An Integrated Data-Driven Approach for Evaluating the Impact of Additional Drill Holes in Mining Exploration

José González
Dassault Systèmes
Santiago, Chile
email:
jose.gonzalez@3ds.com

María González
Dassault Systèmes
Santiago, Chile
email:
maria.gonzalez@3ds.com

Lukas Meszaros
Dassault Systèmes
Santiago, Chile
email:
lukas.meszaros@3ds.com

Andre Orlandi
Dassault Systèmes
São Paulo, Brazil
email:
andre.orlandi@3ds.com

***Abstract*—Efficient resource exploration in mining necessitates a strategic approach to drilling campaigns, where the acquisition of geological data must be both informative and cost-effective. In this paper, we present a new methodology that combines clustering algorithms, global bias assessment, and variogram comparison to evaluate the efficacy of additional drill holes in geological investigations. The methodology begins with the application of clustering algorithms to identify spatial patterns within existing data sets. Through this process, we categorize geological information into meaningful groups, allowing for the identification of regions with similar characteristics. Subsequently, a global bias assessment is conducted to quantify the contribution of each drill hole to the overall dataset. This step aids in discerning the unique information provided by individual holes and highlights potential redundancies. The variogram comparison is then employed to analyze the spatial continuity and variability of geological features. By assessing how variograms evolve with the inclusion of additional drill holes, we can determine when the acquired information reaches a point of diminishing returns. Through the integration of clustering validation parameters (Rand index), global bias between campaigns, and variogram contrast, we can quantify the moment when the data no longer contributes new information to our models. This analysis forms the basis for informed decision-making regarding the optimal number and placement of drill holes.**

Keywords - Drillhole Data; Clustering analysis; Global bias assessment; Variogram; Mining exploration efficiency; Decision-making.

I. INTRODUCTION

In the mining industry, acquiring and interpreting geological data must be informative, cost-effective, and delivered on time at the expected quality to support multidisciplinary analyses required to report mineral resources according to industry standards. Although the importance of high-quality geological data is well-recognized, continuous robustness measurements during general and detailed exploration phases remain limited due to time and resource constraints. Implementing more frequent robustness measurements could enable timely data publication for interdisciplinary analysis and provide decision support for optimizing drilling campaign timelines and costs.

Addressing the challenge of spatial dependence in clustering problems using drillhole data, Romary et al. [10] introduced two algorithms: geostatistical hierarchical clustering and geostatistical spectral clustering, with proximity measures as the key differentiating factor. Notably, objects are clustered based on their spatial connectivity through an undirected graph. Geostatistical hierarchical clustering demonstrated superior performance compared to other algorithms, although its applicability is limited to smaller datasets. Building upon this, Fustos [6] explored two approaches that integrate geostatistical formulations into unsupervised machine learning clustering techniques, aiming to incorporate geological knowledge. The first approach modifies the distance function in hierarchical clustering by integrating the variogram function into the conventional Mahalanobis distance De Maesschalck [2]. The second approach utilizes a mixture of distributions. Both proposals were initially validated using synthetic data and subsequently applied to real-world cases involving geochemical, metallurgical, and geological data. Furthermore, Faraj [5] proposed a workflow for defining domains, employing hierarchical clustering and emphasizing geology, statistics, and spatial continuity through a distribution-based classification.

This paper contributes to this line of research by presenting a practical application case, demonstrating the effectiveness of geostatistical clustering in a real-world scenario.

The rest of the paper is structured as follows. In Section 2, we present the methodology and key indicators applied to the analysis of clustering for drillhole data. In Section 3, we present results applied to a copper deposit, including a case study on the decision matrix for publishing preliminary geological data. Finally, in Section 4, we conclude the work and provide recommendations for future research.

II. METHODOLOGY

To achieve the goals outlined in Section I, we applied unsupervised machine learning clustering, global bias assessment and variography comparison to drillhole data for continuous measurement of data robustness, considering execution time and feasibility of human effort.

The proposed approach involves comparing two sets of geological data: one with data up to a specific time and another with the additional newly acquired data. The comparisons focus on two aspects: the effect on data grouping using the Adjusted Rand Index and the impact on grade spatial distribution using global bias. When comparing these two metrics, we can create a decision matrix that, when contrasted with the difference between the experimental variograms, can support an informed decision-making process by indicating whether the collected data population possesses the desired robustness or not. In Figure 1, we can see the interpretation of the Adjusted Rand Index and global bias values to represent different scenarios of robustness of the geological data from drillhole campaigns. In this figure, the target scenario indicates that the dataset already has satisfactory robustness.

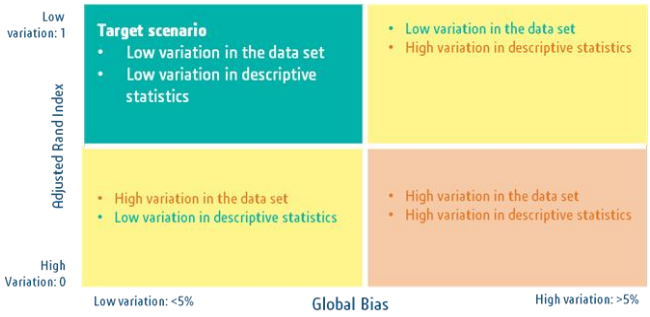

Figure 1. Decision matrix with detailed scenarios based on the Adjusted Rand Index and global bias.

The methodology employed in this study encompasses the selection of variables for clustering algorithms and the subsequent application of non-hierarchical clustering methods, particularly focusing on Clustering for Large Applications (CLARA) Kaufman et al. [9]. The chosen variables include coordinate, numerical, and categorical parameters, with specific attention to numerical variables based on geochemical data with low linear correlations to avoid redundancy. CLARA, an extension of the K-medoids method, is adopted due to its computational efficiency and integration capability with scaled coordinate variables. Furthermore, data treatment involves considerations such as detection limits for geochemical elements and normalization of variable distributions. The process of selecting the optimal number of clusters is addressed, using the elbow method to discern significant decreases in intracluster variance without sacrificing relevance or introducing redundancy. In this way, candidates for the number of clusters to be generated are found and this information is crossed with the number of geological domains to select the ideal number of clusters. In addition, the decision to apply the Manhattan distance metric is justified by the irregular grid distribution of the drillholes and the desire to take advantage of spatial information while mitigating dimensional errors.

The following sub-sections describe the indicators of interest obtained by applying clustering to datasets from drilling surveys conducted before and after incorporating data from new campaigns.

A. *Indicator of interest: Rand index*

The Rand index is a quantity used to compare different subsets of the same data set in different groupings. The Adjusted Rand Index (ARI) is calculated by correcting the unadjusted index for Random overlaps. Its value is a number between 0 and 1, where 1 means that the partitions are identical, while 0 represents that they are completely random (Hubert [8]; Yeung [12]; Warrens [11]).

Rand indices are calculated across distinct instances, facilitating a comparison of cluster alterations based on shared information. This evaluation method allows for an assessment of the impact of new data on reservoir understanding by analyzing changes in the Rand index. An increase in the ARI indicates that the new data does not significantly contribute additional insights into the deposit. On the other hand, a decrease in the ARI indicates that the data acquired is considerably different from the data obtained in previous campaigns.

TABLE 1. ADJUSTED RAND INDEX INTERPRETATIONS.

Adjusted Rand Index Value	Interpretation
ARI < 0.5	Low correlation
0.5 < ARI < 0.8.	Medium correlation
0.8 < ARI. < 0.9	Strong correlation
0.9 < ARI < 1.0	Very strong correlation

B. *Indicator of interest: Decluster Mean Bias*

In addition to the metrics for clustering techniques with machine learning, considering the need to work with regionalized variables, geostatistical metrics are used. One of these metrics corresponds to the declustered mean bias (Chilès [1]; Deutsch [3]; Emery [4]), which allows us to appreciate the change in domain statistics as new information is added through the drilling campaigns. The calculation of the bias is done through the following equation (1) where i and j represent two partitions of the total data according to the amount of information available up to that moment.

$$\text{Bias (\%)} = \left| \frac{m_{zi}}{mzj} - 1 \right| \cdot 100 \qquad (1)$$

where m_z represents the declustered mean of the data. With respect to the declustering means, these are calculated using a cell coherent to the spacing between drillholes up to the given drilling campaign, which will decrease in size as the sampling density for the reservoir increases.

TABLE 2. GLOBAL BIAS INTERPRETATIONS.

Adjusted Rand Index Value	Interpretation
> 5%	High variation
< 5%	Variation within the standards expected in international norms
< 3%	Low variation

C. *Indicator of interest: Variograms*

After performing a variographic map analysis, experimental variograms are calculated in the preferential directions with the data from two different drilling campaigns and the nugget effect, the sill and the ranges are compared. For this purpose, it is necessary to use the same parameters and directions for the variograms of the different drilling campaigns when contrasting.

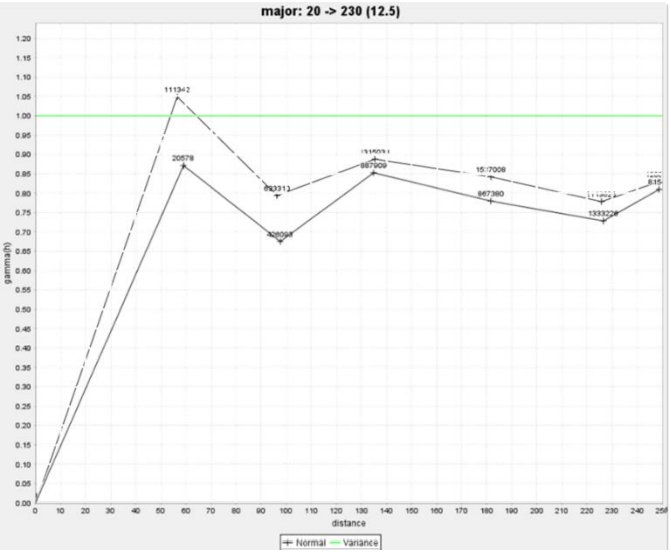

Figure 2. Variograms of cumulative drilling campaigns in the same direction in order to contrast short and long range spatial differences.

III. RESULTS

For Deposit A (whose name has been changed to maintain confidentiality), the main mineral commodity corresponds to copper, therefore the analysis and decision for the methodology are going to be centered to better understand this element. The geochemical variables chosen for the analysis correspond to Copper, Gold, Iron and Magnesium and the categorical variables chosen for the deposits correspond to the lithological domains.

TABLE 3. MAIN STATISTICS FOR DEPOSIT A WITH COMPOSITE SIZE OF 2 M.

	Number of data points	Mean	Min	Max	Standard deviation
Copper grade (%)	60094	0.278	0.001	29.640	0.512
Gold grade (ppm)	60092	0.049	0.005	30.700	0.049
Iron grade (%)	52781	6.244	0.05	37.200	6.244
Magnesium grade (%)	43409	4.069	0.005	25.200	3.908

With the critical variables defined, the decision matrix between the Adjusted Rand Index and global bias is constructed. Of the matrix, it can be observed that in the last 2 campaings, the new information ceases to significantly contribute to the defined criteria. Subsequent to this observation, a more detailed analysis of the drillhole distribution for the second last campaign is conducted finding that with a sub-distribution of 28% of the campaign, geological preliminary data could have been published. This suggests that information obtained after this point, upon reaching the target zone, does not yield additional relevant insights. This analysis underscores the potential for expediting downstream processes by as much as 11 months.

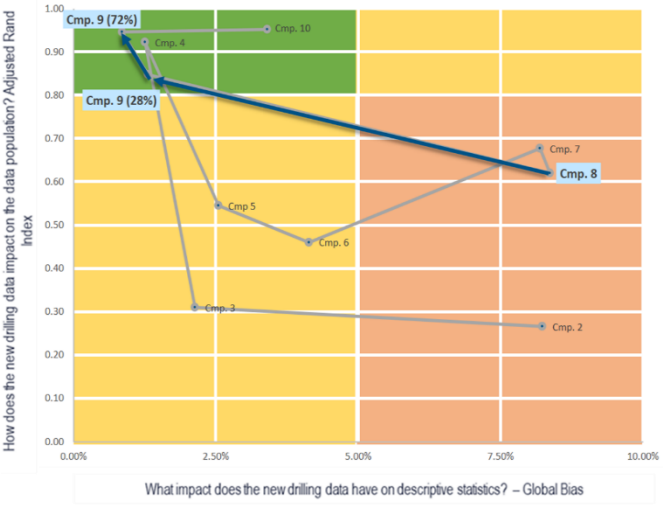

Figure 3. Decision matrix with drilling campaigns for Deposit A. Campaigns of interest highlighted.

IV. CONCLUSIONS AND FUTURE WORK

In this work, we aimed to assess the robustness of additional drill holes for the publication of preliminary geological data in the mining workflow by integrating clustering algorithms, global bias assessment, and variogram comparison. To achieve this, we developed a structured workflow that leverages these techniques to analyze drillhole data and extract meaningful insights summarized in a decision matrix. After applying the proposed workflow, it was possible to identify robust results by up to 11 months by being able to assess and quantify the maturity of the data to integrate it into the resource estimation workflow and

improve results by sharing them earlier to downstream processes.

Based on our analysis, we conclude that:

• Automated robustness measurement achieves expected quality while optimizing time and effort.

• The decision matrix is useful for guiding decisions on geological data publication and influencing drilling campaigns.

• Analysis shows no confirmation bias, ensuring objective quality assessments.

This approach contributes to ensure timely data publication with expected quality for interdisciplinary analyses, improve drilling campaign decisions regarding costs, timing and data quality in both general and detailed exploration phases, and mitigate the risk of rework due to inappropriate data quality.

For future work, it is important to consider that in using clustering techniques, one of the most critical parameter in this context is the number of clusters to be generated, which can be estimated using the elbow methodology. However, it is essential to carry out manual iterations of this process and to contrast the results with the specific geological information of the reservoir. This practice becomes an obstacle to the full automation of the proposed methodology. In addition, by employing the Mahalanobis distance, which includes coordinates, the spatial correlation of variables can be taken into account in a trivial way. However, it would be advisable to evaluate this methodology using a distance function that incorporates this consideration more comprehensively, such as the distance function proposed by Gonzalez [7].

ACKNOWLEDGMENT

We would like to express our gratitude to our team from Dassault Systèmes for their pivotal role in enabling the execution of this study and its successful application to a real-world project. Furthermore, we gratefully acknowledge the invaluable guidance and inspiration provided by Professors Xavier Emery and Nadia Mery of the Universidad de Chile in the field of geostatistical innovation.

REFERENCES

[1] J. Chilès and P. Delfiner, "Geostatistics: Modeling Spatial Uncertainty," 2nd ed., Wiley, New York, 2012.

[2] R. De Maesschalck, D. Jouan-Rimbaud, and D. L. Massart, "The Mahalanobis Distance," Chemometrics and Intelligent Laboratory Systems, vol. 50, pp. 1-18, 2000.

[3] C. Deutsch, "Cell Declustering Parameter Selection," University of Alberta, 2015.

[4] X. Emery, "Geoestadística," Universidad de Chile, 2019.

[5] F. Faraj, "A Simple Unsupervised Classification Workflow for Defining Geological Domains Using Multivariate Data," Mining, Metallurgy & Exploration, pp. 1609–1623, 2021.

[6] R. Fustos, "Descubrimiento de unidades geometalúrgicas por medio de análisis de conglomerados geoestadístico," Doctoral Thesis, Universidad de Chile, 2017.

[7] J. González, "Definition of Resource Estimation Domains Using Machine-Learning Techniques," Universidad de Chile, 2022.

[8] L. Hubert and P. Arabie, "Comparing Partitions," Journal of Classification, vol. 2, no. 1, pp. 193–218, 1985.

[9] L. Kaufman and P. Rousseeuw, "Finding Groups in Data: An Introduction to Cluster Analysis", John Wiley & Sons, pp. 68-88, 1990.

[10] T. Romary, F. Ors, J. Rivoirard, and J. Deraisme, "Unsupervised Classification of Multivariate Geostatistical Data: Two Algorithms," Computers & Geosciences, vol. 85, part B, pp. 96-103, 2015.

[11] M. J. Warrens and H. van der Hoef, "Understanding the Adjusted Rand Index and Other Partition Comparison Indices Based on Counting Object Pairs," Journal of Classification, vol. 39, no. 3, pp. 487–509, 2022.

[12] K. Y. Yeung and W. L. Ruzzo, "Details of the Adjusted Rand Index and Clustering Algorithms," Bioinformatics, vol. 17, no. 9, pp. 763–774, 2001.