processes

Article

Data-Driven Synthesis of a Geometallurgical Model for a Copper Deposit

Yuyang Mu [1,2,†] and Juan Carlos Salas [1,*,†]

[1] Department of Mining Engineering, Pontificia Universidad Católica de Chile, Av. Vicuña Mackenna 4860, Santiago 7820436, Chile; ymu@uc.cl
[2] Zijin (Xiamen) Engineering Co., Ltd., 20th Floor, Block B, Haifu Center, No. 599 Sishui Road, Huli District, Xiamen 361000, China
* Correspondence: jcsalasm@uc.cl
† These authors contributed equally to this work.

Abstract: Geometallurgy integrates aspects of geology, metallurgy, and mine planning in order to improve decision making in mining schedules. A geometallurgical model is a 3D space that is typically synthesized from early-stage small-scale samples and is composed of several metallurgical units, or domains. This work explores the synthesis of a geometallurgical model for a copper deposit using a purely data-driven unsupervised approach. To this end, a dataset of 1112 drill samples is used, which are clustered using different methods, namely, *k*-means, hierarchical clustering (AGG), self-organizing maps (SOM), and DBSCAN. Two cluster validity indices (Silhouette and Calinski–Harabasz) are used to select the final model. To validate the potential of the proposed approach, a simulated economic evaluation is conducted. Results demonstrate that *k*-means exhibits a better performance in terms of modeling and that using the obtained geometallurgical model for mining scheduling increases the project's Net Present Value (NPV) by as much as 4%. Based on these results, the proposed methodology is an appealing alternative for generating geometallurgical models within greenfield, brownfield and ongoing operations.

Keywords: geometallurgy; machine learning; unsupervised learning; cluster analysis; copper deposit

Citation: Mu, Y.; Salas, J.C. Data-Driven Synthesis of a Geometallurgical Model for a Copper Deposit. *Processes* **2023**, *11*, 1775. https://doi.org/10.3390/pr11061775

Academic Editors: Alessandro Navarra, Norman Toro, Roberto Parra and Henrik Saxen

Received: 26 April 2023
Revised: 30 May 2023
Accepted: 5 June 2023
Published: 10 June 2023

1. Introduction

Geometallurgy, a concept defined around 1970, incorporates geological, metallurgical, and mine planning information to improve decision making in mining projects [1–4]. The notion of geometallurgy has modernized lately and currently incorporates the mine-to-mill concept and environmental variables, providing an even more complete view of the mining value chain [5,6].

A geometallurgical model is a 3D spatial volume formed by a set of geometallurgical units or domains. According to Rajabinasa and Asghari [7], each geometallurgical unit is a 3D spatial section of a mine body with similar mining and metallurgical characteristics. A block model can be established from the geometallurgical model in order to aid and improve mine planning decisions and ore processing throughout the life of the mine. Geometallurgical models should be created with combined information from geology, metallurgy, mine planning, the environment, and other variables that relate to the product, subproducts and effects on the operation's sustainability [7–12].

The current practice for geometallurgical modeling is based on the combination of a geological model and metallurgical data. Usually, data stem from metallurgical testing involving samples from geological domains and laboratory tests. As an example, Suazo et al. [13] defines the geometallurgical unit in two stages, firstly grouping the units based on geological similarities, and secondly by incorporating flotation kinetics obtained from laboratory testing. Recently, linear regression and supervised learning have been used

to predict metallurgical variables [6,8,9,14], trying to avoid the expensive and time consuming laboratory testing stage; yet, the idea that geometallurgical modeling is a combination of geology and mineral processing remained unchanged.

A study by Baumgartnert et al. [15] on Canahuire deposits observed that a geology model defined by lithology, alteration, and mineralization shows similar metal recovery as the four geological domain, so sometimes the geological domain does not guarantee a direct relationship between the geological and metallurgical response, as mentioned by Rajabinasab et al. [7]. Furthermore, according to Lishchuk et al. [12], the integration of geology and mineral processing is no longer sufficient since we lack the knowledge for the optimization of the entire value chain. Instead, geometallurgy should now adopt a holistic and interdisciplinary approach involving geology, mineralogy, mineral processing, environmental and economic aspects of mining.

This work explores geometallurgical model synthesis using a purely data-driven approach. Specifically, all the information available from an early-stage mining project, including spatial, geological, processing and intrinsic variables, are used to generate a geometallurgical model, with special consideration of the project's value. The proposed methodology achieves geometallurgical domain definition through clustering of drill samples. Hence, different clustering methods are evaluated, namely hierarchical, *k*-means clustering, DBSCAN, and self organizing maps, and an economic analysis is presented that supports the advantages of using the obtained geometallurgical model. Consequently, the contributions of this work are twofold: (i) a data-driven methodology for the synthesis of a geometallurgical model from early-stage information is proposed; and (ii) a comprehensive case study on a copper deposit is presented, including an economic evaluation to assess the impact of the geometallurgical model on the project value.

The manuscript is organized as follows: Section 2 presents related work, Section 3 details the methodology for synthesizing a purely data-driven geometallurgical model, Section 4 presents an economical analysis illustrating the potential of the obtained model, and Section 5 states concluding remarks and future research.

2. Related Work

The use of data-driven techniques in geometallurgy has seen an upward trend in recent years. In this context, two main lines of research can be identified. First, linear regression and supervised learning techniques have been used to predict metallurgical variables. In this line, Johnson et al. [9] used data from hyperspectral images derived from mill samples and multiple linear regression to predict Au-Cu recovery, grade, and throughput, giving R-correlation values to observed data of between 0.56 and 0.71. Similarly, Rincon et al. [14] obtained a linear model with input data from drill samples with a variable Cu grade, texture, mineralization style and lithology to predict the Axb milling index and operating work index for each production block, giving R-correlation indices of 0.82 and 0.8, respectively, between experimental and predicted values. In both cases, the geometallurgical units were already defined. Regarding the application of supervised learning, Silva et al. [8] realised that a neural network model is used to predict the concentrate yield and the modal mineralogy for a Nepheline Syenite deposit by magnetic separation laboratory tests, with bulk chemical analysis as input data. The model achieved a correlation coefficient of 0.9 for concentrate yields between the predicted and observed data, while most modal mineralogy had Pearson correlation coefficients lower than 0.8, indicating a potential for improvement. Similarly, Niquini et al. [6] used a neural network to predict a series of plant output variables based on the ore grade of Zn, Pb, and two binary variables for ore body identification in a zinc deposit. Results show a correlation coefficient between predicted and real values for the test set higher than 0.9 for four of the output variables (Zn concentrate mass recovery, tailings' mass recovery, metallurgical recovery of Zn from Zn concentrate, metallurgical recovery of Zn from tailings) and lower than 0.3 for two other output variables (Pb concentrate mass recovery and metallurgical recovery of Zn from Pb

concentrate). Since labels are required for any supervised learning technique, these models could be used only for brownfield studies as an alternative to geochemical testing.

In a second line of research, the use of unsupervised learning techniques, mainly cluster analysis, has been adopted for geometallurgical modeling. Rajabinasa and Asghari [7] tested three clustering techniques to define the geometallurgical domains of an iron ore deposit using five intrinsic variables (Fe, FeO, S, magnetic susceptibility, and sample coordinates), concluding that *k*-means and self-organizing maps were best scored using the Silhouette and Calinski–Harabasz indices as measures of clustering quality. Both techniques defined three clusters identified as a poor ore cluster, a medium ore cluster and a rich ore cluster. This study used 356 core samples, did not consider any processing variables and did not consider mining variables. Consequently, it is not strictly a geometallurgical model. In another study, Bhuiyan et al. [16] implemented the analysis on a gold ore deposit with *k*-means clustering analysis using as variables Bond ball mill work index (BWi), point load strength index (PLSI), rock quality designation (RQD) as geomechanical variable, magnetic susceptibility (MAGSUSC) and twenty geochemical variables, resulting in five clusters. Each cluster has unique characteristics dominated by a combination of higher or lower values by BWi, PLSI, RQD and MAGSUSC. However, BWi results are similar between three of the five clusters. Therefore, using only BWi as the clustering variable, two clusters were redefined by *k*-means, and then supervised learning (random forest) was conducted with ten-fold cross-validation taking the results as labels, obtaining a 70% accuracy for the BWi prediction. Even though this study does include variables associated with processing, it does not include mining aspects, spatial variables or ore.

3. Data-Driven Geometallurgical Model Synthesis

3.1. Dataset

To conduct the data-driven modeling, a dataset of 1112 drilling samples from a copper deposit were generated, which includes 29 numerical variables (see Table 1) and three geological categorical variables that were labeled by a geologist (see Table 2). The dataset includes the coordinates of the sampling point (east, north, and elevation), geochemical data in the the form of percentage of total copper (CuT), total iron (FeT), Mo, As, Zn, soluble copper (CuS), SiO_2, and Al_2O_3 present in the ore, size reduction energy consumption of the sample in terms of Wi and SPI, and mineralogy (percentages of minerals or alteration in the surface of a briquette, measured by optical microscopy). Metallurgical test were performed in a laboratory-scale flotation cell in batch mode. In Table 1, the main statistics of the numerical variables are presented, while in Table 2 the categorical variables are detailed.

Table 1. Statistics of numerical variables.

Family	Variable	Min	Max	Median	Mean	Std Dev
Spatial [m]	East	2841.0	4133.9	3294.2	3347.6	261.8
	North	2204.9	6038.2	4149.6	4160.8	794.3
	Elevation	1448.3	2854.6	2221.9	2223.7	238.0
Grade [%]	CuT	0.13	5.20	0.87	1.04	0.59
	FeT	0.02	13.48	1.26	1.72	1.40
	Mo	0.00	0.42	0.03	0.05	0.05
	As	0.00	1.02	0.01	0.04	0.07
	Zn	0.00	2.15	0.02	0.06	0.12
	CuS	0.01	0.40	0.04	0.05	0.04
	SiO_2	8.60	94.20	70.90	69.85	10.09
	Al_2O_3	4.00	76.90	14.40	14.90	7.88
	CuCon	21.70	42.90	31.40	31.29	3.94
Recovery [%]	Copper recovery	53.60	97.00	88.70	87.93	4.97

Table 1. *Cont.*

Family	Variable	Min	Max	Median	Mean	Std Dev
Process [kWh/t]	Wi	10.03	19.77	14.11	14.15	1.60
	SPI	0.46	8.36	3.23	3.15	1.17
Mineralogy [%]	Cc	0	97	0	3	11
	Dg	0	50	3	7	10
	Cv	0	62	3	8	11
	Bn	0	77	2	10	15
	En	0	73	0	5	10
	Cp	0	100	20	29	30
	Py	0	93	40	38	24
Alteration and gangue [%]	PIR	0	57	0	3	7
	QS	0	100	45	51	38
	SVG	0	86	0	4	8
	KSIL	0	83	0	7	14
	PF	0	100	37	37	30
	Clo	0	46	0	1	4
	Qz	0	79	9	15	16

CuCon: Copper in concentrate, Cc: Chalcocite, Dg: Digenite, Cv: Covellite, Bn: Bornite, En: Enargite, Cp: Chalcopyrite, Py: Pyrite, Qz: Quartz, QS: Quertz sericite, SVG: Early sericite gray green alteration, KSIL: Intense potassium alteration, PF: Pyrophyllite alteration, Clo: Chorite, PIR: Pyritic alteration. Mineralogy (%) indicates the amount of each mineral or alteration present in the briquette, which is studied under the microscope. Copper grade is determined by chemical analysis with atomic absorption spectroscopy.

Table 2. Summary of categorical variables.

Variable	Type
Alteration	301, 303, 305, 307, 309, 312, 318
Mineralization	206, 208, 209, 210, 211, 212, 213, 214, 215, 216, 217, 218, 219, 220
Mineral zone	Weak secondary ore, Strong secondary ore, Primary ore

3.2. Data Pre-Processing

Before modeling, a standard preconditioning is applied to the dataset, consisting of one-hot encoding for categorical variables, as suggested by Müller et al. [17], cleaning of dependent variables, and normalization of time series. Cleaning of dependent variables is performed using the "Clean" algorithm from the Pybalu package, which checks the dependency between variables by randomly choosing one variable and removing those with a Pearson correlation coefficient exceeding 0.99. "Clean" also checks and eliminates variables with constant values. As a result, the geological categorical variable "Strong secondary ore" was removed. In terms of normalization, a projection of the $[0, 1]$ interval was implemented. According to García et al. [18], normalization is applied to avoid suboptimal solutions because of different scales in the data.

3.3. Dimensionality Reduction

Reducing the dimensions of the dataset has several benefits, e.g., it can improve the machine learning model performance, reduce processing time and storage requirements, and facilitate data visualization, according to Müller et al. [17]. Two classical methods are proposed in this case study: principal component analysis (PCA) and autoencoders. While the former is a linear transformation that projects each datapoint onto a few principal components, a new coordinate system of lower dimension, the latter are unsupervised neural networks trained to reconstruct their inputs at the output layer, passing through an intermediate layer usually of lower dimension than the inputs. Autoencoders can have

several layers with arbitrary activation functions, where the most commonly used are sigmoid, rectified linear unit (ReLU) and hyperbolic tangent (tanh) functions.

Dimensionality reduction is performed by a group of variables (detailed in Table 3) which are reduced to one dimension. The rationale behind variable grouping is as follows. "Spatial" variables indicate the 3D location of data; "Another Grade" variables are process variables related to extra revenue or penalties; "Process" variables are parameters related to energy consumption; and "Mineralogy" variables are those obtained by laboratory testing that have a relation with process indicators as recovery, or acid consumption, among others. Note that CuT, CuS, CuCon, and recovery of copper were not included in any group, as they have stronger implications in the copper industry value chain than the other variables. Consequently, they should have the same weight than the other groups in the input dataset used for geometallurgical modeling.

Table 3. Grouping of variables.

Type	Name	Variables
Grouped	Spatial	East, North, Elevation
	Another grade	Fe, Mo, As, Zn, SiO_2, Al_2O_3
	Process	Wi, SPI
	Mineralogy	Cc, Dg, Cv, Bn, En, Cp, Py, Qz,QS, SVG, KSIL, PF, Clo, PIR
Independent	Total copper	CuT
	Copper recovery	Rec
	Soluble copper	CuS
	Copper in concentrate	CuCon

Dimensionality reduction using PCA is an standard and well studied procedure. On the other hand, autoencoders need to be hand-crafted, with different activation functions and numbers of neurons defined for each group of variables depending on the nature of the data. Figure 1 presents the architectures used in this work for each group of variables, which were defined based on a trial-and-error procedure.

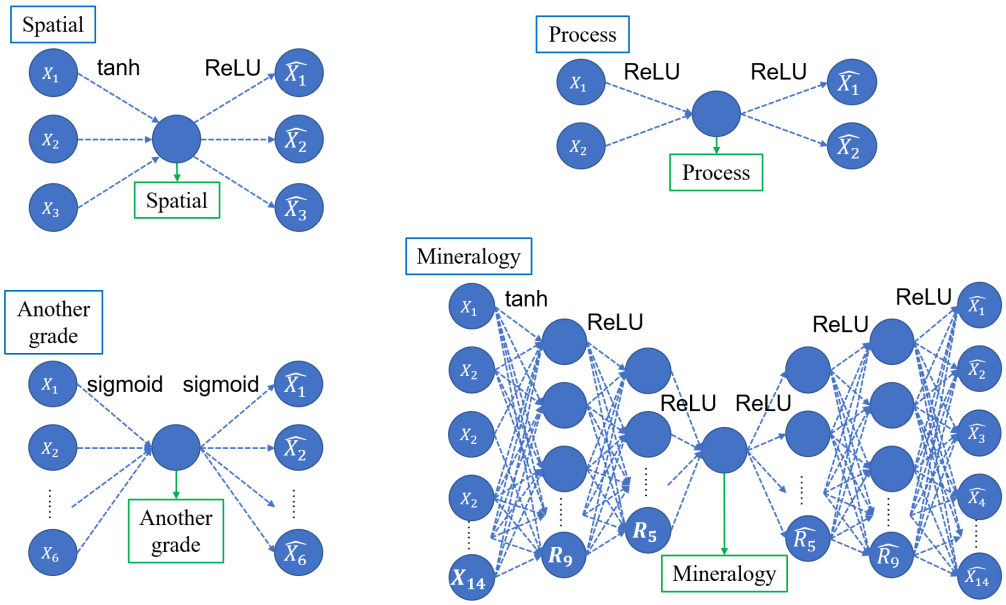

Figure 1. Autoencoder architecture used for Spatial, Process, Another grade, and Mineralogy.

To compare the performance of both dimensionality reduction methods, the mean squared error (MSE) of the reconstructed signal with respect to the original values is

calculated. To this end, an inverse PCA transform is required, while for the autoencoder this metric is given by the loss function. Table 4 presents the statistical information of the original dataset and the reconstruction using PCA. The means of the original dataset (μ_i) and the recovered dataset (μ_f) are quite similar; however, the standard deviation of the PCA inverse transform (σ_f) is smaller than the original dataset (σ_i), indicating that some characteristics of the original dataset are lost. Similarly, Table 5 provides the statistical information for autoencoder-based reconstruction, which is similar to the PCA case with a marginal difference in favor of the autoencoder in terms of standard deviation. In terms of MSE, the autoencoder presents better performance for the groups of variables with larger initial dimensions. Due to this fact, autoencoders are selected in this work as the dimensionality reduction method.

Table 4. Statistical information of PCA-based reconstruction.

Group Name	μ_i	σ_i	MSE	μ_f	σ_f
Spatial	0.484	0.205	0.0102	0.485	0.178
Another grade	0.194	0.258	0.0056	0.194	0.246
Process	0.4	0.154	0.0045	0.4	0.139
Mineralogy	0.177	0.26	0.0184	0.177	0.222

Table 5. Statistical information of autoencoder-based reconstruction.

Group Name	μ_i	σ_i	MSE	μ_f	σ_f
Spatial	0.484	0.205	0.0104	0.483	0.180
Another grade	0.194	0.258	0.0044	0.197	0.257
Process	0.4	0.154	0.0046	0.4	0.138
Mineralogy	0.177	0.26	0.0157	0.169	0.243

3.4. Model Synthesis

The model synthesis is based on an unsupervised clustering process, which divides data into different classes [19] that then conform the different metallurgical domains of the model.

3.4.1. Clustering Methods

In this work, four clustering methods are tested, namely, *k*-means clustering, hierarchical clustering, self-organizing maps and DBSCAN, which has been proven successful when clustering with irregular forms of data [17].

k-means clustering looks to find *k* cluster centers, or centroids, aiming to determine the partition with a minimum within-cluster sum of squares (WCSS). The algorithm begins with *k* initial centers randomly selected. Then, the distance between each datapoint and the center is calculated, and a cluster is assigned for each point considering the nearest center. The process continues until the WCSS does not improve anymore [20]. As a result of using the euclidean distance to assign a datapoint to the nearest center, the concept behind *k*-means is the existence of spherical-like clusters that are expected to be similar in size.

Hierarchical clustering (AGG) is a stepwise fusion process with consideration of linkages among data. Each datapoint is initially considered as one cluster, then the two most similar clusters are merged based on a specific linkage method, until the *k* requested clusters are achieved [21]. This method has four possible types of linkages: (1) 'ward', which minimizes the variance of the clusters being merged; (2) 'average', which uses the average distances of each datapoint of the two clusters; (3) 'complete', which adopts the maximum distances between all datapoints of the two clusters; or (4) 'single', which employs the minimum of the distances between all datapoints of the two clusters [22].

Self-organizing maps (SOM) is an unsupervised classification method based on an artificial neural network. Training starts with a weight vector $W_v(s)$, using the Euclidean distance to measure the similitude with the input dataset $D(t)$, where t indexes the dataset. When the current iteration weight vector is similar enough the input vector, it is labeled as best matching unit (BMU, $d_v(X)$). Then, the weight vectors of the nodes in the neighborhood of the BMU (including the BMU itself) are updated by pulling them closer to the input vector, namely [23],

$$W_v(s+1) = W_v(s) + \theta(u, v, s) * a(s) * (D(t) - W_v(s)), \quad (1)$$

where a is the learning rate, $\theta(u, v, s)$ is the neighborhood function, u is the index in the BMU map and s is the current iteration. This process finishes when new movements do not increase the similarity to the input [23]. In this algorithm, the main hyperparameters are the learning rate and neighborhood function.

Density-based spatial clustering of applications with noise (DBSCAN) is an algorithm that creates groups of datapoints that are closely packed together (datapoints with many neighboring datapoints), marking as outliers those points belonging to a low-density region (datapoints with neighboring points too far away). DBSCAN has two hyperparameters: the maximum distance between two datapoints to be considered neighbors, and the minimum number of points required to form a dense region. The algorithm randomly begins on a datapoint, searches neighborhoods within the maximum distance, and becomes a new cluster if they have the minimum number of points required. Finally, DBSCAN forms clusters and a group of outliers.

3.4.2. Cluster Validity Indices

To evaluate the optimality of a given clustering result, Silhouette and Calinski–Harabasz indices are used. As demonstrated by Liu et al. [24] and Arbelaitz et al. [25], these indices have good performance in assessing the validity of clusters, hence determining the optimal number of clusters. The Silhouette index (SI) measures the cohesion based on the distance between all the datapoints in the same cluster, and the separation is based on the nearest neighbor distance [26].

Given a dataset $X := \{x_1, x_2, ..., x_N\}$ and a partition or clustering of X as a set of K disjoint clusters $C = \{c_1, c_2, ..., c_K\}$. The SI index is calculated as [26]

$$SI = \frac{\sum_{c_I \in C} \sum_{x_i \in c_I} SI(x_i)}{N} \quad (2)$$

with

$$SI(x_i) = \frac{b(x_i) - a(x_i)}{max\{a(x_i), b(x_i)\}} \quad (3)$$

and

$$b(x_i) = \min_{c_J \in C \setminus c_I} \frac{1}{N_J} \sum_{x_i \in c_I, x_j \in c_J} d_e(x_i, x_j), \quad (4)$$

$$a(x_i) = \frac{1}{N_I} \sum_{x_i \neq x_j \in c_I} d_e(x_i, x_j), \quad (5)$$

where $d_e(x_i, x_j)$ indicates the euclidean distance between datapoint i and datapoint j, $a(x_i)$ represents the average distance between datapoint x_i and all other datapoints in the same cluster, $b(x_i)$ indicates the lowest average distance of datapoint x_i from every other cluster to which x_i does not belong, N is the number of datapoints in the dataset, and N_I and N_J refer to the number of datapoints in clusters I and J.

In this formulation, $a(x_i)$ represents cohesion of the datapoint x_i in the cluster c_I, $b(x_i)$ represents separation of the datapoint x_i with other clusters, and the SI value ranges from -1 to 1. A value of SI closer to 1 indicates good classification, while a SI near to -1 suggests misclassification, as mentioned by Rousseeuw [26].

On the other hand, the Calinski–Harabasz index (CH) measures cohesion based on the distance from all the datapoints to centroids. It simultaneously assesses separation based on the distance from each centroid to the global centroid. The CH index is calculated as [27]

$$CH = \frac{N-K}{K-1} \frac{\sum_{c_J \in C} m_J * d_e(\overline{c}_J, \overline{X})}{\sum_{c_J \in C} \sum_{x_i \in c_J} d_e(x_i, \overline{c}_J)}, \quad (6)$$

where N is the number of datapoints in the dataset, K is the number of clusters, M is a tuple containing the number of datapoints per cluster, i.e., $M = \{m_1, m_2, \ldots, m_K\}$, $\overline{X}$ represents the centroid of the entire dataset, $\overline{c_J}$ indicates the center of cluster J, x_i represents the datapoint i, and d_e denotes the euclidean distance between two datapoints. From this definition, it can be seen that the CH index ranges from 0 to positive infinity, and that a good classification maximizes the CH index.

Both indices are suitable for the evaluation of a geometallurgical model as the modeling requires cohesion within a geometallurgical domain and separation between different domains.

3.4.3. Modeling

For model evaluation purposes, several models are synthesized depending on the clustering method and the dataset used as input. Figure 2 demonstrates, schematically, the routes used for forming the input datasets, considering both "clean" data and data after normalization. There are two methods of generating input datasets: directly using the data after preprocessing and using dimensionality reduction based on autoencoders. The datasets are then used for clustering and results are compared using the SI and CH indices.

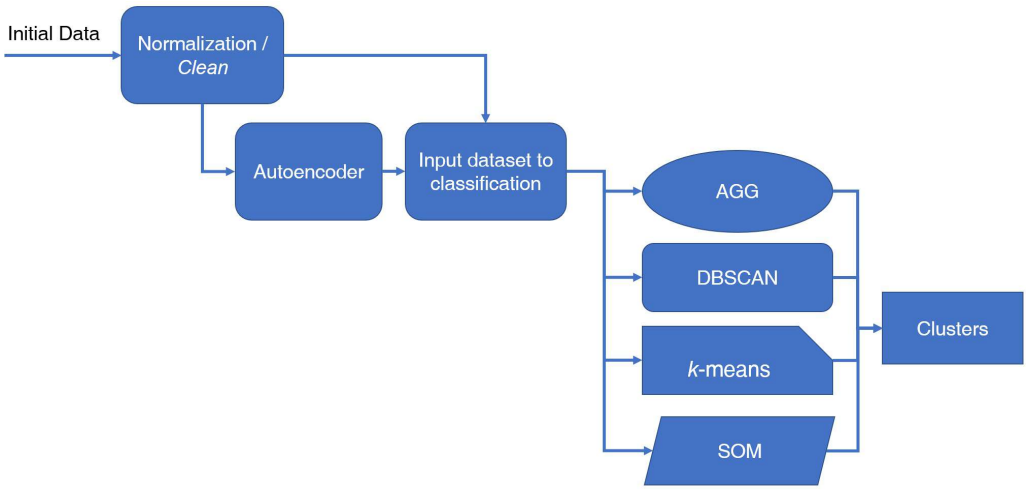

Figure 2. Process flowchart for data analysis.

A total of six datasets are generated, using different combinations of reduced groups of variables and variables without dimensionality reduction, as detailed in Table 6. To include spatial variables, three alternatives exist for each dataset previously defined: considering all spatial coordinates, with dimensionality reduction, and no spatial variables. Therefore, eighteen datasets are finally obtained for modeling purposes, and, consequently, four models per dataset are synthesized.

Table 7 presents the two best results in terms of the Silhouette index (SI) for each clustering method, identifying the dataset used in each case. The results reveal that the *k*-means method outperforms its competitors in terms of clustering quality as measured by the SI.

Table 6. Datasets used for model synthesis.

Dataset Name	Independent Variables	Grouped Variables							
	CuT, CuS, CuCon, Copper Recovery	Geological Variables	Another Grade		Process			Mineralogy	
	Complete	Complete	Complete	Reduced	Complete	Wi Only	Reduced	Complete	Reduced
Original	•	•	•		•			•	
Original/ no geology	•		•		•			•	
Original/ no geology/ no SPI	•		•			•		•	
Partially reduced/ no geology/ no SPI	•			•		•			•
Partially reduced/ no geology	•			•	•				•
Fully reduced	•			•			•		•

• Each bullet represents the type of data included in each data set.

Table 7. Results in terms of the Sillhouette index.

Clustering Algorithm	Dataset Input to Cluster Analysis	Spatial Variables	SI
k-means	Original/no geology/no SPI	None	0.264
k-means	Original/no geology	None	0.263
AGG-Ward	Original/no geology	None	0.249
AGG-Ward	Original/no geology/no SPI	With dimensionality reduction	0.249
SOM	Original/no geology/ no SPI	All spatial coordinates	0.242
SOM	Original/no geology	With dimensionality reduction	0.227
DBSCAN	Original/no geology	None	0.205
DBSCAN	Original/no geology/no SPI	All spatial coordinates	0.205

Similarly, Table 8 presents the two best results in terms of the Calinski–Harabasz index (CH) for each clustering method. A similar trend can be seen, with *k*-means giving the best results. It is interesting to note that, in terms of datasets, the best performers presented in Tables 7 and 8 do not use dimensionality reduction nor include geological categorical variables or spatial variables. This suggests that is feasible to synthesize a geometallurgical model from numerical variables.

To further explore the performance differences throughout the different datasets, the results from the *k*-means method applied to all the datasets are shown in Table 9. It can be seen that the first six rows present similar performance with a variation lower than 1%. For the rest of this study, the best performer is selected as the nominal model. Figure 3 presents the final geometallurgical model, where each cluster is identified as a geometallurgical unit (UGM). The ore body is a convex form, which in the upper center has no drilling sample information. As a result, unit 1 is located south-east and deeper, unit 3 in the south-west and upper, and unit 2 in the north of the ore body.

Table 8. Results in terms of the Calinski–Harabasz index.

Clustering Algorithm	Dataset Input to Cluster Analysis	Spatial Variables	CH
k-means	Original/no geology/ no SPI	None	485.21
k-means	Original/no geology	None	482.46
AGG-Ward	Original/no geology	All spatial coordinates	455.42
AGG-Ward	Original/no geology/ no SPI	All spatial coordinates	452.46
SOM	Original/no geology	None	416.89
SOM	Original/no geology	With dimensionality reduction	415.83
DBSCAN	Original/no geology	All spatial coordinates	288.40
DBSCAN	Original/no geology/no SPI	None	263.55

It can be observed that each geometallurgical domain is clearly separated and has continuity within the domain, even though the input dataset did not include any spatial variable. Nonetheless, it is deemed reasonable that every point in a domain has similar characteristics because the ore body's characteristics were the result of geological evolution.

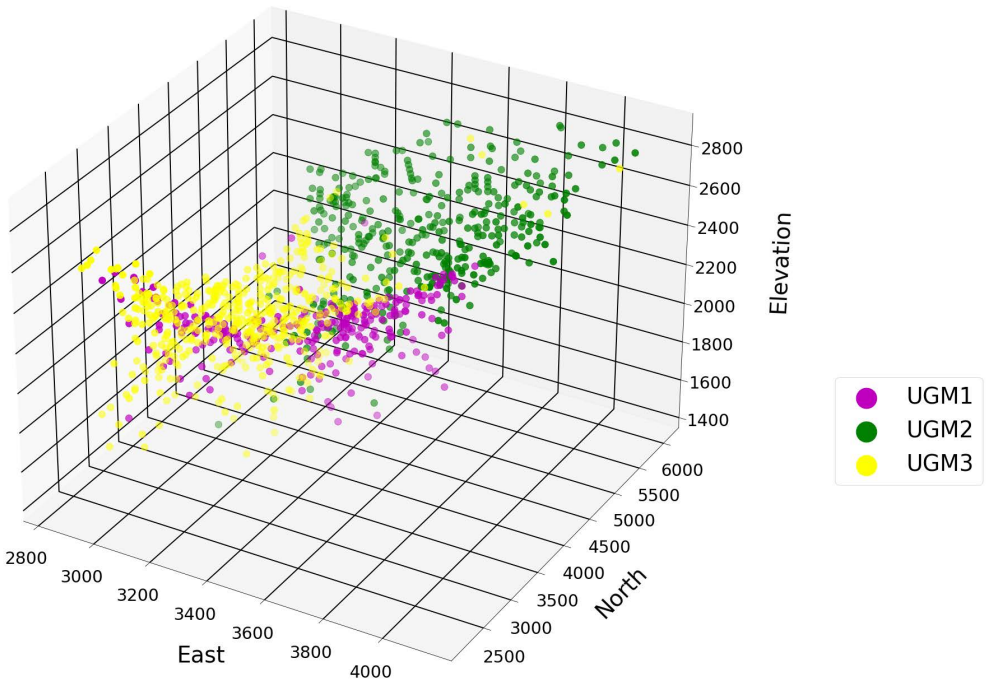

Figure 3. Spatial visualization of the clusters.

To illustrate the behavior of relevant variables among the UGMs, Figures 4–7 present box plots of BWi, SPI, CuCon and As. It can be seen that in terms of BWi, the geometallurgical units are significantly different, while for CuCon the difference between units 1 and 2 is minimal and the same minimal difference is observed for units 2 and 3 in terms of both SPI and As. The results in Figures 4–7 highlight that geometallurgical modeling is a multivariable exercise that can be conducted with the aid of learning-based methods.

Table 9. Results for *k*-means.

Dataset Input to Cluster Analysis	Spatial Variables	SI	CH
Original/no geology/no SPI	None	0.264	485.21
Original/no geology	None	0.263	483.46
Original/no geology	All spatial coordinates	0.263	481.68
Original/no geology/ no SPI	All spatial coordinates	0.263	481.57
Original/no geology	With dimensionality reduction	0.263	483.59
Original/no geology/no SPI	With dimensionality reduction	0.262	484.55
Partial reduced/no geology/no SPI	All spatial coordinates	0.243	431.68
Partial reduced/no geology/ no SPI	With dimensionality reduction	0.241	431.53
Partial reduced/no geology	All spatial coordinates	0.237	422.90
Fully reduced	All spatial coordinates	0.235	417.52
Partial reduced/no geology	With dimensionality reduction	0.230	414.20
Fully reduced	With dimensionality reduction	0.226	410.01
Partial reduced/no geology/ no SPI	None	0.148	300.81
Fully reduced	None	0.147	307.46
Partial reduced /no geology	None	0.147	307.46
Original	With dimensionality reduction	0.142	322.25
Original	All spatial coordinates	0.142	321.54
Original	None	0.142	320.61

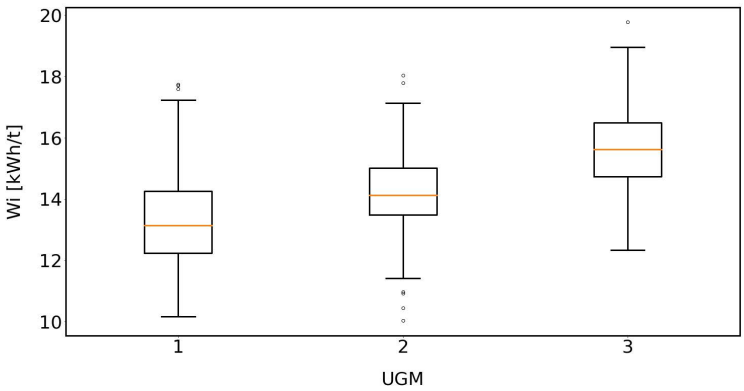

Figure 4. Box plot of relevant variables for the different resulting geometallurgical units (Case Wi).

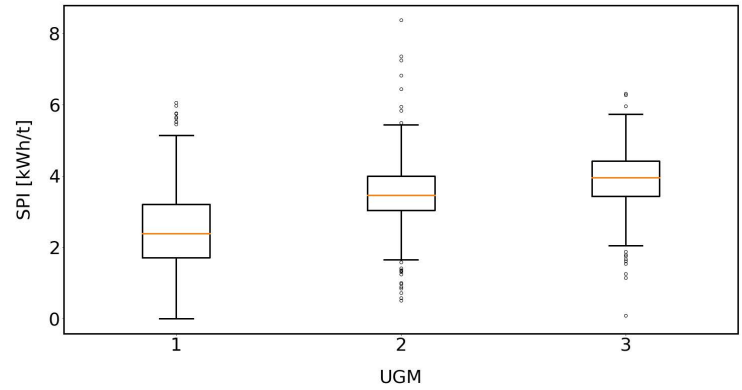

Figure 5. Box plot of relevant variables for the different resulting geometallurgical units (Case SPI).

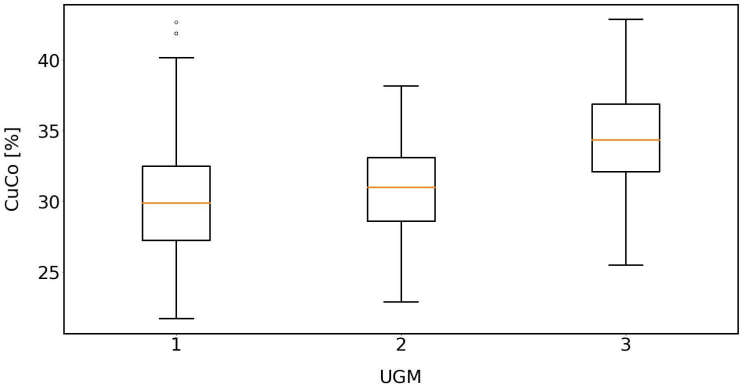

Figure 6. Box plot of relevant variables for the different resulting geometallurgical units (Case CuCon).

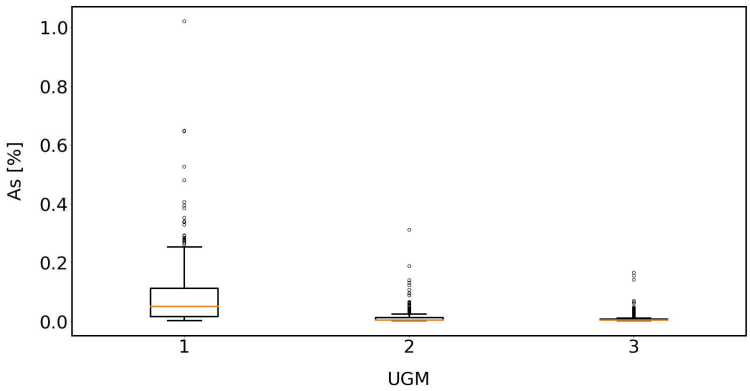

Figure 7. Box plot of relevant variables for the different resulting geometallurgical units (Case As).

4. Economic Evaluation

To illustrate the advantages of having a proper geometallurgical model, a simplified economic evaluation is conducted comparing two scenarios. The first one, which is the base case, considers an average cost and plant throughput for all the deposit. The second scenario considers variable cost and plant throughput depending on the geometallurgical model and the extraction sequence. Evaluation is performed through the estimation of the Net Present Value (NPV) of the project without considering capital investment.

To define the extraction sequence, several steps must be completed, namely, correcting outlier points, creating a block model, and performing interpolations.

4.1. Filtering Outliers

As observed in Figure 3, when spatial consistency is considered, some datapoints belonging to one cluster can be situated within a different cluster, hence behaving as outliers in terms of spatial continuity. To avoid this situation, the data are filtered using the spatial variables (east, north, and elevation) and cluster domain, checking the seven closest points to each datapoint. Seven was found to be the optimal number of points that can effectively correct those outliers and do not change the border points' labels unnecessarily. If six or more points are classified into the same cluster as the reviewed point, the cluster domain of the majority of the points is assigned to the point under analysis. If less than six points are in the same class, these points are considered border points and the cluster domain is maintained.

4.2. Block Model

A cubic block size of 45 m per side is used. Each block is assigned with information from the geometallurgical model. If there is only one point inside the block, the information of this point will be assigned to the block. If there are two points inside the block, an average of the numerical data for the variables and a randomly selected cluster of the data are assigned. If there are more than two points, an average of the numerical data will be adopted for the variables, and the major occurrence of cluster code is selected.

4.3. Interpolation

As there might be blocks without information between two datapoints, interpolation is used. First, blocks needing interpolation are identified by using Alpha shape [28], a computational geometric technique that identifies all blocks within the area with information in the border. In a second step, Gradient Boosting, a machine-learning-supervised regression method, with the drill sample as a training set, is used to predict the information associated with the blocks identified in the first step.

4.4. Final Block Model and Economic Analysis

Figure 3 shows that the deposit has a convex shape with a gangue body in its center, surrounded by the ore body. This means that, conventionally, the mine exploitation would start in a point with less gangue in order to continue towards a singular direction. Hence, to enable using the extraction sequence as an independent variable, an adjustment is made in which only a continuous subset of the original ore body is considered in the block model for economic evaluation.

The selected limits for this model are between 2800 and 3800 east, between 2500 and 5200 north, and between 2000 and 2300 elevation. The adjusted block model is shown in Figure 8.

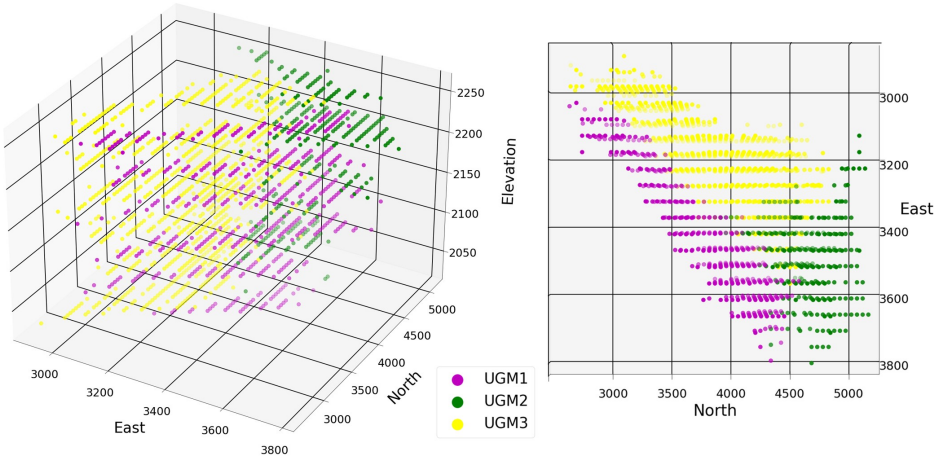

Figure 8. Adjusted block model used in the economic evaluation.

To determine the value of the block model, we use a simulation model developed by Anani et al. [29] to obtain an extraction sequence that uses the economic parameters presented in Table 10 and defined as follows. The long-term copper price used is the average price reported in three mining projects in Chile [30–32]. Mining cost represents the cost variation induced by increasing depth in meters applied to the current deposit, derived from a benchmarking study in Chile [33]. A discount rate of 10% is reasonable for this type of mine project [32,34]. Finally, the GAP is a parameter from the solver Gurobi that indicates the difference between an objective bound and an incumbent solution objective. This value can be determined according to each project. Afum et al. [35] mentioned that a GAP value of less than 5% is tolerable, while Ben-Awuah et al. [36] also suggests that a GAP value of less than 1% is sufficient for a mining project. In our case, a GAP value of 0.4% is used.

Table 10. Parameters used in the economic evaluation.

Parameter	Value
Copper price [US$/lb]	3
Mining cost [US$/t]	3–3.95
Discount rate [%]	10
GAP [%]	0.4

Table 11 presents the different production parameters for the base case and each cluster or geometallurgical unit. The geometallurgical case considers the average value in each geometallurgical domain, while the base case considers the average of the whole dataset. The installed production capacity is defined as 50 kt/d and 40,300 MW of available power. The energy consumption is estimated by the work index of Bond. The energy cost is calculated by energy consumption and local electricity pricing, which is 0.094 US$/kWh [33]. The total processing cost is estimated assuming the energy cost represents 26% of the total cost, as reported in a benchmark study from Chilean copper mine projects [33]. The production ratio is defined as the ratio of the production capacity at each case and the production capacity of the base case.

Table 11. Cost parameters for economic evaluation.

UGM	$\overline{W_i}$ [kWh/t]	Energy [kWh/t]	Energy Cost [US$/t]	Processing Cost [US$/t]	Production Ratio
1	13.31	17.73	1.67	6.41	1.09
2	14.22	18.48	1.74	6.68	0.97
3	15.66	19.66	1.85	7.11	0.88
Base case	14.15	18.42	1.73	6.67	1

Table 12 shows the result of the simulation using the parameters mentioned in Table 11. An NPV of USD 3746 million is obtained for the base case, while the geometallurgical case has an NPV equal to USD 3903 million, 4% higher, demonstrating the relative impact of using the geometallurgical model generated by a machine-learning algorithm.

Table 12. Case NPV.

Case	NPV [MMUS$]
Base	3746
Geometallurgical	3903

5. Conclusions

The generation of a geometallurgical model is an essential part of block modeling and mine planning in any mining project. Since mineral deposits are complex systems, a large database is required to identify relationships between the initial data and the final

extraction sequence and production plan. In this context, we propose a purely data-driven methodology for geometallurgical modeling, based on clustering of drilling samples from an ore body.

Our main conclusions are as follows:

1. The evaluation of different clustering methods indicates that *k*-means outperforms its competitors in terms of the Silhouette and Calinski–Harabasz indices. For the ore deposit under analysis, the optimal number of clusters, hence the number of metallurgical units, is three.
2. In terms of dimensionality reduction, it was found that autoencoders perform better than PCA in the tested dataset; yet, there is no evidence that such a reduction improves the clustering results. It remains to be determined if dimensionality reduction brings any advantage when using larger databases.
3. Additionally, it was found that spatial variables are not determinant in the definition of the clusters or for their spatial continuity.
4. It was also found that geological categorical variables do not contribute to better clustering results, which suggests that it might be possible to synthesize a geometallurgical model from numerical variables.
5. As a measure of the impact that a geometallurgical model can have in the value of a mining project, an economic evaluation comparing the value generated with a mine plan based on average values and on differentiated geometallurgical values was conducted using a block model generated with drilling data and using filtering, block modeling and interpolation techniques. The result indicates that using the geometallurgical model increases the NPV of the project by about 4%. It must be noted, however, that the economic evaluation is performed using cost data for the copper industry in Chile, which might not be applicable to other mining regions.
6. This work has demonstrated that it is feasible to generate a geometallurgical model using a purely data-driven methodology. It is found that using k-means, with encoding for dimensionality reduction, provides a reasonable methodology for clustering data, maintaining spatial continuity.
7. It is envisioned that the proposed methodology for generating geometallurgical models can be further developed, for example, performing cross validation with existing geometallurgical models, or applying it to brownfield projects or ongoing operations, where operational data can be included in the model to improve the clustering analysis and to increase the impact on mining business decisions.

Author Contributions: Conceptualization, Y.M. and J.C.S.; methodology, Y.M. and J.C.S.; formal analysis, Y.M. and J.C.S.; investigation, Y.M. and J.C.S.; data curation, Y.M.; writing—original draft preparation, Y.M. and J.C.S.; writing—review and editing, Y.M. and J.C.S.; supervision, J.C.S.; funding acquisition, J.C.S. All authors have read and agreed to the published version of the manuscript.

Funding: Work was supported in part by ANID under grant ANID PIA ACT192013.

Data Availability Statement: The original data set is not publicly available because of confidentiality agreements.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ellefmo, S.L.; Aasly, K.; Lang, A.; Vezhapparambu, V.S.; Silva, C.A. Geometallurgical concepts used in industrial mineral production. *Econ. Geol.* **2019**, *114*, 1543–1554. [CrossRef]
2. Hoal, K.O. Getting the geo into geomet. *SEG Discov.* **2008**, *73*, 1–15. [CrossRef]
3. McQuiston, F.; Bechaud, L.J. Metallurgical sampling and testing. In *Surface Mining*; Pfeidler, E.P., Ed.; The American Institute of Mining, Metallurgical and Petroleum Engineers Inc.: New York, NY, USA, 1968; pp. 103–121.
4. Sepulveda, E.; Dowd, P.A.; Xu, C.; Addo, E. Multivariate modelling of geometallurgical variables by projection pursuit. *Math. Geosci.* **2017**, *49*, 121–143. [CrossRef]
5. Hunt, J.; Berry, R. Economic geology models 3. Geological contributions to geometallurgy: A review. *Geosci. Can. J. Geol. Assoc. Can.* **2017**, *44*, 103–118.

6. Niquini, F.; Costa, J. Mass and metallurgical balance forecast for a zinc processing plant using artificial neural networks. *Nat. Resour. Res.* **2020**, *29*, 3569–3580. [CrossRef]
7. Rajabinasab, B.; Asghari, O. Geometallurgical domaining by cluster analysis: Iron ore deposit case study. *Nat. Resour. Res.* **2019**, *28*, 665–684. [CrossRef]
8. Silva, C.A.; Ellefmo, S.L.; Sandøy, R.; Sørensen, B.; Aasly, K. A neural network approach for spatial variation assessment—A nepheline syenite case study. *Miner. Eng.* **2020**, *149*, 106178. [CrossRef]
9. Johnson, C.; Browning, D.A.; Pendock, N.E. Hyperspectral imaging applications to geometallurgy: Utilizing blast hole mineralogy to predict Au-Cu recovery and throughput at the Phoenix mine, Nevada. *Econ. Geol.* **2019**, *114*, 1481–1494. [CrossRef]
10. Dominy, S.C.; O'Connor, L.; Parbhakar-Fox, A.; Glass, H.J.; Purevgerel, S. Geometallurgy—A route to more resilient mine operations. *Minerals* **2018**, *8*, 560. [CrossRef]
11. Koch, P.; Rosenkranz, J. Sequential decision-making in mining and processing based on geometallurgical inputs. *Miner. Eng.* **2020**, *10*, 106262. [CrossRef]
12. Lishchuk, V.; Koch, P.H.; Ghorbani, Y.; Butcher, A.R. Towards integrated geometallurgical approach: Critical review of current practices and future trends. *Miner. Eng.* **2020**, *145*, 106072. [CrossRef]
13. Suazo, C.J.; Kracht, W.; Alruiz, O.M. Geometallurgical modelling of the Collahuasi flotation circuit. *Miner. Eng.* **2010**, *23*, 137–142. [CrossRef]
14. Rincon, J.; Gaydardzhiev, S.; Stamenov, L. Coupling comminution indices and mineralogical features as an approach to a geometallurgical characterization of a copper ore. *Miner. Eng.* **2019**, *130*, 57–66. [CrossRef]
15. Baumgartner, R.; Dusci, M.; Gressier, J.; Trueman, A.; Poos, S.; Brittan, M.; Mayta, P. Building a geometallurgical model for early-stage project development-a case study from the Canahuire epithermal Au-Cu-Ag deposit, Southern Peru. In Proceedings of the First AUSIMM International Geometallurgy Conference, Brisbane, Australia, 5–7 September 2011.
16. Bhuiyan, M.; Esmaieli, K.; Ordóñez-Calderón, J.C. Application of data analytics techniques to establish geometallurgical relationships to bond work index at the Paracutu mine, Minas Gerais, Brazil. *Minerals* **2019**, *9*, 302. [CrossRef]
17. Müller, A.C.; Guido, S. *Introduction to Machine Learning with Python: A Guide for Data Scientists*, 3rd ed.; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2016.
18. García, S.; Luengo, J.; Herrera, F. *Data Preprocessing in Data Mining*, 3rd ed.; Springer: Berlin/Heidelberg, Germany, 2015.
19. Koch, I. *Analysis of Multivariate and High-Dimensional Data*, 3rd ed.; Cambridge University Press: Cambridge, UK, 2013.
20. MacKay, D.J.; Mac Kay, D.J. *Information Theory, Inference and Learning Algorithms*, 3rd ed.; Cambridge University Press: Cambridge, UK, 2013.
21. Dubien, J.L.; Warde, W.D. A mathematical comparison of the members of an infinite family of agglomerative clustering algorithms. *Can. J. Stat. Rev. Can. Stat.* **1979**, *7*, 29–38. [CrossRef]
22. Müllner, D. Modern hierarchical, agglomerative clustering algorithms. *arXiv* **2011**, arXiv:1109.2378.
23. Kohonen, T. The self-organizing map. *Proc. IEEE* **1990**, *78*, 1464–1480. [CrossRef]
24. Liu, Y.; Li, Z.; Xiong, H.; Gao, X.; Wu, J. Understanding of internal clustering validation measures. In Proceedings of the 2010 IEEE International Conference on Data Mining, Sydney, Australia, 13–17 December 2010; pp. 911–916.
25. Arbelaitz, O.; Gurrutxaga, I.; Muguerza, J.; Pérez, J.M.; Perona, I. An extensive comparative study of cluster validity indices. *Pattern Recognit.* **2013**, *46*, 243–256. [CrossRef]
26. Rousseeuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. [CrossRef]
27. Caliński, T.; Harabasz, J. A dendrite method for cluster analysis. *Commun. Stat.-Theory Methods* **1974**, *3*, 1–27. [CrossRef]
28. Alphashape. Available online: https://github.com/bellockk/alphashape (accessed on 1 July 2021).
29. Anani, A.; Li, H.; Ortiz, I.A. Conditions under which an integrated approach to the transition mine problem outperforms a disintegrated approach. In Proceedings of the APCOM 2021, Online, 30 August–1 September 2021.
30. Maycock, J.; Luraschi, A.; Mendozam, M.; Bianchin, M.; Rennie, D.; Guzman, C.; Amelunxen, R.; Gingles, M.; Kerr, T.; Betinol, R. *NI 43-101 Technical Report on Feasibility Study Update, Santo Domingo Project*; Capstone Mining Corp.: Atacama, Chile, 2018.
31. *Scoping Study for the Sierra Gorda Project*; Quadra Mining LTD.: Antofagasta, Chile, 2009.
32. Marinho, R.; Nelson, M. *NI 43-101 Technical Report on Quebrada Blanca Phase 2 Feasibility Study 2016*; Teck Resources Limited: Tarapacá, Chile, 2017.
33. *Estudio Benchmarking Gestión Minera*; Encare Benchmarking Gestión: Santiago, Chile, 2018.
34. Rayo, J.; Sánchez, N. *Estudio Factibilidad Rajo Inca: Proyecto Rajo Inca (PRI)*; Corporación Nacional del Cobre de Chile: Santiago, Chile, 2018.
35. Afum, B.; Ben-Awuah, E. MILP framework for open pit and underground mining transitions evaluation. In *Mining Goes Digital: Proceedings of the 39th International Symposium'Application of Computers and Operations Research in the Mineral Industry (APCOM 2019)*; CRC Press: Wroclaw, Poland, 2019.
36. Ben-Awuah, E.; Richter, O.; Elkington, T.; Pourrahimian, Y. Strategic mining options optimization: Open pit mining, underground mining or both. *Int. J. Min. Sci. Technol.* **2016**, *26*, 1065–1071. [CrossRef]