. and Schmullius, C. C., 2006, Evolving standards in land cover characterization. *Journal of Land Use Science*, 1:2-4, 157-168.

Hislop, S., Jones, S., Soto-Berelov, M., Skidmore, A., Haywood, A., and Nguyen, T. H., 2018, Using landsat spectral indices in time-series to assess wildfire disturbance and recovery. *Remote Sensing*, 10:3, 460.

Homer, C., Huang, C., Yang, L., Wylie, B., and Coan, M., 2004, Development of a 2001 national land-cover database for the United States. *Photogrammetric Engineering & Remote Sensing*, 70:7, 829-840.

Huang, C., Davis, L. S. and Townshend, J. R. G., 2002, An Assessment of Support Vector Machines for Land Cover Classification, *International Journal of Remote Sensing*, 23:4, 725–749.

Inglada, J., Arias, M., Tardy, B., Hagolle, O., Valero, S., Morin, D., Dedieu, G., Sepulcre, G., Bontemps, S., Defourny, P. and Koetz, B., 2015, Assessment of an operational system for crop type map production using high temporal and spatial resolution satellite optical imagery. *Remote Sensing*, 7:9, 12356-12379.

Inglada, J., Vincent, A., Arias, M., Tardy, B., Morin, D., and Rodes, I., 2017, Operational high resolution land cover map production at the country scale using satellite image time series. *Remote Sensing*, 9:1, 95.

Instituto da Conservação da Natureza e das Florestas (ICNF), 2018, Áreas ardidas, Retrieved from http://www2.icnf.pt/portal/florestas/dfci/inc/mapas (Accessed 12/02/2020).

Lawrence, R. L., and C. J. Moran, 2015, The AmericaView Classification Methods Accuracy Project: A Rigorous Approach for Model Selection. *Remote Sensing of Environment,* 170, 115-120.

Leinenkugel, P., Deck, R., Huth, J., Ottinger, M., and Mack, B., 2019, The potential of open geodata for automated large-scale land use and land cover classification. *Remote Sensing*, 11:19, 2249.

Li, J., Bioucas-Dias, J. M., and 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.

Liu, W., Yang, J., Li, P., Han, Y., Zhao, J., and Shi, H., 2018, A novel object-based supervised classification method with active learning and random forest for PolSAR imagery. *Remote Sensing*, 10:7, 1092.

Loosvelt, L., De Baets, B., Pauwels, V. R., and Verhoest, N. E., 2014, Assessing hydrologic prediction uncertainty resulting from soft land cover classification. *Journal of Hydrology*, 517, 411-424.

Loosvelt, L., Peters, J., Skriver, H., Lievens, H., Van Coillie, F. M., De Baets, B., and Verhoest, N. E., 2012, Random Forests as a tool for estimating uncertainty at pixel-level in SAR image classification. *International Journal of Applied Earth Observation and Geoinformation*, 19, 173-184.

Lu, Q., Ma, Y., and Xia, G. S., 2017, Active learning for training sample selection in remote sensing image classification using spatial information. *Remote Sensing Letters*, 8:12, 1210-1219.

Mack, B., Leinenkugel, P., Kuenzer, C., and 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. and Fang, F., 2018, Implementation of machine-learning classification in remote sensing: an applied review. *International Journal of Remote Sensing*, 39:9, 2784-2817.

McFeeters, S. K., 1996, The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. *International journal of remote sensing*, 17:7, 1425-1432.

Paris, C., Bruzzone, L., and Fernández-Prieto, D., 2019, A novel approach to the unsupervised update of land-cover maps by classification of time series of multispectral images. *IEEE Transactions on Geoscience and Remote Sensing*, 57:7, 4259-4277.

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. and Duchesnay, E., 2011, Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.

Pflugmacher, D., Rabe, A., Peters, M., and Hostert, P., 2019, Mapping pan-European land cover using Landsat spectral-temporal metrics and the European LUCAS survey. *Remote sensing of Environment*, 221, 583-595.

Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., and Rigol-Sanchez, J. P., 2012, An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 67, 93-104.

Roteta, E., Bastarrika, A., Padilla, M., Storm, T., and Chuvieco, E., 2019, Development of a Sentinel-2 burned area algorithm: Generation of a small fire database for sub-Saharan Africa. *Remote Sensing of Environment*, 222, 1-17.

Rouse, J. W., Haas, R. H., Deering, D. W. and Sehell, J. A., 1974, *Monitoring the vernal advancement and retrogradation (Green wave effect) of natural vegetation*. Final Rep. RSC 1978-4, Remote Sensing Center, Texas A&M Univ., College Station.

Shadman Roodposhti, M., Aryal, J., Lucieer, A., and Bryan, B. A., 2019, Uncertainty assessment of hyperspectral image classification: Deep learning vs. random forest. *Entropy*, 21:1, 78.

Shannon, C. E., 1948, A mathematical theory of communication. *The Bell system technical journal*, 27:3, 379-423.

Thanh Noi, P., and Kappas, M., 2018, Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using Sentinel-2 imagery. *Sensors*, 18:1, 18.

Thonfeld, F., Steinbach, S., Muro, J., and Kirimi, F., 2020, Long-term land use/land cover change assessment of the Kilombero catchment in Tanzania using random forest classification and robust change vector analysis. *Remote sensing*, 12:7, 1057.

Townshend, J. R. G., 1992, Land cover. *International Journal of Remote Sensing*, 13:6-7, 1319-1328.

Tuia, D., Volpi, M., Copa, L., Kanevski, M., and 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.

Vuolo, F., Neuwirth, M., Immitzer, M., Atzberger, C., and Ng, W. T., 2018, How much does multi-temporal Sentinel-2 data improve crop type classification?. *International journal of applied earth observation and geoinformation*, 72, 122-130.

Weigand, M., Staab, J., Wurm, M., and Taubenböck, H., 2020, Spatial and semantic effects of LUCAS samples on fully automated land use/land cover classification in high-resolution Sentinel-2 data. *International Journal of Applied Earth Observation and Geoinformation*, 88, 102065.

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

Wulder, M. A., White, J. C., Goward, S. N., Masek, J. G., Irons, J. R., Herold, M., Cohen, W. B., Loveland, T. R. and Woodcock, C. E., 2008, Landsat continuity: Issues and opportunities for land cover monitoring. *Remote Sensing of Environment*, 112:3, 955-969.

Yu, L., Liang, L., Wang, J., Zhao, Y., Cheng, Q., Hu, L., Liu, S., Yu, L., Wang, X., Zhu, P., Li, X., Xu, Y., Li, C., Fu, W., Li, X., Li, W., Liu, C., Cong, N., Zhang, H., Sun, F., Bi, X., Xin, Q., Li, D., Yan, D., Zhu, Z., Goodchild, M. F. and Gong, P., 2014, Meta-discoveries from a synthesis of satellite-based land-cover mapping research. *International Journal of Remote Sensing*, *35:*13, 4573-4588.

Zha, Y., Gao, J., and Ni, S., 2003, Use of normalized difference built-up index in automatically mapping urban areas from TM imagery. *International journal of remote sensing*, 24:3, 583-594.

UNIGIS PT

NOVA IMS Information Management School