Where did all the Ore came from? A Geospatial Machine Learning Framework for predicting cumulative production flows in 2019 for Copper, Nickel and Zinc operations.

Florian Huber

Master thesis

Supervisors: Dr. Valerio Barbarossa (first), Dr. Mike Buxtion (second)

Institute of Environmental Sciences (CML)

Leiden University

Netherlands

Abstract

Rising material-intensive consumption and the global energy transition are driving unprecedented increases in the extraction of primary industrial materials, a trend that is intrinsically linked to significant environmental (e.g., deforestation, water stress) and social impacts (e.g., community displacement) that undermine human wellbeing in a spatially explicit . Yet, a paucity of site-specific extraction data hampers robust decision making. In response, our study estimates global, site-specific cumulative flows of ore, concentrate, waste rock, and tailings for copper, nickel, and zinc operations by combining a Hubbert and hyperbolic process model—which, on average, explains 60% of production variance when trained on data spanning the production peak—with an optimized machine learning pipeline. The pipeline employs support vector regression enhanced by principal component analysis and geospatial feature engineering, achieving explained variances of 88% for Cumulative Ore Production (COP) , 84% for Cumulative Tailings Production (CTP) and 77% for Cumulative Concentrate Production (CCP) respectively, along with minimal RMSE. Applied to 6808 instances where either Copper or Nickel or Zinc primary and by-production occurs for the year 2019 460 Gt of COP, 460 Gt of CTP and 26 Gt of CCP were estimated. When compared against extensive material flow databases, reveals that more then half of the global ore extraction is undocumented. Although these results must be interpreted with caution due to limited sample sizes, uncertain cumulative distribution properties, and inherent uncertainties in land use–derived features, our integrated machine learning approach—leveraging open-source geospatial data—demonstrates promising potential for approximating global cumulative flow values and inform globa spatial explicit imapct assessment.

Contents

Acronyms **5**

1 Introduction **13**

2 Methods and Data **16**

2.1 Data acquisition . . . 18

2.2 Data Preparation . . . 22

2.2.1 System definition and mine and production data calculation . . . 24

2.2.2 Land use derived feature design . . . 27

2.2.3 Environmental policy stringency . . . 28

2.2.4 Lithographical classes . . . 28

2.2.5 Primary and byproduced commodities . . . 28

2.3 Cummulative production model . . . 29

2.3.1 Hubbert model . . . 29

2.3.2 FEMP Model . . . 30

2.3.3 Model Fitting . . . 30

2.3.4 Statistical Evaluation . . . 31

2.3.5 Outlier Detection . . . 31

2.3.6 Model Comparison Metrics . . . 31

2.3.7 Results Analysis . . . 32

2.3.8 Estimating confidence intervals with using monte carlo simulation . . . 32

2.4 Machine learning model . . . 33

2.4.1 Exploratory Analysis . . . 33

2.4.2 Geography feature design . . . 34

2.4.3 Principle Component Analysis (PCA) . . . 35

2.4.4 Model selection process . . . 35

2.4.5 Model evaluation metrics . . . 36

2.4.6 Hyperparameter optimization . . . 36

2.4.7 Error Diagnostic . . . 37

2.4.8 Data augmentation . . . 37

2.5 Allocation . . . 38

3 Results **41**

3.1 Cumulative production model . . . 41

3.1.1 Global Analysis: The Hubbert model outperforms the FEMP model for most mining instances . . . 41

3.1.2 Local Analysis: Mine instances at the end of their life-cycle show higher captured variance . . . 41

3.2 Machine learning model . . . 43

3.2.1 Unary area feature shows the highest correlation across regression tasks with the target variables. . . . 43

3.2.2 Five distinct clusters of commodity extraction can be distinguished. . . . 45

3.2.3 95% of the feature space variance can be explained by 2/3 of the features. . . . 46

3.2.4 Training loop: Non-linear model show high overfitting and variance for the test sample set . . . 46

3.2.5 Hyperparameter Optimization: Explained Variance is increased slightly for nonlienar models however the overfitting gap is still substantial . . . 47

3.2.6 The standardized error increases explonentially with the observed target variables indicating an imbalanced dataset . . . 48

3.2.7 Creating synthetic train and test samples increases the test and train scores across metrics and target variables . . . 49

3.2.8 486 Gt of cumulative ore flows and 460 Gt of cumulative tailing flows were generated at global Zinc, Copper and Nickel mines to 2019 . . . 52

3.3 Allocation . . . 53

4 Discussion **56**

4.1 Halve of the global COP production in 2019 is undocumented . . . 56

4.2 Limitations and Future Outlook . . . 58

5 Conclusion **60**

Appendix **61**

A Method and data **61**

A.1 Modelling of cummulative production . . . 61

A.2 Machine Learning modelling and analysis . . . 62

A.3 Uncertainty analysis . . . 63

A.3.1 Monte Carlo estimation of confidence intervals 63
A.4 Allocation 65

B Cumulative production model **68**
B.1 Data Coverage 68
B.2 Outlier analysis 70
B.3 Global Analysis 76
B.4 Local Analysis 79

C Machine Learning Model **82**
C.1 Exploratory Analysis 82
C.2 Correlation Analysis 85
C.3 Geography feature design 88
C.4 PCA 90
C.5 Training loop 93
C.6 Hyperparameter optimization 96
C.7 Error Analysis 96
C.8 Data augmentation 100
C.9 Allocation 104

Acronyms

LCA Life Cycle Assessment 25

AMD Acid Mine Drainage 20

COV Coefficient of Variance 31

RMSE Root Mean Squarred Error 8

R2 Coefficient of Determination 36

CP Concentrate production 8

WRP Waste rock production 8

TP Tailings production 9

OP Ore production 8

CCP Cumulative Concentrate Production 1

CWRP Cummulative Waste Rock Production 25

CTP Cumulative Tailings Production 1

COP Cumulative Ore Production 1

EPS Environmental Performance Stringency Index 18

List of Figures

1 Research Framework: The research is structured in five phases that are distinguished by colours. The phase title corresponds to the method chapter that discusses the respective modelling steps and the id of the research question that is adressed. 17

2 Simplified Product System of the allocated flows for the processes Mining, Concentration, and Refining. Tme: Total material extracted, Conc: Concentrate of type j. All flows are allocated to commodity c. Flows have the unit mass per year. 26

3 Examplatory mines for the introduced typology: including peak (1), pre-peak (2) and post peak (3). Row 1 shows the cummulative production with 95% confidence interval. Row 1 the Production with 95% confidence interval. The text includes the fitting results for the metrics R2 and NRMSE for comparison purposes. The black scatter indicate the observed production. The colored lines the simulation fit for the two models. 42

4 Results of the Hyperparameter optimization using Random Search and crossvalidation. The decision metric for which the model selection was optimized was the R2. The confidence Intervals show the standard deviation around the mean. For every model the test and the train sample are evaluated using stratisfied sampling. 48

5 Rresults of the hyperparameter optimization for the five fold crossvalidated metrics R2 and RMSE metrics for the three target variables CTP, COP and CCP on the synthetic dataset.. . 51

6 Spatial distribution of COP in base 10 logarithmic tonnes. The scatters are scaled by the equally weighted area of the instances. The minimize distortion around the poles the map uses the interrupted Goode homolosine projection. 52

7 COP allocations based on different strategies at the country level. 54

8 Per-country aggregates of COP for predicted values and validation data derived from United Nations Environment Programme (2024) Both axes are log-scaled (base 10). ISO-3 country codes label scatter points, with lines indicating parity (black), factor 10 deviation (red), and relative differences to the validation data (blue). Columns represent commodities (Copper, Zinc, Nickel), while rows distinguish allocation types (Primary and Occurrence) 57

9 Histograms of the coverage for the 986 mines in the sample for the four target variables. Coverage is defined as ratio of years potential production years from the project start to 2023. The sample only includes mines which are not closed by 2023. The vertical lines include the 25%, 50% and 75% percentile. 70

10 Heatmap of missing values for Waste rock production (WRP). A missing value is a year of theoretical activity that has eather a NaN value.The periodicity is one year and the x column is in the interval form 1950 to 2023 in ascending order form left to right. White coloring represents years that are not applicable for the specific mine. 71
11 Heatmap of missing values for Ore production (OP). A missing value is a year of theoretical activity that has eather a NaN value.The periodicity is one year and the x column is in the interval form 1950 to 2023 in ascending order form left to right. White coloring represents years that are not applicable for the specific mine. 72
12 Time series of 36 randomly selected mines for OP. Depending on window size and threshold outliers are identified based on the rolling mean method. This graph represents a window-size of 10 and a threshold of two standard deviations. 73
13 Time series of 36 randomly selected mines for OP. Depending on window size and threshold outliers are identified based on the rolling mean method. This graph represents a window-size of 12 and a threshold of two standard deviations. 74
14 Histograms of R2s vor the different Target variables colored by the specific model fitted. Note that for all target variables the hubbert model shows higher frequencies for higher R2s, however the distributions are still heavily left tailed. 77
15 Histograms of NRMSE vor the different Target variables colored by the specific model fitted. The NRMSE defined as the RMSE relative to the maximum observed flow production flow . 77
16 Histograms of Root Mean Squarred Error (RMSE) vor the different Target variables colored by the specific model fitted. 78
17 The fitted model instances were categorized if their parameter set was statistical significant for the model fit for the FEMP model (F), Hubbert model or both (H) and not significant (N) 78
18 Time series of 36 randomly selected mines for OP. The graph shows the observed data (black), the simulated model fit, and the confidence intervals derived by the Monte Carlo method (FEMP: orange, Hubbert: purple). For visualization purposes, the upper confidence interval was replaced by eight times the confidence interval if the upper bound exceeded this threshold. 79
19 Time series of 36 randomly selected mines for Concentrate production (CP). The graph shows the observed data (black), the simulated model fit, and the confidence intervals derived by the Monte Carlo method (FEMP: orange, Hubbert: purple). For visualization purposes, the upper confidence interval was replaced by eight times the confidence interval if the upper bound exceeded this threshold. 80

20 Time series of 36 randomly selected mines for Tailings production (TP). The graph shows the observed data (black), the simulated model fit, and the confidence intervals derived by the Monte Carlo method (FEMP: orange, Hubbert: purple). For visualization purposes, the upper confidence interval was replaced by eight times the confidence interval if the upper bound exceeded this threshold. 81
21 Numerical features for the regression task to predict COP. Sample size = 321 82
22 Log transformed numerical features for the regression task to predict COP. Sample size = 321 83
23 Categorical features for the regression task to predict COP. 84
24 Correlation Heatmap for the Features and the target variables for the COP regression task. . 85
25 Correlation Heatmap for the Features and the target variables for the CTP regression task. . 86
26 Correlation Heatmap for the Features and the target variables for the CCP regression task. . 87
27 Silhouette score dependent on the cluster numbers for the four cumulative target variables. For COP and CCP $n = 6$ shows the highest score wherease for CTP $n = 5$ shows the highest result. High Silhouette scores indicate high average sameness of instances to the mapped cluster. 88
28 The Shillouette optimized clustering of instances for COP with a $n = 6$ clusters and the maximum similarity of an instance generated for the five clusters. 89
29 Distance of instance two the optimal cluster using a KMeans ($n = 10$). The distance with the maximum slope is for $\epsilon^* = 193779248$ and indicates the highest change in the distribution in the distance distribution. 89
30 The graph shows the cumulative explained variance against the ith component of the PCA for COP feature space. The dotted lines indicate the number of features that explain 85% (d= 32, purple) and 95% (d=43, green) of the variance of the dataset. 90
31 The graph shows the cumulative explained variance against the ith component of the PCA for CTP feature space. The dotted lines indicate the number of features that explain 85% (d= 27, purple) and 95% (d=37, green) of the variance of the dataset. 91
32 The graph shows the cumulative explained variance against the ith component of the PCA for CCP feature space. The dotted lines indicate the number of features that explain 85% (d= 28, purple) and 95% (d=38, green) of the variance of the dataset. 92
33 Training loop results for COP prediction. The figure presents four performance metrics—Mean Absolute Error (MAE), Normalized Root Mean Squared Error (NRMSE), R^2, and Root Mean Squared Error (RMSE)—evaluated across various preselected models. Different colors distinguish between training and test iterations, while boxplots summarize the outcomes from a five-fold cross-validation. 93

34 Training loop results for CTP prediction. The figure presents four performance metrics—Mean Absolute Error (MAE), Normalized Root Mean Squared Error (NRMSE), R^2, and Root Mean Squared Error (RMSE)—evaluated across various preselected models. Different colors distinguish between training and test iterations, while boxplots summarize the outcomes from a five-fold cross-validation. 94

35 Training loop results for CCP prediction. The figure presents four performance metrics—Mean Absolute Error (MAE), Normalized Root Mean Squared Error (NRMSE), R^2, and Root Mean Squared Error (RMSE)—evaluated across various preselected models. Different colors distinguish between training and test iterations, while boxplots summarize the outcomes from a five-fold cross-validation. 95

36 This figure shows the standardized errors vs the Logarithmic transformed CCP. The horizontal lines indicate the third standarddeviation from the mean. The standardized errors follow an exponential trend. 97

37 This figure shows the standardized errors vs the Logarithmic transformed CTP. The horizontal lines indicate the third standarddeviation from the mean. The standardized errors follow an exponential trend. 98

38 This figure shows the standardized errors vs the Logarithmic transformed **TCP!** (**TCP!**). The horizontal lines indicate the third standarddeviation from the mean. The standardized errors follow an exponential trend. 99

39 This figure shows the distribution of the target variable COP, CCP and CTP before and after transformation using the SMOGN algorithm. The distributions represent the original (red) distribution of the target vector and the synthetic one in (blue). The right shift of the synthetic distribution shows that the sample includes now more samples in this region. The two verticle dotted lines show the 75% (left) and the 90% (right) quantile. 100

40 Training loop results for COP prediction on the synthetic dataset. The figure presents two performance metrics R2, and Root Mean Squared Error (RMSE)—evaluated across various preselected models with five fold crossvalidation. Different colors distinguish between training and test iterations, while boxplots summarize the outcomes from a five-fold cross-validation. . 100

41 Training loop results for CTP prediction on the synthetic dataset. The figure presents two performance metrics R2, and Root Mean Squared Error (RMSE)—evaluated across various preselected models with five fold crossvalidatio. Different colors distinguish between training and test iterations, while boxplots summarize the outcomes from a five-fold cross-validation. . 101

42 Training loop results for CCP prediction on the synthetic dataset. The figure presents two performance metrics R2, and Root Mean Squared Error (RMSE)—evaluated across various preselected models with five fold crossvalidatio. Different colors distinguish between training and test iterations, while boxplots summarize the outcomes from a five-fold cross-validation. . 101
43 Consistency check for the prediction of COP and CTP. Instances that show a higher **CTP** were colored blue to indicate mass balance inconsistency. 102
44 Consistency check for the prediction of COP and CCP. Instances that show a higher **CCP** were colored blue to indicate mass balance inconsistency. 103
45 Weighted degree distribution of selected commodities. 104
46 Boxplot illustrating the influence of allocation on commodity extraction weights. 104
47 COP estimations for global Copper primary- and by-production for Occurance based allocation. The scatters are scaled by the area of the mining instance. The map uses interrupted Interrupted Goode Homolosine to reduce distortion around the poles and represent areas accurately. 105
48 COP estimations for global Nickel primary- and by-production for Occurance based allocation. The scatters are scaled by the area of the mining instance. The map uses interrupted Interrupted Goode Homolosine to reduce distortion around the poles and represent areas accurately. 105
49 COP estimations for global Zinc primary- and by-production for Occurance based allocation. The scatters are scaled by the area of the mining instance. The map uses interrupted Interrupted Goode Homolosine to reduce distortion around the poles and represent areas accurately. 106

List of Tables

1 This table introduces the features used to predict cumulative flow vlaues, explaines what variable they approximate and discusses the causal mechanisms underlying. 19
2 This Table shows the data source and spatio temporal characteristics of the underlying Variables. Features are used in the predicitive tasts. Intermediate Variables are used to generate features (e.g. Ore prodcution) or for modelling steps (e.g. world bounds) 23
3 Overview of regression tasks, sample sizes, and commodity classifications by primary and byproduct roles 34
4 Global sum of cumulative target variable flows in the year 2019 distinguished by allocation type. The mine count expresses how many mines participated in estimating the global sum. . 55

5 Comparison between the Equations of Mineral Production Forecast (FEMP) and the Hubbert Model 61
6 Machine learning models suitable for small sample sizes with weak-to-moderate correlations and mixed data distributions. Derived from (Burkov, 2019; Géron, 2023) 62
7 Comparison of Allocation Approaches. See for economic and physical allocation (Althaus & Classen, 2005) and exergy allocation (Valero et al., 2015) 67
8 Summary statistics for the missing values of the mines. The statistic represents the percentage of years per mine that do have a non zero and non NaN value. 69
9 Summary statistics for the fitting of production data to the two models: FEMP and Hubbert. Importantly the Pn_pval variables note the proportion of mines that show a significant p-value for the parameter. For instance for TP the p-value of parameter 1 is in 80% of the cases statistical signficant with and significance level of 5%. 76
10 Tailings Production Error Analysis 96
11 Concentrate Production Error Analysis 96
12 Ore Processed Mass Error Analysis 97

1 Introduction

Humanity currently faces unprecedented environmental challenges. Several of the nine planetary boundaries, critical for maintaining Earth's system stability have already been exceeded (Richardson et al., 2023). In particular, rising greenhouse gas emissions have driven a 1.1°C increase in global temperatures relative to the 1850–1900 baseline period (IPCC, 2023), and the rate of species extinction now far surpasses safe thresholds. These trends, coupled with unequal historical and ongoing contributions to unsustainable energy use and production patterns (IPCC, 2023, p. 4), underscore the need for a profound social-ecological transition.

The energy transition is pivotal to mitigating further environmental degradation and involves a mix of demand-side solutions and the adoption of low-emission technologies. These technologies —including renewable energy generation, electrification of industrial processes, and expanded energy storage— are inherently material-intensive. Their successful deployment depends on a significant increase in the production of critical commodities, such as copper, nickel, and zinc (IEA, 2024; Nijnens et al., 2023; Watari et al., 2021). The primary extraction of these commodities is further related to other planetary boundary compartments notably biodiversity integrity (Sonter et al., 2018, 2023), deforestation (Giljum et al., 2022), and water stress in certain watersheds (Meißner, 2021). Also endangerment of local souvereignity of indigneouse communities (J. R. Owen et al., 2023; Schandl et al., 2018) is related with existing and new development of projects. As a result multiple studies warn of an impact shift of the energy transiton both production and consumption induced (Creutzig et al., 2024; Sonter et al., 2023)

While the environmental impacts of CO_2 emissions and climate change have been studied extensively at aggregated levels, the environmental and operational impacts of mining are distributed in a highly heterogeneous manner across country aggregates. This heterogeneity has been emphasized by multiple authors (Giljum et al., 2022; Macklin et al., 2023; Meißner, 2021) and points to the need for spatially explicit impact assessments. In practice, such assessments require detailed mine-specific inventory data, including information on commodity production, water and energy usage, and emissions. However, not only is a comprehensive inventory of these parameters currently lacking, but Maus and Werner (2024) also reports that nearly half of the mining instances are missing from even the most extensive proprietary databases (Standard & Poor, 2025). These data gaps are systematic in nature, exacerbated by factors such as corruption, lax regulatory oversight, and unregulated artisanal mining (Maus & Werner, 2024).

A critical component of understanding mining impacts is the assessment of cumulative production flows (e.g. ore, tailings, waste rock), as many environmental extensions—such as land occupation, impacts of tailing dams, and waste rock storage—are related to the time-integrated production rather than annual

flows (Islam & Murakami, 2021; J. Owen et al., 2023; Werner et al., 2019; Xiao et al., 2025).

The estimation of these cumulative quantities has been approached through two distinct analytical paradigms. The first approach involves instance-specific modelling aimed at increasing the net present value of a mine by identifying optimal block sequences through process-based methods (e.g. linear programming) and uncertainty-based techniques (e.g. Kriging, Monte Carlo Simulation) (Osanloo et al., 2008). In addition, machine learning methods have emerged as promising tools for ore grade and tonnage prediction as well as mine scheduling optimization, particularly in underground mining operations (Chimunhu et al., 2022; Jung & Choi, 2021). The second approach adopts a regional or global focus, integrating spatially explicit inventories gathered from either proprietary sources (with (Standard & Poor, 2025) as a notable example) or open source data collections. Open source efforts have documented over 1400 commodity extracting facilities, capturing a significant portion of production for metals such as copper (134 sites covering approximately 70% of production in 2018), nickel (41 sites, 40%), and zinc (40 sites, 40%) (Maus & Werner, 2024). Complementary data sources include national accounts from Australia (Mudd, 2010) and USGS inventories (Karl et al., 2019), as well as geospatial and machine learning based studies (Werner, Mudd, et al., 2020).

Recent studies have further demonstrated that cumulative production volumes are strongly correlated with land use impacts. For instance, Werner et al. (2019) and Xiao et al. (2025) highlight that the cumulative production of tailings, waste rock, and commodity output is intricately linked to the extent of land occupation, with Xiao et al. (2025) noting a non-linear relationship between these variables. Nevertheless, the substantial number of mining instances not captured in existing inventories (Maus & Werner, 2024) implies that data collection is both resource intensive and subject to continual revision (Jasansky et al., 2023). Consequently, there is a broad consensus on the importance of employing machine learning and open source modelling routines to overcome these challenges (Maus & Werner, 2024; Sonter et al., 2023; Werner et al., 2019).

In summary, while numerous studies have advanced our understanding of mining impacts and production planning at both local and global scales, there remains no global-scale, site-specific assessment of cumulative production flows for metal mines using openly available data. Therefore, the objective of this study is to develop a global predictive geospatial machine learning framework capable of accurately modelling spatially explicit cumulative flows of OP, CP, WRP, and TP in copper, nickel, and zinc mining operations. To address this objective, the following research questions are formulated:

1. **RQ1:** What production processing model best approximates cumulative production flows at the individual mine level, and how does its performance vary across different mine instances?

2. **RQ2:** Which combinations of machine learning algorithms and preprocessing techniques yield the highest predictive accuracy and explained variance for flow-specific regression tasks?

3. **RQ3:** How can cumulative production flows be disaggregated by commodity to reflect site-specific spatial granularity?

2 Methods and Data

The following chapter outlines the research framework see figure 1 designed to address the stated research questions. We start explaining the rationale of the covariate selection ollowed by a detailed description of the data preparation steps and data sources. Subsequent sections cover the modelling of cumulative production, machine learning methodologies, uncertainty analysis, and the allocation of cumulative target variables to individual commodities. To ensure FAIR principles (Wilkinson et al., 2016) all modelling processes leverage open-source Python packages, with complete documentation provided in an open-source GitHub Repository. Further important intermediate data and exploratory analysis steps are accessible via the Supplementary Material. In addition, trained machine learning model pipelines are accessible under Model Repository

This chapter presents the research framework 1 employed to address the outlined research questions. It begins by justifying the selection of covariates and describing the data preparation steps. The subsequent sections detail the modeling of cumulative production, the application of machine learning techniques, uncertainty analysis, and the allocation of cumulative target variables to specific commodities.

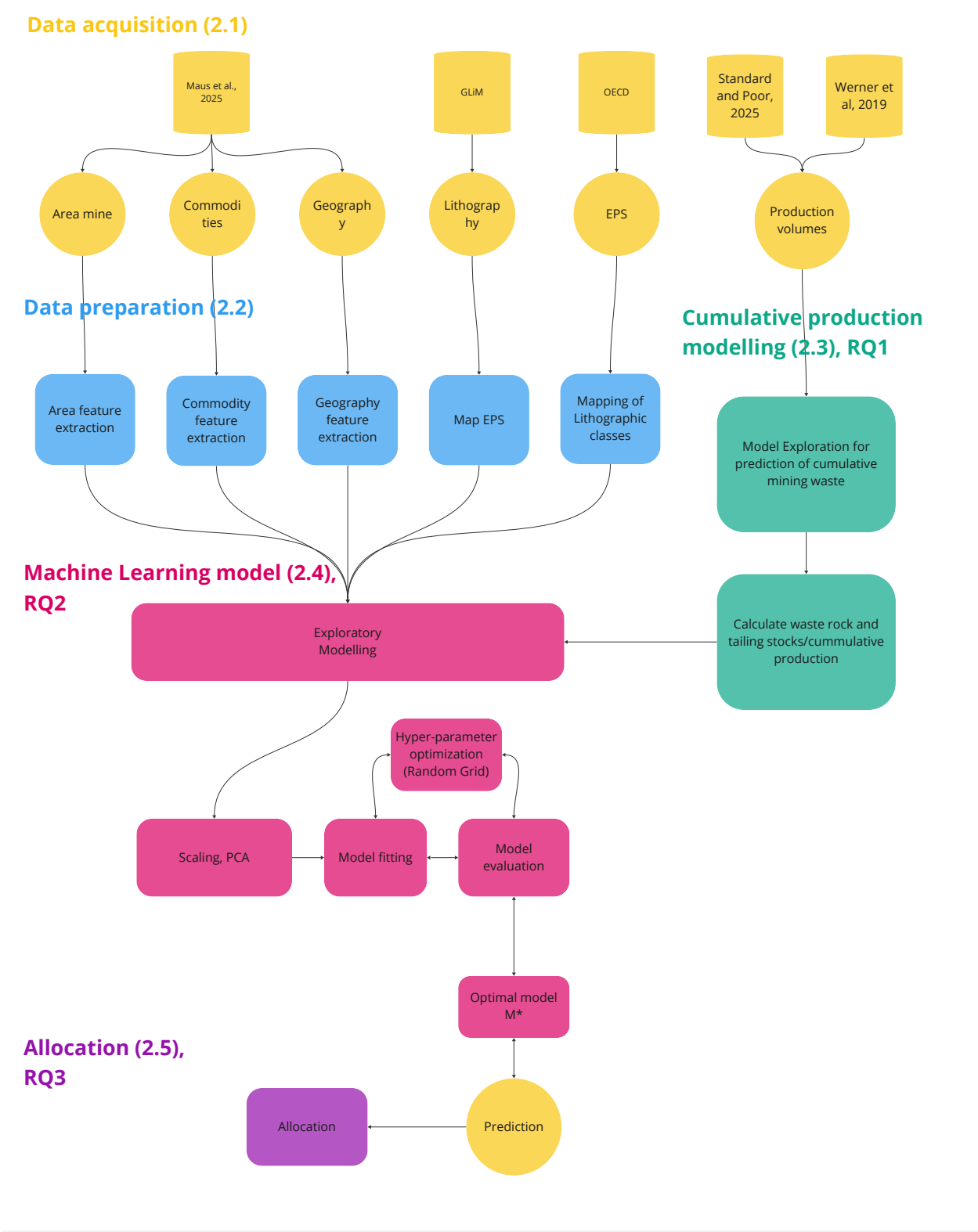

Figure 1: Research Framework: The research is structured in five phases that are distinguished by colours. The phase title corresponds to the method chapter that discusses the respective modelling steps and the id of the research question that is adressed.

2.1 Data acquisition

The most essential nature of a supervised machine learning project is that a function is derived trough an optimization procedure that optimizes the parameter of this function for the specific problem (Burkov, 2019, p.1) [1] The covariates selected for predicting the target variables are presented in tab 1. and targets are used to model the former by fitting the letter to a model. To ground the model in the current evidence and knowledge regarding the prediction of cumulative production volumes a literature review is conducted. For the most covariates proxies must be formulated to include the relation to the target variables. For instance as no global site specific ore grade (not speaking of time serieses) are available commodity classes are used to distinguish at the differences in ore-grades between commodities, because commodities show in comparison significantly different ore grades (Nassar et al., 2022, 2023). The selection of covariates is therefore directly aligned with the research objective of conducting a global high-resolution assessment of mine cumulative flows and associated reference variables. The covariates determine from the literature review are: The union area of associated polygons of a mine, compactness ration between convex hull area and union area of polygons (proxy for deposit size), Primary and Byproduct commodities (ore-grade), Lithographical classes (ore minearology), Environmental Performance Stringency Index (EPS) (Waste relted environmental regulation of mining operations) and the Union centroid (Geography). An explanation of causal relations between covariates and features is conducted in 1.

After identification of relevant variables the connection to their respective data source can be established see tab 2. It also emphasizes the global coverage of the data sources and the relatively extensive temporal span of production data analyzed in this study. By examining over three decades of mine production, this analysis enables the modeling of long-term trends, such as cumulative production functions. Additionally, the study utilizes the most comprehensive mining database to date, as detailed by Maus et al. (2022), in terms of the breadth of mining features included.

[1] This work uses feautres and covariates as synonymous. This reduces not consistency as both represent more statistical (covariates) and data driven (features) modeling schools.

Table 1: This table introduces the features used to predict cumulative flow vlaues, explaines what variable they approximate and discusses the causal mechanisms underlying.

Feature	**Proxy for**	**Mechanism**
Union area, Convex hull area and perimeter and compactness of land use	Deposit size	Cumulative production volume is the strongest predictor both for absolute area of a mine and for the area of tailings and waste rock (Werner, Mudd, et al., 2020). As correlation is causally non-directional it is assumed that land use is a strong predictor for cumulative production volume, tailing stock and tailing flow. That an increased production leads to an increase in mine land cover for e.g. open pits, tailing dams waste rock dumps is a common assumption in the geospatial modeling of mine properties (Werner et al., 2019). However applied mining technologies and process configurations have also an impact on the mine extent realized to extract a certain volume (Werner et al., 2019). Further the production volume together with the percentage of open-cut mining [2] are the most important explanatory variables for Waste rock and Tailings flows (Mudd, 2007). Open-cut mining increases especially the waste rock stocks, because of the larger overburden that has to be removed to mine the deposit. The relationallity however is complicated as larger land use (indicating at least for open-pit extraction larger deposits) can indicate lower ore grades as the extraction is because of the high quantities still economical (Ridley, 2013). So called World class deposits are contained in the highest 10% of the deposit size distribution and shape have a enormous impact on the production and therefore supply of commodities (Ridley, 2013).

Continued on next page

[2] Often mining happens as a mixture of open pit and underground mining, dependent on the deposit type

Table 1 – continued from previous page

Covariate	Proxy for	Causal mechanism
Primary and byproduct commodities extracted	Ore-grade	With decreasing ore-grade factor inputs (e.g. energy consumption) and outputs (e.g. waste rock) are increasing a fact both theoretically postulated (Ridley, 2013) and empirically validated Mudd (2007), Nassar et al. (2022), and Werner et al. (2019). For every commodity mined there is only a small number of minerals exploited reflecting the energy demand of extraction (e.g. in the case of copper sulfide minerals like Chalcopyrite have lower energy consumption and therefore economic cost of extraction then comparable silicates) (Ridley, 2013). Additionally the average concentration in the upper crust is dependent on universal abundance and the condensation processes of the earth (Ridley, 2013). As a result of this the mine waste flows and stocks will be dependent on the ore grade which is mediated by the commodity type (Nassar et al., 2022; Ridley, 2013).
Litographical classes	Ore mineralogy	Mine waste production is also either directly on indirectly (via ore-grade) influenced by the ore mineralogy and physical characteristics. A deposit constitutes of different ore bodies and is dependent on their mineralogy and physical characteristics. (Ridley, 2013). The mineralogy influences processing (e.g. milling, refining) and the removal of deleterious elements (e.g. sulfur in Ni, Cu, Zn mining that causes Acid Mine Drainage (AMD)).

Continued on next page

Table 1 – continued from previous page

Covariate	Proxy for	Causal mechanism
EPS mean and slope	Waste related environmental regulation of mining activities	The EPS aggregates various policy measures on a 0–6 scale, with higher values indicating stricter regulations. The index comprises market-based policies (e.g., taxes, trading schemes) and nonmarket-based policies (e.g., environmental standards, R&D subsidies) (Kruse et al., 2022). Stricter Environmental policies show and statistical significant impact on clean patenting activities (Andersen & Noailly, 2022, p.165-166). Particularly non market-based instruments like environmental standards show a positive influence. As multiple of the analyses patents are related to mining waste disposal or waste water recycling it is reasonable to assume that greater EPS in policy development per country will lead to less mining waste produced.

Continued on next page

Table 1 – continued from previous page

Covariate	Proxy for	Causal mechanism
Union centroid	Latitude, Longitude	Geography (latitude and longitude) plays a fundamental role based on a well-established geographic principle: processes occurring in close spatial proximity are typically more related than those further apart (Miller, 2004). In this study, this principle implies that large deposits and high-grade ore mines may often be found near each other, as they share location-specific mineralogical characteristics such as host-rock type, tectonic setting, and genetic processes of mineral enrichment (Ridley, 2013) Tang and Werner (2023) investigates this theoretical elaboration empirically and finds that mine land use shows statistically significant clustering which is, according to the authors, explained amongst others by geological phenomena (fault zones, orogenic belts) and the mine economics (e.g. depth of mineralization, proximity to infrastructure) Lastly Geography mediated via Terrain restricts the size of waste rock dumps and tailing storage dams. For example, the design of tailing dams is highly terrain dependent (Islam & Murakami, 2021).

2.2 Data Preparation

The data utilized in this study originates from various sources, necessitating an initial step to harmonize all feature inputs with the study's data model. This process includes conducting logical and consistency checks, as well as transforming data types and units as required.

Table 2: This Table shows the data source and spatio temporal characteristics of the underlying Variables. Features are used in the predicitive tasts. Intermediate Variables are used to generate features (e.g. Ore prodcution) or for modelling steps (e.g. world bounds)

Variable name	**Data type, - spatial**	**Spatial coverage**	**Temporal coverage**	**Data Source**
Features				
Union area, Convex hull area and perimeter and compactness of land use	Numerical, Vector (polygon), numerical	Global	Data was derived between 2019 and 2022	Maus et al., 2022
Primary commodities, By-produced commodities	Binary Vector (polygon)	Global	data was derived between 2019 and 2022	Maus et al., 2022
Union centroid	Numerical, Vector (point)	Global	Data was derived between 2019 and 2022	Maus et al., 2022
Litographical classes	Binary, Vector (polygon)	Global	No Temporal resolution indicated	Hartmann and Moosdorf, 2012
Environmental Performance index	Numerical, None	Global	2024	Hartmann and Moosdorf, 2012
Intermediate Variables				
Ore production, Waste rock production, Concentrate production, Tailings production	Numerical Vector (point)	Global	Data was derived between 1990 and 2022	Werner et al., 2019, S&P Capital IQ Pro database Standard & Poor, 2025
World boundaries boundaries level 0	None, Vector (polygon)	Global	Not indicated, long term validity assumption	World Food Programme, 2019
Byproduct Host mapping	Numeric, None	Global	2021	Greffe et al., 2024
Extracted Ore	Numeric, None	Global	Country dependent 1970 (min) - 2022 (max)	United Nations Environment Programme, 2024

2.2.1 System definition and mine and production data calculation

This purpose is to follow a mass balanced approach and make the derivatives of the model output reusable/comparable for/with various socio-metabolical analysis methods like ew-MFA, LCA and PIOts (Pauliuk et al., 2015). This chapter describes the defined system, governing equation and harmonization transformation steps to ensure a consistent data flow.

The production data from Werner et al. (2019) was imported and separated into area-specific and mine-specific datasets. Variable names were transposed to align with a study-specific scheme, ensuring consistent naming conventions during modeling.

The data contained inconsistent mine names, which required the implementation of a fuzzy matching approach using the Python library `fuzzywuzzy`. Matches with a similarity score above 90 (out of 100) were included, as manual inspection confirmed their accuracy. Data Preparation Werner Data, and results are provided in the `fuzzy-werner-a-prod` tab of the Meta Data Repository.

Data from Standard & Poor (2025) was extracted for the period 1991–2022, as concentrate production data begins in 1991. Only mines in the active or closed stages were included, excluding Grassroot or Brownfield exploration stages due to their lack of significant production. To account for the importance of byproduct production in mining (Greffe et al., 2024; Nassar et al., 2015), all commodity production data was extracted, enabling analysis where target commodities like copper are byproducts.

The data was restructured on a site-, site-year-, and site-year-commodity-specific basis. Variables and units were processed analogously to the Werner dataset, ensuring consistency. Further the process was established to be straightforardly scalable to multiple commodities.

To merge the Werner and S&P data, site-year matching was performed using S&P property identifiers. The commodity-specific properties of the S&P dataset were prioritized to identify mines with target commodities such as copper, nickel, and zinc. A fuzzy matching process, as described above, was conducted to map property identifiers between datasets. Production data was converted to tonnes, and years with zero production were replaced with `NaN` values and excluded. For transparency, each property-year value was tagged with its data source, prioritizing S&P due to its reliance on self-reported and verified data, in contrast to the exogenous assumptions in Werner, Mudd, et al. (2020).

An inconsistency was observed where the startup year (defined as the first year of commodity production according to S&P definitions) occurred after reported production years. To address this, the first year of reported production was substituted with the startup year.

Waste rock and tailings production values are provided in Werner, Mudd, et al. (2020) but were calculated for the S&P dataset. Waste rock production was estimated using equation 4 , and tailings production using

equation 7. The latter assumes that the mass of concentrates subtracted from processed ore equals tailings production, consistent with a mass balance approach in Life Cycle Assessment (LCA) and input-output modelling. Water consumption during flotation was excluded, aligning with Werner, Mudd, et al. (2020). A mapping table linking concentrates commodities was established based on literature and S&P data.

The system description aligns with the framework outlined by Heijungs and Suh (2002, p.11). This study focuses exclusively on Cummulative Waste Rock Production (CWRP), COP, CTP, CCP, excluding all other flows unrelated to these objectives. Although the described unit processes are aggregates of multiple operations (e.g., milling and flotation in copper concentrate production), the mass balance assumption for waste production remains valid at this aggregated level, justifying the simplification. This is obviously a great simplification for the true process scheme and spatial temporal heterogneity however this aggregation level was proven useful by a number of recent studies (Greffe et al., 2024; Nassar et al., 2023) The reference flow of a process is the ouflow of valuable material e.g., the ore production in the mining process. From a LCA perspective, processes such as *Concentration* and *Refining* in most mines are multifunctional, requiring waste flows to be allocated based on the reference flows. This study assumes that all concentrates are refined into commodities, designating the ultimate functional flow of the product system as f^{Com}. Ideally the exact concentration of valuable commodity (not necessarily the element concentration because of impurities) per reference flow (e.g. ore production) is known. As this information is impossible to derive site specific on a global scale commodity type based allocation mechanisms must be developed. This assumes the commodity is also refined at the site extraction stage an assumption frequently applied (Greffe et al., 2024; Nassar et al., 2023; Nedal T. Nassar et al., 2022) because the refining process is the most material efficient process and the distribution of efficiencies can be assumed to be narrow meaning low uncertainties. Consequently, system partitioning is applied, and waste flows per process are allocated according to the commodity type extracted (Heijungs & Suh, 2002, p.46).[3] Explicitly incorporating these absolute flows into the system description enhances clarity and ensures consistency with the allocation framework. This system description is further transformable into an input-output system as described in Heijungs and Suh (2002, p.121) and Pauliuk et al. (2015).

[3]To make this very clear fig 2 includes also the refining stage altough stocks and transaction factors are derived fro the mining and concentration stage.

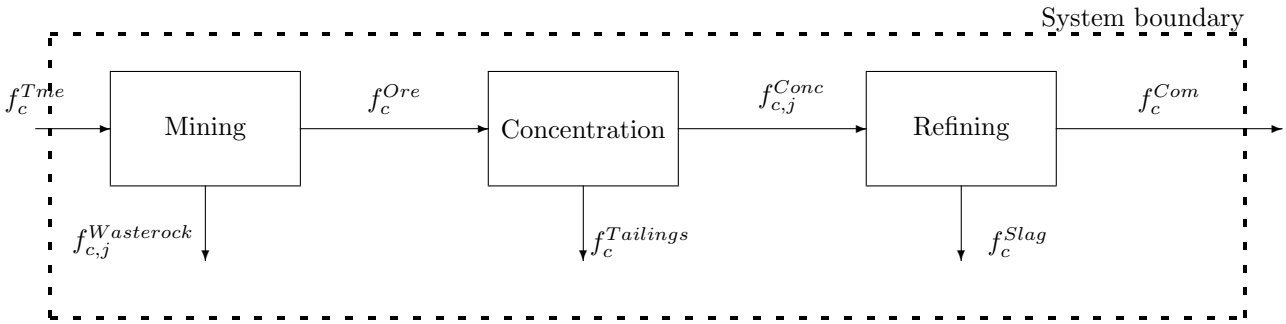

Figure 2: Simplified Product System of the allocated flows for the processes Mining, Concentration, and Refining. Tme: Total material extracted, Conc: Concentrate of type j. All flows are allocated to commodity c. Flows have the unit mass per year.

The absolute flows can be calculated as a summation of the allocated flows where:

$$f^{Waste} = \sum_{c}^{C} f_c^{Waste}, c \in C \tag{1}$$

To allocate a flow we can introduce the allocation factor a_c for a commodity at a mine. This step can be annalogues performed for the all the introduced flows of 2 Therefore allocated flows can be derived from the absolute via:

$$f_c^{COP} = a_c f^{OP} \tag{2}$$

By substitution of the absolute variable in equation we must therefore infer the identity:

$$1 = \sum_{c}^{C} a_c \tag{3}$$

To calculate the absolute waste rock flow per mine the strip ratio is introduced as follows:

$$stripratio = \frac{f^{Waste}}{f^{Ore}} \tag{4}$$

As there are multiple concentrate types produced at a single mine (e.g. coproduction of Zinc and Lead concentrates) this adds further complexity why the allocated and absolute concentrate production must be summed over the different types. J is the set of concentrates produced at a mine.

$$f_c^{Conc} = \sum_{j}^{J} f_{c,j}^{Conc}, j \in J \tag{5}$$

As the system is mass balance consistent on at least two levels (absolute flows and atomic elements)

we can derive two mass balance equations utilized for the calculation of the balancing items in the mining process:

$$f^{Tme} = f^{Ore} + f^{Wasterock} \tag{6}$$

And the concentration process:

$$f^{Tailings} = f^{Ore} - f^{Concentrate} \tag{7}$$

All the flows per mine are time dependent due to ore grade decline, economic price and cost fluctuation etc. This is shown with the following expression explanatory for the cummulative waste rock flow.

$$F^{Waste}(t) = \int_0^t f^{Waste}(t)dt \tag{8}$$

2.2.2 Land use derived feature design

Allocating land used derived polygons to a specific mine instance is a complicated taks because of the spatial heterogeneity of the features, the large (also heterogenouse) extend of mines and different process configurations (Maus et al., 2022; Werner, Bach, et al., 2020). As a result this work uses various proxy variables like Unary area and Compactness ratio to represent this heterogeneity.

The dataset compiled by Maus et al. (2022) integrates mining polygons from various sources, including Global Energy Monitor (2024), Jasansky et al. (2023), Maus et al. (2020), OpenStreetMap contributors (2024), and Werner et al. (2019). To these polygons, commodity clusters were assigned using an optimization algorithm designed to minimize unallocated polygons while maximizing the number of polygons assigned to a single cluster. Known data source mine IDs were also mapped to these clusters following the methodology referenced in Maus et al. (2022).

The allocation of S&P property ids to cluster IDs and subsequently to polygons ensures spatially optimal assignments to mines. However, overlapping polygons necessitated adjustments, including performing unions of all polygons associated with a mine to maintain consistency with the mass balance framework. Additionally, length and area measurements were standardized to kilometers and square kilometers (km^2). To allocate land use among mines linked to a single polygon, an equal allocation mechanism was implemented, which is broadly applicable across datasets with (e.g., Standard & Poor (2025)) and without (e.g., OpenStreetMap contributors (2024)) mine-level data. Alternative allocation methods, such as those based on cumulative production volume, revenue, or exergy, were deemed more data-intensive, requiring extensive

temporal data collection on production volumes and prices. To complement the equal allocation approach, unweighted area metrics were also included in the analysis.

Further, the convex hulls of polygons were calculated, providing metrics for area and perimeter. The ratio of the convex hull area to the unary union area offers insights into the spatial dispersion of polygons. Compactness, defined as the union area normalized by the convex hull area, was introduced as an indicator of this relationship, with weighted forms of these indicators also included. To derive geographic features for each mine, the centroid of the union of polygons was calculated, ensuring a consistent approach while preserving spatial similarity. This method was preferred over using the centroid of the convex hull, as it excludes voids or gaps, focusing solely on land identified as mining areas by the classification algorithm.

2.2.3 Environmental policy stringency

The Environmental Policy Stringency (EPS) indicator, measured for 40 countries from 1990 to 2022, provides a numerical range from 0 (low stringency) to 6 (high stringency) (table 1). The EPS time series for each country can be characterized by its average value over the period and its average annual change. The former reflects the overall policy stringency during the interval, while the latter indicates how much stringency evolved year by year. Both metrics are relevant for mining operations, as higher stringency often implies increased operational or development costs. Additionally, the trajectory of EPS can signal the regulatory landscape influencing long-term investment decisions. To integrate EPS data with mining operations, mines were spatially joined to country boundaries defined by the World Food Programme's global dataset (World Food Programme, 2019), which includes ISO3 country codes. Where country-specific EPS data was unavailable, continent-level means were used for imputation, based on the assumption that environmental stringency trends are similar within continents. This approach aligns with Tobler's first law of geography (Miller, 2004) and is supported by the low standard deviation of continental-level EPS statistics.

2.2.4 Lithographical classes

The GLiM dataset (Hartmann & Moosdorf, 2012) classifies geochemical and mineralogical properties of rocks into twelve lithographical classes. After harmonizing the coordinate reference systems, the dataset was spatially aligned with the union centroids (**??**), allowing each mining operation to be assigned to a lithographical class. These classes are represented as binary variables in subsequent analyses.

2.2.5 Primary and byproduced commodities

Byproduction plays a crucial role in the impact assessment of mines, as multiple commodities can be extracted from the same ore body (Greffe et al., 2024; Nassar et al., 2015). A notable example is the byproduction

of copper, gold, and silver from porphyry deposits. In this study, commodities have been reallocated and encoded as binary variables with suffixes distinguishing primary and secondary production. This distinction is vital, as the primary commodity typically contributes the largest share of the mine's production by mass (e.g., copper in Chile's Escondida mine, iron in Australia's Pilbara region). However, exceptions exist, such as gold extraction with copper byproduction, where the primary product (gold) may have a market price three orders of magnitude higher than the byproduct (copper).

Since the allocation of commodities to mines is not always unique, multiple primary commodities may be associated with a single mine operation, and this complexity was integrated into the algorithm. Additionally, the byproduction data from S&P often includes commodities that were never produced during a mine's lifetime. To address this, the byproduct-primary relationships in this study were harmonized with the byproduct-to-host ratios developed by Greffe et al. (2024), based on a study of 3,422 deposits across 141 countries. For each target commodity, only the byproduct commodities identified by Greffe et al. (2024) were included. This ensures consistency with current literature and prevents the allocation of tailings and waste rock masses to commodities that show no production during a mine's productive phase.

2.3 Cummulative production model

In order to predict cumulative production flows, it is essential to formalize the relationship between the production value of a given year and the cumulative production at that time. This formalization is rooted in process models rather than statistical models for two main reasons: first, a rich literature on depletion models of non-renewable stocks provides strong theoretical grounding (see, e.g., Section **??**); and second, process models have demonstrated high predictive performance across various commodities and applications. Moreover, this approach is particularly promising for cumulative forecasts, as statistical methods tend to be biased toward short-term production fluctuations that are effectively smoothed out in a process-based framework.This study investigates two production models, the *Hubbert Model* and the *FEMP Model*, to approximate cumulative production patterns. This chapter outlines the theoretical foundations, model fitting procedures, statistical evaluations, and data handling techniques employed in the analysis.

2.3.1 Hubbert model

The Hubbert model, originally developed to predict peak oil production, assumes finite reserves (R) and a logistic growth rate dampened by resource availability. Key parameters include the growth rate (k) and the inflection point (t_0). The general assumption is that an exponential growth cannot be sustained indefinetly for a a non-renewalbe resource (Calvo et al., 2017). Therefore the model proposes a symetric production time

relation with exponential growth-peak - and exponential degrowth. Its application to mineral extraction has been validated for commodities such as nickel, copper, and zinc, where declining ore grades over time reduce the amount of tonne extracted by totatal material extracted (Calvo et al., 2017).

Criticisms of the Hubbert model include its omission of factors such as environmental constraints, geopolitical influences, and substitution effects (Calvo et al., 2017; IEA, 2023). Nonetheless, it offers a life-cycle perspective, making it suitable for modeling site-specific production with finite cumulative outputs.

$$f(t) = \frac{R \cdot k \cdot e^{-k(t-t_0)}}{\left(1 + e^{-k(t-t_0)}\right)^2}, \tag{9}$$

where k is the growth rate, t_0 is the inflection point, and R is the ultimate recoverable resource.

The cumulative production $F(t)$, obtained by integrating $F(t)$, is given as:

$$F(t) = \frac{R}{1 + e^{-k(t-t_0)}}. \tag{10}$$

2.3.2 FEMP Model

The FEMP model, proposed by Rodríguez (2013), is based on a geometric sequence of resource depletion. It assumes a constant production-to-reserves ratio (C), leading to a steady decline in production volumes. Unlike the Hubbert model, FEMP predicts an initial production peak and focuses on short-term extraction trends. While the Hubbert model theoretically captures longer-term dynamics, FEMP provides a contrasting perspective for early production phases and is especially suitable for mining instances the dont show distinct production peaks but direct exponential declines.

$$f(t) = -R_0 \cdot ln(1 - C) * (1 - C)^t \tag{11}$$

where R_0 is the initial reserve and C is the depletion factor.

$$F(t) = R_0 \cdot (1 - (1 - C)^t \tag{12}$$

In table 5 both models are extensively compared based on various categories like representation of rate of change and the use cases.

2.3.3 Model Fitting

Both models were fitted to production data for each mine using Python's `scipy.optimize.curve_fit` function. Data preprocessing was an essential step, where only mines with at least 10 years of production data

were included to ensure reliability. To manage computational load efficiently, model fitting was parallelized. Explicitly defining parameter bounds and initial guesses enhanced the likelihood of convergence during fitting. Additionally, error calculations were performed by diagonalizing the covariance matrix obtained from the curve-fitting results to estimate parameter uncertainties. These uncertainties were subsequently used to compute t-statistics and p-values for hypothesis testing.

2.3.4 Statistical Evaluation

The significance of model parameters was tested under the null hypothesis ($H_0 : \theta = 0$) using the t-statistic and the Student-t distribution. A significance threshold ($\alpha = 0.05$) was applied to determine whether parameters were significantly different from zero.

2.3.5 Outlier Detection

Outliers were identified using a rolling mean approach with variable-specific window sizes and a threshold of two standard deviations. This method accounted for temporal dependencies, capturing trends and seasonality in the data. Flagged outliers were labeled to ensure reproducibility in subsequent analyses. Parametrization of the rolling window outlier detection algorithm is available in in the data exploration scripts.

2.3.6 Model Comparison Metrics

Goodness-of-fit was assessed using several metrics for each mine-model combination. The R^2 statistic was calculated to explain the proportion of variance captured by the model. RMSE quantified the average prediction error, providing an absolute measure of model performance. To allow comparisons across mines with varying production scales, the Coefficient of Variance (COV), a normalized form of RMSE, was also computed.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(f_{\text{obs}}(t_i) - f_{\text{pred}}(t_i)\right)^2}, \tag{13}$$

where $f_{\text{obs}}(t_i)$ and $f_{\text{pred}}(t_i)$ are the observed and predicted production values at time t_i, respectively, and n is the number of observations.

To normalize RMSE and account for scale differences, the Coefficient of Variation (COV) was computed as:

$$\text{NRMSE} = \frac{\text{RMSE}}{f_{\text{obs}}^{\max}}, \tag{14}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{f}_i - f_i)^2}{\sum_{i=1}^{n}(f_i - \bar{f})^2} \tag{15}$$

2.3.7 Results Analysis

The evaluation of model fit involved creating histograms of R^2, $RMSE$, and $NRMSE$ to visualize the performance distributions across all mines. Significance analysis was conducted to classify mines into categories based on the significance of the FEMP and Hubbert models. Mines were labeled as significant for FEMP (F), Hubbert (H), both (HF), or neither (N), based on their p-values. Temporal trends were examined by analyzing errors over time to identify potential seasonality or other patterns. Further, a typology is introduced to understand discuss suitabillity of models the models for specific mine instances.

2.3.8 Estimating confidence intervals with using monte carlo simulation

Estimating the parameters of the models comes with considerable parameter uncertainty related amongst others to data gaps in production history, short term variabillity and the phase in which the mining instance is assessed (e.g. Pre-peak production vs. Post-peak production). Monte Carlo simulation is a widely used method for quantifying uncertainties in predictive models, particularly when parameter interactions and nonlinearities are present (Harrison et al., 2010) This approach involves generating probability density functions (PDFs) for model parameters, conducting random sampling, and deriving statistical distributions of outputs (Harrison et al., 2010). For this study, Monte Carlo simulations were applied to the Hubbert model and the FEMP model to estimate confidence intervals for production rates and cumulative production. This approach is justified because we have to assume that the confidence interval around a given production estimate is not symmetrical as the a negative production is not possible. However approaches like gaussian error propagation are only applicable under a normallity assumption of errors (Taylor, 1997, p.73)

First we need to define therefore the parameter set P for every production $f(t, P)$ and the cumulative production function $F(t, P)$ as a set of random variables with the mean and standard deviation according to eq. (16) eq. (17) (hubbert) and eq. (19), eq. (20) (femp). Accordingly $f(t, P)$ and $F(t, P)$ as dependend on random set of variables represent also random variables with parameters according to eq. (22) and confidence interval for $\mu_{f(t)}$ according to eq. (23).

Second a distribution of the random variables is defined. are assumed to follow truncated normal distributions, reflecting the positivity constraints of production functions $(f(t, P), F(t, P)0)$ (Calvo et al., 2017; Rodríguez, 2013). Specifically, parameters are bounded between $a = 0$ and $b = \infty$, with the exception of the C parameter in the FEMP model, which is restricted to $[0, 1]$ This ensures physical consistency and avoids

undefined behaviour. As $P \in \Re^+$ the condition $(f(t,P), F(t,P)0)$ is satisfied.

Further if all parameters of the Hubbert and FEMP models follow a truncated normal distribution constrained to positive values (R^+), the resulting confidence intervals $CI_{f(t,p)/F(t,P)}$ will exhibit two key characteristics. First, the truncation skews the distribution by removing the left tail, causing the confidence intervals to become asymmetrical, with the lower bound closer to the mode and the upper bound extending further due to the longer positive tail. Second, the truncation ensures that all parameter values are strictly positive, which propagates through the model equations, guaranteeing that the confidence intervals for both the production rate $f(t,P)$ and cumulative production $F(t,P)$ remain strictly positive. These properties align with the physical constraints of the models, as negative values for parameters or outputs would be nonsensical.

As the confidence interval are a measure of variabillity around the mean of the random variable $f(t,P), F(t,P)$ but not for the deterministic evaluation of $f(t,P^*), F(t,P^*)$ an assumption of this work is that the developed confidence intervals around can be used as a measure of variabillity variability of $f(t,P^*)$ (see eq. (24) and **??**. This approach is justified as the parameter distributions are also derived using the optimal parameter sets.

2.4 Machine learning model

This chapter outlines the methodological framework for the machine learning modeling process. Initially, exploratory analysis is conducted to identify significant correlations among variables. A Geographical clustering approach is introduced to represent the spatial configuration of mining instances in the training sample in the feature space. Subsequently, techniques such as principal component analysis (PCA), hyperparameter optimization, and error diagnostics are introduced and justified to reduce mulitcollinearity and increase model performance. Finally, data augmentation is applied to achieve a more balanced representation of the dataset, thereby improving the generalizability of the model.

2.4.1 Exploratory Analysis

Summary statistics were calculated for the prediction tasks, including mean, median, and quartiles for numerical variables, and count and percentage of true values for binary variables. The percentage of zero or NaN values was also computed to assess the role of imputation. Numerical variables were imputed by sorting data by active years and applying forward-fill methods, assuming that mining-related variables like area typically increase with operational lifetime. Exceptions, such as EPS_mean and EPS_slope, were imputed using mean values. Binary variables were imputed with their mode.

Normality of numerical variables was assessed using the Shapiro-Wilk test, applied to both raw and log-

transformed variables. Variables were tested according to the H_0 of bing normally distributed, the alternative hypothesis is that the variable is not normally distributed (H_1). If the p-value exceeded a significance threshold ($\alpha = 0.05$) the H_0 cannot be rejected and we can assume that the variable is normal (log-normal) distributed. Normal variables are advantageous for reducing optimization time during training (Burkov, 2019, p.25), and their distribution determines whether normalization or standardization is required.

Regression Task	**Sample Size**	**Primary/Byproduct Copper**	**Nickel**	**Zinc**
CTP	203	88 / 77	24/ 24	57 / 49
CWRP	33	22 / 7	1 / 1	5 / 3
COP	320	124 / 125	41 / 31	64 / 86
CCP	231	101 / 87	35 / 31	63 / 44

Table 3: Overview of regression tasks, sample sizes, and commodity classifications by primary and byproduct roles

To analyze correlations, Pearson correlation was used for numerical-numerical relationships, chi-square tests for binary-binary relationships, and point-biserial correlation for numeric-binary relationships. These methods were applied to test the null hypothesis (H_0) of independence against the alternative (H_1) that the variables are dependent. For all tests, H_0 was rejected if $p < 0.05$. Correlation results informed about the evidence for strong associations and mulitcolinearity (Géron, 2023).

2.4.2 Geography feature design

The exploratory data analysis (see fig. 22) reveals a bimodal distribution in the latitude of mine locations, indicating spatial clustering. Prior research further supports the presence of geographical autocorrelation between mine area and commodity production (Jasansky et al., 2023; Tang & Werner, 2023). To incorporate spatial configuration into the model, various methodologies can be employed, such as global and local autocorrelation (Rey et al., 2023, pp. 157–176) or clustering-based approaches (Rey et al., 2023, pp. 251–258). This study adopts a clustering technique (Géron, 2023, pp. 82–85) that integrates KMeans clustering with a similarity measure based on an RBF kernel, capturing both categorical cluster membership and a continuous measure of cluster fit.

Determining the optimal number of clusters is crucial for the KMeans algorithm. To achieve this, the Silhouette score is computed for cluster counts ranging from 2 to 20, with the optimal number selected based on the highest score for each target variable (Géron, 2023, pp. 270–271). The Silhouette score, defined within $[-1, 1]$, quantifies cluster cohesion, where a value of 1 signifies perfect clustering, -1 indicates incorrect

classification, and 0 represents a cluster boundary.

For each instance, the geographical components, specifically the similarity to the n_i clusters for i target variables, are incorporated into the feature space to enhance the regression model's predictive capabilities.

2.4.3 Principle Component Analysis (PCA)

Dimensionality reduction techniques—including projections, manifold learning, Principal Component Analysis (PCA), and random projections—are widely used to reduce data dimensionality and address multicollinearity (Géron, 2023, pp.237–253). In PCA, the axis that preserves the maximum variance is selected because it minimizes information loss, which can also be viewed as minimizing the mean squared distance between the original data and its projection. This criterion yields the first principal component (the eigenvector corresponding to the greatest variance), while subsequent components are chosen to be orthogonal, thereby ensuring linear independence (Géron, 2023, p.244). Furthermore, the number of components is typically determined by constructing an elbow plot that shows the cumulative explained variance against the number of dimensions, with a common threshold being 95% explained variance, as illustrated in fig. 30 .Moreover, because PCA is sensitive to the scale of the input features, standardization is applied to the data; this prevents features with larger numerical ranges from disproportionately influencing the variance structure and ensures that each feature contributes equitably to the analysis.

2.4.4 Model selection process

This regression problem presents several characteristics that influence the choice of appropriate models.

First, the sample size is comparably low for copper, nickel, and zinc mines. This constraint necessitates the selection of algorithms that can perform effectively with limited data. For instance, deep neural networks are unsuitable due to their high data requirements to achieve convergence.

Second, the problem is likely to exhibit non-linear relationships. Specific feature-target combinations, such as compactness and cumulative concentrate production, demonstrate varying distributions. However, this assumption can only be validated by initially fitting a linear model.

Lastly, model interpretability [4] is an important criterion. The ability to explain the influence of features on the target variable varies across models. Linear regression models offer high interpretability, decision tree-based models provide medium interpretability, and shallow neural networks have very limited interpretability (Géron, 2023).

Table 6 provides a summary of model assessment based on these criteria.

[4] Understood here as post hoc interpretabillity

Training loop A stratified train-test split was implemented to maintain the original proportions of mines producing copper (84.6%), zinc (48%), and nickel (22.2%) as either primary or by-product outputs. To capture the production structure, a categorical feature was encoded from the primary and by-product binary indicators (e.g., 1-0-1-0-0-1), and the dataset was stratified according to the distribution of these combinations.

2.4.5 Model evaluation metrics

The performance of the models is assessed using the Root Mean Squared Error (RMSE), and the coefficient of determination (Coefficient of Determination (R2)). Detailed explanations of these metrics are provided in **??**.

2.4.6 Hyperparameter optimization

Hyperparameters are parameters of the learning algorithm that define its optimization process and determine the best parameter instances for a model (Géron, 2023, p.32). An example is the learning rate in gradient descent optimization. Since no theoretical framework exists to universally determine optimal hyperparameter settings, heuristic methods are necessary for selection (Géron, 2023, p.33).

Common heuristic approaches for hyperparameter tuning include manual adjustment, grid search, random search, and Bayesian optimization (Géron, 2023, pp.91-95). In grid search, a predefined set of hyperparameter values is systematically evaluated for all possible combinations. In contrast, randomized search samples hyperparameters from specified distributions, making it particularly advantageous in large parameter spaces where exhaustive grid search is computationally prohibitive (Burkov, 2019, p.59). Given this advantage, this study employs randomized search.

Tree-based models, such as `RandomForestRegressor` and `SVR`, are inherently stochastic, and randomized search further introduces non-determinism. Therefore, controlling for randomness is essential to ensure reproducibility and robustness in hyperparameter selection. To address this, stratified sampling is applied, as described in the previous chapter.

The selection of hyperparameter ranges for each model is based on a literature review and model descriptions in the `scikit-learn` library. Table 6 provides an overview of hyperparameters and their influence on the learning algorithm. Furthermore, broad parameter distributions were initially defined, and optimal values were cross-examined against the initial intervals. If optimal parameters were found near the boundaries of the search space, adjustments were made accordingly.

The primary statistical criterion for hyperparameter optimization was the R2, as this metric is particularly relevant for global-scale models. The justification for this choice is twofold: (1) in global models of this

resolution, error minimization is secondary to identifying whether a model can effectively capture variance, and (2) preliminary analysis of non-optimized models with default parameters suggests that high variance correlates with low root mean squared error (RMSE). For all target variables, the regression problem appears to be relatively unaffected by the bias-variance tradeoff.

2.4.7 Error Diagnostic

For the best-fitting model obtained through hyperparameter optimization, an error analysis is performed using several regression diagnostic techniques. Key diagnostic methods include examining the distribution of residuals ($Y_{Obs} - Y_{Pred}$), identifying outliers, and assessing the dependency of residuals on the observed target values to detect heteroskedasticity (Bruce et al., 2020, p.176). Since the target variables (COP, CTP, and CCP) are known to be log-normally distributed, it is reasonable to assume that the residuals exhibit a similar distribution, an assumption verified via the Shapiro-Wilk test. This consideration is critical because log-normally distributed residuals yield reduced standard errors compared to normally distributed ones, potentially affecting outlier detection methods such as the IQR approach (Bruce et al., 2020, pp.181-187). Additionally, heteroskedasticity is formally tested with the Breusch-Pagan test to examine any statistically significant relationship between the residuals and the explanatory features, highlighting areas where feature engineering may be required. To further ensure model robustness, these tests are applied to both the training and test sets, and visual inspections are conducted by plotting standard errors against the observed target values.

2.4.8 Data augmentation

As discussed in previous chapters, the presence of overfitting in tree-based methods and the exponential increase in standard errors beyond the 90% interval indicate a fundamental issue with the distribution of training data. Specifically, the scarcity of instances in the higher quantiles of the target variable limits the model's ability to generalize effectively, constraining the achievable R2 values. To mitigate this, data augmentation techniques can be employed to synthetically enhance the representation of underrepresented regions in the data (Géron, 2023, p.500). Among existing techniques, Random Oversampling simply duplicates instances, which can lead to overfitting as the model memorizes these instances rather than learning generalizable patterns. A more sophisticated approach is the Synthetic Minority Over-sampling Technique (SMOTE), which generates synthetic samples by interpolating between a given minority instance and a randomly selected subset of its k-nearest neighbors. This process expands the decision space of the minority class and reduces overfitting by preventing exact duplication of instances (Chawla et al., 2002). However, standard SMOTE does not account for variations in the variable density, leading to potentially unrealistic

synthetic samples in sparsely populated regions. To address this limitation, this study employs the SMOGN algorithm, which extends SMOTE by incorporating Gaussian noise dependent on k-nearest neighbor distances. This ensures that instances further from the seed sample receive more conservative perturbations, thereby preserving the structure of the data while improving the model's robustness in high-variance regions (**branco˙smogn˙nodate**). By integrating SMOGN, we enhance the model's capacity to generalize across the full range of target values, reducing bias and improving predictive performance. The SMOGN instance is parametized using the minimum of the cummulative production variable and the 50% quantile as majority seeds that are undersampled wherase the 90% quantile and the 95% quantile are oversampled. This ensures that these regions are more represented in the training and test set. The sample method is choosen to be "balance" as "extreme" leads to multimodal distributions. The influence of the k for the neighbours designation tested for k= 5, 7, 10 and no significant difference was found why the default value of 5 was adopted.

2.5 Allocation

The cumulative flow predictions are per mine instance. As a mine has multiple Primary and Secondary commodities the produced Ore must be allocated to the extracted commodities. For the partitioning of any inventory data points many strategies exist relating to physical (mass, energy) or economic/social (revenue) as it is an subjective choice (Heijungs & Suh, 2002). Traditional partitioning methods applied especially to the raw material extraction stage are physical, economic, and exergy allocation that have limitations. Physical allocation assigns burdens based on extracted mass, but fails to consider the economic value of commodities, which is crucial in decision-making (Althaus & Classen, 2005). Economic allocation, based on revenue share, incorporates commodity prices but is highly sensitive to market volatility (Valero et al., 2015). Exergy allocation, proposed by Valero et al. (2015), distributes burdens based on exergy savings or costs compared to a high-entropy state, reflecting energy efficiency. However, it requires detailed exergy data, which is often unavailable.

Given the scarcity of detailed time series data on extracted tonnage and prices for individual mines, these traditional methods are impractical. Even on an aggregated level time series averages of cumulative values will be ver heuristically to derive. Therefore a network-based approach is proposed as a more globally applicable alternative. table 7 compares the four methods and their advantage and drawbacks. This method models commodity extraction as a network, with nodes representing commodities and edges indicating their co-occurrence in the same mines (Barabási, 2015). The weight of an edge between two nodes i and j is defined as the number of mines where both commodities are co-extracted:

$$w_{ij} = \sum_{m=1}^{M} x_{im} \cdot x_{jm}$$

where: M is the total number of mines, x_{im} is a binary variable indicating the extraction of commodity i in mine m (1 if extracted, 0 otherwise), x_{jm} is a binary variable indicating the extraction of commodity j in mine m (1 if extracted, 0 otherwise).

Further the self-referential loops signify the total acitivity of a commoditiy in the network (absolute values of extraction). This is adjusted for by subtracting the co-occurances form the diagonal values:

$$w_{ii}^{adjusted} = w_{ii} - \sum_{i \neq j} w_{ij}$$

Therefore the number of singular mining activities of df are established.

The degree of a node i in this weighted network, denoted as d_i, is given by the weighted sum of all edges incident to node i. The degree can be interpreted as the number of links that commodity i has to other commodities and to itself, with the importance of the co-occurrence weighted by the number of mines.

$$d_i = \sum_{j} w_{ij}$$

In this formulation, the self-loop at each node i reflects how frequently commodity i is extracted, i.e., how often the commodity appears in the network.

Using the weighted degree per node, we can determine the importance of a commodity within the primary extraction network. The weighted degree not only considers the number of linkages (co-extractions) but also the importance of each co-occurrence.[5] For example, a commodity like iron or copper, which has a high production volume, will exhibit a high weighted degree due to the large number of mines that extract these commodities. On the other hand, commodities like gold and silver, which have high economic value and are often co-extracted, will also show a high weighted degree despite potentially lower production volumes.

Thus, the weighted degree potentially captures both the importance of a commodity due to large tonnage extractions and high economic value per tonne in the mining network, both important variables for evaluating feasibility of a mining project.

For a single mine, the sum of the weighted degrees of all commodities extracted in that mine can be interpreted as the overall importance of that mine within the extraction network. Let d_i denote the weighted degree of commodity i and the total set of commodities extracted at a mine J. Then, the total importance of mine m, D_m, is:

[5] Importantly also co-occurrence with itself in other words the absolute occurrence of the commodity.

$$D_m = \sum_{i}^{J} d_i$$

This can be used to normalize the individual weighted degrees of commodities within the mine, allowing us to calculate a relative importance factor for each commodity extracted at that mine.

The relative importance of a commodity i in a mine m, denoted as $\alpha_{i,m}$, can be calculated by normalizing its weighted degree with respect to the total importance of the mine:

$$\alpha_{i,m} = \frac{d_i}{D_m}$$

This allocation factor $\alpha_{i,m}$ indicates the relative importance of commodity i in mine m compared to the overall importance of the mine in the extraction network. It is a crucial metric for understanding the role of individual commodities in a mine's extraction profile, particularly when considering byproducts and their economic relevance. We can show by substitution that the identity eq. (3) is valid.

A second allocation approach emphasizes the distribution of cumulative waste or concentrate flows among the primary extracted commodities in a mine, under the assumption that the primary commodity—determined based on revenue considerations—should bear the majority of the environmental burden. Notably, 15.43% of the 23,872 mines analyzed have more than one primary commodity due to the optimization procedure outlined in section 2.2.5. A potential refinement of this method involves clustering mines based on primary-byproduct relationships, thereby eliminating improbable associations. However, as this study does not focus on developing such an allocation mechanism, the cumulative value was equally distributed among the commodities for these mines.

3 Results

3.1 Cumulative production model

This chapter presents the results of the model fitting, the statistical assessment of the results and aims to answer research question item 1 and explains which production function is suitable to model cumulative production flows depending on different categories (e.g. COP). It is separated in Global Analysis to make a general statement about the model suitabillity for the sample of mines and a Local Analysis chapter that introduces an explanatory framework to explain model and mine specific differences. Please find the results of the statistical analysis in S1.

3.1.1 Global Analysis: The Hubbert model outperforms the FEMP model for most mining instances

In a comparative analysis of the FEMP and Hubbert models across the target variables CP, OP, TP, and WRP, the Hubbert model consistently outperforms FEMP on several quantitative measures. For instance, in the CP target, the mean R2 increases from 0.41 under FEMP to 0.60 with the Hubbert model, while the mean RMSE is reduced from approximately 77,000 to 38,000. Similar trends are evident for OP, TP, and WRP, with Hubbert also showing lower standard deviations in R2 and RMSE, which implies more consistent performance.

Despite these improvements, the error distributions reveal some nuances. For example, in the OP target, the RMSE skewness is higher for Hubbert (7.15) than for FEMP (5.57), suggesting that although the majority of observations exhibit lower errors with Hubbert, a few extreme outliers inflate the tail of the distribution. Likewise, for CP, the normalized RMSE (NRMSE) skew increases from 1.14 in FEMP to 2.06 in Hubbert, further indicating the presence of a few large deviations despite overall lower errors.

Although sample sizes vary—particularly with only 50 samples for WRP—these quantitative discrepancies collectively suggest that the Hubbert model not only provides a better fit and lower error metrics but also yields more robust parameter significance, thereby offering a more reliable representation of the production data.

3.1.2 Local Analysis: Mine instances at the end of their life-cycle show higher captured variance

The analysis introduces a mine-specific clustering framework to assess depletion model performance, categorizing trajectories into three types based on data coverage relative to production peaks (fig. 3). **Type**

1 (peak-including) trajectories, exemplified by Broken Hill (Zn, Australia), Maria (Cu, Mexico), and Sleeping Giant (Au, Canada), exhibit well-constrained Hubbert model fits with narrow confidence bands, as symmetric pre- and post-peak data stabilize parameters L, k, and t_0. However, the FEMP model often reduces to $P(t) \approx R_0 C$ for small C as $((1-C)^t \approx 1$, producing constant approximations (e.g., Fast Boulder (Pd, USA), Cannington (Ag, Australia)) due to structural rigidity and $R_0 - C$ degeneracy. As a result, Monte Carlo simulations yield narrow confidence intervals for $P(t)$, because the sensitivity of the production estimate to independent variations in R_0 and C is greatly diminished as both parameters are dependent on each other for small C ($C \approx \frac{1}{R0}$). Examplatory time series can be found for TP (fig. 20), OP (fig. 18 and CP (fig. 19)

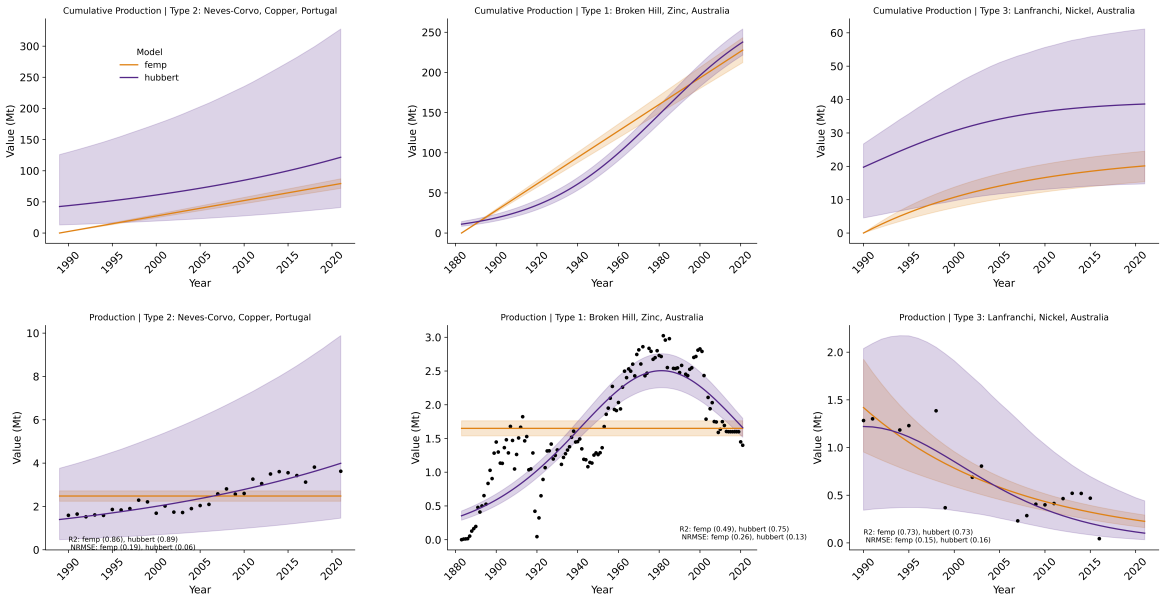

Figure 3: Examplatory mines for the introduced typology: including peak (1), pre-peak (2) and post peak (3). Row 1 shows the cummulative production with 95% confidence interval. Row 1 the Production with 95% confidence interval. The text includes the fitting results for the metrics R2 and NRMSE for comparison purposes. The black scatter indicate the observed production. The colored lines the simulation fit for the two models.

Type 2 (pre-peak) trajectories, such as Neves-Corvo (Cu, Portugal), Garpenberg (Zn, Sweden), and Kevitsa (Ni, Finland), yield unstable Hubbert fits with 95% confidence intervals spanning orders of magnitude, as missing decline-phase data under-constrain t_0 and k, amplifying parametric uncertainty. Conversely, FEMP's exclusion of growth dynamics renders it ill-suited for these cases. **Type 3 (post-peak)** trajectories, including Falcondo (Ni, Dominican Republic) and Manitoba Division (Ni, Canada), favor FEMP's exponential decay assumption, achieving low uncertainty. However, limited temporal coverage (e.g., San Vicente (Zn, Peru), Morro Agudo (Zn, Brazil)) risks misinterpreting late-stage declines as irreversible trends,

underscoring the need for cautious extrapolation.

In summary, our comprehensive analysis demonstrates that the Hubbert model offers a more reliable representation of mine production dynamics than the FEMP model. The Hubbert model consistently achieves higher R2 values, lower RMSE, and reduced coefficient of variation across the production targets (CP, OP, TP, and WRP), reflecting its superior fit and robustness. Furthermore, its statistically significant parameter estimates—evident in well-defined peak-including trajectories such as those observed at Broken Hill (Zn, Australia), Maria (Cu, Mexico), and Sleeping Giant (Au, Canada)—confirm its capacity to capture both the growth and decline phases of production. In contrast, the FEMP model's structural rigidity, which often reduces its output to a near-constant $P(t) \approx R_0 C$ for small C, results in deceptively narrow confidence intervals that fail to reflect underlying parameter uncertainty. Consequently, the Hubbert model is preferred for modeling mine production, as it provides a more accurate and statistically robust framework for forecasting production trends.

3.2 Machine learning model

This chapter answers item 1 and shows how selected covariates relate to the target variable, what preprocessing steps are conducted, why data augmentation is critical for these unbalanced training test datasets and lastly which machine learning pipelines performs the best for the respective regression tasks. Intermediate analysis of the fitting results and the final model output can be accessed via S2

3.2.1 Unary area feature shows the highest correlation across regression tasks with the target variables.

Normality plays a crucial role in improving algorithmic performance for certain machine learning and statistical methods. Many algorithms, including linear regression and principal component analysis, assume normally distributed data to ensure accurate parameter estimation and optimal model performance. When a feature exhibits a heavy-tailed distribution, standardization tend to compress most values into a narrow range, making it difficult for models to capture meaningful variations (Géron, 2023, p.76). Transforming such features, for instance, using a logarithmic or power transformation, can reduce skewness and make the distribution more symmetrical, which improves interpretability and model stability. Additionally, normality enhances the convergence properties of gradient-based optimization techniques, such as those used in neural networks and support vector machines. A well-behaved distribution facilitates smoother gradient updates, reducing the risk of poor convergence behavior and leading to more efficient learning. (Géron, 2023, p.76)

The analysis of variable distributions for different target variables reveals distinct patterns of log-normality.

For the target variable COP, the convex hull-related variables, including convex hull area, convex hull perimeter, and convex hull perimeter weighted, exhibit a log-normal distribution as indicated by p-values exceeding 0.05. Similarly, for CCP, cumulative production, convex hull area, and convex hull perimeter show log-normal characteristics with p-values above the significance threshold. In the case of CTP, only the weighted convex hull perimeter demonstrates log-normality with a p-value greater than 0.05. When considering a stricter significance level of 0.01, no variables maintain log-normality across all target variables, suggesting that while a log transformation may improve normality for some variables, its effectiveness is context-dependent. As a result the absolute area and perimeter related variables are log-normalized as there is indication that this improves the symmetry of the distributed data. The same is true for the cummulative production values (e.g. COP) as the probabillity for log-normal distribution is considerable bigger then .01% across the target variables.

Examining the histograms in fig. 21 fig. 22 for the regression task to predict COP reveals several key aspects that guide data transformation and outlier handling. First, log transformation substantially improves the symmetry of distributions for COP itself, as well as for the Unary area and the Convex hull area, including their weighted variants. Nonetheless, even after log transformation, an outlier investigation remains advisable for Cumprod. In contrast, the count of coallocated mines and the count of polygons associated with a mine do not benefit similarly from log scaling, as they exhibit a majority of small values (fewer than ten coallocated mines and fewer than ten polygons), punctuated by extreme outliers exceeding 75 and 200, respectively. Latitude and, to a lesser extent, Longitude display bimodal patterns, indicating geographic clustering at lower and higher latitudes, as well as a concentration of mines below the equator. To address this spatial heterogeneity, a clustering algorithm using a similarity score (e.g., geodesic distance) can group observations into regionally coherent clusters, thereby reducing within-cluster variance and facilitating outlier detection. Finally, fig. 23 demonstrates that the most prevalent primary commodities for COP prediction include Gold (100), Silver (150), and Copper (100), while byproduct commodities such as Lead (100) and Zinc (50) also occur frequently. Together, these findings underscore the importance of careful feature engineering, outlier handling, and, where relevant, clustering to enhance model stability and predictive performance. For the lithographical classes by far the most frequent observations are mt (Metamorphics), pa (Acid plutnoic rocks and pb (Basic plutionic rocks).

The correlation analysis reveals that COP (fig. 24 is strongly and significantly associated with several features. In particular, COP shows positive correlations with Unary_area (0.64), Unary_area_weighted (0.56), Byprod_Molybdenum (0.40), and Convex_hull_area (0.29), while it is negatively correlated with Byprod_Lead (-0.34), Weight (-0.27), and Byprod_Zinc. This indicates that the selected variables show moderate explainability for the target variable. Interestingly, the correlation between CCP and Unary_area is considerably

smaller (0.36) and the difference to the weigted variable is also considerably smaller (0.35). As Concentrates are way more dependent on a chain of processing efficiencies their relation with geospatial features (e.g. Unary_area) is not as closely as with features earlier in the processing chain. In favor of this argument is the high correlation between CTP and Unary_area of (0.66).

Notably, the area-derived parameters display substantial multicollinearity; for example, Convex_hull_area is highly correlated with Unary_area (0.83), and Weight is negatively correlated with Unary_area (-0.47), reflecting the tendency for larger, more dispersed mining areas to be associated with lower individual mine allocations. Moreover, Primary_Copper is strongly correlated with Byprod_Copper (0.63), Byprod_Molybdenum (0.41), and Byprod_Silver and Primary_Nickel co-occures often wiht Byprod_Cobalt (0.60). Although these primary and byproduct variables are for the same commodity mutually exclusive at the mine level (see copper example). Their positive correlations likely result from economic decisions and geospatial clustering, where mines in similar regions exhibit analogous commodity profiles.

In addition, basic plutonic rock shows notable associations with Primary_Platinum (0.54), Byprod_Nickel (0.48), and Byprod_Chromium (0.35), suggesting that lithological classes may provide useful insights into the types of extracted commodities. Finally, while EPS_mean and EPS_slope are significantly correlated (0.65), indicating that countries with higher EPS values also experience greater growth rates, their correlation with COP is weak and not significant (-0.06). These findings underscore the need to address multicollinearity and consider geospatial clustering in regression models to improve interpretability and predictive performance.

3.2.2 Five distinct clusters of commodity extraction can be distinguished.

As depicted in fig. 27, the silhouette score computed for cluster numbers 2 to 20 reaches its peak at $n^* = 6$ for CCP, and $n^* = 5$ for both COP and CTP. Prior to this peak, the increasing silhouette scores indicate that the respective cluster numbers yield sufficiently distinct and well-separated groupings of mining instances. Beyond the optimal cluster number, however, the silhouette score declines, suggesting that the clusters become overly granular, and points may align with multiple clusters. In fig. 28, the mining instances for the COP variable are shown, colored by their maximum similarity (i.e., similarity to the nearest cluster), together with cluster centers and global boundaries. Notably, six clusters emerge around the Andean region, North America, East Europe and Russia, China, South Africa, and Australia, with the highest similarity observed in close proximity to each cluster center. This spatial pattern is broadly consistent with known global mining distributions (Maus et al., 2022; Werner et al., 2019), and for copper mining in particular aligns with the findings of Jasansky et al. (2023).

3.2.3 95% of the feature space variance can be explained by 2/3 of the features.

Figure fig. 30 demonstrates that the first 32 components account for approximately 85% of the variance in the COP feature space, while 42 components capture 95% of the variance. The characteristic decrease in the slope of the elbow plot indicates that subsequent components contribute exponentially less to the explained variance. Notably, beyond a threshold of 50 components, additional features yield almost no further variance, justifying the adoption of this threshold for the dataset. This dimensionality reduction is attributable to the multicollinearity among features derived from geospatial variables and commodity binaries, which is consistent with the results of the multicollinearity analysis. The other target variable show similar results. For CCP, fig. 32 85% of the variance are explained by 28 components and for CTP, fig. 31 27 components are needed. Regarding the 95% for CCP 38 and for CTP 37 components are needed highligthing that both regression problems need less features to acheive the same variance.

3.2.4 Training loop: Non-linear model show high overfitting and variance for the test sample set

The initial model evaluation includes seven shortlisted models to predict the target variables. For each regression problem (e.g., COP, CTP, CCP), models are fitted using the default hyperparameter configuration of the respective library. Performance is assessed based on the coefficient of determination (R2) and root mean squared error (RMSE, in Mt).

For COP, fig. 33, the best models, in terms of mean test R2, are the Gradient Boost Regressor 0.28 (std 0.12) and the Random Forest Regressor 0.27 (0.16). While the Linear Regression and Ridge Regression models exhibit the highest R2 modem on the test set, their results are skewed due to high negative outliers in the training set with R2 ¡-3. Linear models appear balanced between train and test sets, whereas tree-based models demonstrate substantial overfitting, with Random Forest and Gradient Boost showing gaps of 0.42 and 0.60, respectively. This overfitting is likely due to insufficient regularization and limited training instances.

RMSE analysis further confirms overfitting, particularly for the Gradient Boost Regressor, which exhibits the lowest test RMSE of 228 Mt but a much lower train RMSE 99 Mt Additionally, standard deviations are significantly higher in the test set compared to the training set, with a coefficient of variance of 0.15 for training and 0.49 for testing.

For CTP, fig. 34, results indicate superior performance of linear models. The best models, based on test R2 mean, are Ridge Regression 0.40 (0.29) and Linear Regression 0.40 (0.30), though with substantial variance. More complex models yield lower predictive power of test R2 means smaller then 0.20. Linear

models exhibit balanced performance across train and test sets, with Ridge and Linear Regression achieving mean test R2 of 0.52, whereas tree-based models show considerably higher explained variance in training (e.g., Gradient Boost Regressor, mean R2 of 0.95 (0.01) on the train set, further highlighting overfitting.

In terms of RMSE of the test set, linear models achieve the lowest errors, with both Linear and Ridge Regression models yielding 316 (161) Mt. More complex models, such as Gradient Boosting Regressor and MLPRegressor, exhibit lower mean training errors but fail to generalize well to the test set.

In summary, both linear and tree-based models show promise for COP, whereas linear models are clearly preferable for CTP. However, substantial overfitting is observed in both cases, necessitating hyperparameter optimization in the following chapter. Overfitting is further exacerbated by the relatively small dataset size, which limits the diversity of training examples, leading to poor generalization. Additionally, feature selection may not be optimal, causing the models to learn noise instead of meaningful patterns, further amplifying the issue.

3.2.5 Hyperparameter Optimization: Explained Variance is increased slightly for nonlienar models however the overfitting gap is still substantial

This section evaluates the predictive performance of six regression models—ElasticNet, Lasso, Ridge, Support Vector Regression (SVR), Random Forest Regressor, and Gradient Boosting Regressor—across, wiht optimized hyperparameter set for three target variables: CCP, COP, CTP.

For predicting CCP, all models exhibit relatively low predictive performance. The highest training R2 value is achieved by Gradient Boosting Regressor (R2 mean train = 0.320), but it generalizes poorly, showing a significantly lower test R2 of 0.72. SVR performs slightly better in test prediction mean R2 test of 0.15 , indicating that complex models might overfit the training data.

For CTP, Ridge Regression achieves the highest mean test R2 of 0.41 among the linear models, while SVR performs best overall with R2 = 0.50. Despite moderate training performance of R2= 0.46, SVR generalizes better compared to tree-based models such as Random Forest) R2 = 0.29 and Gradient Boosting R2= 0.25. The large gap between training and test performance in Gradient Boosting R2 = 0.47 suggests overfitting.

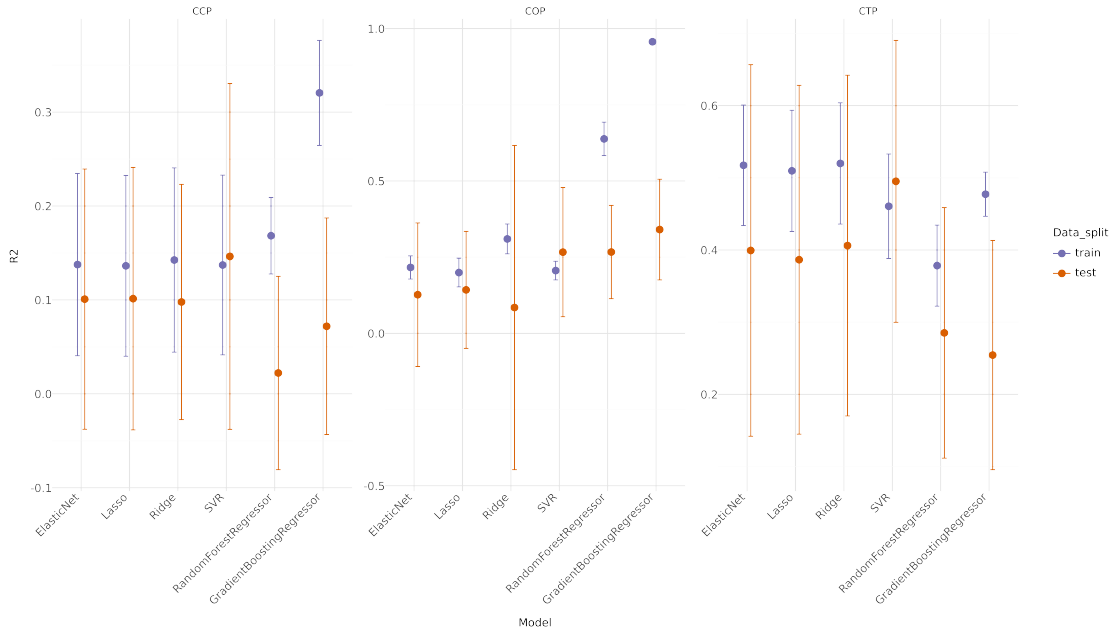

Figure 4: Results of the Hyperparameter optimization using Random Search and crossvalidation. The decision metric for which the model selection was optimized was the R2. The confidence Intervals show the standard deviation around the mean. For every model the test and the train sample are evaluated using stratisfied sampling.

For COP, the highest training mean R2 is achieved by Gradient Boosting R2 = 0.96, but its test performance drops sharply to 0.34, highlighting overfitting. Random Forest shows better test performance stability R2 = 0.26 but still struggles with generalization. Ridge Regression suffers from the highest variance between training and test performance R2 train = 0.31 and R2 test = 0.085, indicating potential sensitivity to feature scaling and data distribution.

To investigate the effect of potential outliers in the following a error analysis for the best fitted and balanced model is conducted. For CCP the best model according to mean R2 test is the SVR, for COP tough substantial overfitting is occurent the GBR is choosen. Wherease fo CTP the SVR Regressor is choosen.

3.2.6 The standardized error increases explonentially with the observed target variables indicating an imbalanced dataset

The standardized errors are analyzed plotting them against the observed values to elaborate on Heteroeskadicity. fig. 36, fig. 37 and fig. 38 illustrate that the optimal models for all three targets systematically fail to predict mines with high production values, as indicated by an exponential increase in standardized errors. For example, the exponential trend starts at values above 10^7 t for CCP and above 10^8 t for both COP and CTP; notably, the SVR model for CTP and the GBR for COP exhibit standard deviations exceed-

ing 2 for Gt mines. Additionally, while the test set for CCP and COP displays a steeper exponential error increase compared to the training data, the opposite is observed for CTP—where training errors and slope are higher, consistent with a higher R^2 in the test set, suggesting strong regularization. These observations imply an imbalance in the training data, particularly at the higher end of the production spectrum, thereby advocating for the application of oversampling algorithms to achieve a more representative distribution and improve model performance.

3.2.7 Creating synthetic train and test samples increases the test and train scores across metrics and target variables

To evaluate the impact of our synthetic data generation approach—which specifically oversamples the 60th, 70th, 80th, and 90th quantiles—we compared the distribution of the target variables before and after augmentation. Table fig. 39 The sample size for CCP increased from 233 to 433 observations, while the logarithmic mean shifted from 14.78 to 15.15 and the standard deviation from 1.55 to 1.72. Similar trends are observed for the other variables, with notable increases in counts (322 to 597 for COP and 205 to 380 for CTP). These changes indicate that oversampling in the upper quantile regions not only augments the dataset by adding synthetic instances in areas where data were previously sparse but also helps in reducing the imbalance between the minority (upper quantiles) and majority segments. Although the overall sample size increases, the effective sample size for the majority region is reduced through selective under-sampling, leading to a more balanced representation across quantiles. Consequently, the synthetic dataset preserves the central tendencies and variability of the original distribution while improving coverage in critical, underrepresented regions—thereby enhancing the robustness of subsequent predictive models.

The introduction of synthetic data by oversampling the 60th, 70th, 80th, and 90th quantiles has led to significant improvements in predictive accuracy and reductions in overfitting across all target variables. For COP (fig. 40) complex models such as the GradientBoostingRegressor saw their mean test R2 increase from 0.28 (std 0.12) to 0.74 (std 0.08), while the MLPRegressor improved dramatically from -0.07 (std 0.48) to 0.83 (std 0.04), indicating enhanced non-linear modeling and more stable predictions. The best model is the MLPR with a mean R2 of 0.83 (0.05) and an RMSE of 153 (19) Mt. However overfitting is still prevelant as the train RMSE is almost factor 2 smaller with 71 (8) Mt. Similarly, for CTP (fig. 41, the GBR's mean R2 improved from 0.09 (std 0.33) to 0.86 (std 0.12), and both RandomForestRegressor and SVR increased from approximately 0.20–0.23 to over 0.83, with corresponding RMSE reductions (e.g., GBR's RMSE decreased from 385 Mt to 242 Mt). For CTP the best performing models on the test set is the RFR with a mean R2 of 0.87 (0.04) and an RMSE of 241 (38) Mt. In the case of CCP (fig. 42), although linear models like ElasticNet exhibited an increase in RMSE from 196 Mt to 33.3 Mt and marginal or negative shifts in R2 (e.g., ElasticNet

from -0.10 to -0.14), non-linear models showed marked gains; for instance, the GradientBoostingRegressor's R2 improved from -0.09 (std 0.22) to 0.66 (std 0.18) while its RMSE decreased from 19 Mt to 17 Mt. The best scoring model in the test set was the GBR with and mean R2 of 0.66 (0.18) and RMSE 17 (3) Mt. Importantly even the best models shows considerable variance for both metrices compared to the other target variables.

In conclusion the data augmentation, meaning the expansion of the 60% - 90% quantiles lead to substantially higher explained variance an reduction in RMSE. This is especially true for the models SVR RFR, GBR and MLPR. Further the overfitting gap (especially for the test R2 could be reduced. Non Linear models show now clearly a better explainabillity of the regression task.

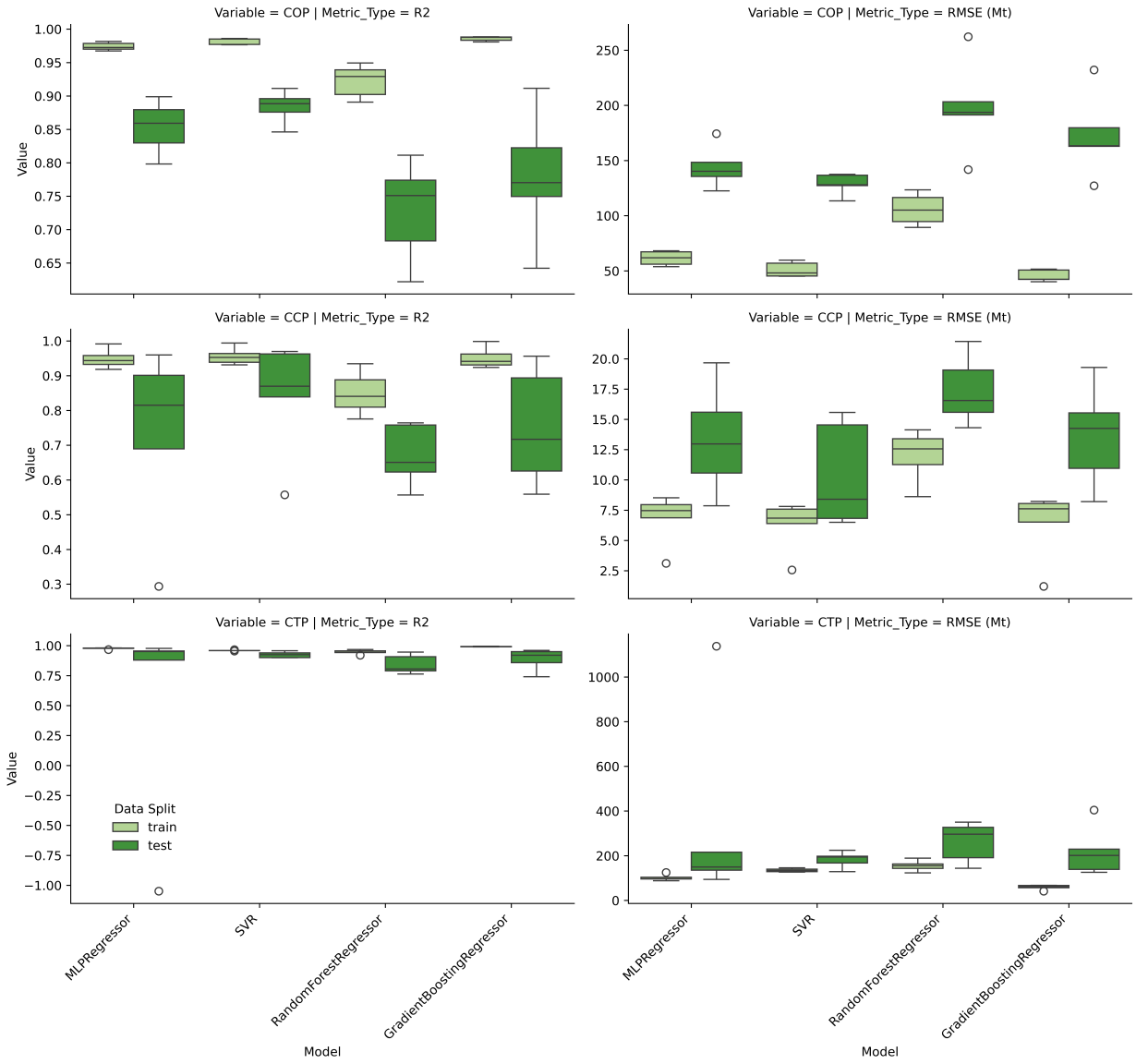

Figure 5: Rresults of the hyperparameter optimization for the five fold crossvalidated metrics R2 and RMSE metrics for the three target variables CTP, COP and CCP on the synthetic dataset..

The hyperparameterization results of the augmented data set were further evaluated, with parameters optimized in terms of RMSE as described in the previous chapter. As shown in fig. 5, the SVR regressor achieved the highest mean train R2 values, reporting 0.88 (0.02) for COP and 0.84 (0.17) for CTP, and 0.75 (0.18) for CCP which indicates a substantial improvement in explained variance compared to the results in fig. 42. Moreover, the mean RMSE values were minimized by the SVR, with 129 (10) Mt for COP and 104 (4) Mt for CTP; however, the relative RMSE error was larger for CTP than for COP. For CTP, although the SVR achieved a mean R2 of 0.92 (0.03), the MLPRegressor, despite exhibiting a higher median of 0.96, was skewed by an outlier in the test set, thereby substantially inflating its variance. Additionally, both

target variables—particularly on the test set—revealed that tree-based methods, such as the GBR for COP (see fig. 5), exhibit high variance and a pronounced sensitivity to outliers, with a notable overfitting gap; for example, the SVR for COP displayed a 7% variance gap between the training and test sets. Overall, the hyperparameter optimization reduced the overfitting gap across all models and target variables, leading to the conclusion that, with respect to RMSE minimization, the SVR regressor is the preferred model for the individual prediction tasks. In the following chapter, these optimized models are employed to generate predictions for COP, CCP, and CTP on unseen data.

3.2.8 486 Gt of cumulative ore flows and 460 Gt of cumulative tailing flows were generated at global Zinc, Copper and Nickel mines to 2019

Using a dataset of 6808 copper, nickel, and zinc producing instances, the best performing models for CTP, COP, and CCP were evaluated. On a $\log_{10}$ scale, the mean values are 6.32 (std. 0.4) for CCP, 7.44 (0.56) for **OCP!** (**OCP!**), and 7.69 (0.75) for CTP. Notably, the CTP predictions exceed the COP predictions on average, which violates the mass balance constraint (eq. (7)). Moreover, significant outliers are present; for example, the maximum COP value reaches 9.34—nearly two orders of magnitude above the mean—while CTP displays extreme outliers with a maximum of 18.03. Spatial analysis further reveals that COP values span a wide range of 5 orders of magnitude (Mt-Gt), with pronounced clusters in regions such as the Americas (notably Mexico, Chile, Peru, and Brazil), as well as in mining districts of South-East China, West-Australia, and East-Kasachstan. This heterogeneity implies that the impacts of mining are highly localized and necessitate a sub-national assessment of social and environmental effects.

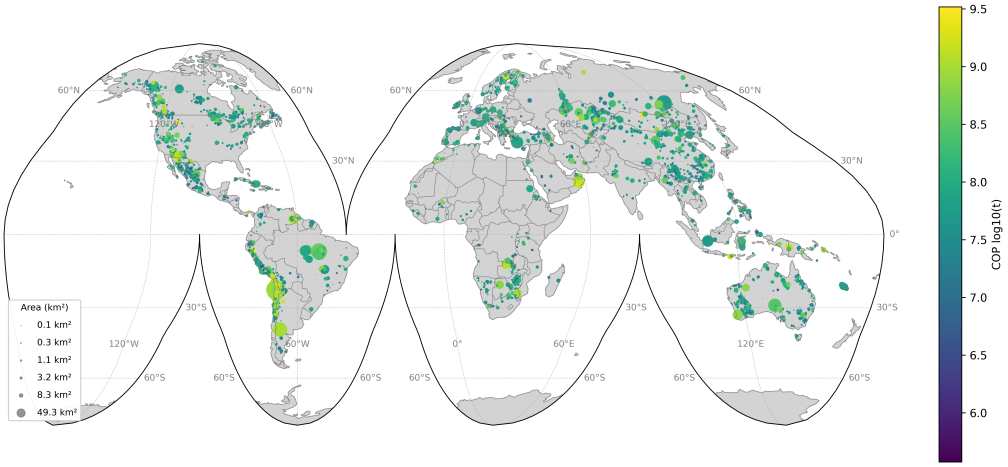

Figure 6: Spatial distribution of COP in base 10 logarithmic tonnes. The scatters are scaled by the equally weighted area of the instances. The minimize distortion around the poles the map uses the interrupted Goode homolosine projection.

Mass balance inconsistency is further evident when comparing predicted quantities. Specifically, numerous instances exhibit higher CTP (and CCP) values relative to COP, contravening the theoretical requirement that COP should be less than both CTP and CCP (eq. (7)). Out of 6732 prediction instances, only 2236 are mass balance consistent; the remaining 4487 cases show anomalously high CTP values, particularly within the lower magnitude range (10^6–10^8 tonnes), with some CTP predictions exceeding 10^{10}. This indicates a bias in the SVR Regressor towards outliers for small and high CTP values, and the potential integration of mass balance considerations into the training process is discussed in detail in the subsequent discussion section.

In response to research question 2, our findings indicate that non-linear models—especially support vector regression (SVR) and gradient boosting regressors—combined with rigorous preprocessing (e.g., log-transformation and principal component analysis) yield superior predictive accuracy for flow-specific regression tasks. Moreover, targeted data augmentation via oversampling the 60th to 90th quantiles markedly enhanced model performance by reducing overfitting, as demonstrated by improvements in test R2 (e.g., COP predictions increased from 0.28 to 0.74) and lower RMSE across multiple models. These results underscore the critical role of integrating advanced feature engineering with synthetic data generation to improve model robustness and generalizability. For this reason the CTP variable is further calculated as the not asmuch outlier dependent models of COP and CCP through eq. (7). The globally estimates CCP with 26 Gt, COP 486 Gt and **TCP!** with 460 Gt.

3.3 Allocation

The estimates of COP, CCP and CTP are allocated to the commodities extracted at a mine using two distinct strategies. The first, an occurrence-based allocation, distributes values in proportion to the commodity's relative frequency within the global mining network—thus reflecting both price and mass considerations—while the second, more conservative approach allocates cumulative flows solely based on the primary extraction, assuming that the majority of the mine's value is attributed to the primary commodity. As demonstrated in fig. 46, commodities such as Silver (byproduct weight of 15 000 versus primary weight of 4 000), Lead, Molybdenum, and Cobalt exhibit higher byproduct weights relative to their primary weights, suggesting that even low-tonnage commodities may be sufficiently valuable to be co-extracted. In contrast, commodities like Iron, Gold, and Uranium are predominantly primary extracted, driven by extensive geological availability or high economic value, whereas Nickel and Copper represent a balanced extraction of both primary and byproduct forms. Moreover, the distribution of weighted degrees per mine (see fig. 45) reveals multimodal patterns for Primary Iron, Primary Gold, and Primary Uranium, indicating that mines may either extract

these commodities in isolation or alongside significant byproducts—thus necessitating mine-specific interpretations rather than global generalizations. Byproduct-dominant commodities, such as Palladium, Cobalt, and Molybdenum, generally exhibit low median weights, although Cobalt shows a notably high variance.

Figure 7: COP allocations based on different strategies at the country level.

At the country level, fig. 7 illustrates that for Copper, similar cumulative ore productions are observed for Australia (AUS), Brazil (BRA) and China (CHN), while stark differences occur in Chile (CHL), Indonesia (IDN) and the USA. In the case of Zinc, Indonesia displays nearly an order of magnitude higher primary allocations compared to the occurrence-based strategy—likely due to the co-occurrence with high weighted commodities such as Byproduct Lead (as also reflected in fig. 45)—with analogous patterns observed for Zambia (ZMB) and South Africa (ZAF). Furthermore, Nickel extraction generally results in a higher primary allocated mass under the occurrence-based approach, except in South Africa (ZAF) and Peru (PER).

Table **??** presents a comparative overview of cumulative commodity flows based on two allocation methods: occurrence-based (Occ) and primary production-based (Prim). For example, Copper's ore processed mass is estimated at 138.15 Gt with 5,670 mines under the Occ allocation, increasing to 244.46 Gt with 3,010 mines under the Prim allocation. This indicates that the primary allocation method concentrates production in fewer, more productive sites. In a similar vein, Zinc exhibits a notable discrepancy in tailings production; the occurrence-based cumulative tailings production is 49.68 Gt compared to only 23.40 Gt derived via mass balance, suggesting that different methodologies yield substantially varying estimates. Overall, these allocation factors imply that global ore flows—amounting to 139 Gt or 245 Gt for Copper, Zinc, and Nickel, depending on the method—reflect both the resolution of individual mining instances and the inherent differences in mine counts, with occurrence-based allocations consistently incorporating approximately twice as many Copper and Nickel mines and nearly three times more secondary Zinc instances than the primary allocation. This is consistent with observation of fig. 7 and **??** as for the case of Nickel the in the occurance based weight distribution shows higher median and more spread. This means that likely in most of the

Table 4: Global sum of cumulative target variable flows in the year 2019 distinguished by allocation type. The mine count expresses how many mines participated in estimating the global sum.

Commodity	Cumulative Target Variable	Allocation Type	Sum (Gt)	Mine Count
Copper	Concentrate Production	Occ	6.26	5672
		Prim	11.37	3012
	Ore Processed Mass	Occ	138.15	5670
		Prim	244.46	3010
	Tailings Production	Occ	190.85	5478
		Prim	247.42	2915
	Tailings Production (MB)	Occ	131.98	5644
		Prim	233.23	2992
Nickel	Concentrate Production	Occ	0.87	1475
		Prim	2.18	850
	Ore Processed Mass	Occ	9.11	1476
		Prim	21.09	851
	Tailings Production	Occ	20.96	1424
		Prim	45.52	840
	Tailings Production (MB)	Occ	8.38	1458
		Prim	19.30	833
Zinc	Concentrate Production	Occ	1.80	3336
		Prim	2.70	1217
	Ore Processed Mass	Occ	25.15	3336
		Prim	32.57	1217
	Tailings Production	Occ	49.68	3230
		Prim	42.56	1217
	Tailings Production (MB)	Occ	23.40	3296
		Prim	29.93	1211

Nickel mines is eather the only primary commodity extracted ore extracted with commodities with a smaller weight. As the allocation method was explicitly designed to be instance based we can in the following also devide the allocation factor per instance and produce mine specific maps for Copper **??**, Zinc fig. 49, Nickel fig. 48.

4 Discussion

This study estimated cumulative production flows (CTP, CWRP, COP, CCP) at a global scale with mine-level spatial resolution by calculating the 2019 cumulative production values and introducing a typology of pre-peak, peak-including, and post-peak operations to address item 1. Our findings indicate that the Hubbert model generally outperforms the FEMP model in fitting production time series, particularly for peak-including instances, as it benefits from complete time series and longer production windows; furthermore, Monte Carlo simulations provided reliable confidence intervals for the Hubbert model, whereas underparameterization in the FEMP function resulted in collapsed intervals. The high explained variance of the Hubbert model at the individual mine level suggests that common criticisms—such as those concerning price volatility, regulatory interventions, and technological advancements (Wellmer & Scholz, 2017), as well as metal substitution and recycling effects (Calvo et al., 2017)—are more relevant to aggregated commodity production than to site-specific assessments where local geological characteristics primarily govern economically viable reserves (Calvo et al., 2017; Dominy, 2002) and industrial constraints limit substitutability (Messner, 2002). Additionally, integrating cumulative production flows with predictive features in machine learning models, augmented by targeted data augmentation, substantially improved performance metrics (with R^2 values of 0.88 for COP and 0.84 for CTP) and 0.75 for CCP, although a decreasing trend in algorithmic performance along the production chain suggests that increasing technological influences may obscure the relationship between polygon-derived land use features and production flows. Notably, our assumption of a bimodal log-normal distribution for the target variables was validated by significant improvements in both RMSE and R2 after oversampling the upper quantiles and undersampling the median using SMOTGN, implying that a minority of mines (approximately 10%) disproportionately contribute to global cumulative production—a finding consistent with studies that report the largest 20% of deposits account for over 92% of known copper (Singer, 2017) and that world-class mines exhibit both high grade and high volume (cf. Ridley, 2013, p. 13).

4.1 Halve of the global COP production in 2019 is undocumented

The comparison between the COP estimates from this study and those derived from the GMF database (United Nations Environment Programme, 2024) reveals two key findings. First, the estimates presented here generally exceed the GMF values by a factor of 2 to 10, particularly for countries with significant historical mining activities, such as Chile (CHL), Canada (CAN), and Peru (PER). Second, as the GMF-based COP increases, the two datasets tend to converge. However, stark discrepancies remain for certain countries, such as Germany (DEU), Hungary (HUN), and Jamaica (JAM), where estimates diverge by

multiple orders of magnitude.

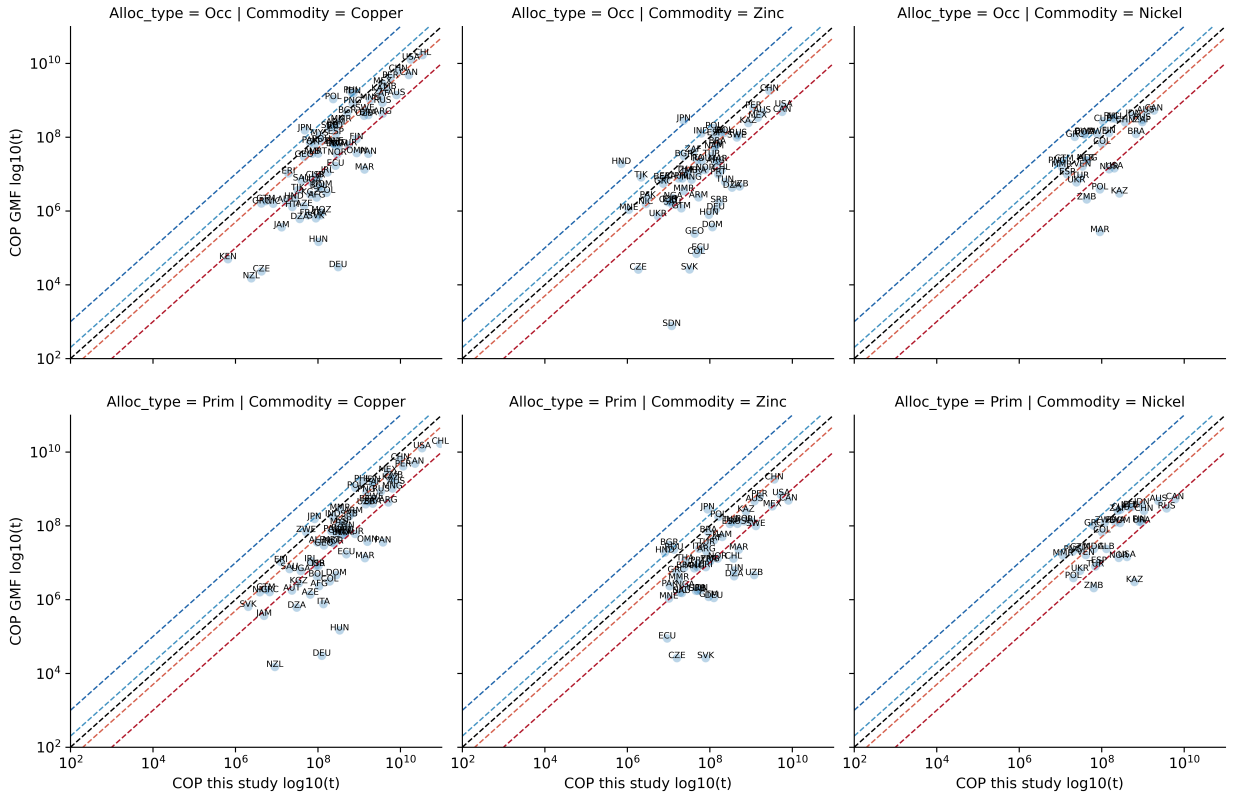

Figure 8: Per-country aggregates of COP for predicted values and validation data derived from United Nations Environment Programme (2024) Both axes are log-scaled (base 10). ISO-3 country codes label scatter points, with lines indicating parity (black), factor 10 deviation (red), and relative differences to the validation data (blue). Columns represent commodities (Copper, Zinc, Nickel), while rows distinguish allocation types (Primary and Occurrence)

These differences arise from both methodological and data-related uncertainties. The most important difference if the size of assed instances where this study quantifies 6808 mines that extract either Copper/Nickel/Zinc derivatives the GMF includes 3064 mining instances covering all metal-ore types. This is an major advantage of this geospatial data driven approach as a substantial substet of the global commodity extraction is unknown. Further GMF database relies on temporally constrained sources, such as the British Geological Survey, which excludes data prior to 1970. In contrast, this study accounts for cumulative extraction over the lifetime of a mine, making temporal alignment challenging. Further differences stem from the GMF's use of linear interpolations to address missing data, whereas this study applies a parabolic production hypothesis. Additionally, GMF estimates rely on national-scale recovery rates and ore grades, whereas this study integrates instance-specific information. Potential overestimation in this study may result from the SVR model's tendency to overpredict high-production mines or the assumption of continuous production

from mine initiation to closure. Indicative can be the outlier sensitivity of the final CTP Regressor. Finally, the overestimation for certain countries may be attributed to historical mining activities not captured in recent national statistics, particularly in regions where metal extraction was significant before 1970, such as Germany, Hungary, and the Czech Republic (Radetzki, 2009). Future research should refine uncertainty assessments, particularly regarding cumulative production fluctuations in 2019.

4.2 Limitations and Future Outlook

This study critically examines the modeling of cumulative production flows while acknowledging several key limitations and avenues for future research. One of the clearest drawbacks of this approach is that it is heavily reliant on high quality times series data for mining instances which is famously scarce. For this reason the train/test subsets are relatively small compared to the global commodity extracting instances. Depending on the allocation principle 3010 (primary) operations extract globally copper containing ore however the sample included only 320 instances therefore approximately 10% and even less for CCP and CTP tasks. Further, the assumption that mine production is continuous from inception until 2019 with convergence towards a finite cumulative extraction simplifies reality, as actual production is influenced by market dynamics and political factors (Calvo et al., 2017; Wellmer & Scholz, 2017). The dataset's bias toward post-1990 reporting, which yields denser time series data and better Hubbert model fits, further constrains the analysis. Moreover, the aggregation of ores, concentrates, tailings, and waste rock into single instances may introduce errors when specific concentrates disproportionately affect tailings production, compounded by a simplified mass balance that omits flotation chemicals and tailings moisture. The construction of confidence intervals via a Monte Carlo method, assuming truncated normal distributions, also requires further validation, particularly for cases like the FEMP model where bimodal parameter distributions are observed. In deriving land use features, assumptions such as equal area partitioning among mines and the use of cluster centroids can obscure true spatial heterogeneity (Werner, Mudd, et al., 2020; Xia et al., 2012). For example it could be already established that variables like CTP show higher correlation with concreate delineated tailing pond area then with the overlal land use of the mine (Werner, Mudd, et al., 2020; Xia et al., 2012). Additionally, the absence of direct proxies for processing chain efficiencies highlights the need to integrate remote sensing data (Maus & Werner, 2024; Werner et al., 2019). Also this work does not distinguish explicitly between Open pit mines and Underground mines however this is crucial as Open pit mines in general show higher amounts of extracted Overburden and therefore cumulativ waste flows relative to the commodity extracted.(Mudd, 2007; Werner, Mudd, et al., 2020) Future research should focus on exploring the heterogeneity of mine samples regarding their cumulative flow values as this work could show that a balanced dataset of "average"

mines and world-class mines is a critical factor for predictive performance. Further for dependent variables where the cofounding variable is known there is a potential the incorporation of physics-informed learning techniques in the learning process (Karniadakis et al., 2021) for instance trough inducing a learning bias (mass balance coherents) in the loss fucntion. For adressing the impact of the mentioned consumption and production driven increases in metal production it will be further of utmost importance to integrate derived estimates into spatio-temporal models of mining mining production that link places of extraction and places of consumption and enable decision making along the value chain. (Sun et al., 2019; Wieland et al., 2022).

As many studies in the field also this work recommends the systematic gathering of primary data by corporate enterprises validated by academic institutions and publically open to inform policy maker along value chaines (Maus & Werner, 2024; Sonter et al., 2023). Lastly, Comparing maps of global Copper, Zinc and Nickel extraction of this study with estimates for in-use building stock (Haberl et al., 2021) and infrastructure stock (Wiedenhofer et al., 2024) it becomes evident where deprived resources from places of extraction accumulate. How places of accumulation affect places of extraction, who benefits from this material accumulation and how is this depreviation/accumulation related with increase/decrease in wellbeing of human population at variouse places are crucial questions for the coming century.

5 Conclusion

This work developed a global, mine-specific geospatial machine learning framework to predict cumulative production flows—namely CTP, CWRP, COP, and CCP—which are crucial for environmental and social impact assessments given that many adverse effects correlate more strongly with cumulative extraction than with annual production. To this end, a process model based on the Hubbert hypothesis of a parabolic production function was first validated, achieving an explained variance of approximately 60%, and then integrated with advanced feature transformation, geospatial feature engineering, and data augmentation techniques to build a Support Vector Regression model that reached R2 values of 0.88 for COP, 0.84 for CTP, and 0.75 for CCP. Applied to a dataset of 6,800 mines extracting Copper, Nickel, or Zinc, the model estimates a global cumulative COP of 486 Gt and a cumulative CTP of 460 Gt. Two allocation procedures were implemented: occurrence-based and primary allocation. For instance, Copper's COP is estimated at 138.15 Gt over 5,670 mines under the occurrence-based method, increasing to 244.46 Gt over 3,010 mines under the primary allocation. Furthermore, a comparison with the GMF database (United Nations Environment Programme, 2024) reveals that our COP estimates for 2019 exceed conventional values by a factor of 2.2, reflecting an substantial gap in the documentation of global primary extraction and the resulting impacts. Overall, the findings also indicate that the Hubbert model outperforms the FEMP model—especially for peak-including operations—and that a minority of world-class mines disproportionately drive global cumulative production (Ridley, 2013; Singer, 2017). These insights are pivotal for refining spatially explicit production inventories and for informing strategies to mitigate environmental impacts in the context of increasing raw material demands (Creutzig et al., 2016; Wiedenhofer et al., 2024).

Appendix

A Method and data

A.1 Modelling of cummulative production

Table 5: Comparison between the Equations of Mineral Production Forecast (FEMP) and the Hubbert Model

Feature	**FEMP**	**Hubbert**
Focus	Annual production-reserves ratio (C).	Peak production and lifecycle of resource extraction.
Model Structure	Geometric and power sequences.	Symmetric exponential growth and decline.
hline **Time Dependency**	Reserves and production evolve with fixed (C).	Time drives production, with t_0 as peak time.
Scenarios	Flexible by adjusting (C).	Implies a predetermined production peak.
Primary Use	Direct exponential decline after mining operation start	Time series that show distinct singular production peak and symmetrical production increase/decrease pre and post peak

A.2 Machine Learning modelling and analysis

Table 6: Machine learning models suitable for small sample sizes with weak-to-moderate correlations and mixed data distributions. Derived from (Burkov, 2019; Géron, 2023)

Model Class	Model Name and Rationale for Inclusion
Linear Models	**Linear Regression**: Handles numeric-binary relations effectively, interpretable coefficients, and low overfitting risk for small datasets. **Ridge Regression**: Adds regularization to manage weak-to-moderate correlations, reducing the risk of overfitting. **Lasso Regression**: Performs feature selection by setting small coefficients to zero, useful for sparse and low-sample datasets. **Elastic Net**: Combines Ridge and Lasso for balanced regularization, ideal for correlated features.
Tree-Based Models	**Decision Tree**: Non-parametric, handles mixed data types and captures non-linear relationships. **Random Forest**: Reduces overfitting compared to single trees, robust to small sample sizes. **Gradient Boosting**: Builds sequential trees to optimize performance, effective with weak correlations.
Kernel-Based Models	**Support Vector Machine (SVM)**: Handles numeric and binary variables, particularly effective with log-normal distributions using kernels.
Bayesian Methods	**Bayesian Ridge Regression**: Incorporates uncertainty estimation and works well for small datasets with weak correlations. **Gaussian Process Regression**: Highly flexible, accounts for uncertainty, and performs well with small sample sizes.
Instance-Based Learning	**K-Nearest Neighbors (KNN)**: Simple, interpretable, and effective for low-sample-size problems, particularly with normalization.
Neural Networks (Shallow)	**Multi-Layer Perceptron (MLP)**: Can approximate non-linear relationships in small datasets when hyperparameters are carefully tuned.

A.3 Uncertainty analysis

A.3.1 Monte Carlo estimation of confidence intervals

A time vector $\mathbf{t}$ is defined as:

$$\mathbf{t} = [0, 1, \ldots, t_{\max} - 1]$$

Hubbert Model For the Hubbert model for the optimal parameter set L^*, k^*, and t_0^* truncated normal distributions are initialized.

$$\mu_L = L^*, \mu_k = k^*, \mu_{t_0} = t_0^* \tag{16}$$

and

$$\sigma_L = \sigma_{L^*}, \sigma_k = \sigma_{k^*}, \sigma_{t_0} = \sigma_{t_0^*} \tag{17}$$

The normal distributions for the parameters are truncated between $a = 0$ and $b = 1$ therefore we can write

$$L \sim TN(\mu_L, \sigma_L, [0, inf]), k \sim TN(\mu_k, \sigma_k, [0, inf]), t_0 \sim TN(\mu_{t_0}, \sigma_{t_0}, [0, inf]) \tag{18}$$

FEMP Model Analogues we can define for the femp model:

$$\mu_{R_0} = R_0^*, \mu_C = C^* \tag{19}$$

and

$$\sigma_{R_0} = \sigma_{R_0^*}, \sigma_C = \sigma_{C^*}, \tag{20}$$

$$R_0 \sim TN(\mu_{R_0}, \sigma_{R_0}, [0, inf]), C \sim TN(\mu_k, \sigma_k, [0, inf]), \sigma_C, [0, 1]) \tag{21}$$

We can generalize the parameter sets as P and define now the mean of the production and cumulative production functions over the $N = 10,000$ samples as

$$\mu_{f(t)} = \frac{1}{N} \sum_{i=1}^{N} f_i(t, P), \quad \sigma_{f(t)} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(f_i(t, P) - \mu_{f(t)}\right)^2} \tag{22}$$

Confidence intervals are derived from the $f(t, P)$ distribution by calculating the upper (95%) and lower (5%) and lower percentiles:

$$CI_{f(t)} = [f_5(t, P), f_{95}(t, P)] \tag{23}$$

The Normalized confidence interval is defined as

$$CI^{Norm}_{f(t,P)} = [\frac{f_5(t, P)}{\mu_{f(t,P)}}, \frac{f_{95}(t, P)}{\mu_{f(t,P)}}] \tag{24}$$

Let P^* be the optimal parameter interval we can write fort the confidence interval

$$CI_{f(t,P^*)} = [\frac{f_5(t, P)f(P^*)}{\mu_{f(t,P)}}, \frac{f_{95}(t, P)f(t, P^*)}{\mu_{f(t,P)}}] \tag{25}$$

This approach assumes that the variability of $f(t, P^*)$ scales proportionally with the normalized uncertainty, which is valid when $f(t)$ exhibits linear or approximately proportional behavior with respect to its parameters near P^*. It further assumes that the underlying distribution of $f(t)$ is well-represented by its mean and that the coefficient of variation ($CV = \frac{\sigma_{f(t)}}{\mu_{f(t)}}$) is small, ensuring stable normalization. This method is not applicable if the mean of $f(t)$ is close to zero, the parameter-response relationship is highly non-linear, or if distributions are multimodal or skewed.

A.4 Allocation

Aspect	Physical Allocation	Economic Allocation	Exergy Allocation	Occurance based (This Study)
Allocation Basis	Mass extracted per commodity	Revenue share based on market prices	Exergy savings/costs from extraction compared to a high-entropy state	Co-occurrence of commodities in the same mines (network structure)
Data Requirement	Tonnes of commodities extracted per mine	Prices of commodities and tonnages extracted	Exergy values of commodities, extraction energy, and entropic states	Binary presence of commodities in mines, co-occurrence data
Advantages	Simple, based on physical quantities	Incorporates economic value of commodities	Reflects the environmental impact of extraction through energy analysis	Captures co-occurrence patterns and network structure, no need for price/exergy data, is not dependent on speculative market fluctuations. Includes still a long term notion of economic viabillity.
Drawbacks	Ignores economic value, does not reflect price fluctuations	Susceptible to price fluctuations in volatile markets	Requires detailed exergy data, which may not be available	Does not directly account for physical or price impacts, can be biased towards commodities that are often extracted. Under represents the importance of large mines with high tonnage for examble Escondilla mine in Chile (Ridley, 2013)
Applicability	Suitable for stable systems with clear mass flows	Ideal when market prices are stable and accessible	Best for evaluating environmental and energy impacts of extraction	Effective when price or exergy data is unavailable, emphasizes co-extraction relationships

Table 7: Comparison of Allocation Approaches. See for economic and physical allocation (Althaus & Classen, 2005) and exergy allocation (Valero et al., 2015)

B Cumulative production model

B.1 Data Coverage

This chapter outlines the analysis of data coverage, temporal trends, and outlier detection for mine production variables across 575 mines with associated target commodities as byproducts or primary products, each reporting at least one nonzero production value. Coverage analysis revealed that OP had the highest average coverage, with 54% of active mine years recorded (standard deviation: 32%), as shown in table 9. Waste rock production, which depends on primary strip ratio recordings, had significantly lower coverage at $6\% \pm 18\%$. Tailings production (TP), derived from concentrate production (CP) and processed ore (OP), exhibited trends consistent with these findings. The lower mean coverage of TP ($31\% \pm 36\%$) compared to OP and CP can be explained as the formar is dependent on the covarage of the letter see eq. (7).

Temporal data trends were analyzed using heatmaps (fig. 10 and **??**), which indicated an overall increase in coverage over time, peaking between 2010 and 2015. The graphs highlight the heterogeneity in the time series where the model is fitted. For some mines, data coverage spans consistently from 1950 to 2023, while for others, consistent coverage cannot be assumed after 1990, with a noticeable increase post-2010. Additionally, data collection for individual mines is not uniform over time, exhibiting multiple gaps. These gaps can be attributed to several factors: (1) limited production due to uneconomic extraction, (2) lack of mining production reporting (both direct and indirect), and (3) fluctuations in data gathering intensity from the services provided by Standard & Poor (2025). Both are concistently observed in public enclosed mining production data (Maus & Werner, 2024).

Reporting errors and outliers—particularly during high-production periods—introduced significant noise into the dataset. To mitigate this, we employed a rolling mean outlier detection method that leverages the parabolic hypothesis of the Hubbert model as a filtering criterion, in accordance with miner production theory (Calvo et al., 2017). This nonparametric approach is preferable to methods such as quartile analysis, which assume static moments and are therefore inconsistent with the dynamics observed in **??**. The method involves selecting an appropriate window size for the rolling mean and a threshold (set at two standard errors) for outlier detection. As shown in **??**, a window size of 10 effectively identified outliers in OP; however, larger window sizes (e.g., 12) tend to capture more long-term deviations. For instance, in the Atlas Toledo copper mine (Copper, Philippines), an initial production value below 5000 kt was flagged as an outlier when using a window size of 12 but not with 10. Similar discrepancies are observed in the San Bartolome mine (Silver, Bolivia), Americano do Brasil (Nickel, Brazil), and New Afton (Gold, Canada), where a smaller window size

is less sensitive to short-term fluctuations. Thus, our analysis indicates that a window size of 12 may be overly sensitive to transient variations, particularly at the beginning and end of the production phase.

For mines like Huckleberry (Copper, Canada) and Sossego (Copper, Brazil), window sizes larger than 12 suppressed high and low production values near the beginning and end of production phases. Conversely, Tocantins (Nickel, Brazil) showed short-term fluctuations that were better captured with a window size of 10. However, window sizes below 10 proved overly sensitive, filtering points that adhered to the general trend.

For WRP, the analysis required smaller window sizes to address frequent data gaps. A window size of 10 provided a balance between sensitivity and long-term trend alignment, as demonstrated for the Alamo Dorado mine (Silver, Mexico) in **??**. This approach minimized the over-identification of outliers, particularly at the beginning and end of time series, and was consistently applied for WRP outlier detection.

Table 8: Summary statistics for the missing values of the mines. The statistic represents the percentage of years per mine that do have a non zero and non NaN value.

	TP	WRP	OP	CP
count	986	986	986	986
mean	0.31	0.06	0.54	0.38
std	0.36	0.18	0.32	0.40
min	0.00	0.00	0.00	0.00
25%	0.00	0.00	0.25	0.00
50%	0.12	0.00	0.56	0.29
75%	0.62	0.00	0.86	0.78
max	1.00	1.00	1.00	1.00

B.2 Outlier analysis

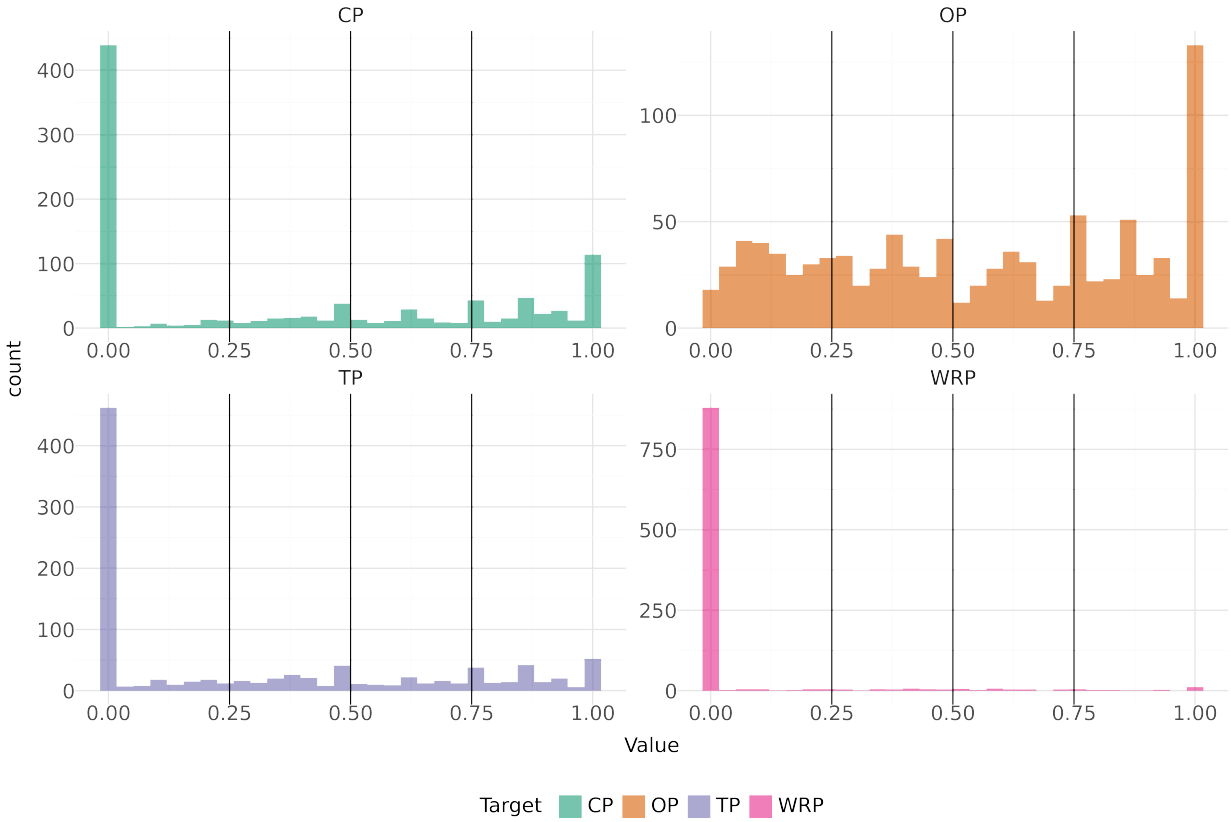

Figure 9: Histograms of the coverage for the 986 mines in the sample for the four target variables. Coverage is defined as ratio of years potential production years from the project start to 2023. The sample only includes mines which are not closed by 2023. The vertical lines include the 25%, 50% and 75% percentile.

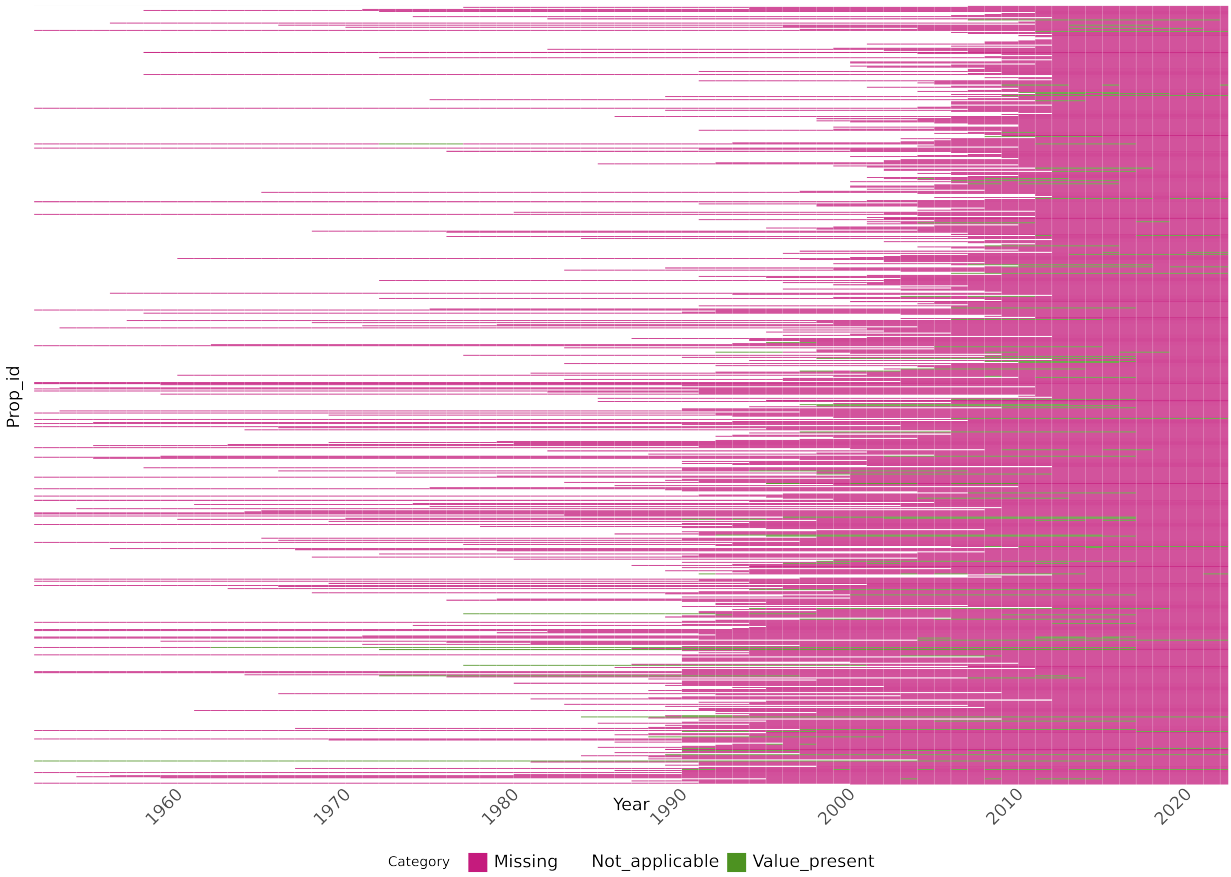

Figure 10: Heatmap of missing values for WRP. A missing value is a year of theoretical activity that has eather a NaN value.The periodicity is one year and the x column is in the interval form 1950 to 2023 in ascending order form left to right. White coloring represents years that are not applicable for the specific mine.

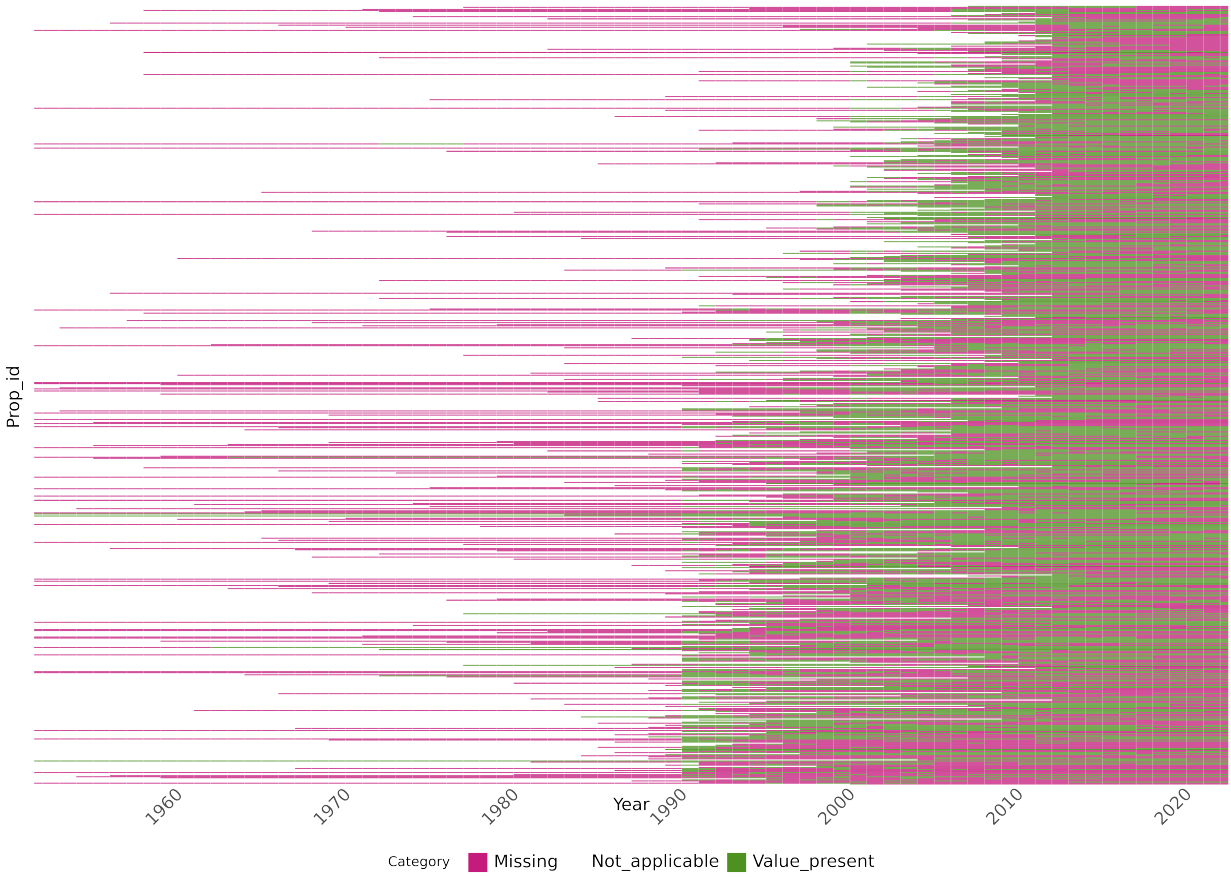

Figure 11: Heatmap of missing values for OP. A missing value is a year of theoretical activity that has eather a NaN value.The periodicity is one year and the x column is in the interval form 1950 to 2023 in ascending order form left to right. White coloring represents years that are not applicable for the specific mine.

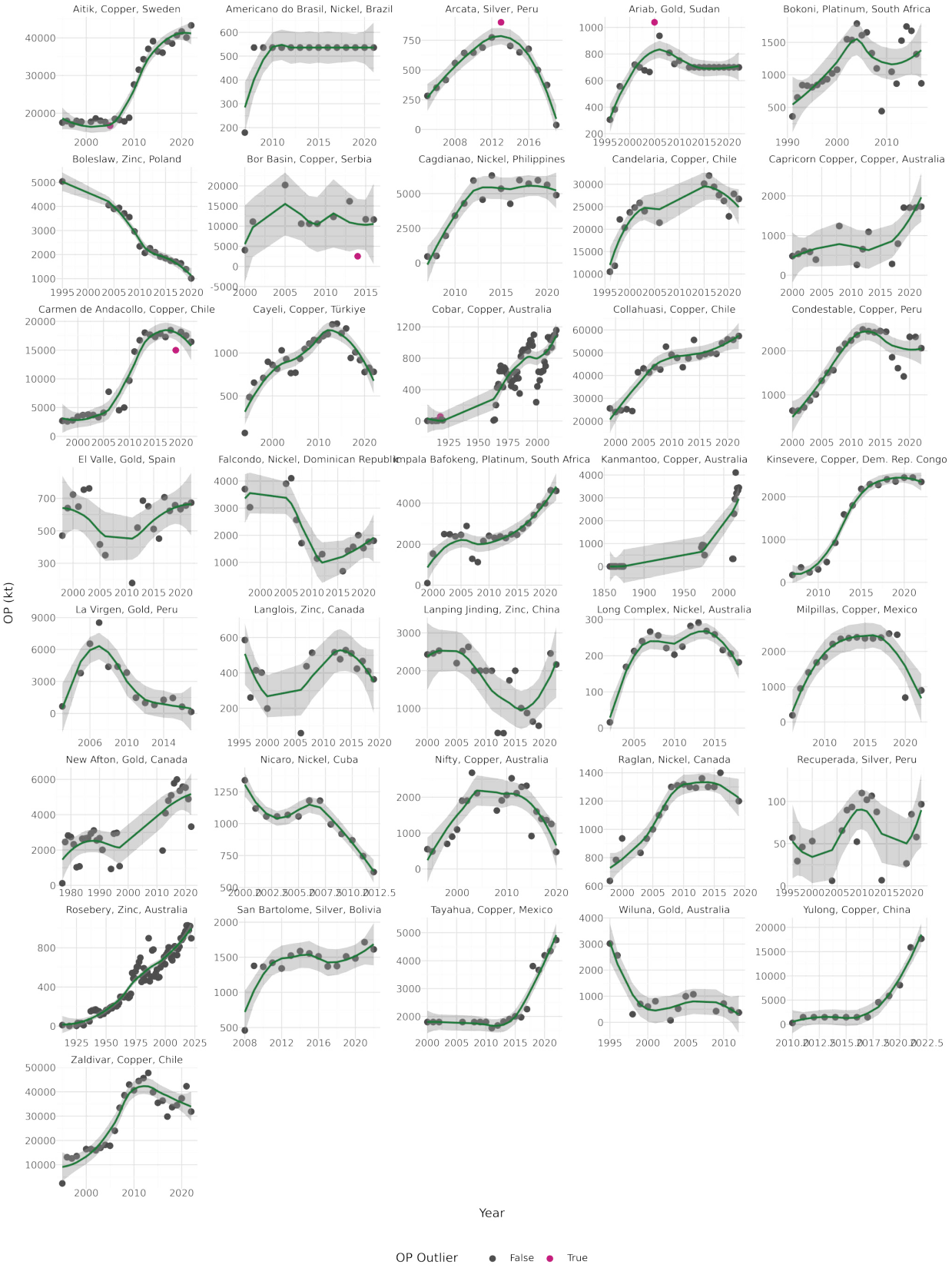

Figure 12: Time series of 36 randomly selected mines for OP. Depending on window size and threshold outliers are identified based on the rolling mean method. This graph represents a window-size of 10 and a threshold of two standard deviations.

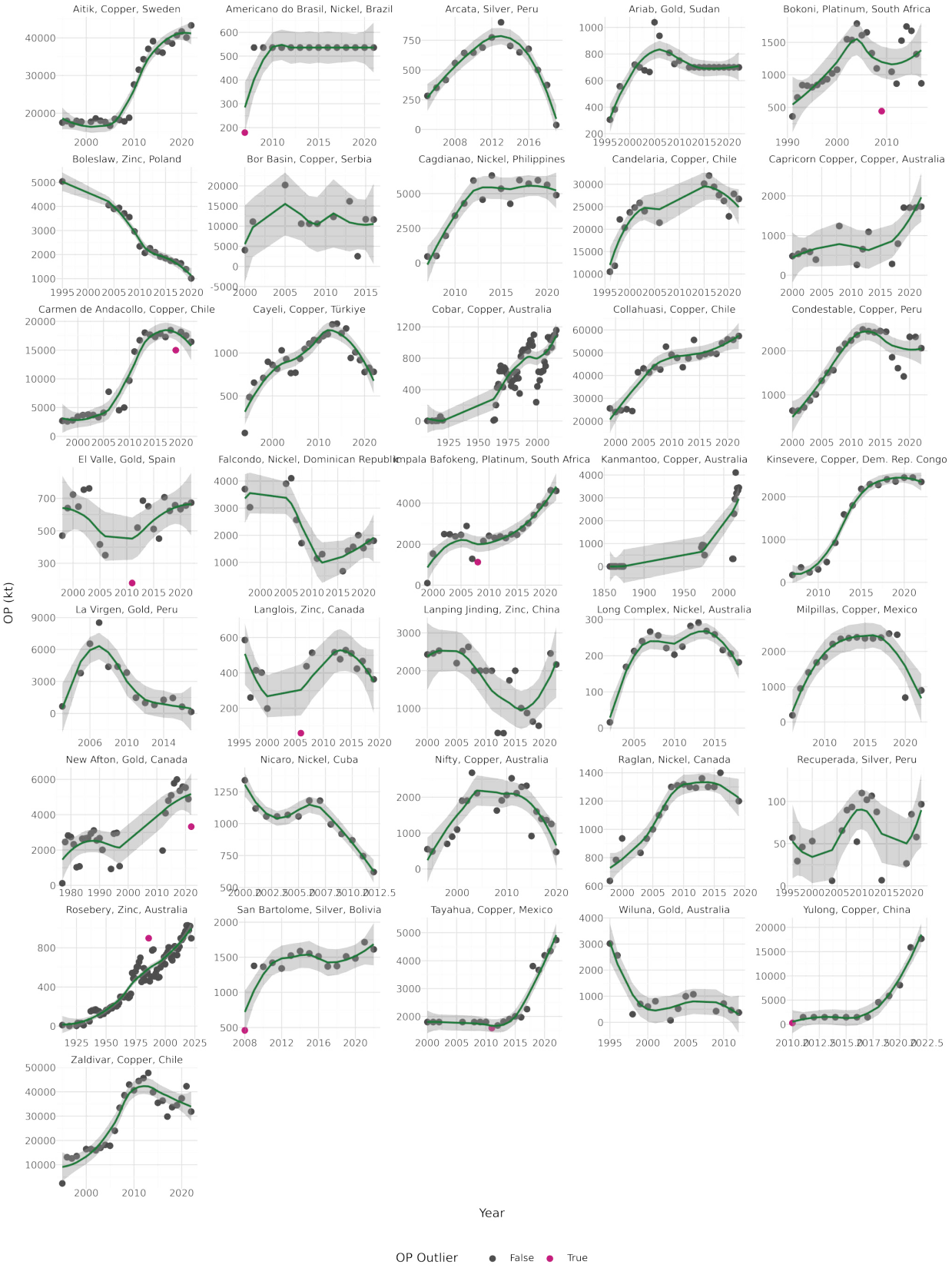

Figure 13: Time series of 36 randomly selected mines for OP. Depending on window size and threshold outliers are identified based on the rolling mean method. This graph represents a window-size of 12 and a threshold of two standard deviations.

B.3 Global Analysis

Table 9: Summary statistics for the fitting of production data to the two models: FEMP and Hubbert. Importantly the Pn_pval variables note the proportion of mines that show a significant p-value for the parameter. For instance for TP the p-value of parameter 1 is in 80% of the cases statistical signficant with and significance level of 5%.

Target_var	CP		OP		TP		WRP	
Model	femp	hubbert	femp	hubbert	femp	hubbert	femp	hubbert
R2_mean	0.41	0.60	0.45	0.62	0.47	0.61	0.42	0.60
R2_std	0.29	0.27	0.30	0.28	0.31	0.29	0.29	0.24
R2_median	0.38	0.68	0.49	0.69	0.48	0.70	0.46	0.62
R2_skew	0.12	-0.68	-0.04	-0.70	-0.07	-0.66	-0.02	-0.43
R2_kurtosis	-1.37	-0.65	-1.35	-0.63	-1.33	-0.81	-1.36	-0.70
RMSE_mean	77465.39	38789.99	2638131.68	1778607.25	2831168.74	1843399.58	12308588.54	11118131.33
RMSE_std	119527.99	59406.80	6680336.02	5649085.46	6411632.98	5915768.04	16813157.62	20315496.47
RMSE_median	37299.77	21098.19	485187.18	254384.70	537573.01	257623.99	6848525.72	5387635.06
RMSE_skew	3.40	4.30	5.57	7.15	5.38	7.78	3.24	4.05
RMSE_kurtosis	14.08	24.89	39.86	66.61	40.12	75.33	12.82	18.61
NRMSE_mean	0.23	0.13	0.26	0.15	0.26	0.14	0.24	0.19
NRMSE_std	0.11	0.07	0.13	0.11	0.14	0.10	0.09	0.11
NRMSE_median	0.22	0.12	0.24	0.12	0.23	0.11	0.23	0.15
NRMSE_skew	1.14	2.06	0.93	1.72	0.89	1.72	0.91	1.29
NRMSE_kurtosis	1.60	6.53	0.75	3.47	0.26	3.00	2.09	2.03
P1_pval	0.82	0.78	0.80	0.80	0.77	0.84	0.78	0.74
P2_pval	0.68	0.79	0.67	0.77	0.60	0.81	0.68	0.68
P3_pval	0.00	0.79	0.00	0.83	0.00	0.86	0.00	0.76
Sample_size	330	330	454	454	264	264	50	50

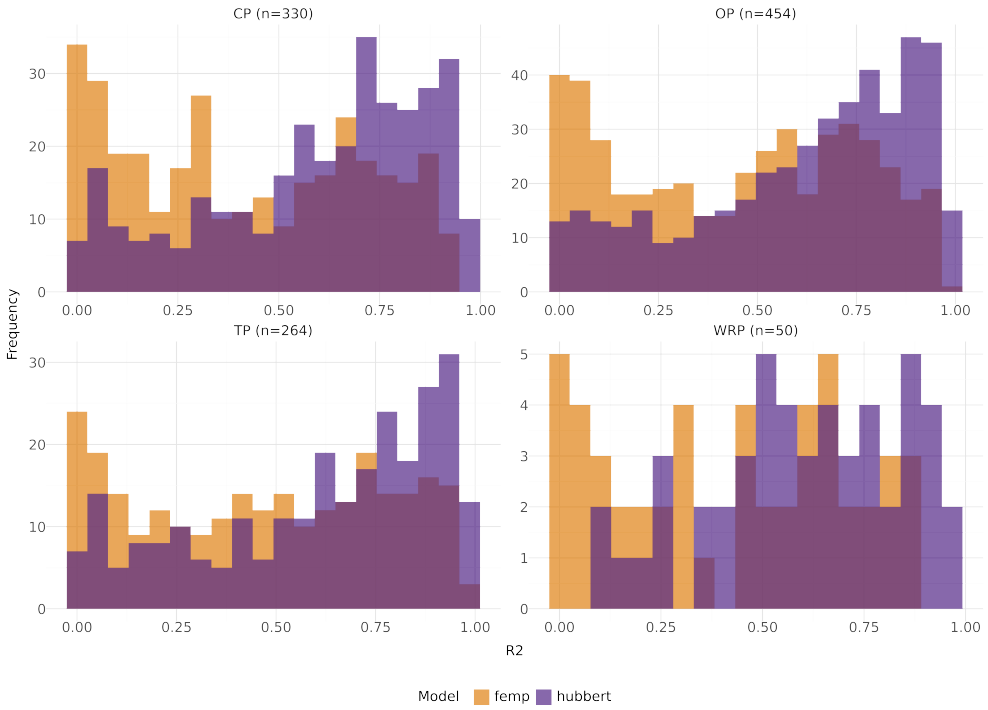

Figure 14: Histograms of R2s vor the different Target variables colored by the specific model fitted. Note that for all target variables the hubbert model shows higher frequencies for higher R2s, however the distributions are still heavily left tailed.

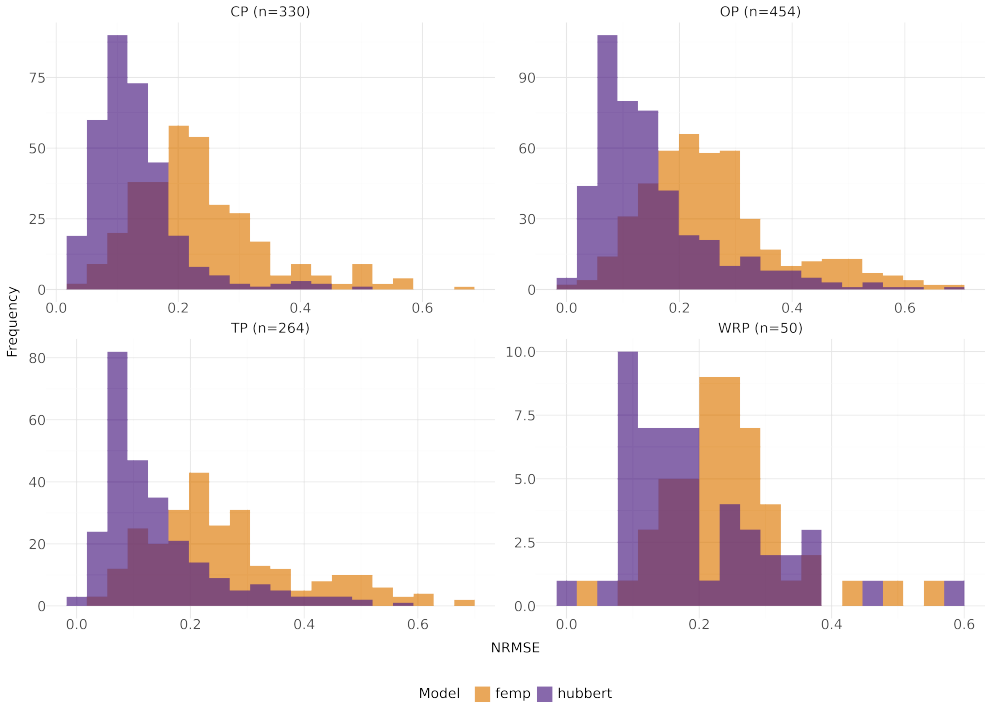

Figure 15: Histograms of NRMSE vor the different Target variables colored by the specific model fitted. The NRMSE defined as the RMSE relative to the maximum observed flow production flow

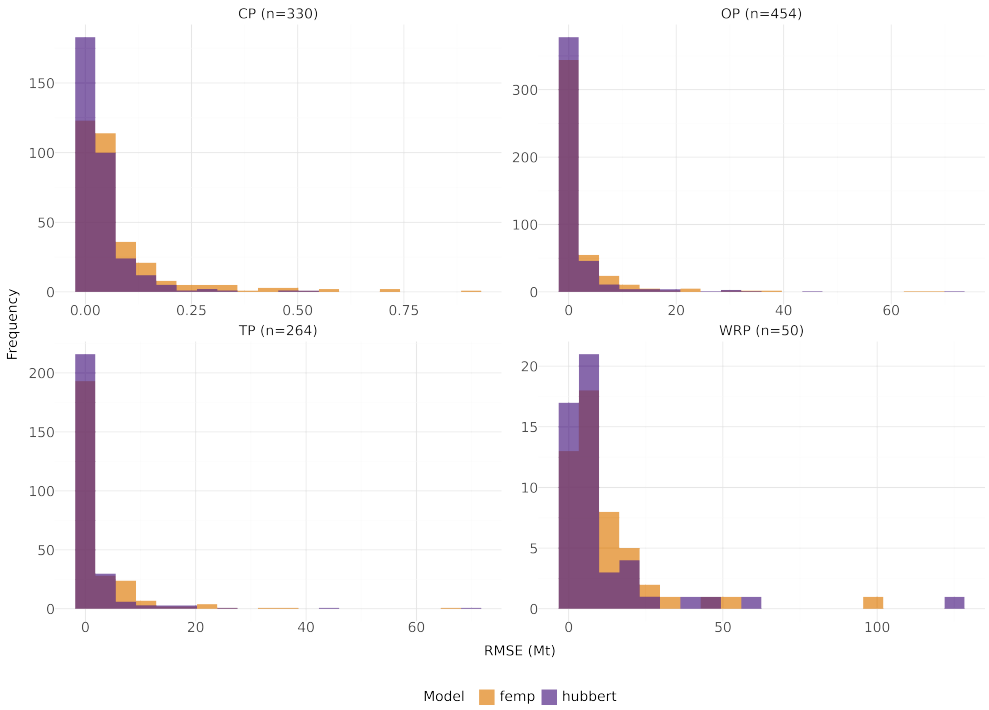

Figure 16: Histograms of RMSE vor the different Target variables colored by the specific model fitted.

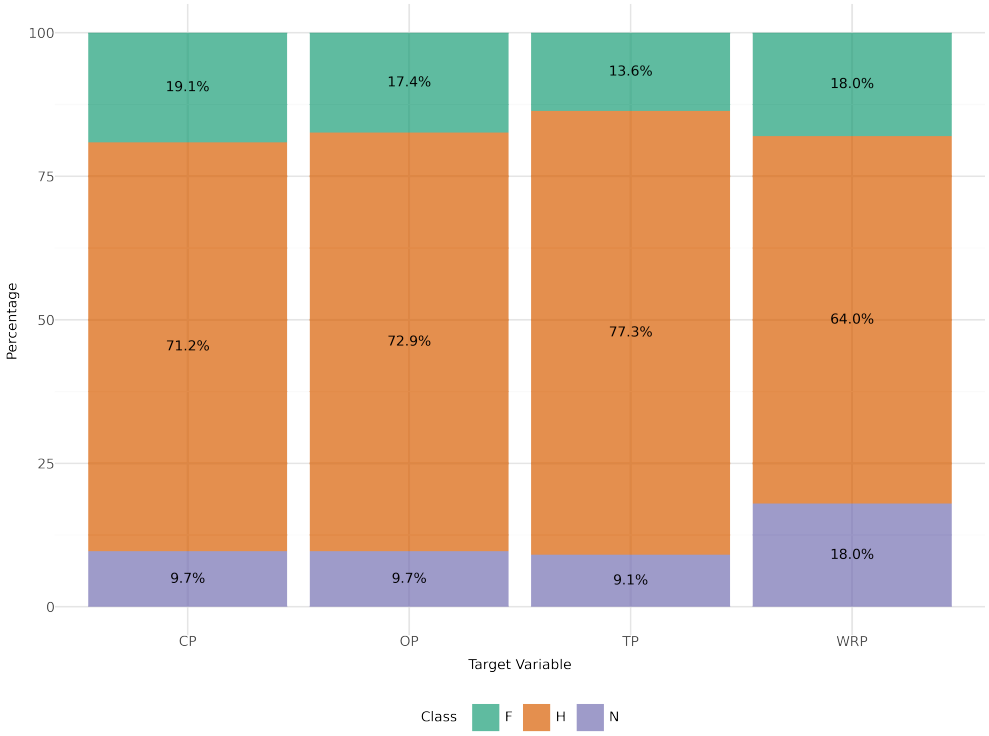

Figure 17: The fitted model instances were categorized if their parameter set was statistical significant for the model fit for the FEMP model (F), Hubbert model or both (H) and not significant (N)

B.4 Local Analysis

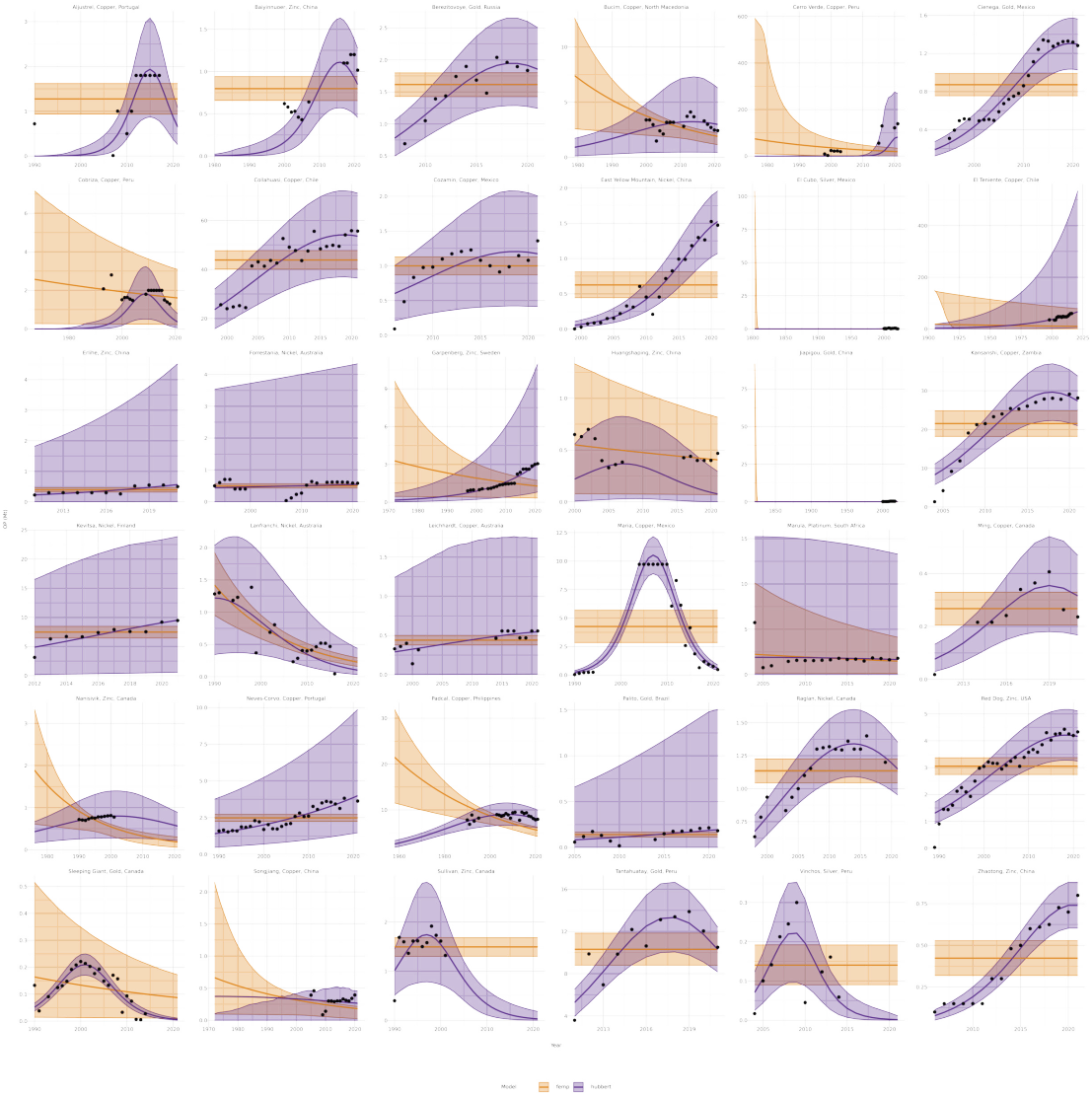

Figure 18: Time series of 36 randomly selected mines for OP. The graph shows the observed data (black), the simulated model fit, and the confidence intervals derived by the Monte Carlo method (FEMP: orange, Hubbert: purple). For visualization purposes, the upper confidence interval was replaced by eight times the confidence interval if the upper bound exceeded this threshold.

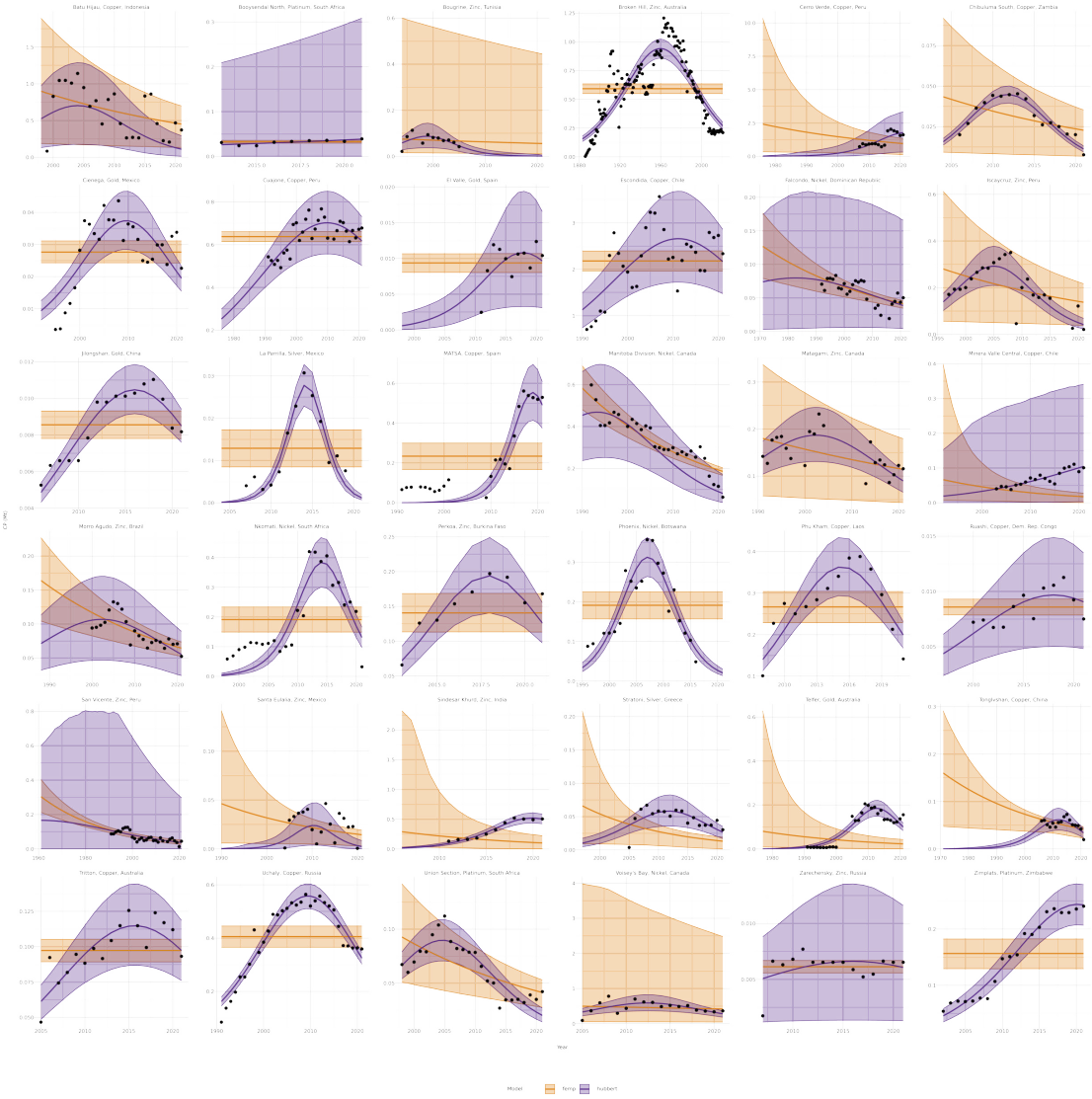

Figure 19: Time series of 36 randomly selected mines for CP. The graph shows the observed data (black), the simulated model fit, and the confidence intervals derived by the Monte Carlo method (FEMP: orange, Hubbert: purple). For visualization purposes, the upper confidence interval was replaced by eight times the confidence interval if the upper bound exceeded this threshold.

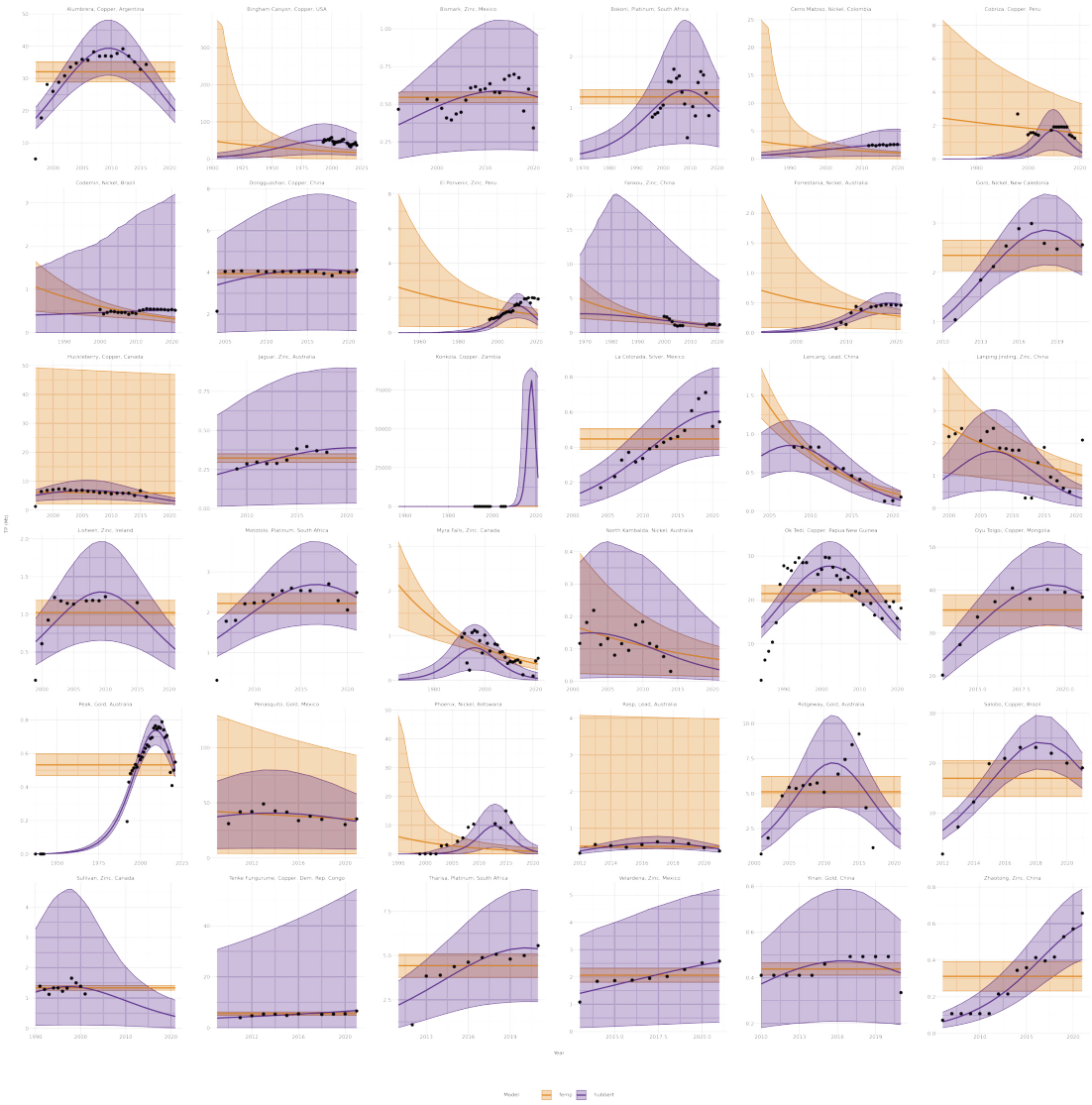

Figure 20: Time series of 36 randomly selected mines for TP. The graph shows the observed data (black), the simulated model fit, and the confidence intervals derived by the Monte Carlo method (FEMP: orange, Hubbert: purple). For visualization purposes, the upper confidence interval was replaced by eight times the confidence interval if the upper bound exceeded this threshold.

C Machine Learning Model

C.1 Exploratory Analysis

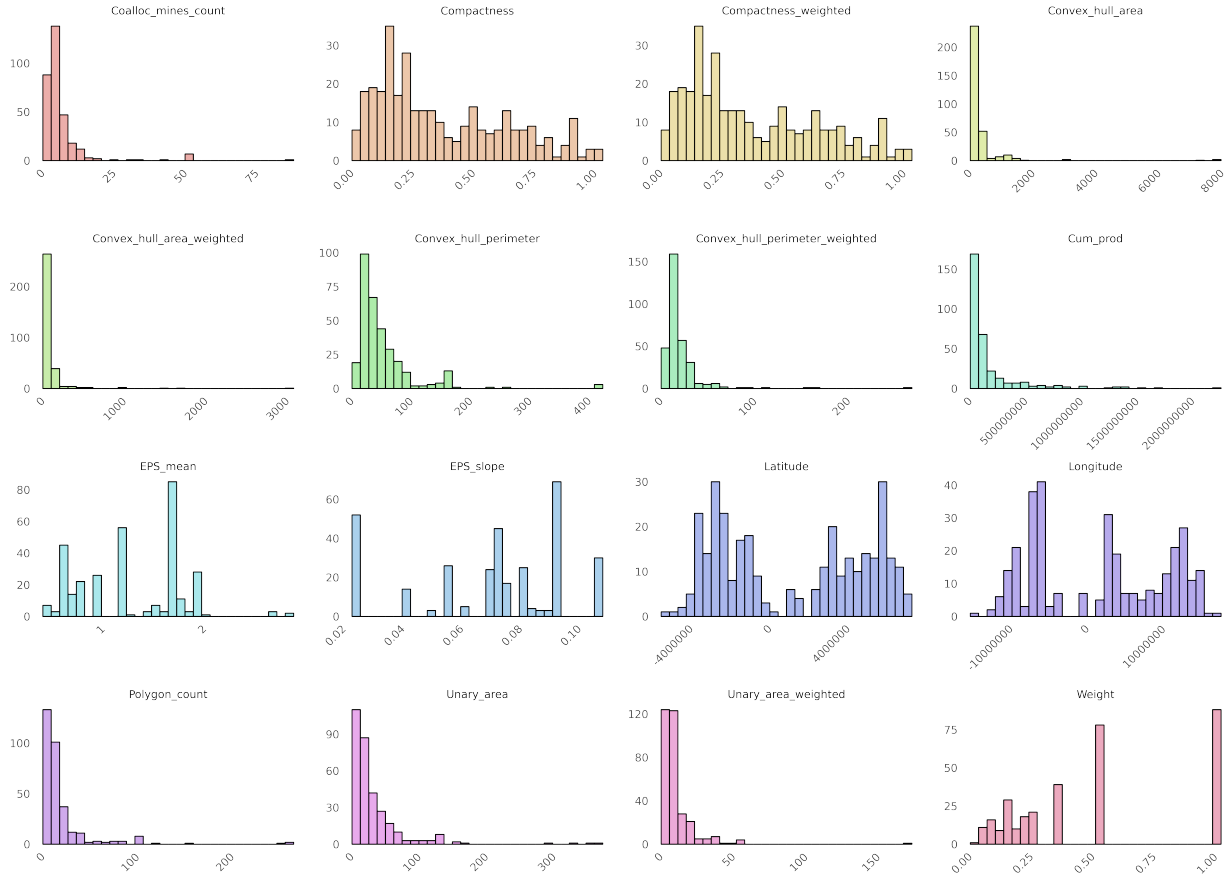

Figure 21: Numerical features for the regression task to predict COP. Sample size = 321

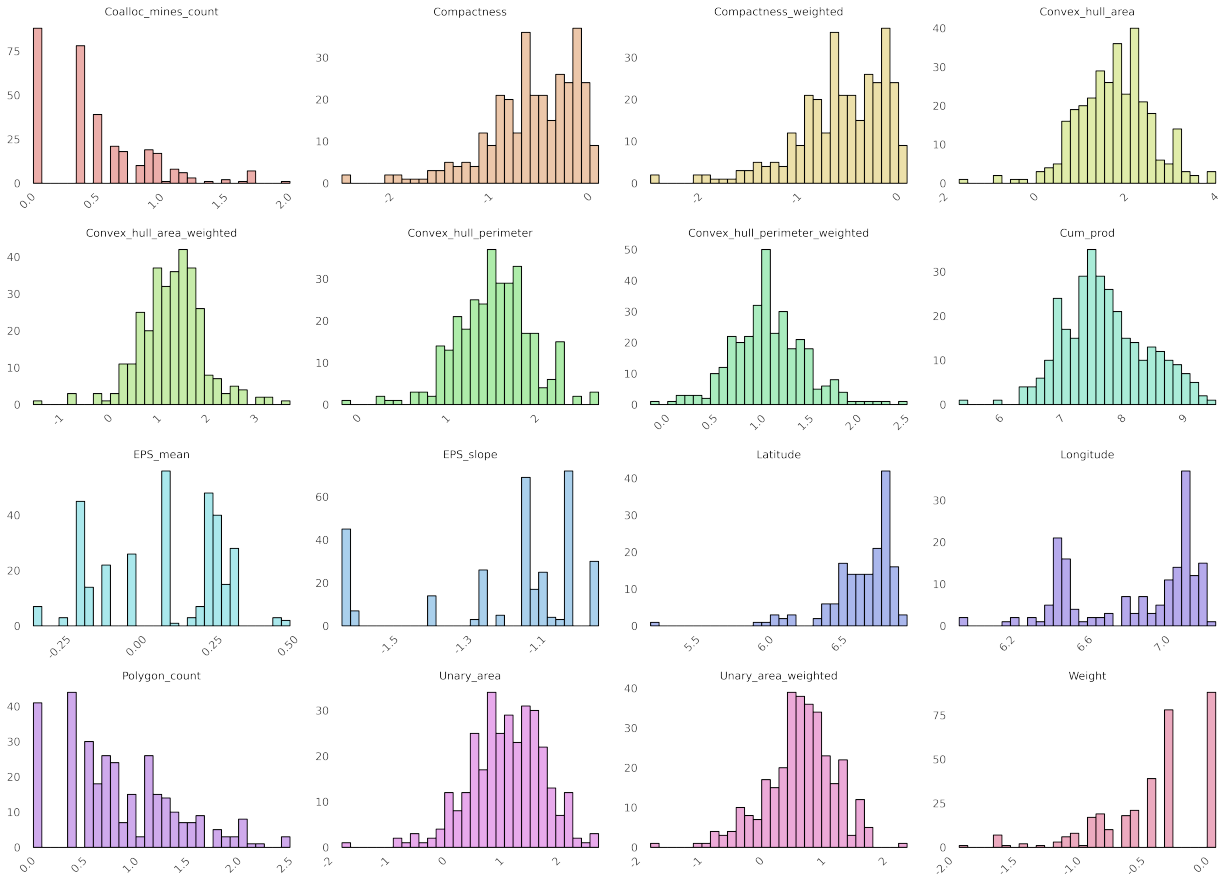

Figure 22: Log transformed numerical features for the regression task to predict COP. Sample size = 321

Figure 23: Categorical features for the regression task to predict COP.

C.2 Correlation Analysis

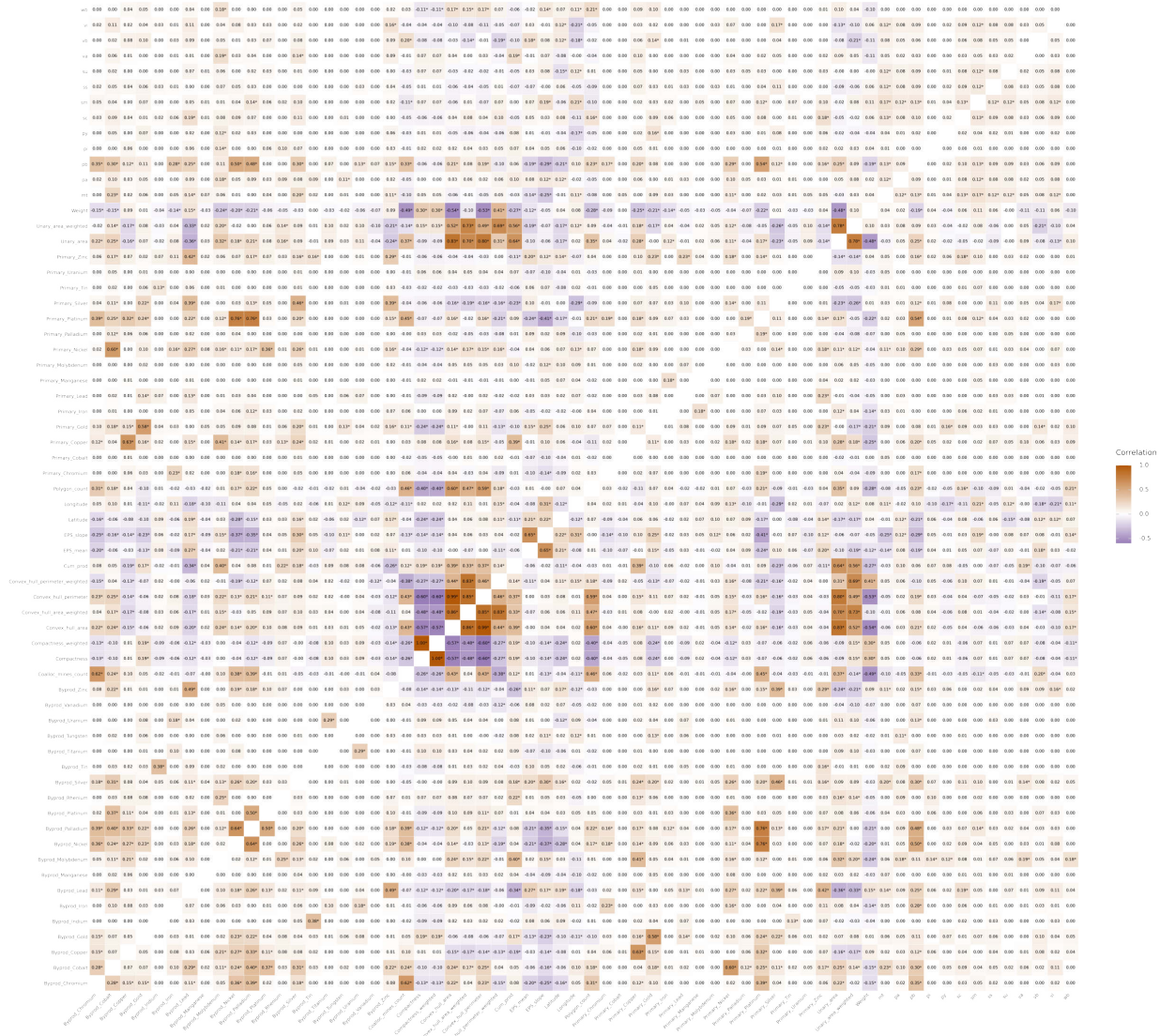

Figure 24: Correlation Heatmap for the Features and the target variables for the COP regression task.

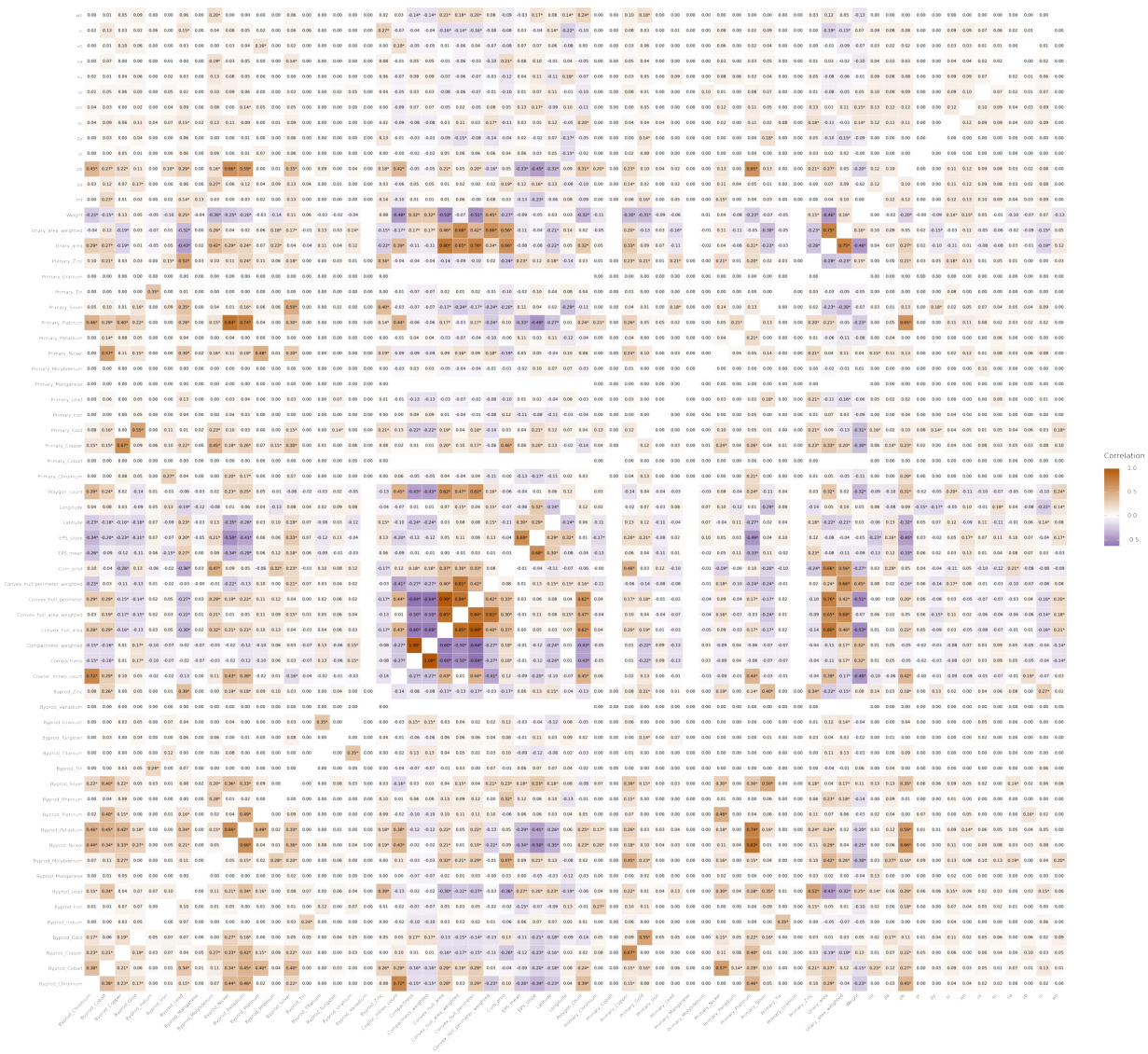

Figure 25: Correlation Heatmap for the Features and the target variables for the CTP regression task.

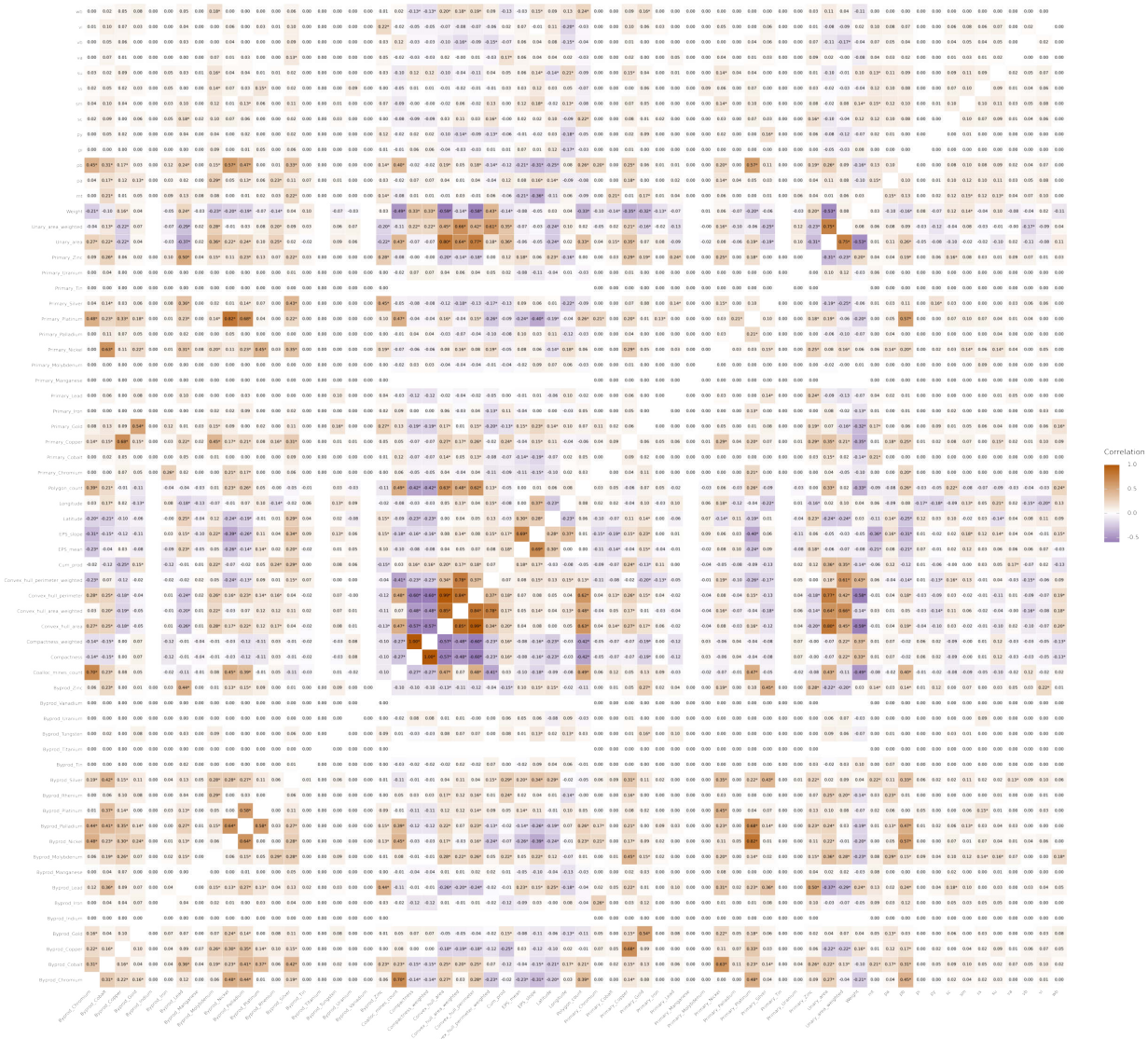

Figure 26: Correlation Heatmap for the Features and the target variables for the CCP regression task.

C.3 Geography feature design

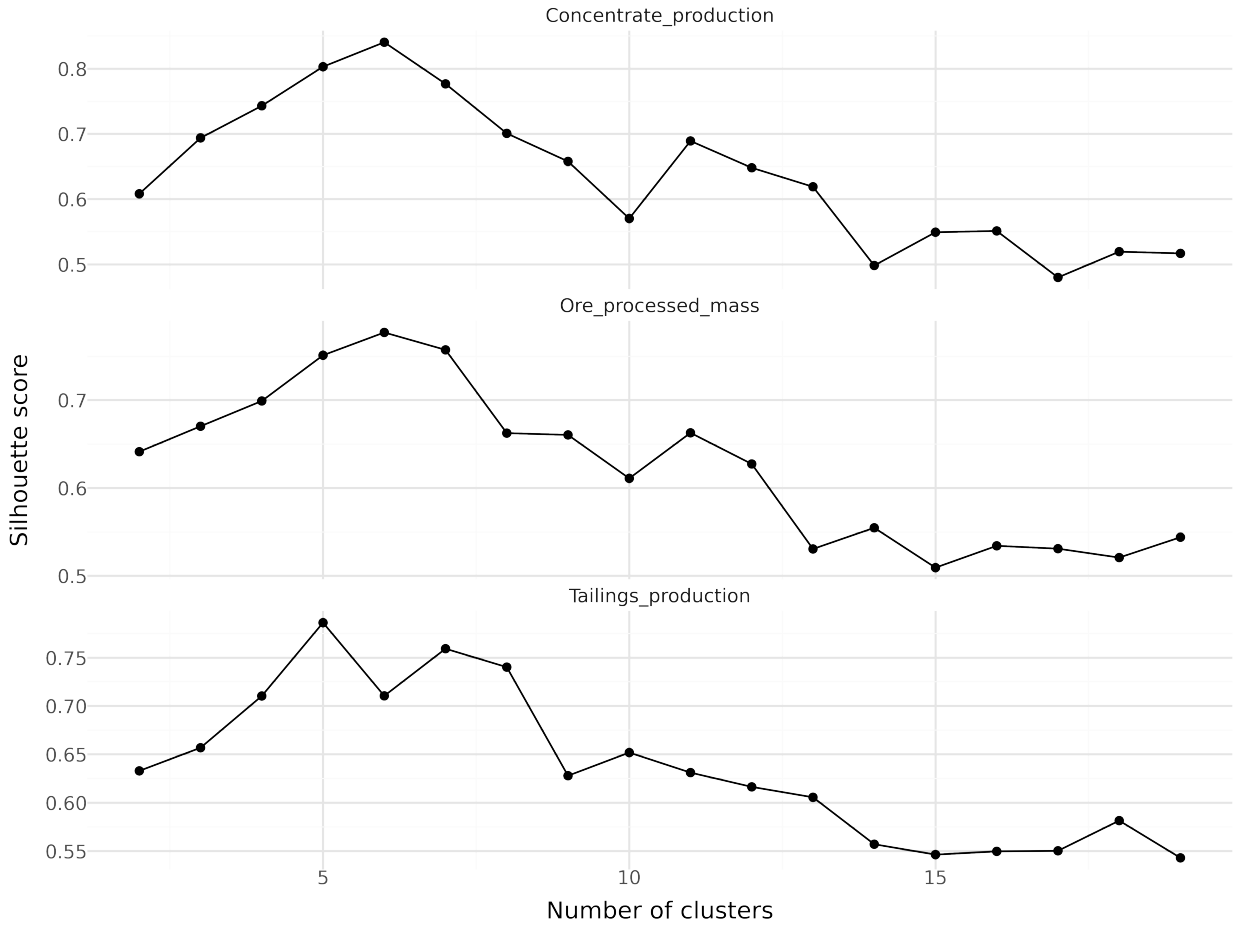

Figure 27: Silhouette score dependent on the cluster numbers for the four cumulative target variables. For COP and CCP $n = 6$ shows the highest score wherease for CTP $n = 5$ shows the highest result. High Silhouette scores indicate high average sameness of instances to the mapped cluster.

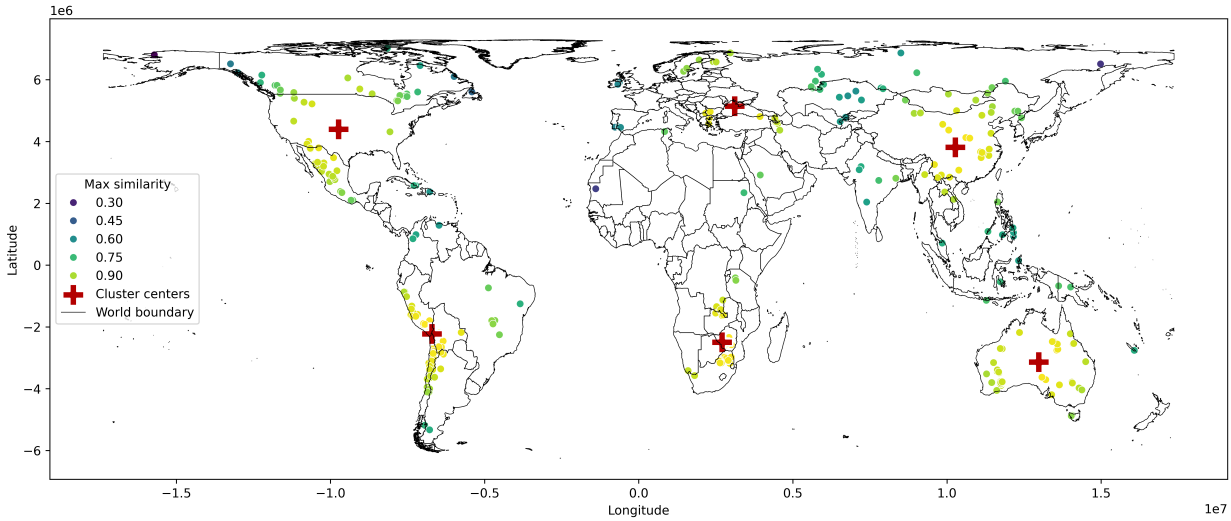

Figure 28: The Shillouette optimized clustering of instances for COP with a $n = 6$ clusters and the maximum similarity of an instance generated for the five clusters.

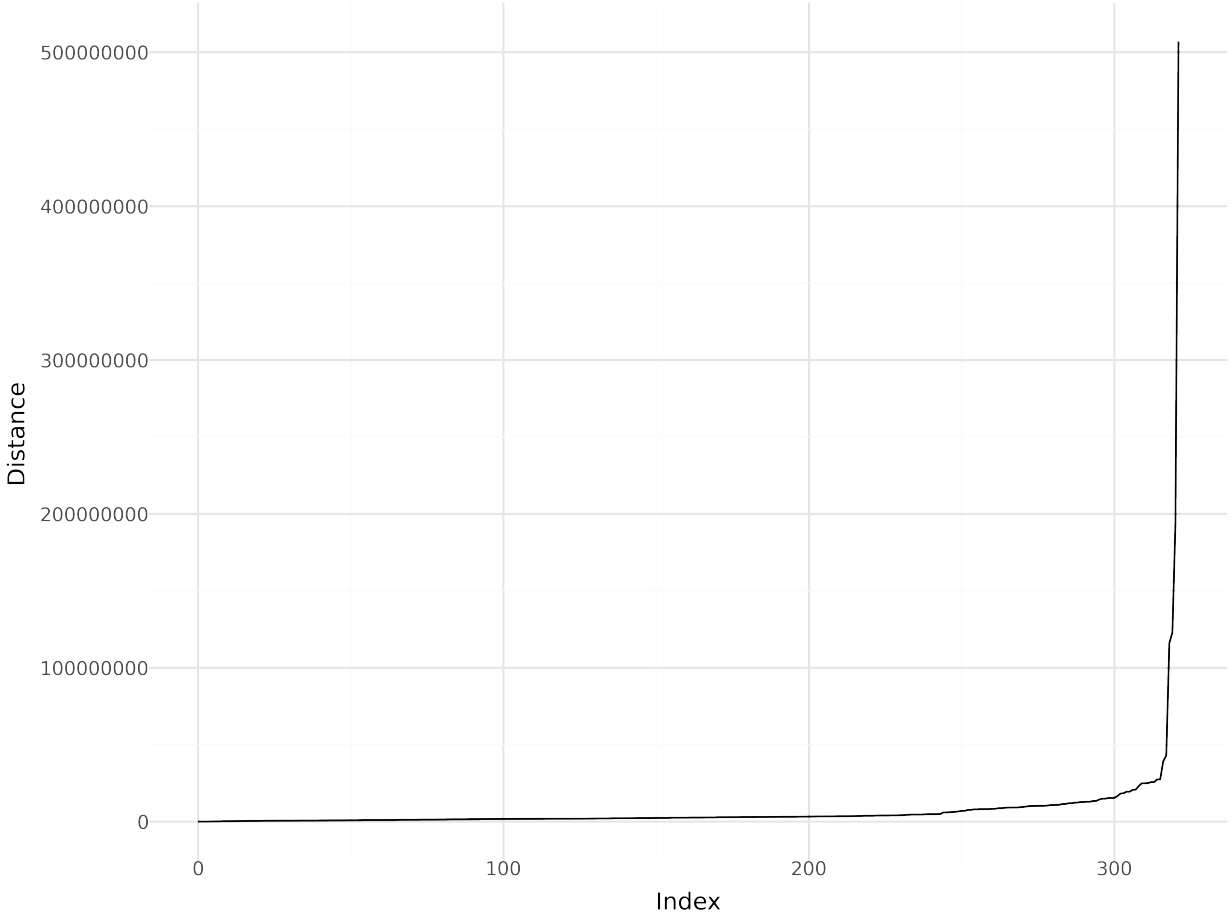

Figure 29: Distance of instance two the optimal cluster using a KMeans ($n = 10$). The distance with the maximum slope is for $\epsilon^* = 193779248$ and indicates the highest change in the distribution in the distance distribution.

C.4 PCA

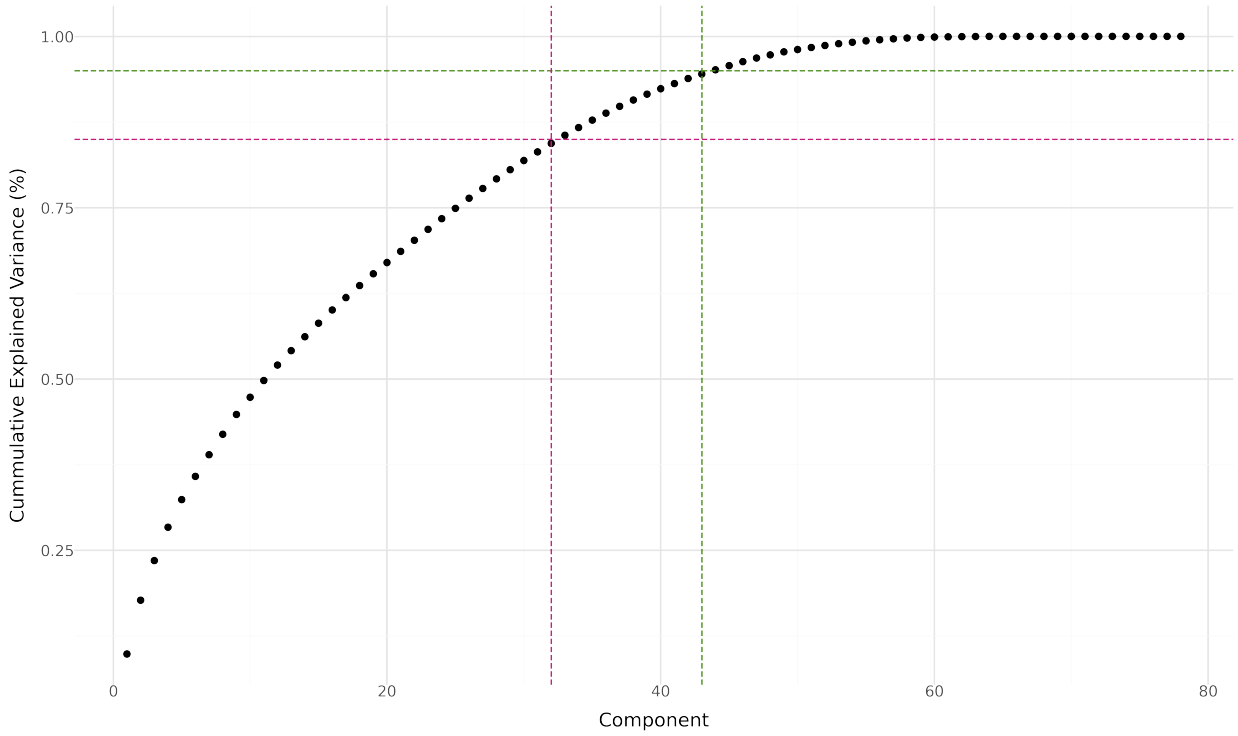

Figure 30: The graph shows the cumulative explained variance against the ith component of the PCA for COP feature space. The dotted lines indicate the number of features that explain 85% (d= 32, purple) and 95% (d=43, green) of the variance of the dataset.

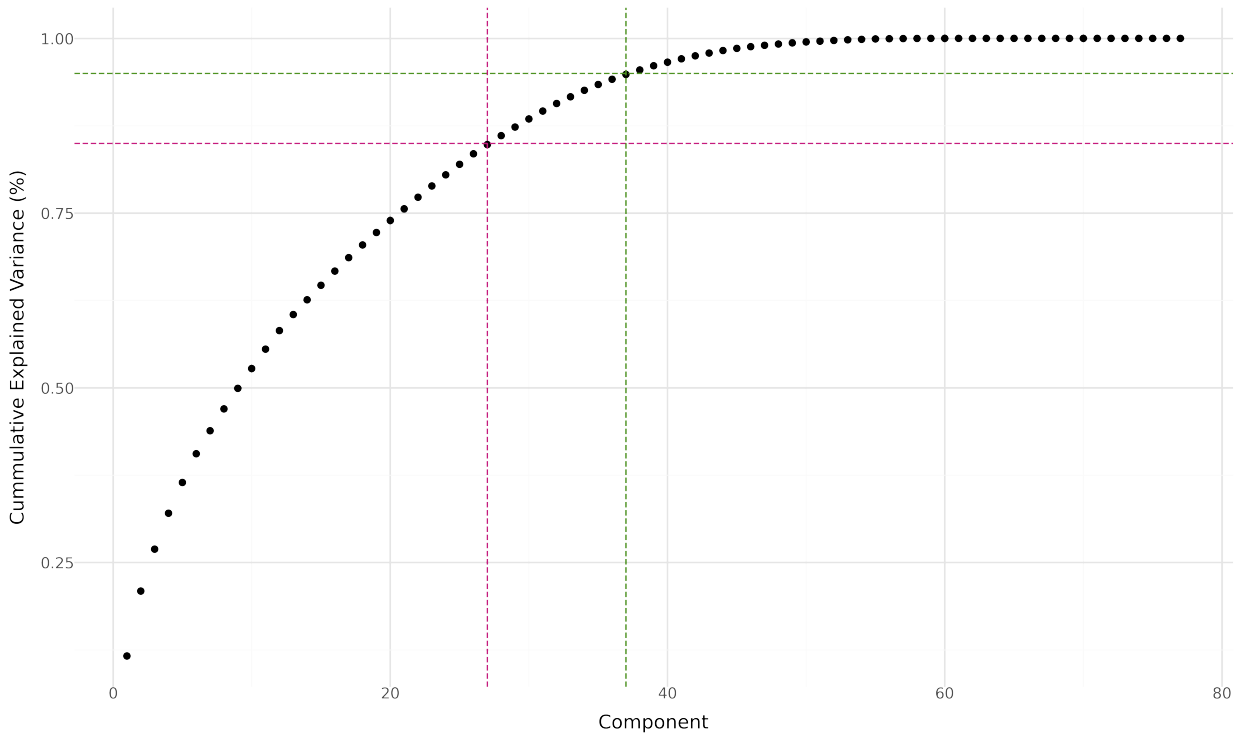

Figure 31: The graph shows the cumulative explained variance against the ith component of the PCA for CTP feature space. The dotted lines indicate the number of features that explain 85% (d= 27, purple) and 95% (d=37, green) of the variance of the dataset.

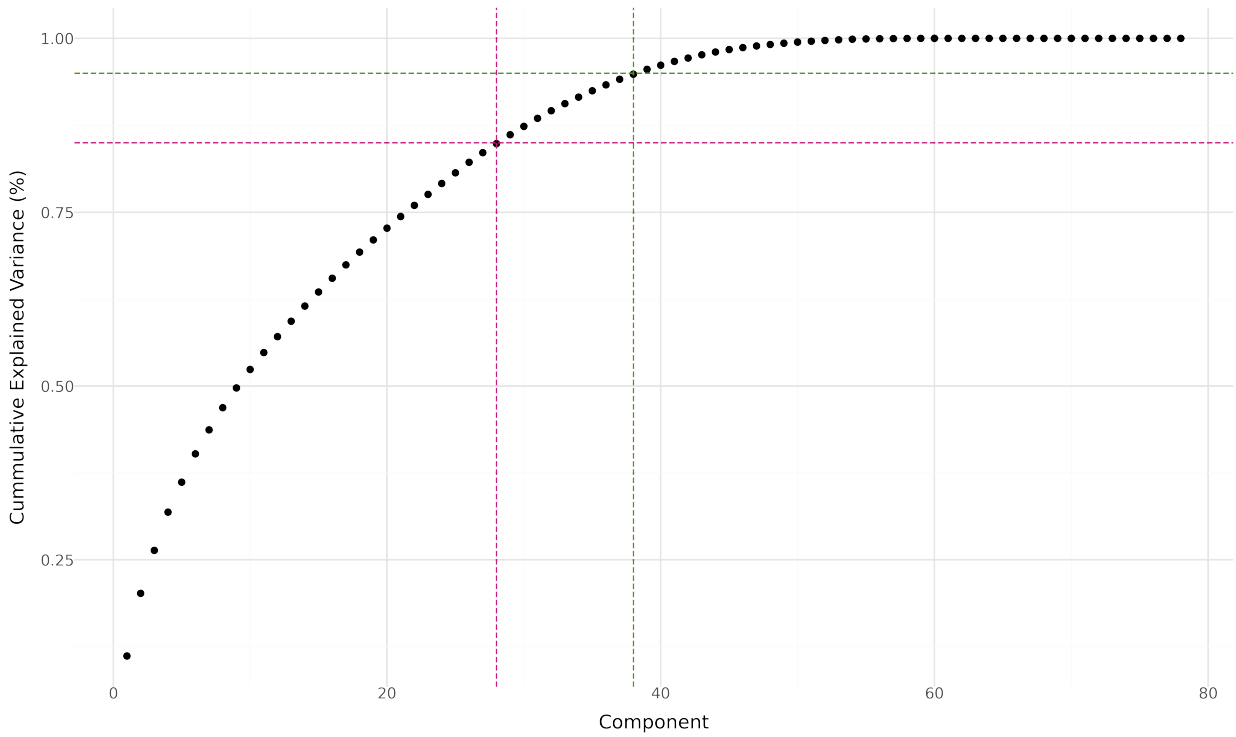

Figure 32: The graph shows the cumulative explained variance against the ith component of the PCA for CCP feature space. The dotted lines indicate the number of features that explain 85% (d= 28, purple) and 95% (d=38, green) of the variance of the dataset.

C.5 Training loop

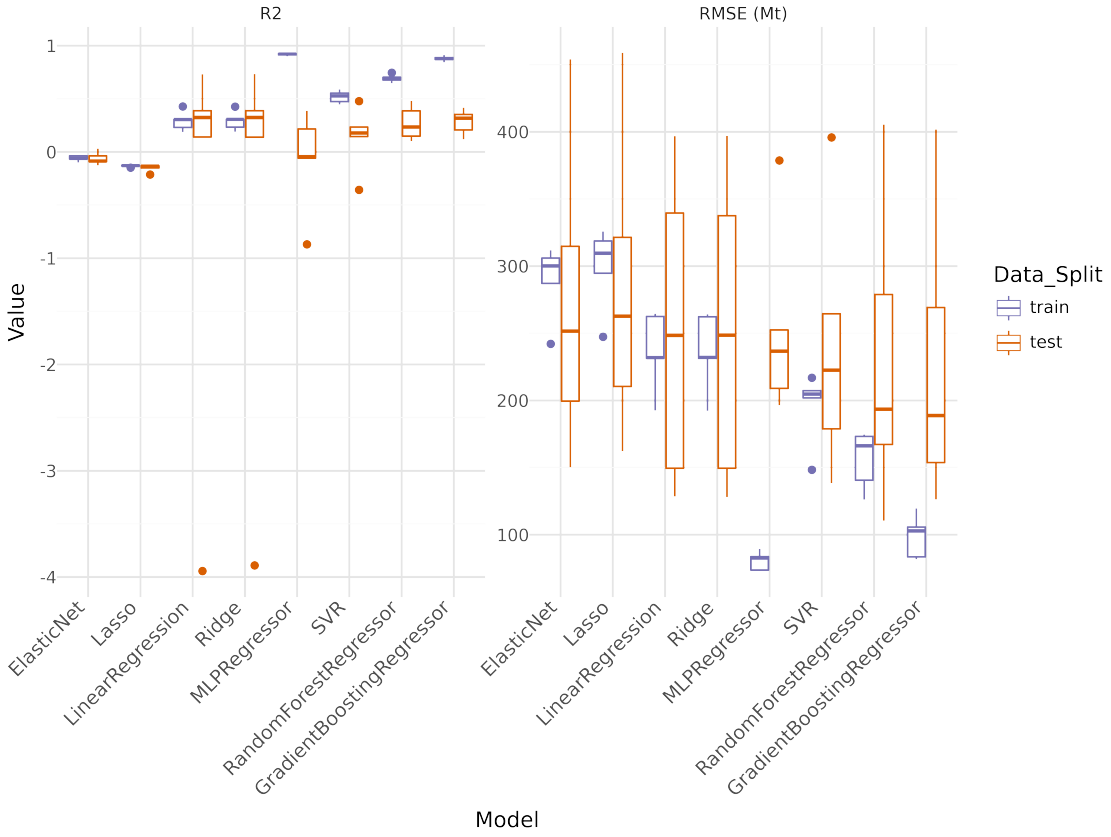

Figure 33: Training loop results for COP prediction. The figure presents four performance metrics—Mean Absolute Error (MAE), Normalized Root Mean Squared Error (NRMSE), R^2, and Root Mean Squared Error (RMSE)—evaluated across various preselected models. Different colors distinguish between training and test iterations, while boxplots summarize the outcomes from a five-fold cross-validation.

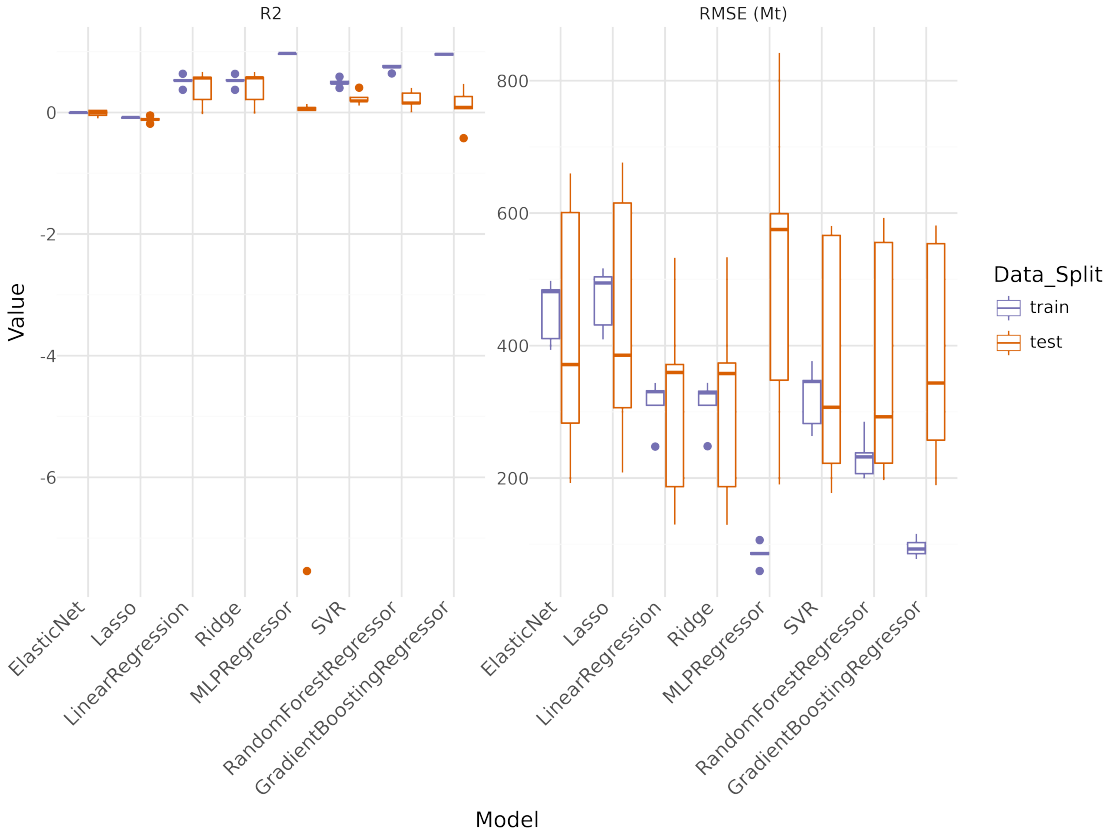

Figure 34: Training loop results for CTP prediction. The figure presents four performance metrics—Mean Absolute Error (MAE), Normalized Root Mean Squared Error (NRMSE), R^2, and Root Mean Squared Error (RMSE)—evaluated across various preselected models. Different colors distinguish between training and test iterations, while boxplots summarize the outcomes from a five-fold cross-validation.

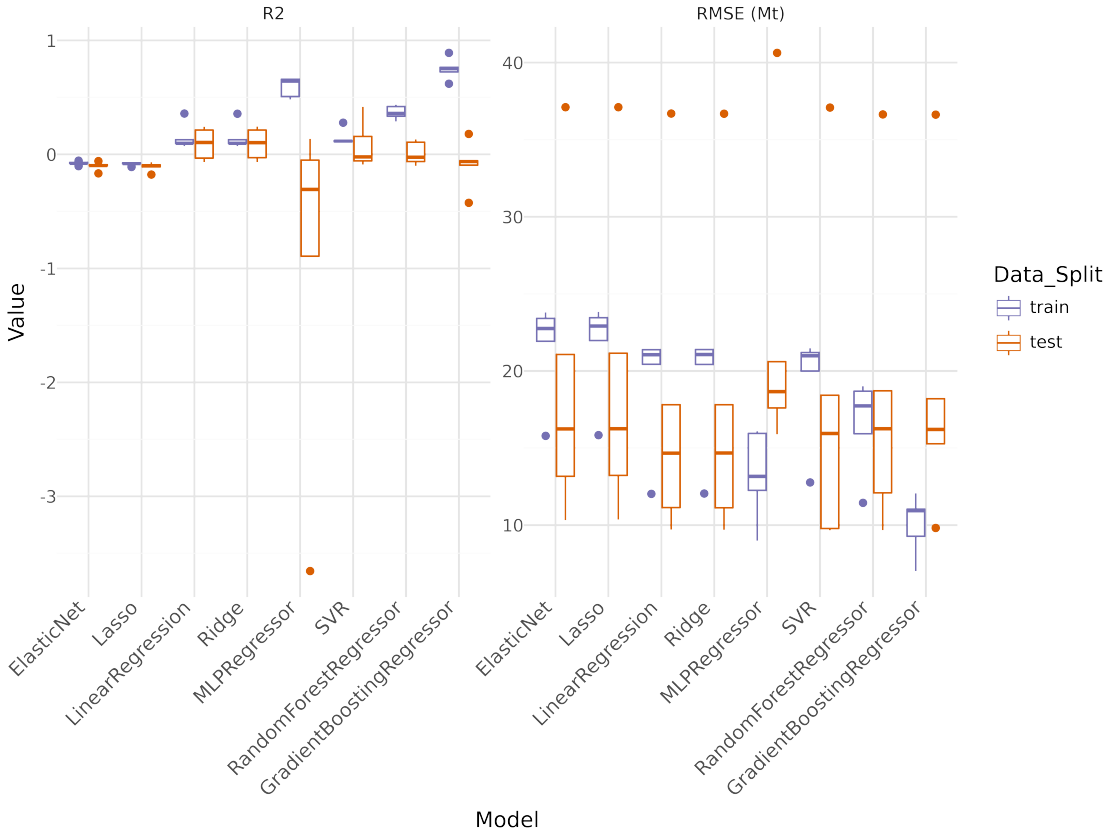

Figure 35: Training loop results for CCP prediction. The figure presents four performance metrics—Mean Absolute Error (MAE), Normalized Root Mean Squared Error (NRMSE), R^2, and Root Mean Squared Error (RMSE)—evaluated across various preselected models. Different colors distinguish between training and test iterations, while boxplots summarize the outcomes from a five-fold cross-validation.

C.6 Hyperparameter optimization

C.7 Error Analysis

Table 10: Tailings Production Error Analysis

Metric	**Train**	**Test**
R^2	0.4700	0.5854
RMSE	3.56×10^8	1.72×10^8
Shapiro p-value	6.78×10^{-23}	7.67×10^{-8}
Durbin-Watson	1.9782	1.9803
Breusch-Pagan Test		
LM Statistic	69.41	38.88
LM-Test p-value	0.7187	1.0000
F-Statistic	1.3284	0.4709
F-Test p-value	0.1037	0.8469

Table 11: Concentrate Production Error Analysis

Metric	**Train**	**Test**
R^2	0.1115	0.0917
RMSE	2.15×10^7	1.04×10^7
Shapiro p-value	2.23×10^{-25}	1.92×10^{-8}
Durbin-Watson	1.5786	1.8343
Breusch-Pagan Test		
LM Statistic	36.02	47.00
LM-Test p-value	1.0000	0.9979
F-Statistic	0.5394	12066.57
F-Test p-value	0.9951	8.29×10^{-5}

Table 12: Ore Processed Mass Error Analysis

Metric	**Train**	**Test**
R^2	0.9470	0.8988
RMSE	5.90×10^7	1.16×10^8
Shapiro p-value	6.78×10^{-23}	3.82×10^{-15}
Durbin-Watson	1.9568	1.7647
Breusch-Pagan Test		
LM Statistic	82.18	53.11
LM-Test p-value	0.3512	0.9861
F-Statistic	1.4401	0.7306
F-Test p-value	0.0314	0.7752

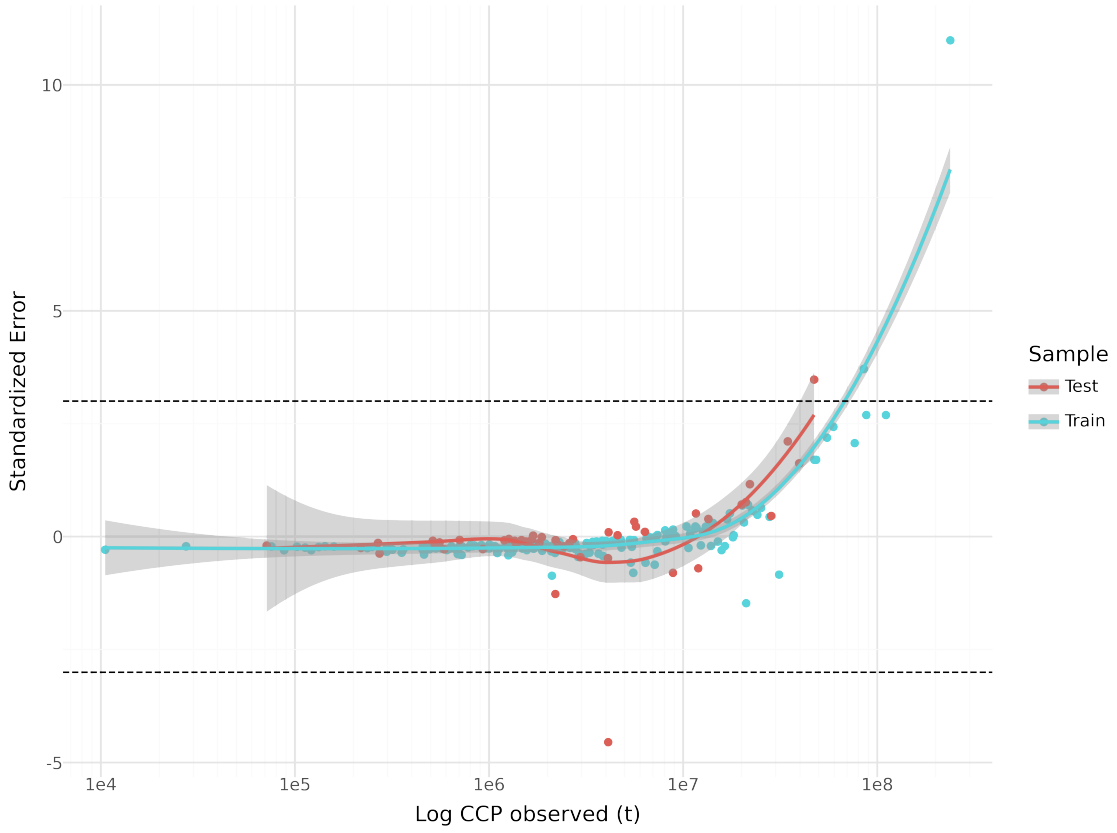

Figure 36: This figure shows the standardized errors vs the Logarithmic transformed CCP. The horizontal lines indicate the third standarddeviation from the mean. The standardized errors follow an exponential trend.

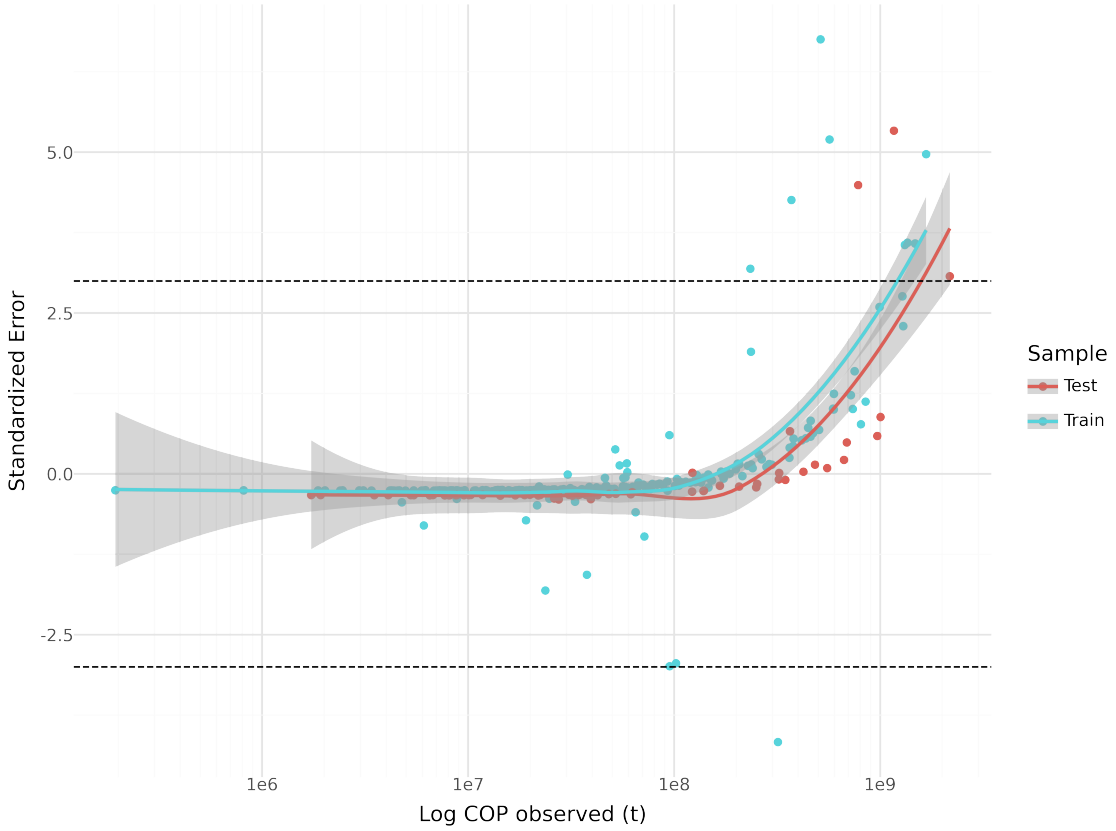

Figure 37: This figure shows the standardized errors vs the Logarithmic transformed CTP. The horizontal lines indicate the third standarddeviation from the mean. The standardized errors follow an exponential trend.

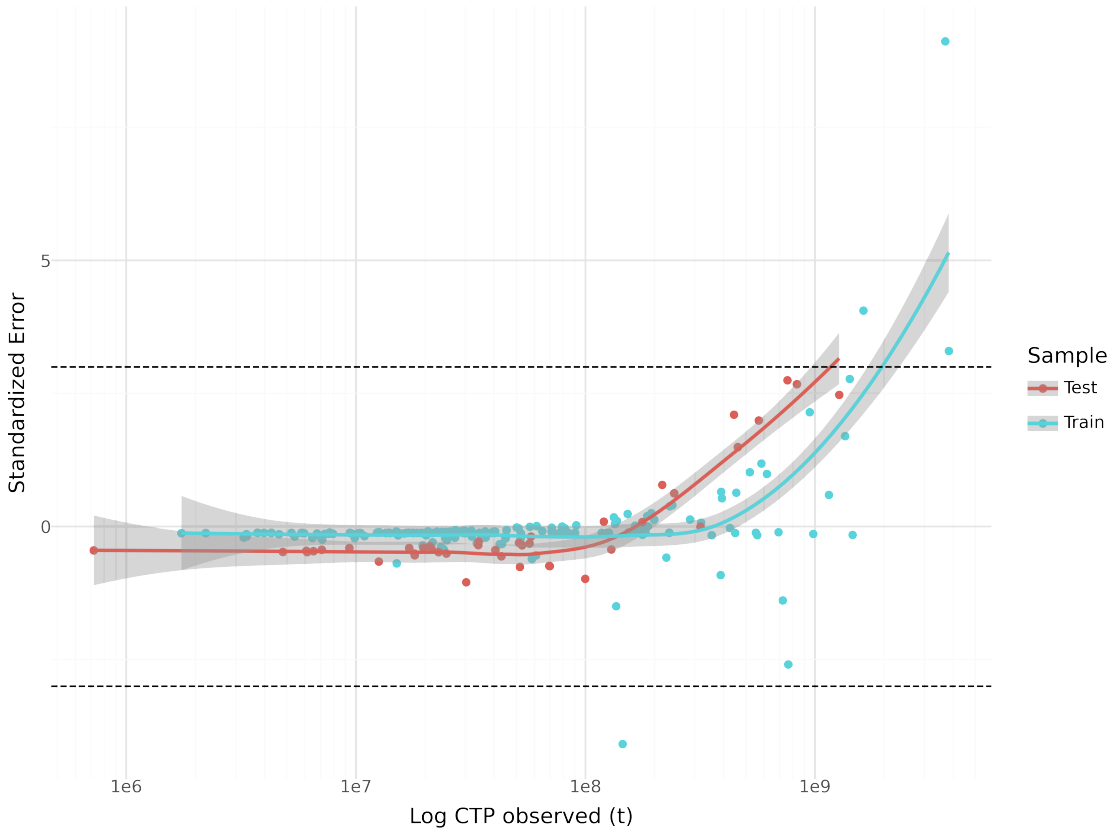

Figure 38: This figure shows the standardized errors vs the Logarithmic transformed **TCP!**. The horizontal lines indicate the third standarddeviation from the mean. The standardized errors follow an exponential trend.

C.8 Data augmentation

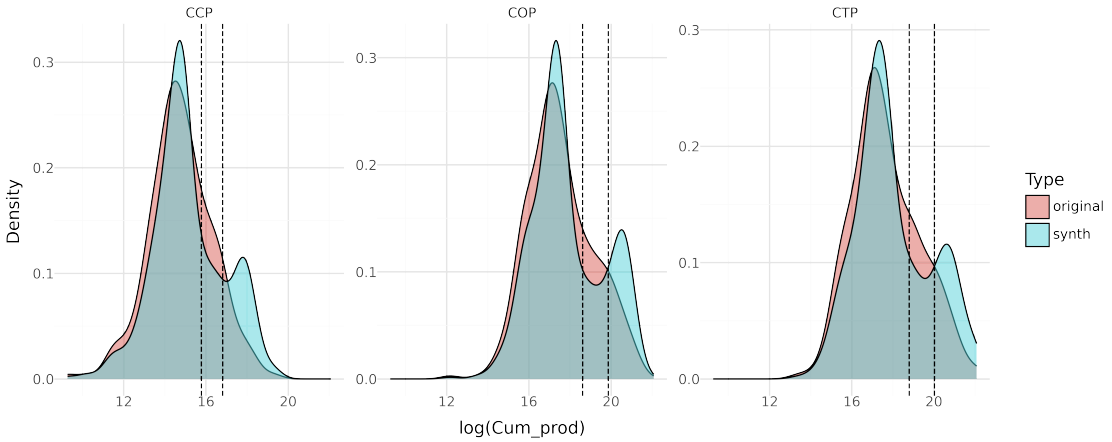

Figure 39: This figure shows the distribution of the target variable COP, CCP and CTP before and after transformation using the SMOGN algorithm. The distributions represent the original (red) distribution of the target vector and the synthetic one in (blue). The right shift of the synthetic distribution shows that the sample includes now more samples in this region. The two verticle dotted lines show the 75% (left) and the 90% (right) quantile.

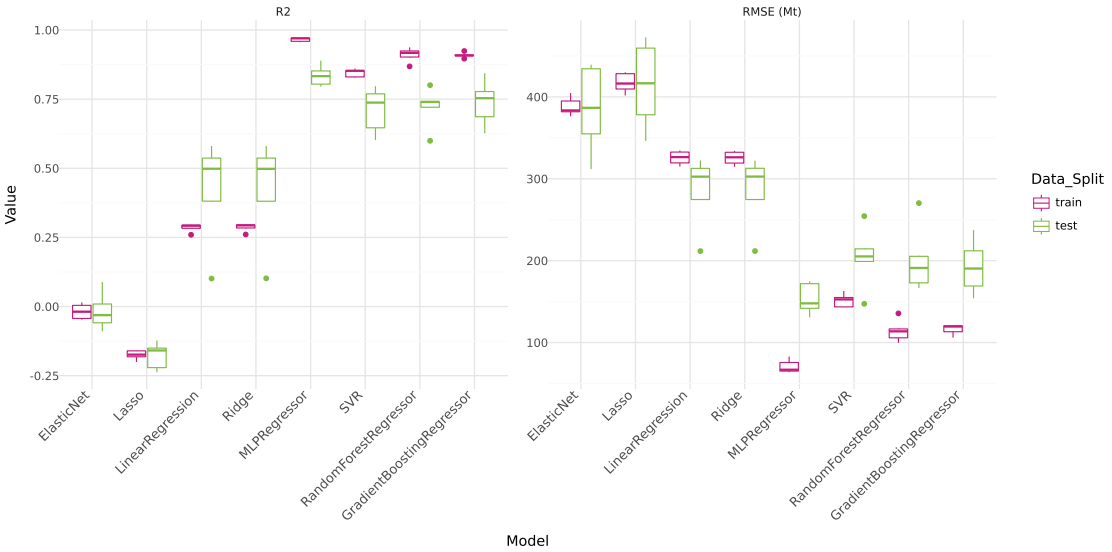

Figure 40: Training loop results for COP prediction on the synthetic dataset. The figure presents two performance metrics R2, and Root Mean Squared Error (RMSE)—evaluated across various preselected models with five fold crossvalidation. Different colors distinguish between training and test iterations, while boxplots summarize the outcomes from a five-fold cross-validation.

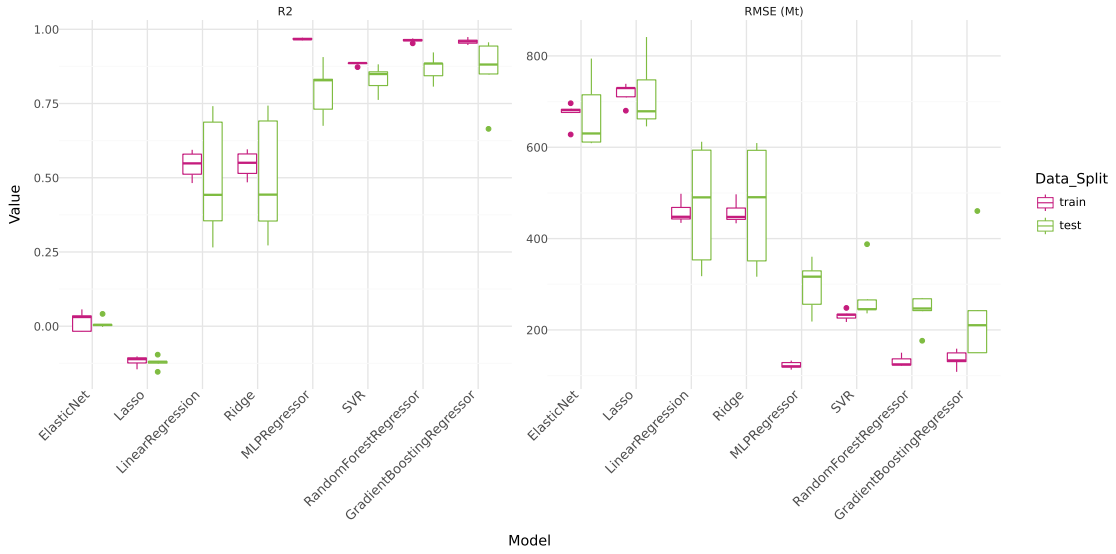

Figure 41: Training loop results for CTP prediction on the synthetic dataset. The figure presents two performance metrics R2, and Root Mean Squared Error (RMSE)—evaluated across various preselected models with five fold crossvalidatio. Different colors distinguish between training and test iterations, while boxplots summarize the outcomes from a five-fold cross-validation.

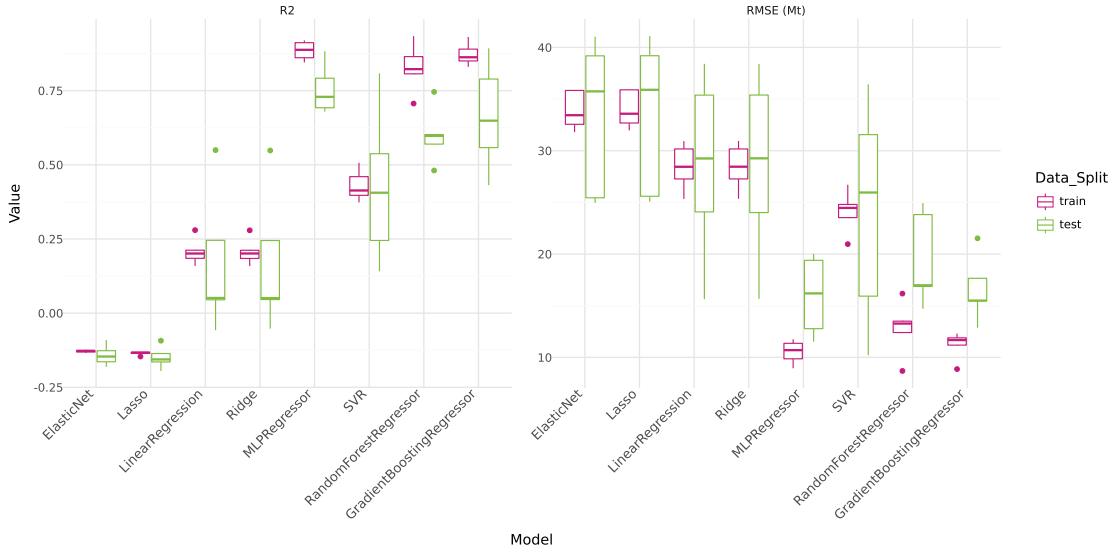

Figure 42: Training loop results for CCP prediction on the synthetic dataset. The figure presents two performance metrics R2, and Root Mean Squared Error (RMSE)—evaluated across various preselected models with five fold crossvalidatio. Different colors distinguish between training and test iterations, while boxplots summarize the outcomes from a five-fold cross-validation.

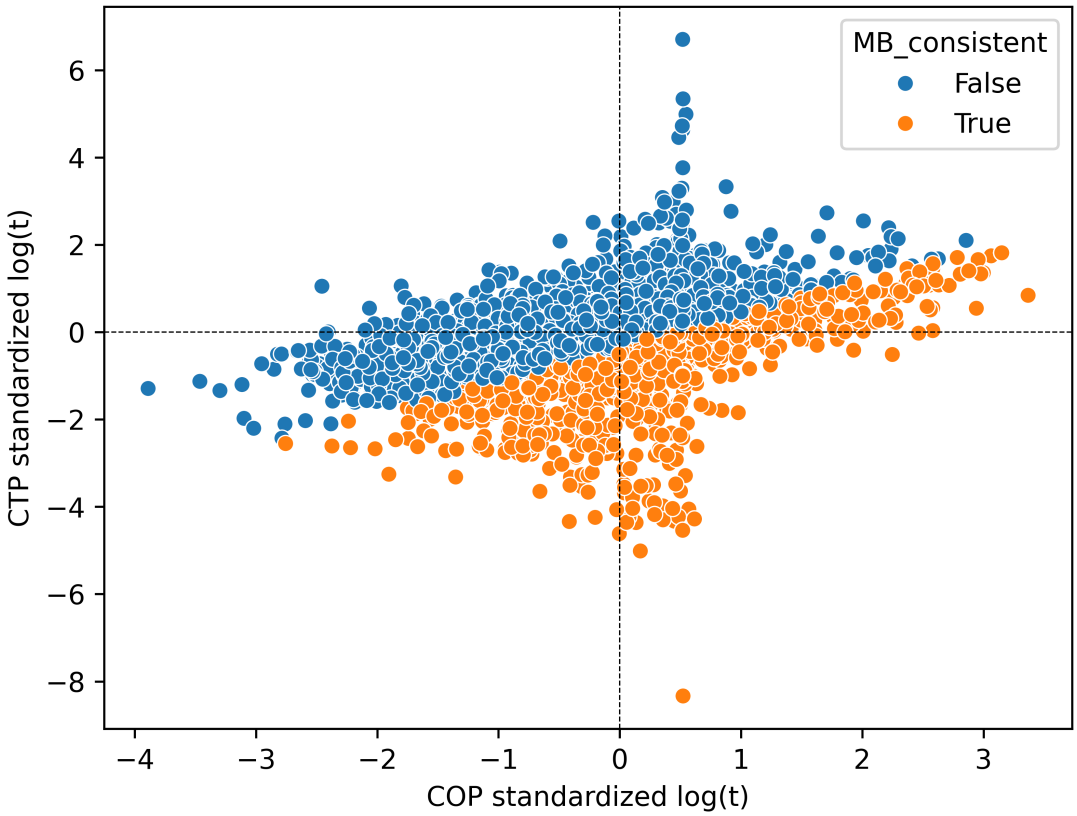

Figure 43: Consistency check for the prediction of COP and CTP. Instances that show a higher **CTP** were colored blue to indicate mass balance inconsistency.

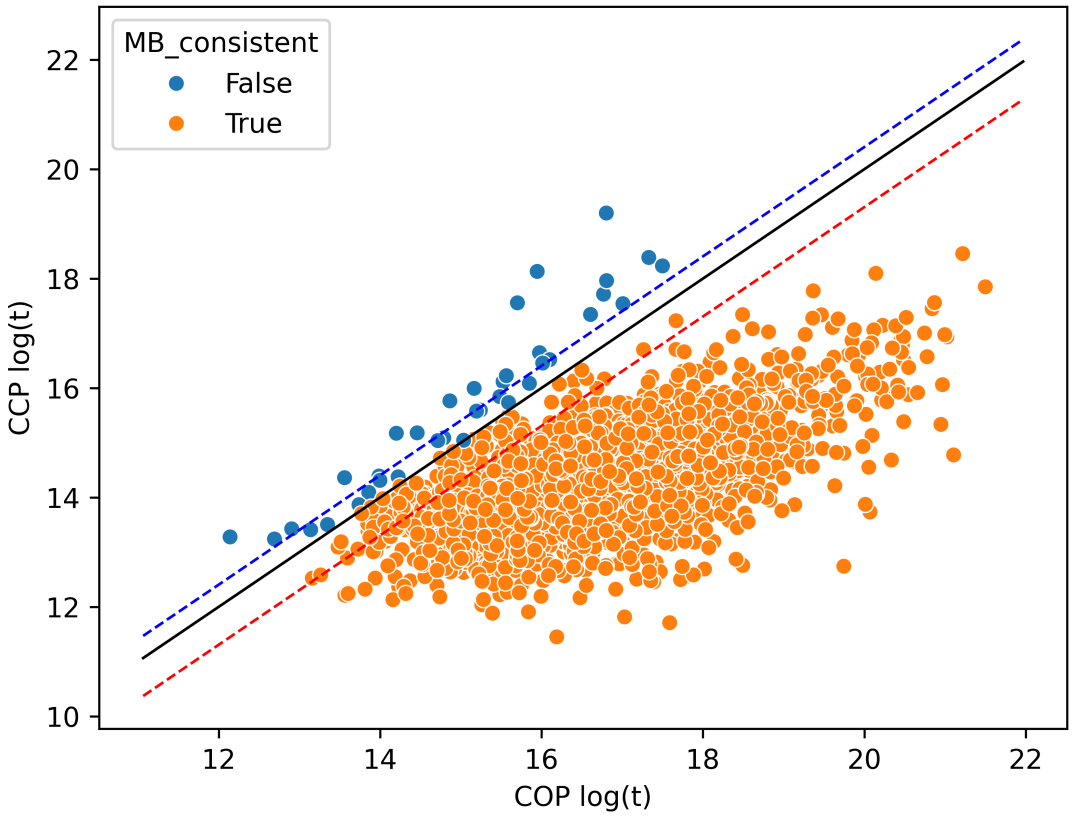

Figure 44: Consistency check for the prediction of COP and CCP. Instances that show a higher **CCP** were colored blue to indicate mass balance inconsistency.

C.9 Allocation

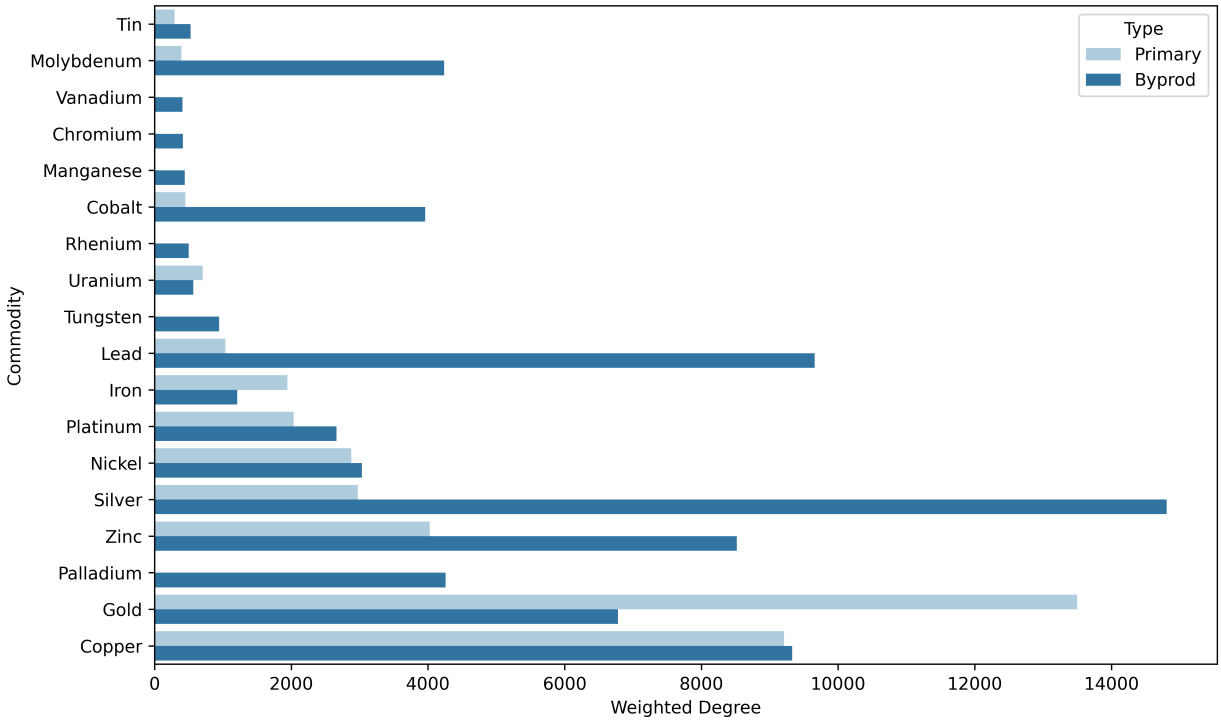

Figure 45: Weighted degree distribution of selected commodities.

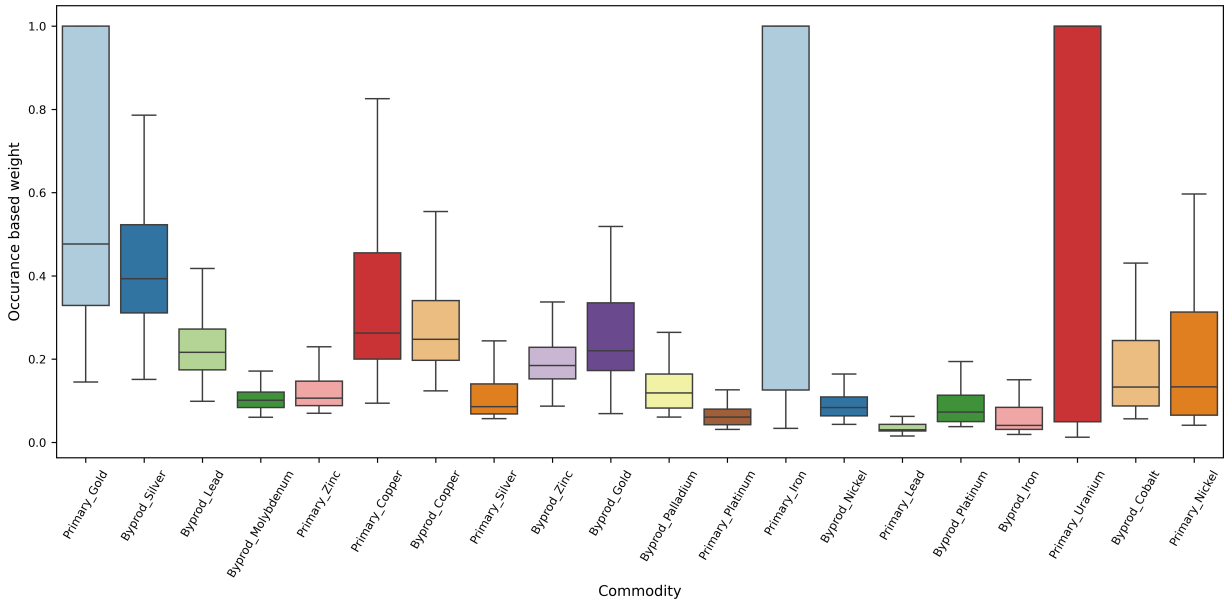

Figure 46: Boxplot illustrating the influence of allocation on commodity extraction weights.

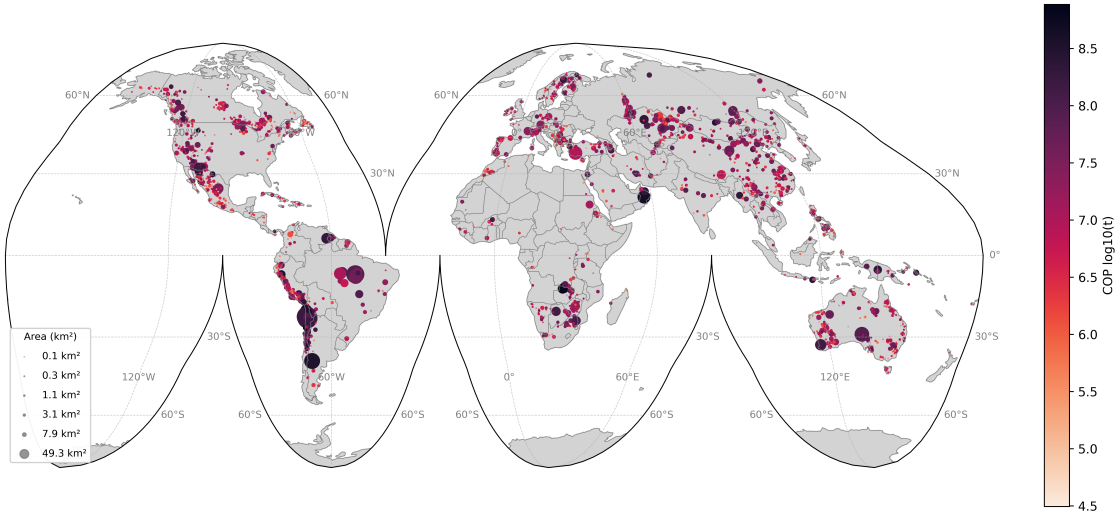

Figure 47: COP estimations for global Copper primary- and by-production for Occurance based allocation. The scatters are scaled by the area of the mining instance. The map uses interrupted Interrupted Goode Homolosine to reduce distortion around the poles and represent areas accurately.

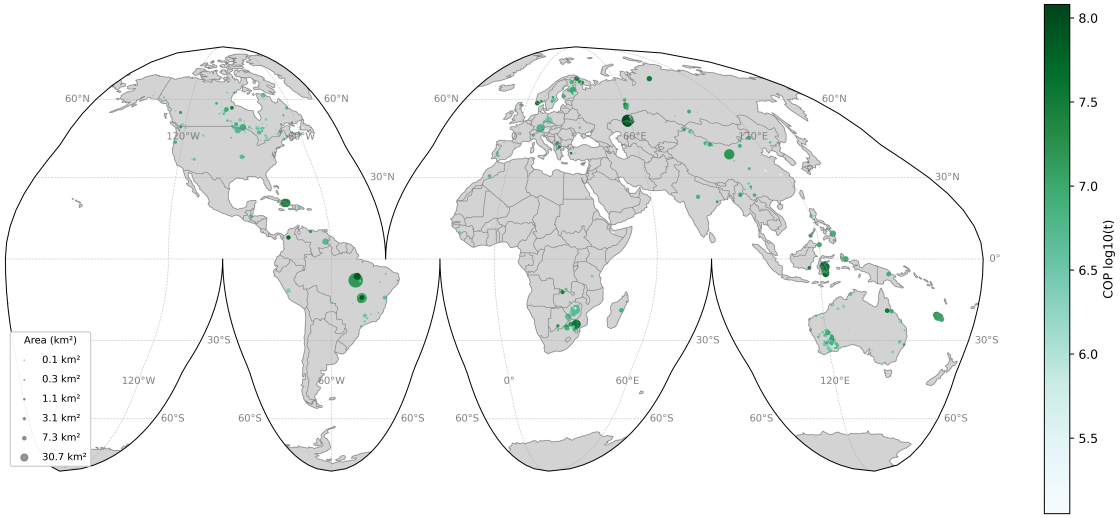

Figure 48: COP estimations for global Nickel primary- and by-production for Occurance based allocation. The scatters are scaled by the area of the mining instance. The map uses interrupted Interrupted Goode Homolosine to reduce distortion around the poles and represent areas accurately.

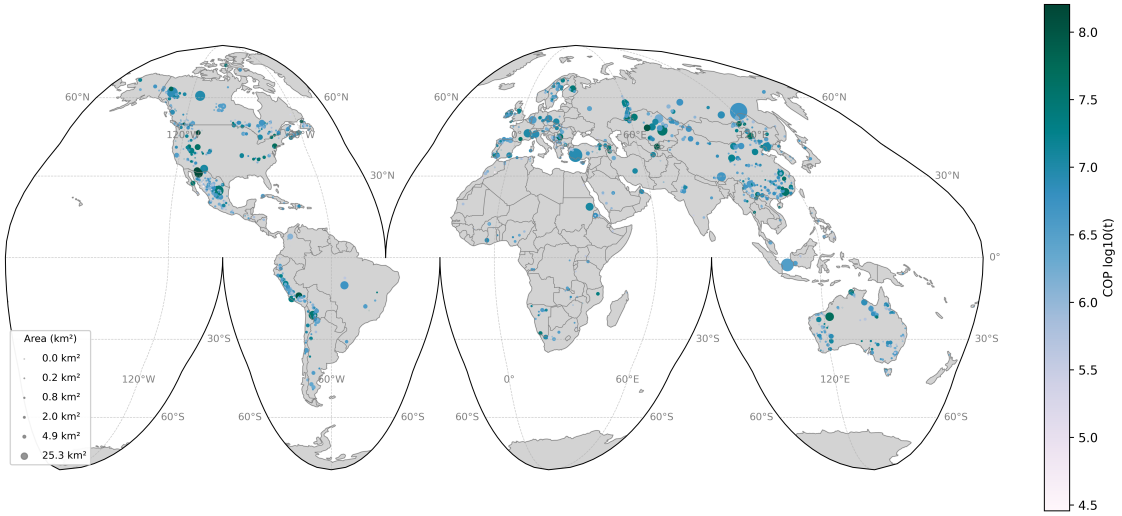

Figure 49: COP estimations for global Zinc primary- and by-production for Occurance based allocation. The scatters are scaled by the area of the mining instance. The map uses interrupted Interrupted Goode Homolosine to reduce distortion around the poles and represent areas accurately.

References

Althaus, H.-J., & Classen, M. (2005). Life Cycle Inventories of Metals and Methodological Aspects of Inventorying Material Resources in ecoinvent (7 pp). *The International Journal of Life Cycle Assessment*, *10*(1), 43–49. https://doi.org/10.1065/lca2004.11.181.5

Andersen, M., & Noailly, J. (2022). Environmental Regulations in the Mining Sector and Their Effect on Technological Innovation. In A. Daly, D. Humphreys, G. Valacchi, & J. Raffo (Eds.), *Global Challenges for Innovation in Mining Industries* (pp. 142–171). Cambridge University Press. https://doi.org/10.1017/9781108904209.007

Barabási, A.-L. (2015). *Network Science*. Retrieved December 8, 2024, from http://networksciencebook.com/

Bruce, P., Bruce, A., & Gedeck, P. (2020). *Practical Statistics for Data Scientists* (Vol. 2).

Burkov, A. (2019). *The hundred-page machine learning book*. Andriy Burkov.

Calvo, G., Valero, A., & Valero, A. (2017). Assessing maximum production peak and resource availability of non-fuel mineral resources: Analyzing the influence of extractable global resources. *Resources, Conservation and Recycling*, *125*, 208–217. https://doi.org/10.1016/j.resconrec.2017.06.009

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

Chimunhu, P., Topal, E., Ajak, A. D., & Asad, W. (2022). A review of machine learning applications for underground mine planning and scheduling. *Resources Policy*, *77*, 102693. https://doi.org/10.1016/j.resourpol.2022.102693

Creutzig, F., Fernandez, B., Haberl, H., Khosla, R., Mulugetta, Y., & Seto, K. C. (2016). Beyond Technology: Demand-Side Solutions for Climate Change Mitigation. *Annual Review of Environment and Resources*, *41*(1), 173–198. https://doi.org/10.1146/annurev-environ-110615-085428

Creutzig, F., Roy, J., & Minx, J. (2024). Demand-side climate change mitigation: Where do we stand and where do we go? *Environmental Research Letters*, *19*(4), 040201. https://doi.org/10.1088/1748-9326/ad33d3

Dominy, S. C. (2002). Errors and Uncertainty in Mineral Resource and Ore Reserve Estimation: The Importance of Getting it Right. *Exploration and Mining Geology*, *11*(1-4), 77–98. https://doi.org/10.2113/11.1-4.77

Géron, A. (2023). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems* (Third edition). O'Reilly.

Giljum, S., Maus, V., Kuschnig, N., Luckeneder, S., Tost, M., Sonter, L. J., & Bebbington, A. J. (2022). A pantropical assessment of deforestation caused by industrial mining. *Proceedings of the National Academy of Sciences*, *119*(38), e2118273119. https://doi.org/10.1073/pnas.2118273119

Global Energy Monitor. (2024). Global Energy Monitor.

Greffe, T., Frenzel, M., Werner, T. T., Mudd, G., Wang, P., Margni, M., & Bulle, C. (2024). Byproduct-to-Host Ratios for Assessing the Accessibility of Mineral Resources [Publisher: American Chemical Society]. *Environmental Science & Technology*. https://doi.org/10.1021/acs.est.4c05293

Haberl, H., Schmid, M., Haas, W., Wiedenhofer, D., Rau, H., & Winiwarter, V. (2021). Stocks, flows, services and practices: Nexus approaches to sustainable social metabolism. *Ecological Economics*, *182*, 106949. https://doi.org/10.1016/j.ecolecon.2021.106949

Harrison, R., Granja, C., & Leroy, C. (2010). Introduction to Monte Carlo Simulation, 17–21. https://doi.org/10.1063/1.3295638

Hartmann, J., & Moosdorf, N. (2012). The new global lithological map database GLiM: A representation of rock properties at the Earth surface. *Geochemistry, Geophysics, Geosystems*, *13*(12), 2012GC004370. https://doi.org/10.1029/2012GC004370

Heijungs, R., & Suh, S. (2002). *The Computational Structure of Life Cycle Assessment* (A. Tukker, Ed.; Vol. 11). Springer Netherlands. https://doi.org/10.1007/978-94-015-9900-9

IEA. (2023). *Sustainable and Responsible Critical Mineral Supply Chains* (tech. rep.).

IEA. (2024). *Global Critical Minerals Outlook 2024* (tech. rep.).

IPCC. (2023, July). *Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (tech. rep.) (doi: 10.59327/IPCC/AR6-9789291691647.001). Intergovernmental Panel on Climate Change (IPCC). Retrieved March 29, 2024, from https://www.ipcc.ch/report/ar6/syr/

Islam, K., & Murakami, S. (2021). Global-scale impact analysis of mine tailings dam failures: 1915–2020. *Global Environmental Change*, *70*, 102361. https://doi.org/10.1016/j.gloenvcha.2021.102361

Jasansky, S., Lieber, M., Giljum, S., & Maus, V. (2023). An open database on global coal and metal mine production [Number: 1 Publisher: Nature Publishing Group]. *Scientific Data*, *10*(1), 52. https://doi.org/10.1038/s41597-023-01965-y

Jung, D., & Choi, Y. (2021). Systematic Review of Machine Learning Applications in Mining: Exploration, Exploitation, and Reclamation. *Minerals*, *11*(2), 148. https://doi.org/10.3390/min11020148

Karl, N. A., Mauk, J. L., Reyes, T. A., & Scott, P. C. (2019). Lithium Deposits in the United States. https://doi.org/10.5066/P9ZKRWQF

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., & Yang, L. (2021). Physics-informed machine learning [Publisher: Nature Publishing Group]. *Nature Reviews Physics*, *3*(6), 422–440. https://doi.org/10.1038/s42254-021-00314-5

Kruse, T., Dechezleprêtre, A., Saffar, R., & Robert, L. (2022, March). *Measuring environmental policy stringency in OECD countries: An update of the OECD composite EPS indicator* (OECD Economics Department Working Papers No. 1703) (Series: OECD Economics Department Working Papers Volume: 1703). https://doi.org/10.1787/90ab82e8-en

Macklin, M. G., Thomas, C. J., Mudbhatkal, A., Brewer, P. A., Hudson-Edwards, K. A., Lewin, J., Scussolini, P., Eilander, D., Lechner, A., Owen, J., Bird, G., Kemp, D., & Mangalaa, K. R. (2023). Impacts of metal mining on river systems: A global assessment. *Science*, *381*(6664), 1345–1350. https://doi.org/10.1126/science.adg6704

Maus, V., Giljum, S., da Silva, D. M., Gutschlhofer, J., da Rosa, R. P., Luckeneder, S., Gass, S. L. B., Lieber, M., & McCallum, I. (2022). An update on global mining land use [Number: 1 Publisher: Nature Publishing Group]. *Scientific Data*, *9*(1), 433. https://doi.org/10.1038/s41597-022-01547-4

Maus, V., Giljum, S., Gutschlhofer, J., da Silva, D. M., Probst, M., Gass, S. L. B., Luckeneder, S., Lieber, M., & McCallum, I. (2020). A global-scale data set of mining areas [Number: 1 Publisher: Nature Publishing Group]. *Scientific Data*, *7*(1), 289. https://doi.org/10.1038/s41597-020-00624-w

Maus, V., & Werner, T. T. (2024). Impacts for half of the world's mining areas are undocumented [Bandiera_abtest: a Cg_type: Comment Number: 7993 Publisher: Nature Publishing Group Subject_term: Environmental sciences, Policy, Industry, Sustainability]. *Nature*, *625*(7993), 26–29. https://doi.org/10.1038/d41586-023-04090-3

Meißner, S. (2021). The Impact of Metal Mining on Global Water Stress and Regional Carrying Capacities—A GIS-Based Water Impact Assessment [Number: 12 Publisher: Multidisciplinary Digital Publishing Institute]. *Resources*, *10*(12), 120. https://doi.org/10.3390/resources10120120

Messner, F. (2002). Material substitution and path dependence: Empirical evidence on the substitution of copper for aluminum. *Ecological Economics*, *42*(1), 259–271. https://doi.org/10.1016/S0921-8009(02)00052-6

Miller, H. J. (2004). Tobler's First Law and Spatial Analysis [Publisher: Routledge _eprint: https://doi.org/10.1111/j.1467-8306.2004.09402005.x]. *Annals of the Association of American Geographers*, *94*(2), 284–289. https://doi.org/10.1111/j.1467-8306.2004.09402005.x

Mudd, G. (2007). Sustainability and Mine Waste Management – A Snapshot of Mining Waste Issues. *Waste Management*.

Mudd, G. (2010). The Environmental sustainability of mining in Australia: Key mega-trends and looming constraints. *Resources Policy*, *35*(2), 98–115. https://doi.org/10.1016/j.resourpol.2009.12.001

Nassar, N. T., Graedel, T. E., & Harper, E. M. (2015). By-product metals are technologically essential but have problematic supply. *Science Advances*, *1*(3), e1400180. https://doi.org/10.1126/sciadv.1400180

Nassar, N. T., Lederer, G. W., Brainard, J. L., Padilla, A. J., & Lessard, J. D. (2022). Rock-to-Metal Ratio: A Foundational Metric for Understanding Mine Wastes [Publisher: American Chemical Society]. *Environmental Science & Technology*, *56*(10), 6710–6721. https://doi.org/10.1021/acs.est.1c07875

Nassar, N. T., Lederer, G. W., Padilla, A. J., Gambogi, J., Cordier, D. J., Brainard, J. L., Lessard, J. D., & Charab, R. (2023). Rock-to-metal ratios of the rare earth elements. *Journal of Cleaner Production*, *405*, 136958. https://doi.org/10.1016/j.jclepro.2023.136958

Nedal T. Nassar, Nassar, N. T., Graham W. Lederer, Graham W Lederer, Brainard, J., Jamie Brainard, Brainard, J. L., A. J. Padilla, Abraham J Padilla, Joseph D. Lessard, & Joseph D Lessard. (2022). Rock-to-Metal Ratio: A Foundational Metric for Understanding Mine Wastes. [MAG ID: 4224657280 S2ID: a55f167b0651bdbdbbae14a9368d864e2f8a257c]. *Environmental Science & Technology*. https://doi.org/10.1021/acs.est.1c07875

Nijnens, J., Behrens, P., Kraan, O., Sprecher, B., & Kleijn, R. (2023). Energy transition will require substantially less mining than the current fossil system. *Joule*, *7*(11), 2408–2413. https://doi.org/10.1016/j.joule.2023.10.005

OpenStreetMap contributors. (2024). OpenStreetMap. https://www.openstreetmap.org

Osanloo, M., Gholamnejad, J., & Karimi, B. (2008). Long-term open pit mine production planning: A review of models and algorithms. *International Journal of Mining, Reclamation and Environment*, *22*(1), 3–35. https://doi.org/10.1080/17480930601118947

Owen, J., Kemp, D., Lechner, A., Ang, M., Lèbre, E., Mudd, G., Macklin, M., Saputra, U., Witra, T., & Bebbington, A. (2023). Increasing mine waste will induce land cover change that results in ecological degradation and human displacement. *Journal of environmental management*, *351*, 119691. https://doi.org/10.1016/j.jenvman.2023.119691

Owen, J. R., Kemp, D., Lechner, A. M., Harris, J., Zhang, R., & Lèbre, É. (2023). Energy transition minerals and their intersection with land-connected peoples [Number: 2 Publisher: Nature Publishing Group]. *Nature Sustainability*, *6*(2), 203–211. https://doi.org/10.1038/s41893-022-00994-6

Pauliuk, S., Majeau-Bettez, G., & Müller, D. B. (2015). A General System Structure and Accounting Framework for Socioeconomic Metabolism. *Journal of Industrial Ecology*, *19*(5), 728–741. https://doi.org/10.1111/jiec.12306

Radetzki, M. (2009). Seven thousand years in the service of humanity—the history of copper, the red metal. *Resources Policy*, *34*(4), 176–184. https://doi.org/10.1016/j.resourpol.2009.03.003

Rey, S. J., Arribas-Bel, D., & Wolf, L. J. (2023). *Geographic data science with Python* (First edition). CRC Press.

Richardson, K., Steffen, W., Lucht, W., Bendtsen, J., Cornell, S. E., Donges, J. F., Drüke, M., Fetzer, I., Bala, G., Von Bloh, W., Feulner, G., Fiedler, S., Gerten, D., Gleeson, T., Hofmann, M., Huiskamp, W., Kummu, M., Mohan, C., Nogués-Bravo, D., ... Rockström, J. (2023). Earth beyond six of nine planetary boundaries. *Science Advances*, *9*(37), eadh2458. https://doi.org/10.1126/sciadv.adh2458

Ridley, J. (2013). *Ore deposit geology*. Cambridge university press.

Rodríguez, S. P. (2013). Methodology and Equations of Mineral Production Forecast—Part I. Crude Oil in the UK and Gold in Nevada, USA. Prediction of Late Stages of Production [Number: 5 Publisher: Scientific Research Publishing]. *Open Journal of Geology*, *3*(5), 352–360. https://doi.org/10.4236/ojg.2013.35040

Schandl, H., Fischer-Kowalski, M., West, J., Giljum, S., Dittrich, M., Eisenmenger, N., Geschke, A., Lieber, M., Wieland, H., Schaffartzik, A., Krausmann, F., Gierlinger, S., Hosking, K., Lenzen, M., Tanikawa, H., Miatto, A., & Fishman, T. (2018). Global Material Flows and Resource Productivity: Forty Years of Evidence [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jiec.12626]. *Journal of Industrial Ecology*, *22*(4), 827–838. https://doi.org/10.1111/jiec.12626

Singer, D. A. (2017). Future copper resources. *Ore Geology Reviews*, *86*, 271–279. https://doi.org/10.1016/j.oregeorev.2017.02.022

Sonter, L. J., Ali, S. H., & Watson, J. E. M. (2018). Mining and biodiversity: Key issues and research needs in conservation science [MAG ID: 2902503737]. *Proceedings of The Royal Society B: Biological Sciences*, *285*(1892), 20181926. https://doi.org/10.1098/rspb.2018.1926

Sonter, L. J., Maron, M., Bull, J. W., Giljum, S., Luckeneder, S., Maus, V., McDonald-Madden, E., Northey, S. A., Sánchez, L. E., Valenta, R., Visconti, P., Werner, T. T., & Watson, J. E. M. (2023). How to fuel an energy transition with ecologically responsible mining. *Proceedings of the National Academy of Sciences*, *120*(35), e2307006120. https://doi.org/10.1073/pnas.2307006120

Standard & Poor. (2025, January). S&P Capital IQ Pro database. Retrieved January 2, 2025, from https://www.capitaliq.spglobal.com/web/client?auth=inherit&OktaLogin=true#dashboard/metalsAndMining

Sun, Z., Tukker, A., & Behrens, P. (2019). Going Global to Local: Connecting Top-Down Accounting and Local Impacts, A Methodological Review of Spatially Explicit Input–Output Approaches [Publisher: American Chemical Society]. *Environmental Science & Technology*, *53*(3), 1048–1062. https://doi.org/10.1021/acs.est.8b03148

Tang, L., & Werner, T. T. (2023). Global mining footprint mapped from high-resolution satellite imagery [Publisher: Zenodo]. https://doi.org/10.5281/ZENODO.6806816

Taylor, J. R. (1997). *An introduction to error analysis: The study of uncertainties in physical measurements* (2. ed). University Science Books.

United Nations Environment Programme. (2024). *Global Resources Outlook 2024: Bend the Trend – Pathways to a liveable planet as resource use spikes* (tech. rep.). International Resource Panel. Nairobi. https://wedocs.unep.org/20.500.11822/44901

Valero, A., Domínguez, A., & Valero, A. (2015). Exergy cost allocation of by-products in the mining and metallurgical industry. *Resources, Conservation and Recycling*, *102*, 128–142. https://doi.org/10.1016/j.resconrec.2015.04.012

Watari, T., Nansai, K., & Nakajima, K. (2021). Major metals demand, supply, and environmental impacts to 2100: A critical review. *Resources, Conservation and Recycling*, *164*, 105107. https://doi.org/10.1016/j.resconrec.2020.105107

Wellmer, F. .-., & Scholz, R. W. (2017). Peak minerals: What can we learn from the history of mineral economics and the cases of gold and phosphorus? *Mineral Economics*, *30*(2), 73–93. https://doi.org/10.1007/s13563-016-0094-3

Werner, T. T., Bach, P. M., Yellishetty, M., Amirpoorsaeed, F., Walsh, S. D., Miller, A., Roach, M., Schnapp, A., Solly, P., Tan, Y., Lewis, C., Hudson, E., Heberling, K., Richards, T., Chia, H. C., Truong, M., Gupta, T., Gupta, T., & Wu, X. (2020). A Geospatial Database for Effective Mine Rehabilitation in Australia [MAG ID: 3081412499]. *Minerals*, *10*(9), 1–21. https://doi.org/10.3390/min10090745

Werner, T. T., Bebbington, A., & Gregory, G. (2019). Assessing impacts of mining: Recent contributions from GIS and remote sensing [MAG ID: 2966257124]. *The Extractive Industries and Society*, *6*(3), 993–1012. https://doi.org/10.1016/j.exis.2019.06.011

Werner, T. T., Mudd, G., Schipper, A. M., Huijbregts, M. A. J., Taneja, L., & Northey, S. (2020). Global-scale remote sensing of mine areas and analysis of factors explaining their extent [MAG ID: 2988983401]. *Global Environmental Change-human and Policy Dimensions*, *60*, 102007. https://doi.org/10.1016/j.gloenvcha.2019.102007

Wiedenhofer, D., Streeck, J., Wiese, F., Verdolini, E., Mastrucci, A., Ju, Y., Boza-Kiss, B., Min, J., Norman, J., Wieland, H., Bento, N., León, M. F. G., Magalar, L., Mayer, A., Gingrich, S., Hayashi, A., Jupesta, J., Ünlü, G., Niamir, L., ... Pauliuk, S. (2024). Industry Transformations for High Service Provisioning with Lower Energy and Material Demand: A Review of Models and Scenarios. *Annual Review of Environment and Resources*, *49*(1), 249–279. https://doi.org/10.1146/annurev-environ-110822-044428

Wieland, H., Lenzen, M., Geschke, A., Fry, J., Wiedenhofer, D., Eisenmenger, N., Schenk, J., & Giljum, S. (2022). The PIOLab: Building global physical input–output tables in a virtual laboratory [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jiec.13215]. *Journal of Industrial Ecology*, *26*(3), 683–703. https://doi.org/10.1111/jiec.13215

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

World Food Programme. (2019). World Administrative Boundaries - Countries and Territories. Retrieved February 20, 2024, from https://public.opendatasoft.com/explore/dataset/world-administrative-boundaries/export/

Xia, J., She, D., Zhang, Y., & Du, H. (2012). Spatio-temporal trend and statistical distribution of extreme precipitation events in Huaihe River Basin during 1960–2009. *Journal of Geographical Sciences*, *22*(2), 195–208. https://doi.org/10.1007/s11442-012-0921-6

Xiao, J., Werner, T. T., Komai, T., & Matsubae, K. (2025). Assessing the Relationship Between Production and Land Transformation for Chilean Copper Mines Using Satellite and Operational Data [Number: 2 Publisher: Multidisciplinary Digital Publishing Institute]. *Resources*, *14*(2), 25. https://doi.org/10.3390/resources14020025