



MONASH  
University

MONASH  
BUSINESS  
SCHOOL

Department of  
Econometrics &  
Business Statistics

☎ (03) 9905 2478  
✉ [BusEco-Econometrics@monash.edu](mailto:BusEco-Econometrics@monash.edu)

ABN: 12 377 614 012

# Predicting Critical Elements in Coal Mine Waste: A Machine Learning and Spatial Statistics Approach for a Low-Emission Future

## Yuhao Long

Master of Business Analytics

33412448

[ylon0012@student.monash.edu](mailto:ylon0012@student.monash.edu)

## Evan Ginting

Master of Business Analytics

33477558

[egin0003@student.monash.edu](mailto:egin0003@student.monash.edu)

Report for

ETC5543

**24 October 2024**



## Contents

<b>Abstract</b>	<b>3</b>
<b>Background and Motivation</b>	<b>3</b>
Critical elements overview . . . . .	3
Critical elements in coals . . . . .	6
Economic concentrations . . . . .	8
Existing economic deposits . . . . .	9
Extraction of critical elements . . . . .	10
<b>Objectives and Significance</b>	<b>12</b>
<b>Data</b>	<b>13</b>
Data source and description . . . . .	13
Data scope . . . . .	14
Initial Data Analysis and wrangling process . . . . .	14
<b>Exploratory Data Analysis</b>	<b>15</b>
<b>Methodology</b>	<b>19</b>
Predicting Critical Elements . . . . .	19
Elements with Above PAAS Level . . . . .	24
<b>Results</b>	<b>24</b>
Predicting Critical Elements . . . . .	24
Elements with Above PAAS Level . . . . .	26
<b>Discussions</b>	<b>33</b>
Predicting Critical Elements . . . . .	33
Elements with Above PAAS Level . . . . .	34
<b>Conclusion</b>	<b>34</b>
<b>Appendix</b>	<b>34</b>

## Abstract

## Background and Motivation

### Critical elements overview

In this technology advancement era where mineral-based technologies are relied by many industrial sectors, critical elements become highly-sought elements in the world (Emsbo, Lawley, and Czarnota 2021). Critical elements can be defined by two main criteria: first, elements that are essential for manufacturing modern technologies, supporting economic frameworks, and ensuring national security; and second, elements with vulnerable supply chains, which can be affected by political issues, geographic concentration of extraction or production, and natural disasters (Sinclair and Coe 2024-04; Fortier et al. 2018; DISR 2023).

According to [Critical Minerals Strategy 2023–2030](#) (Geoscience Australia 2023), Geoscience Australia has identified 15 elements as highly vulnerable to future supply chain disruptions and additional 15 elements as moderate risk, which defined in Table 1 (Coyne and Campbell 2023; Skirrow et al. 2013; IEA 2024a; Fortier et al. 2018; Austrade 2024). Among these critical elements, Rare Earth Elements (REE) represent a significant subset. REEs are a group of 17 chemically similar metallic elements composed of fifteen lanthanides, scandium, and yttrium. They further grouped into light (LREEs) and heavy (HREEs) based on their atomic number, electron arrangement, and chemical properties (Reid 2018). Ultimately, critical elements (including REE) are crucial for many high-tech industries, including electronics, renewable energy, and defense (Huang, Fan, and Tiand 2018).

Global initiatives to reduce carbon emissions by transitioning to clean energy have led to increasing demand for critical elements, which are essential resources for achieving this goal (IEA 2021; Wang et al. 2022). According to IEA (2024b) research, demand for these elements is projected to double, triple, or even quadruple, depending on the scenario, relative to current production levels. Among these elements, lithium is experiencing the most rapid growth due to rising demand for electric vehicle (EV) batteries, while copper leads in terms of production volume. Graphite demand is expected to almost quadruple, and the demand for nickel, cobalt, and rare earth elements (REEs) is projected to double. Furthermore, Fortier et al. (2018) indicates that the growing reliance on critical elements is also driven by their applications across various key sectors, including energy, defense, communications, healthcare, transportation, and agriculture. These dynamics have intensified competition to discover new sources and establish stable, long-term supply chains for these vital resources (Emsbo, Lawley, and Czarnota 2021). The prominent usage of each element and its projected demand are also detailed in Table 1.

Table 1: Summary of Critical Elements: Production, Global Share, Projected Demand, and Usage

Critical Element	Production (Kilotonnes) 1	Global Production (Percentage)	Projected Demand (Kilotonnes) 2	Usage 3	Level
Aluminum and derivative (Al)	20	14	-	Aerospace alloys, Coating in Li-ion batteries	High
Cobalt (Co)	5.9	3	243.03	Li-ion battery cathodes, stainless steels, superalloys	High
Gallium (Ga)	-	-	0.25	Radar, light-emitting diodes (LEDs), photovoltaics films	High
Germanium (Ge)	-	-	0.03	Fiber/infrared optics, Polymerization	High
Lithium(Li)	61	47	615.55	Catalysts, semiconductors Li-ion batteries, aerospace alloys, ceramics	High
Magnesium(Mg)	2.6	10	30.95	Pyrotechnics, nanocomposites in automotive/aerospace	High
Manganese(Mn)	3.3	17	855	Steel, Agricultural fertilizer, lightweight alloys	High
Nickel(Ni)	150 <sup>a</sup>	4.5	2792.68	Cathodes of Li-ion batteries, Non-ferrous alloys	High
Rare-earth elements (REE) <sup>b</sup>	18	6	61.96	Catalysts, magnets, guidance, lasers	High
Silicon (Si)	0.05	1	2025	Solar PVs, Silicon wafers in electronic and photovoltaic cells	High
Tantalum (Ta)	0.057	3	0.44	Micro-capacitors, superalloys	High
Titanium (Ti)	0.85	8.4	22.69	Aerospace and marine alloys, pigment	High
Tungsten (W)	-	-	0.17	Lightning, Cutting and drilling tools, catalysts	High
Vanadium (V)	-	-	35.23	Steel or aerospace alloys	High
Zirconium (Zr)	0.5	36	11.14	Cladding fuel rods, nuclear reactors	High
Antimony (Sb)	4	4	-	Flame retardant, lead-acid batteries	Moderate
Arsenic (As)	-	-	0.55	Microwave communications, pesticides, semiconductors	Moderate
Beryllium (Be)	-	-	-	Satellite communications, lightweight alloys	Moderate

Table 1: Summary of Critical Elements: Production, Global Share, Projected Demand, and Usage (continued)

Critical Element	Production (Kilotonnes) <sup>1</sup>	Global Production (Percentage)	Projected Demand (Kilotonnes) <sup>2</sup>	Usage <sup>3</sup>	Level
Bismuth (Bi)	-	-	-	Pharmaceuticals, lead-free solders, cosmetics	Moderate
Chromium (Cr)	66.1	0.3	823.7	Steel or aerospace alloys, leather tanning	Moderate
Fluorine (F)	-	-	-	Refrigerants, dental care, nuclear processing	Moderate
Graphite (Gr)	-	-	8406.7	Rechargeable batteries, semiconductors and sensors, water filtration	Moderate
Hafnium (Hf)	-	-	0.02	Nuclear reactors, aerospace alloys	Moderate
Indium (In)	-	-	0.17	Flat-panel displays, low-Melting Alloys, semiconductors	Moderate
Molybdenum (Mo)	-	-	104.44	Improving strength and corrosion resistance in steel alloys	Moderate
Niobium (Nb)	-	-	1.97	High-Strength Low-Alloy (HSLA) Steel, superalloys, superconductors, welding	Moderate
Platinum-group elements (PGE) <sup>c</sup>	<0.01	<0.01	0.03	Catalysts, jewelry, thermocouples	Moderate
Rhenium (Re)	-	-	-	Superalloys, catalysts, electrical Contacts, filaments	Moderate
Selenium(Se)	-	-	0.26	Alloying agents, solar cells, glass production	Moderate
Tellurium (Te)	-	-	1.55	Copper or steel alloys, semiconductors, solar cells, thermoelectric Materials	Moderate

<sup>a</sup> Data collected in 2012.  
<sup>b</sup> 17 elements, including lanthanoid, Scandium (Sc), and Yttrium (Y).  
<sup>c</sup> 6 elements, including all transition metals in the d-block.  
Source: <sup>1</sup> Skirrow et al., 2013 & Coyne and Campbell, 2023; <sup>2</sup> IEA, 2024b; <sup>3</sup> Fortier et. al., 2018; Austrade, 2024

As one of the promising top global producers of these critical elements, Australia, with its abundant deposit and technological expertise, plays a pivotal role in the sustainable energy transition and supply chain stability. Australia is the largest producer of lithium, the third largest producer of cobalt, and the fourth largest producer of rare earths. It also produces significant amounts of aluminium,

nickel, and copper, which are essential for low-emission technologies like electric vehicles, solar panels, and wind turbines (DISR 2023). Australia's Government strategy to ensure the fulfillment of this potential has been proactive and multifaceted, especially in the past 5 years. The strategy include a range of incentives, finance facilities, grants and other support for the critical elements sector. Some of the important initiatives as reported in DISR (2023), are:

- 1). The Australian Government's Critical Minerals Facility, with AUD 4 billion budget, supports projects that are aligned with the nation's Critical Minerals Strategy and serve the national interest.
- 2). The Northern Australia Infrastructure Facility (NAIF) allocates up to AUD 500 million of the AUD 5 billion to help finance projects in the Northern Territory, Queensland, and Western Australia.
- 3). The Junior Minerals Exploration Incentive (JMEI) promotes investment in small minerals-exploration firms that focus on greenfield exploration.
- 4). Australian federal, state and territory government authorities are collaborating on the AUD 10 million Critical Minerals National Productivity Initiative to develop pre-feasibility studies of common-user infrastructure for the critical elements sector.
- 5). The Major Projects Facilitation Agency (MPFA) supports developers of projects over AUD 20 million by providing information on Australian Government regulations and approvals, mapping out critical approval processes, and communicating with regulators to address issues.
- 6). The Critical Minerals Production Tax Incentive offers a production incentive worth 10 percent of relevant processing and refining costs for Australia's 31 critical elements. This incentive is available for up to 10 years per project for production between 2027–28 and 2039–40, provided the projects reach final investment decisions by 2030.

### **Critical elements in coals**

The green technology initiatives and commitments by countries worldwide for decarbonisation have led to a growing market of clean energy and technology. As a result, critical elements, including REE, are projected to be in high demand in the future (United States Department of Energy 2017). Currently, China leads the world in REE production, dominating the international trade and global value chain of rare earths (USGS 2024). However, their recent export restrictions on REE have led to a disruption in the global supply chain (Mancheri 2015). In response, there is an increasing focus on identifying alternative sources of critical elements, with coal being explored as a potential new source of critical elements (Hodgkinson and Grigorescu 2021).

In Hodgkinson and Grigorescu (2020) element mapping project on Bowen basin Table 2, the largest coal reserves in Australia, the concentration of element composition is subjective to sample’s lithology rather than the depth grading:

- 1). In coal and its derivatives, although the majority of element concentrations fall below the benchmark when compared to Post-Archaeon Australian Shales (PAAS) standard (McLennan 2011), a widely used geochemical reference material in the average composition of shales, local samples exhibit enrichment in Heavy Rare Earth Elements (HREEs) and scandium(Sc). Additionally, concentrations of the critical element bismuth (Bi) are abnormally elevated, showing levels 4-6 times higher than the crustal average.
- 2). Siltstone and mudstone yield unremarkable findings in terms of elemental enrichment, with most element concentrations failing to meet significant thresholds. However, the concentration of cobalt compounds approaches the crustal average, suggesting potential economic value that warrants further investigation.
- 3). Tuffaceous rock, formed from volcanic ash, is rich in pumice and lithic fragments. Samples reveal elevated concentrations of strategic elements, including Rare Earth Elements (REEs), gallium (Ga), and bismuth (Bi). Moreover, a potential lithium-rich borehole has been identified, with lithium concentrations approximately five times higher than the crustal average.

Table 2: Summary of Hodginkson’s research on critical element mapping in coal mines

Critical Element	PAAS Standard	Average Concentration (ppm)	Highest Concentration (ppm)	Above Crustal Average (Percentage)
Coal Seam & Associate				
Lithium	21.0	13.7	25	22
REE	184.0	115.8	205	11
Cobalt	17.0	16.9	30	44
Nickel	47.0	11.2	40	0
Tantalum	1.0	0.33	1	33
Vanadium	97.0	85.6	140	22
Zirconium	193.0	102.1	160	0
Gallium	17.0	12.3	25	44
Bismuth	0.2	0.59	1	67
Chromium	92.0	14.33	51	0
Niobium	12.0	7.9	32	11
Molybdenum	1.0	1.33	3	78
Siltstone & Mudstone				
Lithium	21.0	17.2	28	33
REE	184.0	138.8	189	17

**Table 2: Summary of Hodginkson's research on critical element mapping in coal mines (continued)**

Critical Element	PAAS Standard	Average Concentration (ppm)	Highest Concentration (ppm)	Above Crustal Average (Percentage)
Cobalt	17.0	<b>39.3</b>	<b>134</b>	67
Nickel	47.0	25.8	<b>73</b>	33
Tantalum	1.0	0.2	1	17
Vanadium	97.0	<b>102</b>	<b>225</b>	33
Zirconium	193.0	116.8	<b>243</b>	17
Gallium	17.0	15	<b>22</b>	33
Bismuth	0.2	0.18	<b>0.4</b>	67
Chromium	92.0	51.33	<b>266</b>	17
Niobium	12.0	4.67	8	0
Molybdenum	1.0	0.67	1	67
<b>Tuffaceous Rocks</b>				
Lithium	21.0	<b>21.5</b>	<b>105</b>	12
REE	184.0	<b>244</b>	<b>441</b>	88
Cobalt	17.0	9.25	<b>19</b>	12
Nickel	47.0	3.75	30	0
Tantalum	1.0	<b>1.1</b>	<b>2</b>	100
Vanadium	97.0	27.9	70	0
Zirconium	193.0	153.75	<b>282</b>	25
Gallium	17.0	<b>32.25</b>	<b>37</b>	100
Bismuth	0.2	<b>0.71</b>	<b>1.2</b>	100
Chromium	92.0	0	0	0
Niobium	12.0	9.63	<b>18</b>	25
Molybdenum	1.0	1	<b>5</b>	50

Source: Adapted from Hodginkson et al., 2020

## Economic concentrations

Before the recent surge in demand, extracting critical elements (including REE) from coal was considered costly, however, methods have been developed that reduce the cost and also environmentally friendly. Coal and coal byproducts are substantially enhanced with trace metals and have been proposed as a potential source (Eterigho-Ikelegbe, Harrar, and Bada 2021). A report by US Department of Energy suggests that extracting REEs from coal material already mined for other purposes, either as a dual product or byproduct, could be more cost-effective than a dedicated REE mining. Despite the challenge of processing large volumes to obtain economic concentrations, cost may be reduced due to the pre-processed (mined, crushed, and washed) state of the materials and transported to areas with existing infrastructure (United States Department of Energy 2017; Hodgkinson and Grigorescu 2021).



Based on research, the elements and metals in coal that are considered to have the best chance for economic recovery are: REE, Ag, Au, PGEs, Be, Se, V, Ga, Sb, Sc, Mo, W, Re, Ge, U, Y, Nb, Zr, Al (Dai and Finkelman 2018).

Efficient extraction of critical elements from coal or coal ash, and finding coal sources with highly elevated concentrations of these elements, are essential prerequisites. The value of the element in any coal sources largely determined by their concentration, which varies based on geological and geochemical conditions. Therefore, to economically recovering critical elements from coal is by identifying sources with the highest critical elements levels and accessibility first before proceeding to extraction and recovery stages (Eterigho-Ikelegbe, Harrar, and Bada 2021). Reid (2018) in his report suggests that critical elements average level in coal is typically just 35 ppm, which is insufficient for economic extraction. While Seredin and Dai (2012) in their reports suggests that a cutoff grade for Rare Earth Oxides (ROE) for coal seam has been suggested as 800-900 ppm. Table 3 provide suggested cut-off grade for 18 critical elements. Additionally, Talan and Huang (2022) reports that coal sources are often relatively enriched in the heavy rare earths (HREEs, Ho, Er, Tm, Yb, Lu) and critical rare earths (CREEs, Y, Nd, Dy, Eu, Tb) compared to traditional mineral deposits. A cut-off grade of 115–130 ppm rare earth element on a whole mass basis may be considered economical. However, the U.S. Department of Energy set a criterion in their work assessing raw coal with total rare earth element content greater than 300 ppm on a whole dry coal basis.

### **Existing economic deposits**

Currently, there are no coal mines or coal basins extracting critical elements (including REE) at a commercial states. Most sources found conclude that we are still in the research and development process (Hodgkinson and Grigorescu 2020; Dai and Finkelman 2018; Eterigho-Ikelegbe, Harrar, and Bada 2021; Honaker et al. 2019; Talan and Huang 2022). A report by Eterigho-Ikelegbe, Harrar, and Bada (2021) suggests that the research and development are primarily at the laboratory and pilot scale stages. For example, scholars like Honaker and coworkers who have significant experience in this subject, have constructed a 0.23 t/h solid feed pilot plant for testing different coal-based feedstocks (Honaker et al. 2019). However, commercial-scale extraction processes have not yet been widely implemented. Talan and Huang (2022) report support this finding as well. They propose a potential approach for commercial-scale extraction of critical elements, indicating ongoing effort to bridge the gap between laboratory research and commercial application. Additionally, they note that further validation and techno-economic analysis are required before these processes can be scaled up commercially. Furthermore, scholars who focusing their research on China's coal resources also suggest further research is needed to scale up the extraction process (Qin et al. 2015b,a; Liu et al. 2024; Zhao et al. 2019; Zou et al. 2023). Given that China is the leader in critical elements

**Table 3: Selected critical elements with their suggested cut-off grade (ppm)**

Critical Element	Suggested Cut-off Grade
U	1000
Ge	300
V	1000
Se	500–800
Ga	100 (50)
REE	1000
Y	300
Sc	100
Nb	300
Zr	2000
Mo	1000
Re	1
W	1000
Au + Pt + Pd	2
Ag	10
Be	300
Sb	1000
Cs	150

*Source:* Adapted from Dai and Finkelman, 2018

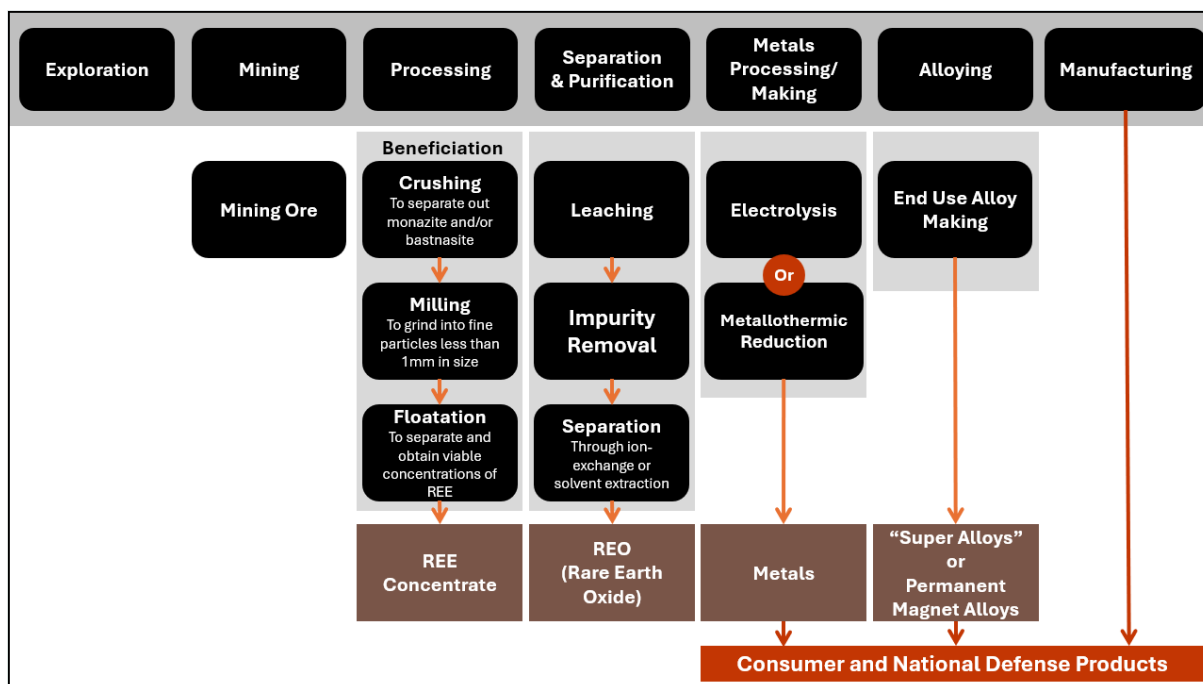
production, this fact indicates that their production of critical elements does not primarily come from coal and its byproducts.

We can conclude that significant research and development efforts have been made towards commercial extraction of critical elements from coal and coal byproducts. However, it is still long way from actual large-scale implementation. Most scholars indicate the necessity of considering sustainability factors. Effective waste management and treatment, along with the evaluation of extraction costs and technical feasibility, are important. At the end of the day, these elements are going to be utilised in green technology initiatives, such as renewable energy systems and electric vehicles, which demand high-purity materials. Ensuring that their extraction processes are environmentally and economically sustainable is crucial to align with the goals of these green technologies initiatives. Furthermore, ongoing research is vital to bridge existing technology gaps and develop robust, scalable methods that can transition from pilot projects to commercial operations.

### **Extraction of critical elements**

There are several stages required for critical elements can be extracted from coal and coal byproducts. Coal and coal byproducts is first prepared by crushing and milling to release the organic from inorganic elements to generate fine concentrates. These concentrates are then undergo further separation techniques such as physical (froth flotation, electrostatic, magnetic, gravity) and chemical (leaching).

Later, it will further processed into purification to extract the REE into pure oxide form (e.g. REO). REO can be standalone products, however, market demand dictates their conversion into metals. Thus, in the refining stage, using an electrolysis or a metallothermic reduction process, REOs will be converted into high-purity rare earth metals. These metals may then be alloyed with other elements to make them harder and stronger for the end-use market (Eterigho-Ikelegbe, Harrar, and Bada 2021; United States Department of Energy 2017; Talan and Huang 2022). The figure 1 visualise this process.



**Figure 1:** REE Value Chain (Source: US Department of Energy, 2017)

Recently, water-leaching is a prevailing approach since it mitigate the flaw from traditional techniques. Water leaching can be considered as less environmentally detrimental compared to strong acid/alkaline leaching, as well as cost effective for solvent selection. The crucial stages on this preparation workflow are low temperature activation and water leaching. During the stage of low temperature activation, the chemical reaction within coal fly ash (CFA) will be facilitated by complexation agents (ammonium salts or weak acids) in covered alumina crucibles, which help liberate critical elements from the matrix of the CFA. After the activation and cool down to ambient temperature, the tablets are placed in water for the leaching and dissolve process. Water acts as the leaching solvent, extracting these soluble elements into the leachate. The configuration in temperature and mass ratio of solvent will be the vital determinant for optimized recovery. Take Lithium example, it can achieve a stable leaching efficiency of 90% through ammonium fluoride leaching at 150°C with a  $SiO_2/NH_4F$  mass ratio of 1:1.35 Xu et al. (2021).

Another innovation is Hydrophobic-Hydrophilic Separation (HHS), designed to leverages the disparity of affinity (water-repellent & water-friendly) properties of substances to achieve separation. It can treat as a complementary application for small particle delamination without size limit, providing flexible and extensible purpose in the segregation of ultrafine coal Hodgkinson and Grigorescu (2021).

In precious Hodgkinson and Grigorescu (2020) element mapping project on Bowen basin, the largest coal reserves in Australia, the concentration of element composition is subjective to sample's lithology rather than the depth grading:

- 1). In coal and derivative, albeit majority of element concentrations is inferior of the benchmark against earth crust average, local samples exhibit enrichment in HREE and Scandium in respect to McLennan (2011) Post-Archaean Australian Shales (PAAS) standard, while abnormal 4-6 times higher than crustal average in moderate critical element, Bismuth (Bi).
- 2). Siltstone and mudstone has a lackluster finding to classify enrichment for majority of elements concentration, except for the concentration of Cobalt compound barely meet crustal average, whose ubiquitous economic value may warrant further examination.
- 3). As the sediment from volcanic ash, tuffaceous rock is rich in pumice and lithic fragments. The sample display a series of elevated concentrations of strategic elements including REE, Ga and Bi. Besides, a potential Lithium-rich borehole is found, with approximate 5 times higher than crustal average.

## Objectives and Significance

The primary objective of this study is to develop a predictive model using machine learning and spatial statistics to identify and quantify critical minerals—such as copper, lithium, nickel, cobalt, and rare earth elements—in coal mine waste. By analysing existing coal data, this study aims to provide valuable insights into the potential recovery of these essential minerals from coal deposits, contributing to waste reduction and supporting the transition to low-emission technologies. The outcome of this study will help assess the economic value of these minerals and promote sustainable practices within the mining and energy sectors.

The study utilises multiple datasets, some of which are publicly available while others are confidential. The public dataset was released by the Australian Coal Industry's Research Program (ACARP) in 2021, whereas the confidential data comes from Matrix Geoscience's clients. Although the datasets originate from different sources, both contain information on the concentration of various critical elements sampled from coal mines and power plants across Australia. Additionally, the study employs data on elemental concentrations from Post Archean Australian Shale (PAAS), which represents the

background crustal abundance of elements. This comparison helps determine whether the elemental concentrations in the coal samples exceed typical crustal abundance levels.

This study is significant because it addresses a critical gap in understanding the presence and potential recovery of valuable minerals from coal mine waste. By employing advanced machine learning and spatial statistics techniques, the project has the potential to revolutionise how the mining industry views coal waste—transforming it from an environmental challenge into a resource opportunity. The findings could contribute to more sustainable mining practices, reduce waste, and support the global transition to low-emission technologies by securing a local supply of critical minerals. Furthermore, this research could pave the way for further exploration of coal mine waste as a valuable resource in other regions, creating economic and environmental benefits on a broader scale.

## Data

### Data source and description

As the escalating research into this flourishing market, the largest national scientific research institution, CSIRO (Commonwealth Scientific and Industrial Research Organisation) in Australia Federation has initiated the early-stage research into the potential of element extraction in coal mines. Launched in 2001, The C29030 project in Australian Coal Industry's Research Program (ACARP) (Hodgkinson and Grigorescu 2021) has gauged the first comprehensive map of element concentration in Australian coal and provided the open dataset for testing result. This element mapping targets on 50 elements that has rarely studied from nearly 90 samples in Queensland and New South Wales, which are carefully selected in 13 coal mines and power plants to represent 6 basins, covering a range of coal measures and seams of varying ages. The dataset is a wide format excel spreadsheet, with the information of detected element name and their corresponding concentration in parts per million (ppm) for each labelled sample(e.g.2648CR01), as well as the sample's coordinate in a separate record.

Simultaneously, private-sector analyses are also promptly emerging, utilising a variety of advanced analytic techniques, including four-acid digestion with ICP-AES and fusion-based acid digestion with ICP-MS. Although the report format is not standardised (long or wide format), but all of them cover basic information for element name and concentration as well as additional borehole composite index in ppm, which is sufficient for further deep-dive investigation.

Eventually, PAAS file attached to (McLennan 2011) Post-Archean Australian Shale normalised score, representing the average composition of shale formed after the Archean eon, which is a geochemical reference baseline in the comparison of element enrichment in this report.

## Data scope

The 'Element\_Value\_ppm' represents concentration values of elements (measured in parts per million - ppm) obtained from a series of lab tests on real rock samples collected from nature. In this project, the rocks was analysed between two test, either ME-MS81 or ME-4ACD1. Both tests are whole rock analytical methods that involve a fusion decomposition process, which breaks down all minerals in the sample. The key difference between them is their sample preparation methods: ME-MS81 determine elements concentrations from a fused sample with ICP-MS (Inductively Coupled Plasma Mass Spectrometry) instrument analysis, while ME-4ACD1 determines concentrations by adding base metals through a four-acid digestion with ICP-MS. These tests were conducted by ALS, a global company providing laboratory analysis services for the minerals industry. More information can be found on their website [url].

From a cost perspective, the ME-MS81 test costs AUD 35.84 per sample, whereas ME-4ACD1 test costs AUD 9.96 per sample, which is nearly four times price difference. Given the lower cost of the ME-4ACD1, it will be more economically feasible for future rock samples. Hence, the goal of the predictive analysis is to determine if predicting critical elements, particularly REE, HREE, and LREE, using elements tested in this test is practical. ME-4ACD1 tests covers only eleven elements, which are Cadmium (Cd), Molybdenum (Mo), Lead (Pb), Silver (Ag), Scandium (Sc), Zinc (Zn), Lithium (Li), Cobalt (Co), Copper (Cu), Nickel (Ni), and Thallium (Tl). Therefore, our predictive model will utilise these elements to predict REE, HREE, and LREE. However, this data scope applies only to the predictive modelling part.

## Initial Data Analysis and wrangling process

To prepare for exploratory data analysis and predictive modeling, different data formats must be collated into a standardised structure, which involves converting the data into a long format and trimming the trivial variables to ensure consistency and readiness for further analysis. The cleaning data will include element symbol, element value in ppm and project region only, but other specific cleaning requirements still need to be addressed. First and foremost, The private projects are unable to disclose the location in detail, therefore; they will be replaced to Confidential A, B and C respectively. Furthermore, partial element values are recorded as character due to their undetected status(nd) and subtle presence(< sign). a value of -999.00 will be assigned to represent their negative relationship without compromising data integrity. Another significant issue in data wrangling is the sample duplication, mainly because cross detection methods are applied in their analysis. To resolve this, we will remain the element concentration value for the most precise observation range.

After the preliminary tidying, the data is organised and concise but some information is still lacking in subsequent study. Since the analytic technique will only detect single element, the critical element

groups like REE and its subcategories are absent in calculation. the element list need to be expanded by the aggregation of the concentration values from their corresponding symbols. Additionally, the data must be joined with PAAS file and normalised value should be derived by using formula  $\text{Normalized value} = \frac{\text{Element value(ppm)}}{\text{PAAS score(ppm)}}$ , along with Boolean flag for enrichment (above background,  $\geq 1$ ). The final step is replacing outliers and missing values by the median value within their respective project. If no project-specific values are available the global median of the element will be applied. The median replacement is a mild imputation to ensure the robustness and consistency of data structure and in turn, maintain data integrity without distorting the subsequent modelling.

In a nutshell, The final cleaning data include total 8 variables and 11032 observations. Figure 2 shows the correct data type for all variables, with no missing value left.

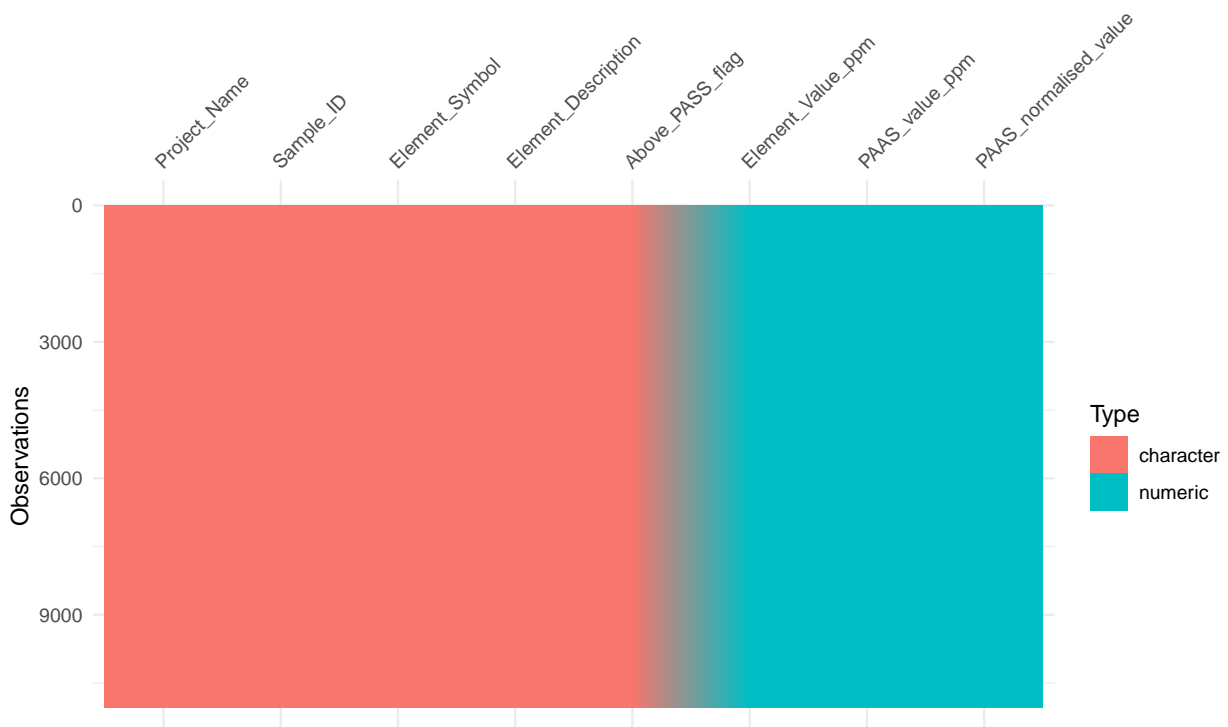


Figure 2: *n at-a-glance of final data wrangling result*

## Exploratory Data Analysis

The Exploratory Data Analysis (EDA) was carried out on critical elements concentration data, which provide information about the concentration of critical mineral across different areas in Australia that has been normalised using PAAS (Post-Archean Australian Shale) standard. The primary objective of this analysis is to gain a deeper understanding of the data's structure and key characteristics. Through this, we aim to identify significant trends, correlations, and outliers that may influence the outcomes of the study.

The EDA for critical elements concentration data starts with a descriptive statistics for every critical elements included in the data. These summary statistics can be found in the Table 4. The descriptive statistics features include the minimum value of the critical elements, maximum value, the mean, the median, the standard deviation, and the kurtosis of the variable.

**Table 4: Descriptive Statistics of Elements' Normalised Value**

Element Sym- bol	Element Description	min	max	mean	median	sd	kurtosis
Ag	Silver	2.000	13.200	4.142	3.600	2.389	5.783
Al	Aluminium	0.000	8.500	0.845	0.438	0.982	34.192
Au	Gold	0.556	12.222	3.521	3.333	2.064	7.764
Ba	Barium	0.013	18.182	1.108	0.750	2.027	43.446
Be	Beryllium	0.033	6.000	1.126	0.667	1.134	7.156
Bi	Bismuth	0.787	25.197	3.685	2.402	3.302	17.818
Cd	Cadmium	0.102	7.347	1.271	0.918	1.115	9.839
Ce	Cerium	0.056	5.938	0.956	0.985	0.595	21.060
Co	Cobalt	0.118	7.882	0.906	0.588	1.133	17.661
Cr	Chromium	0.012	10.807	0.441	0.205	1.048	54.642
Cs	Caesium	0.024	6.870	1.068	0.923	0.910	12.858
Cu	Copper	0.040	10.200	1.681	1.880	1.160	13.140
Dy	Dysprosium	0.114	5.286	1.522	1.517	0.847	5.566
Er	Erbium	0.087	4.957	1.401	1.320	0.815	6.383
Eu	Europium	0.114	7.773	1.676	1.648	0.969	8.535
Fe	Iron	0.000	9.689	0.474	0.180	1.066	45.053
Ga	Gallium	0.076	3.076	1.293	1.291	0.684	2.148
Gd	Gadolinium	0.132	6.605	1.589	1.513	0.901	7.420
Ge	Germanium	0.088	43.750	7.481	0.344	13.031	3.653
HREE	Ho+Er+Tm+Yb+Lu	0.036	0.995	0.282	0.267	0.157	6.712
Ho	Holmium	0.125	4.750	1.352	1.306	0.767	5.962
In	Indium	0.400	7.200	1.434	1.000	1.117	9.581
LREE	La+Ce+Pr+Nd+Pm+Sm	0.089	4.214	1.020	1.064	0.540	6.992
La	Lanthanum	0.100	2.540	0.871	0.930	0.480	2.566
Li	Lithium	0.250	14.250	2.386	2.000	1.974	9.533
Lu	Lutetium	0.000	6.250	1.540	1.438	0.919	7.067
MREE	Eu+Gd+Tb+Dy+Y	0.088	4.384	1.361	1.344	0.756	5.952
Mn	Manganese	0.002	18.717	0.594	0.109	1.884	66.528
Mo	Molybdenum	0.067	13.733	2.971	2.667	2.195	10.237
Nb	Niobium	0.042	3.567	0.622	0.647	0.402	16.601
Nd	Neodymium	0.065	4.423	1.145	1.221	0.645	4.693



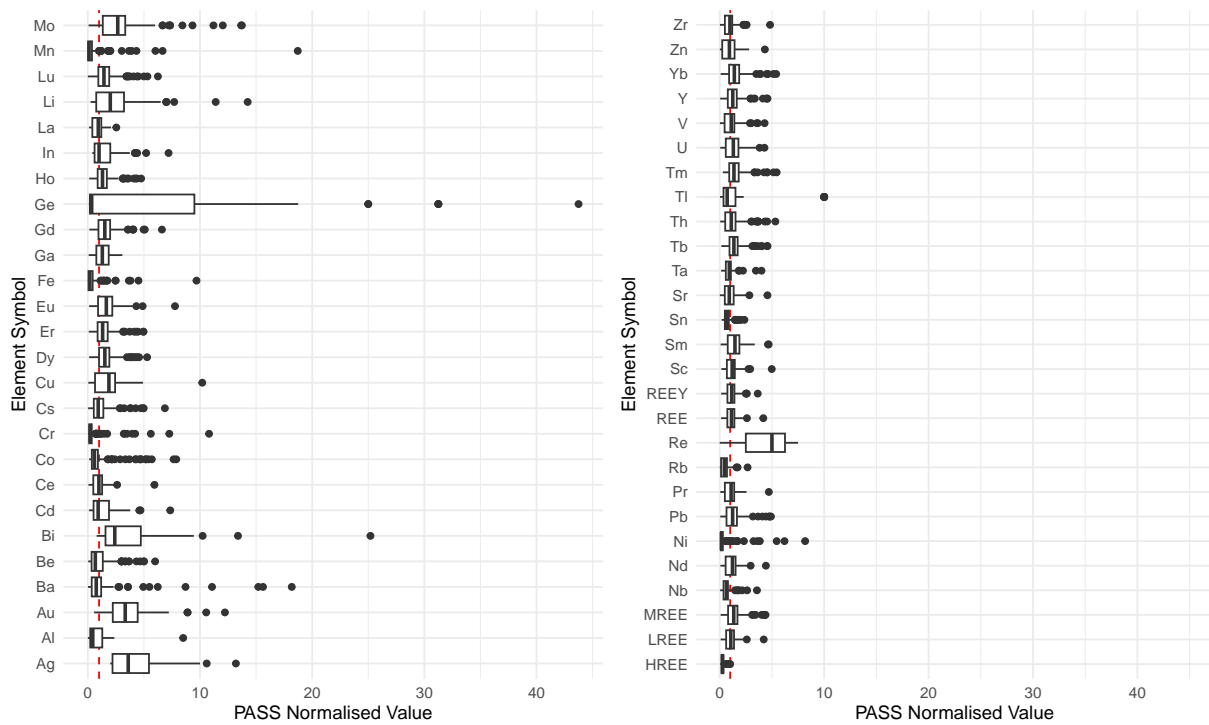
**Table 4: Descriptive Statistics of Elements' Normalised Value (continued)**

Element Sym- bol	Element Description	min	max	mean	median	sd	kurtosis
Ni	Nickel	0.023	8.182	0.386	0.159	0.869	44.099
Pb	Lead	0.052	4.909	1.276	1.235	0.856	6.139
Pr	Praseodymi	0.056	4.704	1.033	1.092	0.596	7.452
REE	La+Ce+Pr+Nd+Sm+Eu+Gd+Tb+Dy+Ho+Er+Tm+Yb+Lu	0.134	4.174	1.089	1.130	0.539	6.574
REEY	La+Ce+Pr+Nd+Sm+Eu+Gd+Tb+Dy+Ho+Er+Tm+Yb+Lu+Y	0.138	3.641	1.120	1.151	0.528	4.600
Rb	Rubidium	0.001	2.670	0.472	0.441	0.396	5.920
Re	Rhenium	0.000	7.500	4.286	5.000	3.134	1.792
Sc	Scandium	0.162	4.985	1.137	1.191	0.610	8.482
Sm	Samarium	0.089	4.689	1.421	1.460	0.785	4.173
Sn	Tin	0.182	2.364	0.726	0.655	0.350	6.862
Sr	Strontium	0.006	4.571	0.955	0.916	0.608	6.965
Ta	Thallium	0.133	4.000	0.865	0.920	0.482	14.453
Tb	Terbium	0.156	4.578	1.378	1.359	0.760	5.576
Th	Thorium	0.044	5.329	1.172	1.107	0.853	6.382
Tl	Tantalum	0.030	10.000	1.942	0.715	3.123	5.782
Tm	Thulium	0.303	5.455	1.453	1.364	0.816	7.679
U	Uranium	0.054	4.286	1.289	1.304	0.805	3.219
V	Vanadium	0.019	4.299	1.038	1.103	0.666	5.959
Y	Yttrium	0.045	4.568	1.278	1.250	0.764	6.443
Yb	Ytterbium	0.091	5.409	1.462	1.400	0.854	6.361
Zn	Zinc	0.014	4.324	0.921	0.930	0.699	4.134
Zr	Zirconium	0.021	4.821	0.923	0.979	0.592	9.671

The most important statistics from this table is kurtosis. Kurtosis measures the combined weight of the tails of a distribution relative to its centre. In this way, we can use kurtosis as an indicator of the presence of outliers. A high kurtosis values is indicative of outliers. The summary statistics suggest that Al, Ba, Bi, Ce, Co, Cr, Cs, Cu, Fe, Mn, Nb, and Ni are having many outliers due to high kurtosis value. Validating the outliers will be easier with data visualisation, which will be presented in the next step.

The next step was analysing the distribution of each element's normalised value. By utilising boxplots, it will help identifying the outliers of each critical elements, while also identifying which element's has median value above the PAAS standard. Obtaining these information will help to understand

whether certain element is naturally abundance in the nature, thus it is most likely economically efficient to extract. The figure 3 provide these informations.



**Figure 3:** *Distribution of Critical Elements (Box-Plot)*

The boxplots shows some critical points that can be summarised as follow:

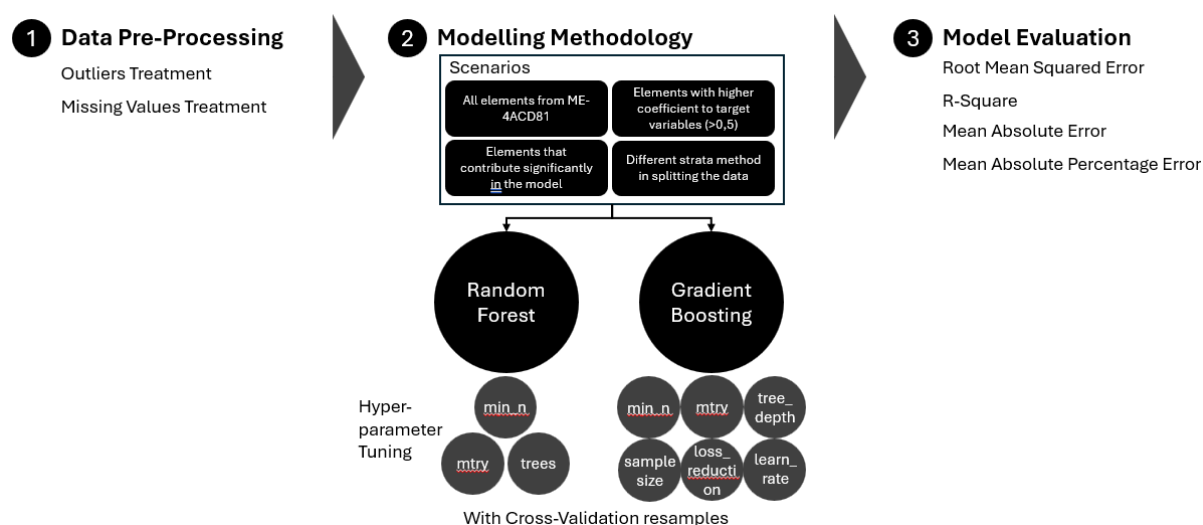
- **Medians:** Most of the elements have their medians close to 0, indicating that the majority of the values are low or concentrated around a lower range. Considering that the threshold to define whether a sample is below/above background is 1 (vertical red dashed line), some elements that have median above the threshold are Re, Ag, Au, Mo, Bi, Li, Cu, Eu, Dy, Gd, Sm, Lu, Yb, Tm, Tb, MREE, Er, Ho, U, Ga, Y, Pb, Nd, Sc, REEY, REE, Th, V, Pr, LREE
- **Spread and variability:** The elements exhibit varying degrees of spread. For example, Ge shows a wide range with its box stretching from a low value near 0 to a higher value around 10, indicating a large variability. On the opposite side, elements like Mn, Fe, Cr, Co, Rb, Ni, HREE have very narrow IQR, indicating less variability.
- **Outliers:** Several elements have significant outliers, as indicated by the dots outside the whiskers of the box plots. For instance, Ge, Bi, Mn, and Ba show notable outliers far from the main data range. Additionally, some elements that was reported from table 4, which have high kurtosis are Al, Ba, Bi, Ce, Co, Cr, Cs, Cu, Fe, Mn, Nb, and Ni. These outliers suggest the presence of some unusually high or low values for these elements, which could be of interest for further investigation.

- Symmetry and skewness: Some elements like Ge, Bi, and Ag appear to have a right skew, with longer whiskers or outliers extending to the right, indicating that the distribution of their normalised values has a tail on the higher end. Elements like Ga and Sm show a more symmetrical distribution with whiskers extending fairly equally on both sides of the box.
- Comparison across Elements: Ge stands out with a particularly large spread and median, making it an outlier among the elements. Conversely, many rare earth elements (REE, REEY, MREE, LREE, HREE) have relatively low medians and a small spread, indicating that their normalised values are generally low and clustered.

## Methodology

### Predicting Critical Elements

The predictive modeling process in this project follows a structured methodology, consisting of key steps aimed at selecting the most accurate and robust model for predicting the target variable. Each step contributes to refining the model's performance, from data preprocessing and feature selection to model training and evaluation. The overall framework of this methodology is outlined in the figure 4.



**Figure 4:** The mindset of model build-up in predictive analysis

The methodology for this project follows a three-stage process. First, data preprocessing is conducted, addressing outliers and missing values to ensure clean data. Next, several modeling scenarios are explored using Random Forest and Gradient Boosting, with hyperparameter tuning applied to optimise model performance. These scenarios include using all elements, elements with high correlation to target variables, and different stratification methods for data splitting. Finally, model performance is evaluated using key metrics such as Root Mean Squared Error (RMSE), R-squared, Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE).

### **Data Pre-Processing**

On one hand, outlier is well known as the observations/data points that deviate significantly from other observations in a dataset. On the other hand, missing value is observation when there is no data stored for certain variables. In predictive modelling, outliers and missing value can have an effect to the regression model when predicting the target variable in a negative way. For example, outliers value will disrupt the normal pattern of the dataset, while missing value will introduce uncertainty to the model. Thus, if not treated, they might lead the model to make inaccurate predictions. With this concern, outlier and missing value will be treated, aimed to reduce the negative effect to the model performance.

For outliers, Tukey (1977) introduce the approach to detect outlier in a dataset. He defined that outlier can be detected by the 1.5 Interquartile Range (IQR). An observation is considered an outlier if it lies below  $Q1 - 1.5 \times IQR$  or above  $Q3 + 1.5 \times IQR$ , where IQR is the interquartile range between the first and third quartiles. This method is particularly effective for detecting outliers in data with a skewed distribution. This method will be applied to the dataset for this project in detecting outliers. Next, in treating the outliers, [Barnett] stated that, as a simple approach, outliers can be replaced with sample median for robust estimation. This, with the additional issue regarding lab test error in determining the elements' concentration values, the outliers will be replaced with median value of where the samples were obtained from. For example, for outliers detected in "Sc" data from "Confidential Project A", will be replaced with median value of "Sc" of "Confidential Project A".

For missing values, Hastie, Tibshirani, and Friedman (2009) stated that, there are number of ways in treating missing value (assuming random missing value), which are remove observations with any missing value; rely on the learning algorithm to handle missing values during the training process; and fill in all missing values prior to training process. The last option was chosen and the missing value are filled with the median value of where the samples were obtained from, but if the element does not exist in that particular project, then the global median will be applied. For example, for missing value detected in "Sc" data from "Confidential Project A", then those missing values will be replaced with median value of "Sc" in general.

### **Stratified sampling and Initial split**

Unlike simple random sampling, stratified sampling is deployed to preserve the diversity of project's key feature across the training and test datasets, which leads to unbiased predictive modeling. However, some projects have limited borehole samples measured in the laboratory and their patterns might not be adequately presented in the initial data split. Thus, the project-based aggregation in terms of the classification within correlation matrix offers an effective solution to mitigate this bias. To properly address the hidden pattern in the neighbor similarity against global matrix, The approach

incorporates completed pairwise method to neglect “NA” in correlation calculation and all element in the list (Menon 2020). Based on this strata division, the raw data will be split to training and testing set with 2/3 proportion.

### **Prediction Model**

In this project, the tidymodels package will be used to implement predictive modeling in R. Tidymodels is a comprehensive framework built on tidyverse principles, designed for modeling and machine learning [tidymodels]. Two prediction methods are employed in this analysis: Random Forest (RF) and Gradient Boosting Tree (GBT), both of which are well-suited for handling complex, non-linear relationships in the data. These methods have been selected for their robustness in prediction and their ability to minimise overfitting through ensemble learning techniques.

Random Forest is an ensemble learning method that constructs multiple decision trees using subsets of the training data through a technique called ‘bagging’ (Bootstrap Aggregating). Bagging is a variance reduction technique for an estimated prediction function. In regression, this involves fitting the same regression tree multiple times to different bootstrap samples of the training data and averaging the results. Random Forest is a significant modification of bagging, as it builds a large collection of de-correlated trees and then averages their outputs. The concept behind Random Forest is to improve the variance reduction achieved by bagging by reducing the correlation between the trees, while minimizing any increase in variance. This is accomplished by randomly selecting input variables during the tree-growing process (Hastie, Tibshirani, and Friedman 2009; Breiman 2001). The engine setting in tidymodels will be set to randomForest.

Whereas random forests build an ensemble of deep independent trees, GBT build an ensemble of shallow trees in sequence with each tree learning and improving on the previous one. The main idea of boosting is to add new models to the ensemble sequentially. In essence, boosting attacks the bias-variance-tradeoff by starting with a weak model (e.g., a decision tree with only a few splits) and sequentially boosts its performance by continuing to build new trees, where each new tree in the sequence tries to fix up where the previous one made the biggest mistakes (i.e., each new tree in the sequence will focus on the training rows where the previous tree had the largest prediction errors) (Boehmke and Greenwell 2019). The engine setting in tidymodels will be set to xgboost.

### **Hyperparameter tuning**

Building an effective machine learning model is a complex and time-intensive process that requires selecting the appropriate algorithm and optimizing the model’s architecture through hyperparameter tuning. Machine learning models have two kinds of parameters: model parameters, which are initialised and updated during the training process (such as weights in neural networks), and hyperparameters, which cannot be learned from the data and must be set before training begins.

Hyperparameters define key aspects of the model's structure and learning process. Examples include the penalty parameter in support vector machines, the learning rate in neural networks, the activation function, and optimizer types. Hyperparameter tuning is essential for improving model performance, as it involves finding the optimal combination of settings that maximises predictive accuracy (Yang and Shami 2020). Tree-based model have more hyperparameter that can be tuned, however, for this project, seven key hyperparameters were the top priority for tuning, which are:

- 1). `mtry` (number of predictors to sample): Controls the number of predictors considered at each split in a decision tree. A smaller `mtry` introduces more variability among trees, while a larger `mtry` can lead to more accurate splits but risks in overfitting problem.
- 2). `min_n` (minimal terminal node size): Determines the minimum number of observations in the ending node. Smaller values allow trees to grow deeper, capturing more details in training data, while larger values lead to simpler models that may generalize better.
- 3). `trees` (number of trees): Define the the total number of decision trees in the ensemble. More trees generally improve performance but also increase computational cost.
- 4). `tree_depth` (tree depth): Controls the depth of the individual trees.
- 5). `learn_rate` (learning rate): It determines the influence of each tree on the final prediction and regulates the pace at which the algorithm progresses along the gradient descent path (or learns).
- 6). `sample_size` (sample size): Define the the total number of decision trees in the ensemble. More trees generally improve performance but also increase computational cost.
- 7). `loss_reduction` (minimum loss reduction): The reduction in the loss function required to split further

To determine the best combination of these hyperparameters in the grid, k-fold cross-validation was employed. In k-fold cross-validation, the dataset is randomly divided into K equal-sized groups, known as “folds.” For each resampling iteration, the model is trained on K-1 folds, while the remaining fold is used for validation. In standard k-fold cross-validation (without repetition), the process is repeated K times, ensuring that each fold is used once for validation (Kuhn and Wickham 2020). Cross-validation provides a reliable estimate of the optimal tuning parameter value.

### **Model evaluation**

When assessing a model's performance on a particular dataset, it is essential to have a method for determining how closely its predictions align with the actual observed data. This involves measuring the degree to which the predicted value for an observation approximates the true value for that observation. In this project, several metrics will be used, including Root Mean Squared Error (RMSE),

Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and R-Squared. These metrics provide a comprehensive understanding of how well the model fits the data, capturing both the average magnitude of error and the variability explained by the model (James et al. 2013; Hastie, Tibshirani, and Friedman 2009).

**Root Mean Squared Error (RMSE)** RMSE is a standard way to measure the error of a model in predicting continuous outcomes. It represents the square root of the average of the squared differences between the observed and predicted values. RMSE gives more weight to large errors, making it particularly useful when large errors are undesirable (James et al. 2013; Hastie, Tibshirani, and Friedman 2009). The formula for RMSE is:

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

where  $y_i$  is the actual value,  $\hat{y}_i$  is the predicted value, and  $n$  is the number of observations.

**Mean Absolute Error (MAE)** MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It's a linear score, meaning all individual differences are weighted equally (James et al. 2013; Hastie, Tibshirani, and Friedman 2009). The formula for MAE is:

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

MAE is easy to interpret because it provides the average error in the same units as the target variable.

**Mean Absolute Percentage Error (MAPE)** MAPE expresses the error as a percentage of the actual values, making it useful for comparing models on different scales or datasets. However, MAPE can be problematic when actual values are very close to zero (James et al. 2013; Hastie, Tibshirani, and Friedman 2009). The formula for MAPE is:

$$MAPE = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

**R-squared (Coefficient of Determination)** R-squared explains the proportion of the variance in the dependent variable that is predictable from the independent variables. It provides insight into how well the model captures the variability in the data, with a higher R-squared indicating a better fit (James et al. 2013; Hastie, Tibshirani, and Friedman 2009). The formula for R-squared is:

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

where  $\bar{y}$  is the mean of the actual values.

**Additional Considerations** These evaluation metrics each have strengths and weaknesses. RMSE penalises large errors more than MAE, making it more sensitive to outliers. MAPE is useful for interpretation as it gives a percentage error, but it can be misleading when actual values are very low.

R-squared is a useful measure of fit, but it can sometimes be artificially high in overfitting scenarios. Therefore, it is essential to use a combination of these metrics to gain a well-rounded understanding of the model's performance. For this project, we are going to look for a model that produce the smallest RMSE, MAE, and MAPE, while produce the largest R-Squared, because this implies the model produce the least error and cover the most variability from the data.

### **Elements with Above PAAS Level**

## **Results**

### **Predicting Critical Elements**

Table ?? shows the model performance evaluation for predicting REE, HREE, and LREE across different scenarios. For REE prediction, the predictive model was built under five different scenarios. From the error evaluation metrics, the RMSE value is ranging from 47.57 to 66.45, MAE value is ranging from 33.85 to 38.18, and MAPE value is ranging from 26.40% to 29.44%. A model with the lowest value indicates the least error, which turn out the random forest default tuned model where all elements are included is the lowest of all. Additionally, from the variability metric, the R-Square value is ranging from 0.37 to 0.62. A model with the highest value indicates a better model in capturing variability of the data, in this case the same model also score the the highest of all. This results concludes the best model in predicting REE is the random forest default tuned model where all elements are included.

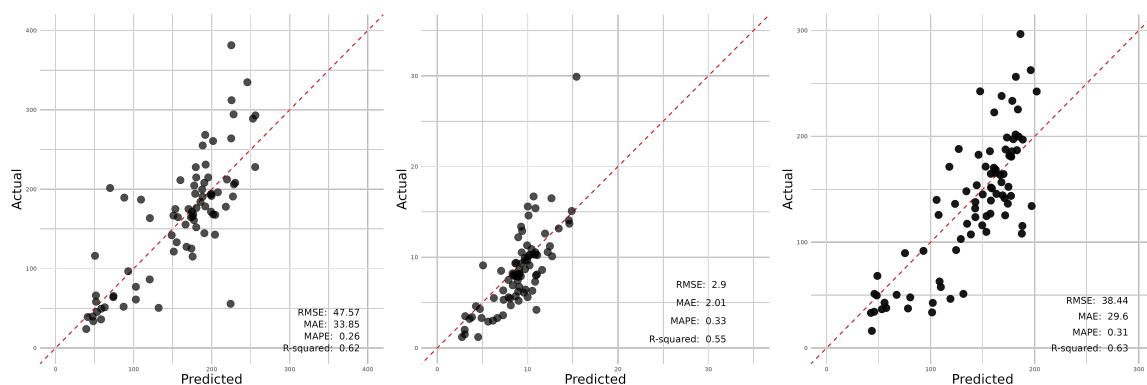
For HREE prediction, the predictive model was built under four different scenarios. A scenario where we construct the model with elements which has correlation coefficient above 0.5 is absent because there is only one element that fall under this category (Sc). It was decided not to perform this scenario for HREE for the reason that Sc correlation coefficient to HREE is only 0.57, with just moderately strong coefficient, Authors thought predicting HREE with one element will: first, oversimplified the model; second, ignore the fact that HREE is less abundant in the nature which further reduces the likelihood that a single moderately correlated element could serve as a reliable predictor. Next, From the error evaluation metrics, the RMSE value is ranging from 2.9 to 3.33, MAE value is ranging from 2.01 to 2.26, and MAPE value is ranging from 28.86% to 36.93%. The Grouped Project - Stratified has the lowest value of RMSE and MAE, but the random forest default tuned model is the lowest of MAPE. Moreover, R-Square value is ranging from 0.46 to 0.55. In this case, Grouped Project - Stratified has the highest score. In Overall, the latter model has better metrics in three evaluation metrics, making it the best model in predicting HREE.

For LREE prediction, five different scenarios was performed to get the optimal performance. From the error evaluation metrics, the RMSE value is ranging from 38.43 to 59.47, MAE value is ranging from 26.62 to 32.45, and MAPE value is ranging from 25.07% to 31.38%. The model that was constructed



using elements with correlation coefficient above 0.5 score the lowest of all, making it the model with least error rate. Furthermore, R-Square value is ranging from 0.44 to 0.63. In this case, the same model score the highest. It can be concluded that the random forest which constructed with elements that has correlation coefficient value above 0.5 is the best model in predicting LREE.

With the best model identified, figure ?? shows the predicted vs actual values of each model. The red dashed diagonal line portray the ideal scenario where the predicted values perfectly match the actual values. Points lying on this line indicate perfect prediction, while deviations from the line reflect prediction errors.



**Figure 5:** Predicted vs Actual Values of the Best Models

The plot that compares predicted vs. actual values of REE reveals moderate prediction performance by the model. With an RMSE of 47.57 and MAE of 33.85, the model demonstrates an average prediction error of roughly 47 units and an average absolute error of about 34 units, indicating that predictions are fairly close to the actual values but still have notable deviations. A MAPE of 26% suggests that, on average, the model's predictions deviate by 26% from the actual REE values, which is a slight improvement in accuracy compared to other similar models. The R-squared value of 0.62 indicates that 62% of the variance in the actual values is captured by the model, showing a reasonable fit but still leaving significant room for error. In terms of key insights, the model performs relatively well for mid-range REE values, as the points are closely aligned along the diagonal line. However, the scatter becomes more pronounced at higher REE values (beyond 300), where the model underestimates the actual values, as observed by the upward spread of points above the line. This suggests that while the model captures general patterns, it faces challenges in predicting high REE values with precision, due to outliers and high range (max - min) value of REE as describe in Table 4.

From HREE plot, the model demonstrates reasonable predictive performance with some room for improvement. The RMSE of 2.9 indicates that, on average, the model's predictions deviate by about 2.9 units from the actual HREE values. The MAE of 2.01 suggests that the average absolute error is 2.01 units, indicating that the model tends to have relatively small errors overall. The MAPE of

33% shows that the model's predictions are, on average, 33% off from the actual values, which reflects moderate accuracy. The R-squared value of 0.55 reveals that 55% of the variance in the actual HREE values is captured by the model, indicating that the model captures the general pattern but leaves a significant portion of variability unexplained. Some key insights from the plot, most data points cluster around lower values, particularly between 0 and 10 for both predicted and actual values, suggesting that the model predicts these ranges fairly well. However, for higher actual HREE values (above 20), the model tends to underestimate, as indicated by points above the diagonal line. Additionally, there are a few outliers, particularly the one near 30, where the model considerably underpredicts. Overall, the model performs well for low to mid-range HREE values but struggles with higher values, leading to underestimation and some degree of error at the upper range.

Similar to other model where they have captured the general trend with some noticeable deviations, the same case observed from LREE plot. The RMSE of 38.44 and MAE of 29.6 indicate moderate errors in the predictions. MAE is slightly lower than RMSE, meaning that large errors (outliers) are not excessively frequent, but the general prediction error is still notable. A MAPE of 31% highlights that the model's predictions are, on average, 31% away from the actual values, indicating moderate accuracy. The R-squared value of 0.63 suggests that 63% of the variance in the actual data is explained by the model, which shows a fair fit but leaves room for unexplained variance. Key insights from the plot include more accurate predictions in the lower to mid-range values of LREE, with a clustering of points near the diagonal line, while the model tends to underestimate higher actual LREE values, as shown by the increasing spread above 200. This underestimation at higher values signals some challenges in predicting extreme values accurately, which might be resulted from outliers and high range of LREE as describe in Table 4.

### Elements with Above PAAS Level

PAAS (Post-Archean Australian Average Shale) represents the precise composition of the upper crust, which is derived from the average of 23 shales in Post-Archean age, with the exclusion of aberrant materials and the correction of quadratic equation. (McLennan 2018) PAAS normalisation refers to the process of tested element concentrations normalization against the PAAS score, providing a standardized way to element abundance across different samples. The The formula can be written as:

$$\text{PAAS Normalized} = \frac{\text{Element Concentration in Sample (ppm)}}{\text{PAAS Concentration (ppm)}}$$

Figure ?? showcase the overall distribution of PAAS levels for all elements in global project, which can give a panorama viewpoint for the enrichment and universal of each element's distribution. A ratio greater than 1 indicates that the element concentration in the sample is enriched relative to PAAS standard(Represented as green triangle), implying potential for economic exploitation, vice versa.

Only a few elements, such as Rhenium(Re), Germanium(Ge), and HREE, cannot meet the detectable sample sizes because of not found or subtle presence, while the remaining elements are well-represented, with sufficient sample volumes observed in abundant concentration.

Notably, elements like Silver(Ag) and Gold(Au) tend to be elevated the greatest proportion of samples in enrichment. Only one valid samples is detected below average. However, their spatial disparity is volatile. Even within the same region might receive completely reverse result, which can pose the challenge in mass production.

In turn, the elements of Bismuth(Bi) and Lithium(Li) present promising prospects in future growth. First and foremost, The extraction from coal mine is a supplementary solution for ore reserves and driven by market's demand-supply dynamics. With the rising awareness of environmental sustainability and conservation, Bismuth, as a non-toxic alternative to lead (Pb), has gained widespread attention, especially its alloy glasses for radiation protection purposes. (Al-Buriah et al. 2021). The global shifts towards EVs has also accounted for a surge of demand in lithium batteries. Besides, Compared with other elements with above-average levels, their samples show higher levels of enrichment (more than 5 times the PAAS standard), with some specific projects detecting concentrations exceeding 10 times the PAAS standard.

Meanwhile, some instances of the element enrichment are misled by the operational misconduct. Despite scarce numbers of valid samples, Germanium is still outstanding in the graphic for it's extreme enriched samples displaying 30~40 times higher than PAAS standard. Considering it is a relatively rare in most natural environments and all these abnormal samples come from the same project (Wandoan), this seems more like a human error in statistical reporting from the laboratory. Molybdenum(Mo) and Barium(Ba) are also not the typical elements enriched in coal seams. Their apparent enrichment might be incurred by the contamination of drilling fluids.

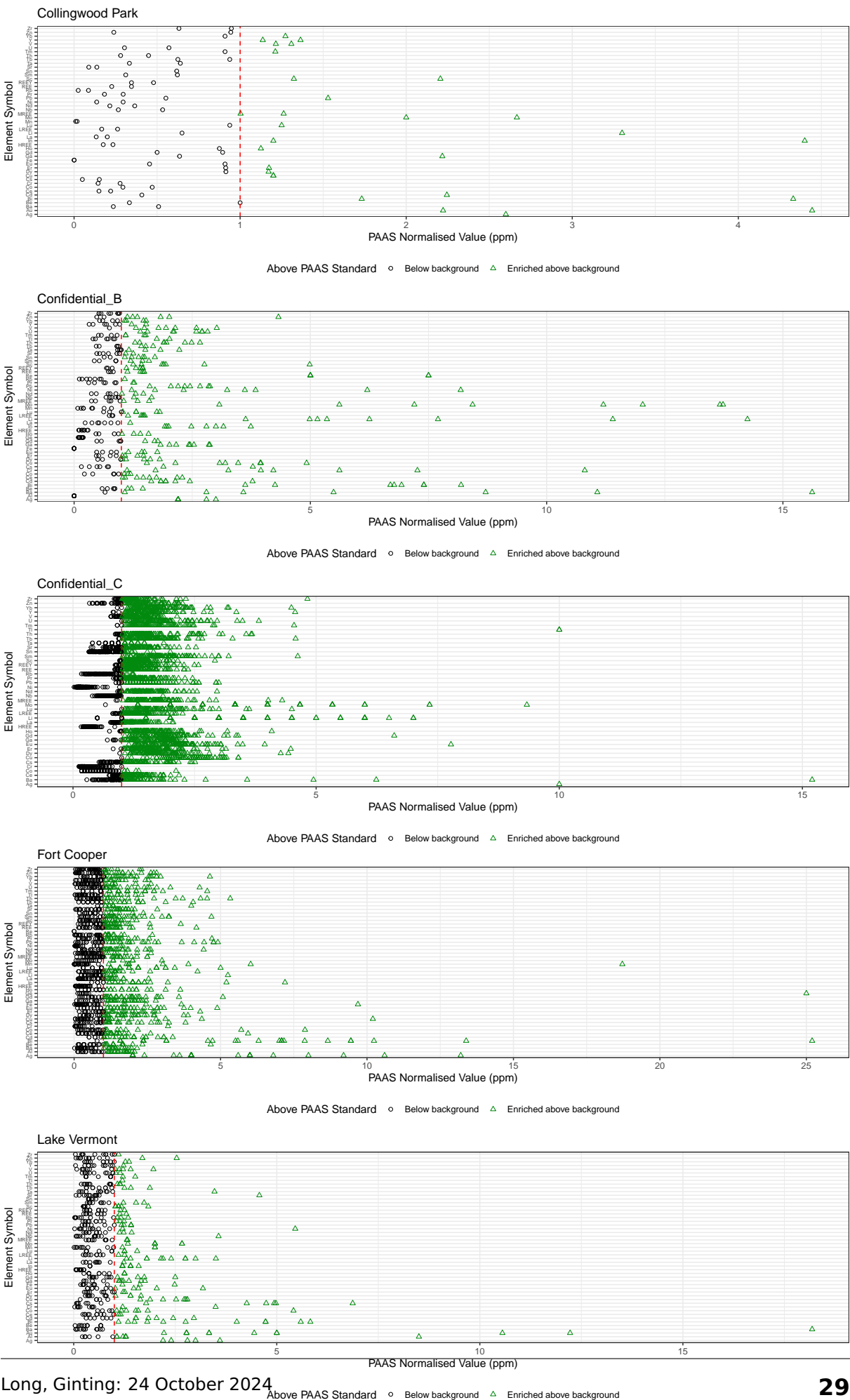
Understanding the spatial variability and geochemical behavior across projects enable the subsequent exploration focus on project regions with the highest potential yield. Figure ?? is the project breakdown for the distribution of element concentration with reference to PASS levels. Confidential\_A and Fort Cooper Coal Measure(FCCM) are the primary potential enrichment areas. All measured elements exhibit a clear and widespread enrichment pattern, consistently maintaining an enrichment range of 1 to 5 times to the PAAS standard.

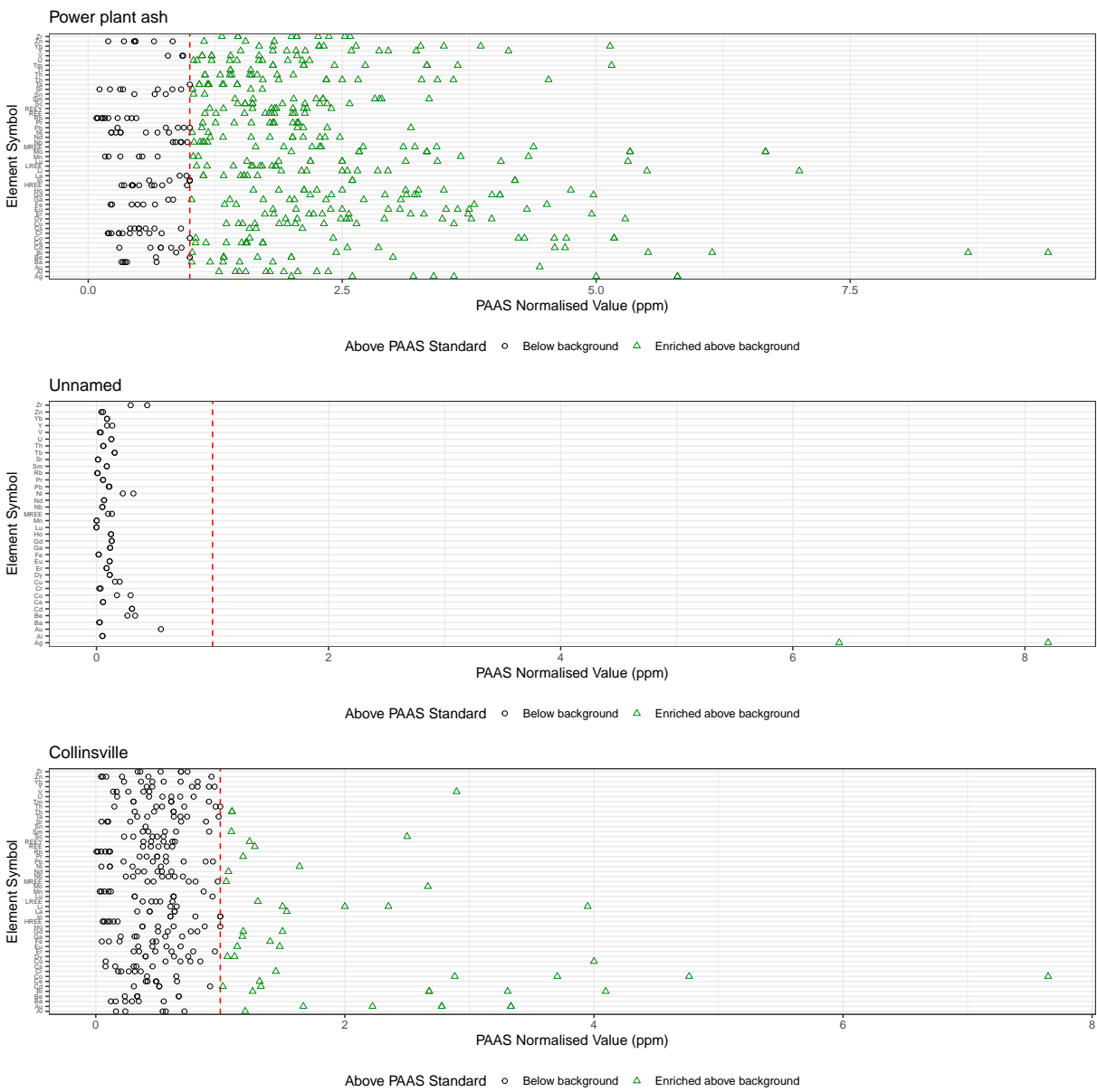
Moreover, PowerPlant Ash and Confidential\_B also have certain potential for development. Although their elements are not as pronounced as those in the two regions mentioned above, a significant number of samples still show higher enrichment levels above the PAAS standard, particularly for critical elements e.g. Aluminum(Al) and Titanium(Ti) in PowerPlant Ash, LREE and MREE series

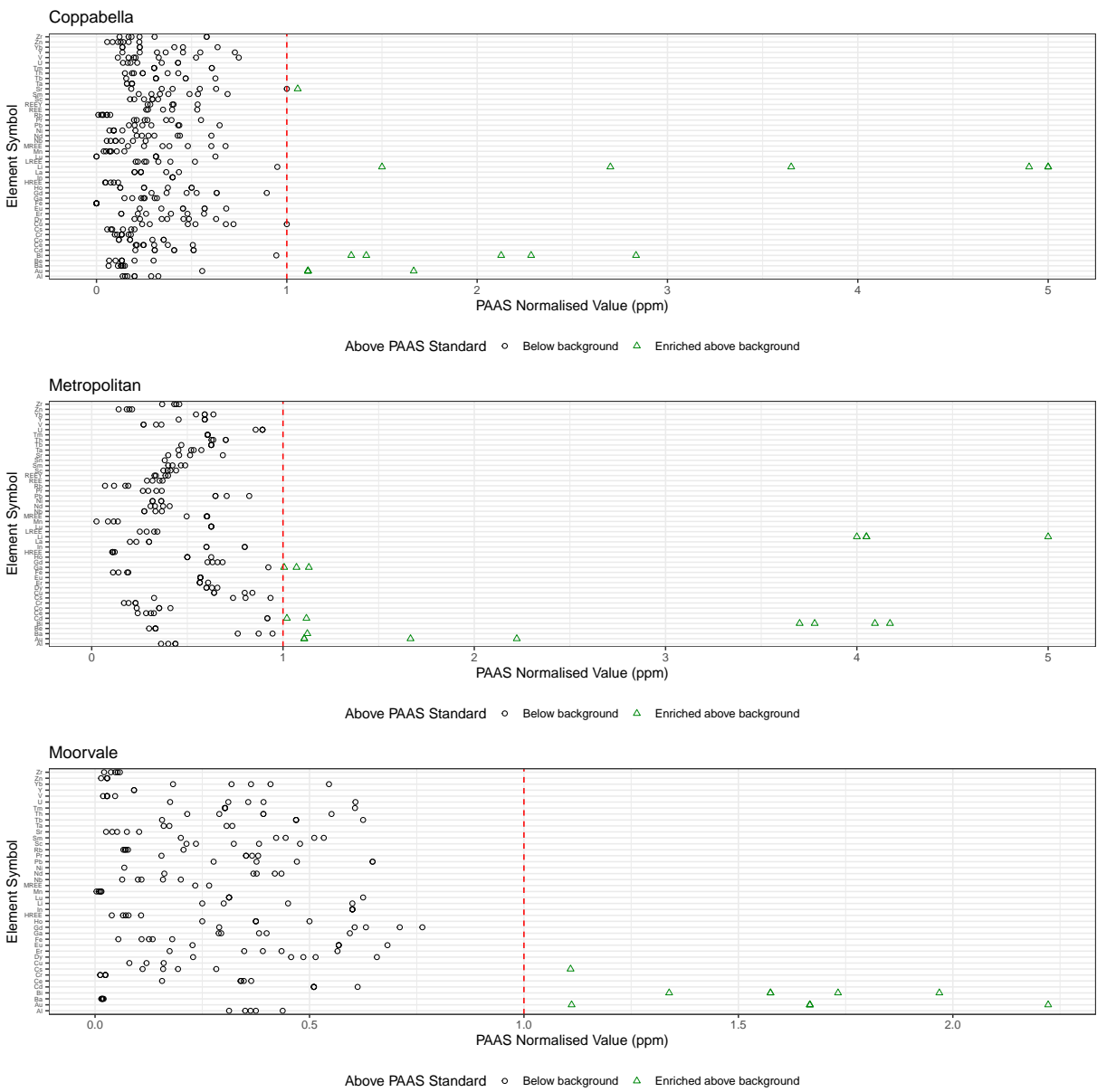
in Confidential\_B. The variance among their samples is substantial, suggesting that the enrichment might be localized.

In the end, Also some impoverished projects are found in project breakdown, including Unnamed, Coppabella, Newland and Metropolitan etc 8 projects. These projects show lower-than-average concentrations across the majority of measured elements, indicating limited economic potential under current conditions.

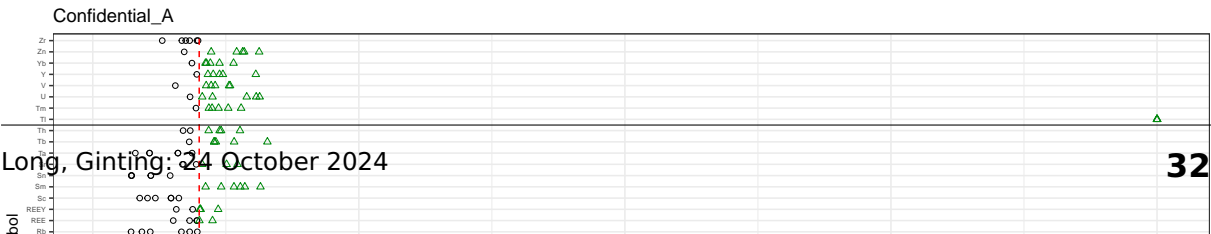
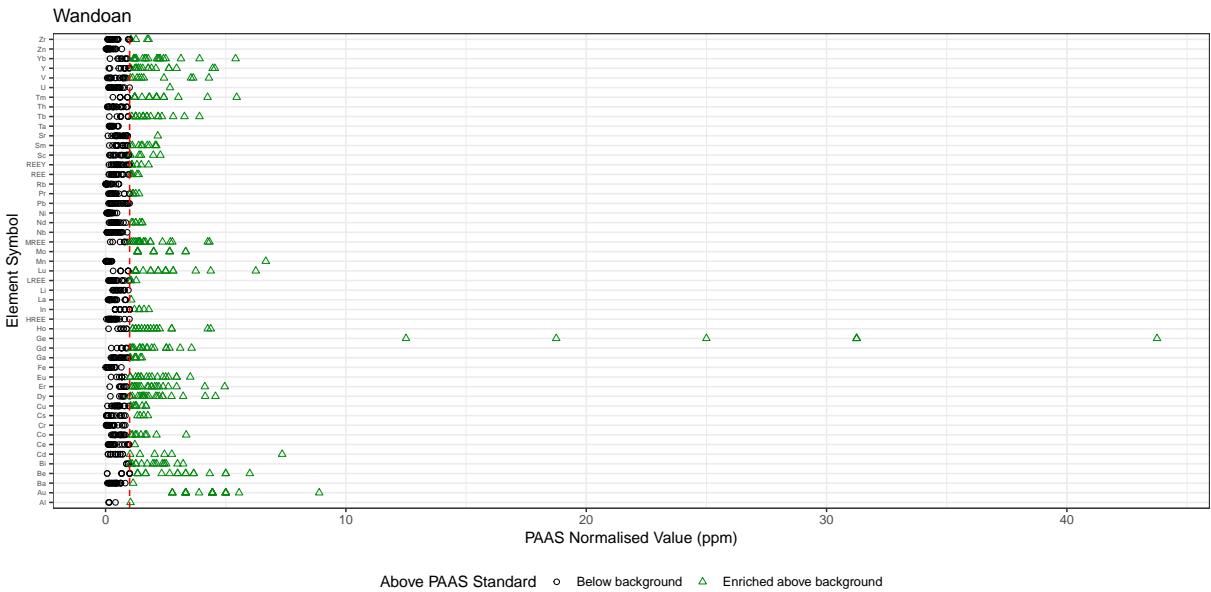
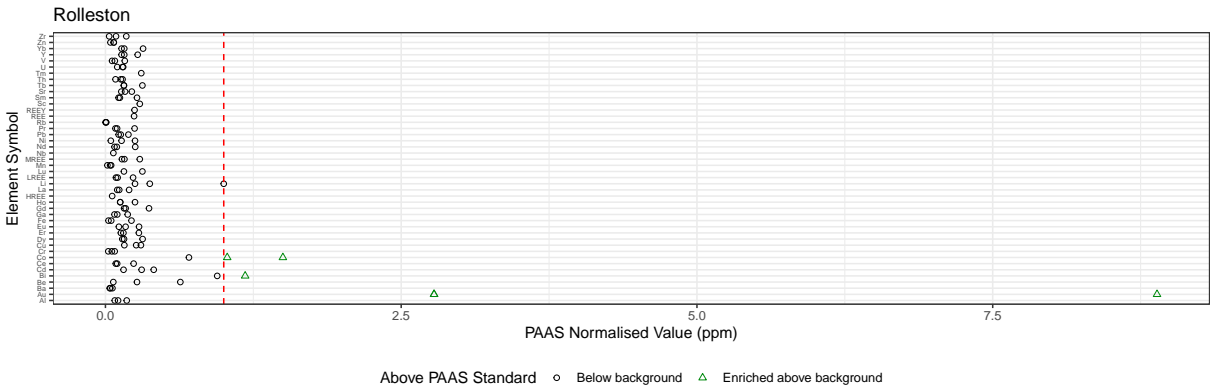
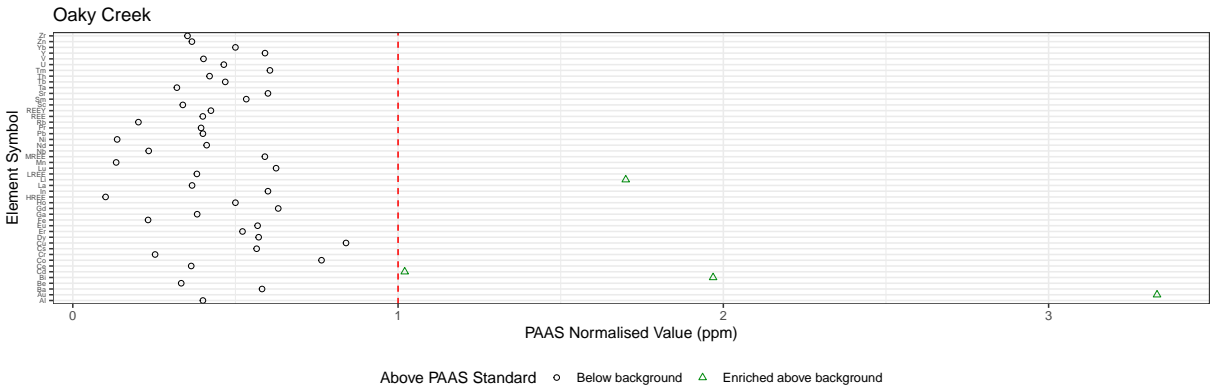
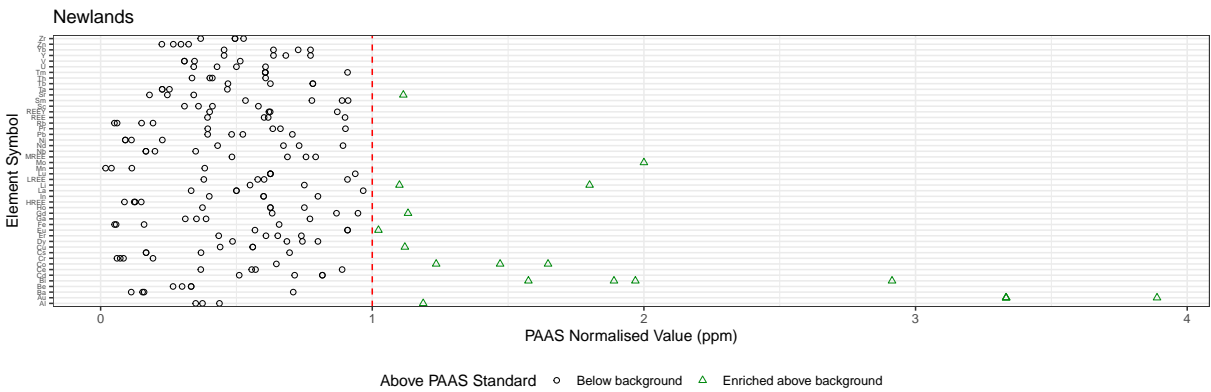
Predicting Critical Elements in Coal Mine Waste: A Machine Learning and Spatial Statistics Approach for a Low-Emission Future







Predicting Critical Elements in Coal Mine Waste: A Machine Learning and Spatial Statistics Approach for a Low-Emission Future





## Discussions

### Predicting Critical Elements

As all the best models have been thoroughly analyzed, it is worth discussing whether these models can be considered ‘good models’ for predicting REE, HREE, and LREE. Starting with the advantages, all the best models have RMSE and MAE values that are relatively close to the actual values, with some deviations. However, when these figures are compared to the standard deviation of each target variable, they fall significantly below the standard deviation. As shown in Table 4, the standard deviations of REE, HREE, and LREE are 78.9, 4.81, and 71, respectively. This indicates that the model’s prediction errors are smaller than the natural variability in the target variables, meaning the models perform better than a simple mean-based prediction. Moriasi et al. (2007) reported that if RMSE and MAE values are less than half of the standard deviation, the error can be considered low. In this case, the MAE values meet this criterion, while the RMSE values are slightly higher, which might be due to the model’s higher sensitivity to outliers. Additionally, it is recommended to use the RMSE to Standard Deviation Ratio (RSR) to assess model performance, categorising models as ‘very good’ if RSR is below 0.5, ‘good’ if RSR is between 0.50 and 0.60, ‘satisfactory’ if RSR is between 0.60 and 0.70, and ‘unsatisfactory’ if RSR is above 0.70. The RSR of all the best models falls within the ‘good’ category, which suggests that these models can be considered good based on RMSE performance.

On the downside, the R-squared values of 0.62 (REE), 0.55 (HREE), and 0.63 (LREE) indicate that the models are only moderately strong in capturing the variability of the data. Ideally, an R-squared value above 0.8 would be desirable, but an R-squared above 0.5 is still acceptable (Moriasi et al. 2007). Furthermore, the MAPE values of 26% (REE), 33% (HREE), and 31% (LREE) are relatively high, although they remain below the 50% threshold.

There are several key considerations that justify classifying these models as good models. First, none of the elements from the ME-4ACD81 test show a high correlation with the target variables, making it challenging to produce highly accurate models. Second, a cost-benefit trade-off exists. Given that the ME-4ACD81 test is almost four times cheaper than the ME-MS81 test, these models are satisfactory, especially for businesses operating on tight budgets. Third, linked to the second point, the affordability of the ME-4ACD81 test allows for more samples to be tested, resulting in more data. With additional data, the model can be retrained, potentially improving its predictive accuracy. Therefore, it is reasonable to use samples tested via the ME-4ACD81 test. Finally, it should be noted that some errors were reported by the lab when inputting sample concentration values, leading to many outliers in the data. However, due to time constraints, the authors were unable to obtain corrected data.”

### **Elements with Above PAAS Level**

With the escalating demand in the Australian market and the presence of enriched concentrations, previous study suggests Bismuth and Lithium are the top promising elements for economic extraction. However, due to the various measurement protocols applied in element mapping across projects (ME-4ACD81 and ME-MS81 analysis etc.), there is a significant amount of missing data for some elements, attributed to insufficient measurement precision and scope, instead of absence in coal mines. Re-examining these project's samples using a consistent testing method can help improve data integrity. In general, the abnormal findings on enrichment aligns with the recent cutting-edge analyses from both local and international research, including [sun2010] and (Hodgkinson and Grigorescu 2020).

The project-wise breakdown serves as an insightful guideline for prospective extraction efforts by pinpointing enrichment area. A comprehensive and in-depth investigation resource is imperative to prioritize on Confidential\_A and Fort Cooper Coal Measure (FCCM), where the distribution of enriched elements is pervasive and stable-concentrated. To optimise resource input but still yield high-quality output, A targeted study in specific enriched elements may prove to be more effective in other potential area, such as PowerPlant Ash and Confidential\_B. Conversely, some projects that only received lackluster results should rule out in further exploration.

### **Conclusion**

### **Appendix**

## Reference

- Austrade (2024). *Australian Critical Minerals Prospectus*. <https://international.austrade.gov.au/en/news-and-analysis/publications-and-reports/critical-minerals-prospectus>.
- Boehmke, B and B Greenwell (2019). *Hands-On Machine Learning with R*. Chapman & Hall/CRC The R Series. CRC Press. <https://books.google.com.au/books?id=aXC9DwAAQBAJ>.
- Breiman, L (2001). Random Forests. *Mach. Learn.* **45**(1), 5–32.
- Al-Buriahi, M, I Olarinoye, S Alomairy, I Kebaili, R Kaya, H Arslan, and BT Tonguc (2021). Dense and environment friendly bismuth barium telluroborate glasses for nuclear protection applications. *eng. Progress in nuclear energy (New series)* **137**, 103763–.
- Coyne, J and H Campbell (2023). Developing Australia’s Critical Minerals and Rare Earths: Implementing the Outcomes from the 2023 Darwin Dialogue.
- Dai, S and RB Finkelman (2018). Coal as a promising source of critical elements: Progress and future prospects. *International Journal of Coal Geology* **186**, 155–164.
- DISR (2023). *Critical Minerals Strategy 2023–2030*. <https://www.industry.gov.au/publications/critical-minerals-strategy-2023-2030>.
- Emsbo, P, C Lawley, and K Czarnota (2021). Geological surveys unite to improve critical mineral security.
- Eterigho-Ikelegbe, O, H Harrar, and S Bada (2021). *Rare earth elements from coal and coal discard – A review*.
- Fortier, S, N Nassar, G Lederer, J Brainard, J Gambogi, and E McCullough (2018). Draft critical mineral list—Summary of methodology and background information—U.S. Geological Survey technical input document in response to Secretarial Order No. 3359.
- Geoscience Australia (2023). *Critical Minerals Strategy 2023–2030*, 8–14.
- Hastie, T, R Tibshirani, and J Friedman (2009). *The elements of statistical learning: data mining, inference and prediction*. 2nd ed. Springer. <http://www-stat.stanford.edu/~tibs/ElemStatLearn/>.
- Hodgkinson, JH and M Grigorescu (2020). Strategic elements in the Fort Cooper Coal Measures: potential rare earth elements and other multi-product targets. *Australian Journal of Earth Sciences* **67** (3), 305–319.
- Hodgkinson, JH and M Grigorescu (2021). *C29030 ACARP - Mapping of Elemental Coal Content for Exploration and Mining*.
- Honaker, RQ, W Zhang, J Werner, A Noble, GH Luttrell, and RH Yoon (2019). Enhancement of a Process Flowsheet for Recovering and Concentrating Critical Materials from Bituminous Coal Sources. *Mining, Metallurgy and Exploration* **37**(1).

- Huang, Z, M Fan, and H Tiand (2018). Coal and coal byproducts: A large and developable unconventional resource for critical materials – Rare earth elements. *Journal of Rare Earths* **36** (4), 337–338.
- IEA (2021). *The Role of Critical Minerals in Clean Energy Transitions*. <https://www.iea.org/reports/the-role-of-critical-minerals-in-clean-energy-transitions>.
- IEA (2024a). *Critical Minerals Data Explorer*. <https://www.iea.org/data-and-statistics/data-tools/critical-minerals-data-explorer>.
- IEA (2024b). *Global Critical Minerals Outlook 2024*. <https://www.iea.org/reports/global-critical-minerals-outlook-2024>.
- James, G, D Witten, T Hastie, and R Tibshirani (2013). *An Introduction to Statistical Learning : with Applications in R*. eng. 1st ed. 2013. Springer Texts in Statistics, 103. New York, NY: Springer New York. Chap. 8.1 8.2.
- Kuhn, M and H Wickham (2020). *Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles*. <https://www.tidymodels.org>.
- Liu, C, T Zhao, G Wang, and D Chen (2024). A review of China’s resources of lithium in coal seams. *Frontiers in Energy Research* **12**.
- Mancheri, NA (2015). World trade in rare earths, Chinese export restrictions, and implications. *Resources Policy* **46**, 262–271.
- McLennan, SM (2018). “RARE EARTH ELEMENTS IN SEDIMENTARY ROCKS: INFLUENCE OF PROVENANCE AND SEDIMENTARY PROCESSES”. eng. In: *Geochemistry and Mineralogy of Rare Earth Elements*. Originally published 1989. Vol. 21. Germany: Walter de Gruyter GmbH, pp.169–200. <https://ebookcentral.proquest.com/lib/monash/reader.action?docID=5600203>.
- McLennan, SM (2011). Relationships between the Trace Element Composition of Sedimentary Rocks and Upper Continental Crust. *G3* **2**(4).
- Menon, S (2020). *Stratified sampling in Machine Learning*. Medium. <https://medium.com/analytics-vidhya/stratified-sampling-in-machine-learning-f5112b5b9cfe>.
- Moriasi, DN, JG Arnold, M Liew, RL Bingner, RD Harmel, and TL Veith (2007). Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations. *Transactions of the ASABE* **50**, 885–900.
- Qin, S, Y Sun, Y Li, J Wang, C Zhao, and K Gao (2015a). Coal deposits as promising alternative sources for gallium. *Earth-Science Reviews* **150**, 95–101.
- Qin, S, C Zhao, Y Li, and Y Zhang (2015b). Review of coal as a promising source of lithium. *International Journal of Oil, Gas and Coal Technology* **9** (2). doi: 10.1504/IJOGCT.2015.067490, 215–229.
- Reid, I (2018). *Non-energy uses of coal*. <https://www.sustainable-carbon.org/report/non-energy-uses-of-coal-report-ccc291/>.

- Seredin, VV and S Dai (2012). Coal deposits as potential alternative sources for lanthanides and yttrium. *International Journal of Coal Geology* **94**, 67–93.
- Sinclair, L and NM Coe (2024-04). Critical mineral strategies in Australia: Industrial upgrading without environmental or social upgrading. *Resources policy* **91**, 104860.
- Skirrow, RG, DL Huston, TP Mernagh, JP Thorne, H Dulfer, and AB Senior (2013). Critical Commodities for a High-Tech World: Australia's Potential to Supply Global Demand.
- Talan, D and Q Huang (2022). A review study of rare Earth, Cobalt, Lithium, and Manganese in Coal-based sources and process development for their recovery. *Minerals Engineering* **189**, 107897.
- Tukey, J (1977). *Exploratory Data Analysis*. Addison-Wesley series in behavioral science v. 2. Addison-Wesley Publishing Company. <https://books.google.com.au/books?id=UT9dAAAAIAAJ>.
- United States Department of Energy (2017). *Report on Rare Earth Elements from Coal and Coal Byproducts*.
- USGS (2024). *Mineral Commodity Summaries: Rare Earths*. <https://pubs.usgs.gov/periodicals/mcs2024/mcs2024-rare-earths.pdf>.
- Wang, P, WQ Chen, X Cui, J Li, W Li, C Wang, W Cai, and X Geng (2022). Critical mineral constraints in global renewable scenarios under 1.5 °C target. *Environmental Research Letters* **17**(12), 125004.
- Xu, H, C Liu, X Mi, Z Wang, J Han, G Zeng, P Liu, Q Guan, H Ji, and S Huang (2021). Extraction of Lithium from Coal Fly Ash by Low-Temperature Ammonium Fluoride Activation-Assisted Leaching. *Separation and Purification Technology* **279**, Article 119757.
- Yang, L and A Shami (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing* **415**, 295–316.
- Zhao, L, S Dai, VP Nechaev, EV Nechaeva, IT Graham, D French, and J Sun (2019). Enrichment of critical elements (Nb-Ta-Zr-Hf-REE) within coal and host rocks from the Datanhao mine, Daqingshan Coalfield, northern China. *Ore Geology Reviews* **111**.
- Zou, J, H Chen, L Zhao, X Wang, H Li, T Li, and H Wang (2023). Enrichment of critical elements (Li-Ga-Nb-Ta-REE-Y) in the coals and host rocks from the Daping Mine, Yudongnan Coalfield, SW China. *Ore Geology Reviews* **152**, 105245.