D2.4 Report on Accuracy assessment

***MAIL**:* Identifying Marginal Lands in Europe and strengthening their contribution potentialities in a CO2 sequestration strategy

MAIL project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 823805; [H2020 MSCA RISE 2018]

Project title	Identifying Marginal Lands in Europe and strengthening their contribution potentialities in a CO2 sequestration strategy
Call identifier	H2020 MSCA RISE 2018
Project acronym	MAIL
Starting date	01.01.2019
End date	31.12.2021
Funding scheme	Marie Skłodowska-Curie
Contract no.	823805
Deliverable no.	D2.4
Document name	MAIL_D2.4.pdf
Deliverable name	Report on Accuracy assessment
Work Package	WP2
Nature[1]	R
Dissemination[2]	PU

[1]**R** = Report, **P** = Prototype, **D** = Demonstrator, **O** = Other
[2]**PU** = Public, **PP** = Restricted to other programme participants (including the Commission Services), **RE** = Restricted to a group specified by the consortium (including the Commission Services), **CO** = Confidential, only for members of the consortium (including the Commission Services).

Editor	Luis A. Ruiz Juan Pedro Carbonell-Rivera Jesús Torralba Pérez Elke Krätzschmar Charalampos Georgiadis
Authors	Anastasios Stammas Abdulrashid Hassan Georgios Spanos Zoi Touloudi
Contributors	All Project Partners
Date	23/07/2021

MAIL Consortium

Aristotle University of Thessaloniki (AUTH) Greece	Industrieanlagen Betriebsgesellschaft MBH (IABG) Germany
Gounaris N. – Kontos K. OE (HOMEOTECH) Greece	Centrum BadanKosmicznychPolskiejAkademiiNauk (CBK PAN) Poland
Universitat Politecnica de Valencia (UPV) Spain	Fundacion Centro De Servicios Y Promocion FOrestral Y de su Industria De Castilla y Leon (CESEFOR) Spain

ABBREVIATIONS

Term	Explanation
ERR	Error Rate
FN	False Negative
FP	False Positive
Ha	Hectares
ISO	International Standard Organisation
MCC	Matthews Correlation Coefficient
ML	Marginal Land
nML	Non-Marginal Land
PA	Producers Accuracy
OA	Overall Accuracy
S2GLC	Sentinel-2 Global Land Cover
TN	True Negative
TP	True Positive
UA	Users Accuracy

Contents

EXECUTIVE SUMMARY

Accuracy assessment is the final step in the analysis of classification data which enables the accuracy verification of the results. It is carried out once the interpretation/classification has been concluded. As a result, we are interested in assessing the accuracy of thematic maps or classification images, which is known as thematic or classification accuracy. The accuracy assessment examines the agreement between the classification data and the reference data or true class. A true class is what is seen on the ground or from high or very high-resolution images. Uncertainty or lack of information about the true value is associated with accuracy and precision. Accuracy is a relative measure of quality and the exactness of an estimate and accounts for systematic errors, also known as bias.

The objective of this task was the methodological development of an accuracy assessment technique for the evaluation of the result of the first phase "hard constraints" of m/sm MLs Development methodology **MAIL** task 2.3 known as Marginal and Non-marginal land classification map. Statistically robust and transparent approaches for assessing the accuracy and estimating area to ensure the integrity of the classified map were employed in carrying out the accuracy assessment of the classification map and estimating area based on the validation/reference sample data.

The accuracy assessment was quantified by creating an error matrix that compared the classification layer with the validation reference data using point-based and area-based assessments. The point-based assessment was developed by stratified random points and the area-based by the intersection of the two layers. The accuracy metrics that used to evaluate the assessment are the overall accuracy, user's accuracy, producer's accuracy, error rate, Matthew's correlation coefficient and F1-score.

The binary classification measures of accuracy were applied on the test countries of Greece, Spain, Germany, and Poland, as well as on a merged layer including all of them. The point-based overall accuracy was 71.52% in Greece, 82.87% in Spain, 60.61% in Germany, and 90.97% in Poland giving an overall accuracy for all testing sites of 67.98%, while the area-based accuracy assessment produced an overall accuracy of 70.75% in Greece, 83.42% in Spain, 59.79% in Germany, and 90.56% in Poland concluding to an overall accuracy for the fusion of the testing sites of 67.73%. Both accuracy assessment techniques showed very little differences of 0.5 to 1% among the respective testing countries.

Comparing the predicted area of the D2.3 methodology for the class of ML with the actual reference polygons that were provided for each test country by the respective project partners, we quantified the deviation of the predicted ML area for the given sites, from the actual truth. More specifically, the classification underestimated the ML area in Greece by 341 ha and in Poland by 225 ha. Conversely, the actual ML area was overestimated by the classification methodology in Spain by 158 ha and in Germany by 8469 ha.

Analysing the error matrix with the S2GLC map for the provided validation data areas, the majority of the correctly classified as ML samples were over the herbaceous vegetation land cover, with 51.45% in Greece, 44.20% in Spain and, 96.58% in Poland while in Germany most of them were over moors and heathland 49.53% and natural material surfaces 42.27%.

The result of the analysis of the ML validation data areas with the S2GLC map showed that the majority of the ML areas 75.13% in Greece, 39.49% in Spain, and 89.88% in Poland were over the herbaceous vegetation land cover whereas in Germany they were over moors and heathland 47.33% and natural material surfaces 38.52%.

1 INTRODUCTION AND OBJECTIVES

MLs are not clearly defined and may differ in nature and extent throughout the European Union, mostly affected by latitude. Therefore, ML detection is complicated due to its many forms. At this task m/sm MLs will be assessed through field stratified random sampling, based on the stratification that will arise by ***MAIL*** Task 2.3 (e.g., abandoned fields, degraded grasslands, etc.). This assessment was based on sampling units with regional form (polygons). The sample size depends on the number of polygons that were classified as m/sm MLs. Each ML, according to the area that implement the classification will have a specific number of polygons that were classified as m/sm MLs (population - N) and through stratification to subpopulations (strata – N_i, N_{ii},…). The sample size (n) will be given by the implementation of statistical methods according to the binomial distribution. The minimum sampling size of N will be defined by the consortium. The results will be used to refine the proposed methodology and adjust the dependencies between indicators of T2.3.

The main objectives of Task 2.4 (Accuracy assessment of m/sm MLs detection) are:

- Definition of the statistical methods, limits that are going to be used for the accuracy assessment.
- Accuracy assessment of m/sm MLs detection/classification.

2 LITERATURE REVIEW

Many uncertainties surround the meaning and interpretation of map quality, making it a difficult variable to assess objectively and severely limiting the ability to assess the extent to which remote sensing's potential as a source of land cover data is being realized. As a result, while a thematic map offers an unquestionable simplification of reality, it has defects and is just one model or representation of the depicted theme (Woodcock & Gopal, 2000). Accuracy is a difficult property to calculate and express, despite its apparent simplicity in concept. The term accuracy is commonly used in thematic mapping from remotely sensed data to express the degree of 'correctness' of a map or classification. A thematic map derived from a classification can only be considered accurate if it depicts the land cover of the area in an impartial manner (Foody, 2002).

For several years, accuracy assessment has been a source of heated discussion and study in the field of remote sensing. This is partly due to the fact that commonly used standard methods like the kappa coefficient are not always sufficient. Furthermore, the kappa coefficient shares features with other accuracy measures in terms of compensating for chance agreement and allowing the importance of variations in accuracy to be evaluated. However, there are several issues with the evaluation and reporting of classification accuracy that make it difficult to interpret accuracy statements (Foody, 2002).

2.1 Overview of accuracy assessment

In any case, accuracy assessment is an important part of any project that uses spatial data, and there are many explanations for this, including: (1) the need to self-evaluate and learn from errors, (2) the need to quantitatively compare methods and algorithms, and (3) the need to use the information obtained from spatial data analysis in some decision-making process (Congalton, 2001). It is also worth noting, that there is no distinct way to assess map accuracy, just as there is no single way to produce a map as well (Congalton & Green, 2019). Any map's or spatial data set's accuracy is a function of both positional and thematic accuracy; these two types of map accuracy evaluation are described below.

Positional accuracy refers to the precision with which map features are located, and it measures the distance between a geographical feature on a map and its true or

reference position on the ground (Bolstad et al., 2005). More specifically, positional accuracy deals with the accuracy with which a point in imagery is mapped with reference to its physical location on the ground. The ability to determine the same exact position on the image and on the ground is important for any accuracy comparison. Topography, or the natural and artificial physical features of an environment, is the most important factor affecting positional accuracy, while sensor characteristics and viewing angles may also have an effect. Conventionally, positional accuracy has been calculated in terms of the Root Mean Square Error (RMSE). The RMSE is calculated as the sum of the squares of the differences between the location of a point on one data layer and the same point on another data layer, usually the ground, applying the same data that was utilized to register the layers together. The equation (Barnston, 1992) is:

$$RMSE_{fo} = [\sum_{i=1}^{N} (z_{f_i} - z_{o_i})^2 / N]^{1/2}$$

Where:

f = forecasts (expected values or unknown results)

o = observed values (known results)

$(Z_{fi} – Z_{oi})^2$ = differences, squared

N = sample size

As a result, this metric is not an independent indicator of positional accuracy. Alternatively, collecting an independent sample of points from which to compute the RMSE would be more efficient and representative of true accuracy (Congalton, 2007).

Thematic accuracy deals with the labels or attributes of the features of a map, and measures whether the mapped feature labels are different from the true feature label. Thematic accuracy refers to the accuracy of a mapped land cover category at a particular time compared to what was actually on the ground at that time. Land cover classifications must be tested using data that are considered to be accurate in order to perform a meaningful evaluation of accuracy. Thus, it is critical to have some understanding of the accuracy of the reference data before comparing them to the remotely sensed map (Congalton, 2007).

Although these two types of accuracy can be evaluated independently, they are inextricably linked, and failing to consider both is a serious mistake. All accuracy

assessment workflows follow a three-step process: (1) creating an accuracy assessment sample, (2) gathering data for each sample, and (3) interpreting the results (Congalton & Green, 2019).

The accuracy of the final map in a remote sensing project is the result of the accumulation of several errors along the way (Figure 1). Each of the major error sources may contribute to the total error budget separately, and/or through a mixing process. For many applications, it is critical to define and comprehend (1) error sources, as well as (2) the appropriate mechanisms for regulating, minimizing, and/or disclosing the severity of such errors to end-users (Congalton, 2001).

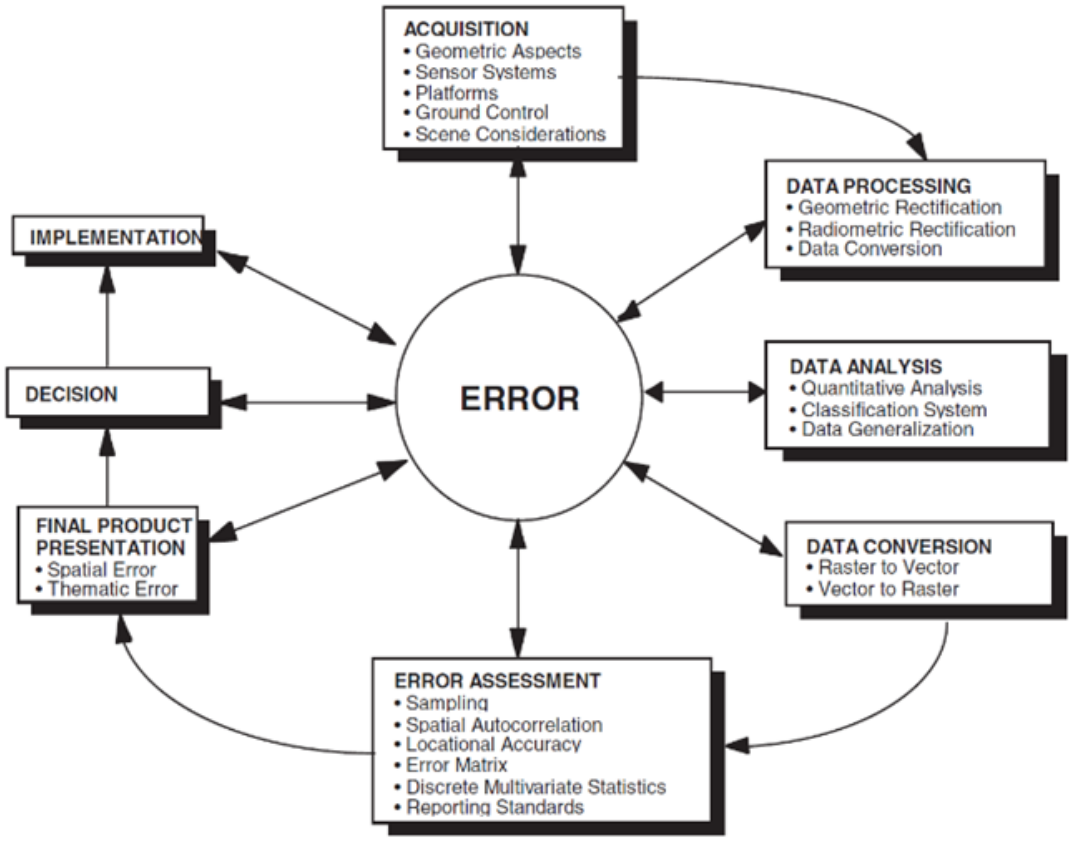


Figure 1. Error sources and accumulation of error in a typical remote sensing project (Lunetta et al., 1991).

When conducting an accuracy assessment, there are various aspects to be considered in addition to the actual analysis techniques. In practice, if these aspects are not taken into account, the techniques are of little use. Thus, in order to produce a correct error matrix, the following factors must be regarded (Campbell, 1981; Congalton, 1988a, 1988b, 1991; Congalton & Green, 1999; Hay, 1979; Stehman, 1992; Van Genderen & Lock, 1977):

- Classification scheme;
- Sampling scheme;
- Spatial autocorrelation;
- Sample size;
- Reference (ground) data collection and;
- Sample unit

Each of these factors contributes valuable information to the assessment analysis and omitting any one of them may result in serious flaws in the assessment process.

2.2 Classification Scheme

When designing a project that makes use of remotely sensed data, it is critical to devote enough attention to the classification scheme that will be used. Classification schemes are a method of processing spatial data in a logical and orderly manner. Any mapping project needs classification schemes because they summarize the total number of objects considered to a manageable number and help sort out the chaos in the row data.

The classification scheme allows the map maker to characterize landscape features, in such way that they are also quickly discernible by the user. There are two essential components to a classification scheme: firstly a set of labels (e.g., water bodies, coniferous forest, etc.); and secondly a set of rules or definitions, such as a dichotomous key for label assignment (e.g. in the climax stage, shrubby formations with sparse trees have a canopy closure of 15 percent or less, and tree heights of over 5-7 m.)(Congalton & Green, 2019).

Any classification system should be mutually exclusive and fully comprehensive. To put it another way, any region that needs to be classified should appertain to only a single category or class and all the image regions should be entailed in the classification. Finally, the use of a hierarchical classification system, if feasible, would be highly beneficial. Particular categories within the classification system may be summarized into more general ones if such a scheme is used.

2.3 Sampling Design

The sampling design dictates how the subset of the map will be chosen, which establishes the basis for the accuracy assessment (Finegold et al., 2016). For the sampling design, the following are considered prerequisites: Being familiar with the

distribution of thematic classes over the study area, concluding on the types and the number of samples to be acquired, and selecting a sampling scheme for picking the samples. One of the most difficult and critical components of any accuracy assessment is the design of an appropriate and effective sample to collect accurate validation and map accuracy data since the design will determine both the cost and the statistical rigor of the assessment (Congalton & Green, 2019).

2.3.1 Sampling scheme

A sampling scheme is an essential part of any accuracy assessment therefore, the right scheme must be mindfully selected in order to generate an error matrix that is indicative of the entire classified image. Poor sampling scheme selection may introduce major biases into the error matrix, potentially overestimating or underestimating the true accuracy. Furthermore, depending on the analysis techniques to be applied to the error matrix, the appropriate sampling scheme may need to be chosen (Congalton, 1991).

Simple random sampling, systematic sampling, cluster sampling, stratified random sampling, and stratified systematic unaligned sampling are some of the sampling schemes that can be used to obtain accurate assessment results.

Each sample unit in the research area has an equal probability of being chosen in a simple random sample. Usually, a random number generator is utilized to obtain random x and y coordinates for collecting samples. Randomness has some interesting statistical properties that are useful for further analysis of the data (Congalton & Green, 2019).

Systematic and cluster sampling can also be beneficial in practice. *Systematic sampling* is a technique in which sample units are chosen at a predetermined and constant interval throughout the study region. The first sample is, typically, chosen randomly, and subsequent samples are taken at predetermined intervals. The main benefit of systematic sampling is the ease with which it can be done uniformly across the entire study area (Congalton & Green, 2019). *Cluster sampling* has proved especially useful in assessing remotely sensed data since it allows quickly collecting information on many samples. While gathering an abundance of sample units in contiguity to one another has some clear advantages, cluster sampling should be used mindfully and cautiously (Congalton & Green, 2019).

Simple random sampling and *stratified random sampling* are similar techniques. However, a level of previous knowledge of the study area is required to split it into smaller regions or strata, which are then randomly sampled. The main benefit of stratified random sampling is that it includes all strata (i.e., map classes), no matter how small they are. This is particularly critical when it comes to ensuring that enough samples, from rare but important map classes, are included (Congalton & Green, 2019).

Finally, *stratified systematic unaligned sampling* aims to combine the benefits of randomness and stratification with the efficiency of a systematic sample while avoiding the shortcomings of systematic sampling's periodicity. This can be considered as a fusion method, which gives more randomness to the stratum than just a random start (Congalton & Green, 2019).

In any case, each scheme has its own set of pros and cons, hence it is critical to comprehend them all and implement the one that is best suited for the situation. The analysis must then be done in accordance with the chosen sampling scheme. Generally, *stratified random sampling* is regarded by some authors as the most suitable sampling method (Congalton, 2001).

2.3.2 Spatial Autocorrelation

Spatial autocorrelation is a measure of the positive or negative impact that a feature at a specific location has on its immediate surroundings. When determining which sampling scheme to use, spatial autocorrelation is an important factor to consider. If there is a positive correlation between samples, it is important to place the samples far enough apart to eliminate this correlation for the precision of the accuracy estimates. This is particularly true for sampling schemes such as cluster sampling and systematic sampling (Congalton, 2001).

When the presence, absence, or degree of one characteristic influences the presence, absence, or degree of the same characteristic in neighbouring units, is known as spatial autocorrelation (Cliff, 1973). This condition is especially significant in accuracy assessment if an error in one location is observed to affect errors in adjacent locations, either in a positive or a negative way (Campbell, 1981).

2.3.3 Sample size

Evaluating the entire mapped area in most projects is a waste of resources, therefore in order to evaluate the classification accuracy, a sample of cases is used. This sample is called the testing set (Foody, 2009). When it comes to developing and interpreting classification accuracy estimates, the sample size is crucial.

Accuracy assessment necessitates the collection of a sufficient number of samples per map class, such that, the result is a statistically accurate representation of the map's accuracy, but as small as possible to reduce the budget (Finegold et al., 2016). To calculate the appropriate sample size, most researchers initially used an equation based on the binomial distribution or a standard approximation to the binomial distribution. These methods are statistically sound for determining the sample size required to calculate the overall accuracy of a classification or even the overall accuracy of a particular category. The equations are dependent on the proportion of sample units that are correctly labelled as well as a margin of error. On the other hand, though, these methods were not intended for selecting a sample size for producing an error matrix.

When it comes to building an error matrix, it is not simply a matter of right or wrong (binomial distribution). On the contrary, it is a matter of determining which errors or categories are being confused. In an error matrix with *n* land cover categories, there is one correct answer for each category and (*n-1*) incorrect answers. To accurately reflect this challenge, enough samples must be collected. Thus, the binomial distribution cannot be used to determine the sample size for an error matrix. The use of a multinomial distribution is suggested instead (Tortora, 1978).

The multinomial distribution can and should be used to calculate the required sample size for each project. However, for maps of less than 1 million acres in size and less than 12 classes, as a "rule of thumb" it is suggested to collect a minimum of 50 samples for each mapping class (Congalton, 1988a). Each class should have 75 to 100 accuracy assessment sites for larger area maps or more complex maps. These guidelines were developed empirically over several projects, and the multinomial equation proved that they strike a good balance between statistical validity and practicality.

As a result, practical considerations are often a major factor in determining sample size. The number of samples for each category, for example, may be modified based

on the relative significance of that category within the mapping project's objectives or the inherent variability within each of the categories (Banko, 1998). Due to budget restrictions or other factors, it is often preferable to focus the sampling on the categories of interest and increase the number of samples taken in those categories, while decreasing the number of samples taken in less influential categories (Finegold et al., 2016).

Summing up, it might seem tempting to create a sample that includes a large number of samples from the most accurate categories and a small number of samples from the most confusing categories. Even though this technique would ensure a high level of accuracy, it would not be indicative of the map's accuracy. It is recommended to make sure that the sampling effort is well-planned and executed and to note that both smaller and larger sample sizes may be problematic (Foody, 2009). In any case, it is important to record the whole process so that potential map users will understand how the assessment was completed.

2.4 Response Design

The response design for the accuracy assessment is the protocol that encompasses all steps that determine whether the map and the reference data are in agreement. Under the assumption that the reference data sources are adequately more accurate than the map classification being evaluated, the response design provides the scheme for the comparison of the classification with the reference map (Olofsson et al., 2014).

2.4.1 Reference Data Collection

It is worth noting that accurate ground or reference data must be obtained in order to properly determine the accuracy of the remotely sensed classification. The accuracy of the ground data, on the other hand, is rarely known, and the amount of effort required to collect the necessary data is rarely understood. While no reference data set can be entirely accurate, the reference data must be as accurate as possible; otherwise, the assessment would not be correct. As a result, any accuracy assessment must carefully evaluate the compilation of ground or reference data (Congalton, 2007).

It should also be mentioned that the reference data are commonly referred to as "ground truth" data. While it is true that the reference data are regarded as more accurate than the map being evaluated, this does not mean that they are flawless or

represent "the truth." Consequently, the word "ground truth" is unsuitable and, in some cases, inaccurate (Congalton & Green, 2019).

Field measurements, existing maps or even higher resolution satellite imagery may be used as reference data. Aerial photography is commonly used to assess the accuracy of maps produced with moderate-resolution satellite imagery, such as SPOT and Landsat TM, ground visits are frequently used to assess the accuracy of maps created with high-resolution airborne imagery, and manual image analysis is usually used to evaluate the accuracy of various classification algorithms (Congalton & Green, 2019).

The following factors, or a combination of them can cause errors in the reference data: (1) variations in registration between reference data and remotely sensed map classification, (2) data entry errors, (3) classification scheme errors, (4) variations in land cover between the date of remotely sensed imagery collection and the date of the reference data, and (5) errors in marking reference data (Congalton & Green, 2019).

2.4.2 Spatial Unit

The spatial unit used in accuracy assessment must also be considered. The spatial unit is used to compare map and reference data. In the sampling procedure, the spatial unit may either match the map's resolution or require the aggregation of pixels to pixel blocks (Finegold et al., 2016). Individual pixels, clusters of pixels, or polygons can be suitable sample units, depending on the application. Polygon sampling is the most widely used method to date. To keep the accuracy assessment's spatially explicit character, the user should aim for reference data with the same or higher level of detail (Finegold et al., 2016). It has been referred that, although the pixel is the most common spatial unit, depending on the project's requirements, a grouping of pixels, such as a 3x3 block or a polygon, might be chosen as the sample unit. (Congalton, 2001).

2.5 Analysis

Depending on the detection technique, every project has different accuracy requirements and type of assessment strategy. Accuracy assessment is more than an indication of the map’s accuracy; it also provides sample data that can be employed in order to minimize the bias in pixel counting and to reduce the standard error in the estimated area.

2.5.1 Error Matrix

An error matrix is the most common way to describe the classification accuracy of remotely sensed data (sometimes called a confusion matrix or a contingency table). Many researchers have recommended using an error matrix to define accuracy, and it should be accepted as the standard reporting convention. An error matrix is a square array of numbers laid out in rows and columns that expresses the number of sample units (pixels, clusters of pixels, or polygons) assigned to a particular category in comparison to the actual category as checked on the ground. The reference data (ground truth) is represented by the columns, while the classification produced by remotely sensed data is represented by the rows (Table 1). The number of rows and columns in such matrices is equal to the number of categories whose classification accuracy is being measured (Lillesand et al., 2015).

An error matrix is a very useful way to reflect accuracy since it clearly describes the accuracies of each category as well as the inclusion (commission errors) and exclusion (omission errors) errors present in the classification. When an area is entailed in the wrong category, it is called a commission error. While, when a field is left out of the group to which it corresponds, it is called an omission mistake. Any error on the map is an omission from the right category and a commission to the wrong one (Congalton & Green, 2019).

Table 1. Example of an error matrix.

Satellite Image Classification	**Reference Data (Ground Truth)**			
	A	B	C	Row Totals
A	AA	AB	AC	
B	BA	BB	BC	
C	CA	CB	CC	
Column Totals				Sample Total

A variety of descriptive and analytical statistical methods can be initiated from the error matrix. Overall accuracy, for example, is calculated by dividing the total correct (i.e., the sum of the major diagonal) by the total number of pixels in the error matrix. This value was part of the older, site-specific evaluation and is the most widely mentioned accuracy assessment statistic (Congalton, 1991).

Additionally, individual category accuracies can be calculated in a similar way. Nevertheless, in this situation, there is the option to divide the number of correct pixels in that category by the total number of pixels either in the corresponding row or the corresponding column. The total number of correct pixels in a category is typically divided by the total number of pixels in that category as determined by the reference results (i.e., the column total). Generally, errors of commission arise when a pixel is incorrectly included in a category being evaluated, whereas errors of omission occur when a pixel is left out of the category being evaluated. Since the classification's producer is interested in how accurately a specific area can be categorized, the omission error, often referred to as producer's accuracy, indicates the likelihood of a reference pixel being correctly classified. On the other hand, commission error, also known as user accuracy, indicates the likelihood that a pixel classified on the map/image represents that category on the ground (Story & Congalton, 1986).

The error matrix for the accuracy assessment should be rigorously generated, in order to be considered trustworthy. A key assumption in all of the above analyses is that the error matrix is genuinely representative of the entire classification. If the matrix is improperly generated, then all the analysis is meaningless.

2.5.2 The kappa coefficient

Another widely applied way of measuring a map's accuracy is the Kappa coefficient (Cohen, 1960), which is a gauge of the proportional improvement by the classifier over a purely random sample to classes (Agyemang et al., 2011). It shows the extent to which the correctly classified values of an error matrix are attributed to a "true" versus a "chance" agreement. In other words, it is a means of comparing the observed agreement to an arbitrary expected agreement, if the observer ratings were independent. Additionally, it denotes the proportionate reduction in error caused by a classification process, as opposed to an error caused by a completely arbitrary classification (Cohen, 1960; Munoz & Bangdiwala, 1997).

The leverage of the Kappa analysis technique is that it produces two statistical tests of significance. The first one offers the possibility to test if a given land cover map generated from remotely sensed data is significantly better than if it had been generated by haphazardly assigning labels to regions. The second one compares between any two matrices to check if they are statistically significantly different. In consequence, it is possible to determine if an algorithm differs from another and

conclude which one performs better, based on a chosen accuracy measure (e.g. overall accuracy) (Congalton, 2001). Nonetheless, the validity of the aforementioned conclusions is arguable and numerous articles have questioned the use of the Kappa coefficient analysis (Foody, 2002; Pontius Jr & Millones, 2011; Stehman, 1997).

2.5.3 Binary Classification

Binary classifiers are statistical and computational models that divide an unknown dataset into two segments: positives (P), and negatives (N). The dataset is hence classified to P or N. In order to assess the functionality of a classifier, its prediction performance needs to be evaluated (Saito & Rehmsmeier, 2015). This method was deemed necessary due to a difference in the numbers of positive and negative occurrences. Typically, such an imbalance in classes, with the negatives outnumbering the positives, is naturally evident in various scientific areas with unequal class distributions (Chawla et al., 2002, 2004; Kubat et al., 1998; Rao et al., 2006). For a model's class predictions, the labels {T (True), F (False)} are used to differentiate between the actual class and the predicted one. In this project, the ML Hard Layer is the classification model that was trained in the training phase, to predict the true classes of the ML (P) and nML(N) (Tharwat, 2020). The binary classifier then classifies all data instances as either positive or negative and ultimately generates four types of outcomes:

1. True Positives (TP): is the correctly classified positive sample when the sample is positive, and it is also classified as positive (i.e., ML and the detection methodology classified it as ML).
2. True Negatives (TN): is the correctly classified negative sample when the sample is negative, and it is also classified as negative (i.e., nML and the detection methodology classified it as nML).
3. False positives (FP): is the incorrectly classified positive sample, when the sample is negative, but it is classified as positive (i.e., nML, that the detection methodology classified it as ML, "What it says is ML, is actually nML").
4. False negatives (FN): is the incorrectly classified negative sample, when the sample is positive, but it is classified as negative (i.e., ML, that the detection methodology classified it as nML. "Worst prediction").

The 2x2 error matrix formulated by the above-mentioned outcomes is called a confusion matrix. It is based on a pixel-by-pixel comparison of the thematic map and

the reference points for the accuracy assessment sample, and the class labels allocated by the map and reference data are cross-tabulated. All the basic evaluation measures based on the binary classification are calculated from the confusion matrix. The correct predictions are represented by the green diagonal, while the incorrect predictions by the pink diagonal (Figure 2).

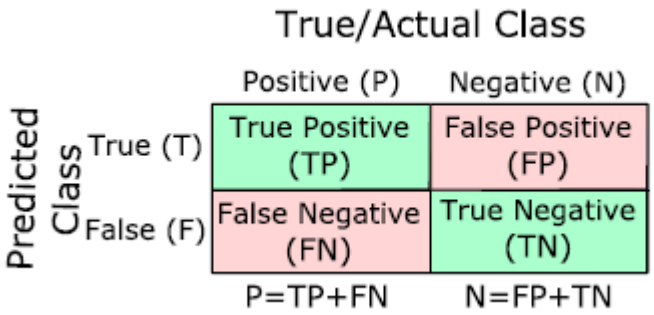


Figure 2. Confusion matrix. The output of the predicted class is either True or False.

The most commonly used performance measures, based on binary classification, are accuracy and error rate (ERR) (He & Garcia, 2009). Sensitivity and specificity are two other popular metrics (Altman & Bland, 1994). Sensitivity is equal to the true positive rate and recall, while specificity is equal to one minus the false positive rate. Precision is another measure and it is equivalent to positive predictive value. Quality Factor, also known as the Jaccard metric or Tanimoto similarity coefficient describes the quality of the positive class. The Matthews correlation coefficient (MCC) (Baldi et al., 2000) and the F1-score (Goutte & Gaussier, 2005) are also useful but are used less frequently.

2.6 Concluding remarks

Despite recent developments, the existing state of accuracy assessment suggests that several issues remain to be solved. Therefore, despite the fact that the topic has progressed significantly, there is still room for improvement. The commonly used methods for accuracy assessment and reporting are often inaccurate, which is a major source of concern. Despite the apparent objectivity of quantitative accuracy measurements, it is important to view accuracy assessment statements with caution. An apparently objective accuracy assertion may be misinterpreted due to a variety of factors (Foody, 2002).

There are several aspects of classification accuracy assessment that must be considered. The first is that the accuracy of any estimation is just as good as the

information used to determine the "true" land cover types present in the test sites. The accuracy assessment process should, as possible, entail an estimation of the errors inherent in the reference data. It is not unusual for image interpretation errors, spatial misregistration, data entry errors, and changes in land cover between the date of the classified image and the date of the reference data to affect the accuracy of the reference data. The second point to be noted is that the accuracy assessment procedure should be structured to reflect how the classification is intended to be used. For example, a single pixel misclassified as "wetland" in the middle of a "corn" area may be insignificant in the development of a regional land use plan, but it may be inappropriate if the classification is used to determine land taxation or implement wetland protection legislation. The third point is that remotely sensed data are usually just a small subset of the many different types of resident data found in a GIS (e.g. it is likely the propagation of errors through the multiple layers of information in a GIS) (Lillesand et al., 2015). Finally, it has been noted that high overall map accuracy is not always representative of the high detection accuracy of individual classes (GFOI, 2013). Therefore, both producer's and user's accuracy need to be computed and taken into consideration for all individual classes.

Summing up, it is important to note the following established principles:

- Examining the map visually is essential, but it is not sufficient. "It appears to be accurate" is not a statement of fact.
- A classification is not complete until it has been evaluated. Only then, can judgments based on that information be considered valid.
- Quantitative accuracy assessment is a powerful tool for evaluating spatial data, both descriptively and analytically.
- Choosing the right reference data is crucial.
- The accuracy assessment is based on the error matrix.
- Any form of accuracy assessment should be reported to the user.
- Suggestions for improving the classification can be made by interpreting accuracy in classes.
- What is true and realistic in a specific area may not be true or feasible in regional or global projects.
- We are unable to foresee all the issues that could occur when dealing with large areas.

- Each project has its own set of accuracy requirements and assessment strategies.
- High overall map accuracy does not always imply high individual class detection accuracy.

3 METHODOLOGY

For the task of the accuracy assessment, two different approaches were applied (see Chapter 3 and Annex I: Initial Approach). In the initial approach, an attempt was made to evaluate the Final Map from ***MAIL*** deliverable 2.3 using Esri's WGS84 Imagery map from ArcGIS online as a reference, on specific testing sites in Germany, Greece, Poland and Spain. Although this method attempts to evaluate the accuracy of the final product of the detection methodology applied in ***MAIL*** deliverable 2.3, the actual assessment proved to be very challenging. The interpretation of marginality and suitability (Marginal Lands, Potential Marginal Lands and Unsuitable Lands), on which the Final Map layer is based, can be subjective, depending on the needs and differences among countries. Furthermore, the whole evaluation would be based on the interpreter's manual assignment of agreement based on satellite imagery, potentially introducing the interpreter's uncertainty and bias.

For these reasons, this approach is not complete and lies in Annex I and a more objective one was developed. The steps applied for the evaluation of the ''ML and nML Classification" layer on the dictated test countries, for which validation data were obtained, are presented in the following flowchart (Figure 3) and are described in the subchapters **Error! Reference source not found.** to 3.3.

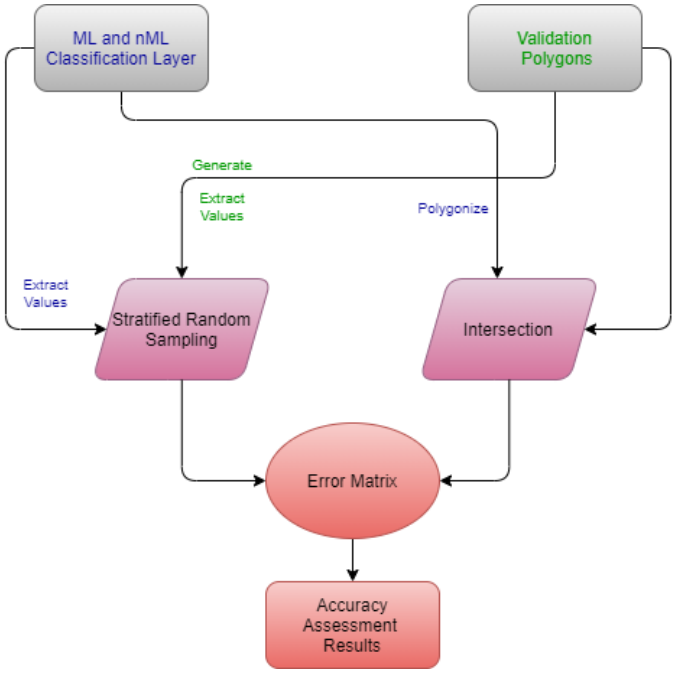


Figure 3. Workflow of the accuracy assessment methodology.

3.1 Datasets And Validation Sites

To properly evaluate the results after the implementation of “hard” thresholds in ***MAIL*** task 2.3, “ground truth” validation data were acquired from the test sites of the four different test countries of Greece, Spain, Germany, and Poland.

The result of the analysis of this “hard” thresholds phase is an intermediate layer, named “**Marginal and Non-marginal land classification**” layer, which included all potential Marginal Lands (ML) and all LULC types that are not fulfilling the ***MAIL*** definition of Marginal Lands ("impervious", "croplands", "forest", "protected areas" "water bodies", "permanent snow", "marshes", "peatbogs" and "changed”) like non-marginal land (nML). For more details refer to ***MAIL*** deliverable 2.3. The two classes are a full representation of the total study area; no region pertains to two classes and no region is excluded from the classification.

ML: areas derived after the implementation of the hard constraints threshold.

nML: all the area that has been excluded in the process of generating the hard layer.

3.1.1 Validation data

To properly evaluate the ML and nML classification layer, experts from each country provided polygons consistent with the ML and nML ***MAIL*** classification definition that would serve as reference data for each country. This was through experts’ prior knowledge of the land use and land cover, landscape and terrain and visual interpretation of ML and nML areas from high-resolution imagery to freely available Google Earth, Google Earth Engine and High-resolution GIS software basemaps, etc. The order of selection for analysis of each test country was based on their climatic condition. Greece and Spain have predominantly Mediterranean climates, while Germany and Poland have predominantly temperate climates. Additionally, in order to gain an overall insight into the performance of the ML and nML classification methodology, the testing sites from the four countries were merged and the same accuracy assessment procedure was followed. The colour code for each test country introduced in Table 2 will be used throughout the document to easily identify each country in this report. Figure 4 represents the ML classification layer for each country.

Table 2. Total provided area of ML and nML test sites for each test country.

Test Site	ML	nML
Greece	7988 ha	5274 ha
Spain	1649 ha	2199 ha
Germany	352 ha	20,913 ha
Poland	539 ha	2463 ha
Merged	10,529 ha	30,849 ha

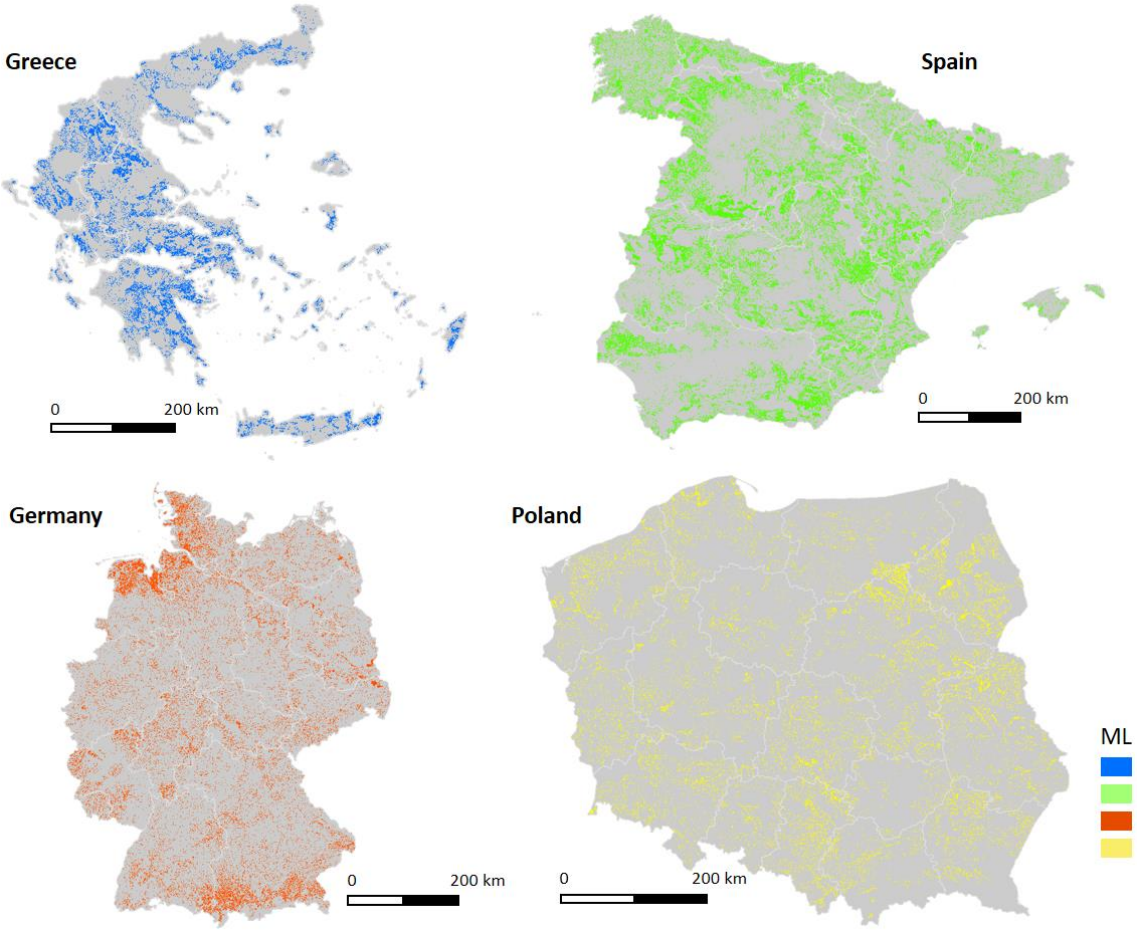


Figure 4. ML classification layer for Greece, Spain, Germany and Poland.

Greece

In Greece, 28 different areas consisting of separate ML and nML polygons were provided for various areas spread across the country, like, Kozani, Serres, Florina, Drama, Kastoria, and Pella, as presented in Figure 5. A total of 7988 ha of ML and 5274 ha of nML areas were provided (Table 2). According to the project partner, these areas were selected based on their land use and specific site knowledge. In many cases in the ML area polygon or the surrounding nML there were successful afforestation areas with conifers (e.g., region of Kozani), which supports their identification as ML by the partner since the state had already considered them as marginal and suitable for afforestation sometime in the past. The majority of these lands are either grasslands or partially scrublands. Moreover, some of these lands are used as pastures but they are neither regularly managed, nor they are calculated in any future afforestation project, therefore they can be considered as marginal and suitable for afforestation.

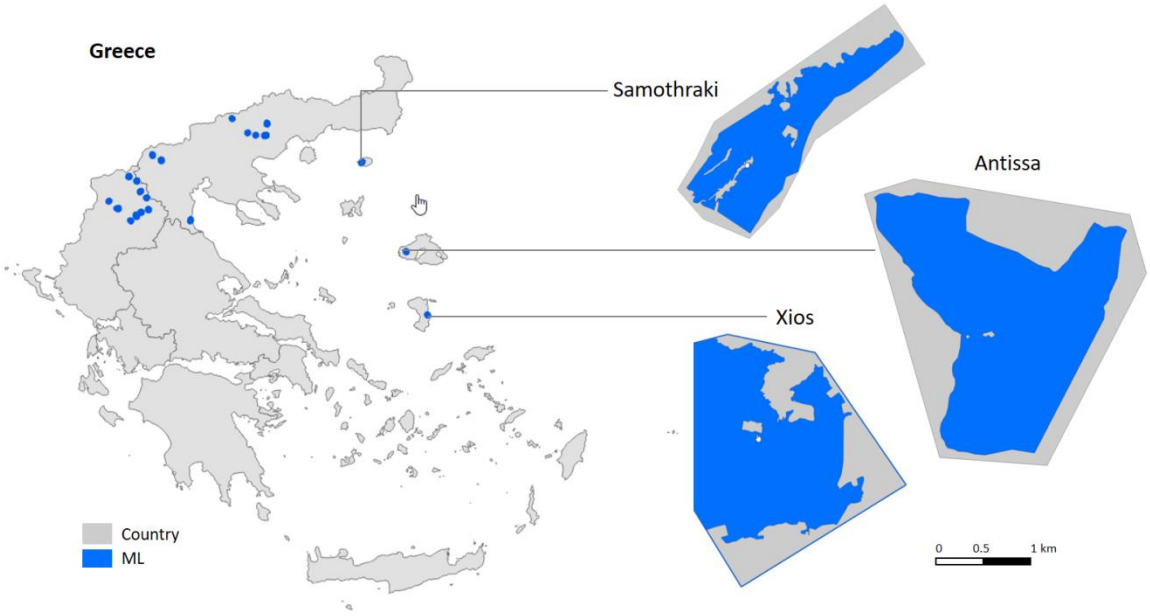


Figure 5. ML and nML validation data locations in Greece. Highlighting three from the provided polygon in Samothraki, Antissa, and Chios Regions.

Spain

Accordingly in Spain, 10 different areas consisting of individual ML and nML polygons were provided, covering various regions of the mainland, like Cuenca, Soria, Segovia, León, Valladolid, Zamora, Burgos, and Ávila (Figure 6), contributing to a total of 1649

ha of ML and 2199 ha of nML areas (Table 2). According to the project partner, the landscape varies among the polygons, since they tried to cover all the regions. These regions are grasslands, scrublands, and bare land based on a regional layer of land cover and land use. Protected areas where afforestation is not recommended by the regional government due to conservation issues were excluded as well as areas where the slope and elevation exceed 35% and 1800 m. Also, from these regions, the arable use layers were removed. One of the principles under which these areas were selected is that these areas are not currently used and there is no economic activity.

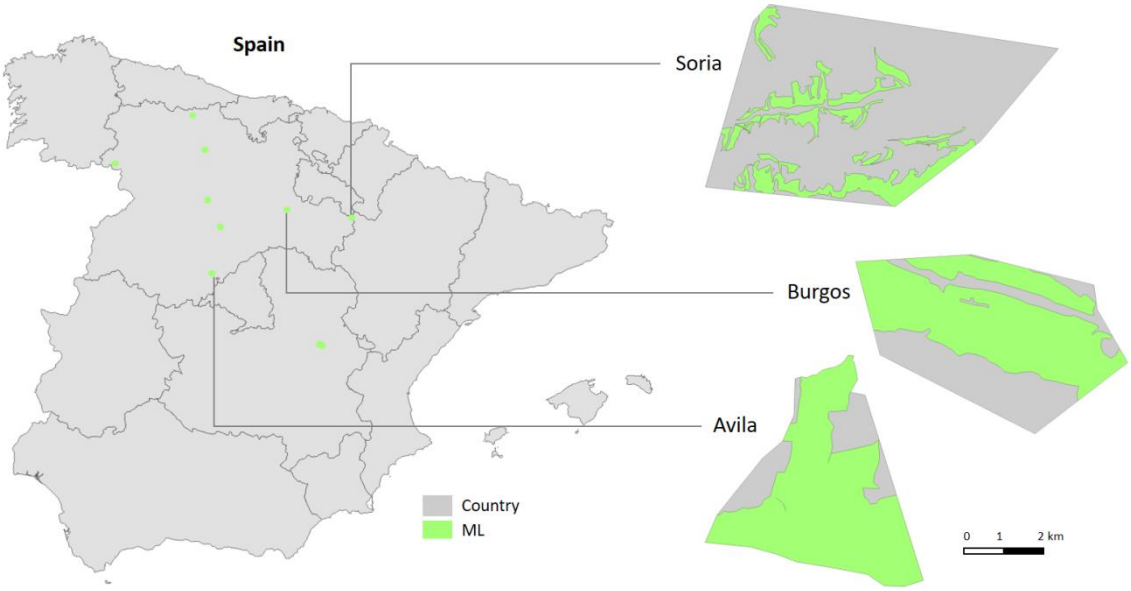


Figure 6. ML and nML validation data locations in Spain. Highlighting three from the provided polygon in Soria, Burgos, and Ávila Regions

Germany

For Germany three different areas each consisting of ML and nML polygons totalling 352 ha of ML and 20,913 ha of nML areas were provided from certain regions like Nochten-Reichwalde, Grafenschau, and Sallgast in the Federal State of Saxony (Figure 7). Obtaining large reference data of ML polygons in Germany proved to be challenging due to the fact that most of the low profit, low fertile, unused or abandoned lands, designated as “Odlands” (wastelands or marginal), reside in areas designated as protected, which is out of the scope of the ***MAIL*** definition.

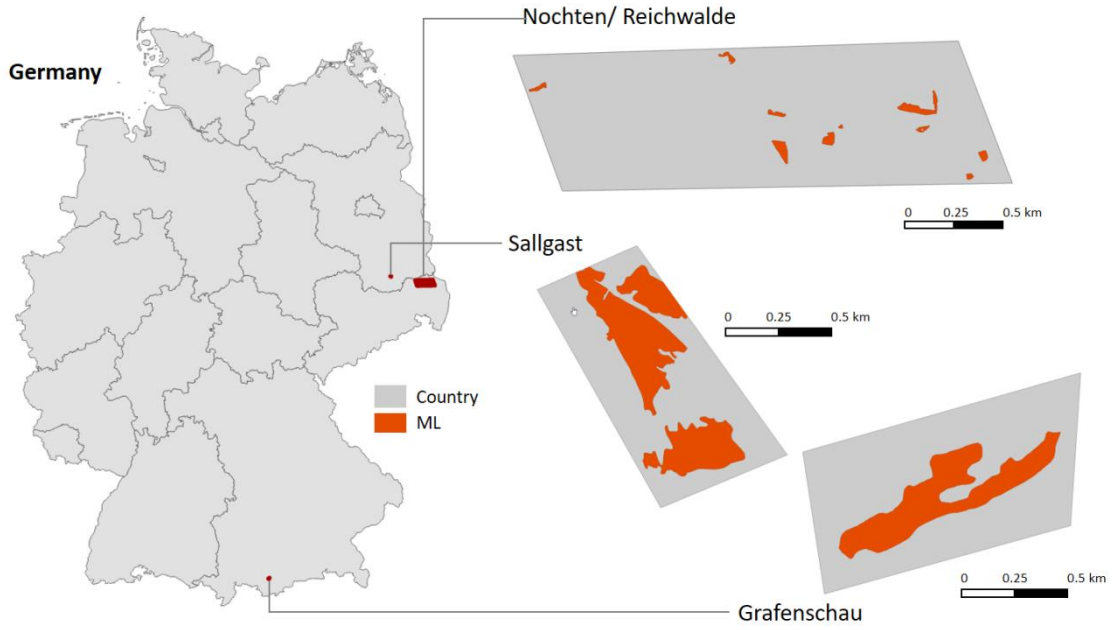


Figure 7. ML and nML validation data locations in Germany. Highlighting the 3 provided polygon in Nochten-Reichwalde, Grafenschau, and Sallgast.

Poland

ML in Poland are mostly caused by the abandonment of agricultural areas (fields). The agriculture fields, however, are typically fragmented, relatively small (usually long and narrow) and in some cases pertain to bigger regions classified as protected. As a result, 12 different areas, consisting of various small, narrow and fragmented ML and nML polygons could be gathered, providing a total area of 539 ha of ML and 2,463 ha of nML in the region of Świętokrzyskie Voivodeship (Figure 8).

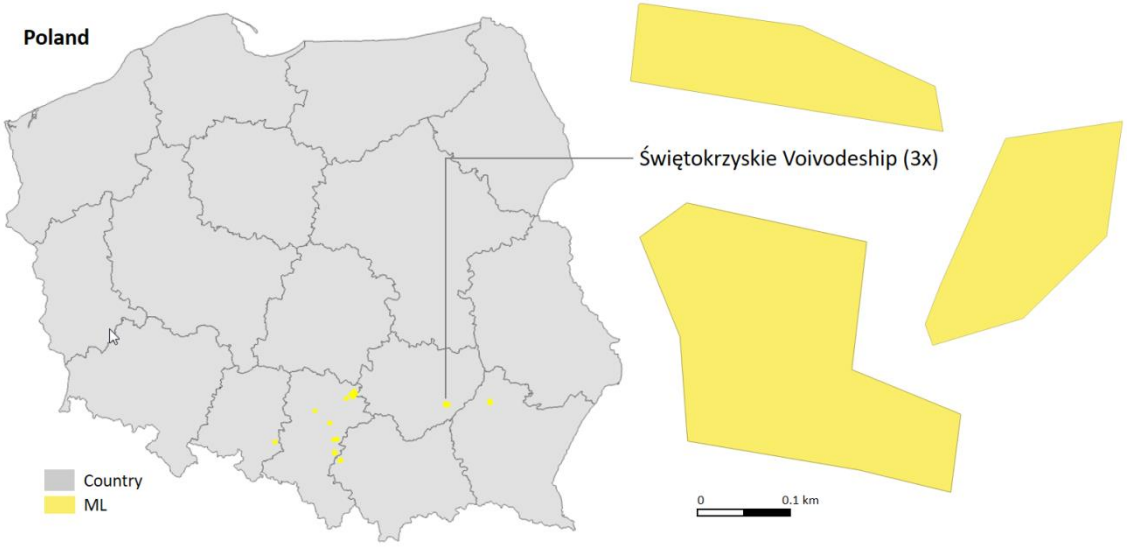


Figure 8. ML and nML validation data locations in Poland. Highlighting some of the provided polygons in the region Świętokrzyskie Voivodeship.

3.1.2 Spatial Unit

When comparing layer and validation data, a spatial unit is used, which, for sampling data in this type of classification, can be a pixel or a pixel block. In the sampling procedure, the spatial unit can either match the resolution of the map or involve the aggregation of pixels to a pixel block. In this task, two different units will be utilized leading to two different approaches that in the end will be compared with each other: the *pixel*, which is the same size as the basemap of the detection methodology, 10 x 10 m, and the *area* of the reference *polygons*, which proved to vary significantly from polygon to polygon due to the difficulties described previously in gathering reference data.

3.1.3 Sample Size

For this project, the testing set comprises of the validation sites that were provided by the ***MAIL*** partners for each country. It was decided that each project partner would provide a minimum area of 1,000 ha meeting the definition of ML and nML as defined in ***MAIL*** deliverable 2.1, for the objectives of the accuracy assessment. The whole area that was provided by each country's experts will be utilized, but since each partner was able to provide areas of different extent for ML and nML validation data, the sample size for the accuracy assessment in each country varies as well.

3.2 Analysis

3.2.1 Point-based assessment

To comply with the international standard organization (ISO) 2859 Series and 3952-1 2005 "Guideline of defining sample size and devising sampling Methods" and to have a full representation of the total provided validation data, a **stratified random sampling design**, was chosen as the most suitable sampling scheme. The two validation data classes were used as strata for determining the sample size, conforming with the equal probability sampling designs, simple random, stratified random and systematic designs, in which the validation data strata also correspond to the layer classes.

Here, the usual Binomial Probability Theory of determining the number of samples is not going to be used but an arbitrary number of one sample point per hectare was deemed suitable. This was chosen because the number of sample points had to be large enough to produce sufficiently precise estimates of the ML and nML classification layer accuracy, but not too large to raise the probability of spatial autocorrelation issues or over-estimation of the significance of any non-zero difference (Foody, 2009). Thus, the number of points allocated to each ML and nML class is proportional to the total size in hectares of validation data provided, as shown in Table 3.

Table 3. Sample size and allocation of sample points for each test country.

	ML		nML	
	Area [ha]	Allocated Points	Area [ha]	Allocated Points
Greece	7988	7988	5274	5274
Spain	1649	1648	2199	2199
Germany	352	352	20,913	20,913
Poland	539	539	2463	2463
Merged	10,529	10,529	30,849	30,849

The stratified random sampling technique was applied using the ArcGIS Pro, tool "create random points". This way, the predefined number of sample points for each

class was distributed over the whole ML and nML classes polygons, representing fully the testing sites.

A subset of the ML and nML classification layer was created based on the provided validation data for each country using the “Extract by Mask” ArcGIS Pro tool. The cell values of the classification layer were then extracted based on the previously generated set of point features (sample points) using the ‘’Extract values to points’’ tool and their values were recorded as a new field in the attribute table. Finally, the attribute table of the extracted points for each test country was exported for further analysis. An example of the sample points is shown in Figure 9.

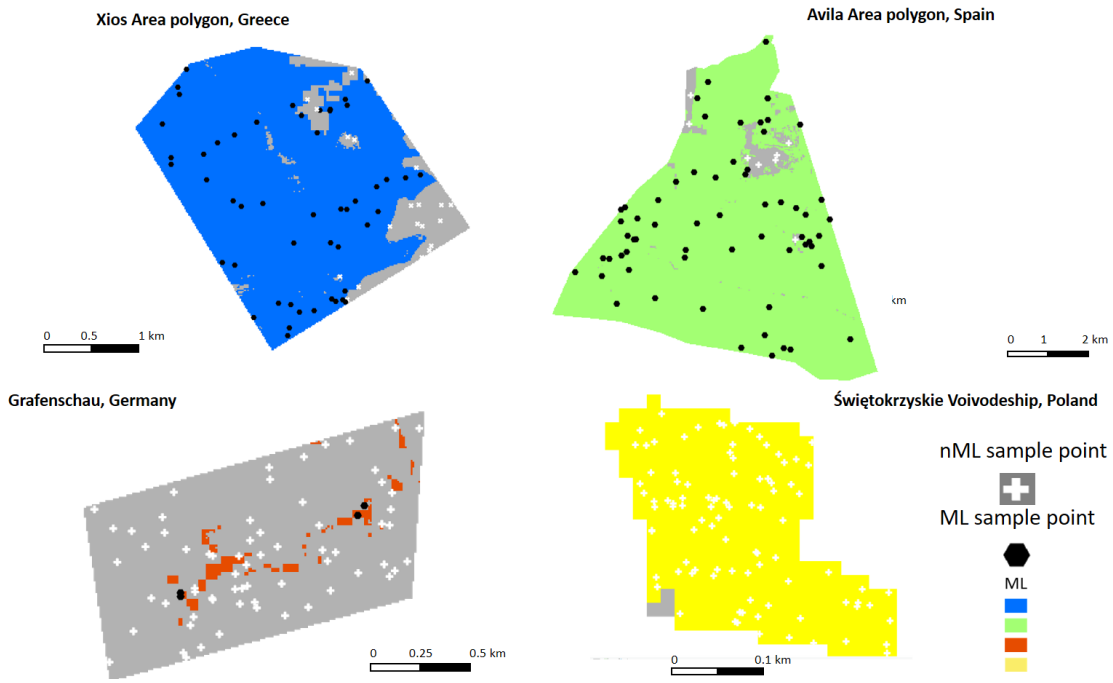


Figure 9. ML and nML land classification layer for selected polygons in Greece, Spain, Germany and Poland showing extracted values to sample points.

3.2.2 Area-based assessment

In a slightly different approach, instead of creating random points in the test sites in order to assess the accuracy of the classification map, the whole area of the validation sites will be incorporated in the accuracy assessment. In other words, instead of comparing specific pixels, we will compare geometries.

In this case, the ML and nML classification map was converted from raster to polygon and these polygons were then overlaid on top of the validation polygons. By applying a

simple “intersect” algorithm on these two layers the geometries of agreement and disagreement become evident and the features or the portions of the features which overlap in all layers and/or feature classes will be written to the output feature class as unique features. The outcome of this procedure is a shapefile with four features, corresponding to the four agreement/disagreement scenarios between the classified map and the reference polygons, and the respective area of each feature is readily computed using the “Add Geometry Attributes” ArcGIS tool.

This procedure was applied to each test country separately and for all of them combined as well.

3.3 Evaluation Metrics

An error matrix with four outcomes based on the comparison of the classification layer and the validation sample was generated for the accuracy assessment of the detection methodology applied in the ***MAIL*** task 2.3 for each test country. In the main diagonal of the error matrix reside the correct classification while in the off-diagonal the omission and commission errors are nested (Congalton, 1991). The error matrix is the basis of all the evaluation measures of accuracy obtained by the point-based and the area-based assessment methods.

The different measures calculated are listed and explained as follows:

Overall Accuracy (OA) is the ratio between the correctly classified samples to the total number of Samples. It essentially tells us what percentage of the reference sites were correctly mapped out of all of them. The overall accuracy is commonly given as a percentage, with 100 percent accuracy indicating that all reference sites were correctly categorized (Congalton, 1991):

$$OA = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FN} + \mathrm{FP}} * 100$$

User's accuracy (UA) is the proportion of the area mapped as a particular category that is actually that category “on the ground” (Congalton, 1991). The user’s accuracy can therefore be considered as a measure of the reliability of the map. If a user employs the final map in order to locate a particular spatial unit, the user's accuracy gives the conditional probability of that map location actually representing the mapped unit. It is calculated by dividing the correct classified pixels in a class by the total number of pixels that were classified in that class (row total) and multiplying by 100 (Banko, 1998).

The probability of *commission error* is the complementary measure to user's accuracy, it represents the features of a category on the map that were misclassified and for each class, it is calculated as (Finegold et al., 2016):

$$\text{error of commission } [\%] = 100\% - \text{User's Accuracy } [\%]$$

Producer's accuracy (PA) is the proportion of the area that is a particular category on the ground and it is also mapped as that category (Congalton, 1991). The producer's accuracy measures how well a given area is classified and provides the producer of the final classification map with the conditional probability of a particular location of spatial unit appearing as that on the map. It is computed by dividing the number of correct pixels in one class by the total number of reference pixels for this class (column total) and multiplying by 100 (Banko, 1998).

Producer's accuracy is the complement of the probability of *omission error*, which represents the proportion of actual features on the ground that are omitted from the classification map (Finegold et al., 2016).

$$\text{error of omission}[\%] = 100\% - \text{Producer's Accuracy } [\%]$$

From the binary classification perspective, two similar with UA and PA measures of accuracy are *sensitivity or recall* (also referred as true positive rate or hit rate) and *precision* (also called Positive Predictive Value).

Sensitivity is calculated as all the positive correctly classified samples divided by the total number of positive samples and can be interpreted as the proportion of positive samples that were correctly classified (Sokolova et al., 2006).

$$Sensitivity\ or\ Recall = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Precision denotes the proportion of positive samples that were correctly classified to the total number of positive predicted samples (Sokolova et al., 2006). How "precise" is the model when predicting a certain class.

$$Precision = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Since our investigation is targeted at the detection of ML, the ML can be considered as our positive sample and the nML as the negative sample. Consequently, sensitivity equals UA and precision equals PA for the ML samples.

Error rate (ERR) or misclassification rate is the number of misclassified samples from both ML and nML classes (Bradley, 1997), which is how often was detection rate incorrect.

$$ERR = \frac{\mathrm{FP} + \mathrm{FN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FN} + \mathrm{FP}}$$

F1-score is also known as F-measure called, and denotes the harmonic mean of precision and recall (Sokolova et al., 2006). The value of the F1-score ranges from zero to one, and high values of the F1-score show high classification performance.

$$F1\text{-}score = 2 * \frac{\mathrm{PREC} * \mathrm{REC}}{\mathrm{PREC} + \mathrm{REC}}$$

Matthew's correlation coefficient (MCC) denotes the correlation between the observed and predicted classifications, and it is calculated directly from the error matrix. A coefficient of +1 shows a perfect prediction, and −1 denotes total disagreement between prediction and true validation values and zero signifies that is not better than a random prediction (Matthews, 1975).

$$MCC = \frac{\mathrm{TP} * \mathrm{TN} - \mathrm{FP} * \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})\,(\mathrm{TP} + \mathrm{FN})\,(\mathrm{TN} + \mathrm{FP})\,(\mathrm{TN} + \mathrm{FN})}}$$

Kappa is a measure of agreement between the predictions and the actual class. It can also be a comparison of the overall accuracy to the expected random chance accuracy. Because of class imbalance and having just 2 classes and to intelligently nullify the dominance of one class while evaluating the layer and validation sample data it is necessary to derive the kappa in this task. The Kappa coefficient is computed from the following equations (Jensen, 1996; Sim & Wright, 2005):

$$Kappa = \frac{\mathrm{accuracy} - \mathrm{expAccuracy}}{1 - \mathrm{expAccuracy}}$$

$$k = \frac{N \sum_{i=1}^{r} x_{ii} - \sum_{i=1}^{r} (x_{i+} * x_{+i})}{N^2 - \sum_{i=1}^{r} (x_{i+} * x_{+i})}$$

Where:

r = number of rows or columns in the error matrix

N = total number of observations in error matrix

X_{ii} = major diagonal element for class i

x_{i+} = total number of observations in row i (right margin)

x_{+i} = total number of observations in column i (bottom margin)

The statistical significance of any given classification matrix can also be determined by utilizing the Kappa coefficient as a basis. According to Cohen (1960), Kappa can be considered as the chance-corrected proportional agreement and takes values from +1 (perfect agreement) to -1 (complete disagreement). Using these values as references, Munoz & Bangdiwala (1997) and Viera & Garrett (2005) developed some guidelines for interpreting the Kappa, by quoting Landis & Koch (1977) as shown in Table 4.

Table 4. Kappa interpretation guidelines of Landis & Koch (1977).

Kappa statistic	Strength of Agreement
< 0	Poor
0.01 – 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81 – 1.00	Almost Perfect

3.4 Land Cover Analysis

After the confusion matrix is tabulated by the agreement/disagreement samples and the various accuracy assessment measures have been computed, a final analysis on the land cover types of the ML areas took place. For this extra step, we exploited the Sentinel-2 Global Land Cover (S2GLC) product. The S2GLC project was founded by the European Space Agency (ESA) and was implemented by a consortium led by CBK PAN. The S2GLC 2017 product represents a high-resolution land cover classification map for most of the European continent. The classification is based on multi-temporal Sentinel-2 imagery obtained through 2017 and the dataset is delivered in the native

Sentinel 2 spatial resolution of 10 m distinguishing among 13 land cover classes (Malinowski et al., 2020).

The TP, FP, and FN sample points of each country, corresponding to correctly, misclassified and omitted ML areas, were then overlaid on the S2GLC map. Following the same principle, the validation ML polygons that were digitized by the respective experts from each country were overlaid with the S2GLC map and the land cover types were extracted. The integration of the S2GLC map into this assessment will allow us to conduct a brief analysis and gain an insight into the land cover types that are prevalent or associated with MLs, but also to identify potential land cover types that are problematic in being identified as ML by the detection methodology.

4 RESULTS

The results of the accuracy assessment of the “ML and nML classification” layer from ***MAIL*** deliverable 2.3 conducted on each of the test countries and the merged layer under investigation will be presented in this chapter. For each test site, the generated measures of accuracy derived both from the point-based and the area-based approaches will be included, which ultimately will provide an insight on how well the detection methodology was able to detect MLs.

4.1 Greece

The general point-based and area-based results and error matrices for Greece are displayed in Table 5 and Table 6. Table 5 shows the error matrix for Greece with four outcomes based on a pixel-by-pixel comparison of the classification layer and validation samples for the two classes of ML and nML. From the 7,988 samples for ML and 5,274 for nML, 5,902 points were correctly classified for the ML and 3,583 for the nML. Table 6 displays the area-based error matrix for Greece which shows that 5,877 ha of ML and 3,504 ha of nML are classified correctly and 1,769 ha of ML and 2,110 ha of nML are classified incorrectly.

Table 5. Point-based error matrix (Greece).

True/Actual Class (Reference)				
Predicted Class (Classified)		**ML**	**nML**	**Total**
	ML	5,902	1,691	7,593
	nML	2,086	3,583	5,669
	Total	7,988	5,274	

Table 6. Area-based error matrix (Greece).

True/Actual Class (Reference)				
Predicted Class (Classified)		**ML**	**nML**	**Total**
	ML	5,877	1,769	7,646
	nML	2,110	3,504	5,614
	Total	7,987	5,273	

From the correctly classified and incorrectly classified samples, the TP, TN, FP, and FN samples have been derived and the different accuracy metrics such as Producer's Accuracy, User's Accuracy, F1-score (Table 7), Overall Accuracy, Error Rate, Kappa and Matthew's correlation coefficient (Table 8) have been computed.

Table 7. Class statistics (Greece).

Assessment	Class	PA (%)	UA (%)	F1-score (%)
Point-based	ML	73.89	77.73	75.76
	nML	67.94	63.20	65.49
Area-based	ML	73.58	76.86	75.19
	nML	66.45	62.42	64.37

Analysing the results from the error matrix of point-based assessment, the values for UA, PA and F1-score obtained for the ML class are 77.73%, 73.89% and 75.76%, respectively, indicating that the percentage of ML class correctly detected to the total number of reference ML areas was good and the detection methodology was relatively precise in detecting ML class. On the other hand, the results of the area-based error matrix show a slight decrease compared with the point-based assessment. UA, PA and F1-score for ML are 76.86%, 73.58% and 75.19%.

Table 8. Overall statistics (Greece).

Assessment	OA (%)	ERR (%)	Kappa	MCC
Point-based	71.52	28.48	0.41	0.41
Area-based	70.75	29.25	0.40	0.40

Table 8 shows that overall accuracy of 71.52% and 70.75% was achieved by the detection methodology from the point-based and area-based methods with an overall error rate of 28.48% for the first and 29.25% for the second. The MCC of 0.41 and 0.40 and kappa of 0.41 and 0.40 of the above tables show a positive correlation and moderate agreement between the reference and classified classes.

4.2 Spain

Table 9 and Table 10 show the error matrix for Spain with four outcomes based on a point-based and area-based comparison of the classification layer and the reference "ground truth" data for the two classes of ML and nML.

Table 9. Point-based error matrix (Spain).

True/Actual Class (Reference)				
Predicted Class (Classified)		**ML**	**nML**	**Total**
	ML	1,396	406	1,802
	nML	253	1,793	2,046
	Total	1,649	2,199	

Table 10. Area-based error matrix (Spain).

True/Actual Class (Reference)				
Predicted Class (Classified)		**ML**	**nML**	**Total**
	ML	1,409	398	1,807
	nML	240	1,800	2,040
	Total	1,649	2,198	

From the 1,649 random samples for ML and 2,199 for nML, 1,396 points were correctly classified for the ML and 1,793 for the nML. For the area-based assessment 1,409 ha of ML and 1,800 ha of nML are correctly classified. As a result, the total number of correctly classified samples for ML and nML for point-based assessment is 3,189 and for area-based assessment is 3,209 ha (Table 9 and Table 10).

From the correctly and incorrectly classified samples, of point-based and area-based assessments, the values of TP, TN, FP, and FN samples have been derived and the values of the various accuracy metrics were calculated (Table 11). Analysing the results from the error matrix, the values for precision (UA), recall (PA) and F1-score obtained for the class of ML, are 77.47%, 84.66% and 80.90%, respectively for the point-based and 77.98%, 85.45% and 81.54% for the area-based approach, confirming that in both cases the total sample of reference ML samples was sufficient, and the detection methodology was quite precise in detecting the ML class.

Table 11. Class statistics (Spain).

Assessment	Class	PA (%)	UA (%)	F1-score (%)
Point-based	ML	84.66	77.47	80.90
	nML	81.54	87.63	84.48
Area-based	ML	85.45	77.98	81.54
	nML	81.89	88.24	84.95

Table 12 shows that the overall accuracy of the whole detection methodology with the provided validation data was 82.87% for the point-based method and 83.42% for the area-based method, with an overall error rate of 17.13% and 16.58% respectively, which are rather positive results considering the high accuracy and low error rate. The MCC of 0.66 in the first case and 0.67 for the second and kappa of 0.65 and 0.67 in the tables above show a relatively positive correlation and moderate agreement between the validation and predicted classes.

Table 12. Overall statistics (Spain).

Assessment	OA (%)	ERR (%)	Kappa	MCC
Point-based	82.87	17.13	0.65	0.66
Area-based	83.42	16.58	0.67	0.67

4.3 Germany

Table 13 and

Table 14 show the error matrix for Germany, with four outcomes based on a point-based and an area-based comparison of the classification layer and the validation data for the two classes of ML and nML.

Table 13. Point-based error matrix (Germany).

True/Actual Class (Reference)				
Predicted Class (Classified)		**ML**	**nML**	**Total**
	ML	317	8,436	8,753
	nML	35	12,477	12,512
	Total	352	20,913	

Table 14. Area-based error matrix (Germany).

True/Actual Class (Reference)				
Predicted Class (Classified)		**ML**	**nML**	**Total**
	ML	311	8,509	8,820
	nML	40	12,403	12,443
	Total	351	20,912	

From the 352 random samples for ML and 20,913 for nML, 317 samples were correctly classified for the ML and 12,477 for the nML. Also, quite close to these results the area-based error matrix shows that from 351 ha of ML and 20,912 ha of nML, the correctly classified hectares were 311 for ML and 12,403 for nML. The evident difference in ML and nML class is due to the provided validation data in subchapter 3.1.1. As a result, the total number of correctly classified samples for ML and nML were 12,794 for point-based assessment and 12,714 ha for area-based assessment.

Table 15. Class Statistics (Germany).

Assessment	Class	PA (%)	UA (%)	F1-score (%)
Point-based	ML	90.06	3.62	6.96
	nML	59.66	99.72	74.66
Area-based	ML	88.60	3.53	6.78
	nML	59.31	99.68	74.37

From the correctly classified and incorrectly classified samples, the values of TP, TN, FP, and FN samples were derived and the values of the different calculated accuracy metrics from the subchapter 3.3 are displayed below (Table 15 and Table 16).

Analysing the results of the point-based error matrix, the values for UA, PA and F1-score obtained in the class ML, are 3.62%, 90.06% and 6.96%, and that of the nML of 99.72%, 59.66%, and 74.66% respectively. And for area-based assessment UA is 3.53% for ML and 99.68% for nML, PA is 88.6% for ML and 59.31% for nML, and F1-score is 6.78% for ML and 74.37% for nML. The very low precision, F1-score of the ML class for Germany might be as a result of an error in the validation data or class imbalance as the nML class validation data were larger than the ML class, this can also be seen from the very low kappa of 0.04, in both assessments, low agreement between

the predicted and the actual class. It was not able to completely nullify the influence of the large nML class, the MCC further proofs that by showing a very low correlation between the predicted and the actual class.

Table 16. Overall statistics (Germany).

Assessment	OA (%)	ERR (%)	Kappa	MCC
Point-based	60.61	39.84	0.04	0.13
Area-based	59.79	40.21	0.04	0.13

Table 16 shows that the overall accuracy of the whole detection methodology with the provided validation data for Germany was 60.61% for the point-based assessment method and 59.79% for the area-based assessment method.

4.4 Poland

Table 17 and Table 18 show the error matrices for Poland with four outcomes based on the point-based and area-based comparison of the classification layer and the Validation data for the two classes of ML and nML. From the 539 random samples for ML and 2,463 for nML, 292 points were correctly classified for the ML and 2,439 for the nML. From the 538 ha of ML and 2,461 ha of nML, the area which is correctly classified for ML covers 284 ha and for nML 2,432 ha.

Table 17. Point-based error matrix (Poland).

True/Actual Class (Reference)				
Predicted Class (Classified)		**ML**	**nML**	**Total**
	ML	292	24	316
	nML	247	2,439	2,686
	Total	539	2,463	

Table 18. Area-based error matrix (Poland).

True/Actual Class (Reference)				
Predicted Class (Classified)		**ML**	**nML**	**Total**
	ML	284	29	313
	nML	254	2,432	2,686
	Total	538	2,461	

The total number of correctly classified point samples for ML and nML is 2,739 and of incorrectly classified points is 271, while for the area-based assessment 2,722 ha were correctly and 283 ha incorrectly classified.

From the correctly classified and the misclassified samples for Poland the values of TP, TN, FP, and FN samples and the values of the accuracy metrics explained in 3.3 were derived respectively in Table 19 and Table 20. Analysing the results from the error matrix of point-based assessment, the values for UA, PA and F1-score obtained in the class ML, are 92.41%, 54.17%, and 68.3%, and that of the nML of 90.8%, 99.03%, and 94.74% respectively. The analysis of the results of area-based assessment for the same metrics are 90.74%, 52.79% and 66.75% for ML and for nML are 90.54%, 98.82% and 94.5%. Indicating that the percentage of ML class correctly detected to the total sample of ML validation samples was adequate and the detection methodology was very precise in detecting ML class in both assessments.

Table 19. Class statistics (Poland).

Assessment	Class	PA (%)	UA (%)	F1-score (%)
Point-based	ML	54.17	92.41	68.30
	nML	99.03	90.80	94.74
Area-based	ML	52.79	90.74	66.75
	nML	98.82	90.54	94.50

Table 20 shows that the overall accuracy of the whole detection methodology with the provided validation data was 90.97% for the point-based assessment and 90.56% for the area-based assessment with an overall error rate of 9.03% and 9.44% respectively, which are positive results considering the high accuracy and the low error rate. The MCC of 0.67 and 0.65 and kappa of 0.64 and 0.62 show a relatively positive correlation and agreement between the reference and predicted classes.

Table 20.Overall statistics (Poland).

Assessment	OA (%)	ERR (%)	Kappa	MCC
Point-based	90.97	9.03	0.64	0.67
Area-based	90.56	9.44	0.62	0.65

4.5 Merged

Merging the data of Greece, Spain, Germany and Poland, two fusion error matrices were tabulated based on the point- and area-based assessment results, comparing the classification layer and validation samples for the two classes of ML and nML, and are displayed below (Table 21 and Table 22). From the 10,529 stratified random samples for the ML and the 30,849 for the nML, 7935 points were correctly classified as ML and 20,195 as nML. Overlaying the reference polygons on the prediction map, out of the 10,528 ha of ML, 7,882 ha were classified correctly and out of the 30,848 ha of nML, 20,141 ha were correct.

Table 21. Point-based error matrix (Merged).

True/Actual Class (Reference)				
Predicted Class (Classified)		**ML**	**nML**	**Total**
	ML	7,935	10,654	18,589
	nML	2,594	20,195	22,789
	Total	10,529	30,849	

Table 22. Area-based error matrix (Merged).

True/Actual Class (Reference)				
Predicted Class (Classified)		**ML**	**nML**	**Total**
	ML	7,882	10,707	18,589
	nML	2,646	20,141	22,787
	Total	10,528	30,848	

From the correctly classified and incorrectly classified samples, of both approaches, the values of TP, TN, FP, and FN samples were derived, and the computed measures of accuracy are displayed in Table 23. Analysing the results from the error matrix, the values of UA, PA and F1-score that was achieved for the class of ML, are 42.69%, 75.36% and 54.5%, respectively based on the points evaluation and 42.4%, 74.87% and 54.14% based on the area evaluation.

Table 23. Class statistics (Merged).

Assessment	Class	PA (%)	UA (%)	F1-score (%)
Point-based	ML	75.36	42.69	54.50
	nML	65.46	88.62	75.30
Area-based	ML	74.87	42.40	54.14
	nML	65.29	88.39	75.10

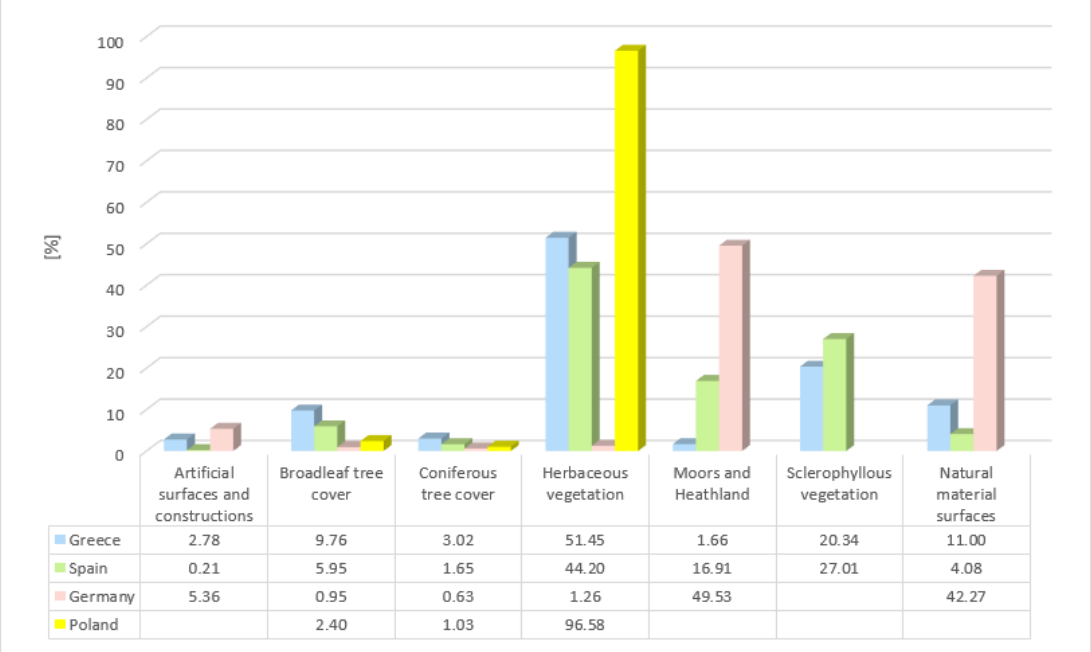
The Overall Accuracy for the point-based assessment applied on the merged sites is 67.98%, which is slightly higher than area-based assessment which is 67.73% (Table 24). The Error rate of the first assessment is 32.02% while in the second one is 32.27%. MCC and Kappa in both assessments seem to be influenced by German's results, having a rate of 0.33 and 0.32 for kappa and 0.36 and 0.35 for MCC.

Table 24.Overall statistics (Merged).

Assessment	OA (%)	ERR (%)	Kappa	MCC
Point-based	67.98	32.02	0.33	0.36
Area-based	67.73	32.27	0.32	0.35

4.6 Land Cover Analysis

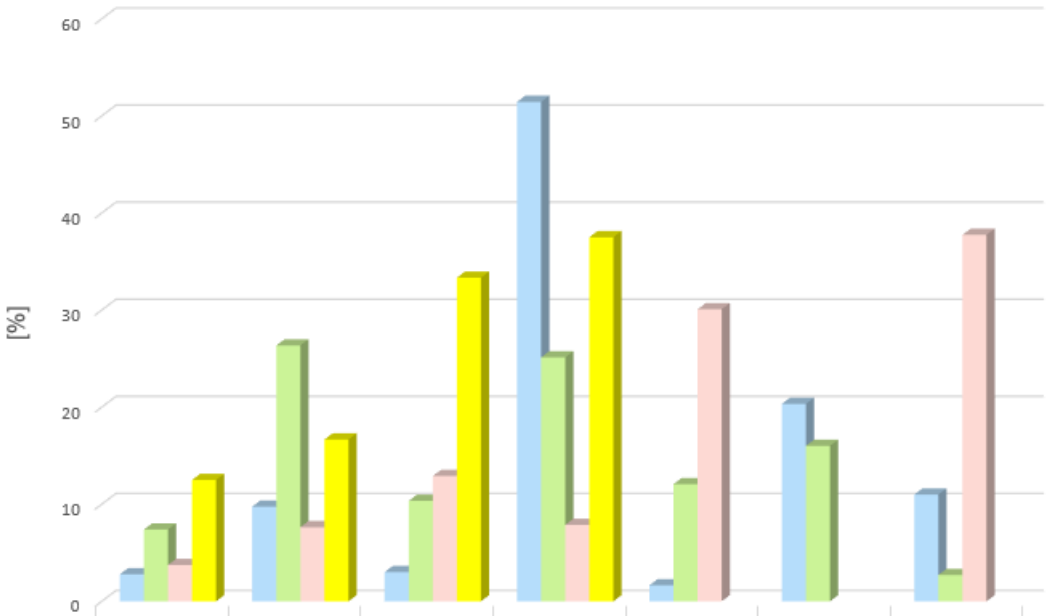
Initially, the class statistics of the ML areas derived from the confusion matrix (TP, FP, FN) were overlaid with the S2GLC product map. This way, we can identify which types of land cover were correctly classified as ML, which were mistakenly classified as ML, and which even though they were validated as ML from the project partners, the detection methodology failed to classify them as marginal.



	Artificial surfaces and constructions	Broadleaf tree cover	Coniferous tree cover	Herbaceous vegetation	Moors and Heathland	Sclerophyllous vegetation	Natural material surfaces
Greece	2.78	9.76	3.02	51.45	1.66	20.34	11.00
Spain	0.21	5.95	1.65	44.20	16.91	27.01	4.08
Germany	5.36	0.95	0.63	1.26	49.53		42.27
Poland		2.40	1.03	96.58			

Figure 10.Distribution of TP samples from error matrix over S2GLC (in %).

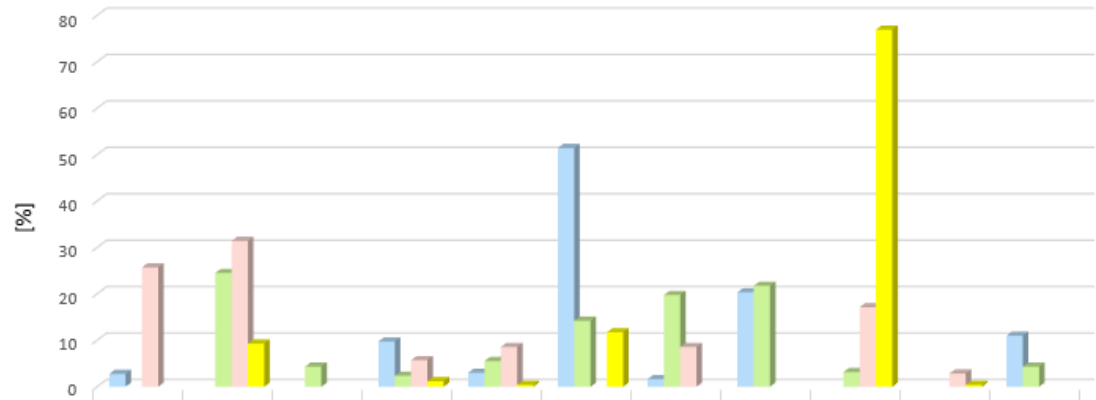
The results of the analysis of ML error matrix with S2GLC in Figure 10 show that the majority of TP samples, which are the correctly classified samples of ML based on the provided validation data, are associated with herbaceous vegetation land cover for Greece (51.45%), Spain (44.20%), and for Poland (96.58%). For Germany, the majority of the TP points are on the land cover classes of moors and heathland (49.53%) and on natural material surface (42.27%).



	Artificial surfaces and constructions	Broadleaf tree cover	Coniferous tree cover	Herbaceous vegetation	Moors and Heathland	Sclerophyllous vegetation	Natural material surfaces
Greece	2.779	9.758	3.016	51.449	1.656	20.343	10.999
Spain	7.389	26.355	10.345	25.123	12.069	16.010	2.709
Germany	3.746	7.634	12.909	7.871	30.085		37.755
Poland	12.500	16.667	33.333	37.500			

Figure 11. Distribution of FP samples from error matrix over S2GLC (in %).

The FP (Figure 11. Distribution of FP samples from error matrix over S2GLC (in %).), which are the nML samples incorrectly classified as ML, are also within the herbaceous land cover class for Greece (51.45%) and Poland (37.5%), however for Poland a significant percent is also within the coniferous tree land cover (33.3%). For Spain, the FP samples reside on the broadleaf tree (26.35%) and herbaceous vegetation (25.12%) land cover, while for Germany they are found on other land cover types like natural material surfaces (37.75%) and moors and heathland (30.08%), which are quite consistent with the TP samples.



	Artificial surfaces and constructions	Cultivated areas	Vineyards	Broadleaf tree cover	Coniferous tree cover	Herbaceous vegetation	Moors and Heathland	Sclerophyllous vegetation	Marshes	Peatbogs	Natural material surfaces
Greece	2.78			9.76	3.02	51.45	1.66	20.34			11.00
Spain		24.51	4.35	2.37	5.53	14.23	19.76	21.74	3.16		4.35
Germany	25.71	31.43		5.71	8.57		8.57		17.14	2.86	
Poland		9.31		1.21	0.40	11.74			76.92	0.40	

Figure 12. Distribution of FN samples from error matrix over S2GLC (in %).

The FN, which are ML samples incorrectly classified as nML, seem to be distributed across all land cover classes with Greece noting a pick in the herbaceous land cover class (51.45%) and Poland exhibiting the majority of the FN samples on the marshes land cover type (76.92%), as shown in Figure 12. This misclassification might be attributed to various reasons, as discussed in chapter 2 “Literature review”, like an inherent error in the input data of the detection methodology. Another potential source of error could be the difference in the dates between the reference data collection and the generation of the S2GLC product, which was produced in 2017 and is the basemap that produced the ML and nML classification layer.

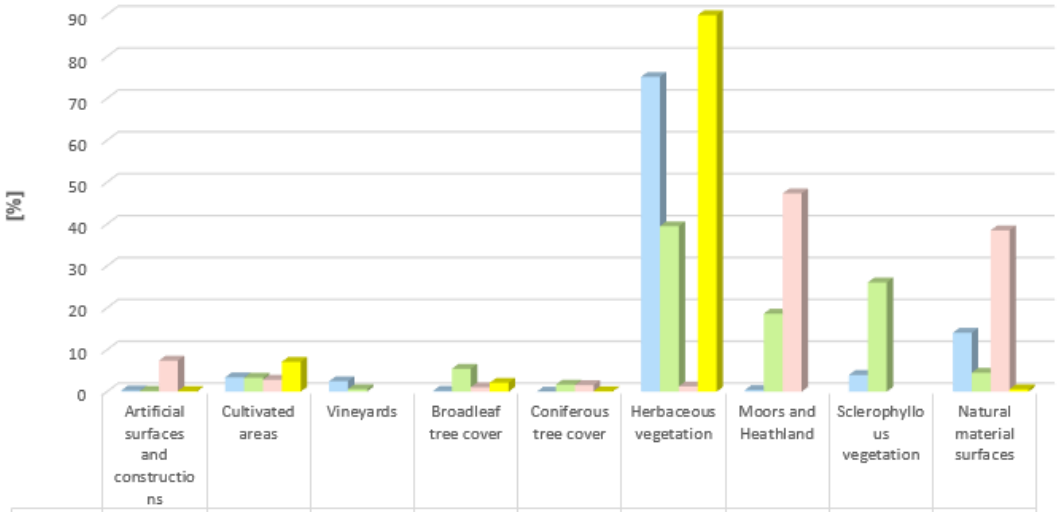
Further exploiting the S2GLC product and the validation data that were offered from the respective project partners, the S2GLC map was intersected with the ML validation polygons and the land cover types of the ML polygons were extracted. The total area of each land cover is presented in the following table in ha, grouped by country (

Table 25).

Table 25. Land cover class sizes of provided ML validation data areas over the S2GLC.

Land Cover Class sizes of ML area over S2GLC (in ha)					
Code	**Class Name**	**Greece**	**Spain**	**Germany**	**Poland**
62	Artificial surfaces and constructions	21.89	3.09	25.42	0.53
73	Cultivated areas	269.3	54.56	9.75	24.91
75	Vineyards	196.71	9.84		
82	Broadleaf tree cover	14.47	90.19	3.66	7.41
83	Coniferous tree cover	2.6	27.78	5.36	0.45
102	Herbaceous vegetation	5,911.83	649.53	4.33	313.56
103	Moors and Heathland	31.34	306.42	162.28	
104	Sclerophyllous vegetation	314.41	428.97		
121	Natural material surfaces	1,105.87	74.39	132.05	2.01
	Total	**7,868.42**	**1,644.77**	**342.85**	**348.87**

The result of the analysis of the ML class validation data areas with land S2GLC, revealed that the majority of the ML areas were over the herbaceous vegetation land cover in Greece (75.13%), Spain (39.49%), and Poland (89.88%) (Figure 13). In addition, another 26% of Spain's reference ML areas is found on sclerophyllous vegetation. Likewise, the validation data provided for Germany lie predominantly on moors and heathland (47.33%) and on natural material surfaces (38.52%) land cover.



	Artificial surfaces and constructions	Cultivated areas	Vineyards	Broadleaf tree cover	Coniferous tree cover	Herbaceous vegetation	Moors and Heathland	Sclerophyllous vegetation	Natural material surfaces
Greece	0.28	3.42	2.50	0.18	0.03	75.13	0.40	4.00	14.05
Spain	0.19	3.32	0.60	5.48	1.69	39.49	18.63	26.08	4.52
Germany	7.41	2.84		1.07	1.56	1.26	47.33		38.52
Poland	0.15	7.14		2.12	0.13	89.88			0.58

Figure 13. Distribution of land cover class sizes of provided ML validation data areas over S2GLC (in %).

5 DISCUSSION

5.1 Interpretation of the accuracy estimates

The comparison of the two types of accuracy assessment techniques is carried out below with the calculated metrics from both error matrices Table 27 as well as Table 26 which compares the hectares of ML for each country.

Table 26. Comparison of predicted and reference ML classes.

Area of ML in ha	Greece	Spain	Germany	Poland	Merged
Predicted (area-based)	7,646	1,807	8,820	313	18,589
Reference	7,987	1,649	351	538	10,529

Table 26 represents the area of the predicted ML from the area-based assessment and of the reference data provided by the project partners. The predicted area from the area-based assessment was chosen since, unlike the stratified random sampling, the intersection of the reference and classified polygons, make use of the total area under study. This table shows that the model in some cases overestimates and in some underestimates the true area of ML. In the case of Greece, the model underestimated the ML area by 341 ha. Same as Greece, Poland's true ML area is underestimated by 225 ha. On the other hand, Spain's and Germany's actual ML areas were overestimated by the 2.3 methodology model by 158 and 8,469 ha, respectively. Germany exhibits a significant and disproportional deviation from the reality, which affects the prediction of the merged countries, which overall present an overestimation of the true area of 8,060 ha.

Table 27. Results summary for ML of each country.

Metric of accuracy	Greece		Spain		Germany		Poland		Merged	
	Point-based	Area-based	Point-based	Area-based	Point-based	Area-based	Point-based	Area-based	Point-based	Area-based
OA (%)	71.52	70.75	82.87	83.42	60.61	59.79	90.97	90.56	67.98	67.73
UA (%)	77.73	76.86	77.47	77.98	3.62	3.53	92.41	90.74	42.69	42.40
PA (%)	73.89	73.58	84.66	85.45	90.06	88.60	54.17	52.79	75.36	74.87
F1-SCORE	75.76	75.19	80.90	81.54	6.96	6.78	68.30	66.75	54.50	54.14
ERR (%)	28.48	29.25	17.13	16.58	39.84	40.21	9.03	9.44	32.02	32.27
KAPPA	0.41	0.40	0.65	0.67	0.04	0.04	0.64	0.62	0.33	0.32
MCC	0.41	0.40	0.66	0.67	0.13	0.13	0.67	0.65	0.36	0.35

Table 27 presents the summary of the results of all the metrics examined, for both assessments (point-based and area-based) for each country. It indicates the results of the ML as these are the areas of interest. Concerning the ML areas in hectares, it is visible that Greece is tested -with a significant difference- the biggest number of hectares of ML followed by Spain, Germany and Poland. This difference of hectares between the countries is remarkable as the reference area of ML of Greece is almost four times bigger than Spain's, 14 times bigger than Germany's and 20 times bigger than Poland's.

"There is no general rule as to which level of accuracy is good and which is not. Judgment on the data validity depends on the purpose of the map and thus needs to be dealt with on a case-by-case basis."(Finegold et al., 2016).Table 27 illustrates the OA of each country, showing that Poland had the higher percentage of OA which means that Poland had the greatest correctly classified proportion out of all reference polygons. Poland is followed by Spain, Greece and German with the last one having the lowest OA. The OA may have been significantly influenced by the ML and nML classification layer detection methodologies' input data, as well as the quality of the validation data areas provided. The different input data utilized in the detection methodology for the delineation of forests, croplands, protected areas, impervious, changed areas, and other land covers also had an overall accuracy less than 100 percent. As a result, this potential error is expected to have influenced the overall accuracy result of the accuracy assessment. Furthermore, as the S2GLC, which is the detection methodology's basemap, was produced in 2017, and the validation data for the assessment were all acquired in 2021, there is a substantial gap in the elapsed time between the produced classified layer and the validation data. As a result, some land use and/or land cover may have changed, which introduces an additional source of error, thus significantly influencing the result of the overall accuracy of the accuracy assessment.

The producer's accuracy is based on the producer's classification point of view, while the user's accuracy shows the reality on the ground. In this case, there is a substantial contrast between the countries. The PA shows a high percentage for Germany and a moderate one for Poland, while UA an extremely low percentage for Germany and great one for Poland. Greece and Spain's PA and UA are almost similar and above 70% in both cases.

Regarding the ERR, Poland has a minimum error rate followed by Spain, Greece and Germany. German reaches a 40% of the ERR which is bigger than the other countries' but still is a fair rate.

The F1-score is a metric that implies how accurate a model is and depending on UA and PA, in this case, the accuracy assessments of Spain and Greece are showing a substantial accuracy. Poland is following with a fair percentage of accuracy, and last Germany with a very low percentage of accuracy.

According to Table 4 and Table 27 kappa shows a substantial agreement for Spain and Poland and a moderate agreement for Greece. On the other hand, Germany shows a slight agreement confirming once again the deficiency in the country's data.

Last but not least, MCC of Spain and Poland have an equal and quite high correlation between predicted and actual classes, followed by Greece and Germany with the last showing a very low correlation.

From the analysis, it is evident that the results of the merged countries are influenced by the lack of ML data in Germany. The inconsistency observed in the values of the accuracy assessment measures between the four countries under investigation was due to the validation data. Different issues occurred with the validation data:

- Different methods are used by different experts in each country in acquiring data.
- Large differences in the amount of validation data areas provided by each country.
- Class imbalance.
- Determining sample size for the accuracy assessment.

These differences in accuracy measures in each country can be lowered by deriving a standard procedure for acquiring the validation data areas for accuracy assessment or by being acquired from field measurements of the test areas. The use of an equal amount of validation data areas in each country will also prevent this inconsistency significantly. Also, class imbalance concerns, if avoided, could improve value discrepancies by using nearly equal or equal extents of ML and nML class areas. For example, in Germany, the ML class area was 352 ha while the nML class area was 20,913 ha in this case, the dominating effect of the majority nML class will have a

significant effect on the final assessment measures. Compared to Greece, the difference was not as large as 7,988 ha of ML and 5,274 ha of nML was provided, the dominating effect of the majority ML class will not be as significant.

By considering the proportional allocation of the total sample size, defining a standard procedure or methodology of determining sample size will also standardize the accuracy assessment procedure and bring conformity to the values.

5.2 Reasons for difference in land cover class peculiar with ML

The land cover analysis conducted between the ML validation sites and the S2GLC layer identified that the ML areas were over different land cover types. The majority of the ML reference sites in Greece, Spain, and Poland were in the herbaceous land cover, while in Germany they were over the moors and heathland, as well as over the natural material surfaces land cover categories of the S2GLC map. The differences are attributed to the different types of land use and land cover areas that are associated with ML provided by each project partner in their respective country for the accuracy assessment. This could be a result of the difference in climate or in which areas are protected by national laws. More specifically, Greece and Spain have mostly Mediterranean climate, while Poland and Germany mostly temperate. For instance, most of the ML areas in Poland are a result of the abandonment of agricultural fields, which is not the case in Greece, Spain, or Germany.

According to the S2GLC documentation, the herbaceous land cover is defined as ‘*’Lands covered by herbaceous vegetation including both natural low productivity grassland and managed grassland used for grazing and/or mowing”*, hence these are potential areas that can be reforested for carbon sequestration. While most of the areas provided as ML validation data for Greece and Spain were similar with the ones provided for Germany, a significant proportion of the later was defined as protected areas by national law and the provided ML sites were from previous mines and extraction sites.

A peculiarity with the ML of Germany is that many areas of herbaceous vegetation are under one form of protection (mostly for Flora and Fauna), thus cannot be reforested. This is why we identify most of ML areas are under moors and heathland, which according to the S2GLC definition are ‘*’Low growing vegetation with closed cover and with predominately shrub and bushy vegetation”*. According to literature, such areas might also get classified in the near future as protected areas, consequently, any

reforestation can hardly occur there. This is according to the 2002 Federal Nature Conservation Act that created a new statutory requirement for the Länder (states) to set up a network of interlinked biotopes covering at least 10 percent of their area (Section 21 of the Act). The next predominant ML land cover type provided by the German partners is the natural material surfaces, which usually consist of former large mines or mineral extraction sites (e.g., Nochten area in Saxony), for which reforestation strategies of significant extent are already planned.

6 CONCLUSIONS

In order to quantify the ML and nML classification layer data that would allow users to understand the performance of the detection methodology, as well as provide guidelines for the validation of the products, based on provided validation dataset from project partners, a point-based and an area-based accuracy assessment method was performed for the ML and nML classification layer products that cover validation sites in Greece, Spain, Germany and Poland in the EU. The results indicated that from the provided validation data areas Poland had the highest overall accuracy for both the point-based and the area-based accuracy assessment methods of (90.97% and 90.56%) followed by Spain (82.87% and 82.42%), then Greece (71.52% and 70.75%) and Germany (60.61% and 59.79%). The results of the area-based assessment indicated that the ML area in Greece was underestimated by 341 ha and in Poland by 225 ha, while in Spain they were overestimated by 158 ha and in Germany by 8,469 ha. Generally, the ML areas for all testing sites were overestimated by 8,060 ha.

Analysis of the error matrix and provided validation data areas with the S2GLC, showed that in Greece, Spain and Poland majority of the ML areas were over the Herbaceous land cover class while in Germany it was the Moors and Heathland, and Natural Material Surfaces land cover classes.

7 REFERENCES

Agyemang, T. K., Heblinski, J., Schmieder, K., Sajadyan, H., & Vardanyan, L. (2011). Accuracy assessment of supervised classification of submersed macrophytes: the case of the Gavaraget region of Lake Sevan, Armenia. *Hydrobiologia*, *661*(1), 85–96. https://doi.org/10.1007/s10750-010-0465-7

Altman, D. G., & Bland, J. M. (1994). Diagnostic tests. 1: Sensitivity and specificity. *BMJ: British Medical Journal*, *308*(6943), 1552.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. F., & Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, *16*(5), 412–424. https://doi.org/10.1093/bioinformatics/16.5.412

Banko, G. (1998). *A Review of Assessing the Accuracy of Classifications of Remotely Sensed Data and of Methods Including Remote Sensing Data in Forest Inventory*. IR-98-081.

Barnston, A. G. (1992). Correspondence among the Correlation, RMSE, and Heidke Forecast Verification Measures; Refinement of the Heidke Score. *Weather and Forecasting*, *7*(4), 699–709. https://doi.org/10.1175/1520-0434(1992)007<0699:CATCRA>2.0.CO;2

Bolstad, P., Jenks, A., Berkin, J., Horne, K., & Reading, W. H. (2005). A Comparison of Autonomous, WAAS, Real-Time, and Post-Processed Global Positioning Systems (GPS) Accuracies in Northern Forests. *Northern Journal of Applied Forestry*, *22*(1), 5–11. https://doi.org/10.1093/njaf/22.1.5

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(7), 1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2

Campbell, J. B. (1981). Spatial Correlation Effects upon Accuracy of Supervised Classification of Land Cover. *Photogrammetric Engineering and Remote Sensing*, *47*, 355–363.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004). Methods for balancing machine

training digital text categorization from documents Towards scalable with non-uniform class. *ACM SIGKDD Explorations Newsletter*, *6*(1), 1–6. https://doi.org/10.1145/1007730.1007733

Cliff, A. D. (1973). *Spatial autocorrelation*.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Congalton, R. G. (1988a). A comparison of sampling schemes used in generating error matrices for assessing the accuracy of maps generated from remotely sensed data. *Photogrammetric Engineering and Remote Sensing (USA)*.

Congalton, R. G. (1988b). Using spatial autocorrelation analysis to explore the errors in maps generated from remotely sensed data. *Photogrammetric Engineering and Remote Sensing (USA)*.

Congalton, R. G. (1991). A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment*, *37*(1), 35–46.

Congalton, R. G. (2001). Accuracy assessment and validation of remotely sensed and other spatial information. *International Journal of Wildland Fire*, *10*(4), 321. https://doi.org/10.1071/WF01031

Congalton, R. G. (2007). Thematic and positional accuracy assessment of digital remotely sensed data. *In: McRoberts, Ronald E.; Reams, Gregory A.; Van Deusen, Paul C.; McWilliams, William H., Eds. Proceedings of the Seventh Annual Forest Inventory and Analysis Symposium; October 3-6, 2005; Portland, ME. Gen. Tech. Rep. WO-77. Washington, DC: US Department*, *77*.

Congalton, R. G., & Green, K. (1999). Basic analysis techniques. In *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*. CRC Press: Boca Raton, FL, USA.

Congalton, R. G., & Green, K. (2019). *Assessing the accuracy of remotely sensed data: principles and practices*. CRC press.

Finegold, Y., Ortmann, A., Lindquist, E., D'Annunzio, R., & Sandker, M. (2016). Map accuracy assessment and area estimation: a practical guide. *Rome: Food and Agriculture Organization of the United Nations*.

Foody, G. M. (2002). Status of land cover classification accuracy assessment. *Remote Sensing of Environment*, *80*(1), 185–201.

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

Global Forest Observations Initiative GFOI. (2013). *Integrating remote-sensing and ground-based observations for estimation of emissions and removals of greenhouse gases in forests* (Issue January). Group on Earth Observations.

Goutte, C., & Gaussier, E. (2005). A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. *European Conference on Information Retrieval*, 345–359.

Hay, A. M. (1979). Sampling Designs to Test Land-Use Map Accuracy. *Photogrammetric Engineering and Remote Sensing*, *45*, 529–533.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284.

Jensen, J. R. (1996). *Introductory digital image processing: a remote sensing perspective.* (Issue Ed. 2). Prentice-Hall Inc.

Kubat, M., Holte, R. C., & Matwin, S. (1998). Machine learning for the detection of oil spills in satellite radar images. *Machine Learning*, *30*(2), 195–215.

Landis, J. R., & Koch, G. G. (1977). *The Measurement of Observer Agreement for Categorical Data* (Vol. 33, Issue 1).

Lillesand, T. M., Kiefer, R. W., & Chipman, J. W. (2015). *Remote sensing and image interpretation. Hoboken* (p. 18). NJ: Wiley.

Lunetta, R., Congalton, R. G., Fenstermaker, L., Jensen, J., Mcgwire, K., & Tinney, L. R. (1991). Remote sensing and geographic information system data integration-Error sources and research issues. *Photogrammetric Engineering and Remote Sensing*, *57*(6), 677–687.

Malinowski, R., Lewiński, S., Rybicki, M., Gromny, E., Jenerowicz, M., Krupiński, M., Nowakowski, A., Wojtkowski, C., Krupiński, M., & Krätzschmar, E. (2020). Automated Production of a Land Cover/Use Map of Europe Based on Sentinel-2 Imagery. *Remote Sensing*, *12*(21), 3523.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, *405*(2), 442–451. https://doi.org/10.1016/0005-2795(75)90109-9

Munoz, S. R., & Bangdiwala, S. I. (1997). Interpretation of Kappa and B statistics measures of agreement. *Journal of Applied Statistics*, *24*(1), 105–112. https://doi.org/10.1080/02664769723918

Olofsson, P., Foody, G. M., Herold, M., Stehman, S. V., Woodcock, C. E., & Wulder, M. A. (2014). Good practices for estimating area and assessing accuracy of land change. *Remote Sensing of Environment*, *148*, 42–57. https://doi.org/10.1016/j.rse.2014.02.015

Pontius Jr, R. G., & Millones, M. (2011). Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. *International Journal of Remote Sensing*, *32*(15), 4407–4429.

Rao, R. B., Krishnan, S., & Niculescu, R. S. (2006). Data mining for improved cardiac care. *Acm Sigkdd Explorations Newsletter*, *8*(1), 3–10.

Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*, *10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Sim, J., & Wright, C. C. (2005). The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. *Physical Therapy*, *85*(3), 257–268. https://doi.org/10.1093/ptj/85.3.257

Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006). Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation. *AAAI Workshop - Technical Report*, *WS-06-06*, 24–29. https://doi.org/10.1007/11941439_114

Stehman, S. V. (1992). Comparison of systematic and random sampling for estimating the accuracy of maps generated from remotely sensed data. *Photogrammetric Engineering and Remote Sensing*, *58*(9), 1343–1350.

Stehman, S. V. (1997). Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment*, *62*(1), 77–89.

Story, M., & Congalton, R. G. (1986). Accuracy Assessment: A User's Perspective. *Photogrammetric Engineering and Remote Sensing*, *52*, 397–399.

Tharwat, A. (2020). Classification assessment methods. *Applied Computing and Informatics*.

Torralba, J., Crespo-Peremarch, P., & Ruiz, L. A. (2018). Evaluación del uso de LiDAR discreto, full-waveform y TLS en la clasificación por composición de especies en bosques mediterráneos. *Revista de Teledetección*, *52*, 27. https://doi.org/10.4995/raet.2018.11106

Tortora, R. D. (1978). A Note on Sample Size Estimation for Multinomial Populations. *The American Statistician*, *32*(3), 100–102. https://doi.org/10.1080/00031305.1978.10479265

Van Genderen, J. L., & Lock, B. F. (1977). Testing land-use map accuracy. *Photogrammetric Engineering and Remote Sensing*, *43*(9).

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistic. *Fam Med*, *37*(5), 360–363.

Woodcock, C. E., & Gopal, S. (2000). Fuzzy set theory and thematic maps: accuracy assessment and area estimation. *International Journal of Geographical Information Science*, *14*(2), 153–172. https://doi.org/10.1080/136588100240895

8 ANNEX I: INITIAL APPROACH

8.1 Datasets

8.1.1 Testing sites

The testing sites included in the ***MAIL*** project present a variety of geomorphology and land cover types. They are located in 4 different countries, in south Europe on the Iberian Peninsula (Spain) and on the Balkan peninsula (Greece), and in north European plain (Germany and Poland).

The total number of testing sites in each country is as follows:

- 3 testing sites in Spain (Sierra de Espadán, Nogueruelas and Terras Atlas).
- 2 testing sites in Greece (Thessaloniki and Komotini).
- 2 testing sites in Germany (Nochten/Reichwalde and Welzow).
- 1 testing site in Poland.

Testing sites in Spain

The testing site of **Sierra de Espadán** covers a total area of 763.22 km² and is located nearly 40 km north of the city of Valencia, in the eastern Spain province of Castellón. The altitude ranges between 250 and 1,000 m above sea level. The area is mainly characterized by forest (the Natural Park of Sierra de Espadán), heathland and cultivated areas. The Natural Park of Sierra de Espadán is a Mediterranean forest with soft and rounded hills, presence of abandoned farming with artificial terraces, and mountain peaks up to 1,100 m of altitude. The Natural Park displays a heterogeneous landscape dominated by pure and mixed native coniferous and deciduous forests, with species of *Pinus* and *Quercus* (Torralba et al., 2018)

The testing site of **Nogueruelas** covers a total area of 42 km² and is located north of the municipality of Nogueruelas (Teruel) about 65 km from the city of Teruel. This is an eminently forested area located in the heart of Sierra de Gúdar. The altitude of the study area ranges between 600 and 1,800 m above sea level. The slopes in the study area are gentler than in the environment due to the fact that the mountain is located in areas of high plateaus, with the appearance of gentle slopes. The area is mainly characterized by forest and heathland.

The testing site of **Terras Atlas** covers a total area of 1,100 km² and is located nearly 210 km northeast of the city of Madrid. The altitude of the study area ranges between 900 and 1600 m above sea level. The area is mainly characterized by forest, heathland and cultivated areas.

Testing sites in Greece

The testing site of **Thessaloniki** covers a total area of 96.63 km² and is located nearly 15 km east of the city of Thessaloniki. The altitude varies significantly from 70 m (the relatively flat lowland area in the southeast which includes cultivated areas) to 1,100m (the mountainous area in the northwest which includes low vegetation areas and natural material surfaces) above sea level. The area is mainly characterized by heathland, forest and cultivated areas.

The testing site of **Komotini** covers a total area of 79.93 km² and is located nearly 15 km south of the city of Komotini. The altitude varies significantly from 50 m (the relatively flat lowland area in the southwest which includes cultivated areas) to 500 m (the mountainous area in the northeast which includes low vegetation areas and natural material surfaces) above sea level. The area is mainly characterized by heathland, forest and cultivated areas.

Testing sites in Germany

The testing site of **Nochten/Reichwalde** covers a total area of 1,042.15 km² and is located nearly 60 km northwest of the city of Dresden. The altitude does not vary considerably from 120 to 160 m above sea level. The area is mainly characterized by heathland, forest, lakes, opencast mining areas and military training areas. Only a very small extent of protected areas exists. In the context of the MLs definition especially post-mining areas are relevant.

The testing site of and **Welzow** covers a total area of 224.43 km² and is located nearly 65 km northwest of the city of Dresden. The altitude varies from 40 to 140 m above sea level. The area is mainly characterized by heathland, forest, cultivated areas and opencast mining areas.

Testing sites in Poland

The testing site of Poland covers a total area of 480.77 km² and is located nearly 190 km south of the city of Warsaw. The altitude does not vary considerably and ranges between 150 (the relatively flat lowland area in the southeast which includes mainly

cultivated areas) and 220 m (the flat lowland area in the northwest which includes also cultivated areas) above sea level. The area is mainly characterized by cultivated areas and forest.

8.1.2 Final Layer

The geodatabase, Final Map Layer (layer FINAL_ML, see **MAIL** deliverable .2.3), including the estimated MLs of the testing sites, was reclassified to 3 classes: (I) Marginal Lands, (II) Potential Marginal Lands and (III) Unsuitable Lands.

According to the methodology that was followed in deliverable 2.3 of the ***MAIL*** project, the surface area and the percentage of the area that was classified as ML in the assessed testing sites (Greece, Germany and Poland), presents a significant variability (Table 28).

Table 28. Total surface of testing sites and estimated MLs.

	Testing site (km²)	**Marginal lands (1)** (km²)	**Potential marginal lands (2)** (km²)	**Unsuitable lands (3)** (km²)	**SUM (1),(2),(3)** (km²)	**Percentage** (%)
GREECE - Thessaloniki	96.63	7.02	0.14	2.06	9.22	9.53
GREECE - Komotini	79.93	3.15	0.04	0.72	3.91	4.89
GERMANY - Nochten/Reichwalde	1,042.15	0.97	0.02	0.16	1.15	0.11
POLAND	480.77	0.36	0.09	0.17	0.62	0.13

The class ‘Marginal Lands’ includes: Herbaceous vegetation, moors and heathland, sclerophyllous vegetation, natural material surfaces. The class ‘Potential Marginal Lands’ (Supra-marginal agricultural lands) includes cultivated areas and vineyards. The class ‘Unsuitable Lands’ includes artificial surfaces and constructions, broadleaf tree cover and coniferous tree cover.

8.1.3 Reference Data

Ground truth data can be collected in the field (field-verified ground reference locations e.g., ground truth with GPS). However, this is time consuming and expensive. Ground truth data can also be derived from interpreting high-resolution imagery, existing

classified imagery, or GIS data layers (pixels as references visually identified from the imagery e.g., aerial photo interpretation). In the ***MAIL*** project, the Imagery (WGS84) was used as reference data/basemap using ArcGIS Online from Esri.

The web map (WGS84), used as basemap using ArcGIS Online from Esri, features satellite imagery for the world and high-resolution aerial imagery for many areas. It uses WGS84 Geographic, version 2 tiling scheme. World Imagery (WGS84) provides one meter or better satellite and aerial imagery in many parts of the world and lower resolution satellite imagery worldwide. The map includes 15 m TerraColor imagery at small and mid-scales (~1:591 M down to ~1:72 k) and 2.5 m SPOT Imagery (~1:288 k to ~1:72 k) for the world. The map features 0.3 m resolution imagery in the continental United States and parts of Western Europe from Maxar. Additional Maxar sub-meter imagery is featured in many parts of the world. In the United States, 1 meter or better resolution NAIP imagery is available. In other parts of the world, imagery at different resolutions has been contributed by the GIS User Community. In select communities, very high-resolution imagery (to 0.03 m) is available down to ~1:280 scale.

“ArcGIS Online Basemaps” features a variety of basemaps that can be accessed from ArcGIS Online. This includes basemaps from Esri and OpenStreetMap. The basemaps (World Imagery, World Street Map, National Geographic World Map, World Topographic Map, Streets, Navigation etc.) can be used as foundation layers to support a range of web maps or web mapping applications. In the ***MAIL*** project, apart from the Imagery (WGS84) that was used as basemap using ArcGIS Online from Esri, the World Topographic Map was also used to identify the testing sites.

8.2 Methodology

The basic steps of the methodology applied to testing sites is as follows in the workflow below.

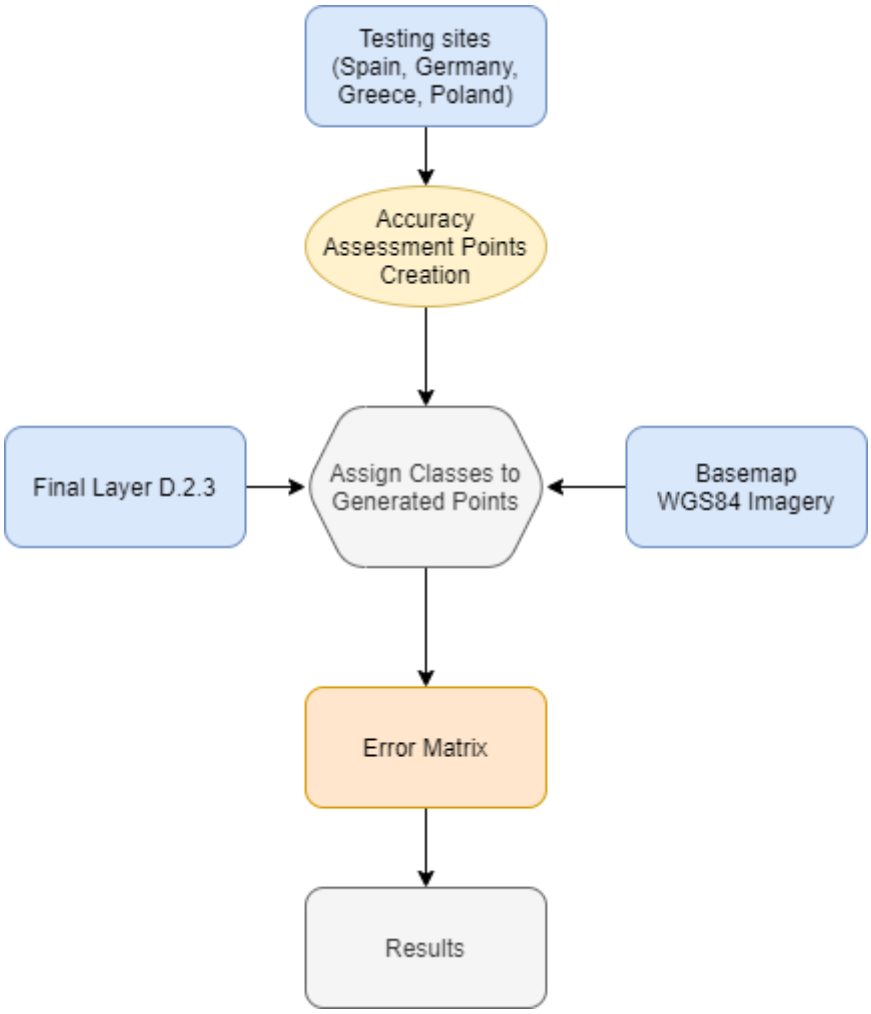


Figure 14. Initial approach workflow.

8.2.1 Sampling Strategy

The sampling scheme, the choice and distribution of samples, is an important part of the accuracy assessment. Selection of the proper scheme is critical to generate an error matrix that is representative of the entire map. The sample must be selected without bias. Failure to meet this important criterion affects the validity of any further analysis performed because the resulting error matrix may overestimate or underestimate the true accuracy.

The sampling strategy that is applied here is stratified random sampling. In this process, the study area is split into strata and random samples are generated within each stratum. Strata can be adjusted based on prior knowledge about the study area in order to divide the area into groups or strata and then each stratum is randomly sampled. Then the map is being stratified into map classes.

In the ***MAIL*** project, the creation of randomly sampled points for the accuracy assessment was made using the Create Accuracy Assessment Points tool in ArcGIS pro. This tool creates a set of random points and assigns a class to them based on reference data. The points creation is dispersed randomly inside each class, with the number of points proportionate to the relative area of each class. The total number of random points that will be generated, depending on sampling strategy and the number of classes, can be selected. The default number of randomly generated points is 500.

8.2.2 Evaluation

The randomly generated points for the accuracy assessment of MLs (Marginal Lands, Potential Marginal Lands, Unsuitable Lands and Excluded Areas) of the testing site were compared to the same locations (the pixels as references visually identified from the aerial imagery) using the web map (WGS84) in ArcGIS pro (Figure 15).

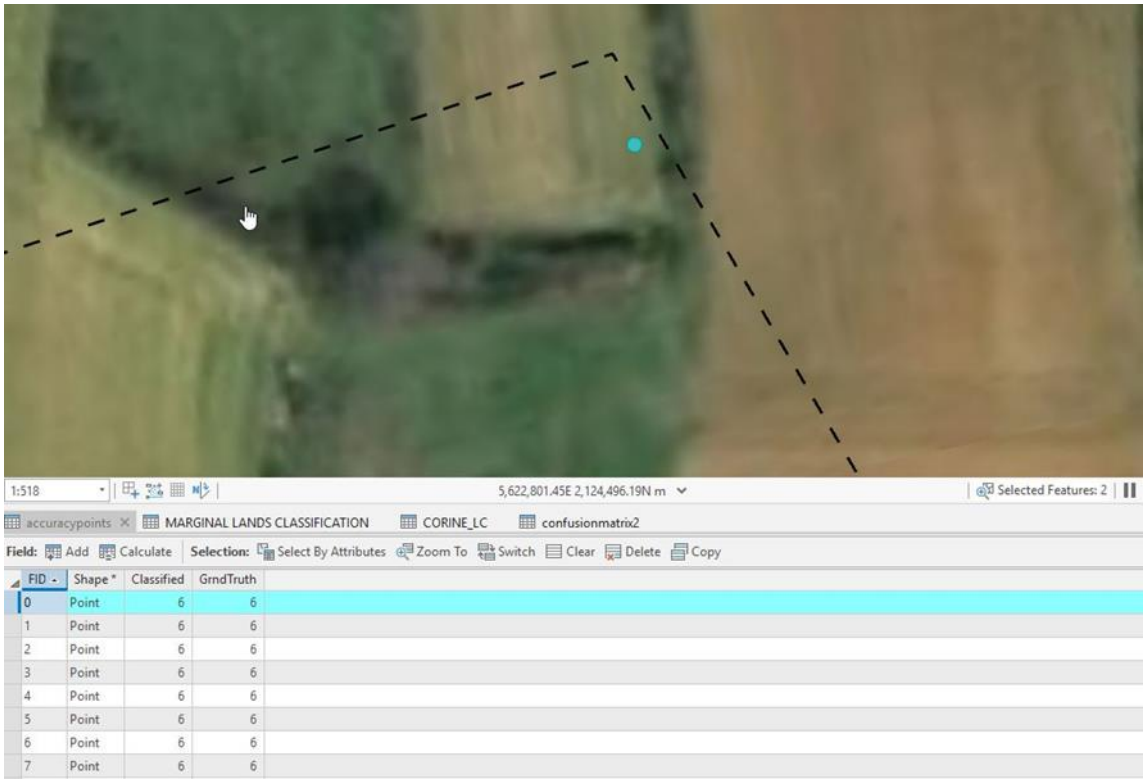


Figure 15. Manually assign classes (GroundTruth) to the generated points for the accuracy assessment of ML of the testing site in Greece (Komotini) (Basemap: ArcGIS Online from Esri).

The creation of error matrix for the Accuracy Assessment of ML of the testing sites consisted of the computation of the percent accuracy for each informational class

(Marginal Lands, Potential Marginal Lands, Unsuitable Lands and Excluded Areas), the user's accuracy and the producer's accuracy, the overall-total accuracy (average summary value) and the calculation of Kappa (Figure 16).

Field: Add Calculate Selection: Select By Attributes Zoom To Switch Clear Delet

OBJECTID *	ClassValue	C_1	C_2	C_3	C_6	Total	U_Accuracy	Kappa
1	C_1	11	0	0	9	20	0.55	0
2	C_2	6	1	0	3	10	0.1	0
3	C_3	0	0	6	4	10	0.6	0
4	C_6	27	0	0	449	476	0.943277	0
5	Total	44	1	6	465	516	0	0
6	P_Accuracy	0.25	1	1	0.965591	0	0.905039	0
7	Kappa	0	0	0	0	0	0	0.424919

Figure 16. Creation of error matrix for the accuracy assessment of ML of the testing site in Greece (Komotini) (Basemap: ArcGIS Online from Esri).

The error matrix Table for each testing site was converted to an Excel file in order to further clarify the information about the classes (Marginal Lands, Potential Marginal Lands, Unsuitable Lands and Excluded Areas) for the Accuracy Assessment of the testing site.

8.3 Results

The evaluation of the Accuracy Assessment of the testing sites consists of the interpretation of the error matrix table (the user's accuracy, the producer's accuracy and the overall-total accuracy) and for each testing site is as follows:

8.3.1 Greece - Thessaloniki

	Thessaloniki								
OBJECTID	ClassValue	Marginal Lands	Potential Marginal Lands	Unsuitable lands	Excluded areas	Total	U_Accuracy (%)	U_Accuracy	Kappa
1	Marginal Lands	20	1	0	15	36	55.56	0.555555556	0
2	Potential Marginal Lands	2	8	0	0	10	80.00	0.8	0
3	Unsuitable lands	0	0	4	7	11	36.36	0.363636364	0
4	Excluded areas	51	4	0	397	452	87.83	0.878318584	0
5	Total	73	13	4	419	509		0	0
6	P_Accuracy (%)	27.40	61.54	100.00	94.75				
7	P_Accuracy	0.273972603	0.615384615	1	0.947494033	0		0.842829077	0
8	Kappa	0	0	0	0	0		0	0.391248449

Figure 17. The error matrix table in the excel file of the testing site in Greece (Thessaloniki).

The error matrix of this testing site (Thessaloniki) indicates an overall accuracy of 84.28%. However, producer's accuracies range from just 27.40% ("Marginal Lands") to 100% ("Unsuitable Lands") and user's accuracies vary from 36.65% ("Unsuitable Lands") to 87.83% ("Excluded areas"). At this point, it is important to appreciate the need for considering overall, producer's, and user's accuracies simultaneously. In this

example, the overall accuracy of the classification is 84.28% (a quite good accuracy). However, as the primary purpose of the classification is to map the locations of the "Marginal Lands" category, we note that the producer's accuracy of this class is not good at all (27.40%) and the user's accuracy for this class is only 55.56%. That is to say, only 27.40% of the MLs have been correctly identified as "Marginal Lands" and 55.56% of the areas identified as "Marginal Lands" within the classification are truly of that category. The only highly reliable category associated with this classification from both a producer's (94.75%) and a user's (87.83%) perspective is "Excluded areas".

8.3.2 Greece - Komotini

	Komotini								
OBJECTID	ClassValue	Marginal Lands	Potential Marginal Lands	Unsuitable lands	Excluded areas	Total	U_Accuracy (%)	U_Accuracy	Kappa
1	Marginal Lands	11	0	0	9	20	55.00	0.55	0
2	Potential Marginal Lands	6	1	0	3	10	10.00	0.1	0
3	Unsuitable lands	0	0	6	4	10	60.00	0.6	0
4	Excluded areas	27	0	0	449	476	94.33	0.943277311	0
5	Total	44	1	6	465	516		0	0
6	P_Accuracy (%)	25.00	100.00	100.00	96.56				
7	P_Accuracy	0.25	1	1	0.965591398	0		0.90503876	0
8	Kappa	0	0	0	0	0		0	0.424919256

Figure 18. The error matrix table in the excel file of the testing site in Greece (Komotini).

The error matrix of this testing site (Komotini) indicates an overall accuracy of 90.50%. However, producer's accuracies range from just 25.00% ("Marginal Lands") to 100% ("Unsuitable Lands" and "Potential Marginal Lands") and user's accuracies vary from 10.00% ("Potential Marginal Lands") to 94.32% ("Excluded areas"). In this example, the overall accuracy of the classification is 90.50% (a good accuracy). However, as the primary purpose of the classification is to map the locations of the "Marginal Lands" category, we note that the producer's accuracy of this class is not good at all (25.00%) and the user's accuracy for this class is only 55.00%. That is to say, only 25.00% of the MLs have been correctly identified as "Marginal Lands" and 55.00% of the areas identified as "Marginal Lands" within the classification are truly of that category. The only highly reliable category associated with this classification from both a producer's (96.56%) and a user's (94.33%) perspective is "Excluded areas".

8.3.3 Germany - Nochten/Reichwalde

	Germany								
OBJECTID	ClassValue	Marginal Lands	Potential Marginal Lands	Unsuitable lands	Excluded areas	Total	U_Accuracy (%)	U_Accuracy	Kappa
1	Marginal Lands	6	0	0	4	10	60.00	0.6	0
2	Potential Marginal Lands	0	9	0	1	10	90.00	0.9	0
3	Unsuitable lands	1	3	3	3	10	30.00	0.3	0
4	Excluded areas	52	0	0	447	499	89.58	0.895791583	0
5	Total	59	12	3	455	529		0	0
6	P_Accuracy (%)	10.17	75.00	100.00	98.24				
7	P_Accuracy	0.101694915	0.75	1	0.982417582	0		0.879017013	0
8	Kappa	0	0	0	0	0		0	0.349623482

Figure 19. The error matrix table in the excel file of the testing site in Germany (Nochten/Reichwalde).

The error matrix of this testing site (Nochten/Reichwalde) indicates an overall accuracy of 87.90%. However, producer's accuracies range from just 10.17% ("Marginal Lands") to 100% ("Unsuitable Lands") and user's accuracies vary from 30.00% ("Unsuitable Lands") to 90.00% ("Potential Marginal Lands"). In this example, the overall accuracy of the classification is 87.90% (a good accuracy). However, as the primary purpose of the classification is to map the locations of the "Marginal Lands" category, we note that the producer's accuracy of this class is not good at all (10.17%) and the user's accuracy for this class is 60.00%. That is to say, only 10.17% of the MLs have been correctly identified as "Marginal Lands" and 60.00% of the areas identified as "Marginal Lands" within the classification are truly of that category. The only highly reliable category associated with this classification from both a producer's (98.24%) and a user's (89.58%) perspective is "Excluded areas".

8.3.4 IV. Poland - Staszow

	Poland								
OBJECTID	ClassValue	Marginal Lands	Potential Marginal Lands	Unsuitable lands	Excluded areas	Total	U_Accuracy (%)	U_Accuracy	Kappa
1	Marginal Lands	0	0	0	10	10	0.00	0	0
2	Potential Marginal Lands	1	1	0	8	10	10.00	0.1	0
3	Unsuitable lands	0	0	6	4	10	60.00	0.6	0
4	Excluded areas	7	0	0	492	499	98.60	0.985971944	0
5	Total	8	1	6	514	529		0	0
6	P_Accuracy (%)	0.00	100.00	100.00	95.72				
7	P_Accuracy	0	1	1	0.957198444	0		0.943289225	0
8	Kappa	0	0	0	0	0		0	0.316095669

Figure 20. The error matrix table in the excel file of the testing site in Poland (Staszow).

The error matrix of this testing site (Poland) indicates an overall accuracy of 94.33%. However, producer's accuracies range from just 0.00% ("Marginal Lands") to 100% ("Unsuitable Lands" and "Potential Marginal Lands") and user's accuracies vary from 0.00% ("Marginal Lands") to 98.60% ("Excluded areas"). In this example, the overall accuracy of the classification is 94.33% (a good accuracy). However, as the primary purpose of the classification is to map the locations of the "Marginal Lands" category, we note that the producer's accuracy of this class is not good at all (0.00%) and the user's accuracy for this class is also 0.00%. That is to say, 0.00% of the MLs have

been correctly identified as “Marginal Lands” and 0.00% of the areas identified as “Marginal Lands” within the classification are truly of that category. The only highly reliable category associated with this classification from both a producer’s (95.72%) and a user’s (98.60%) perspective is “Excluded areas”.

9 ANNEX II

10 ANNEX III