www.multiminer.eu

@company/multiminer

@MultiMiner_EU

Grant Agreement No: 101091374

Project Start Date: 1st January 2023

Project Duration: 42 Months

Deliverable No 4.1

Part 2: Review of accuracy assessment methods

The MultiMiner project is funded by the European Union's Horizon Europe research and innovation actions programme under Grant Agreement No. 10109137474

Grant agreement	**101091374**
Project title	Multi-Source and Multi-Scale Earth Observation and Novel Machine Learning for Mineral Exploration and Mine Site Monitoring
Project acronym	MultiMiner
Project coordinator	GTK Geological Survey of Finland
Project start date	Jan 1, 2023
Project duration	42 months
Related work packages	WP 4 – Site Demonstrations
Related task(s)	Task 4.2 – Thematic accuracy assessment methods
Lead organization	EFTAS
Contributing partners	GTK, BGR, VTT, CGS
Submission date	October 31, 2025
File name	D4.1_Part2_Review of accuracy assessment methods_V2.3_final.docx
Organisation responsible for deliverable	EFTAS, GTK
Author name(s)	Alireza Hamedianfar, Matthieu Molinier, Maarit Middleton, Veronika Kopackova-Strnadová, Martin Kýhos, Michaela Frei, Andre Kalia, Martin Schodlok, Oleg Antropov, Sebastian Teuwsen, Luca Kleinewilinghöfer, Susanne Kocjan
Date	August 27, 2024
Version	2.3
Status	Complete
Dissemination Level	Public

Revision History

Version	Date	Modified By	Comments
V0.08	28.02.2024	EFTAS	Clean version
V1.0	17.04.2024	EFTAS	Clean version
V2.0	21.05.2024	EFTAS	Clean version and merge of subchapters
V2.1	07.06.2024	EFTAS	Clean version
V2.2	30.06.2024	EFTAS	Final version, submitted
V2.3	27.08.2024	GTK	Updated version, submitted

Disclaimer

The contents of this publication are the sole responsibility of its author and do not necessarily reflect the opinion of the European Union.

The MultiMiner project is funded by the European Union's Horizon Europe research and innovation actions programme under Grant Agreement No. 10109137474

Content

List of Tables

List of Figures

The MultiMiner project is funded by the European Union's Horizon Europe research and innovation actions programme under Grant Agreement No. 10109137474

List of Appendices

Appendix 1: Definition of the thematic accuracy assessment procedures for MultiMiner application demonstrations (SENSITIVE)

List of Acronyms

Table 1. A list of acronyms used in this document.

Abbreviation	Meaning
AMD	Acid mine drainage
AUC	Area Under the Curve
DEM	Digital elevation model
EASA	European Union Aviaton Safety Agency
EDAP	Earthnet Data Assessment Project
EO	Earth Observation
ESA	European Space Agency
FPR	False positive rate
in situ	in the natural or original position or place
IoU	Intersection over union
LOOCV	Leave-One-Out Cross-Validation
MAE	Mean Absolute Error
ML	Machine Learning
MSE	Mean Squared Error
NASA	The National Aeronautics and Space Administration
PA	Producer’s accuracy
PIMA	Portable infrared mineral analyser
PSU	Primary Sampling Units
QA4EO	Quality Assurance Framework for Earth Observation
R^2	Coefficient of determination
RMSE	Root Mean Squared Error
ROC	Receiver Operating Characteristics
RPD	Ratio of Performance to Deviation
SAR	Synthetic Aperture Radar
SD	Standard Deviation of sample
SEP	Standard Error of prediction
SORA	Specific Operations Risk Assessment
SSU	Secondary Sampling Units
TPR	True positive rate
TSF	Tailings storage facility
UA	User’s accuracy
UAV	Unmanned areal vehicle

The MultiMiner project is funded by the European Union’s Horizon Europe research and innovation actions programme under Grant Agreement No. 10109137474

Executive Summary

The deliverable 4.1 Part 2 “Review of accuracy assessment methods” is a contribution of MultiMiner Task 4.1: In Situ data standardization and databases to Deliverable D4.1: “Field work protocols for in situ data collection and review of accuracy assessment methods”. It includes a description of the data collection standards and protocols to be used in accuracy assessment of the multi-scale and multi-source data products and test site specific ground truth activities. Accuracy assessment is the final step of the data processing chain which results in assessment for the quality of the application demonstrations done in the MultiMiner project. The accuracy assessment of quantitative image classification provides measures of the relative quality. Based on these measures notifications about the exactness of the estimations and the systematic errors of the outcomes can be made.

For a statistically robust and transparent approach in the MultiMiner project, chapter 1 describes the principles of thematic accuracy assessment methods, which are important to ensure the integrity of the resulted products. Methods of quantitative image classification accuracy assessment are presented. Sample size calculation, sampling design methods and strategies concerning imbalance/ rare classes and mitigation of limited reference data are discussed.

Chapter 2 provides an overview of the quality of EO data used for MultiMiner applications. As the quality of satellite remote sensing (RS) data is largely assumed to be guaranteed by the providers (Chapter 2.1), the acquisition of manner aircraft and drone data especially acquired for the MultiMiner project is examined in detail with regard to the quality.

The definition of the thematic accuracy assessment procedures for MultiMiner application demonstrations are the focus of Appendix 1 (SENSITIVE). Applications are defined as followed: mineral exploration, vegetation monitoring, water quality and AMD, ground moisture monitoring, integrated TSF monitoring, 3D mine site monitoring (consisting of dam stability monitoring and open pit stability monitoring) and dust monitoring. With the description regarding accuracy assessment, the collected in situ data and used EO data, a sampling design evaluation for each application is performed in order to use reasonable methods to determine the respective accuracies. The sampling strategies ensure the spatial and temporal representativeness of the collected data for calibration and demonstration of the varied multi-source and multi-scale data products. Appropriate thematic accuracy assessment concepts are developed for quantitative and non-quantitative attributes. Quantitative accuracy assessment measures (e.g., the contingency table and derived overall, producer’s and user’s accuracies, kappa, precision-recall curves, ROC curves and their AUCs, RMSE) for each data processing product varies greatly because of the nature of the data products produced by the project.

The goals of the report are to

- review the existing thematic accuracy assessment concepts and methods with a special emphasis on the approaches with a limited number of *in situ* data and accuracy assessment methods of weakly supervised and unsupervised classification results,
- and to make a plan how to apply them for the MultiMiner application demonstrations.

The MultiMiner project is funded by the European Union’s Horizon Europe research and innovation actions programme under Grant Agreement No. 10109137474

1. Principles of thematic accuracy assessment methods

1.1. Introduction to thematic accuracy assessment methods

Over the years, RS technology has become an important source of data to provide useful information for various applications in mineral exploration (Sabins, 1999; Bedell, 2004; Pour et al., 2021) and mine site monitoring (Pour et al., 2023; Li et al., 2015; Mckenna et al., 2020). The information derived from RS data can be divided into thematic maps or statistics. Machine learning methodologies have become common tools for remotely sensed data interpretation into lithological, mineral, alteration and structural maps (e.g., Shirmard et al., 2022). Similarly, environmental impacts, safety and mining operations can be monitored with RS for instance for dust emissions (Batbold et al. 2022), water quality (Modiegi et al., 2020), open pit stability (Song et al., 2023), dam stability (Lumbroso et al., 2020; Yan et al., 2024), acid mine drainage (Jackisch et al., 2018), surface mineralogy of tailings storage facilities (Shang et al., 2014) and vegetation status (McKenna et al., 2020). It is essential that the thematic maps derived from remotely sensed data are accurate because they can be used as a major of information for decision making in daily mining operations (Jensen and Jensen, 2012; Wickham et al., 2013). It is important to note that information extracted from remotely sensed data are never free of errors. Thus, maps derived from RS data must be presented together with their uncertainty estimates. It is also essential to identify the origins of the uncertainties and mitigate them as much as possible. Therefore, thematic maps should always be subjected to a detailed accuracy assessment before they are being used for planning and decision making in mineral exploration and mining (Congalton et al., 2019; Borengasser et al., 2007; Moud et al. 2021). The common steps for evaluating the accuracy of RS derived thematic information includes: 1) Identifying the expected accuracy level, 2) determining whether the target variable is discrete or continuous, 3) and specifying the sampling design and the framework (Jensen, 2009). The accuracy and reliability of the RS data, along with the products derived from it, are contingent on the specific context of their application. In terms of widely recognized and frequently implemented best practices, we recommend referring to the seminal works by Congalton et al. (2019) and Olofsson et al. (2014).

The MultiMiner project is funded by the European Union’s Horizon Europe research and innovation actions programme under Grant Agreement No. 10109137474

1.2. Quantitative image classification accuracy assessment

1.2.1. Confusion matrix

To assess the accuracy of RS thematic maps, it is common to overlay in situ reference data on the classification image derived from unsupervised or supervised classification. In the confusion matrix table, the rows indicate the classification labels, and the columns indicate the labels from in situ data (Table 2). To perform thematic accuracy assessment using a confusion matrix, independent test data should be compiled of in situ measurements or observations. A classification map derived from remotely sensed data are compared pixel by pixel with the test data. The result of this step is a confusion matrix table that summarizes the agreement or disagreement between test data and classification. The confusion matrix is not effective at properly evaluating the classified map properties, unless the samples were collected using a random sampling design (Jensen, 2009).

Table 2. Multiclass confusion matrix. UA = user's accuracy and PA = producer's accuracy, N = number of samples.

Classes	**C_1**	**C_2**	**C_3**		**n**	**Row total**	**UA**
C_1	$C_{1,1}$	$C_{2,1}$	$C_{1,3}$	....	$C_{1,n}$	C_{1+}	$C_{1,1}/ C_{1+}$
C_2	$C_{2,1}$	$C_{2,2}$	$C_{2,3}$	....	$C_{2,n}$	C_{2+}	$C_{2,2}/ C_{2+}$
C_3	$C_{3,1}$	$C_{2,3}$	$C_{3,3}$	....	$C_{3,n}$	C_{3+}	$C_{3,3}/ C_{3+}$
.	.	.	.	....	....	....	
n	$C_{n,1}$	$C_{n,2}$	$C_{n,3}$	....	$C_{n,n}$	C_{n+}	$C_{3,3}/ C_{n+}$
Column total	C_{+1}	C_{+2}	C_{+3}	....	C_{+n}	N	
PA	$C_{1,1}/ C_{+1}$	$C_{2,2}/ C_{+2}$	$C_{3,3}/ C_{+3}$	....	$C_{n,n}/ C_{+n}$		

1.2.2. Classification accuracy metrics

Producer's and user's accuracies, kappa coefficient

The information stored in accuracy assessment table is used to compute the confusion matrix metrics such as overall accuracy, producer's and user's accuracies, and kappa coefficient value. Dividing the number of correctly classified pixels by all samples/pixels in the confusion matrix determines the overall accuracy of classification (Congalton et al., 2019; Story et al., 1986). The Cohen's kappa coefficient determines whether the classification map and the test data are in agreement. The kappa coefficient value is within the range of 0 and 1. When the value is close to 1, it is indicative of a reliable image classification result and great agreement between the classified categories and the reference sample. The kappa values that are close to zero indicate that there is little or no agreement between the classified targets and referenced samples (Congalton et al., 2019; Stehmann et al., 1998). There is an added value of using kappa compared to overall accuracy since it considers the possibility of the agreement occurring by chance, and can is thus a more realistic accuracy measure.

Class-related accuracies are known as user's and producer's accuracies. Dividing the number of correct pixels in the class by the total number of pixels in the corresponding row or column is the method used to calculate the accuracy of individual classes. The producer's accuracy or recall refers to the total number of correctly classified pixels in a class divided by the total number of test data pixels in that class. This metric is used to determine how many positive test pixels were accurately classified as positive. The user's accuracy is calculated by dividing the total number of correctly classified pixels in a class by the total number of pixels that were classified in that individual class. The likelihood that a pixel classified in the classification image actually represents that class on the ground is determined by the user's accuracy or precision (Foody, 2004; Jensen, 2009).

ROC curves and their AUC

To evaluate the quality and performance of a binary classification, which outputs a continuous variable i.e. propability of belonging to a class, the two-dimensional Receiver Operating Characteristics (ROC) curve and its Area Under the Curve (AUC) value are commonly used (see example in **Error! Reference source not found.**). Different threshold settings are used to create the ROC curve plot by plotting the true positive rate (TPR) against the false positive rate (FPR). Correctly identified positive samples by the model are the TPR, which is the proportion of actual positive samples. The FPR, on the other hand, is the ratio of actual negative samples that are incorrectly identified as positive by the model. The X axis represents the FPR, and the Y axis represents the TPR. A better classification model has curves that are more top-left-side and a higher TPR and lower FPR for each threshold. A random classifier is represented by the diagonal line, while a perfect classifier with TPR=1 and FPR=0 can be found in the top-left corner.

The AUC is a measure of the model's overall performance, identifying how the model will rank a randomly chosen positive sample higher than a randomly chosen negative sample. If a model is perfect, its AUC would be 1, while if it is random, its AUC would be 0.5. When comparing the performance of multiple models, the AUC provides a single value that determines the overall performance of a model (Bradley, 1997).

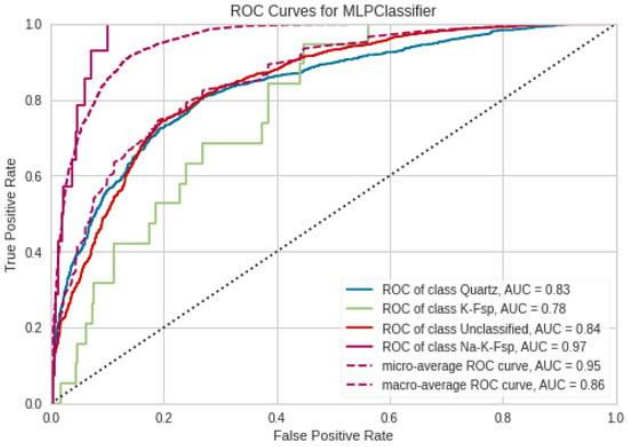

Figure 1. Example of a ROC curves for accuracy assessment of quartz and K-feldspar identified from drill core hyperspectral shortwave infrared imaging data (Rotem et al., 2023).

The MultiMiner project is funded by the European Union's Horizon Europe research and innovation actions programme under Grant Agreement No. 10109137474

Precision-recall curves

A precision-recall curve is a substitute for the ROC curve. The y-axis of this curve shows the precision (positive predictive value) versus the recall (sensitivity) for different thresholds in the x-axis. Precision-recall is a valuable indicator of the effectiveness of prediction when the classes are highly imbalanced. Using the precision-recall curve, it is possible to see the trade-off between precision and recall at different thresholds. Both high recall and high precision can be attributed to a high AUC, with high precision being linked to a low false positive rate and high recall being linked to a low false negative rate. If both scores are high, then the classifier has an ability to produce accurate results (high precision), with a high majority of positive results (high recall) (Hackeling, 2017; Pedregosa et al., 2011).

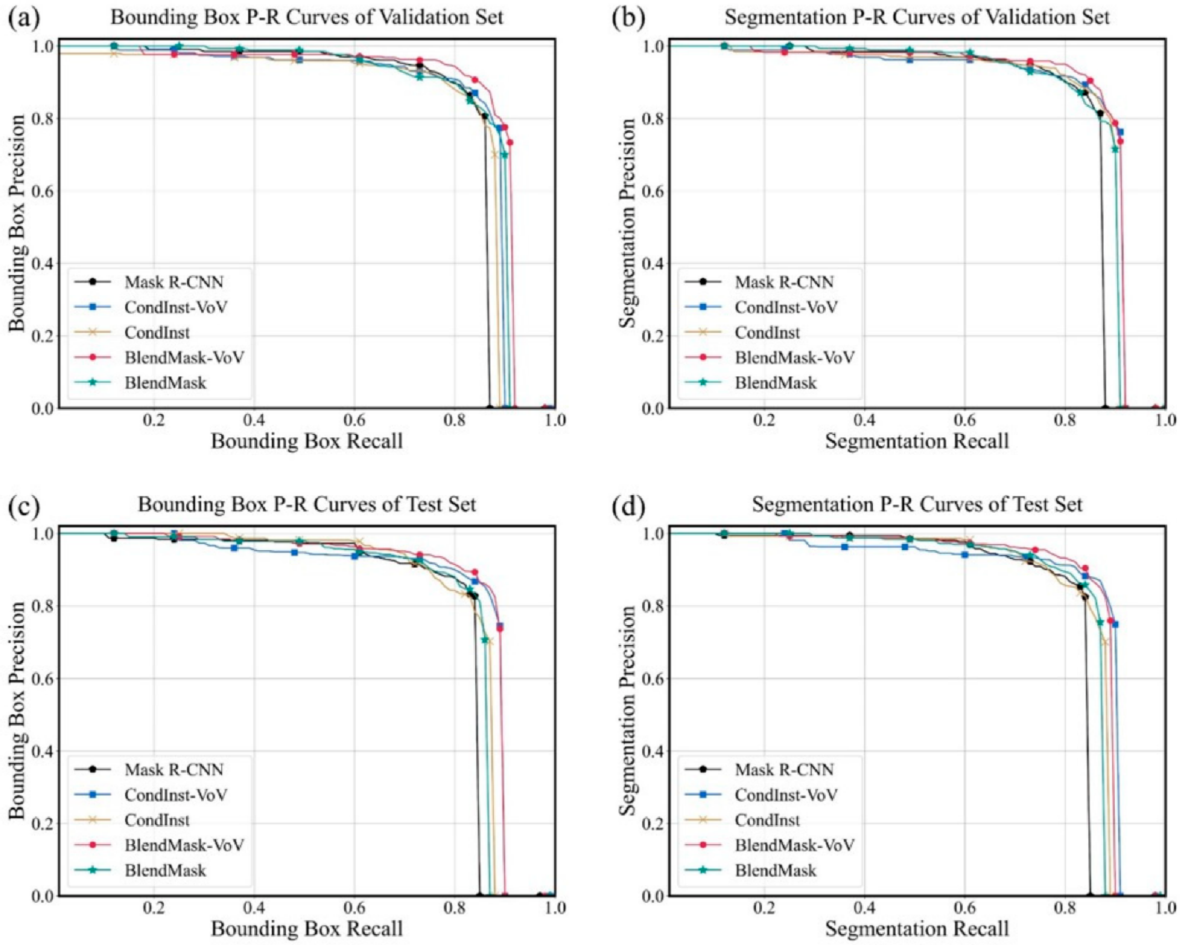

Figure 2. Example precision-recall curves for assessment of mapping valley fill faces by mountain top removal in coal mining from LiDAR data in (Maxwell et al., 2020).

The MultiMiner project is funded by the European Union's Horizon Europe research and innovation actions programme under Grant Agreement No. 10109137474

F1-score

F1-score, also known as f-measure, equally considers the contribution of precision and recall values. F1-score is the harmonic mean of precision and recall values. In case of a multi-class image classification, the F1-score values for each class are calculated based on a one-vs-all technique. The F1-score ranges between 0 to 1, and a higher value specify a better performance of model in a particular class prediction. A high F1 score is commonly associated with a balanced performance that demonstrates the ability of a model to achieve both high precision and recall. If the F1 score is low, it suggests that the model is struggling to achieve a balance between recall and precision (Goutte and Gaussier, 2005). The performance of the model cannot be accurately determined by calculating F1-score based on an arithmetic mean when the precision and recall values have significantly variable values. The F1-score can avoid overestimation and achieve balanced measurements, as the harmonic mean emphasizes the reciprocal of values (Sokolova et al., 2006). The F1-score equation is as follows.

$$F1 - score \; = 2 \; * \frac{\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}}$$

Intersection over union (IoU)

Intersection over union (IoU) is an accuracy measure to determine to what extent the prediction region overlaps with the reference data. This metric is usually used in spatial object detection and segmentation and ranges between 0 and 1. This performance metric quantifies the degree of overlap between pixels in predicted a bounding box or segmented region and the labelled bounding box or segmented ground truth region (Maxwell et al., 2021). The IoU calculation is done by determining the ratio of the overlap between the predicted and ground truth bounding boxes to the union of both bounding boxes. Perfect overlap is indicated by a value of 1, and no overlap is indicated by a value of 0. The equation for IoU is provided below.

$$IoU = \frac{Intersection\ of\ \text{predicted and ground truth bounding boxes}}{\text{Union of predicted and ground truth bounding boxes}}$$

1.2.3. Accuracy metrics for regression results

A regression method can be used to estimate or predict a response variable by using a set of covariates. The targets in a classification problem are categorical, while in a regression task the response variable is continuous. The regression model is created or conditioned based on a set of input variables that correspond to known responses, such as done when constructing classification models. Estimating or predicting ground soil moisture using predictor data extracted from a remotely sensed imagery is an example of regression modelling. The ground soil moisture classes are not relevant in this case, as the response is a continuous variable. Several accuracy measure metrics such as Mean Squared Error, Mean Absolute Error, Root Mean Squared Error, and Coefficient of determination, i.e. R^2, are commonly used to evaluate the performance of machine learning models in regression problems.

The MultiMiner project is funded by the European Union's Horizon Europe research and innovation actions programme under Grant Agreement No. 10109137474

Mean Absolute Error (MAE)

The mean absolute error (MAE) is defined by the average deviation of absolute actual and predicted values in the dataset. Additionally, the MAE is a measure of absolute discrepancies between the actual and predicted values. The average of the residuals in the dataset is taken into account by this measure.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|$$

Mean Squared Error (MSE)

The Mean Squared Error (MSE) describes the average difference in squared error values between the actual and predicted values in the dataset. The variance of the residuals is measured by MSE.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Root Mean Squared Error (RMSE)

Root Mean Squared Error defines the standard deviation of residuals and is calculated using the square root of MSE. The RMSE measures whether the predicted values from a model are in accordance with the actual observed values in the dataset.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$

One plus/minus standard deviation from mean

The measurement of plus/minus one standard deviation from mean describes the range of values obtained by addition/ subtraction one standard deviation to/from the mean.

It represents a value relatively near to the center of distribution. For a normal distribution, approximately 68% of the data falls within this range, indicating that these values are near the average and represent typical variations from the mean.

Ratio of Performance to Deviation (RPD)

The ratio of Performance to Deviation (RPD) is defined as the ratio of standard deviation of the sample data (SD) to the standard error of prediction (SEP). This metric is used to evaluate the accuracy and reliability of predictive models. Higher RPD values indicate better model performance, with values above 2 generally considered good and values above 3 considered excellent.

$$RPD = SD/SEP$$

Coefficient of determination (R-squared)

Coefficient of determination or R-squared (R^2) is statistical metric used to evaluate the goodness of fit of a regression model. This metric highlights the overall effectiveness of the regression model and measures how much of the variation in the dependent variable is due to the independent variables in the model.

The MultiMiner project is funded by the European Union's Horizon Europe research and innovation actions programme under Grant Agreement No. 10109137474

In summary, if the MAE, MSE, and RMSE values are low, the regression model will be accurate (Jensen, 2009). However, it is desirable to have a high R-squared value. In addition, RMSE and R-squared quantify how well a linear regression model fits a dataset. The RMSE indicates the accuracy of a regression model in predicting the response variable values in absolute terms, while R- Squared determines how effectively the predictor variables can explain the variability in the response variable.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\widehat{y_i} - y_i)^2}{\sum_{i=1}^{n}(\bar{y} - y_i)^2}$$

In the abovementioned equations, y_i is the observed value, $\bar{y}$ is the mean of the observed values, $\widehat{y_i}$ is the model predicted value, and *n* is the number of samples.

1.3. Accuracy budget table

An accuracy budget table is a tool to track and manage the accuracy of measurements, calculations or processes within a project. It outlines the expected level of accuracy for each component in a process and helps ensure the achievement of accuracy requirements. The budget table may include information such as the desired accuracy level, allowable tolerances, measurement uncertainties and error sources (Moud et al., 2021).

1.4. Sample size

Identifying the optimal number of test samples for each class is an important step before calculating the overall accuracy as well as class specific accuracies. In order to conduct an accuracy assessment for a classification result it is usual to adopt a design-based framework (Stehmann and Foody, 2019), which assist in determining suitable sample size for accuracy evaluation (Foody, 2004). There are two equations or approaches that can be used to identify the sample size: 1) binomial distribution or the normal approximation, and 2) multinomial distribution that can provide a sample size estimation when multiple classes are included in the thematic image classification result (Congalton and Green, 2019; Pour et al., 2023). When the thematic map is based on binary classification, it is recommended to use the binomial function, which is a specific form of the multinomial function (Adami et al., 2012; Stehman, 2012). Thus, in a binary classification, the probability of correctly classified labels can be calculated based on the binomial distribution sampling design (Radoux et al., 2020). The binomial distribution is transformed into a multinomial distribution when more than two response variables/targets are characterized (Park, 2013). Therefore, multinomial distribution is recommended to evaluate multi-class classification maps with the required sample size (Congalton and Green, 2019; Jensen, 2009).

The following equation demonstrates the use of binomial probability theory for specifying the sample size for land-use classification result, which was suggested by Fitzpatrick-Lins (1981):

$$N = \frac{Z^2(p)(q)}{E^2}$$

The MultiMiner project is funded by the European Union’s Horizon Europe research and innovation actions programme under Grant Agreement No. 10109137474

where p is the percentage of accuracy that the entire map is expected to achieve. q = 100 – p, E is the acceptable error, and the standard normal deviation of 1.96 corresponds to Z = 2 with the 95% two-sided confidence level.

Whilst Anderson, (1976) recommends obtaining approximately 85% overall accuracy for thematic classification map, Pontius Jr & Millones, (2011) emphasised that it would not be necessary to link the achieved accuracy level to 85% because research question or the study influence on the obtained accuracy.

To assess classification accuracy of a multi-class thematic map, some analysts use equations based on a multinomial distribution to determine the required total sample size (*N*). The following equation demonstrates the sample size measured by multinomial distribution approach (Congalton and Green, 2019):

$$N = \frac{B\Pi_i(1-\Pi_i)}{{b_i}^2} \text{ (Anderson et al., 1976)}$$

where Π_i demonstrates the proportion of the population in the *i*th class among all classes that is closest to 50%, b_i indicates the expected precision (e.g., 5%) for this class. B specifies the upper percentile of the chi square (χ^2) distribution, and k indicates the number of classes.
When choosing a stratified sampling, a sample size of at least 30 units per strata is recommended to achieve an acceptable standard error (Olofsson et al., 2014; Lohr, 1999).

Because of the core idea of MultiMiner project to apply weakly supervised methods using as low number of in situ measurements as possible, a smaller sample size should be considered and accepted if accuracy is sufficient.

1.5. Sampling design methods

After identifying the total number of samples and the number of samples for each class, it is important to design a sampling plan including the geographic locations of the ground truth samples. In order to have a reliable accuracy assessment approach, it is vital to collect samples randomly to avoid the unnecessary bias. It is worth noting that bias in the thematic accuracy assessment can cause the confusion matrix to overestimate or underestimate the actual accuracy of the produced classification map. Conducting different sampling design schemes can be done using random sampling methods. Hence, it is essential to choose a suitable random sampling design. Different variance equations are required for different sampling schemes due to their respective sampling models. These widely used sampling design methods are:

- Simple random sampling
- Systematic sampling
- Stratified random sampling
- Cluster sampling
- Post-stratification

The MultiMiner project is funded by the European Union's Horizon Europe research and innovation actions programme under Grant Agreement No. 10109137474

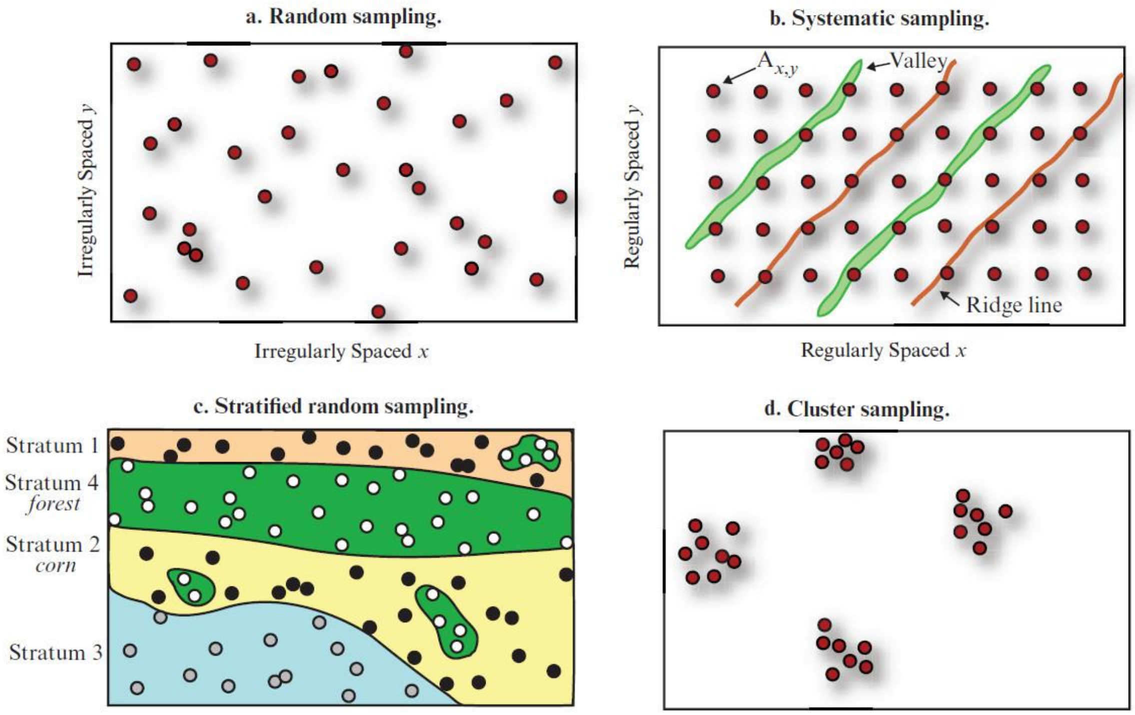

Figure 3. Illustrations of sampling designs methods: a) random sampling; b) systematic sampling; c) stratified random sampling; and d) cluster sampling. Sub-figures adopted from (Jensen, 2009).

Simple Random Sampling

Simple random sampling is a basic sampling technique which selects individuals in a population with equal probabilities. Furthermore, the choice of one sample is not a factor in selecting any other sample. This technique has the potential to undersample classes with limited sample numbers, which is one of its main disadvantages (van Oort, 2007).

Systematic Sampling

Systematic sampling is a statistical sampling technique that assures that random samples are selected from the population, with each sample selected based on a fixed sampling interval. The sampling result of simple random sampling and systematic sampling could lead to a small sample size for a rare class or classes unless the overall number of samples in the population is very large (Jensen, 2009).

Stratified Random Sampling

Stratified random sampling is a probability sampling technique than involves dividing the population into homogenous groups, known as strata. In stratified random sampling, the strata are formed based on shared attributes of the members or on characteristics which corresponds to a RS derived thematic map. The random selection of data from an entire population is done in a stratified manner, which means that unlike in simple random sampling, each possible sample is equally likely to occur. The main benefit of stratified random sampling is that it ensures samples being assigned to every strata even to ones with a limited spatial extent. A ready thematic map is necessary to allocate samples to the different land-cover strata, which is a drawback of stratified random sampling. The remote sensing data acquisition and ground reference test collection can thus seldomly occur on the same day (Jensen, 2009).

The MultiMiner project is funded by the European Union's Horizon Europe research and innovation actions programme under Grant Agreement No. 10109137474

Cluster sampling

Cluster sampling involves using at least two sample sizes. If two sampling units are utilized, the large unit is referred to as the primary sampling unit (PSU). Pixels, block of pixels, linear cluster of pixels (Edwards Jr et al., 1998), an aerial photograph, or satellite image are all viable sources of the PSU. The second sample size is smaller, and this is called a secondary sampling unit (SSU). Initiating the process first involves selecting the sample of PSUs. One-stage cluster sampling is performed when all SSUs within each sampled PSU are chosen. However, a two-stage cluster sampling is performed when a subsample of SSUs is chosen. The primary reason for using cluster sampling is to decrease the cost of obtaining data in the field (Warner et al., 2009).

Post-stratification

Post-stratification is a procedure of dividing the observations into group of strata after completion of the sampling process. Post-stratification may apply to a collected in situ samples prior to availability of a remote-sensing-based stratification. Post-stratification can be performed on any sample set, but the post-stratified estimation can only be useful if the size or proportion of each stratum relative to the population is known (Gregoire and Valentine, 2007). Forestry surveys can be facilitated by post-stratification because field sampling and analysis, as well as interpretation of RS data, can be carried out independently (Köhl et al., 2006).

1.6. Dealing with class imbalance or rare classes

In some cases, the class of interest on a study site have rare occurrence, i.e. it is limited in geographical extent. An example in mineral exploration would be a mineralized body or in mine site monitoring limited dispersion of AMD. In such circumstances the number of labelled data can be greatly imbalanced with only a few labels representing the class of interest. Also, if the coverage of the class is very small relative to the large pixels of the RS data, dense labeling could not applied. Secondly, extensive collection of ground truth data may not be allowed e.g. by the landowner, government agencies or, may not be possible, for example, for industrial safety reason or exceptionally rugged terrain.

In these cases of missing or extremely limited ground truth data, the use of abovementioned quantitative accuracy assessment methods is not possible. Then, alternative accuracy assessment approaches should be employed. The following examples of mineral identification demonstrate what can be done in the absence of quantitative accuracy assessment. Chirico et al. (2023) utilized PRISMA satellite hyperspectral data to delineate hydrothermal dolomitization and supergene alteration. The authors examined the spectral features of PRISMA hyperspectral data against reference spectral reflectance features and created mineral maps by applying band ratios of satellite hyperspectral data. The authors just made a qualitative comparison of the spectral signatures of the PRISMA data to laboratory reflectance spectra without a quantitative assessment of the final mineral map. In a study by Bierwirth et al. (2002), HyMap airborne hyperspectral data was utilized to map alteration minerals which are linked to gold prospects. The authors only validated the mineral abundance maps qualitatively by analysing the portable infrared mineral analyser (PIMA) data of field samples. In addition, during rock analysis the presence of anomalous gold was confirmed was available and it did not allow employing a fully quantitative accuracy assessment (Althnian et al., 2021). However, comparison of the areas between the ground detected and EO data interpreted class extents may be possible.

The MultiMiner project is funded by the European Union's Horizon Europe research and innovation actions programme under Grant Agreement No. 10109137474

1.7. Strategies to mitigate small number of reference in situ data

Unlike supervised machine learning (ML) models, the weakly supervised methods (Katsaragakis et al., 2020) require dealing with a small number of in situ reference data. The small amount of available data affects model training procedures but should not change accuracy assessment principles. A strict division between the definitions of a small or large number of reference data is not to be performed. In practice the small amount of reference data available brings additional constraints to accuracy assessment. In addition, it should be considered that an increase in system complexity, observation rarity and area coverage correlate positively with the small data problem (small data problem: when a dataset has a small sample size, covers the distribution poorly, or when the sample number will probably not be sufficient to find meaningful features with ML (Safonova et al., 2023).

The advantages of small datasets are a fast training time (Safonova et al., 2023; Althnian et al., 2021), reduced memory requirements (Safonova et al., 2023; Katsaragakis et al., 2020, Wang et al., 2023) and small operational costs (Safonova et al., 2023). Disadvantages of small datasets are the lack of generalisability and transferability due to overfitting (Liu et al., 2017). Additionally, the dataset may be biased (Althnian et al., 2021).

Part of a full independent unbiased accuracy assessment procedure for ML models is to carefully define the split between training set (*TRAIN)* for training the weakly-supervised model and testing test (*TEST)* for accuracy assessment (Figure 4), to ensure that both datasets are balanced and representative. The inherent challenge of weakly supervised method lies in designing, implementing and training a model that can, with a fraction of the training set (or coarser/lower quality reference labels), reach a similar accuracy on the independent testing set.

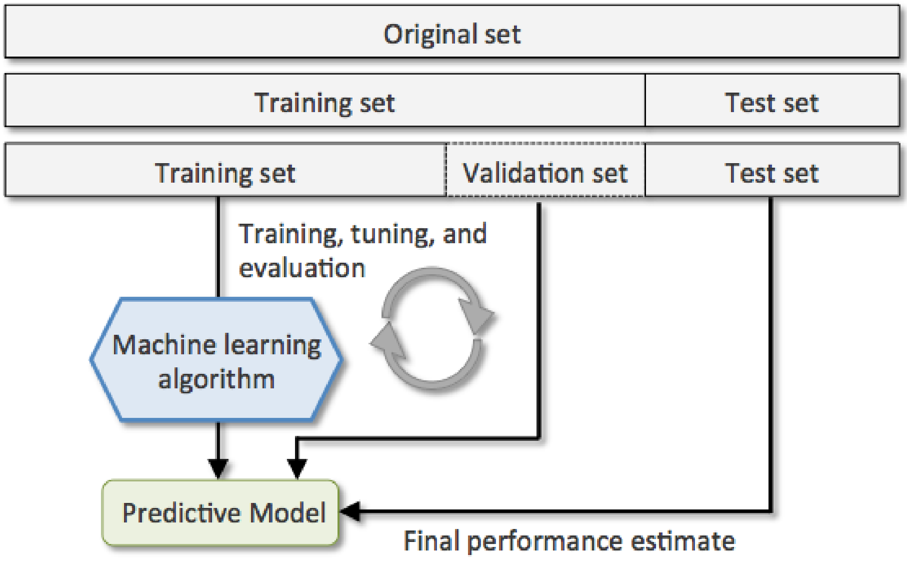
Original set
Training set
Test set
Training set
Validation set
Test set
Training, tuning, and evaluation
Machine learning algorithm
Predictive Model
Final performance estimate

Figure 4. Validation procedure for machine learning algorithm (NaveganTeX, 2019).

An approach to deal with a low number of training data in the model training phase is the technique of transfer learning when a pre-trained model is based on a large and relevant dataset is used. Improving performance and generalisability with reduced data requirements is the advantage. Whereas the disadvantages are risking the transformation into a different domain and an unnecessarily model upsizing. Another technique could be the use of a mix of supervised and

The MultiMiner project is funded by the European Union's Horizon Europe research and innovation actions programme under Grant Agreement No. 10109137474

unsupervised learning to enhance generalisability. The few-shot learning as a type of meta-learning which is used to teach the model to generalise for new tasks with only a few samples per class. Other less common applications such as process-aware, ensemble or active DL technique are summarised by Safonova et al. (2023).

In the model training phase, the splitting of the training-testing sets for model validation can be done in two ways:

1. The repeated random subsampling validation (Figure 5 A), in which the model evaluation is repeated while subsampling is used to estimate model performance instability, is not commonly recommended because spatial and temporal data comes with high autocorrelation.
2. The k-fold cross validation (Figure 5 B) is a better choice when dealing with small datasets (Safonova et al., 2023). Especially Leave-One-Out Cross-Validation (LOOCV) is recommended in case of very small dataset. The special cases of k-fold cross validation are:
   - Stratified cross-validation. In the stratified cross-validation, the distribution of classes in each fold is almost the same as in the original dataset (Diamantidis et al., 2000) (Figure 6). This method deals better with spatial autocorrelation and overfitting, but usually requires more samples.
   - Leave *p* out cross validation. In leave *p* out cross validation, *p* describes the number of testing data left out randomly (usually smaller than the size of one fold in k-fold CV).
   - Leave-One-Out Cross-Validation (LOOCV). In LOOCV, only one observation is left out from the training set (Cheng et al., 2017).

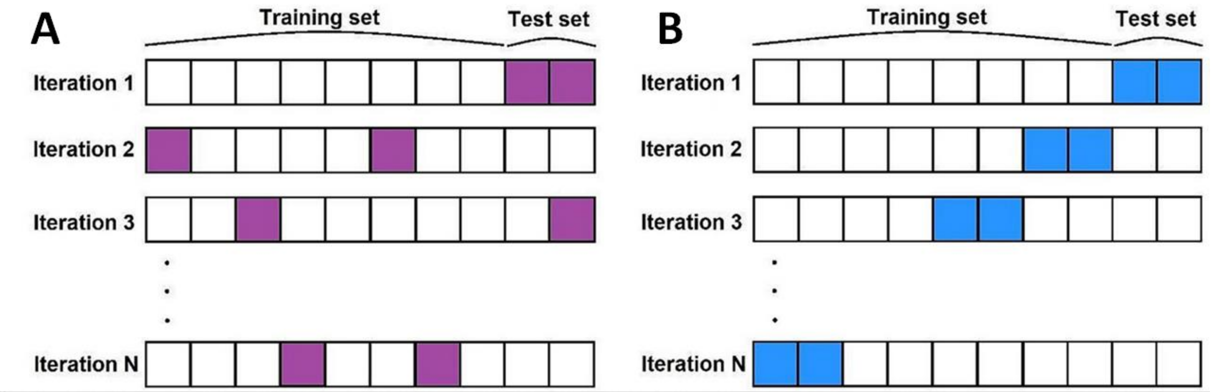

Figure 5. Model validation schemes. ***A*** *Validation by repeated random sampling, in which the model evaluation is repeated while subsampling is performed to estimate the instability of model performance.* ***B*** *K-fold cross validation, in which the data is first divided into groups and then the model performance is evaluated recursively on the basis of one of the groups. If the grouping is based on the spatial coordinates, this is referred to as spatial cross-validation. This figure is adapted and modified from Safonova et al. 2023, CC BY 4.0 license (https://creativecommons.org/licenses/by/4.0/).*

The MultiMiner project is funded by the European Union's Horizon Europe research and innovation actions programme under Grant Agreement No. 10109137474

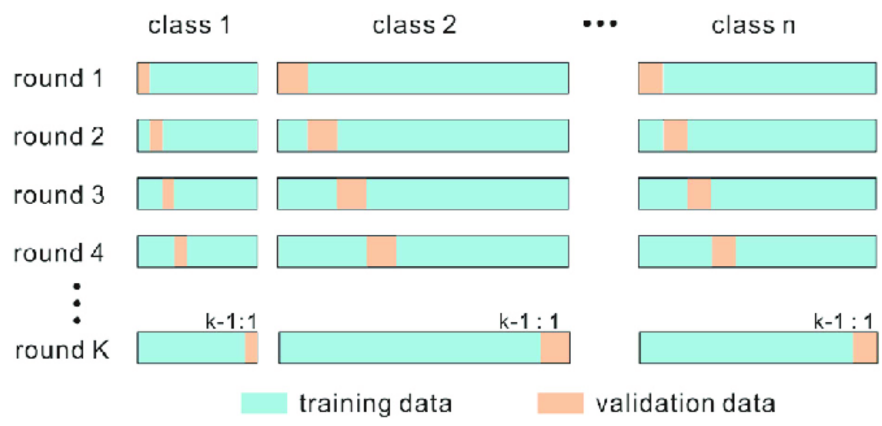

Figure 6. Schematic diagram of stratified k- fold cross-validation. This figure is adapted from Duan 2023, used under CC BY 4.0 license (https://creativecommons.org/licenses/by/4.0/).

Spatial-based methods, which use patches or windows as inputs such as deep learning methods, should be applied with care when only a small amount of reference data is available. This is especially true if samples are spatially close to each other. When the distance between a sample in training set and a sample in test set is shorter than the size of an input patch, then the patches used in training and testing overlap spatially. This leads to overfitting, as some pixel values from the test patch (supposedly independent) have already been "seen" by the model (in the input patch of the nearby training sample). This spatial overlap between training and testing patches leads to over optimistic accuracy assessment, and a loss of generalization ability of the model (Molinier and Kilpi, 2019).

Several alternative sampling schemes have been proposed to reduce these effects, but only strict spatial separation of training and test samples with a buffer larger than the size of input patches can avoid the overlap effect (Molinier and Kilpi, 2019). If a larger sample area cannot be used, then training and test sets should be spread at least so that training patches and test patches are spatially disjoint. Some authors suggest even stricter conditions to avoid spatial autocorrelation or recommend to be aware of the spatial area of applicability for a model trained on a small geographic area (Meyer et al., 2021).

The MultiMiner project is funded by the European Union's Horizon Europe research and innovation actions programme under Grant Agreement No. 10109137474

2. Quality of EO data

Regarding the accuracy assessment a closer look at EO data (used for developing the MultiMiner applications) quality should be performed, as EO data is a basic information tool. Details on EO data and pre-processing used for developing MultiMiner applications are presented in deliverable 3.1 “Timely mine site monitoring algorithms design and input data requirements”.

The EO data used for developing MultiMiner applications are listed in Table 3.

Table 3. EO data used for developing MultiMiner applications. Table will to be updated during the project.

Platforms	**Sensor/Instrument and data**			
	Multispectral	**Hyperspectral**	**SAR**	**Other**
Spaceborne platform	Copernicus Sentinel-2, MODIS, Planet, WorldView-3	EnMAP, PRISMA	Copernicus Sentinel-1, TerraSAR-X	Sentinel-5
Manned aircraft platform		VNIR/SWIR/TIR (AisaFenix/AisaOWL)		Open access LiDAR DEMs
Drone platform	Photogrammetric, drone RGB+Multispectral	VNIR-SWIR camera (HySpex Mjolnir VS-620)	Multi-band Interferometric SAR system (Explorer RD350)	Gamma spectrometer, LiDAR, FLIR (Teledyne)
Close-range sensing	Imaging spectrometers: VNIR, SWIR, LWIR with AisaFENIX and AisaOWL (Specim, Spectral Imaging Ltd.)			

2.1. Quality of satellite remote sensing data

EO data provider take various measures to ensure the quality of their EO products. ESA has established an ongoing Earthnet Data Assessment Project (EDAP) as a collaboration between ESA and NASA to ensure the quality of EO data acquisition (ESA, 2024; Hunt, 2022). The approach for assessing the quality of data products is based on the Quality Assurance Framework for Earth Observation (QA4EO) principle (QA4EO Task Team, 2010). In addition to best practice guidelines, documents are also provided that describe the methods, algorithms, reference datasets and tools for assessing of the data quality (ESA, 2024).

The MultiMiner project is funded by the European Union’s Horizon Europe research and innovation actions programme under Grant Agreement No. 10109137474

2.2. Quality of manner aircraft and drone data

A detailed description of UAV data acquisition is available in the Field Guidebook (D4.1 Part 1, Chapter *3.3 UAV data acquisition*). The following chapter briefly outlines the quality of manned aircraft and drone data.

For UAV data acquisition a number of variables are involved and some of them are not within the scope of the operator´s control. The first control of the data should be carried out as early as possible, as it is very likely that not all initial parameter settings are optimal and need adjustments or recalibration. During the flight procedure quality control is limited (TRuStEE, 2018).

Operational safety is ensured by annual maintenance by the manufactur, preflight and postflight checks, as well as documented battery management. Flight planning and flight procedure are made under the appropriate categories of the European Union Aviaton Safety Agency (EASA) guidelines (EASA, 2017).

To ensure good data quality, the following steps of a protocol must be observed. For flight planning the site and risk assessments are just as important as the planning of the flight and the required equipment. The flight preparation should include examination of the weather forecast, valid site permissions and notifications and preparation of the equipment. The flight procedure includes in situ site and weather assessment, pre-flight controls, the flight itself and post-flight controls. Ancillary measurements to be taken into account are e.g. ground control points. Tasks to perform after flight should include reporting the taken measurements, reporting the field work and data, and quality control and validation of the product (TRuStEE, 2018).

Data should be backed up on a server and entered into a metadata database.

In the event of an accident, a possible data back-up and a report to EASA should be performed. To support drone operations, a risk-based 10 step approach called Specific Operations Risk Assessment (SORA) is available (EASA, 2024).

The MultiMiner project is funded by the European Union’s Horizon Europe research and innovation actions programme under Grant Agreement No. 10109137474

3. References

Adami, M., Mello, M. P., Aguiar, D. A., Rudorff, B. F., & de Souza, A. F. (2012). A web platform development to perform thematic accuracy assessment of sugarcane mapping in South-Central Brazil. *Remote Sensing, 4*(10), 3201-3214.

Althnian, A., AlSaeed, D., Al-Baity, H., Samha, A., Bin Dris, A., Alzakari, N., . . . Kurdi, H. (2021). Impact of Dataset Size on Classification Performance: An Empirical Evaluation in the Medical Domain. *Applied sciences, 11*(2). doi:https://doi.org/10.3390/app11020796

Anderson, J., Hardy, E., Roach, J., & Witmer, R. (1976). *A land use and land cover classification system for use with remote sensor data* (Vol. 671). U.S. Geological Survey Circular.

Batbold, C., Yumimoto, K., Chonokhuu, S., Byambaa, B., Avirmed, B., Ganbat, S., . . . Matsuki, A. (2022). Spatiotemporal Dispersion of Local-Scale Dust from the Erdenet Mine in Mongolia Detected by Himawari-8 Geostationary Satellite. *Scientific Online Letters on the Athmosphere, 18*, 225-230. doi:https://doi.org/10.2151/sola.2022-036

Bedell, R. (2004). Remote Sensing in Mineral Exploration. *SEG Discovery, 58*, 1-14. doi:https://doi.org/10.5382/SEGnews.2004-58.fea

Bierwirth, P., Huston, D., & Blewett, R. (2002). Hyperspectral Mapping of Mineral Assemblages Associated with Gold Mineralization in the Central Pilbara, Western Austrailia. *Economic Geology*, 819-826. doi:https://doi.org/10.2113/gsecongeo.97.4.819

Borengasser, M., Hungate, W., & Watkins, R. (2007). *Hyperspectral Remote Sensing - Principles and Applications.* CRC Press.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learnin algorithms. *Pattern Recognition, 30*(7), 1145-1159.

Cheng, H., Garrick, D., & Fernando, R. (2017). Efficient strategies for leave-one-out cross validation for genomic best linear unbiased prediction. *Journal of Animal Science and Biotechnology, 8*.

Chirico, R., Mondillo, N., Laukamp, C., Mormone, A., Martire, D., Novellino, A., & Balassone, G. (2023). Mapping hydrothermal and supergene alteration zones associated with carbonate-hosted Zn-Pb deposits by using PRISMA satellite imagery supported by field-based hyperspectral data, mineralogical and geochemical analysis. *Ore Geology Reviews*. doi:https://doi.org/10.1016/j.oregeorev.2022.105244

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

Diamantidis, N., Karlis, D., & E., G. (2000). Unsupervised stratification of cross-validation for accuracy estimation. *Artificial Intelligence, 116*, 1-16. doi:https://doi.org/10.1016/S0004-3702(99)00094-6

Duan, X. (2023). Automatic identification of conodont species using fine-grained convolutional neural networks. *Front. Earth Sci., 10*. doi: https://doi.org/10.3389/feart.2022.1046327

The MultiMiner project is funded by the European Union's Horizon Europe research and innovation actions programme under Grant Agreement No. 10109137474

EASA. (2017). *Introduction of a regulatory framework for the operation of drones.* Retrieved from https://www.easa.europa.eu/en/document-library/notices-of-proposed-amendment/npa-2017-05#group-easa-downloads

EASA. (2024, 02/). *Specific Operations Risk Assessment (SORA)*. Retrieved from https://www.easa.europa.eu/en/domains/civil-drones-rpas/specific-category-civil-drones/specific-operations-risk-assessment-sora#Risk%20assessment%20of%20the%20intended%20operation%20%E2%80%93%20SORA

Edwards Jr, T., Moisen, G., & Cutler, D. (1998). Assessing map accuracy in a remotely sensed, ecoregion-scale cover map. *Remote Sensing of Environment*, 73-83. doi:https://doi.org/10.1016/S0034-4257(96)00246-5

ESA. (2020). *Newcomers earth observation guide.* Retrieved from https://business.esa.int/newcomers-earth-observation-guide#ref_9

ESA. (2024, 01/ 26). *Earth Online - EDAP Best Practice Guidelines*. Retrieved from https://earth.esa.int/eogateway/activities/edap/edap-best-practice-guidelines

ESA. (2024, 01/ 26). *Earth Online - EDAP Reference data, methods and tools*. Retrieved from https://earth.esa.int/eogateway/activities/edap/edap-reference-data-methods-tools

ESA. (2024, 01/ 26). *Earth Online-EDAP Overview*. Retrieved from https://earth.esa.int/eogateway/activities/edap

ESA. (2024, 01/ 26). *Quality Management and Assurance*. Retrieved from https://www.esa.int/Enabling_Support/Space_Engineering_Technology/Quality_Management_and_Assurance

Foody, G. (2013). Ground reference data error and the mis-estimation of the area of land cover change as a function of its abundance. *Remote Sensing Letters, 4*, 783-792.

Foody, G. (n.d.). *Accuracy assessment methods and challenges.* Retrieved from http://www.gofcgold.wur.nl/documents/jena08/1410_AA/05jena-accuracy.pdf

Foody, G. M. (2004). Thematic map comparison: Evaluating the statistical significance of differences in classification accuracy. *Photogrammetric Engineering and Remote Sensing, 70*(5), 627-634.

Goutte, C., & Gaussier, E. (2005). A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. *Advances in Information Retrieval: 27th European Conference on IR Research. Proceedings 27*, pp. 345-359. Santiago de Compostela, Spain: ECIR 2005.

Gregoire, T., & Valentine, H. (2007). *Sampling strategies for natural resources and the environment.* Chapman & Hall/CRC Taylor & Francis Group.

Hackeling, G. (2017). *Mastering Machine Learning with scikit-learn: Apply effective learning algorithms to real-world problems using scikit-learn.* Packt Publishing Ltd.

Hunt, S. (2022). *Earth Observation Mission Quality Assessment Framework.* ESA- EDAP. Retrieved from https://earth.esa.int/eogateway/documents/20142/37627/Mission-Quality-Assessment-Guidelines-v2.2.pdf/033c703e-02f8-d993-9859-560aeb61d2a0?t=1676561363850

The MultiMiner project is funded by the European Union's Horizon Europe research and innovation actions programme under Grant Agreement No. 10109137474

Jackisch, R., Lorenz, S., Zimmermann, R., Möckel, R., & Gloaguen, R. (2018). Drone-Borne Hyperspectral Monitoring of Acid Mine Drainage: An Example from the Sokolov Lignite District. *Remote sensing, 10*(385). doi:10.3390/rs10030385

Jensen, J. (2009). *Remote Sensing of the Environment: An Earth Resource Perspective.* Pearson Education.

Jensen, J. R., & Jensen, R. R. (2012). Introductory geographic information systems. Pearson Higher Ed.

Katsaragakis, M., Papadopoulos, L., Konijnenburg, M., Catthoor, F., & Soudris, D. (2020). Memory Footprint Optimization Techniques for Machine Learning Applications in Embedded Systems. *IEEE International Symposium on Circuits and Systems (ISCAS).* Seville, Spain: IEEE. doi:10.1109/ISCAS45731.2020.9181038

Köhl, M., Magnussen, S., & Marchetti, M. (2006). *Sampling Methods, Remote Sensing and GIS Multiresource Forest Inventory.* Springer.

Li, Y., Zhao, H., & Fan, J. (2015). Application of Remote Sensing Technology in Mine Environment Monitoring., *04008, 22*, p. 6. doi: https://doi.org/10.1051/matecconf/20152204008

Liu, B., Wei, Y., Zhang, Y., & Yang, Q. (2017). Deep Neural Networks for High Dimension, Low Sample Size Data., (pp. 2287-2293). Melbourne. doi:https://doi.org/10.24963/ijcai.2017/318

Lohr, S. (1999). Sampling: Design And Analysis. *CRC Press*.

Lumbroso, D., Roca, M., Petkovšek, G., Davison, M., Liu, Y., Goff, C., & Wetton, M. (2020). DAMSAT: An eye in the sky for monitoring tailings dams. *Mine Water and the Environment, 40*. doi:10.1007/s10230-020-00727-1

Maxwell, A., Pourmohammadi, P., & Poyner, J. (2020). Mapping the Topographic Features of Mining-Related Valley Fills Using Mask R-CNN Deep Learning and Digital Elevation Data. *Remote Sensing, 12*(3), 547. doi:10.3390/rs12030547

Maxwell, A., Warner, T., & Guillén, L. (2021). Accuracy assessment in convolutional neural network-based deep learning remote sensing studies- Part1: Literature review. *Remote Sensing, 13*(13), 2450. doi:https://doi.org/10.3390/rs13132450

McKenna, P., Lechner, A., Phinn, S., & Erskine, P. (2020). Remote Sensing of Mine Site Rehabilitation for Ecological Outcomes: A Global Systematic Review. *Remote sensing, 12*(3535). doi:https://doi.org/10.3390/rs12213535

Meyer, H., & E., P. (2021). Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Ecology and Evolution*, 1620-1633. doi:https://doi.org/10.1111/2041-210X.13650

Modiegi, M., Rampedi, I., & Tesfamichael, S. (2020). Comparison of multi-source satellite data for quantifying water quality parameters in a mining environment. *Journal of Hydrology, 591*(125322). doi:https://doi.org/10.1016/j.jhydrol.2020.125322

Molinier, M., & Kilpi, J. (2019). Avoiding Overfitting When Applying Spectral-Spatial Deep Learning Methods on Hyperspectral Images with Limited Labels. *IEEE International Symposium on Geoscience and Remote Sensing (IGARSS) .*

The MultiMiner project is funded by the European Union's Horizon Europe research and innovation actions programme under Grant Agreement No. 10109137474

Moud, F., Ruitenbeek van, F., Hewson, R., & Meijde van der, M. (2021). An approach to accuracy assessment of ASTER derived mineral maps. *Remote Sensing*. doi:doi.org/10.3390/rs13132499

NaveganTeX. (2019). *Cross Validated.* Retrieved from https://stats.stackexchange.com/questions/410118/cross-validation-vs-train-validation-test

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

Park, K. (2013). Generating Thematic Maps from Hyperspectral Imagery Using a Bag-of-Materials Model. *The Ohio State University*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *12*, 2825-2830.

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.

Pour, A., Parsa, M., & Eldosouky, A. (2023). Introduction to mineral exploration. *Geospatial Analysis Applied to Mineral Exploration*, 1-16. doi:doi.org/10.1016/B978-0-323-95608-6.00002-0

Pour, A., Ranjbar, H., Sekandari, M., El-Wahed, M., Hossain, M., Hashim, M., . . . Muslim, A. (2023). Remote sensing for mineral exploration. *Geospatial Analysis Applied to Mineral Exploration*, 17-149. doi:doi.org/10.1016/B978-0-323-95608-6.00002-0

Pour, A., Zoheir, B., Pradhan, B., & Hashim, M. (2021). Editorial for the Special Issue: Multispectral and Hyperspectral Remote Sensing Data for Mineral Exploration and Environmental Monitoring of Mined Areas. *Remote sensing, 13*(3), 519. doi:https://doi.org/10.3390/rs13030519

QA4EO Task Team. (2010). *Quality Assurance for Earth Observation Principles.* Retrieved from http://qa4eo.org/docs/QA4EO_Principles_v4.0.pdf

Radoux, J., Waldner, F., & Bogaert, P. (2020). How response designs and class proportions affect the accuracy of validation data. *Remote Sensing, 12*(2), 257.

Rotem, A., Vidal, A., Pfaff, K., Tenorio, L., Chung, M., Tharalson, E., & Monecke, T. (2023). Interpretation of Hyperspectral Shortwave Infrared Core Scanning Data Using SEM-Based Automated Mineralogy: A Machine Learning Approach. *Geosciences, 13*(7), 192. doi:https://doi.org/10.3390/geosciences13070192

Sabins, F. F. (1999). Remote sensing for mineral exploration. *Ore Geology Reviews*, 157-183. doi:https://doi.org/10.1016/S0169-1368(99)00007-4

Safonova, A., Ghazaryan, G., Stiller, S., Main-Knorn, M., Nendel, G., & Ryo, M. (2023). Ten deep learning techniques to address small data problems with remote sensing. *International Journal of Applied Earth Observation and Geoinformation*. doi:https://doi.org/10.1016/j.jag.2023.103569

The MultiMiner project is funded by the European Union’s Horizon Europe research and innovation actions programme under Grant Agreement No. 10109137474

Shang, J., Morris, B., Howarth, P., Levesque, J., Staenz, K., & Neville, B. (2014). Mapping mine tailing surface mineralogy using hyperspectral remote sensing. *Canadian Journal of Remote Sensing, 35*, S126-S141. doi:10.5589/m10-001

Shirmard, H., Farahbakhsh, E., Müller, R. D., & Chandra, R. (2022). A review of machine learning in processing remote sensing data for mineral exploration. *Remote Sensing of Environment*, 112750. doi:https://doi.org/10.1016/j.rse.2021.112750

Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006). Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation., (pp. 1015-1021). Australasian Joint Conference on Artificial Intelligence.

Song, Z., Li, X., Huo, R., & Liu, L. (2023). Intelligent early-warming platform for open-pit mining: Current status and prospects. *Rock Mechanics Bulletin*(100098). doi:https://doi.org/10.1016/j.rockmb.2023.100098

Stehman, S. (2012). Impact of sample size allocation when using stratified random sampling to estimate accuracy and area of land-cover change. *Remote Sens. Lett, 3*, 111-120.

Stehman, S. (2013). Estimating area from an accuracy assessment error matrix. *Remote Sensing of Environment, 132*, 202-211.

Stehman, S., & Foody, G. (2019). Key issues in rigorous accuracy assessment of land cover products. *Remote Sensing of Environment*, 111199. doi:https://doi.org/10.1016/j.rse.2019.05.018

Stehmann, S., & Czaplewski, R. (1998). Design and Analysis for Thematic Map Accuracy Assessment. *Remote Sensing of Environment, 64*(3), 331-344. doi:10.1016/S0034-4257(98)00010-8

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

TRuStEE. (2018). *Standardized protocol for UAV data acquisition.* TRuStEE - Training on Remote Sensing for Ecosystem modelling . Delalieux, S.

van Oort, P. A. J. (2007). Interpreting the change detection error matrix. Remote Sensing of Environment, 108(1), 1–8. https://doi.org/10.1016/J.RSE.2006.10.012

Wang, H., Duentsch, I., Guo, G., & Ali Khan, S. (2023). Special issue on small data analytics. *International Journal of Machine Learning and Cybernetics, 17*, 1-2. doi:https://doi.org/10.1007/s13042-022-01699-0

Yan, Y., Yu, H., & Wang, Y. (2024). Alarming a tailings dam failure with a joint analysis of InSAR-derived surface deformation and SAR-derived moisture content. *Remote Sensing of Environment, 300*(113910). doi:https://doi.org/10.1016/j.rse.2023.113910