MADITRACE

Methodology for analytical data flow analysis

Deliverable D2.5

Final Version

Authors: Nathan Bodereau (BRGM), Théophile Lohier (BRGM), Alban Moradell-Casellas (BRGM), Quentin Dehaine (GTK), Róbert Arató (MUL), Claire Aupart (BRGM), Yuan Shang (GTK), Delphine Losno (BRGM/UGent), Laura Suarez Criado (UGent), Juha Timperi (MO Group), Anne-Marie DESAULTY (BRGM)

Disclaimer

The content of this report reflects only the author's view. The European Commission is not responsible for any use that may be made of the information it contains.

Funded by
the European Union

Document information

Grant Agreement	n°101091502
Project Title	Material and digital traceability for the certification of critical raw materials
Project Acronym	MaDiTraCe
Project Coordinator	Daniel Monfort, BRGM
Project Duration	1 January 2023 – 30 June 2026 (42 months)
Related Work Package	WP2
Related Task(s)	T2.5 : Methodology for analytical data flow analysis
Lead Organisation	BRGM
Contributing Partner(s)	MUL, GTK
Due Date	30 April 2026
Submission Date	5 June 2026
Dissemination level	Public report, BRGM/RP-75373-FR

History

Date	Version	Submitted by	Reviewed by	Comments
04/05/26	1	T. Lohier	D. Montfort, W. Kloppmann	
26/05/26	2	T. Lohier	S. Gaboreau, head of Mineral Characterization & Traceability Unit (BRGM)	
02/06/26	Final	D. Montfort	Adeline Paul (LGI)	

Funded by the European Union

Table of contents

Funded by the European Union

Funded by the European Union

List of figures

Funded by the European Union

Funded by
the European Union

List of tables

Funded by the European Union

Summary

Several analytical techniques were implemented throughout the project to characterize samples of critical raw materials at different stages of their value chains. Considerable volumes of data were thus generated, with each technique producing datasets with different properties. E.g., the lithium isotopic analyses carried out as part of the MaDiTraCe project, combined with data from the literature, allowed the construction of a univariate dataset including numerous samples. XRF and LA-ICP-MS analyses produced multivariate datasets with a smaller number of samples, particularly for lithium. Finally, LIBS analyses and cathodoluminescence produce very high-resolution maps that require specific processing.

Task 2.5 aims to develop a methodology to determine the origin of critical raw material samples based on data generated by various analytical techniques. Given the diversity of these datasets, it would be simplistic to assume that a single workflow could produce relevant results in all cases. However, it is possible to build a common methodological framework for evaluating the different approaches implemented. The workshop held in April 2025 at BRGM highlighted the essential components for developing this methodological framework and establishing a common evaluation methodology.

Different workflows for the traceability of critical raw materials were implemented in task 2.5 and applied to three representative use cases: lithium, cobalt, graphite, and cobalt. For isotopic data, the main challenge was to find a model capable of discriminating samples based on univariate data. For data acquired with LIBS and cathodoluminescence, the main difficulty was finding a feature construction method that captures relevant information for traceability. Finally, for XRF and LA-ICP-MS data, a linear classification approach showed promising results, and an extension of this method allowing the estimation of uncertainty associated with classifier predictions was developed and implemented for the lithium case.

Furthermore, a harmonized evaluation procedure based on a leave-one-out cross-validation was applied to quantify the discriminatory power of the different approaches and analytical techniques. Special attention was given to quantifying uncertainties in order to provide the certifying authorities with as much information as possible. The developed methodological framework thus allows for defining a more or less strict rejection threshold depending on the probability of belonging to a certain origin.

Funded by
the European Union

Keywords

Chemical analysis, Material fingerprint, critical raw materials, geochemistry, traceability, machine learning, lithium, natural graphite

Abbreviations and acronyms

Acronym	Description
CART	Classification And Regressin Tree
CRM	Critical Raw Material
DPP	Digital Product Passport
EMPA	Electron Probe MicroAnalysis
KDE	Kernel Density Estimation
KNN	K-Nearest Neighbours
LA-ICP-MS	Laser Ablation Inductively Coupled Plasma Mass Spectrometry
LA-MC-ICP-MS	Laser Ablation Multi-Collector Inductively Coupled Plasma Mass Spectrometry
LDA	Linear Discriminant Analysis
LIBS	Laser-Induced Breakdown Spectroscopy
LOO-CV	Leave-One-Out Cross-Validation
MFP	Material Fingerprinting
ML	Machine Learning
mXRF	Mobile X-ray Fluorescence
NB	Naïve Bayes
PCA	Principal Component Analysis
PGE	Platinum Group Element
PLS-DA	Partial Least Squares Discriminant Analysis
PoA	Proof-of-Authority
POC	Particulate Origin Carbon
pXRF	Portable X-ray Fluorescence
Q3	Third Quartile
RF	Random Forest
SEM-EDX	Scanning Electron Microscopy- Energy Dispersive X-Ray
SSI	Self-Sovereign Identity
SVM	Support Vector Machine
WP	Work Package
XRD	X-Ray Diffraction
XRF	X-Ray Fluorescence

Funded by the European Union

1 Introduction

Deliverable 2.5 is developed as part of work package 2, which focuses on developing material fingerprinting (MFP) methods to trace the origin of critical raw materials (CRM) and enhance the transparency and traceability of complex CRM supply chains. The targeted CRM include cobalt, lithium, and natural graphite, key components of lithium-ion batteries and permanent magnets.

The task 2.5 is dedicated to developing and applying reliable computational workflows that proceed multisource (e.g., chemistry, mineralogy) and multi-format data of different structure (e.g., spectra, images, bulk analyses) from the tasks 2.1 and 2.2 (i.e., on-site and in-lab analyses respectively) and to training machine learning models capable of detecting deposit-specific patterns in products across the supply chain. When implementing MFP in certification schemes, these models should be efficient to predict origin of CRMs typically providing a binary prediction accompanied by a probability score. In addition, the strategy also ensures the ability to routinely integrate new deposit data without compromising previously acquired information.

The validated models are destined to be incorporated into a generic certification scheme for CRM supply chains from the deposit to the manufactured and recycled products in the framework of the task 3.3 of the WP 3 (i.e., technological solutions: IT methods).

The document lists all workflows developed as part of the task 2.5 and is structured as follows: the first part provides a brief introduction of machine learning, including key terminology used in data science and an overview of the models developed in this task. A presentation of a workshop held in March 2024 at BRGM summarizes the key issues related to data treatment. The following chapters compiles all strategies developed for each CRM (i.e., lithium, graphite and cobalt).

Funded by
the European Union

2 Concept of machine learning

2.1 AI, machine learning – definition, glossary, and models

2.1.1 Brief introduction to machine learning and paradigms

Artificial intelligence refers to the ability of machines to automatise tasks that would require human intelligence (Figure 1). Machine learning is a domain of AI that focuses on developing algorithms which learn from data and apply them to build models able of making predictions for decision (i.e., prediction of class or values). A trained model results from applying an algorithm to data, enabling it to generalize and make predictions on new samples. The methods vary widely but can be categorized according to the type of dataset and the investigated questions. In supervised learning, models are trained with labelled samples and aim to predict these labels. This includes classification, where the goal is to predict categories (e.g., identifying the origin of a mineral sample), and regression, where the objective is to predict a continuous variable (e.g., estimating elemental concentrations). Unsupervised methods refer to classifying or to predicting a value without considering labels. Algorithms group samples by estimating their similarities, i.e., clustering, or identify data points that deviate from the dataset, i.e., anomaly detection.

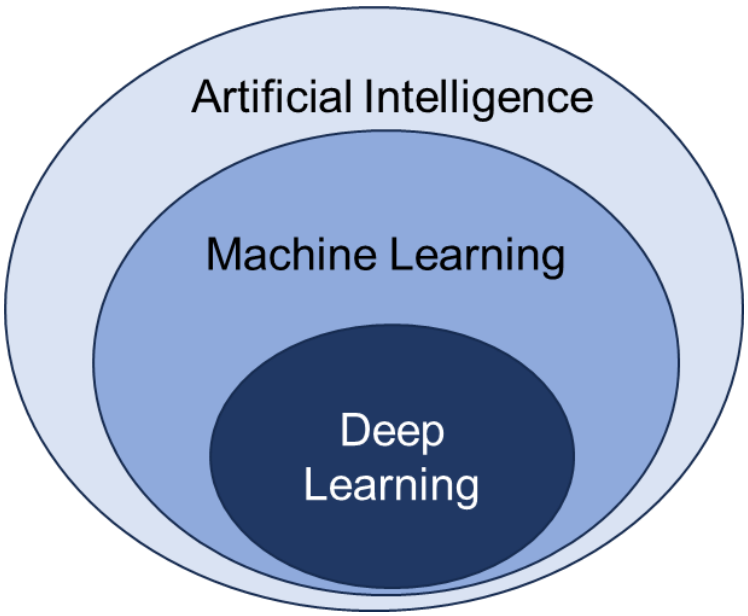

Figure 1. Different domains of AI.

Deep learning is a specialized subset of machine learning with multi-layer neural networks able to automatically extract complex structures and pattern from large datasets without requiring specific parameter setting. More recently, generative AI is a branch of AI enabling models to create text or image by learning from various existing datasets and generating new outputs that compile and summarize them.

Funded by
the European Union

2.1.2 Glossary

2.1.2.1 Validation purposes

Accuracy: a metric used to evaluate classification models, representing the ratio of correctly predicted responses to the total responses.

Cross-validation: technics of validation of the model by splitting the dataset into training (which will feed the model) and testing subsets (where the model will do a prediction).

k-fold cross-validation: specific type of cross-validation where the dataset is divided into k equal parts called folds. The model is trained using each time a different fold as the test set and the remaining folds as the training set.

Confusion matrix: A table comparing true and predicted labels and showing the number of good and erroneous predictions per label.

Leave-one-out: the dataset is split, using all but one sample for training, and the trained classifier predicts the response for the left-out sample. This process is repeated for each sample (n samples = n runs), and the final output is the mean accuracy across all runs.

Loss function: A mathematical function that measures the error between predicted and true values. The learning phase consists of finding the parameters that minimize this function.

2.1.2.2 Data preparation and transformation

Class: A category that a classification model aims to predict (e.g., origin of a sample).

Features: a measurable variable used as input for a mode. Features represent the information the model uses to learn patterns (e.g., elemental concentration, pixel intensities, temperature).

Data preprocessing: Steps necessary to clean, normalize and transform data before model training to improve machine learning.

Standardized data: Data transformed so that each feature has a mean of 0 and a standard deviation of 1.

Normalized data: data rescaled to a fixed range, typically $[0,1]$ or $[-1,1]$. This is useful when features have different units or magnitudes.

Data projection: A transformation that maps datasets into a lower-dimensional space to visualize or simplify complex datasets.

Dimensionality reduction: a method that reduces the number of samples or features without losing the main information.

Outlier: data point that significantly deviates from the rest of the dataset (e.g., measurement errors, rare events or meaningful anomalies.

One-vs-rest: strategy that aim to train a classification model to only distinguish one class from the others. It results in binary models that predict if the sample belongs to the class or not. It enables to further add other classes without compromising previous training.

2.1.2.3 Parameter and learning

Hyperparameter: extra parameters that are not optimized during the training phase. They can be used to control the learning process, such as learning rate, or to define the model architecture, such as the number of trees in forests or the number of neurons in neural networks.

Kernel: A mathematical function used in algorithms such as Support Vector Machines (SVMs) to transform data into a higher-dimensional space where classes become more easily separable.

Overfitting: a situation where the model fits too closely to the training data and includes noise and outliers resulting in poor predictive performance.

Underfitting: the opposite of overfitting. The model is too simple to capture the underlying structure of the data, leading to poor performance on both training and test sets.

2.1.2.4 Ensemble methods

Bagging: techniques where multiple models are trained independently on different subsets of the dataset (bootstrap). Predictions are combined (e.g., by averaging or voting) to reduce overfitting.

Boosting: An ensemble technique where models are trained sequentially. Each new model focuses on correcting the errors of the previous ones. Boosting reduces bias and often achieves high accuracy.

2.1.3 Models

2.1.3.1 Supervised models

- **Logistic regression (LogReg)**

LogReg is a supervised model used for binary classification: the sample belongs or not to the class. Contrary to a linear relationship, it models the probability to belong to a class through a log-odds (sigmoid function) (Figure 2). It assumes that samples are independent, there are linear relationships between features and the log odds, and the target must be binary, meaning that it can take only two values (0 or 1, no or yes). The advantages of the model are its high interpretability, a probabilistic output and it requires a low computational purpose. It however cannot capture non-linear relationships, performs poorly on small datasets and it is sensitive to outliers.

Funded by the European Union

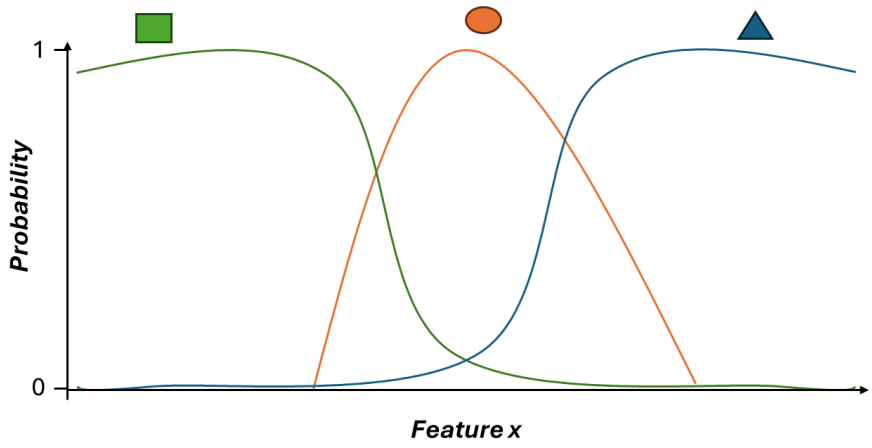

Figure 2. LogReg : log-odds probabilities of three classes considering a feature x.

- **Naïve bayes**

It is a probabilistic classifier based on applying Bayes' rule considering that features are conditionally independent given the class C. This means that the presence of one feature does not affect the presence of another when calculating its probabilities P (Figure 3). The probability type depends on the nature of the data and could be gaussian if the feature follows a normal distribution (Gaussian), multinomial, especially for classification by counting the number of occurrences in features, or Bernoulli, for binary features (yes/no). It is a flexible and fast model, with no parameters to set, that could perform with high-dimension dataset and multiclass problems. It however performs poorly if features are correlated or not normally distributed.

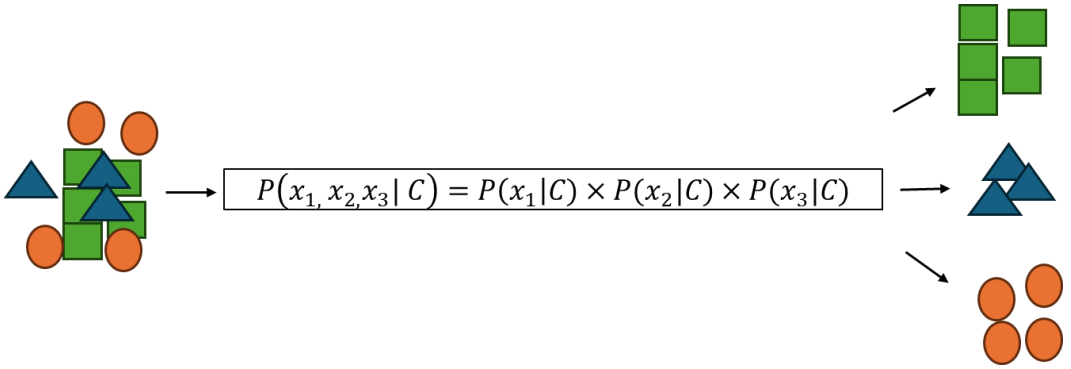

Figure 3. Visualisation of a naïve bayes

- **Linear discriminant analysis (LDA)**

LDA considers linear relationship between features by projecting the dataset into a smaller dimensional space while maximising the separation between classes. The new axes obtained are called canonical discriminant functions and are linear combinations of the original features. LDA makes several key assumptions: (1) features follow gaussian distribution within each class considering the mean and the variance, (2) the covariance matrix (linear relationship between features) is identical across classes and (3) decision boundaries are linear. It can be compared to principal component analysis (PCA), especially for the visualisation, but it includes class awareness. While LDA enables to well interpret decision boundaries, the limits of such a model are that the covariance matrices do not change between classes, the strategy cannot capture nonlinear relationships, and class definition is sensitive to outliers as the model relies on the means and variances of feature per class. Contrary to LDA, Quadratic Discriminant Analysis (QDA) does not assume similar relationship across classes and is more flexible with non-linear boundaries. But this method requires more parameters to be set which could lead to overfitting especially for small dataset. The interpretability is also more difficult than LDA as nonlinear boundaries are more complex.

- **K-nearest neighbours (KNN)**

The KNN is an approach which translates the original dataset into a new distance-based space where the distance between points reflects the similarities between samples. The commonly used distance is Euclidean, even if other metrics could be applied (e.g., Manhattan or Minkowski). The hyperparameter k refers to the number of nearest neighbours to consider for each sample (Cover & Hart, 1967). The assigned class is the most frequent one among these neighbours. It is a simple way to understand and to implement and it does not require assumptions about the data distribution. The limits of the models are related with the time of computing as it requires calculating distances for all training points. The performance also depends on the choice ok: if k is smaller, the model risks overfitting, while a larger k would lead to remove important patterns and lead to underfitting.

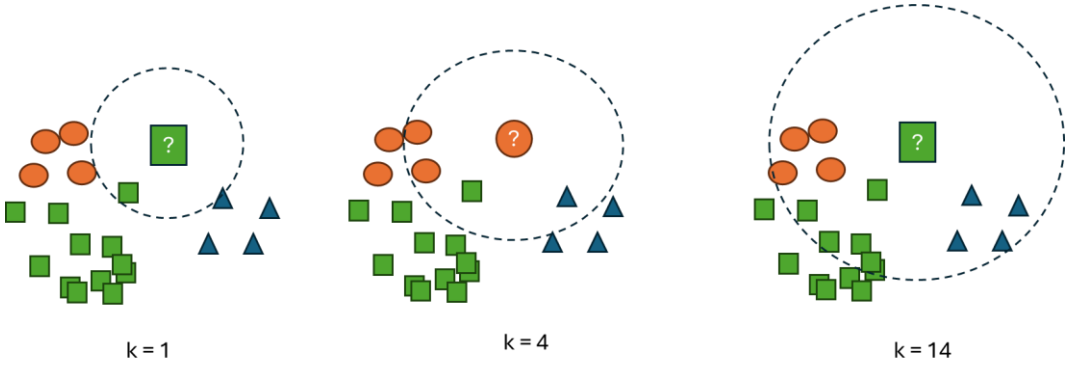

Figure 4. KNN predictions of an unknown sample considering k.

Funded by the European Union

- **Support vector machine (SVM)**

SWM projects data into another dimensional space using a kernel function, making it easier to sperate classes. An optimal hyperplane is calculated to maximise the margin between closest points of each class. The key hyperparameters of the models are: (1) the type of kernel (e.g., linear, polynomial, radial basis function, sigmoid...); (2) its related coefficients; and (3) the regularisation parameter (C) that balances the trade-off between the margin maximisation and the prediction of the model. SWM has the advantage to display robust prediction especially when datasets could lead to overfitting (small number and heterogeneous samples). However, it is computationally intensive especially with large dataset, it requires high processing memory, and the kernel transformation generally limits its interpretation.

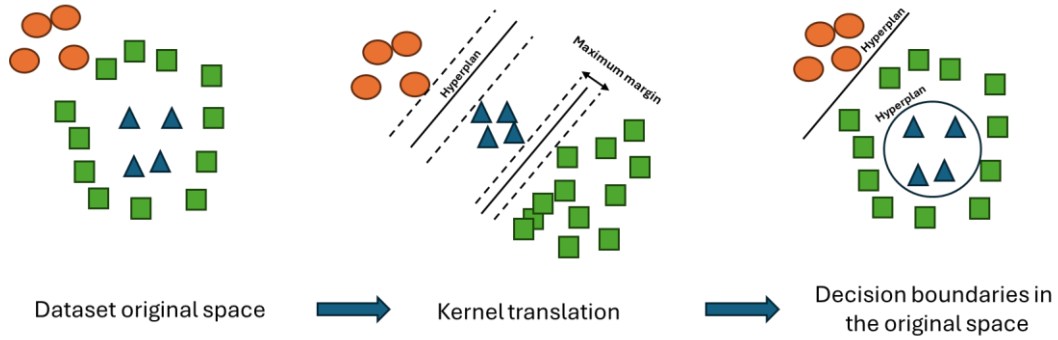

Figure 5. Visualization of a SVM concept.

- **Classification and regression trees (CART)**

Decision trees handle non-linear relationship by recursively splitting the dataset based on feature values used as thresholds, aiming to isolate well-define (i.e., pure) classes at the leaf nodes. Each split is chosen to maximize homogeneity, using criteria such as entropy or Gini impurity. Key parameters include maximum depth, minimum number of samples per leaf, and the splitting criterion. One major advantage of decision trees is their interpretability: the model's logic can be easily visualized and understood, making it useful for explaining predictions. However, they are sensitive to overfitting, especially with deep trees. Additionally, their splits are axis-aligned, limiting their ability to capture complex diagonal patterns.

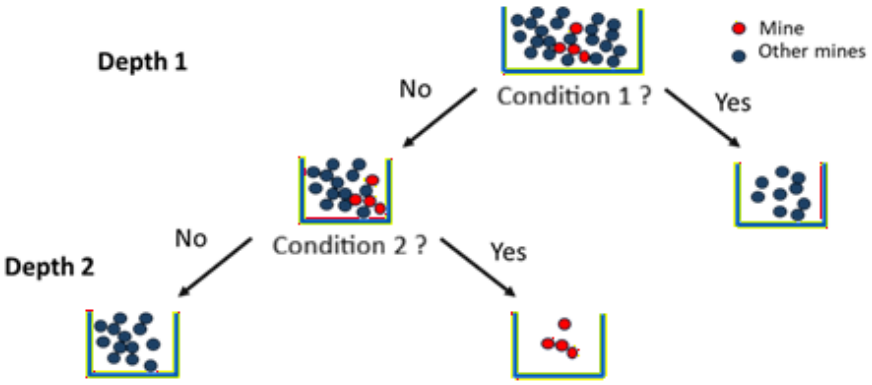

Figure 6. Schema of a classification tree.

In order to address these limitations, Random forest reduces overfitting by training multiple trees on random subsets of the data and aggregating predictions through averaging or voting (Breiman L. , 1996).

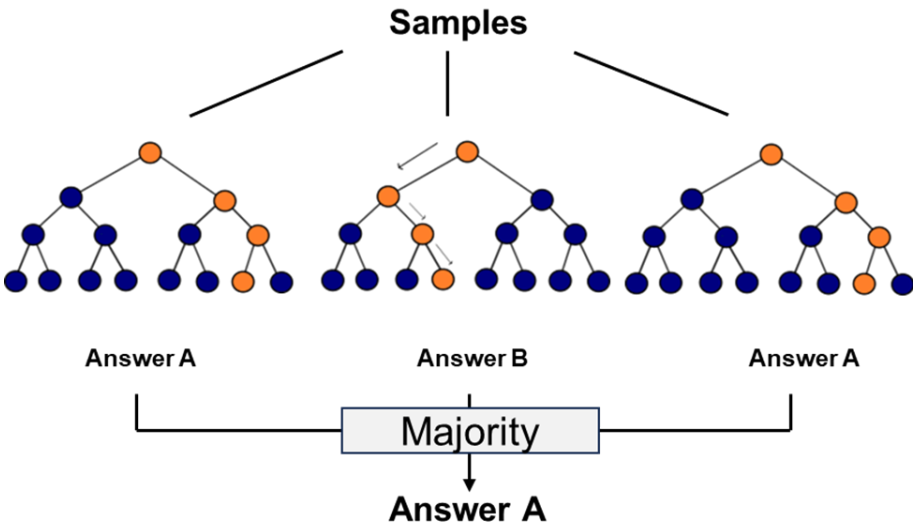

Figure 7. Concept of random forest.

The number of trees is a key hyperparameter. In contrast, **Gradient boosting** trains trees sequentially, which means that each model is correcting the residual errors of the previous one with the goal of gradually reducing bias and variance. It is an appropriate strategy to reduce underfitting.

2.1.3.2 Unsupervised models

- **Principal component Analysis (PCA)**

The goal of PCA is to describe a dataset using a small number of uncorrelated variables but retaining as much information as possible. This reduction is applied by transforming the dataset into a new set of continuous variables called principal components. A major part of the information is summarized into the first components, while the non-essential part is

Funded by the European Union

observable in the latter. The smaller the number of components, the better the PCA representative of our dataset

- **Clustering**

This is a family of unsupervised learning algorithms designed to group similar samples based on their intrinsic patterns, without relying on labels. These methods group the samples by measuring similarity, typically through a distance metric such as Euclidean, Manhattan or cosine distances.

The most widely used is the K-means, which aims to partition data into k compact and spherical clusters, defined by a centroid, by minimizing the within cluster variance. It is simple and efficient to interpret but assumes convex cluster shapes and requires setting the number of clusters k. K-medoids improves the robustness by using a sample as the centroid of the cluster, making it less sensitive to noise and outliers. These approaches are completely deterministic and assign each sample to a cluster.

Probabilistic approaches that allow for more nuanced classification have been developed. Notably, the fuzzy c-means approach calculates a degree of membership to the different clusters for each sample (Kamel & Selim, 1994). We can also cite Gaussian mixture models that describe clusters through multivariate Gaussian distributions, allowing the estimation of a cluster membership probability for each sample.

2.2 April 2025 - Workshop for machine learning and data processing as part of MaDiTraCe

2.2.1 Introduction

A workshop was held at BRGM, Orléans, on 23-24 April 2025, bringing together scientists from several research areas (Figure 8), including geochemists, geologists, geophysicists, software developers, data scientists and circular economy experts.

Its objectives were:

- To pool all machine learning approaches developed in the task 2.5 into a common framework.
- To present complementary strategies developed within other projects that could be beneficial to the task.
- To list technical issues such as missing data, sample distribution representativity and discuss potential solutions.

Figure 8. Participants attending the 2.5. task workshop at BRGM.

To design an appropriate program, pre-workshop meetings were organised with each laboratory and company. The aim of these discussions was to provide an overview of the various analyses carried out as part of Work Package 2, as well as the classification models developed, the issues encountered (such as the handling of missing data and the consideration of limits of quantification), and methods for optimising the processing of high-resolution data. A summary of questions raised in the framework of these meetings is yielded in Figure 9.

This workshop then has been subdivided into three sessions where PhD candidates and researchers presented results of their own research. The abstract of each presentation is available in the following sections. The perspectives to implement them as part of the task 2.5. are displayed in a blue rectangle.

Funded by the European Union

- Session 1 - Data processing, feature creation and data completion:

Analysis of the dataset before classification to ensure the model will be well trained; data completion (low number of samples, limit of quantification), how to create the greatest number of features available? How to deal with high resolution datasets?

- Session II – classification models

Three studies using respectively supervised, unsupervised and semi-supervised models were presented.

- Session III - Uncertainties propagation and Digital Product Passports

The first presentation introduced ideas of concepts developed to propagate uncertainties in the experimental approach as part of signal treatment. The second presentation was as a perspective to how integrate models developed in task 2.5 into the framework of task 3.4.

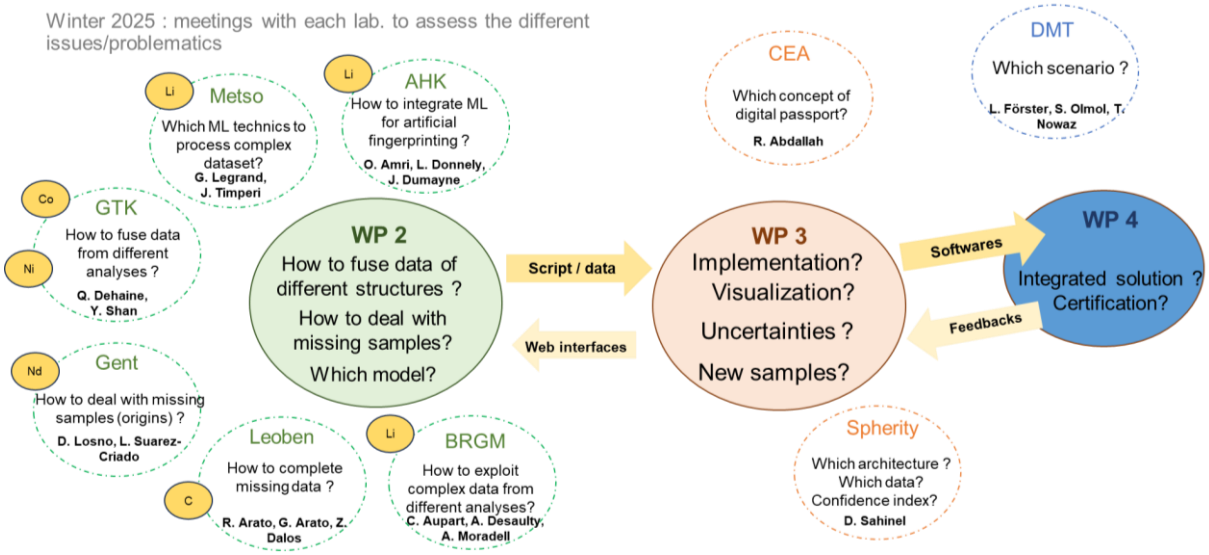

Figure 9. Graphical abstract of the pre-meetings with questions raised.

2.2.2 Session I – post-analysis processing, feature creation and data completion

Neural networks for X-ray diffraction computed tomography. Application to Mineralogical Identification – Titouan Simmonet, Ph.D. (BRGM)

More details in (Simonnet, et al., 2024).

Understanding the chemical and mechanical behaviour of compacted materials (e.g. soil, subsoil, engineered materials) requires a quantitative description of the material's structure, and in particular of the nature of the various mineralogical phases and their spatial relationships. Natural materials, however, are composed of numerous small-sized minerals, frequently mixed on a small scale.

Recent advances in synchrotron-based X-ray diffraction (XRD) tomography now make it possible to obtain tomographic volumes with nanometer-sized voxels, with a XRD pattern for each of these voxels. Here, we use neural network approaches to identify and quantify minerals in a material. Training such models requires the construction of large-scale learning bases, which cannot be made up of experimental data alone. Algorithms capable of synthesizing XRD patterns to generate these bases have therefore been developed. An example of the analysis of an XRD-CT data set obtained on a concrete core are also presented.

Take-home message and implementation possibilities as part of MaDiTraCe:

Neural network generally required a well-constructed training dataset with labelled and variable data. In the framework of the quantification of mineral phases, the authors faced lack of labelled data. To overcome this scientific concern, the model has been trained with a synthetic training data of cement where each phase proportion has been taken from literature.

- The authors have faced the same problem of training model with limited datasets.
- Whereas neuronal networks are not required as most of our analytical results are not of high dimensionality, completion of data sets by synthetic data for portions of the supply chain where samples are underrepresented or absent could be an option.

Funded by the European Union

Deep Internal Learning: Deep Learning with a Single Input – Herbert Rakotonirina, Ph.D (BRGM)

More details in (Rakotonirina, Guridi, Honeine, Atteia, & Van Exem, 2024).

Deep Internal Learning is a paradigm where deep neural networks are trained on only a single input example, without relying on any external training data, by exploiting the input's own internal structure. This presentation introduces the concept and highlights three key methods that exemplify its power: (1) Deep Super-Resolution, a zero-shot approach that reconstructs a high-resolution image from one low-resolution input by leveraging internal patterns; (2) Deep Image Prior (DIP), which uses an untrained neural network's architecture as an implicit prior to perform image restoration tasks (such as denoising or inpainting) with no external data; and (3) a DIP-based spatial interpolation technique that infers missing spatial data solely from a given sample. Notably, none of these methods require any pre-collected training dataset, and the spatial interpolation approach is especially relevant in geoscience contexts where ground-truth observations are often scarce.

Take-home message and implementation possibilities as part of MaDiTraCe:

This presentation highlighted a strategy based on unsupervised deep-learning algorithms designed to predict pollutant concentrations from low-resolution 3D data (i.e., images combined with core-acquisition measurements). The authors demonstrated that the model is capable of being trained on a single sample while still delivering reliable spatial-scale predictions, including for unlabelled pixels and for subsets of core data intentionally removed to test robustness.

- This strategy could be involved as part of image processing (i.e., LIBS, Cathodoluminescence) to capture intrinsic patterns that would benefit to identifying the deposit or origin.
- It would be a good asset for deposits for which data are limited to one sample.

Detecting chemical complexes from LIBS-based elemental imaging by coupling clustering and decision tree – Nathan Bodereau, Ph.D. (BRGM)

More details in sections 3.3**.**

Compiling data from different devices can be challenging as it depends on their structures which could vary from a simple bulk analysis to vast and complex datasets. Here, we focus on the elemental images from Laser Induced Breakdown Spectroscopy (LIBS) performed on Li-rich minerals and concentrates. Combining high-resolution elemental images, that we

seek to summarize, highlights elemental combinations related to sample specificities, such as minerals, linked to their provenance.

We developed an approach to detect these chemical complexes for each sample by coupling clustering and decision trees: The clustering allows identifying the main chemical complexes by grouping pixels by similarities while the decision tree, trained with cluster labels, is used to synthetize and standardize these complexes. Based on this work, we detected several elemental complexes that were not immediately obvious and constitute indicators to classify samples according to their origins.

Take-home message and implantation possibilities as part of MaDiTraCe:

This study highlights a first strategy applied to LIBS imaging of spodumene and lepidolite concentrates, which will be further refined. For each sample, this preliminary approach used an unsupervised clustering algorithm to group pixels with similar elemental signatures based on their LIBS-derived concentrations. Supervised decision-tree models are then trained on these clusters to summarize and interpret the chemical information in a structured way.

- The decision trees reveal specific elemental combinations, particularly involving trace elements that occur in heterogeneous regions of the samples. These associations appear to be characteristic of the geological origin;
- A comparison of decision trees that share similar structures is expected to support the discrimination of deposits, providing a consistent and interpretable framework for comparing samples across different sources;
- A new workflow is proposed in Section 3.2.

2.2.3 Session II – classification models

Lithium mines traceability with portable EDXRF: comparison of data exploitation by LDA, PCA, and RF – Alban Moradell-Casellas, Ph.D. (BRGM)

A case study of the chemical data treatment from on-site instrument analysis (Portable EDXRF) for the lithium traceability was presented. Two supervised models, Linear Discriminant Analysis (LDA) and Random Forest (RF), and an unsupervised model, Principal Component Analysis (PCA) are used to (1) highlight the most discriminant features and (2) to distinguish samples according to their origins. Model comparisons are proposed, based on their performances, to better understand their respective contributions and limitations for the development of geochemical traceability systems. As a result, all these classification methods displayed similar configurations with several groups of samples related to their provenances. LDA et PCA however show limitations as they are not designed to capture non-linear relationships between input variables and output classes. Reversely, RF follows bimodal distributions that include any kind of relationship, and the risk of overfitting is avoided as samples are randomly selected to form sub-datasets to build the trees.

Funded by the European Union

Take-home message

- *Mine **origins are well differentiated** by the LDA → less need of non-linear correlation resilient classification models*
- *PCA can give insights on the **variability within a mine** (natural variability vs processes)*
- ***Origin prediction** with these models is promising*
- *Bad classification for **mines with few samples** in the database*

A fuzzy clustering approach for sample classification: application to riverine radiocarbon in a nuclearized watershed – Nathan Bodereau, Ph.D. (BRGM)

More details in (Bodereau, et al., 2022).

Fuzzy logic is a reliable method to cluster multidimensional data as it estimates the contribution of each cluster on samples which makes it powerful compared to traditional clustering where every sample is assigned to only one group. This approach has been applied to classify riverine particulate organic carbon-14 (Δ^{14}C-POC), at the outlet of the Rhône River, linked to the origins of the material and nuclear power plant releases. Fuzzy C-means Analysis was performed based on water discharges of the main tributaries. The classification summarized the river hydrology into five clusters with Mediterranean/Cevenol flood, oceanic pluvial flood, snowmelt flow, low-water level and baseflow clusters. Δ^{14}C-POC distribution is varying among each cluster confirming the impact of hydrology on ^{14}C activities. In addition, the estimation of probabilities of sample belonging to each cluster made this methodology interesting as a solution to estimate the contribution of each hydrology on Δ^{14}C-POC.

Take-home message and implementation possibilities as part of MaDiTraCe

An unsupervised fuzzy-clustering approach was applied in this study to determine how different hydrological sources contribute to the ^{14}C signal measured at the main Rhône River outlet. This strategy made it possible not only to identify which tributary are contributing per event but also to estimate the probability of each tributary to the event.

- **Fuzzy clustering characteristics: each sample retains a degree of membership to multiple clusters, allowing the method to represent mixed contributions more realistically.**
- **Transferability of the approach: this methodology could also be applied to estimating the degree of membership of a sample to a deposit.**

Funded by the European Union

Features Leverage in Graph Models for Mineral Prospectivity Mapping –Thi Hai Yen Vu, Ph.D. (BRGM, Orléans Univ.)

More details in (Vu, Dao, Nguyen, Vrain, & Breuillard, 2025).

Mineral Prospectivity Mapping (MPM), the process of identifying areas with high potential for mineral deposits, can be divided into two main categories: knowledge-driven and data-driven. Knowledge driven techniques rely on expert opinion on geological data, while data-driven techniques employ machine learning (ML) models to predict the probabilities of mineral occurrences based on known geological datasets. Recently, with the advancement of ML methods, data-driven MPM has gained significant improvements. Notably, graph-based approaches overcome the disadvantages of previously used approaches (pixel-based, image-based) and have demonstrated better performances. However, the graph construction in current methods is based solely on spatial distances between pixels, regardless of their geological attributes. In the cited paper, we introduce a novel graph construction approach that combines spatial distances with other distances obtained from feature mining. Our experiments show that this combination outperforms existing graphs and can considered as a promising approach to integrate feature mining into data-driven models in MPM.

Take-home message and implementation possibilities as part of MaDiTraCe

The authors developed a framework including semi-supervised deep-learning algorithms applied to geological maps with limited labelled cell samples. The main goal was to map lithological diversity and geological structures (e.g., fractures) at the scale of the Armorican massif (France). A first step aimed to identify patterns per cell with a graphic. Similarities between pixels were calculated through a neighbours-based distances approach: Graph-based Neural networks. They displayed good abilities to map the lithologies of the Armorican massif at a high resolution.

- **Graph Neural Networks could be performed to capture patterns per samples across the supply chain and group common patterns, e.g., chemical composition distances.**

2.2.4 Session III – Uncertainties propagation and Digital Product Passports

Study of estimation uncertainty while using an autocalibration method for microgrid-based full Stokes imagers – Benjamin Le Teurnier, Ph.D. (BRGM)

More details in (Le Teurnier, Li, Boffety, Hu, & Goudail, 2020).

Full Stokes polarimetric images can be obtained from two acquisitions with a microgrid polarization camera equipped with a rotating retarder. When the delay of the retarder is imperfectly known, leading to an estimation bias, it can be calibrated from the measurements, but this requires three image acquisitions and may cause divergence of

Funded by the European Union

estimation variance at a low signal-to-noise ratio. With a study of the estimation uncertainties, we determine closed-form equations allowing to decide in which experimental conditions this autocalibration is possible and useful, and to quantify the performance gain obtained in practice. These results are validated by simulations and real-world experiments.

Take-home message and implementation possibilities as part of MaDiTraCe

The authors developed a method based on uncertainty estimation to define when to use the autocalibration of the delay. When a closed form expression of STD is not possible to compute, they compute a theoretical bound. Comparison with Monte Carlo simulation and real-world experiments enable to define when this bound is a reliable criterion.

- **Evaluating the uncertainty surrounding an estimate can help in the definition of a rejection criterion.**

Building Trust for Artificial Fingerprinting as component of DPPs : Session 3 – Implementation regarding WP3 & WP4 – Doruk Şahinel, Ph.D. (Spherity)

Digital Product Passports (DPP) build the digital backbone for the circular economy in Europe, enabling traceability for critical raw material supply chains. The MaDiTraCe project recognizes that DPPs can be more than just a regulatory tool, and ongoing research explores integrating material and artificial fingerprints into the DPP framework. While these fingerprints offer strong potential for anchoring physical products – and their geographical provenance – to their digital counterparts with high confidence, challenges remain in standardizing their representation. To certify artificial fingerprints, the project aims to create a verifiable metadata document describing the fingerprinting operation using W3C verifiable credentials. This includes details about the learning algorithms, model used, training data origin, and application area, stored as a credential in the DPP for auditability and transparency. Artificial fingerprint claims follow SSI principles, and the concept will be enhanced by PoA credentials for AI agent authorization.

Take-home message and implementation possibilities as part of MaDiTraCe

Digital Product Passports (DPP) aim at linking information on tracking, transport, and processing of CRMs. With its identity credential as scientific authority, the institute certifies raw material origin.

- **The results produced by the statistical methods developed in task 2.5 could be integrated in the DPP in form of verifiable credentials as supplementary information for the certification.**
- **It is essential to develop methodologies that are able to give a confidence level on the estimated origin of the raw material.**

2.3 Conclusion of the workshop

The presentations highlighted several machine learning approaches that could be implemented on data from geochemical analyses to improve the traceability of CRMs. They also served as a basis for discussions around the challenges inherent to this issue: representativeness of the available samples, risk of overfitting exacerbated in cases where the number of variables is of the same order of magnitude as the number of samples, processing of data acquired at very high resolution, consideration of uncertainty in the decision-making process.

At the end of these discussions, a generic workflow was proposed (Figure 10). For each step, a set of issues was defined and initial levers were identified, including:

- Completing the isotopic analyses performed on lithium with data from the literature.
- The use of a Leave One Out cross-validation procedure to better detect potential over-fitting.
- The use of one-versus-the-rest classifiers to include deposits for which few samples are available and which are therefore poorly characterized.
- The implementation of dimensionality reduction techniques for very high-resolution data.

Funded by the European Union

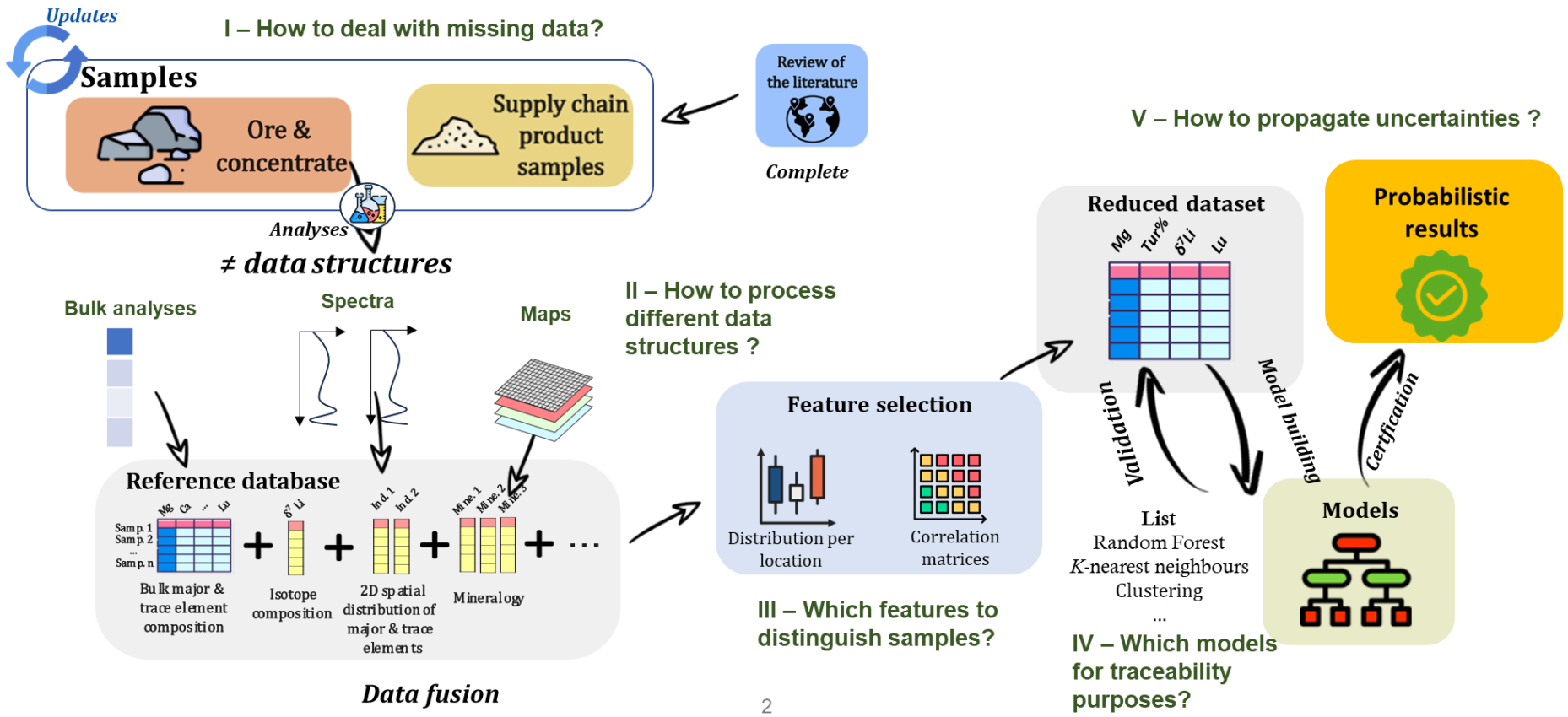

Figure 10. Overall workflow for task 2.5.

Funded by the European Union

3 Li – workflows and results

3.1 Classification based on a univariate dataset: assessing the discriminatory power of δ^7Li

Authorship: Théophile LOHIER, Quentin DEHAINE, Nathan BODEREAU, Alban MORADELL-CASELLAS, Anne-Marie DESAULTY

3.1.1 Context and main goal

Lithium (Li) isotope fingerprints are a valuable tool for determining the origin of lithium. (Desaulty, et al., 2022) demonstrated that the diversity of Li isotopic signatures of different deposits is reflected in the secondary products and can be used to certify lithium origin. The comparison of the δ^7Li of new samples with distributions built from reference samples can help in the estimation of the new sample origin. However, significant inter-class overlaps were highlighted in δ^7Li reference distributions, and samples with δ^7Li signature falling in these overlapping areas cannot be classified. In this section we seek to extend the graphical approach by training a statistical model with δ^7Li data from Li deposits with a double objective: 1) to be able to propose an origin for any new sample, including those whose signature falls between two reference distributions, and 2) to quantify the uncertainty surrounding this estimation. We explored the ability of four machine learning algorithms to achieve these objectives for two resolutions: a coarse resolution where deposits are grouped according to their geographical origin and geological type and a finer resolution where each deposit is considered individually.

3.1.2 Data acquisition

Models were trained with updated literary δ^7Li dataset from (Desaulty, et al., 2022), 55 documents (papers and reports) were examined cumulating 446 samples. In addition, 41 samples were analysed by MC-ICP-MS increasing the number to 537 (Figure 11).

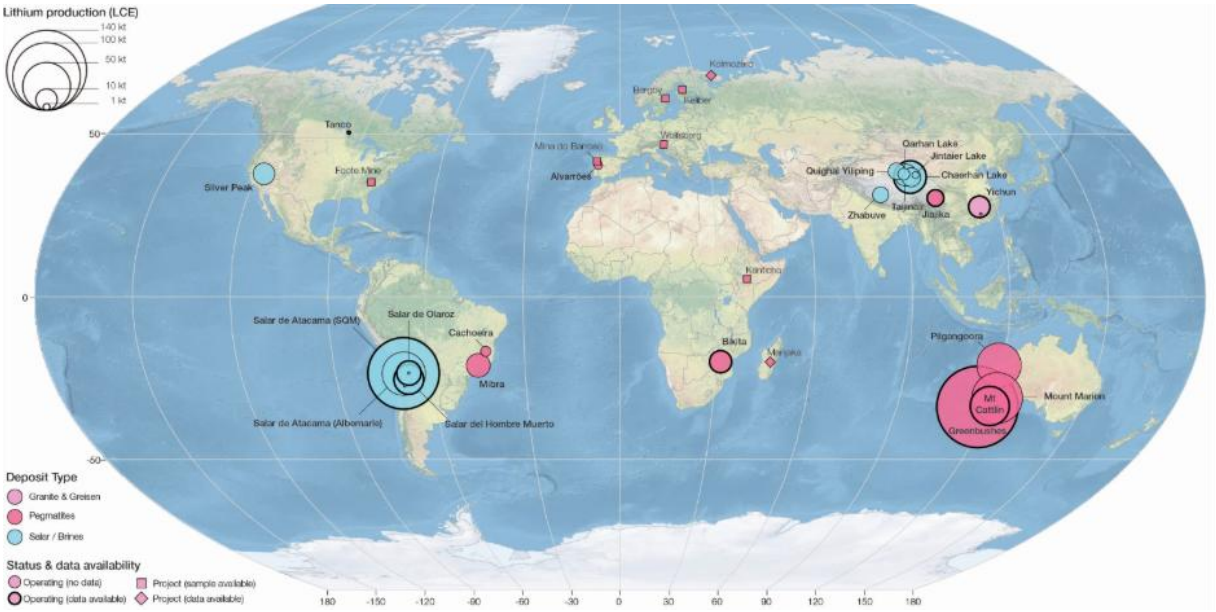

Figure 11. World map of major lithium producers by deposit type and location of the deposits and mines investigated.

3.1.3 Strategy

We aim at exploring the capacity of machine learning algorithms to exploit the δ^7Li signature of samples to determine their origin. We seek to enhance the graphical approach previously developed so that the probability that a sample belongs to a deposit can be quantified on a harmonized scale.

3.1.3.1 Dataset preparation

We have built two datasets from the compiled data. In the first, a rough description of the origin of the samples is constructed based on the type of deposit - brine or hard rock - and the country in which it is located. This results in five classes: Brines from China (Brine China), Brines from South America (Brine SA), Brines from the United States (Brine US), Hard rocks from China (Hard rock China), Hard rocks coming from other countries (Hard rock others). This last class groups all the samples coming from minor deposits, that should be identified as not belonging to the four main deposits. In the second dataset, we select the samples from the most represented deposits, and we seek to determine if we are able to determine their origin from their δ^7Li signature.

3.1.3.2 Method description

For both datasets, we explore the ability of different machine learning algorithms to determine sample origins from their δ^7Li signature. We use the datasets to train four classifiers: a naïve Bayes (NB), a K-nearest neighbours (KNN), a Support Vector Machine (SVM), and a Random Forest (RF) (see section 2.1.3.1 for detailed description of these algorithms). To achieve optimal accuracy, these algorithms must be configured and the number of hyperparameters that need to be set will vary depending on the algorithm complexity. The simplest classifier, NB, can be used without any finetuning. KNN can be configured with only one parameter, while SVM requires the configuration of a kernel and the configuration of a regularization parameter. Finally, RF is the most complex algorithm, and many configurations can be investigated. Parameters include for example the number of trees, the maximum depth of the tree, or the splitting rules. For algorithms requiring hyperparameters setup, we explored different configurations and selected the one that maximizes the predictive power. The predictive power of the different classifiers is assessed for both datasets using Leave-One-Out cross-validation.

On the other hand, we assess the ability of the different algorithms to capture the uncertainty surrounding the estimated origin. Practically, the classifiers return, for each sample, a label that is the predicted origin and the probability of belonging to each of the deposits. We investigate how these probabilities can help to differentiate the samples that should be treated with caution from the samples that can be attached to a deposit with high level of confidence.

3.1.4 Results

3.1.4.1 Dataset description

Looking at Figure 12, the isotopic signatures of the hard rock and brine classes appear quite distinct. However, there is a significant overlap between the distributions of isotopic signatures of South American brines and Chinese brines, and to a lesser extent with that of US brines. Similarly, the distributions of isotopic signatures of Chinese hard rocks and hard rocks from other countries are overlapping. There is also a slight overlap between the distribution of the latter and the distributions of the brines.

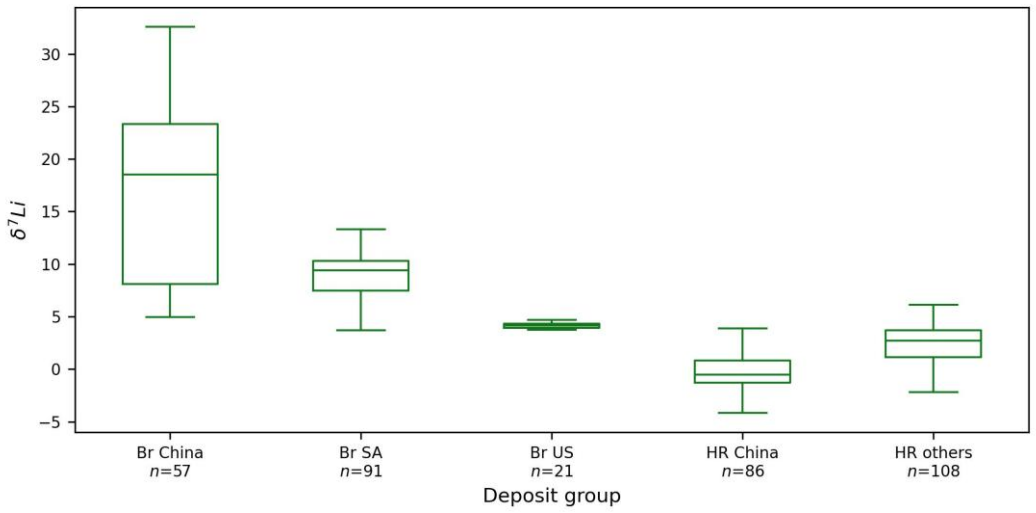

Figure 12. Distribution of δ7Li signature among the five coarse classes (Br = Brine, HR = Hard Rock).

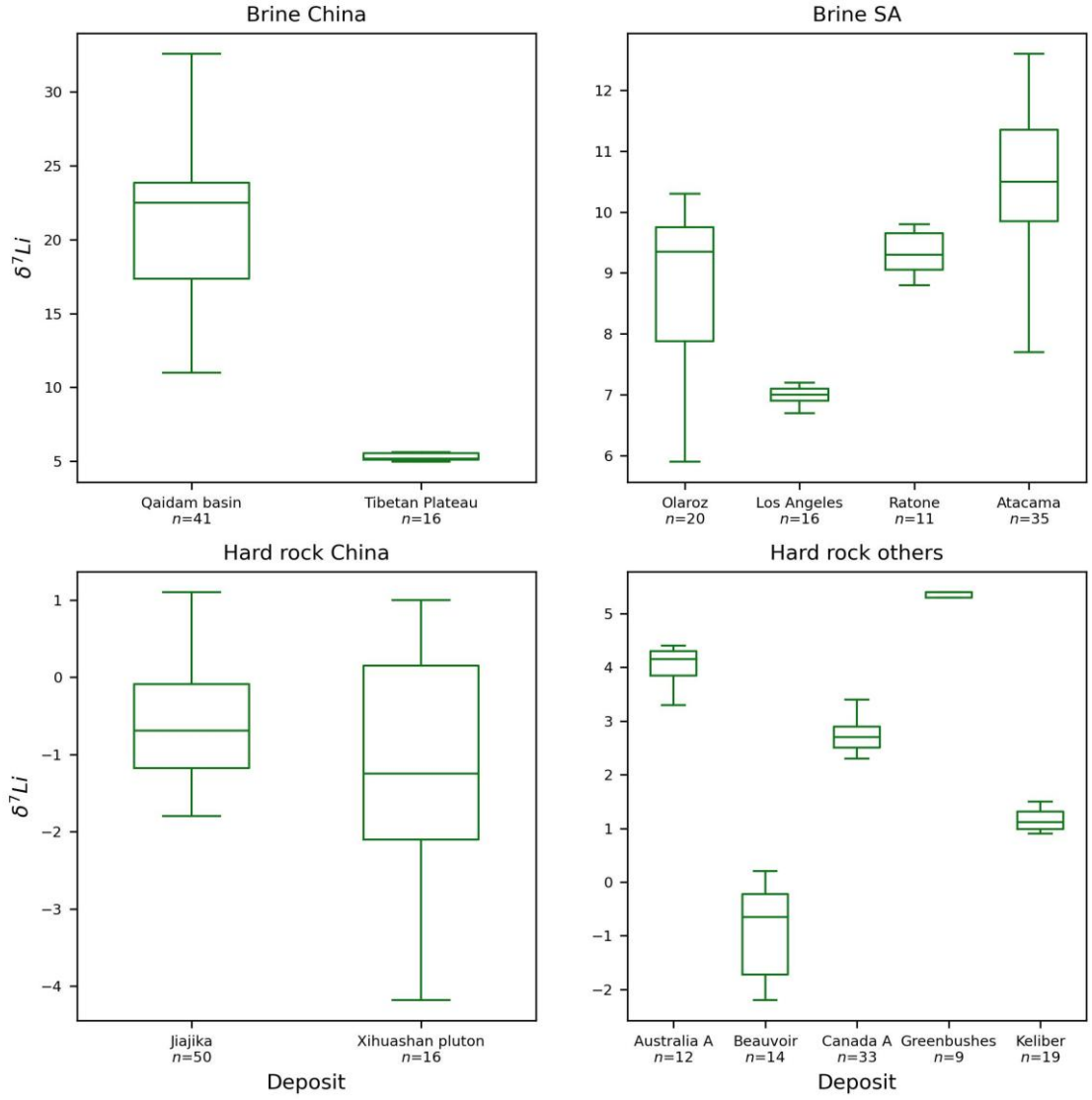

Figure 13. Distribution of δ^7Li signature among the most represented deposits in the dataset for the four coarse classes.

Funded by the European Union

The distributions of isotopic signatures of the two deposits in the Brine China class are clearly distinct, as are those of the four deposits in the Hard rock others class. However, there are significant overlaps between the distributions of isotopic signatures of the deposits in the Brine SA class. Finally, the distributions of isotopic signatures of the two deposits in the Hard rock China class are visually indistinguishable (Figure 13).

3.1.4.2 Coarse origin determination

In this section we detail the performance of the four algorithms for classifying samples across the five coarse classes: Brine China, Brine SA, Brine US, Hard rock China, Hard rock others. Moreover, we discuss the capabilities of the best classifier to provide an estimate of the uncertainty surrounding the predicted origin.

Overall, the KNN shows the best discriminatory power with 81% of samples correctly classified. In particular, it is significantly better than the other classifiers at identifying samples belonging to the brine US class. The RF produces similar results with an accuracy of 78%. It has a better classification rate for the Brine China and Hard rock others classes but lower rates for the Brine SA, Brine US, and Hard rock China classes. The SVM shows good performance for samples belonging to the Brines SA and Hard rock others classes but fails to capture the signature of the Brine US class. The NB has an overall accuracy of 71% and shows lower classification rates for all classes (Figure 14).

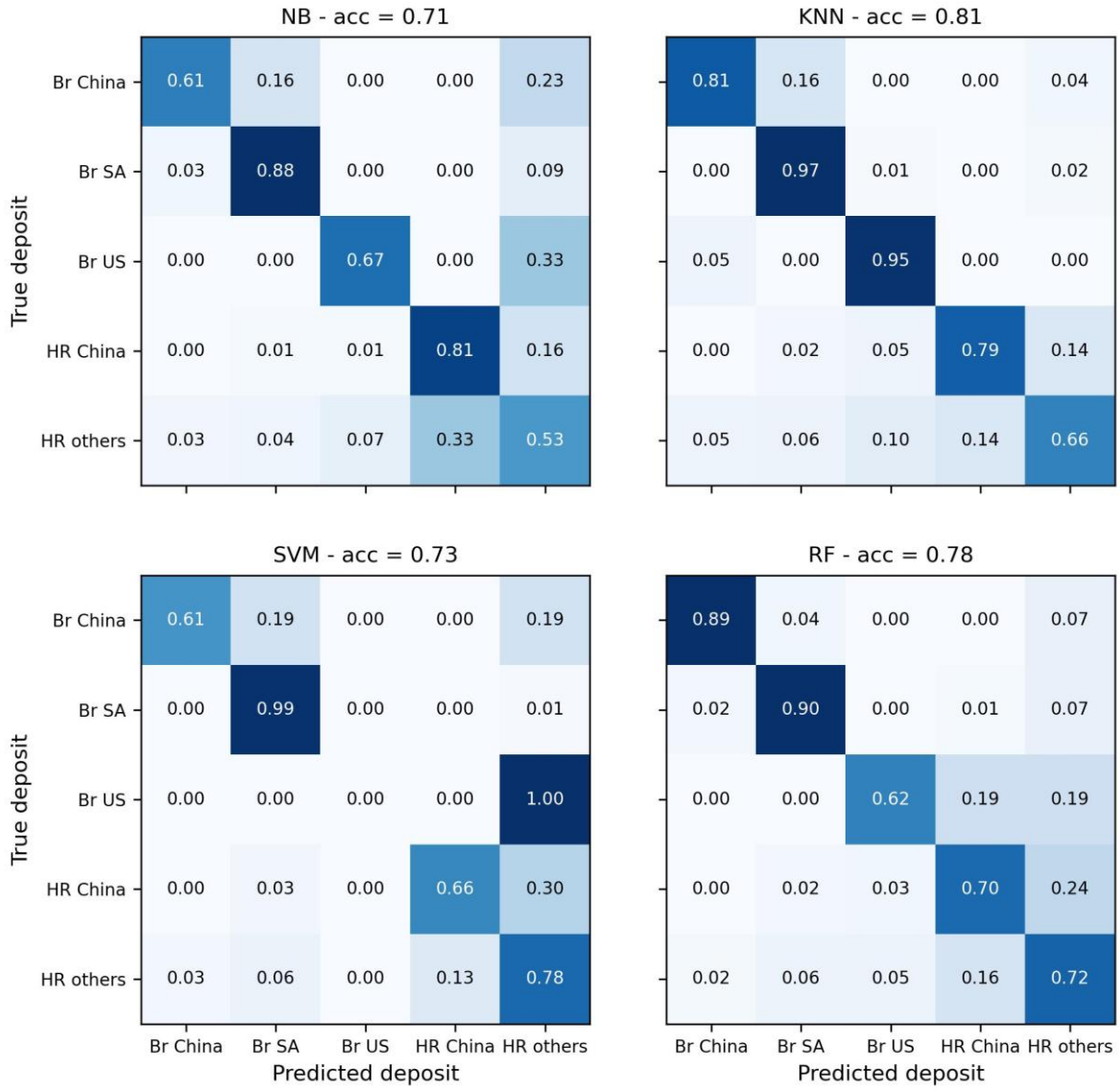

Figure 14. Confusion matrices describing the capacity of the four machine learning models to classify the samples across the five coarse origins.

Looking at the membership probabilities for the different classes for the incorrectly classified samples, we can see that the majority of samples belonging to the Brine China class that were incorrectly classified are assigned to the Brine SA class with a probability between 0.6 and 0.8. For these samples, the membership probability to the Brine China class ranges between 0.2 and 0.4. Similarly, several samples belonging to the Hard rock China class that were incorrectly classified are assigned to the Hard rock others class, with a membership probability to the Hard rock China class ranging from 0.1 to 0.4. Finally, the samples belonging to the Hard rock others class that were incorrectly classified are assigned to the Hard rock China or Brine US classes and almost all show a non-zero membership probability to the Hard rock others class (Figure 15).

Funded by the European Union

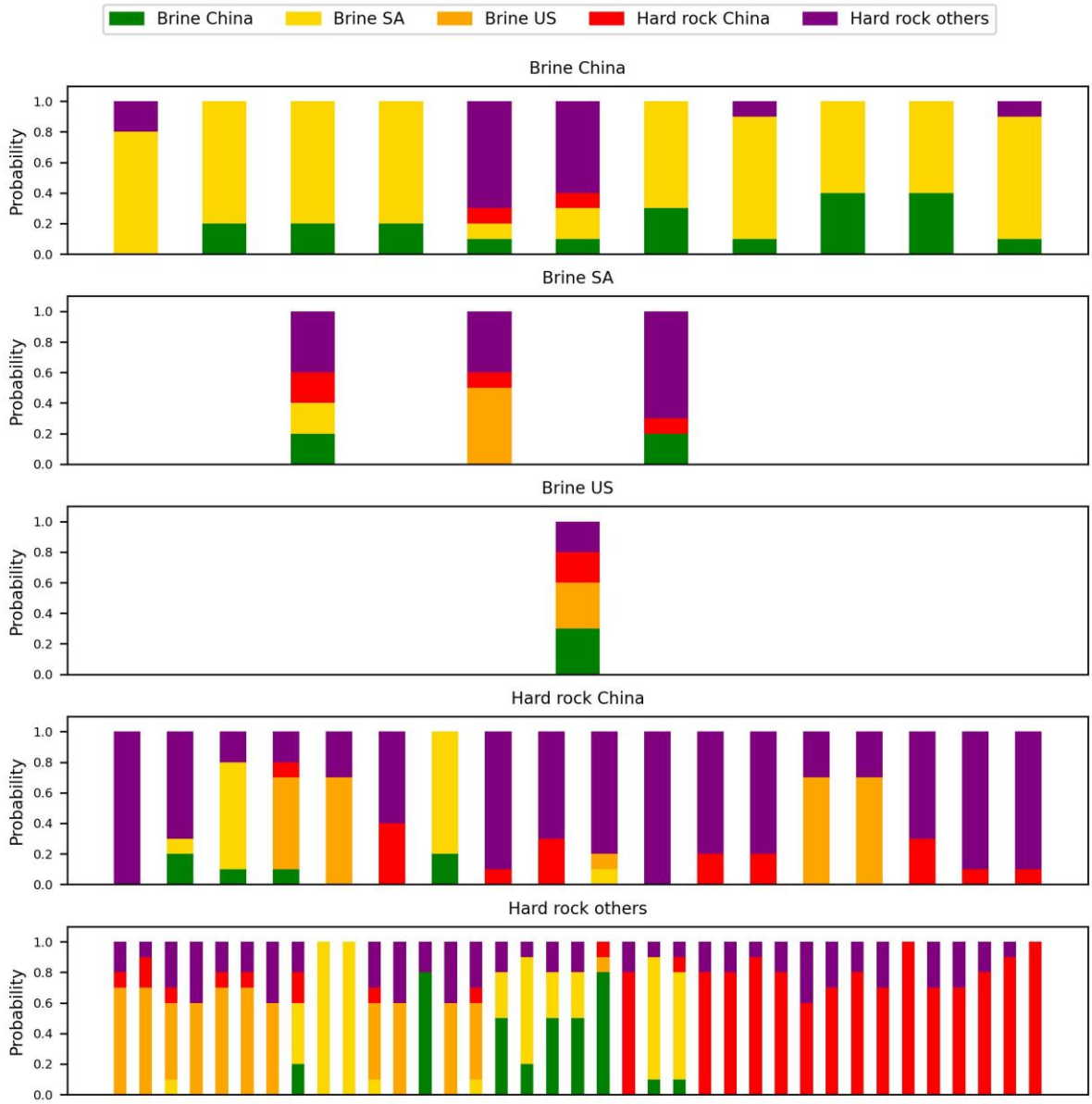

Figure 15. Probability of belonging to each class for samples incorrectly classified by the KNN.

3.1.4.3 Deposit determination

In this section we assume that the coarse class is known and we detail the capabilities of the four machine learning algorithms to classify the samples across the deposits. As in the previous section, we discuss the capabilities of the best classifier to provide an estimate of the uncertainty surrounding the predicted origin.

As highlighted by Table 1, the four models succeed in discriminating the samples coming from the Qaidam basin from those coming from the Tibetan Plateau. They also succeed in predicting the origin of the samples belonging to the class Hard rock others, KNN showing slightly better performance than the other classifiers. However, none of the classifiers succeeds in discriminating the samples coming from the deposits belonging to the class Hard rock China. The apparently good accuracy results from the uneven number of samples between the two deposits.

Table 1. Accuracy of the four machine learning algorithms for the classification of samples across the deposits belonging to each of the coarse classes.

Classifier	Brines China	Brines SA	Hard rock China	Hard rock others
NB	0.96	**0.62**	0.74	0.92
SVM	**1.00**	0.59	**0.82**	0.93
KNN	**1.00**	0.54	0.76	**0.94**
RF	**1.00**	0.57	0.71	0.90

As highlighted in Figure 16 most of the samples coming from the Huashan deposit are assigned to the Jiajika deposit. Finally, all classifiers manage, with slightly different levels of accuracy, to identify samples coming from the salars of Los Angeles and of Atacama. However, most of the samples coming from the salars of Olaroz and Ratone are incorrectly assigned to the salar of Atacama (Figure 16).

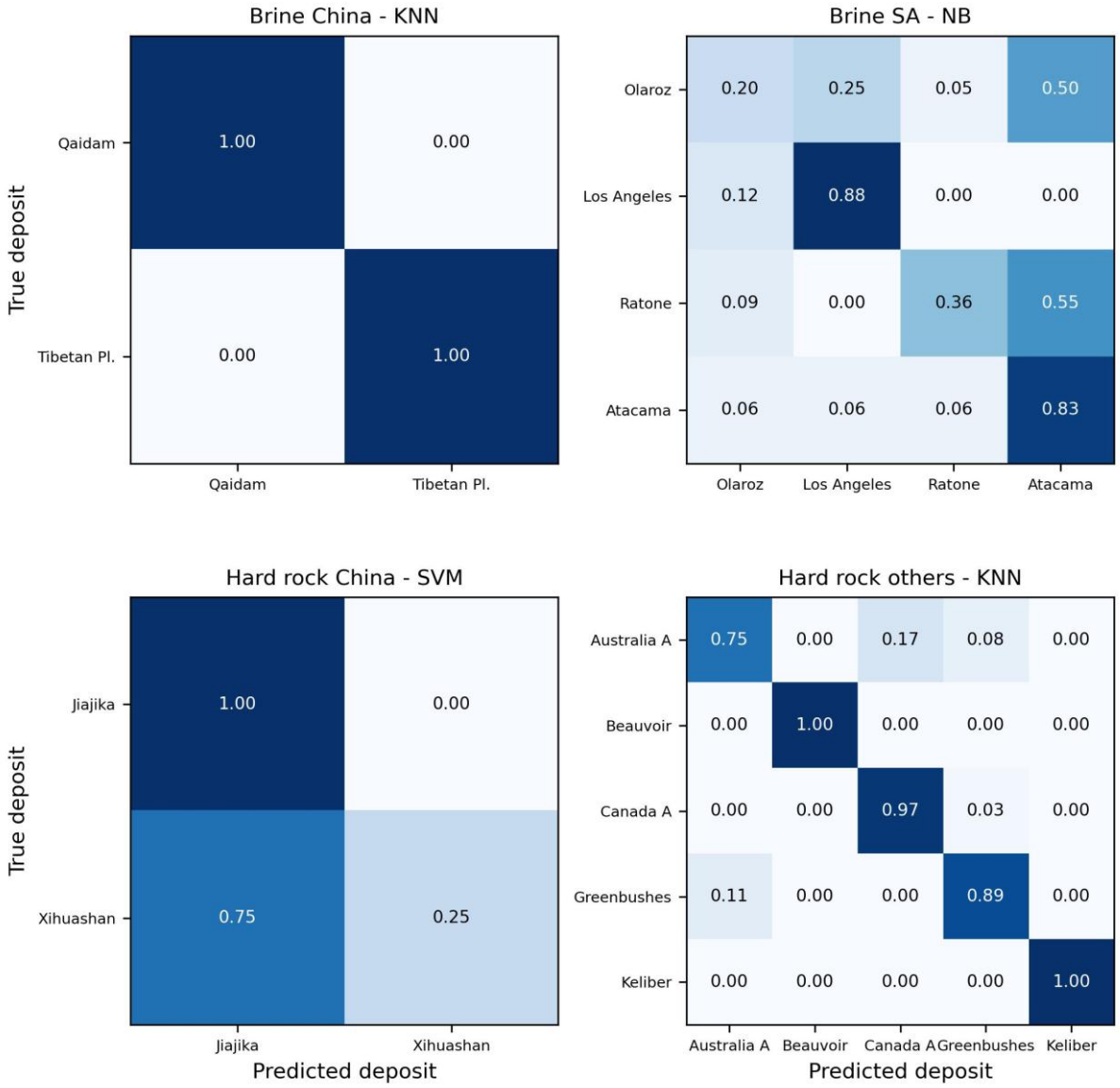

Figure 16. Confusion matrices describing the capacity of the best machine learning algorithm to classify the samples across the deposits of the four coarse origins.

Funded by the European Union

3.2 Classification based on a multivariate dataset: assessing the discriminatory power of X-ray fluorescence (XRF)

Authorship: Théophile LOHIER, Alban MORADELL-CASELLAS, Daniel MONTFORT, Anne-Marie DESAULTY

3.2.1 Context and main goal

In this work we propose an approach for the traceability of lithium hard rock deposits ores and ore concentrates based on widespread and easy-to-implement geochemical analysis; coupled with classification methods adapted to the specifications and needs of the approach. We have analysed around a hundred samples from lithium ores to mineral concentrates coming from world-class deposits, prospects and mines with portable and mobile X-ray Fluorescence (XRF) instruments for the analysis of major, minor and trace elements. Portable and mobile XRF instruments are already widely deployed across industrial facilities and are relatively low-cost, making them particularly suitable for large-scale and routine applications. A lithium ore or mineral concentrate could be characterized with these devices at mine sites, transit places (ports, freight trains stations ...) and at refineries entrance with little preparation and person training. We use and compare two handheld portable XRFs and a mobile XRF which analyses under vacuum, in terms of analytical accuracy and origin verification of the materials. The representativity and the sample preparation effects on the analytical results are studied. We compile for this purpose an analytical database which serves as the basis for statistical treatment.

3.2.2 Data Acquisition

A total of 93 samples coming from 14 origins of lithium deposits around the world (Figure 17) were analysed, as well as 8 samples from unknown origins. The types of these samples range from ores to mine concentrates, with intermediate processing products (rougher or cleaner flotation recoveries) referred to as pre-concentrates (see Table 2 for details). The number of samples available per deposit is uneven, ranging from one to thirty-six.

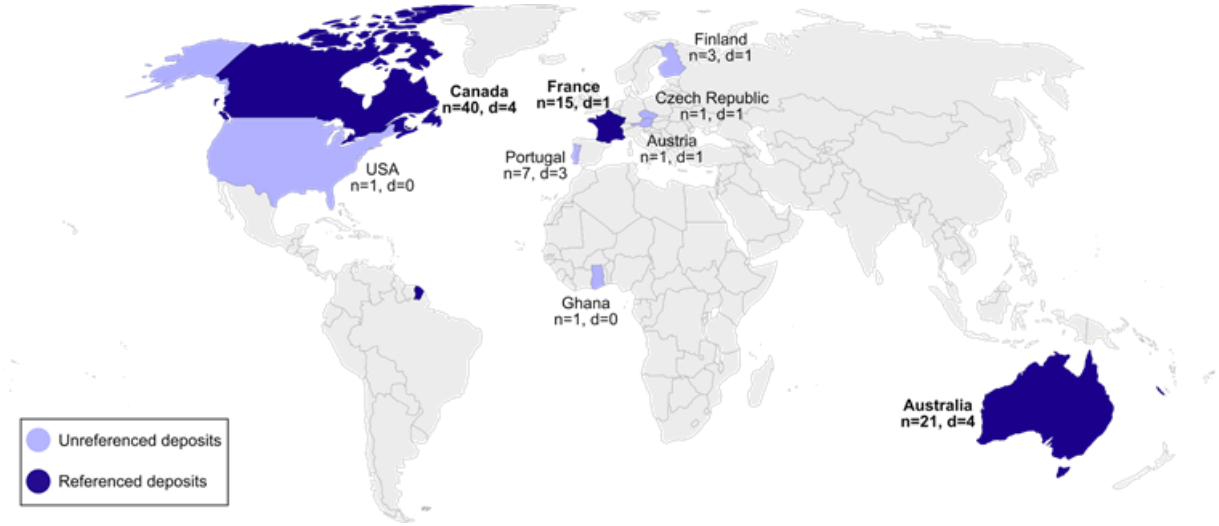

Figure 17. World map of the origin of samples investigated.

The main lithium-bearing minerals present in the samples are spodumene ($LiAlSi_2O_6$), lepidolite, $(K(Li,Al)_3(Si,Al)_4O_{10}(F,OH)_2)$ zinnwaldite $(KLiFeAl(AlSi_3)O_{10}(F,OH)_2)$, petalite

($LiAlSi_4O_{10}$) and minor cookeite ($LiAl_4(Si_3Al)O_{10}(OH)_8$) They are associated with alkali and plagioclase feldspars, quartz, muscovite, biotite and accessory minerals such as chlorites, garnets, apatite, amphiboles, tourmaline or Nb/Ta/W/Sn oxides (respectively columbite, tantalite, wolframite and sphalerite).

Table 2. List of Lithium samples investigated. Samples belonging to referenced deposits used for the model training are highlighted in yellow.

Sample ID	Country	Deposit/Mine	Nature	Li-bearing mineral	Sample ID	Country	Deposit/Mine	Nature	Li-bearing mineral
Au-1	Australia	A	Concentrate	Spodumene	Ca-35	Canada	A	Concentrate	Spodumene
Au-2	Australia	A	Concentrate	Spodumene	Ca-36	Canada	A	Concentrate	Spodumene
Au-3	Australia	A	Concentrate	Spodumene	Fr-1	France	A	ore	Lepidolite
Au-4	Australia	A	Concentrate	Spodumene	Fr-2	France	A	Concentrate	Lepidolite
Au-5	Australia	A	Concentrate	Spodumene	Fr-3	France	A	ore	Lepidolite
Au-6	Australia	A	Concentrate	Spodumene	Fr-4	France	A	Pre-concentrate	Lepidolite
Au-7	Australia	A	Pre-concentrate	Spodumene	Fr-5	France	A	Pre-concentrate	Lepidolite
Au-8	Australia	A	Concentrate	Spodumene	Fr-6	France	A	Pre-concentrate	Lepidolite
Au-9	Australia	A	Concentrate	Spodumene	Fr-7	France	A	Pre-concentrate	Lepidolite
Au-10	Australia	A	Pre-concentrate	Spodumene	Fr-8	France	A	Pre-concentrate	Lepidolite
Au-11	Australia	A	Pre-concentrate	Spodumene	Fr-9	France	A	Pre-concentrate	Lepidolite
Au-12	Australia	A	Concentrate	Spodumene	Fr-10	France	A	Pre-concentrate	Lepidolite
Ca-1	Canada	A	Concentrate	Spodumene	Fr-11	France	A	Pre-concentrate	Lepidolite
Ca-2	Canada	A	Concentrate	Spodumene	Fr-12	France	A	Pre-concentrate	Lepidolite
Ca-3	Canada	A	ore	Spodumene	Fr-13	France	A	Pre-concentrate	Lepidolite
Ca-4	Canada	A	Pre-concentrate	Spodumene	Fr-14	France	A	Concentrate	Lepidolite
Ca-5	Canada	A	Pre-concentrate	Spodumene	Fr-15	France	A	Concentrate	Lepidolite
Ca-6	Canada	A	ore	Spodumene	Li76	?	?	Concentrate	Spodumene
Ca-7	Canada	A	Pre-concentrate	Spodumene	Li77	?	?	Concentrate	Spodumene
Ca-8	Canada	A	Concentrate	Spodumene	Li74	Australia	?	Concentrate	Spodumene
Ca-9	Canada	A	Concentrate	Spodumene	Li75	Australia	?	Concentrate	Spodumene
Ca-10	Canada	A	Concentrate	Spodumene	Li78	Australia	?	Concentrate	Spodumene
Ca-11	Canada	A	Concentrate	Spodumene	Li43a	Australia	B	Concentrate	Spodumene
Ca-12	Canada	A	ore	Spodumene	Li43b	Australia	B	Concentrate	Spodumene
Ca-13	Canada	A	ore	Spodumene	Li43c	Australia	B	Concentrate	Spodumene
Ca-14	Canada	A	ore	Spodumene	Li43d	Australia	B	Concentrate	Spodumene
Ca-15	Canada	A	ore	Spodumene	Li43e	Australia	B	Concentrate	Spodumene
Ca-16	Canada	A	Pre-concentrate	Spodumene	Li69	Australia	C	Concentrate	Spodumene
Ca-17	Canada	A	Pre-concentrate	Spodumene	Li33	Austria	A	Concentrate	Spodumene
Ca-18	Canada	A	Pre-concentrate	Spodumene	Li21	Canada	B	Concentrate	Spodumene
Ca-19	Canada	A	Pre-concentrate	Spodumene	Li28	Canada	C	Concentrate	Spodumene
Ca-20	Canada	A	Pre-concentrate	Spodumene	Li34	Canada	D	Concentrate	Spodumene
Ca-21	Canada	A	Pre-concentrate	Spodumene	Li72	Canada	?	Concentrate	Spodumene
Ca-22	Canada	A	Pre-concentrate	Spodumene	Li50	Czech Republic	A	Tailing conc.	Zinnwaldite
Ca-23	Canada	A	Pre-concentrate	Spodumene	Li20	Finland	A	Concentrate	Spodumene
Ca-24	Canada	A	Pre-concentrate	Spodumene	Li57	Finland	A	Concentrate	Spodumene
Ca-25	Canada	A	Concentrate	Spodumene	Li60	Finland	A	Concentrate	Spodumene
Ca-26	Canada	A	Concentrate	Spodumene	Li73	Ghana	?	Concentrate	Spodumene
Ca-27	Canada	A	Concentrate	Spodumene	Li49	Portugal	A	ore	Lepidolite
Ca-28	Canada	A	Concentrate	Spodumene	Li58	Portugal	B	ore	Spodumene
Ca-29	Canada	A	Concentrate	Spodumene	Li66	Portugal	B	ore	Spodumene
Ca-30	Canada	A	Concentrate	Spodumene	Li67	Portugal	B	Concentrate	Spodumene
Ca-31	Canada	A	Concentrate	Spodumene	Li79	Portugal	B	ore	Spodumene
Ca-32	Canada	A	Concentrate	Spodumene	Li80	Portugal	B	Concentrate	Spodumene
Ca-33	Canada	A	Concentrate	Spodumene	Li59	Portugal	C	ore	Petalite
Ca-34	Canada	A	Concentrate	Spodumene	Li71	USA	?	Concentrate	Spodumene

The high inherent heterogeneity of these deposits in terms of structures, textures and mineralogy makes it difficult to assess the representativeness of a sample. Over the course of a mine's operational life, the extracted material is likely to be subject to significant variations. The inclusion of different beneficiation steps (ores, pre-concentrates and concentrates) in this work aims to provide the model with a wider range of mineral and thus chemical compositions for each deposit. This should facilitate the model's capacity to differentiate deposits based on their unique characteristics rather than their inherent variability.

Funded by the European Union

3.2.3 Strategy

We aim at building a statistical approach allowing to determine if a sample is coming from a known deposit, with the associated degree of probability for each referenced deposit. Classical supervised classification algorithms seek to assign each sample to a known class and do not provide an effective way to deal with heterogeneous under-represented samples that do not belong to known classes. In our use case, we will inevitably face samples coming from unreferenced deposits with properties that have never been seen by the classification model. We aim at developing a method able to deal with these samples from heterogeneous unreferenced deposits.

To this end, we decided to develop a robust and credible traceability method, using data sets comprising at least ten samples to train the classification model. These three deposits originate from a Canadian spodumene pegmatite (Canadian deposit "A", referred as Ca-1 to Ca-36), an Australian spodumene pegmatite (Australian deposit "A", corresponding to Au-1 to Au-12 samples) and a lepidolite rare metal granite deposit from France (Fr-1 to Fr-15); (Table 2). These deposits will be referred to as referenced deposits. Other samples and deposits, not used for model learning, will be designated as unreferenced samples or deposits.

The concentrations of each element are standardized (centred and reduced) before model training. The standardization is then applied to samples not used in the training before their projection in the latent space.

3.2.3.1 Method description

The LDA method is a classical classification algorithm that enables to build a latent space in which the samples from the different classes are grouped in clusters, ensuring that the distance between samples from the same class is minimal while the distance between clusters is maximal. We propose to take advantage of the latent space built with LDA to evaluate the similarities between a sample of unknown origin (or unreferenced samples) and the referenced deposits. First, we build a latent space with LDA using the samples coming from the most represented deposits ("Ca", "Au" and "Fr"). Then, we compute the position of the centroids of these deposits in the latent space. Finally, we compute the Euclidian distance between the unknown samples and each of the centroids. The main issue with this approach is that the compactness of the clusters is likely to vary significantly depending on the number and on the consistency of the available samples used to perform LDA. Because of this instability, it is very difficult to define a distance threshold below which an unknown sample would be systematically attached to a referenced deposit. And it is even more difficult to estimate the uncertainty associated with this decision. To overcome this limitation, we propose to couple LDA with bootstrapping. Bootstrapping is a resampling procedure that enables, from a limited number of samples, to estimate the distribution of an estimator. Here we use bootstrapping to (i) characterize the distribution of the distance between samples belonging to a deposit and the centroid of the deposit in the latent space built with LDA; (ii) characterize the distribution of the distance between an unknown sample and the referenced deposit centroids.

In practice, we construct a training subset including 80% of the samples of each referenced deposit. This data set is used to build a latent space with LDA and the centroid position of each deposit is computed. The subset of samples from reference deposits not used for the construction of the latent space is then projected in the latent space and the Euclidian distance between each sample and the deposit it is belonging to is estimated. In parallel, the distance between the unknown samples and the centroid of each deposit is estimated. Repeating this step allows us to estimate the distribution of both the distance between the reference samples and the centroid of the deposit they belong to, and the distance between

the unknown samples and the centroids of the different deposit. Based on distributions of these two distances, we propose an acceptance criterion for reference deposits. This criterion relies on the comparison of the third quartile (Q3) of the distributions. A difference in the Q3 smaller or equal to 0 means that the distance of the unknown sample to the centroid deposit is smaller than the highest distances observed for the samples known to belong to the deposit in at least 75% of the latent spaces built with LDA. In this case, illustrated in the left panel of Figure 18, the unknown sample will be accepted. On the contrary, a difference in Q3 significantly higher than 0 means that the distance of the unknown sample to the centroid deposit is higher than the highest distances observed for belonging samples (Figure 18, centre panel).

The use of the third quantile enables us to exclude abnormally high distances to deposit centroids, resulting from outlying samples and/or dubious latent representations. Regarding the distribution of distances of the unknown samples, the choice of the percentile also determines the confidence level for the rejection. The use of higher percentiles will lead to systematic rejection of uncertain samples, while lower percentiles will allow for broader acceptance (Figure 18, right panel).

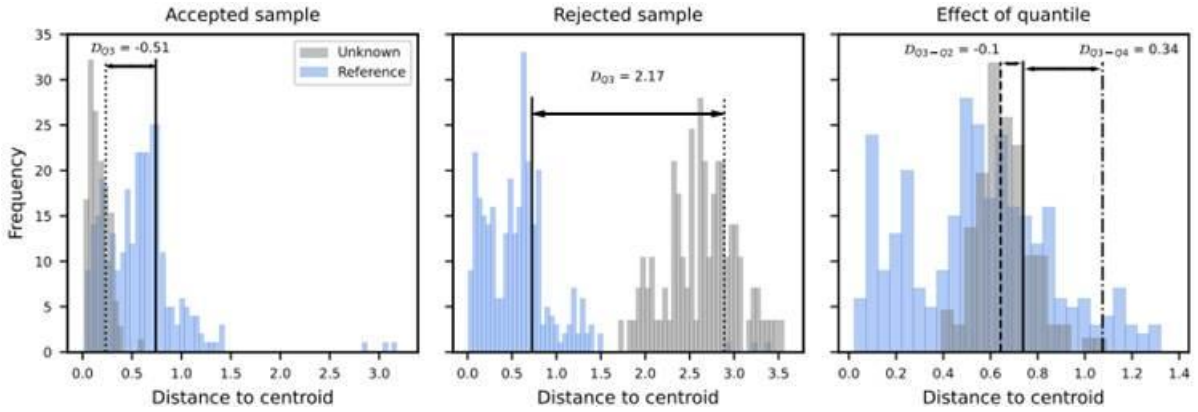

Figure 18. Comparison of the distribution of the distances between samples belonging to a deposit and the centroid of the deposit (in blue) with the distribution of the distances between the deposit centroid and an unknown sample (in grey).

To explore the impact of the confidence level on the model estimations, we propose to investigate the variations in differences between the Q3 of reference and unknown sample distance distributions rather than varying the percentile used in the approach. This investigation is performed using the ROC (Receiver Operating Characteristic) curve that depict the proportion of samples belonging to a reference deposit that are correctly accepted (true positive rate) against the proportion of samples not belonging to the reference deposit that are incorrectly accepted (false positive rate) according to the maximum difference in Q3 allowed.

3.2.3.2 Evaluation

To evaluate the ability of our approach to identify the samples belonging to referenced deposits, we use leave-one-out (LOO) cross validation. Each sample belonging to a referenced deposit is sequentially removed from the training set, and the distribution of its distance to the centroids of the referenced deposit is estimated with the bootstrapped-LDA procedure. This distribution is compared to the distance distribution of samples belonging to the reference deposit using the difference between Q3 (Figure 18).

Each time a sample belonging to a referenced deposit is processed, all the samples belonging to unreferenced deposits are also processed: the distribution of the distances between these samples and the centroid of the reference deposit are estimated and they

Funded by the European Union

are compared to the reference distributions using the Q3 difference. For each sample belonging to unreferenced deposits, we therefore obtain as many replicas as there are samples belonging to the referenced deposit.

For each referenced deposit, we compute the true positive rate (TPR), which is the proportion of samples belonging to a referenced deposit that are correctly accepted, and the false positive rate (FPR), which is the proportion of samples belonging to unreferenced deposits that are incorrectly accepted, for various Q3 thresholds. We use the ROC curve to investigate the effect of the Q3 threshold and select the value leading to the best trade-off between TPR and FPR for each deposit.

Based on the best Q3 thresholds obtained, we produce confusion matrices that depict the classification results. In this representation, the diagonal describes the proportion of samples that are correctly attached to each deposit. We include an artificial deposit, labelled unreferenced, that includes all the samples not belonging to one of the referenced deposits. The last row of the matrix therefore gives the proportion of samples that are incorrectly attached to one of the referenced deposits, while the last column gives the proportion of samples that are incorrectly rejected.

Besides this synthetic representation, we provide a more exhaustive one, which describes the Q3 differences between the distance distribution of each sample and the reference distribution of the deposit they belong to during the LOO cross-validation procedure. These values are compared to the Q3 differences obtained for samples belonging to the other referenced deposits as well as for the samples not belonging to any referenced deposits. It enables us to identify the samples that are incorrectly accepted or rejected, and the samples that are associated with high uncertainty.

Given the small number of samples available for training the LDA, it could be beneficial to reduce the number of input variables. We investigated the impact of the number of chemical elements used for the construction of the latent spaces on the classification results. We build several subsets of elements, based on their feature importance, computed by running bootstrapped-LDA on a dataset including all the samples belonging to referenced deposit. For each of the subsets, we run the full procedure and evaluate the overall proportion of correctly accepted and rejected samples.

3.2.4 Results

3.2.4.1 Classification reliability

For mXRF analysis, the number of elements used to build the latent space has a strong impact on the performance of the classifier with a significantly higher True Negative Rate (TNR) when the number of elements is between 10 and 12 (Figure 19.A). For pXRF analysis, increasing the number of elements used in LDA does not have a large impact on the quality of the classification model, even if the TNR slightly decreases when the number of elements included increases over eleven (Figure 19.B). For pXRF in REE-mode, it has a larger impact with the TPR (True Positive Rate) decreasing when the number of elements included increases over eleven (Figure 19.C).

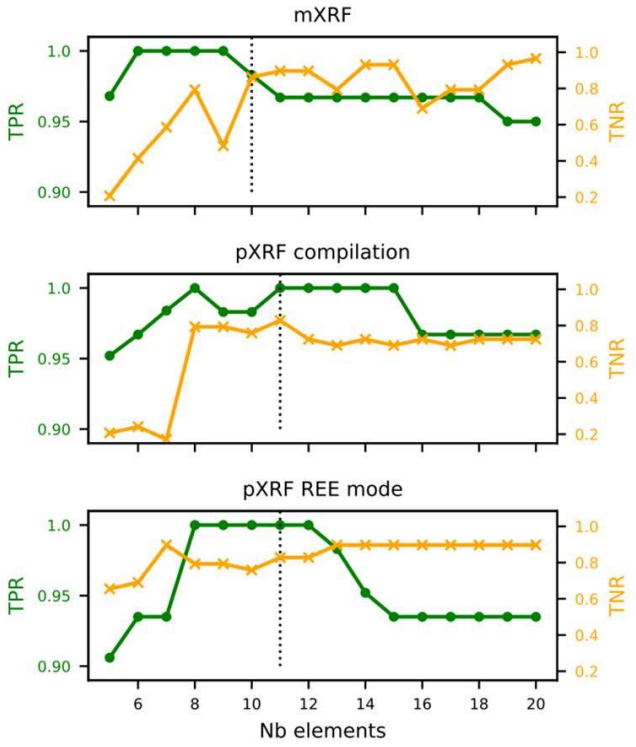

Figure 19. Impact of the number of elements used to build the LDA latent space on the classifier performance, evaluated using the true positive rate (TPR) and the true negative rate (TNR) for mxrf (A), pXRF (B) and pXRF in REE-mode (C) analysis.

Sample origin prediction of the models is shown in the Figure 20. All the Australian and Canadian referenced deposit samples are correctly attached to their respective origins with the three methods. Samples belonging to the French deposit are also correctly predicted with pXRF analysis, but one of the samples, standing for 7% of the samples, is falsely rejected, using mXRF analysis.

Regarding the rejection of samples not belonging to referenced deposits, the three methods give similar results. With mXRF, four samples not belonging to the referenced deposits are incorrectly accepted, accounting for 14% of the unreferenced samples (Figure 20). Results obtained with pXRF compilation and pXRF in REE-mode are similar with two samples incorrectly accepted in Canadian and French deposits (corresponding to 7% of samples), and one sample incorrectly accepted in Australian deposit (3%).

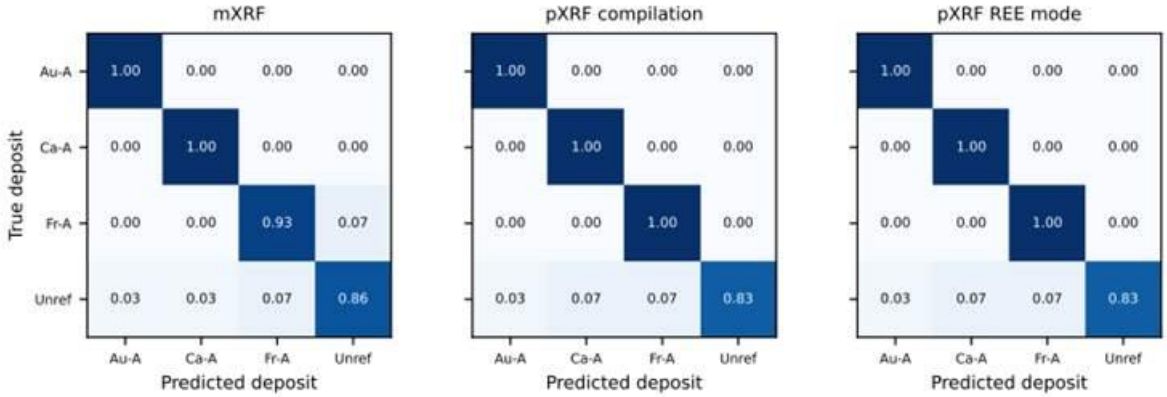

Figure 20. Confusion matrices resulting from models trained with mXRF, pXRF mode compilation and pXRF REE-mode analysis.

Funded by the European Union

3.2.4.2 Evaluating the uncertainty

In order to further explore these results and to understand, if possible, the uncertainty of the method and causes of incorrect predictions, we studied sample distances to referenced deposit centroids (Figure 21, Figure 22 and Figure 23).

Overall, the difference between the third quartile (Q3) of the distribution of sample distances to the centroid of the deposit they belong to and the reference Q3 is less than or close to zero. On the other hand, for the samples belonging to a different referenced deposit, this difference is systematically higher than one with median values across samples higher than 2.0, confirming the ability of the LDA to separate samples belonging to referenced deposits.

The difference between the Q3 of the sample distance distribution to the referenced deposit centroids and the reference Q3 gives an idea of the uncertainty surrounding the rejection: the greater the difference, the lower the uncertainty. The variation of these differences over the folds of the LOO CV (cross-validation), described by the error bars in red, also gives some insights on the uncertainty with important variations being symptomatic of a fuzzy characterization of the referenced deposit.

For example, with mXRF, the differences in Q3 for samples not belonging to the Canadian deposit are significantly higher than the rejection threshold, and the variations of the differences over the folds are small for most of the samples. This indicates a good qualification of the reference deposit and a strong rejection of the unreferenced samples. In contrast, for the Australian deposit, the error bars are much larger, indicating a less accurate qualification of the deposit. Moreover, for several unreferenced samples, the difference in Q3 is close to the rejection threshold (e.g. Li43a-e), which suggests strong similarities with the samples belonging to the referenced deposit.

With regards to the Canadian referenced deposit, Sample Li21 (Canada deposit B) is a recurrent false positive with mXRF and pXRF. The Q3 of the distribution of the distances to the Canadian centroids is systematically smaller than the reference Q3, suggesting a strong similarity with the samples belonging to this deposit. This spodumene concentrate sample is sourced from a distinct deposit, with is situated in close proximity (within a few kilometres) to the primary deposit. All Canadian deposits used for this study are from Archean age, in the Superior province of Québec, meaning that this sample is geologically closely related to the reference deposit.

Sample Li59 (Portugal deposit C) is also predicted with the two methods as the French referenced deposit. The median Q3 difference is very close to the rejection threshold, indicating significant differences with most of the samples belonging to this deposit. Nevertheless, this result is also correlated to the geological similarity between the two deposits. In this case, the sample belongs to a Portuguese deposit. Lithium deposits formation both in Portugal and France (pegmatites and rare metal granites) is related to the Hercynian orogeny. This suggests that, for this sample as well, the recurring false positive could be related to the similar geological context of both deposits.

Lastly, the Q3 of the distance distribution to the centroid of the Australian deposit of sample Li74 (Australian deposit of unknown origin) is consistently smaller than the reference Q3 when the latent space is built from the pXRF analysis, whereas mXRF manages to distinguish the two. All Australian lithium deposits are formed within Archean age Pilbara and Yilgarn cratons. This reinforces further the hypothesis that was discussed in the preceding paragraphs.

The geographic and geologic proximities of these recurring false positive samples (Li21 and Li59) and close or inferior to zero Q3 distance to the referenced deposits (Li21 and Li74) with the referenced deposits suggest that the models is sensible to the geological origins of these deposits, which can be expected to lead to elemental concentrations similarities. Whilst the model is having more difficulties in differentiating between closely related origins, this highlights the existence of geochemical fingerprints of different districts/regions.

No other sample consistently falls below the acceptance threshold with pXRFs and mXRF or present close distribution distances with reference deposits. These other false positives are therefore solely attributable to the limitations of the statistical model.

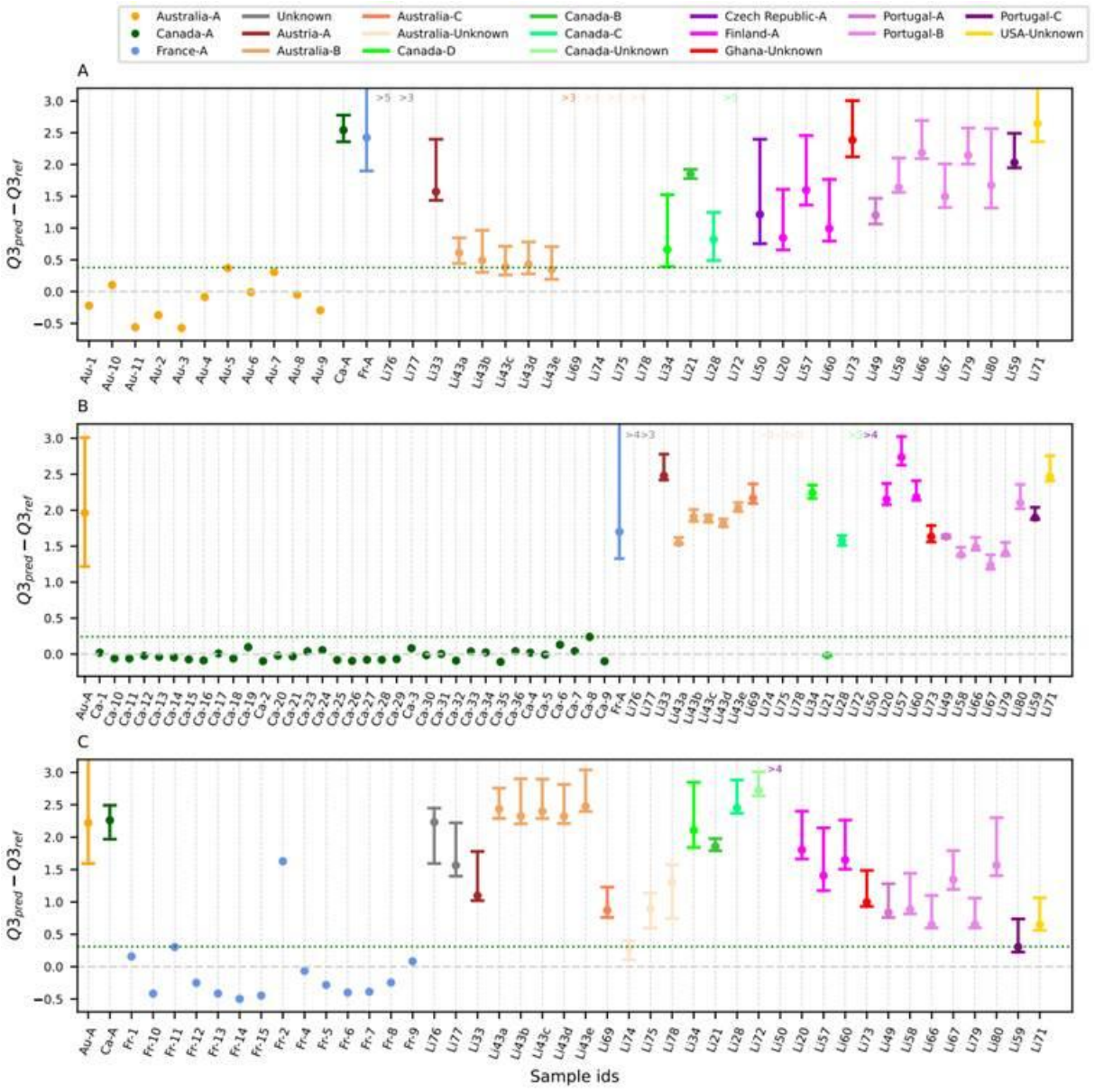

Figure 21. Comparison of reference and unknow sample distance distributions to the centroid of the Australian deposit (A), Canadian deposit (B), and French deposit (C) in the LDA space built from mXRF analysis.

Funded by the European Union

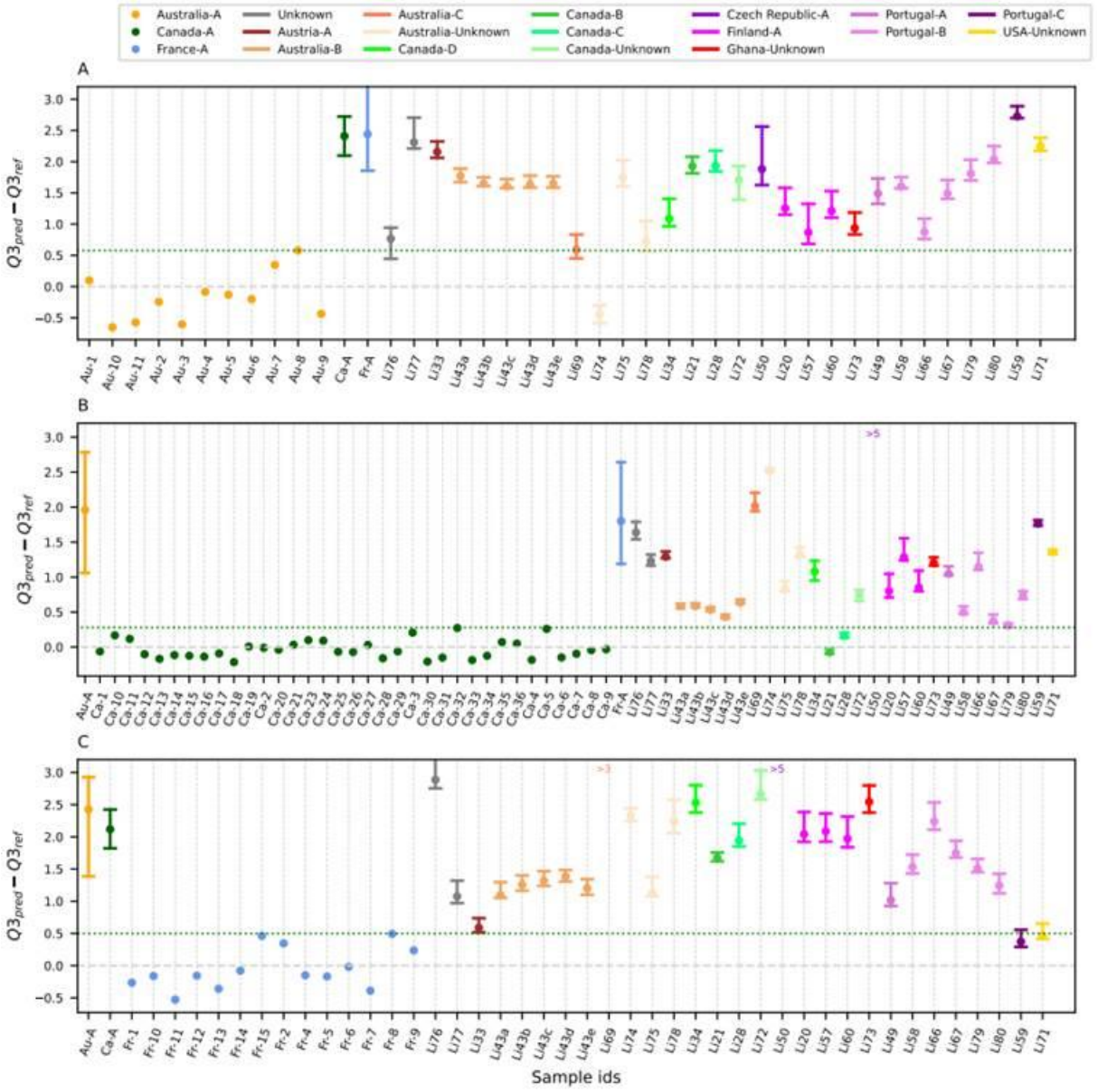

Figure 22. Comparison of reference and unknow sample distance distributions to the centroid of the Australian deposit (A), Canadian deposit (B) and French deposit (C) in the LDA space built from pXRF compilation of mode analysis.

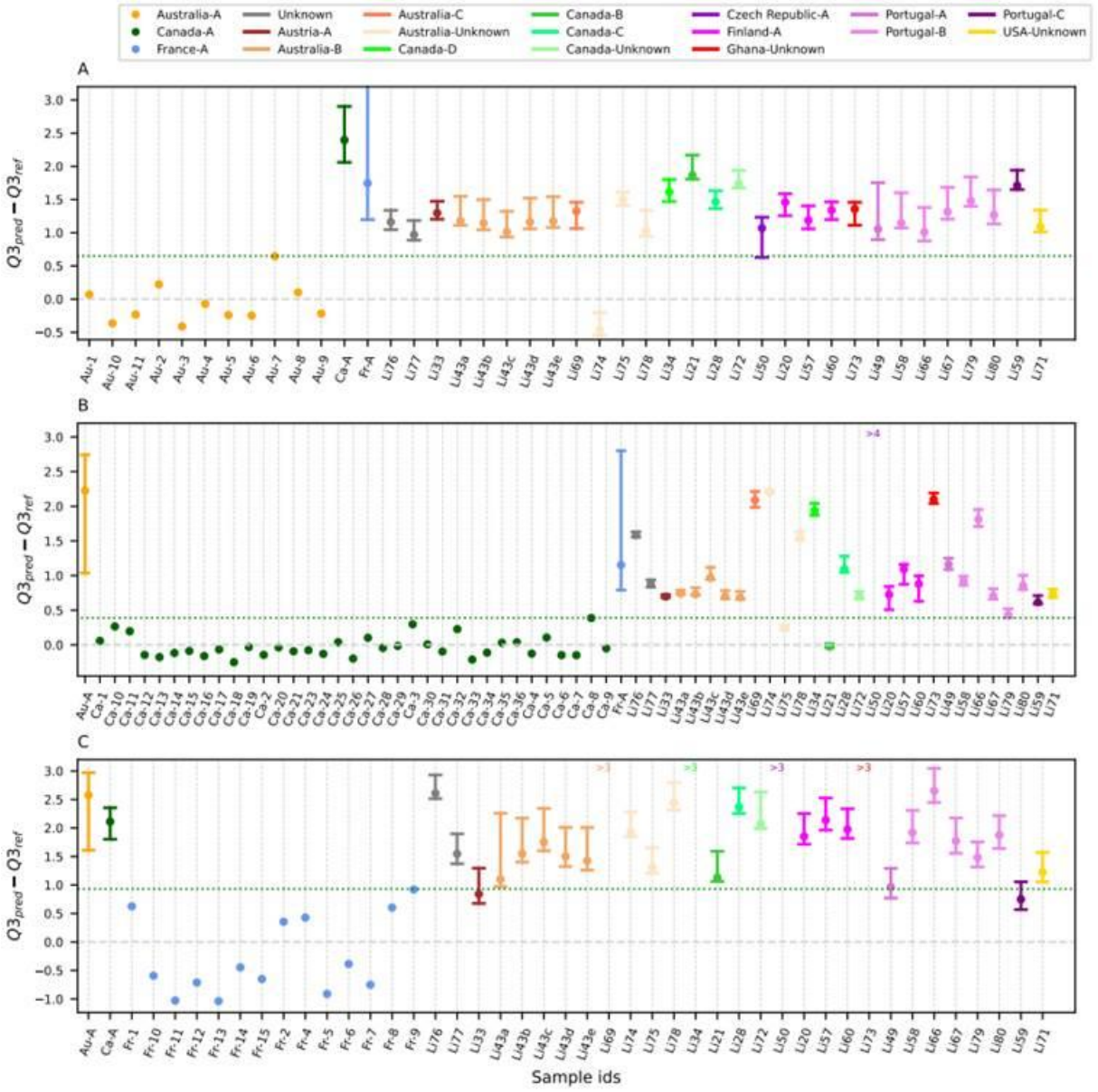

Figure 23. Comparison of reference and unknow sample distance distributions to the centroid of the Australian deposit (A), Canadian deposit (B) and French deposit (C) in the LDA space built from pXRF REE-mode analysis.

3.2.5 Conclusion

The two types of instruments (pXRFs and mXRF) used in this study give in terms of origin prediction very similar results. However, the better separation of one of the unreferenced deposits (not in use for model training) with mXRF may indicate that this technique is more efficient to distinguish other deposits. The comparison of the model built from pXRF REE-mode, and the one built from the compilation of modes demonstrate that REE-mode and thus a more limited number of elements can be used to achieve convincing results in a reduced amount of time. Nevertheless, the models were built using only three referenced deposits, which consequently requires fewer elements. Needing to identify more referenced deposits (and closely related ones) would perhaps require more elements and lower detection limits which could be achieved with instruments like the mXRF used in this work or with benchtop and laboratory XRF instruments.

Funded by the European Union

The method developed allows the prediction of the origin of the upstream part of the lithium supply chain, in the case of hard rock deposits. However, it is important to understand the level of certainty that can be attributed to the verification of provenance. The uncertainty surrounding the acceptance of an unknown sample can be quantified at two levels. The overall accuracy of the approach, estimated through the cross-validation procedure, provides an idea of the epistemic uncertainty of the classification model. For example, with mXRF analyses, the error rate associated with sample acceptance is less than 5% for the Australian and Canadian deposits and less than 10% for the French deposit. Here, the choice of classification model has been guided by the characteristics of the dataset. However, the developed framework is flexible and can be used to assess the ability of other types of models to discriminate the origin of lithium salts. The developed approach also allows, through the Q3 difference, to estimate the uncertainty at the sample level. A sample associated with a Q3 difference less than or equal to zero can thus be accepted with a high level of confidence. A Q3 difference approaching the acceptance threshold indicates a lower level of similarity with the other samples from the deposit and should alert the decision-maker. Furthermore, the rejection of samples belonging to unreferenced deposits and of unwanted mixes of different origins is key in order to detect possible frauds. The developed approach can identify those samples and estimate, through the Q3 difference, the uncertainty surrounding their rejection. A high Q3 difference indicates significant differences in the characteristic properties of the reference deposits, resulting in a rejection with a good level of confidence. Conversely, when the Q3 difference approaches the acceptance threshold, the rejection of the sample becomes more questionable. Here the acceptance threshold has been calibrated for each reference deposit using ROC curves to maximize the classifier's discrimination power. However, this threshold can be relaxed to reduce the risk of false rejection or tightened to reduce the risk of false acceptance, depending on the decision-makers' objectives.

Funded by
the European Union

3.3 Classification based on a spatialized multivariate dataset: application to Laser-Induced Breakdown Spectroscopy (LIBS)

Nathan BODEREAU, Théophile LOHIER, Alban MORADELL-CASELLAS, Damien DEVISMES, Lina JOLIVET, Florian TRICHARD, Claire AUPART, Nicolas GILARDI, Daniel MONFORT-CLIMENT, Anne-Marie DESAULTY.

This part of the work partly relies on data produced in the framework of the project LITHOS 'Cost-effective processing and refining of lithium into lithium hydroxide from strategic European multi-mineral lithium hard-rock projects' (Grant number 101138112) and was supported by the European Union's Framework Programme for Research and Innovation Horizon Europe.

3.3.1 Context and main goal

In this study, we propose a strategy to classify hard-rock lithium minerals and concentrates of known origins constituting the first step of the supply chain using LIBS mapping. Six deposits were considered: Australia-1, Australia-2, Canada-1, Finland, France and Portugal-1. The approach relied on tracking colocation of elements linked to impurities within the powders. Analyses were conducted using ABLASCAN, an innovative device developed by ABLATOM, capable of performing millions of measures in just 15 minutes, thereby producing so high-resolution chemical maps (Figure 24).

The main goal of this analysis is to provide spatial information of intrinsic patterns of deposits. While no spatial pattern is expected to be observed on 2D plan because the samples have been grinded, element superposition per pixels are likely to present key patterns that could be used for sample origin determination.

This section describes a strategy developed to process high-resolution chemical maps, detect key element combinations representative of the deposit of origin and build a classification model to predict sample origin. The objective is double:

- Find a model capable of processing high-resolution maps and classify them according or their origins
- Find a new set of key patterns based on element colocations representative of auxiliary mineral which are in turn representative of the origin

3.3.2 Data acquisition

Thirty-four elements were mapped including major, minor, and trace elements as summarized in Figure 24. Carbon (C), oxygen (O), and hydrogen (H) were not considered as contaminations with resin is expected. Also, the sensor used to measure sulphur (S) was changed during the project, compromising the consistency of the measurements, and it was therefore excluded from subsequent analyses. Finally, hafnium (Hf) was systematically below detection limits.

Funded by the European Union

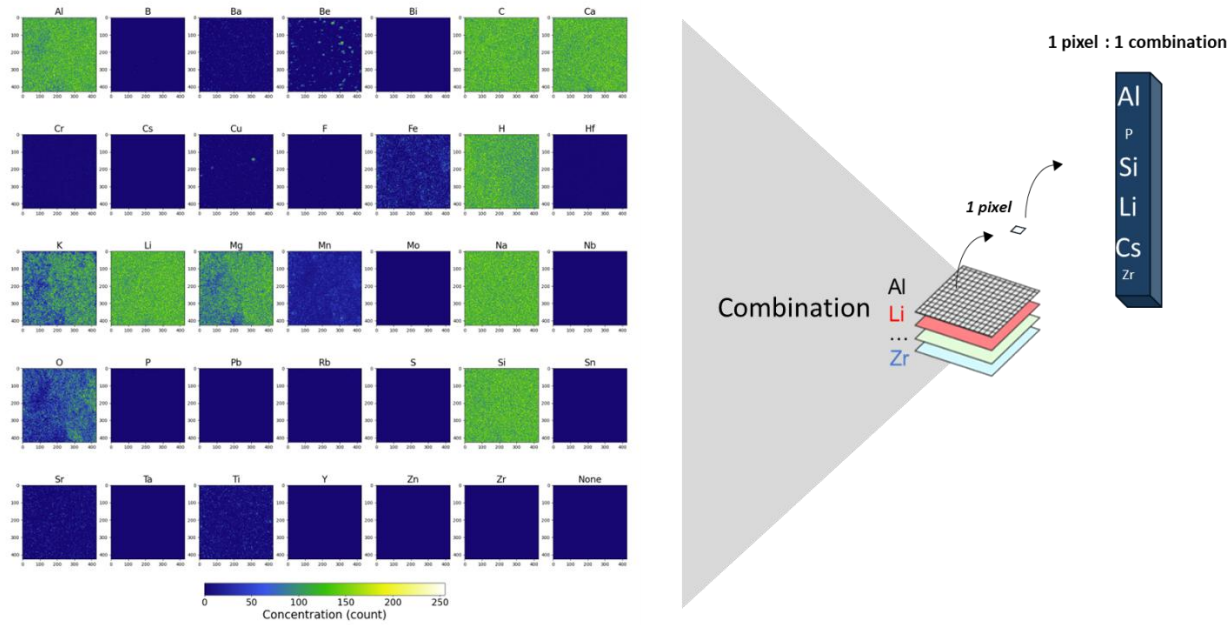

Figure 24. Example of visualisation of chemical maps per investigated element for a spodumene concentrate.

3.3.3 Strategy

The data processing workflow comprises three main steps: image reduction, model training and feature comparison (Figure 25).

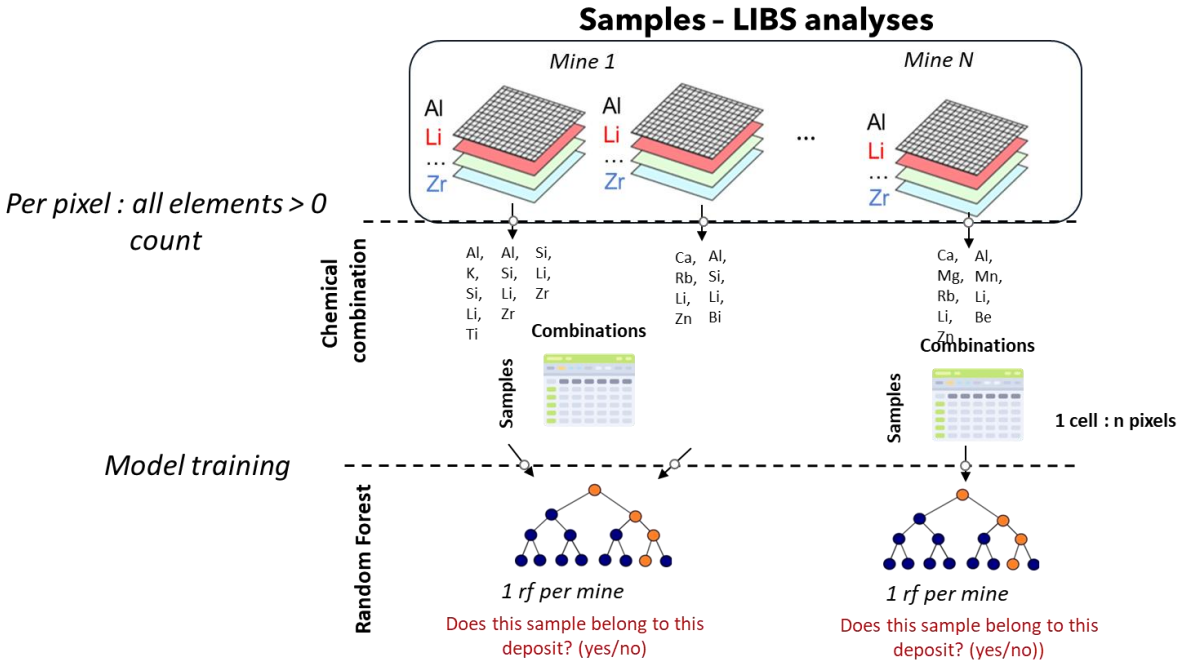

Figure 25. Strategy developed to LIBS maps' processing and classification.

3.3.3.1 Image reduction

Image processing usually begins with dimensionality reduction for an easier representation and to remove data redundancy (Green, Berman, Switzer, & Craig, 1988). Here, a threshold at 0 count was first applied to form a list of present elements per pixel. The results are then summarized into a dataframe where the rows represent samples, the columns, the combinations, and the values correspond to the number of pixels associated with each combination.

In a next step, simulations have been performed by considering another threshold based on the number of pixels covered by combination. It prevents the model against combinations with a negligible number of pixels. We have varied this threshold from 0 to 20,000 pixels by steps of 200 pixels. The threshold is then fixed to a value for which the number of combinations is acceptable and the accuracy of the model still high.

3.3.3.2 Classification model

Random forest has been selected for this task considering the limited number of training samples for the different deposits. Decision trees are also a family of supervised machine learning technique which consists of recursively splitting the dataset in a way to well separate one class from the others. Starting at a rooting node, the model applies successive feature-based thresholds until each class is isolated as a leaf (Morgan & Sonquist, 1963). While decision trees are simple and interpretable, they are also sensitive to noise and can easily overfit. Considering the limited number of training samples for the different deposits ($n<10$ for three out of the six deposits), Random Forest were adopted to address this limitation. This consists in an ensemble of decision trees trained on different subsets of randomly selected samples (called bagging) and its final prediction is the result of the vote of the different trees. This ensemble strategy reduces overfitting and improves predictive performance (Breiman L. , 1996).

One versus rest strategy was selected which consists in training one random forest for each deposit to distinguish it from all the remaining ones. A prediction of 'True' indicates that the sample belongs to the target deposit, while 'False' indicates otherwise. A key advantage of this method is that adding a new deposit does not interfere with previously trained models, which contrasts with traditional multiclass classification algorithms (Rifkin & Klautau, 2004).

The performance of classification has been assessed using the Leave-One-Out Cross-Validation (LOO-CV) strategy: each sample is left out once, all remaining samples are used for training, and the excluded sample is then predicted. The process is repeated for every sample, and the final performance corresponds to the mean accuracy.

3.3.3.3 Feature selection

In random forests, feature ranking generally includes biases associated to the effect of dataset bootstrapping especially when the prediction is a categorical variable (Strobl, Boulesteix, Zeileis, & Hothorn, 2007). For instance, some features are considered more important just because they have been randomly selected many times by bootstrapping. Furthermore, the features that present high importance do not necessarily discriminate a class from the others but has been part of a sequence of decisions including several features (Breiman L. , 2001).

To truly understand which elemental combination has been significant, visualisation of each tree of each forest should be required to know the different pathways combining several combinations. Several techniques have however been developed among which the most

Funded by the European Union

known is the SHapley Additive exPlanations (SHAP) (Lundberg & Lee, 2017). It consists in measuring the contribution of each combination to the final prediction following Shapley's theory. At each training, prediction receives a local importance coefficient, known as a SHAP value, which reflects how much a given combination pushes the model toward or away from a specific class.

3.3.4 Results

3.3.4.1 Classification reliability

Classification LOO-CV accuracies of the 6 training deposits are plotted in confusion matrix (Figure 26). All samples overall show good predictions with an average of 88.9 %. One misclassified sample is observed per deposit, except for Canada-1 where all sample were correctly predicted and for Portugal-1 where two samples were misclassified.

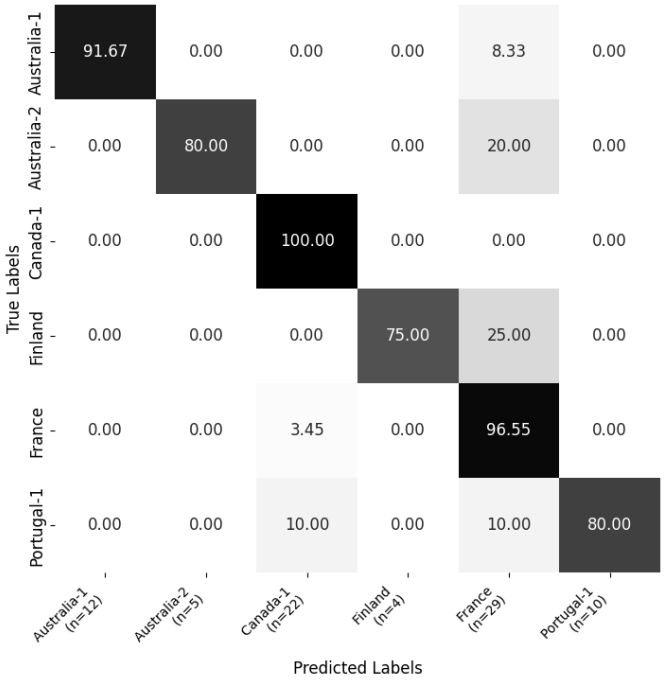

Figure 26. Confusion matrix (in %) of lithium deposits where at least four samples were available, obtained by LOO cross-validation.

On the other hand, we investigate the probabilities of belonging or not to each deposit (Figure 27). The median probabilities of True of random forest of for each class are 69.5%, 60% and 76.5 % for Australia-1, Australia-2 and Portugal-1. These scores reach 80.0 and 90 % for Canada-1 and France deposits: the higher the number of samples, the higher the scores. With these median probabilities, a certification could be assessed to know if a lithium salt belongs to a deposit or not.

Only the Finland-trained model displays a probability of 39 % which could be explained by the limited numbers (n=4). Moreover, the Finnish Li60 sample has a probability to belong to its true deposit that is close to zero.

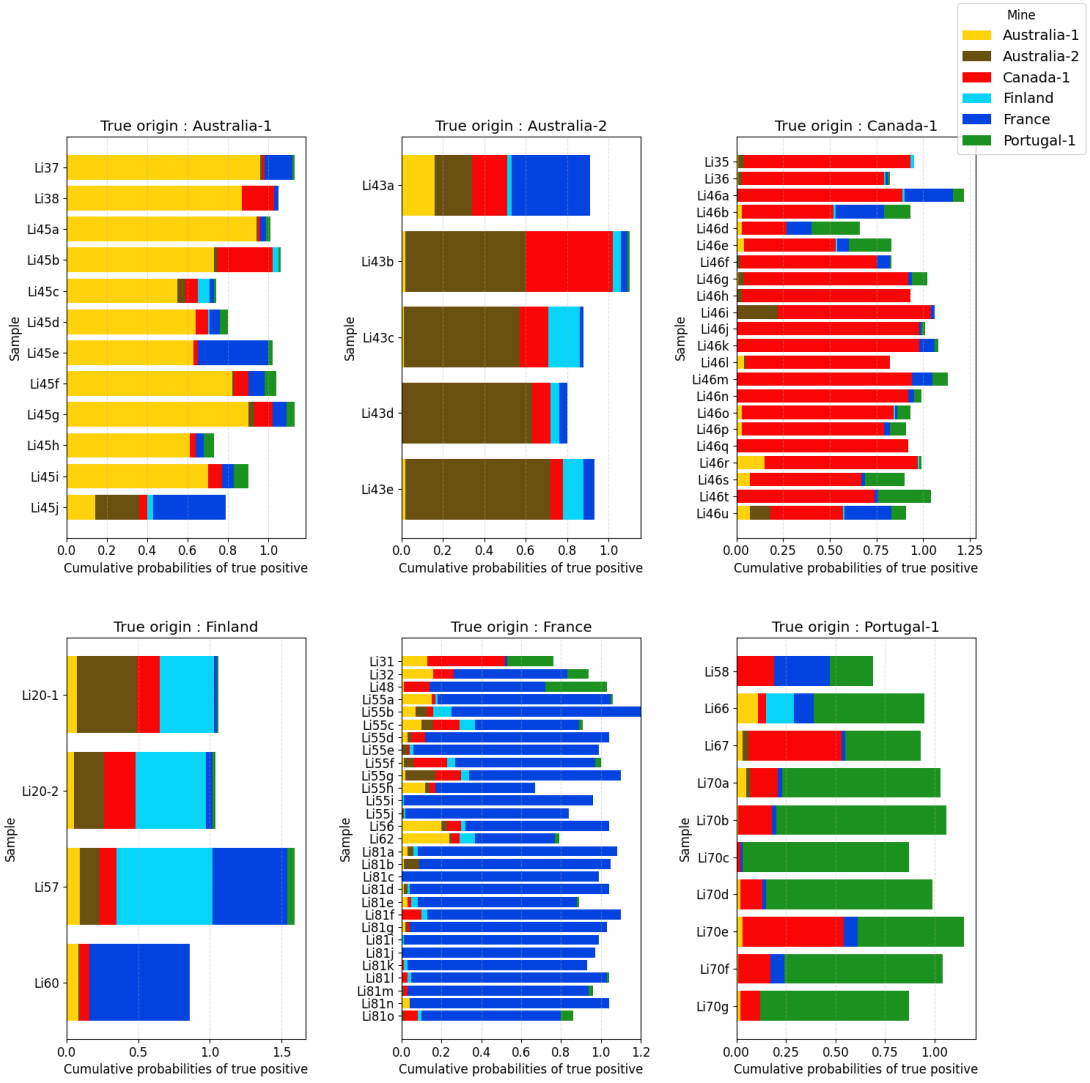

Figure 27. Predicted deposit probabilities per sample. Each subplot is related to the deposit. Finland group excepted, each model overall shows high probabilities to belong to their deposit.

Finally, we studied the models' ability to reject samples that do not belong to any of the deposits used during training. To do this, samples from underrepresented deposits (and therefore not included in the training) were submitted to the model. We compare the highest membership probability to the median membership probability of the samples actually belonging to the deposit.

For example, the median probability of belonging to the France deposit for samples actually belonging to this deposit is around 90% (Figure 27). For the samples belonging to the Austria and Portugal-2 deposits, the highest probability is the probability of belonging to the France deposit. This probability is around 40% (Figure 28), which is much lower than the median probability for the France deposit. We can therefore conclude that these samples do not belong to any of the referenced deposits with a good level of confidence. A similar behaviour is observed for samples from the Canada-2 and Middle Europe deposits, with maximum probabilities around 40% to belong to Australia-1.

Funded by the European Union

However, for samples belonging to the Australia-3 and Portugal-3 deposits, the probabilities of belonging to the Finland deposit are of the same order as the probabilities for samples actually belonging to that deposit. These samples would therefore be wrongly assigned to the Finland deposit. These classification errors are likely due to the poor performance of the Finnish model, which was built from too few samples (n = 4).

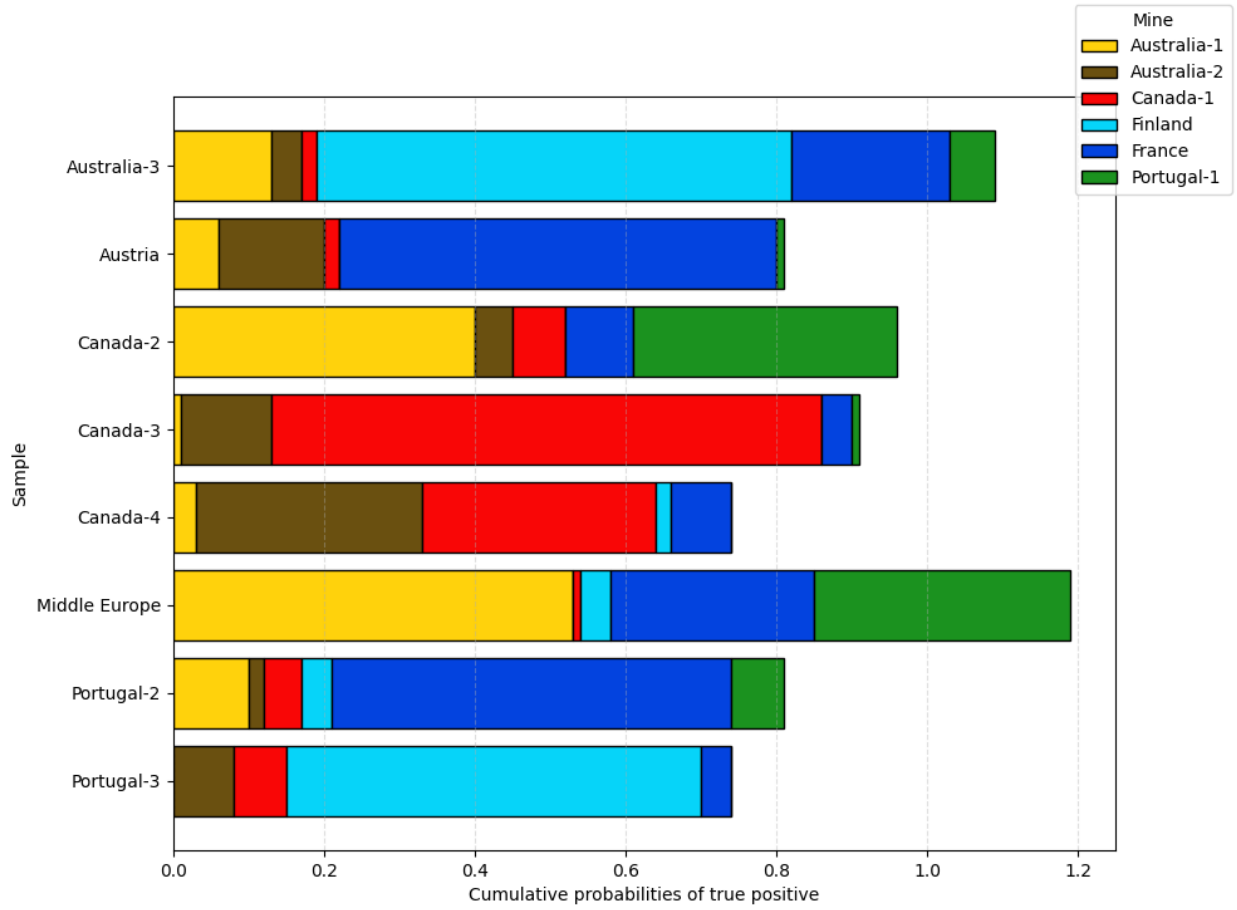

Figure 28. Predictions of probabilities for samples belonging to underrepresented deposits not used in the model training.

3.3.5 Conclusion

This study presents a machine learning framework using Random Forest classification to discriminate lithium deposits based on LIBS elemental images from spodumene, lepidolite concentrates and mineral powders. The one-vs-rest approach enables robust, scalable identification of 6 deposits, leveraging spatial co-occurrence patterns of elements as key features. Leave-One-Out Cross-Validation shows high classification accuracy across the six deposits with a few outliers. Overall, the approach proves effective and interpretable for tracing deposit origins, with potential extension to salt derived from cathodes and batteries to support supply chain traceability.

Funded by
the European Union

3.4 Exploration of a spatialized multivariate dataset: cathodoluminescence images characterization using k-means clustering

Authorship: Claire AUPART, Théophile LOHIER, Alban MORADELL-CASELLAS, Nathan BODEREAU, Daniel MONTFORT, Anne-Marie DESAULTY

3.4.1 Context and main goal

Optical cathodoluminescence is an imaging technique whose value for source tracing of naturally occurring minerals has been highlighted during the last decade (e.g. (Augustsson & Reker, 2012), (Baele, Decrée, & Rusk, 2019)). The technique consists in producing luminescence by exciting the electronic structure of matter using an electron beam. The intensity and colour of luminescence is directly controlled by the atoms present (material chemistry) and their relations in space (material structure). Depending on their history (geological context, formation process, possible transformations and alterations, etc.), rocks and minerals of similar types will develop more or less marked differences in compositions and atomic structure. These variably pronounced differences produce variably pronounced luminescence variations which can be used for mineral traceability. Notably, the colour and luminescence intensity of the studied phases will vary. Rocks and minerals of similar origin are more likely to have a similar history leading to similar luminescence, while samples of different origins will invariably have a different history thus, expectedly, different luminescence. Theoretically, the main limit to using these differences for discriminating origins is our capacity of detecting them.

The studied samples are Li-bearing minerals ores and concentrates. Initial data treatment using simple statistics has highlighted the potential of cathodoluminescence for traceability (cf. deliverable D2.3). It has been based on the comparison of the median luminescence of the medium lightness pixels of each sample, thus characterizing the median luminescence of the main Li-bearing mineral. However, observations of cathodoluminescence images have shown the occurrence of luminescent accessory minerals (Figure 29) that could be used to improve origin identification. Their presence, absence and proportions should vary from one origin to the other. The small pixel proportion represented by accessory minerals is however a challenge for their automatic identification and quantification.

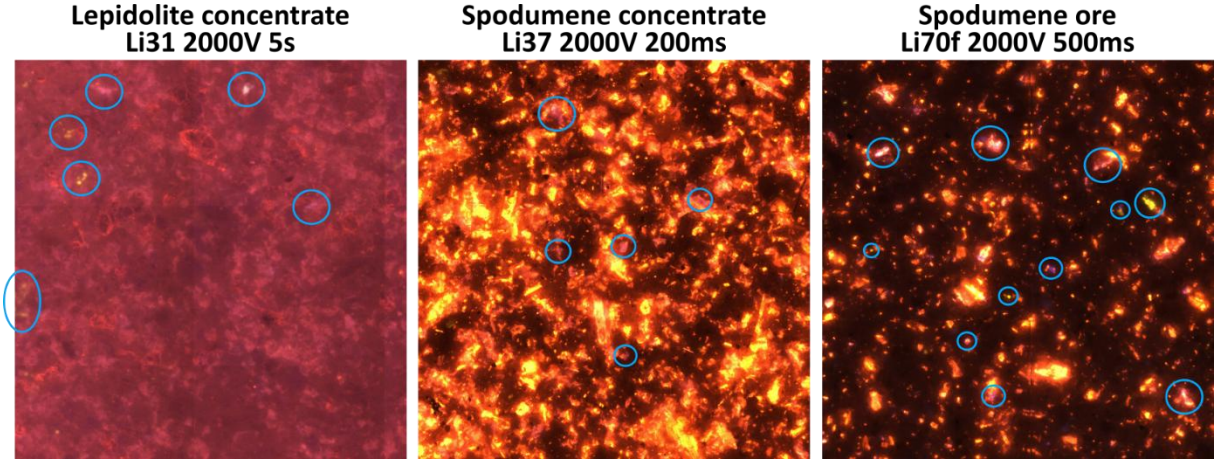

Figure 29. Luminescent accessory minerals in a few different samples (blue circles). Images are 3x3 mm².

Funded by the European Union

The aim of data treatment developed here is the automatic identification on cathodoluminescence images of accessory luminescent minerals groups in Li-minerals ores and concentrates samples.

3.4.2 Data Acquisition

A total of 95 samples has been imaged under the cathodoluminescence microscope, including 77 spodumene samples (mostly concentrates), and 17 Li-mica samples (lepidolite and zinnwaldite). Every sample has been imaged at least twice (both sides of a pellet, see deliverable D2.3 for more details).

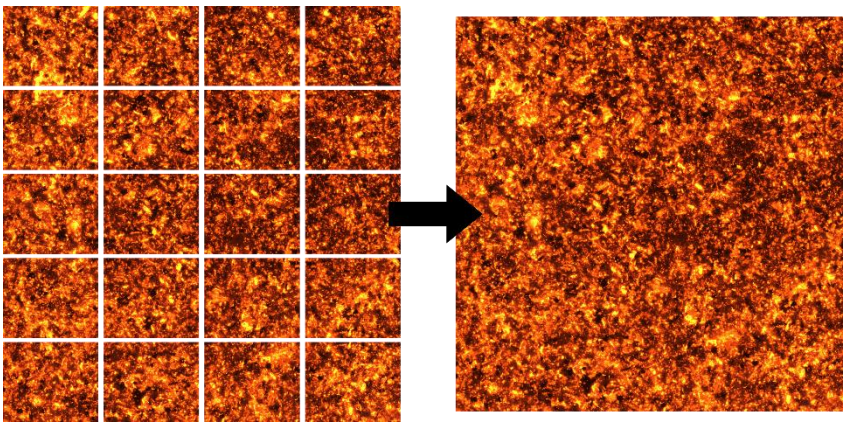

Figure 30. Example of the imaging of sample Li20 (spodumene concentrate) at 4500V with an exposure time of 100 ms. On the left, the raw images as saved by the cathodoluminescence equipment control software. On the right, the reconstructed 6x6 mm² mosaic image.

For each sample pellet, several mosaics of approximately 6 x 6 mm² (Figure 30) have been made at different exposure times (between 50 ms and 5 s) and different electron beam voltages (2000 or 4500 V). Both sides of each pellet have been analysed, between 4 and 7 mosaics were acquired per pellet face. Exposure time was adapted to get a range luminescence intensities allowing comparison between different samples and including each sample optimum luminescence for a given electron beam voltage (Figure 31).

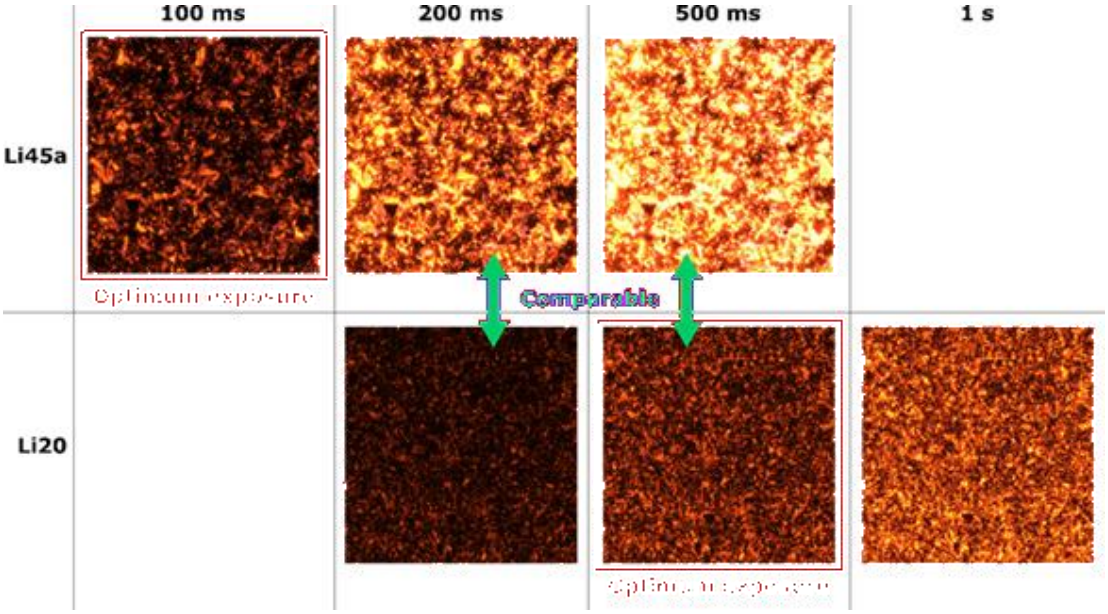

Figure 31. Cathodoluminescence images of two spodumene concentrate samples with different optimum exposure times. Images have been acquired at 2000V and with several exposure times to include optimum exposure images while keeping comparable images.

Optical cathodoluminescence images are retrieved under the form of RGB (red, green, blue proportions) images that are converted to a HSL (hue, saturation and lightness) system, more adapted to luminescence description (Figure 32).

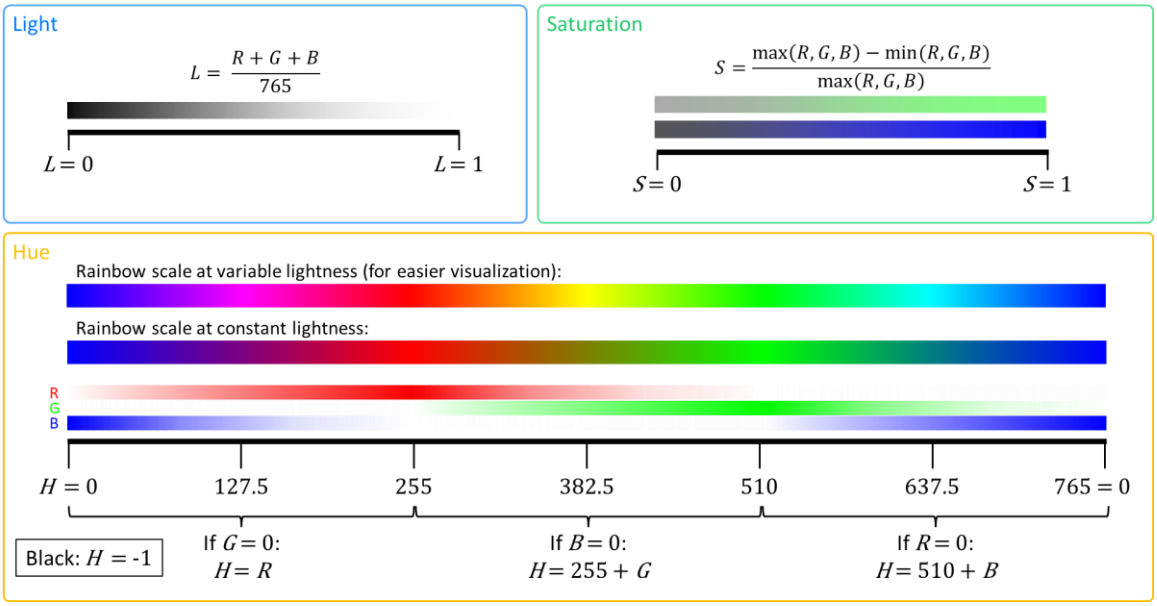

Figure 32. Light, saturation and hue system used for image treatment.

3.4.3 Strategy

Done on optimum luminescence intensity images.

The first step is to filter out all the darkest pixels. Then K-means clustering is used to group pixels with similar properties (hue, saturation and brightness). The approach to identify

Funded by the European Union

pixels with luminescent accessory minerals is to make the method create numerous groups so that even poorly represented groups are differentiated.

3.4.4 Preliminary results

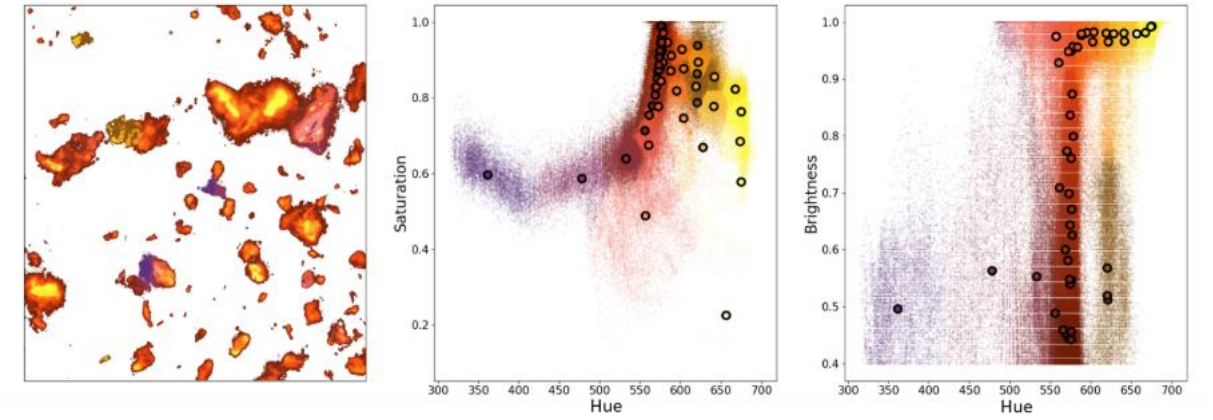

Figure 33. K-mean clustering approach results on a 1 mm² zone of sample Li70f (spodumene ore). Each colour corresponds to a cluster whose centroid is indicated by a black circle.

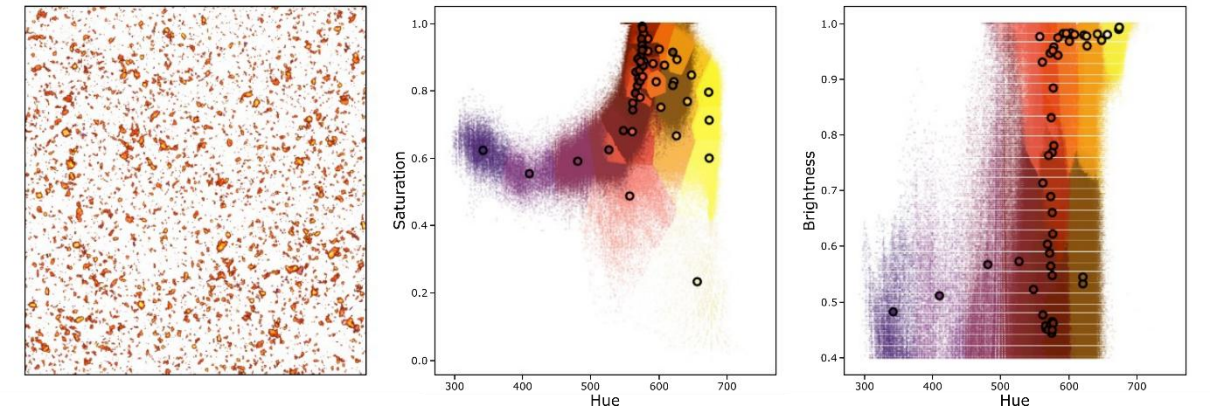

Figure 34. K-mean clustering approach results on full imaged area (6x6 mm²) of zone of sample Li70f (spodumene ore).

Outlook

The explored processing method satisfyingly identifies accessory mineral sub-groups in cathodoluminescence images of individual samples. The next steps are:

- to define parameters linked to the mineral groups presence and relative abundance that can be used to compare samples between each other
- repeat the procedure for all MaDiTraCe samples analysed with cathodoluminescence
- Check the contribution of accessory minerals to help define Li-mineral ores and concentrates origin groups using cathodoluminescence
- Combine parameters obtained from Li-bearing minerals main luminescence with parameters obtained for accessory minerals to refine origin-specific Li-mineral ores and concentrates cathodoluminescence signatures.

4 Natural graphite (C) - workflow and results

4.1 Context and main goal

Currently, natural graphite makes up ca. 50% in the global graphite production, while the global share of natural graphite in battery production is over 80% (Zhang, Liang, & Dunn, 2023). Due to growing energy storage demand and transitions in the energy sector, global graphite demand is expected to increase by a factor of two by 2050 (IEA, 2025). While synthetic graphite has notable advantages over natural graphite in terms of purity, natural graphite is expected to gain significance on the market due to its smaller environmental footprint, lower cost and more favourable processability (Zhao, et al., 2022). While three larger groups of natural graphite can be distinguished, namely semi-graphite, hydrothermal graphite (vein graphite), and flake graphite, flake graphite is best suited for battery production, and it is dominant on the natural graphite market. Therefore, the main focus in the MaDiTrace project is also on flake graphite. High-grade flake graphite ore usually contains up to 40% graphite mixed with various kinds of other minerals, which are characteristic for each graphite deposit (Dallos, et al., under review). While graphite content is increased above 99.95% (battery grade) during processing, the product still retains a certain amount of other minerals and/or industrial products. Given that the studied materials are essentially mineral mixtures, they were studied with a special focus on graphite flakes, on the non-graphitic constituents and the mixture of both.

4.2 Data acquisition

Our sample set comprises 150 samples from 14 countries. In this deliverable, we specifically focus on the 62 concentrates, which are produced after grinding and flotation of raw ore.

Deliverable 2.4 extensively discusses the series of laboratory methods, which were applied for graphite traceability. In summary, graphite flakes were directly analysed by carbon stable isotopy and Raman spectroscopy, non-graphitic constituents were studied by SEM-EDX (and EMPA), while the mixture of both (i.e., graphite flakes mixed with non-graphitic constituents) was analysed by solution ICP-MS, XRF, LA-ICP-MS and LIBS. The data analysis characteristics of these methods are summarized in Table 3.

Table 3. Data analysis characteristics of the methods applied for natural graphite traceability. Large between-group/within group variance means high classification potential between deposits.

Method	# of variables	# analyses/sample	Classification potential	Potential for machine learning
Carbon stable isotopy	1	1	low	low
Raman spectroscopy	6	10-20	low	low
ICP-MS	44	3	high	low
XRF	20	4	high-medium	low
LA-ICP-MS	54	15	high	high
LIBS	32	100	high-medium	high
SEM-EDX	41	500-5000	high	high

Funded by the European Union

Carbon stable isotopy yields only one numerical variable ($\delta^{13}C$), which shows significant overlaps between most flake graphite concentrates, due to the Paleoproterozoic age (and the globally dominant simple biota at that time) of the largest graphite deposits worldwide. Similarly, Raman spectroscopy results in only a few graphite-structure-related numerical output parameters, which are nearly identical in most samples (low classification potential) as most globally traded graphite products comprise well crystallized graphite flakes. Therefore, data analysis aspects are not discussed for these methods.

Solution ICP-MS and XRF are bulk methods, which result in multivariate datasets. Despite their high classification potential, the downside of these methods in terms of data analysis is that only a few (3-4) analyses were done per sample due to practical limitations. The resulting data table therefore comprises only a few entries per sample with many variables, rendering the use of these datasets impractical for machine-learning-based classification methods. As presented in Deliverable 2.4., linear discriminant analysis shows that differentiation between deposits is possible with both methods.

For machine-learning-based classification, the underlying dataset ideally fulfils the following criteria: a large number of variables, many repeated analyses of the same sample and high classification potential. Therefore, in the following sections we discuss the three methods in detail, which fulfil these criteria, namely LIBS, LA-ICP-MS and SEM-EDX (and EPMA).

4.3 Strategy

4.3.1 LIBS

10000 individual shots were made in a rectangular field (i.e., 100 lines, 100 shots per line) on each of the graphite concentrate pellets studied. The resulting spectra were averaged over each line, resulting in 100 averaged spectra per sample. Several approaches were tested on the resulting dataset, which are summarized in Figure 35. Outliers exceeding 3SD in each line were rejected before averaging. In order to correct for optical and surface effects in the session, raw intensities were normalized by the total intensity in each individual spectrum. The generalization potential of the model was tested by a modified LOO cross-validation approach. Further details can be found in (Arató, et al., 2025).

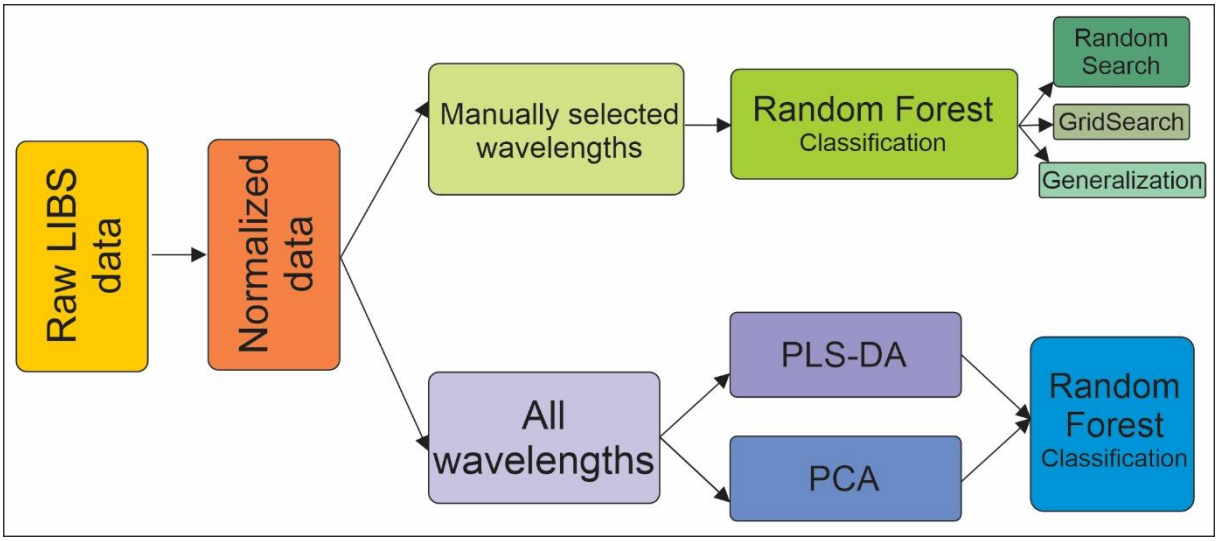

Figure 35. Data analysis approach applied to the LIBS dataset acquired on natural graphite concentrates (Arató, et al., 2025). PLS-DA=partial least squares discriminant analysis, PCA=principal component analysis.

4.3.2LA-ICP-MS

15 lines were ablated in each sample (pressed pellets), each line comprising several 1000 shots. Data were averaged per line, resulting in 15 entries (concentration in ppm) per sample with 54 numerical variables (measured isotopes). Values below detection limit resulted in missing trace element values. These were imputed by linear regression based on the linear correlation between major and trace elements. For the calculation of correlation coefficients, centred log-ratio transformation (Aitchison, 1982) was applied, due to the known 100% constraint of compositional data. Similarly to LIBS, the dataset was subjected to RF classification with LOO cross validation. In addition to LA-ICP-MS data, $\delta^{13}C$ values were also included in the dataset (the same $\delta^{13}C$ value for all 15 LA-ICP-MS in a sample).

4.3.3 SEM-EDX

As described in Deliverable 2.4., non-graphitic constituents were separated from physically purified graphite concentrates and their chemically purified counterparts by density separation. The resulting mineral concentrates were embedded in epoxy, and the resulting mineral concentrates (each mineral) were analysed by EDX, resulting in hundreds to thousands of EDX spectra per sample. Each spectrum is assigned a mineral label. Mineral labels (categorical variable) were label-encoded, and all analyses were merged into a single data frame. Missing values were replaced by 0 (other approaches are being tested). While the data frame comprises 41 numerical variables usually 3 to 8 elements were measured per mineral. Train-test splitting was performed in an 80:20 ratio, whereas LOO cross validation remains to be tested on this dataset.

4.4 Results

4.4.1 LIBS

Each spectrum comprised over 8000 datapoints, resulting in a considerable data size. Dimension reduction was tested by partial least squares discriminant analysis (PLS-DA) and principal component analysis (PCA), where the prior approach significantly outperformed the latter. On the other hand, manual selection of the 32 most intense spectral lines yielded even higher classification scores and was therefore the preferred approach. KNN, SV and RF classifiers as well as a stacked classifier combining all three classifiers was tested. The latter three approaches yielded similar classification accuracy, while the simplicity and robustness over a wide range of hyperparameters made the random forest classifier our primary choice. With the “standard” 80:20 train-test splitting approach, 93 % classification accuracy was achieved. However, the generalization test of the model was achieved by LOO cross validation on deposits where at least four samples from different years were available (Figure 36). This resulted in a decrease to 65% classification accuracy. This highlights the limitations of classification based on highly heterogeneous sample material. Our year-to-year samples from Madagascar probably stem from different mines or different parts of the same mine or might even be incorrectly labelled by the producer, thereby resulting in a highly different chemical signature in each sample and incorrect classification.

Funded by the European Union

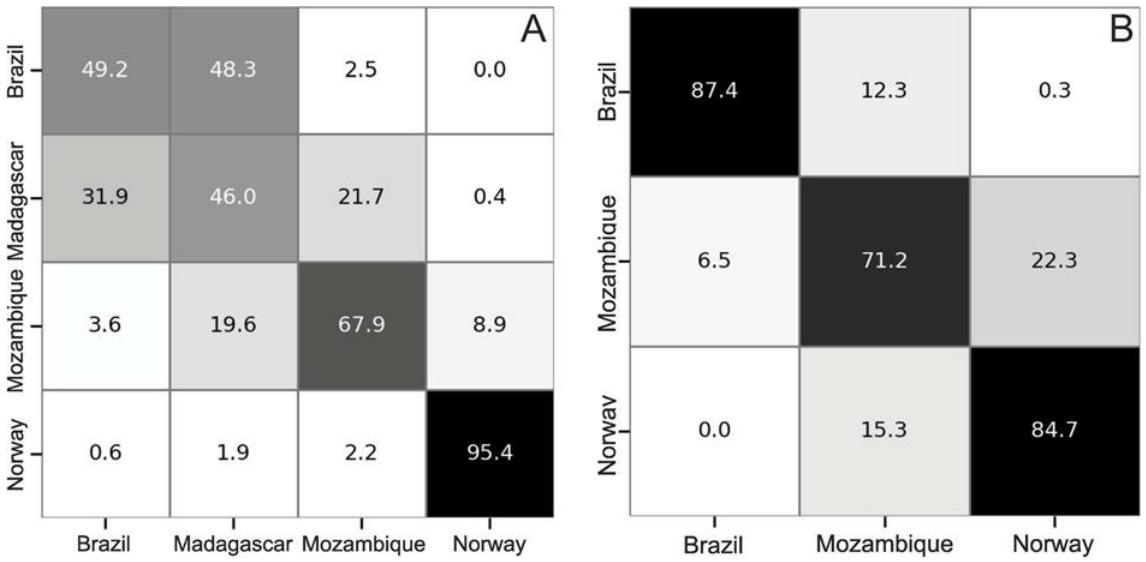

Figure 36. Confusion matrix of graphite deposits where at least four samples from different years were available, obtained by LOO cross-validation.

4.4.2 LA-ICP-MS

RF classification based on LA-ICP-MS significantly outperforms that of the LIBS-based dataset owing to the larger number of variables and the resulting higher classification accuracy. Looking at Figure 37, the number of correctly classified spectra is seen in the diagonal in percent (Bal=Balama, Mozambique; Kai=Kaisersberg, Austria; Ped=Pedra Azul, Brazil; Ska=Skaland, Norway; Vat=Vatomina, Madagascar). Notably, the classification of samples from Madagascar fails in a similar manner. Excluding $\delta^{13}C$ from the dataset results in a noticeable decrease (~3%) in classification accuracy.

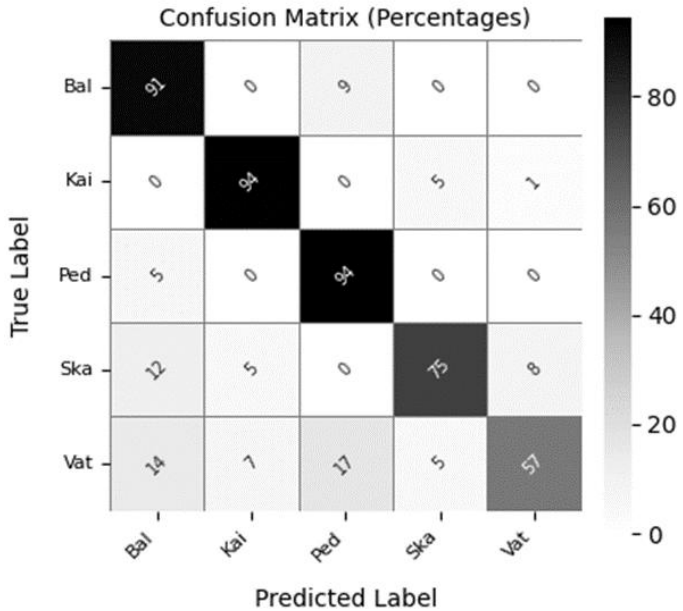

Figure 37. LA-ICP-MS-based confusion matrix of graphite deposits represented by at least four samples, obtained by LOO cross validation.

4.4.3 SEM-EDX

The preliminary results on this dataset show promising classification accuracies on concentrates and chemically purified concentrates alike (Figure 38). Notably, LOO cross validation will be tested, and the dataset will be extended by data acquired on signal crystals by other methods, such as EPMA and LA-ICP-MS.

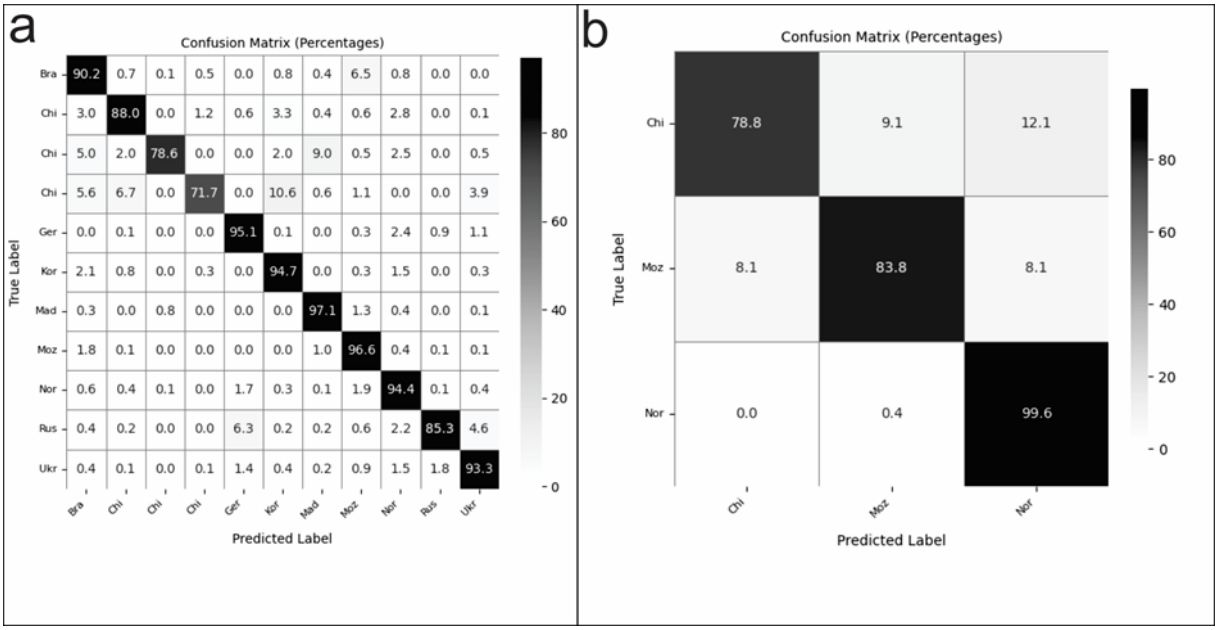

Figure 38. Results of random forest classification based on the SEM+EDX dataset obtained on mineral concentrates separated from graphite products. a) concentrates b) chemically purified samples.

4.5 Conclusion

Methods, which result in datasets with many variables and many repeated measurements from the same sample as well as the resulting variables show significant differences between deposits, are best suited for multivariate classification methods. For natural graphite, random forest classification was applied for the three methods fulfilling these criteria. LA-ICP-MS excels at low detection limits for many trace elements and therefore yields a higher classification accuracy than LIBS. On the other hand, the analysis of single crystals separated from graphite concentrates carries a huge potential even for chemically purified products. Currently, results from SEM-EDX were considered, however, in the future the analysis of single crystals can potentially combine the advantages of all applied methods.

Funded by the European Union

5 Cobalt (Co) - workflow and results

Authorship : Yuan Shang and Quentin Dehaine (GTK)

5.1 Context and main goal

Cobalt (Co) is globally mined as a by-product of nickel (Ni) and copper (Cu), and with some small volumes as primary cobalt or as a by-product of platinum-group metals (PGM) production. Sediment hosted Cu-Co, Ni-Co laterites and magmatic Ni-Cu (-Co) sulphide ores are the three primary deposits that Co has been produced. Due to the challenges of obtaining sediment hosted Cu-Co ore samples from Democratic Republic of the Congo (DRC) and Ni-Co laterite samples from the major operations in southeast Asia countries, in the MaDiTraCe project, we have mainly focused on cobalt sourced from magmatic Ni- Cu (-Co) sulphide ores.

Magmatic sulphide ores form when an immiscible sulphide liquid segregates from a silicate magma and scavenges chalcophile and siderophile elements (e.g., Ni, Cu, Co, Au, PGE, Se, etc.) very efficiently. Because trace elements partition strongly and systematically between silicate melt and later sulphide minerals, their concentrations and ratios preserve information about the source of the magma, ore forming processes (e.g., R-factor, sulphide fractionation, crustal contamination), the degree of partial melting and post-magmatic modification (hydrothermal overprint) (Barnes, Holwell, & Le Vaillant, 2017). Therefore, trace elements composition and their ratios are essentially the "DNA" of the deposit.

Sulphur isotopes, expressed as $\delta^{34}S$ values, are a valuable fingerprinting tool for magmatic sulphide ores because high temperature magmatic processes produce minimal isotope fractionation, preserving the sulphur source signature. Most mantle derived magmatic sulphide deposits display uniform $\delta^{34}S$ values close to 0‰ (approximately −2 to +2‰), characteristic of uncontaminated komatiite hosted and many intrusion related Ni–Cu–PGE systems. Wider $\delta^{34}S$ ranges and systematic shifts reflect assimilation of crustal sulphur, commonly required to trigger sulphide saturation in economically significant deposits. At magmatic temperatures, consistent $\delta^{34}S$ values among pyrrhotite, pentlandite, and chalcopyrite further support a magmatic origin. When integrated with trace element systematics, sulphur isotopes effectively distinguish primitive mantle melts from crustally contaminated magmatic sulphide ore systems ((Barnes, Holwell, & Le Vaillant, 2017), (Naldrett, 2004)).

In MaDiTraCe project, for the purpose of traceability – to trace Co in battery product back to its source ore – we first need to establish the geochemical signature of ores in specific locations, and we can then distinguish these ores spatially, based on their geochemical signature. To this end, we collected representative magmatic sulphide ore samples on a global scale, including some of the most economically significant ores, such as Norilsk and Pechenga in Russia, Leinster in Australia, Reglan and Sudbury in Canada, Jinchuan in China and Kevitsa in northern Finland (Figure 39). We obtained the trace element composition by LA-ICP-MS (Laser ablation inductively coupled plasma mass spectrometry) analysis and S isotopic composition by LA-MC-ICP-MS (Laser ablation multi-collector ICP-MS) analysis at Espoo Research Laboratory of GTK in Finland. Detailed description for the analytical methods is provided in Deliverable 2.4. We aim here to apply multivariate data analysis to

differentiate magmatic Ni-Co sulphide ores globally based on a comprehensive dataset of trace elements composition (and ratios) and S isotopes ($\delta^{34}S$).

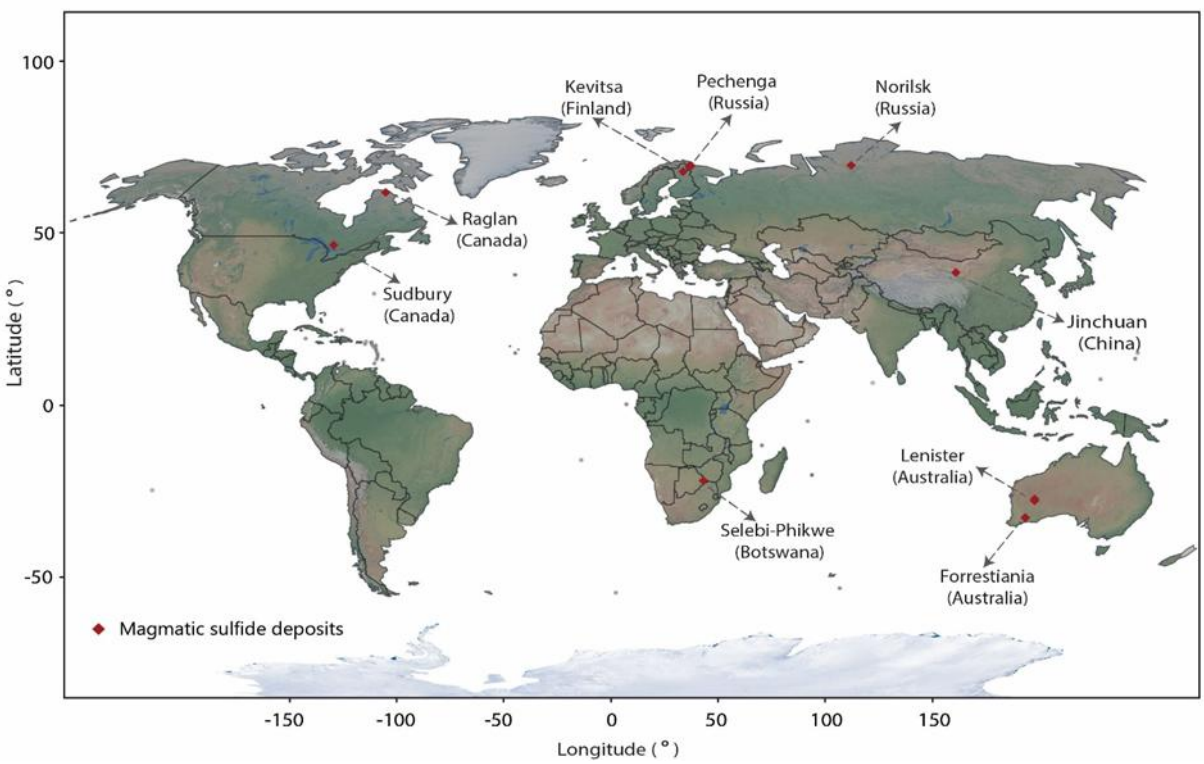

Figure 39. Location of magmatic sulphide ore samples collected.

5.2 Data acquisition

For trace element composition, we have two types of datasets:

1) bulk elemental composition analysed on samples in the form of pressed pellets (24 ore samples and 2 mineral concentrates samples);

2) elemental composition in major sulphide minerals of the ores, including chalcopyrite, pentlandite, pyrrhotite and pyrite (26 ore and 4 mineral concentrate samples).

Both datasets were obtained with in-situ analysis by LA-ICP-MS. The measured elements include: Si, Ti, V, Cr, Mn, Fe, Co, Ni, Cu, Zn, Ga, Zn, Ga, Ge, As, Se, Nb, Mo, Ru, Rh, Pd, Ag, Cd, In, Sn, Sb, Te, Ba, W, Re, Os, Ir, Pt, Au, Hg, Tl, Pb, Bi and U. The raw data reduction was done with the software Iolite (Paton, Hellstrom, Paul, Woodhead, & Hergt, 2011). After data reduction and removal of the elements below detection limits, less elements are available for further data analysis.

The trained data for bulk elements composition include 590 points, with input elements V, Cr, Co, Ni, Zn, As, Se, Ba, Pb and Bi and the element ratio (Ni/Co). For the trace element dataset in sulphide minerals, we obtained 1991 datapoints in total. However, since the elemental compositions in each sulphide mineral are different, we need to treat and analyse each sulphide mineral separately. In addition, because pyrite only appeared in a few of ore samples, so we excluded data from pyrite for further data analysis. In the end, 420 points

Funded by
the European Union

are available for chalcopyrite with Si, Mn, Co, Ni, Cu, Zn, Ga, Se, Rh, Pd, Ag, Cd, In, Pb and Bi as input elements for model training. 664 points are available in pyrrhotite with Si, Mn, Co, Ni, Ge, Se, Mo, Ag, Pb and Bi as input elements for model training. 667 points are available in pentlandite with Si, Mn, Co, Ni, Ge, Se, Ru, Ag, Pb and Bi for model training.

To achieve a comprehensive elemental composition in the minerals, we also measured the major (e.g. Fe and S) and some minor and trace elements (e.g., Ti, Cr, Mn, Cu, Co, Ni, As, Zn etc.) composition in sulphide minerals with EPMA (Electron Probe Micro-analysis). In total, 628 datapoints were obtained from EPMA for the sulphide minerals.

For S isotopes, we analysed 16 ore samples and 4 mineral concentrate samples with 686 datapoints obtained after data reduction.

5.3 Strategy

We used the Linear Discriminant Analysis (LDA) to investigate the differences in trace elements and S isotopic compositions in different ore and mineral concentrate samples. The method attempts to classify observations described by values on continuous variables into groups. The group membership is defined by a categorical variable X, which is predicted by the continuous variables. These variables are called covariates and are denoted by Y. For the linear fitting, it assumes that the within-group covariance matrices are equal, and the covariate means for the groups defined by X are assumed to differ. The method estimates the distance from each observation to each group's multivariate mean using the Mahalanobis distance. The observations are classified into the closest group (https://www.jmp.com/support/help/en/19.0/#page/jmp/overview-of-the-discriminant-platform.shtml#). Please see also **section 2.1.3.1-Supervised models** for the general description of the LDA.

In our case, the input continuous variables are elemental compositions (e.g., Ni, Co, Se, Ag, Pd, Pb, Bi and element ratios such as Ni/Co and Pb/Bi) from samples of different origins, i.e., the Y covariates, are used to classify the samples into categories. X is the origin of the samples, which defines the groups of the variables to be classified. The output is a 2-dimensional biplot called" Canonical Plot". The biplot axes are the first two canonical variables, which provide maximum separation among the groups (sample origin). The observations and the multivariate means of each group are represented as points in the biplot. A 95% confidence level ellipse is plotted for each mean. If the two groups differ significantly, the confidence ellipses tend not to intersect. An ellipse denoting a 50% contour is also plotted (Figure 40). The linear fitting discriminant analyses (LDA) were implemented by using the statistical software JMP 19 and detailed method description is provided in the platform webpage (https://www.jmp.com/en/software/new-release/new-in-jmp).

5.4 Results

5.4.1 Discrimination analysis for bulk element composition in the sulphide ores

We first applied the LDA to bulk elemental composition for the studied samples (Figure 40). The modelled result was below satisfaction level, with 85 data points misclassified

(misclassified percentage of 14%). As seen from the discriminant map (Figure 40), except the Western Australia samples (Leinster and Forrestania) and Selebi-Phikwe sample are relatively well separated, the other samples are mixed in the map.

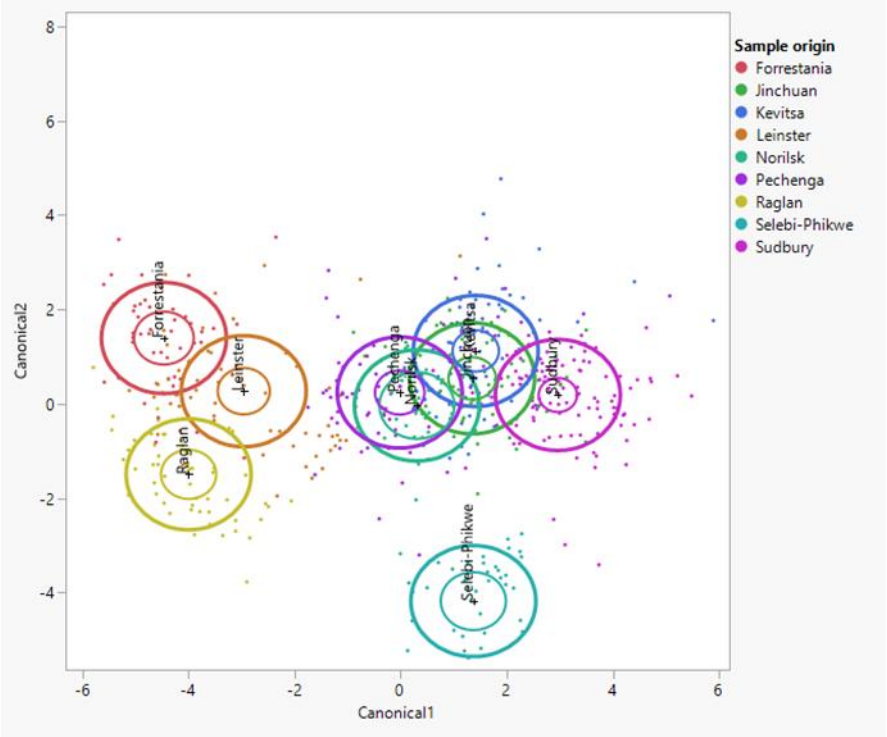

Figure 40. Discrimination map obtained from LDA on bulk trace element composition from magmatic sulphide ores. Thick lines indicates the median values along the canonical variables and the thinner one indicates the 95% confidence interval.

5.4.2 Discrimination analysis for element composition in Pentlandite

Since pentlandite (Pn) is the most valuable mineral in magmatic sulphide ore for Co and Ni, here we show the result obtained from LDA based on the trace element composition in Pn (Figure 41). We included both the trace element compositions (e.g., Si, Mn, Co, Ni, Ge, Se, Ru, Ag, Pb and Bi) and elemental ratio (Ni/Co and Pb/Bi) obtained from LA-ICP-MS and major and minor element compositions obtained from EPMA (e.g., S, Fe, Ni and Co). The classification model from the LDA analysis is much improved with only 4 misclassified datapoints out of 290 counts (Figure 41). It shows that most of the samples can be well differentiated although some overlap still existed (e.g., Pechenga and Selebi-Phikwe). Even for the two samples from Western Australia, i.e.,Leinster and Forrestania that are both associated with komatiite type of magma, can be well separated.

Funded by the European Union

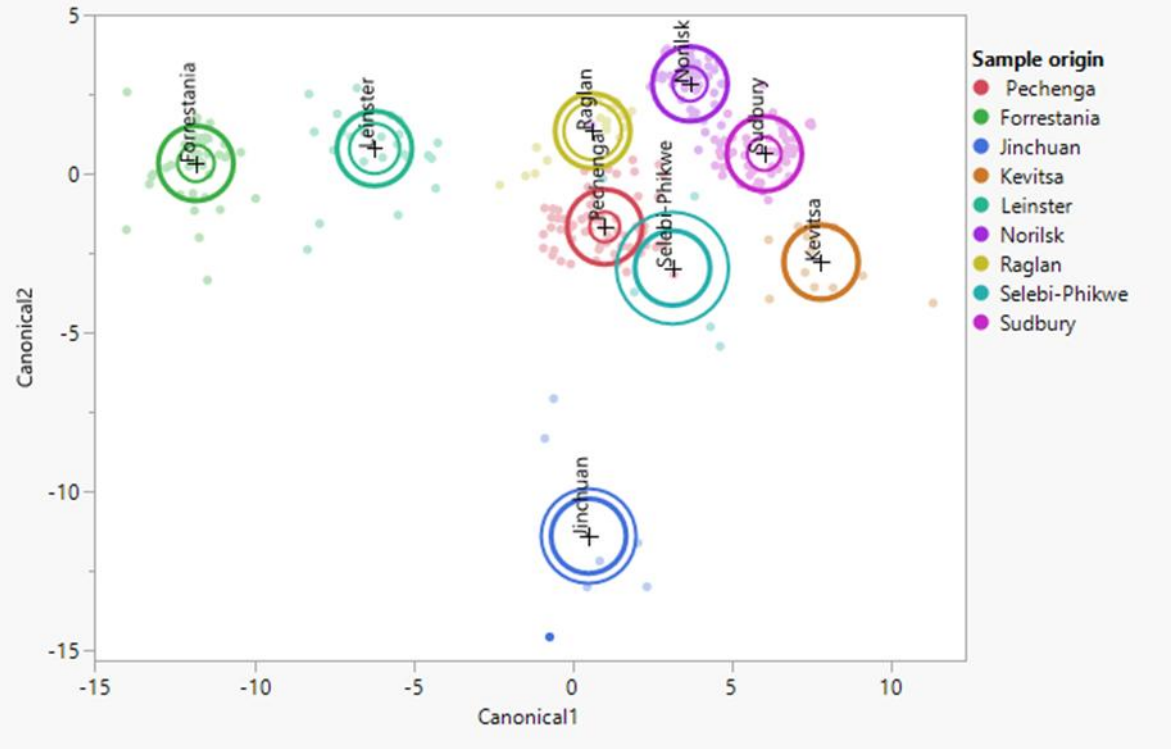

Figure 41. Discrimination map obtained from LDA on major and trace element composition in pentlandite from magmatic sulphide ores. Thick lines indicates the median values along the canonical variables and the thinner one indicates the 95% confidence interval.

5.4.3 Discrimination analysis by combing the trace elements and S isotopic signatures together

Our last effort was to combine the trace elements and S isotopic signatures in pentlandite to investigate whether this will better discriminate samples from different origins. The S isotopic signature ($\delta^{34}S$) in pentlandite of different ore samples is shown in Figure 42. It is obvious that Norilsk samples show significantly heavier $\delta^{34}S$ values in Pn, with a median $\delta^{34}S$ at 11.6 ‰. Ore samples from Leinster display the lowest $\delta^{34}S$ values, with a median $\delta^{34}S$ value at -2.2 ‰ in Pn. $\delta^{34}S$ for other samples show large overlaps and fall within the range -3 to 6 ‰ (Figure 42).

Funded by
the European Union

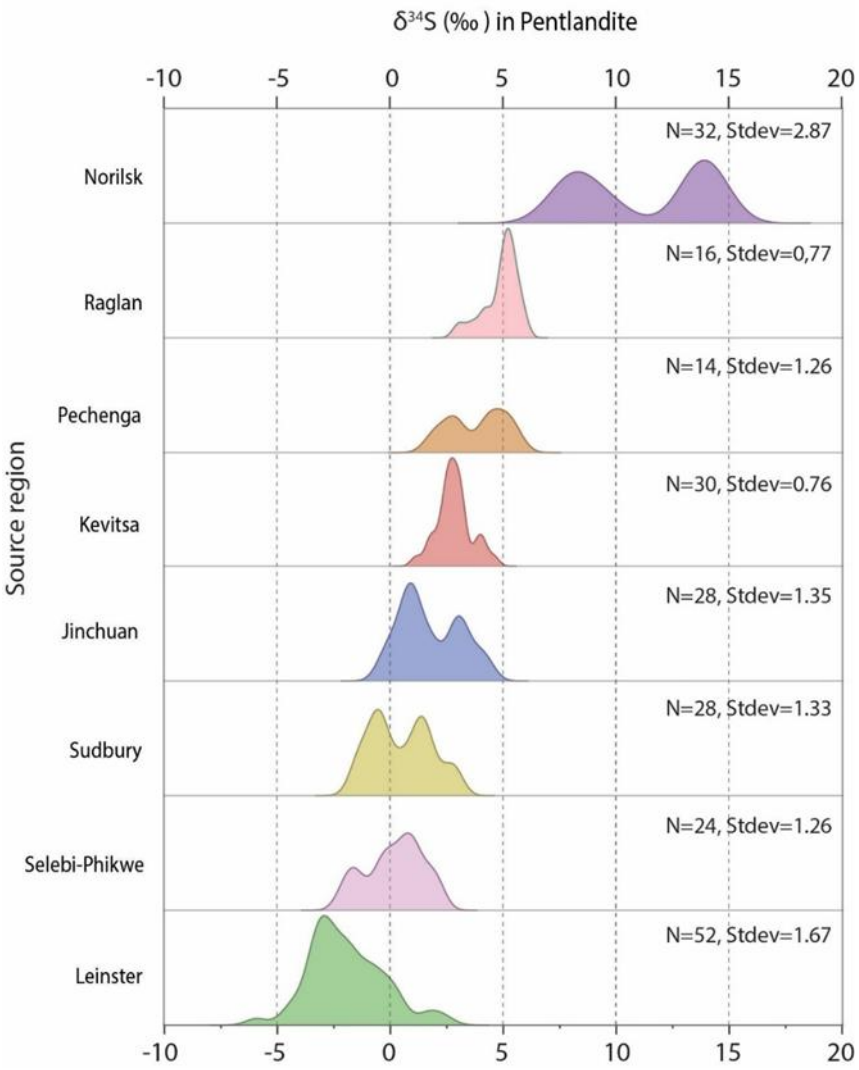

Figure 42. Kernel Density Estimation (KDE) for the distribution of δ34S (‰) in the pentlandite of magmatic Ni-Cu (-Co-PGE) deposits.

The same multivariate data analysis method (LDA) was applied to the combined dataset of elemental composition and S isotopic data, and a better discrimination result was obtained with the addition of sulphur isotope values to the dataset. While the number of the trained data points decreased to 115, the model shows zero misclassifications. Although in the map there are still overlaps between ore samples of Jinchuan, Sudbury and Selebi-Phikwe due to the reason that only the first two linear discriminants are considered for practical reasons (plotting). Leinster, Norilsk, Raglan, Pechenga and Kevitsa can be well separated even in two dimensions (Figure 43).

Funded by the European Union

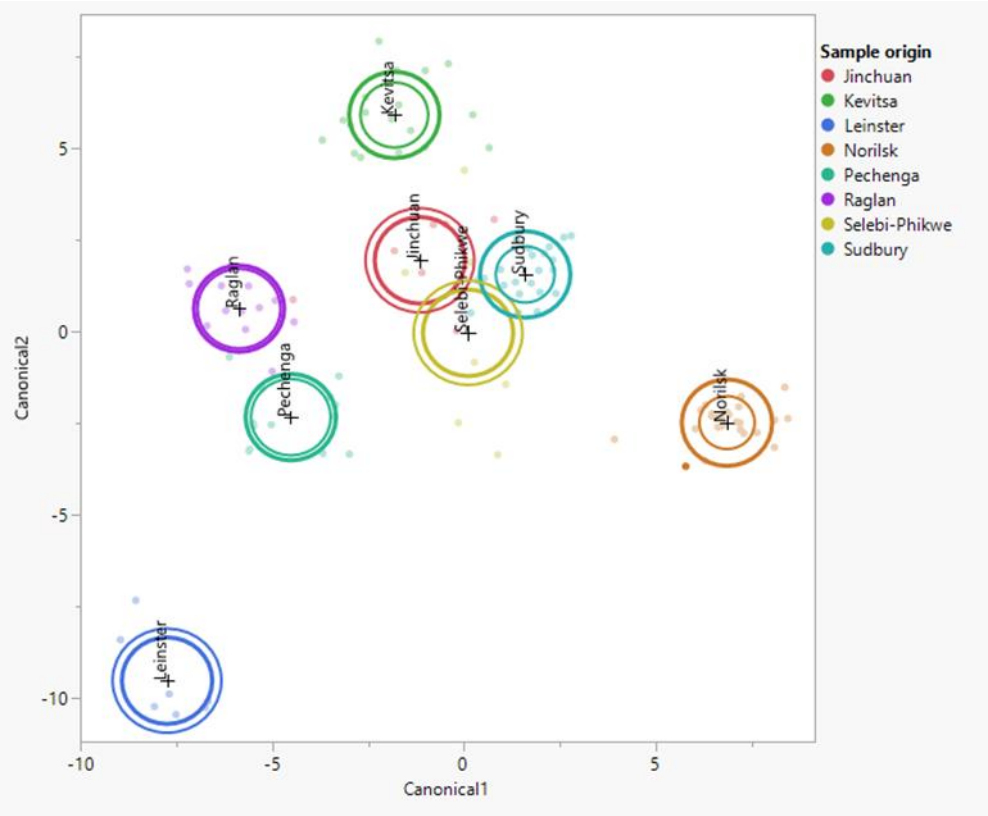

Figure 43. Discrimination map obtained from LDA on element composition and S isotopic signature ($\delta^{34}S$) of pentlandite from magmatic sulphide ores (Shang, et al., in prep.).

6 Conclusion

The work carried out in task 2.5 has highlighted the key steps for building workflows for the traceability of critical raw materials based on data from different types of geochemical analyses. From the initial extraction of the relevant origin-specific features to the final quantification of uncertainty, including model comparison, this deliverable provides several levers to improve CRM traceability by combining geochemical analyses and machine learning.

Feature construction from multivariate data is a key step in this process, whether it involves high-resolution data, such as those generated by LIBS and cathodoluminescence, or bulk data, such as those generated by LA-ICP-MS or XRF. For the latter, it has been shown that linear combinations of the different variables, constructed using LDA, allow for proper separation of the different deposits. Furthermore, a new method based on LDA has been developed to account for poorly characterized deposits, that is, those for which few samples are available. In the case of Cobalt, where all deposits are sufficiently well characterized, it has been demonstrated that nonlinear approaches such as RF can achieve performance at least equivalent to LDA.

For techniques generating large volumes of data, several dimensionality reduction methods have been tested. Standard methods such as PCA and PLS-DA yield mixed results, and in the case of graphite, a manual method was preferred. A clustering approach was tested on cathodoluminescence analyses of lithium samples, producing promising results. However, in this case, the small number of available samples did not allow for the implementation of supervised classification approaches. Finally, a method based on the presence of chemical element assemblages was developed to take advantage of the spatialized data obtained using LIBS.

Since the data from Li isotopic analyses are univariate, the feature extraction phase is not relevant. In this case study, the focus was therefore on the discriminative power of different ML approaches and their ability to capture the uncertainty surrounding the estimated origin. The KNN method notably allows discrimination between salars and hard rocks, as well as between salars themselves, with an accuracy of around 90%. Furthermore, this approach enables distinguishing samples from two Chinese salars and hard rock deposits outside China. The main limitation of this approach is that, in its original form, it does not allow the integration of underrepresented deposits.

The work carried out shows that, regardless of the analysis technique and data processing methods used, determining the origin of a sample is subject to uncertainty. This results from the limited availability of samples to fully characterize the intra-deposit variability. A major source of potential attribution errors are deposits lacking in the reference data base. This has been addressed by including non-referenced deposits in the data analyses and investigate the possibility to identify them as such. The importance of qualifying this uncertainty within the framework of CRM traceability was highlighted during the workshop, and part of the work carried out under task 2.5 focuses on this issue. This deliverable notably demonstrates how the KNN, RF, and bootstrapped LDA methods can capture uncertainty indispensable for decision-making during certification.

Funded by
the European Union

7 Bibliography

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological), 44(2)*, 139-160.

Arató, R., Quarles, D., Obbágy, G., Dallos, Z., Arató, M., Gopon, P., & Melcher, F. (2025). Towards a chemical fingerprint of graphite by laser-induced breakdown spectroscopy. *Journal of Analytical Atomic Spectrometry, 40(9)*, 2526-2537.

Augustsson, C., & Reker, A. (2012). Cathodolumenescence spectra of quartz as provenance indicators revisited. *Journal of Sedimentary Research, 82(8)*, 559-570.

Baele, J.-M., Decrée, S., & Rusk, B. (2019). Cathodoluminescence applied to ore geology and exploration. *Ore deposits: Origin, exploration, and exploitation*, 131-161.

Barnes, S. J., Holwell, D. A., & Le Vaillant, M. (2017). Magmatic sulfide ore deposits. *Elements, 13(2)*, 89-95.

Bodereau, N., Delaval, A., Lepage, H., Eyrolle, F., Raimbault, P., & Copard, Y. (2022). Hydrological classification by clustering approach of time-integrated samples at the outlet of the Rhône River: application to Δ14C-POC. *Water research, 220*, 118652.

Breiman, L. (1996). Bagging predictors. *Machine learning, 24(2)*, 123-140.

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

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory, 13(1)*, 21-27.

Dallos, Z., Arató, R., Obbágy, G., Burhanuddin, B., Burtscher, M., & Melcher, F. (under review). Analytical characterization of graphitic raw ores and natural graphite concentrates; linking residual mineral phases to source provenance. *Available at SSRN 6065626*.

Delaval, A., Duffa, C., Pairaud, I., & Radakovitch, O. (2021). A fuzzy classification of the hydrodynamic forcings of the Rhone River plume: An application in case of accidental release of radionuclides. *Environmental Modelling & Software, 140*, 105005.

Desaulty, A.-M., Monfort Climent, D., Lefebvre, G., Cristiano-Tassi, A., Peralta, D., Perret, S., . . . Guerrot, C. (2022). Tracing the origin of lithium in Li-ion batteries using lithium isotopes. *Nature communications, 13(1)*, 4172.

Green, A. A., Berman, M., Switzer, P., & Craig, M. D. (1988). A transformation for ordering multispectral data in terms of image quality with implications for noise removal. *IEEE Transactions on geoscience and remote sensing, 26(1)*, 65-74.

IEA. (2025). *World Energy Outlook 2025*. Paris: IEA.

Kamel, M. S., & Selim, S. Z. (1994). A relaxation approach to the fuzzy clustering problem. *Fuzzy Sets and Systems, 61(2)*, 177-188.

Le Teurnier, B., Li, X., Boffety, M., Hu, H., & Goudail, F. (2020). When is retardance autocalibration of microgrid-based full Stokes imagers possible and useful? *Optics Letters, 45(13)*, 3474-3477.

Lundberg, S. M., & Lee, S.-I. (2017). Consistent feature attribution for tree ensembles. *arXiv preprint arXiv:1706.06060*.

Morgan, J. N., & Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American statistical association, 58(302)*, 415-434.

Naldrett, A. J. (2004). *Magmatic Sulfide Deposits*. Heidelberg: Springer Berlin.

Paton, C., Hellstrom, J., Paul, B., Woodhead, J., & Hergt, J. (2011). Iolite: Freeware for the visualisation and processing of mass spectrometric data. *Journal of Analytical Atomic Spectrometry, 26(12)*, 2508-2518.

Ponce-Bobadilla, A. V., Schmitt, V., Maier, C. S., Mensing, S., & Stodtmann, S. (2024). Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development. *Clinical and translational science, 17(11)*, e70056.

Rakotonirina, H., Guridi, I., Honeine, P., Atteia, O., & Van Exem, A. (2024). Spatial interpolation and conditional map generation using deep image prior for environmental applications. *Mathematical Geosciences, 56(5)*, 949-974.

Rifkin, R., & Klautau, A. (2004). In defense of one-vs-all classification. *Journal of machine learning research, 5*, 101-141.

Shang, Y., Dehaine, Q., Liu, X., Myllyperkiö, M., Bertelli, M., Kinnunen, P., . . . and Lahaye, Y. (in prep.). Sulphur isotope forensics: Bridging the traceability gap from magmatic sulphide ore to battery product.

Simonnet, T., Grangeon, S., Claret, F., Maubec, N., Fall, M. D., Harba, R., & Galerne, B. (2024). Phase quantification using deep neural network processing of XRD patterns. *IUCrJ, 11(5)*.

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics, 8(1)*, 25.

Vu, T. H., Dao, T. B., Nguyen, V., Vrain, C., & Breuillard, H. (2025). Features Leverage in Graph Models for Mineral Prospectivity Mapping. *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing*, (pp. 597-604).

Zhang, J., Liang, C., & Dunn, J. B. (2023). Graphite flows in the US: insights into a key ingredient of energy transition. *Environmental Science & Technology, 57(8)*, 3402-3414.

Zhao, L., Ding, B., Qin, X.-Y., Wang, Z., Lv, W., He, Y.-B., . . . Kang, F. (2022). Revisiting the roles of natural graphite in ongoing lithium-ion batteries. *Advanced Materials, 34(18)*, 2106704.

Funded by the European Union