ANALYZING SIMILARITIES AMONG PIT LIMIT SCENARIOS IN BLOCK MODEL DOMAIN USING VECTORIZATION AND UNSUPERVISED MACHINE LEARNING

Dimitar Kaykov – dimitar.kaykov@gmail.com

ABSTRACT
Analyzing the influence of various factors on the boundaries of an open pit remains a complex and time-consuming process. Moreover, visual inspection between different contours continues to be a commonly used method to this day despite being impractical for comparing multiple pit alternatives. Applying Hamming and Jaccard distance measures are viable alternatives, on the other hand, as they demand less computational time and provide a relatively sufficient discriminative power between the geometry of two arbitrary pit limits in block model domain. However, they do not allow for their further segregation based on their profitability and have limited discriminative power in terms of pit limit variability. Therefore, to facilitate the multi-criterial comparison between pairs of design alternatives, it is possible to reduce the dimensionality of the block model space while minimizing information loss and establishing the similarity between the considered pairs of pit limits in the latent space. To achieve this, vectorization was applied to the block model, followed by dimensionality reduction using several techniques – PCA and UMAP. The similarity of the resulting low-dimensional representations was then evaluated through visual comparison based on 2D scatter plots, relevant dendrograms, as well as a plan view of pit limits variability.
Key words: *block model, pit limits, dimensionality reduction, clustering*

РЕЗЮМЕ
Анализирането на влиянието на различни фактори върху границите на открития рудник е сложен и трудоемък процес. Визуалната проверка като метод за сравнение все още остава често използван метод, въпреки че не е ефективен при необходимост от сравняване на множество варианти. От друга страна използването на мерките за разстояние на Hamming и Jaccard не изискват значително изчислително време и същевременно предоставят задоволително разграничение между два произволни крайни контура, представени като блокови модели. Въпреки това, тези разстояния не позволяват контурите да бъдат сравнявани въз основа на тяхната рентабилност и ограничено разграничават изменчивостта на границите на рудника. С цел да се осъществи сравнението на двойка варианти за крайния контур на открития рудник по няколко критерия, би могло да се намали размерността на пространството на блоковия модел с минимална загуба на информация. От своя страна, сходството между двойките разглеждани варианти за крайния контур може да се установи в латентно пространство. За целта бе използвана векторизация на блоковия модел, последван от намаляването на неговата размерност посредством методите PCA и UMAP. Сходството на контурите на рудниците чрез така получените резултати бе сравнено чрез 2D точкова диаграма, съответстващите им дендрограми и планов изглед на изменчивостта на границите на рудника.
Ключови думи: *блоков модел, граници на открития рудник, намаляване на размерността на пространството, клъстерен анализ*

Introduction

Optimization of open-pit mine remains a critical domain for improving operational efficiency and economic outcomes in mining engineering. In recent years, the computational efficiency of pit limits optimization algorithms has significantly improved in terms of flexibility (Morales Valera et al., 2015) and time complexity (Poniewierski, 2017). In turn, this allows for the consideration of multiple pit scenarios under uncertain conditions regarding ore grade variability, market price volatility, geotechnical uncertainty and other operational

factors. While traditional Design of Experiments (DOE) frameworks have long supported quantitative analysis in multiple engineering domains, e.g. mechanical, chemical, materials engineering, etc. (Montgomery, 2017), Design of Computer Experiments (DOCE) presents a compelling and practical framework for integrating machine learning methods into traditional and emerging modeling workflows. Moreover, it can be considered as a complementary tool to the application of high-performing optimization algorithms. Therefore, this paper explores the use of experimental design to construct efficient approximations of a complex computational model (pit optimization algorithm), enabling scalable, predictive insights across various pit design scenarios under the influence of multiple uncertain conditions. In particular, it highlights the untapped potential of unsupervised machine learning techniques, such as clustering and dimensionality reduction, uncovering latent structures within high-dimensional datasets for enhancing pit design sensitivity studies and further exploration of the solution space. This approach can prove to be suitable when working with small datasets, where a rigorous metamodel is not feasible due to lower reliability of the regression model or when generating additional scenarios for expanding the dataset is time-consuming. These approaches promise to augment current methodologies with greater flexibility, reduced computation time and deeper interpretability of the influence of external or internal factors.

Computer experiments in engineering design

The integration of machine learning models in engineering design through computer experiments often centers on creating metamodels. They can be regarded as simplified yet sufficiently accurate analogues of complex computational models that simulate real-world behavior under varying conditions. These metamodels are typically developed using experimental designs (DOE or DOCE) through deterministic computer experiments. Carefully selected combinations of input values serve as inputs to the base computational model and the resulting outputs inform the construction of a regression-based metamodel. This model mimics the original system's behavior but with significantly reduced complexity. Hence, these predictive models allow designers to evaluate input-output relationships beyond observed data, offering foresight into performance outcomes. Moreover, it is suitable for integration into optimization procedures as a surrogate model. As a result, metamodels significantly decrease computational and manual effort while enhancing design flexibility by quantifying the influence of individual parameters and reducing bias in decision-making. However, building an effective metamodel isn't always feasible, particularly when faced with numerous uncertain factors, whose combinations grow exponentially. As a result, spatial filling issues related to the experimental design can undermine model accuracy, if not addressed properly. To ensure reliable predictive power and generalizability, well-distributed samples across the design space are used as inputs (Montgomery, 2017).

In open-pit mine optimization, the use of Design of Experiments (DOE) can provide a structured framework for quantifying the influence of key input variables (e.g., slope angles, operational costs, commodity prices, etc.) on both the geometric configuration of the pit limit and its associated Net Present Value (NPV), as well as selecting a feasible pushback direction (Poblete et al., 2021). The prevailing research focus tends to favor stochastic methods by analyzing the variability of optimization results under different inputs (Whittle et al., 2007; Kumral, 2010) or by employing a stochastic optimization algorithm (Albor, Dimitrakopoulos, 2009; Dimitrakopoulos, Lamghari, 2022), which have been more extensively researched within the domain. A general review of current literature reveals that the application of unsupervised machine learning techniques as a supplementary framework for pit design (grounded in computer experiments) remains relatively underexamined, although these methods have been successfully used for feature extraction in other mining-related problems (Li et al., 2019a; Li et al., 2019b). Accordingly, the current paper seeks to employ such methods within the context of ultimate pit design, aiming to uncover latent patterns among the array of design alternatives and thereby augment established approaches in mine planning and pit design.

Complementing Computer Experiments via Unsupervised Machine Learning Methods

Unsupervised machine learning encompasses two primary methodological families: clustering and dimensionality reduction. Though conceptually distinct, they are frequently applied in sequence to preprocess

raw data for supervised learning or to extract interpretable features. Clustering methods analyze object similarity using quantitative metrics such as Euclidean, Manhattan or Hamming distances. These metrics infer similarity inversely with distance. However, clustering becomes ineffective in high-dimensional contexts due to the "curse of dimensionality", where traditional distance metrics lose interpretive value and data sparsity hinders meaningful analysis (https://www.cs.cornell.edu/). Dimensionality reduction techniques address these challenges by transforming high-dimensional data into lower-dimensional latent representations. This facilitates both computational efficiency and interpretability. In particular, non-linear data structures benefit from manifold learning methods, which preserve local and global relationships through topological representations, often relying on geodesic distance rather than linear measures.

Principal Component Analysis (PCA) reduces dimensionality by transforming correlated features into orthogonal components that capture the most variance. It uses a linear transformation based on the eigenvalues and eigenvectors of the covariance matrix (Bishop, 2006). Components are ranked by their eigenvalues which enables the adoption of a significantly lower dimensional vector with negligible loss of information regarding the variability of the data. As PCA is based on linear correlations, its limits when used with complex, nonlinear data are the inability to entirely capture such structures in a lower space representation. In such cases alternatives like Kernel PCA or more advanced methods are preferred.

Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) is a relatively new method for dimensionality reduction when working with multidimensional data, combining both functionality and speed. The UMAP method has theoretical foundations based on topology and therefore provides an opportunity to consider the neighborhood between individual points lying on complex manifolds. For this reason, UMAP is often used in tasks related to large multidimensional datasets (e.g., when working with natural language text data or images). Unlike t-SNE (t-distributed Stochastic Neighbor Embedding), UMAP has the ability to preserve both the local and global structure of the data through a set of tunable hyperparameters. UMAP also has the advantage of being faster to run compared to t-SNE, which makes it a highly preferred method (https://umap-learn.readthedocs.io/en/latest/).

Methodology

By applying a 3-level factorial design (Screening design), each variable is evaluated across three distinct levels, enabling the exploration of non-linear relationships and interactive effects (https://blog.minitab.com). This approach allows advanced scenario analysis, which in turn provides a comprehensive understanding of how variations in input parameters translate into changes in pit depth, contour and wall geometry, while simultaneously tracking the economic viability of each scenario. The integration of geotechnical constraints within the DOE framework further refines the simulation outcomes, ensuring that modeled pit geometries are both physically and economically feasible. It should be stated that the selected values of the set of parameters are chosen for illustrative purposes (Kaykov et al., 2023).

Table 1. Assumed 3-level Screening design for input parameters of the pit optimization software

Input variable (DOE value in brackets)	Min (-1)	Base case (0)	Max (1)
A. Bias of predicted Cu grade (Actual vs. Assumed)	-15 %	0%	+15%
B. Bias of predicted Au grade (Actual vs. Assumed)	-15 %	0%	+15%
C. Cu Price, USD/t	6 000	8 000	10 000
D. Au Price, USD/g	35	55	75
E. Cu recovery during processing	0.80	0.85	0.90
F. Au recovery during processing	0.65	0.73	0.80
G. Pit bottom width, m	60	90	120
H. Pit working width, m	90	120	150
J. Overall slope angle, °	35	40	45
K. Mining costs, USD/t	3.40	3.70	4.00
L. Processing costs, USD/t	10.63	12.50	14.38
M. Ore recovery during mining	0.90	0.93	0.95

The screening design method employed in the program is referred to being multipurpose and flexible, used specifically for detecting quadratic effects, despite having a lower detection capability. However, compared to a Central Composite design or a Box-Behnken design, this configuration requires less runs to achieve similar results. Ultimately, the three-level Design of Experiments (DOE) framework strengthens decision-making by clarifying parameter sensitivities and uncovering optimal configurations that maximize Net Present Value (NPV) across the classical scenario spectrum – Optimistic, Realistic and Pessimistic. The software used for the optimization problem is MiningMath v2.3.52 due to is adopted Direct Block Scheduling approach (https://miningmath.com/2024/08/29/direct-block-scheduling-algorithms/). Output files were subjected to detailed analysis using Python v3.13.2 (https://www.python.org/downloads/) and the packages Pandas (https://pandas.pydata.org/), Matplotlib (https://matplotlib.org/), Scikit-learn (https://scikit-learn.org/stable/index.html). Figure 2 illustrates the assumed screening experimental design, generated via Minitab (https://www.minitab.com/en-us/) with a total of 25 runs, one of which is the base-case scenario (i.e., the central point of the design).

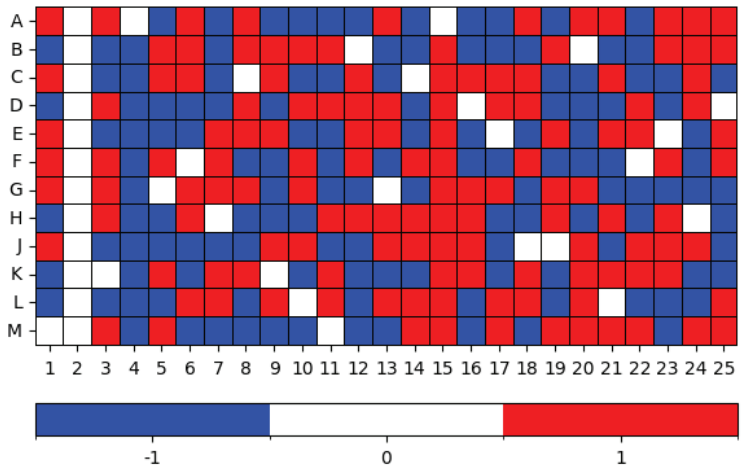

Figure 1. Visual representation of the adopted experimental design used for modelling 25 scenarios

Vectorization is extensively used in other similar data-intense areas, such as image processing, which in terms of its principle analytic framework is very close to the one used in geological block modelling. Given a block model matrix B with n rows and m columns (features), the vectorization's transformation produces the following output:

$$vec(B) = [b_{11},,\cdots b_{1m},\cdots b_{n1},\cdots b_{nm}],\ where\ B = \begin{bmatrix} b_{11} & \cdots & b_{1m} \\ \vdots & \ddots & \vdots \\ b_{n1} & \cdots & b_{nm} \end{bmatrix} \quad (1)$$

The result from vectorization is a high dimensional vector which represents the state of the ultimate pit, as provided via the solver program. It should be noted that the features of this vector (b_{ij}) have a dimension size equal to the product of the number of blocks of the model and the number of features (channels) per block (geochemical composition, geotechnical features and other relevant attributes). Hence, if the initial block model matrix B is $n \times m$, then $vec(B)$ would be represented in $R^{n \times m}$. A collection of multiple pit limit scenarios with their corresponding vectorized block model representations can be tabulated in a new $s \times (n \times m)$ matrix (where s is the number of considered scenarios) and used as a structured dataset for further analysis.

After the initial stage, the following preprocessing step leads to the use of PCA to extract these primary features which are highly responsible for explaining the variance of the vectorized block model values. The established components can be either used as compound features for further interpretation or be further reduced to 2D space through a manifold learning method (UMAP).
To distinguish the discrimination power of the dimensionality reduction approaches to established engineering practices, two more conventional distance measures were initially applied – Hamming and Jaccard distance, which were used to estimate the dissimilarity between pit limit pairs in block model domain. It is worth mentioning that both distance measures rely only on bitwise operations, based on an auxiliary value, signifying whether each block belongs to the ultimate pit limit for the scenario considered. Despite being computationally efficient, however, they do not possess adequate discriminative power to differentiate certain scenarios. Figure 2. Illustrates the applied methodology used for pit limits scenario analysis based on unsupervised machine learning.

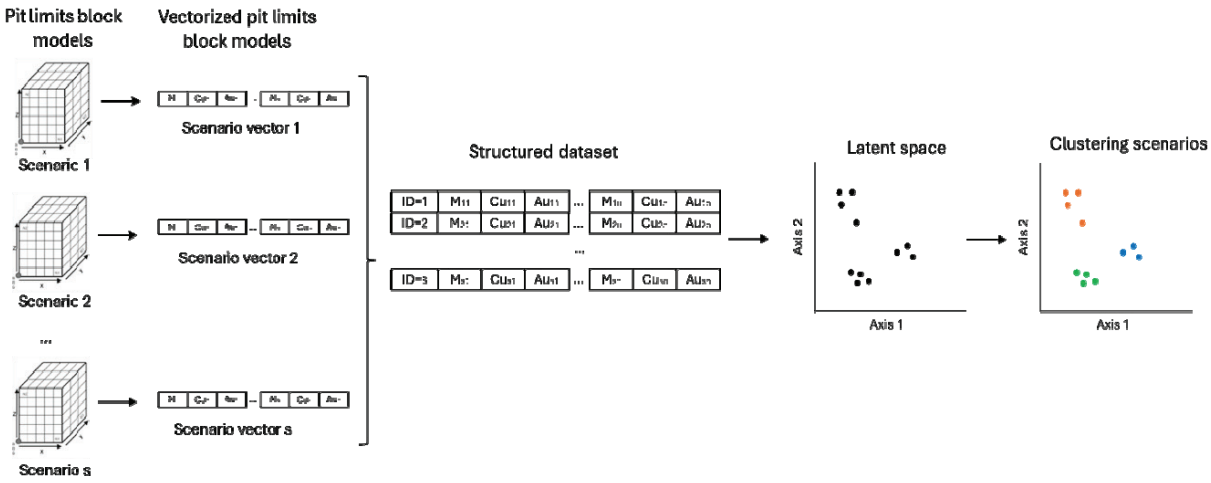

Figure 2. Applied methodology for advanced pit limits scenario analysis

Results

This study used the open-sourced Marvin deposit's block model as a benchmark for the applied methodology, which consists of a total of 53 271 initially unmined blocks (https://miningmath.com/docs/knowledgebase/formatting-data/datasets/). Each block in the utilized model matrix has the following features: binary variable signifying whether a block is part of the ultimate pit limit (i.e. whether it is mined or not), Cu grade (%) and Au grade (g/t). Hence, after vectorization each pit limit scenario is represented as a vector with 159 813 features. These features can successfully be reduced to 10 principal components which contribute to explaining 92.28% of the total variance, as their respective cumulative contributions are as follows:

[0.3600, 0.5346, 0.6392, 0.7260, 0.7811, 0.8292, 0.8629, 0.8905, 0.9092, 0.9228] (2)

The first two components contribute to explaining 53.46% of the total variance and therefore are insufficient for independent analysis. Regardless, they were used for visualization to illustrate the 2D projection of the nonlinear manifold. Figure 3 shows the latent space representation of each pit scenario, after the application of UMAP with their 10-dimensional vector based on PCA.

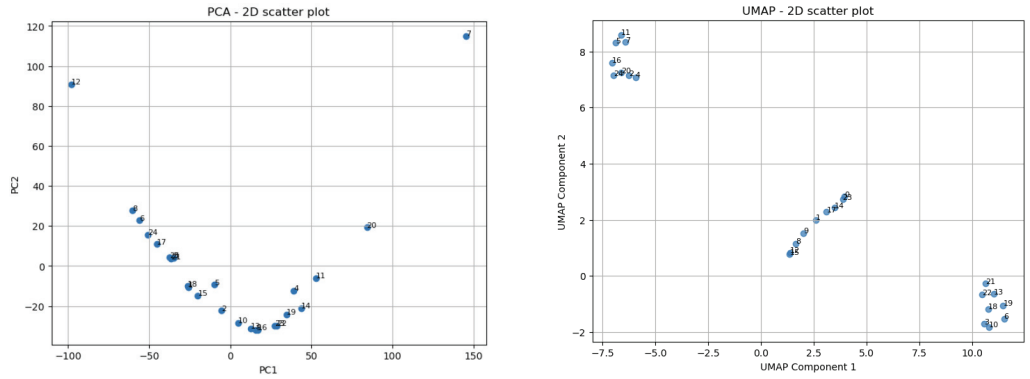

Figure 3. Visual representation of the 2D latent space obtained from the use of PCA (left) and UMAP (right)

It should also be noted that UMAP is a stochastic model and hence, its results may vary from run to run. Regardless, three primary clusters can be observed after the use of UMAP, which can be interpreted as optimistic, pessimistic and contours in between the spectrum. This confirms that the latent space representation of the scenarios considered maintains the same logical coherence of the assumptions for the experimental design. For the sake of comparison, all scenarios were compared in a pairwise manner via Hierarchical Clustering Analysis (HCA), due to its flexibility of merging individual observations, as well as its complementary visualization tool. Hence, results obtained are provided through a dendrogram. In addition to the latent space representation, dendrograms based on the distance matrix for both the Hamming and Jaccard distances for the raw block model data are also built (Figure 4). While the latent space representations were compared via the Manhattan distance, it should also be noted that all dendrograms employ the Complete linkage method for finer discrimination among scenarios.

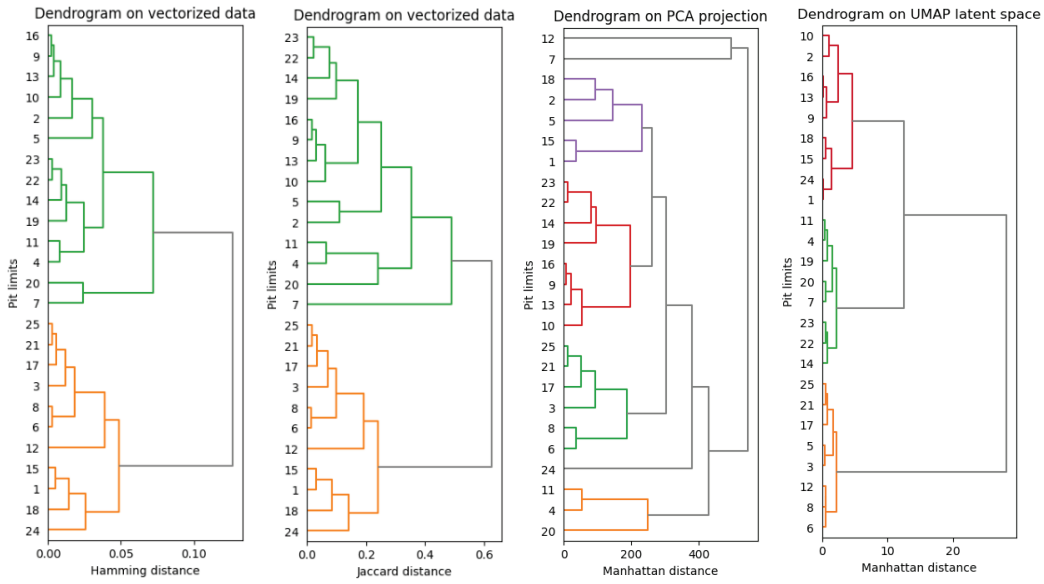

Figure 4. Dendrograms, signifying similarity among considered scenarios before and after the application of dimensionality reduction methods

Figure 5 illustrates the geometric aspects of all obtained clustering solutions based on the two major pit clusters. Cluster 1 corresponds to a more favorable geological bias with respect to grades estimation, as well as overall more favorable commodity market prices. The opposite is true for Cluster 2. It is important to note that the Root Mean Square Deviation was calculated in relation to the Base case scenario (the central point in the experimental design) regarding the pit elevation.

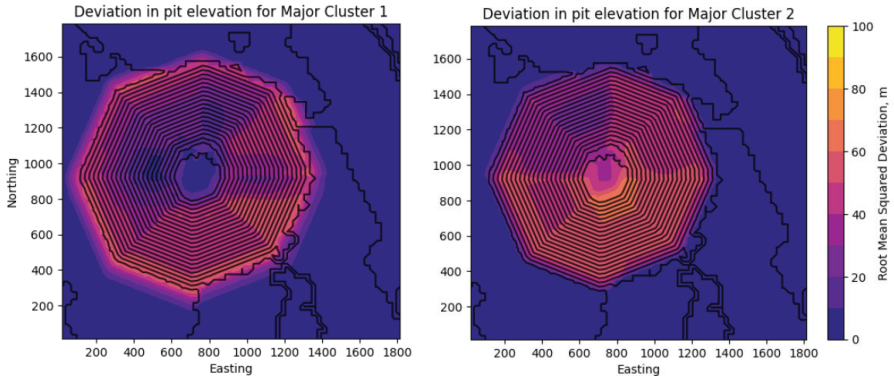

Figure 5. Deviation from Base case (Scenario 2) in pit elevation for major pit limit clusters obtained from HCA based on Hamming and Jaccard distances

Results also show that Hamming and Jaccard distances, applied on the binary values of the block models considered, both have an adequate discriminative power to differentiate pit geometries based on volumetric differences. However, when designing alternatives which need to reflect economic and geological uncertainty, this approach appears to provide less satisfactory results. As computationally efficient as this approach based on sparse vectors may be, naturally, it fails to distinguish certain similarities among pit limits under economic or geotechnical considerations. Furthermore, clustering based on PCA alone may provide a different perspective due to the detected "outlier" pits, as it does not account for the intrinsic manifold structure upon which the solutions reside. Hence, a latent space representation of the set of considered scenarios, that respects ore grades, allows for a more compound comparison. As observed, UMAP seems to have good agreement with the geometric-based approach based on Hamming and Jaccard distances in terms of pairwise comparison. At the same time, it provides an improved way of grouping pit limits according to the specifics of the blocks' geochemical features. Table 2 represents the conditions under which these similar pit limits occur, as well as their mean NPV and its standard deviation. It should be noted that the input variables for the optimization problem were binned into five major factors, which assume the mean of all DOE values across the factor-related variables and pit limits in the corresponding cluster.

Table 2. Influence of mean binned input parameters (factors) on NPV for pit limit scenario clusters obtained via UMAP and HCA

Pit clusters	**Cluster 1**	**Cluster 2**	**Cluster 3**
Pit limit scenario IDs	1, 2, 9, 10, 13, 15, 16, 18, 24	4, 7, 11, 14, 19, 20, 22, 23	3, 5, 6, 8, 12, 17, 21, 25
Geological estimation bias (A and B)	0.06 fair	-0.25 unfavorable	0.19 favorable
Commodity market state (C and D)	0.39 very favorable	-0.69 very unfavorable	0.25 favorable
Processing and metallurgical factors* (E, F and L)	-0.04 fair	0.08 fair	-0.04 fair
Geotechnical and design parameters* (G, H and J)	-0.26 unfavorable	0.00 fair	0.29 favorable
Mining and operational factors* (K and M)	0.22 favorable	-0.19 unfavorable	-0.06 fair
Mean NPV, 10^3 USD	1545.58	626.66	1256.11
Standard deviation of NPV, 10^3 USD	547.79	269.40	273.45

* inverse transformation from raw DOE scores, matching increase in utility to project

The clusters obtained from UMAP are plotted in Figure 6. Once more, the Root Mean Square Deviation of pit limit elevation was calculated in comparison to the Base case scenario.

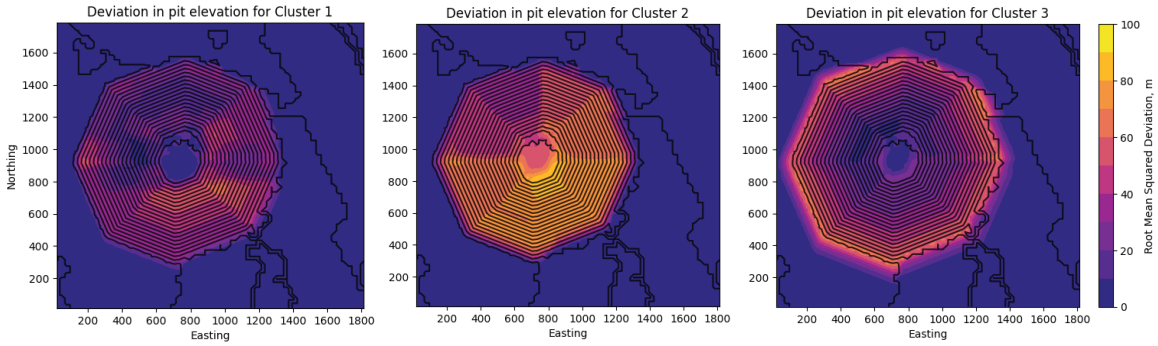

Figure 6. Deviation from Base case (Scenario 2) in pit elevation for pit limit clusters obtained from UMAP-based latent space

Based on the results derived from HCA and UMAP, the well-established practical knowledge of pit limit variability holds true: elevated commodity prices, underestimation of reserves and favorable geotechnical conditions tend to result in expanded pit limits and a higher threshold for the stripping ratio, whereas lower prices and reserves overestimation typically lead to shallower pit designs. Intermediate price scenarios, mixed with unfavorable geotechnical conditions, often result in the inhibited development of specific pit sectors, while cost reduction in mining and processing operations may lead to accessing otherwise infeasible sections.

Discussion and Future Work

As demonstrated, dimensionality reduction techniques are powerful tools that are crucial for the human interpretation of multidimensional data and highly complex concepts related to geological features and design considerations of pit limits under uncertainty. However, a significant drawback of using dimensionality reduction methods is the inability to completely explain the latent space and its emerging properties. Due to the abstract nature of latent spaces, it may become challenging to associate extracted features with the humanly interpretable geochemical parameters or the pit's physical design attributes. Regardless, these methods provide a crucial perspective on the underlying patterns among pit design alternatives under different uncertain conditions. Moreover, the derived latent space values can be used not only for visual inspection and group certain scenarios based on their overall similarity, but also for regression modelling and prediction tasks. It should be noted that applying vectorization is not the only valid method that could potentially work well in a block model context. As vectorization does not take into consideration spatial patterns among adjacent blocks, the use of Convolutional Neural Networks (CNNs), as well as Variational Autoencoders (VAEs), could prove to be promising alternatives. These approaches are capable of learning intrinsic spatial features which are inherent in geological data. Moreover, apart from geochemical features, certain geotechnical properties can also be embedded in the analysis. However, these advanced dimensionality reduction methods need to be compared in terms of their robustness, as well as time complexity, scalability and interpretability under varying data conditions and operational constraints.

Conclusions

What differentiates the generally adopted understanding of pit design from the applied approach in this study is the incorporation of a quantifiable measure of pit limit variability, as well as the expected NPV and its variance across plausible economic scenarios. Hence, the segregation of pit-sensitive regions compared to the base case pit model provides further granularity for pit design needs. In addition, the methodological coherence with conventional models, combined with the empirical validation through scenario analysis, reinforces the robustness and practical relevance of the unsupervised machine learning framework and the feature extraction

results obtained by PCA and UMAP. Should variability in ore grades and rock type densities be required to be taken into account, applying dimensionality reduction becomes the more efficient means of embedding these features into pairwise comparisons, compared to bitwise algebra based on pit geometry. Last but not least, applying a 3-level experimental design with as little as 25 observations based on a Screening design proves to be sufficient for an overview of pit limit variability, deeming the adopted methodology potentially viable for prefeasibility and feasibility studies.

References

1. Albor, F., & Dimitrakopoulos, R. (2009). Stochastic mine design optimization based on simulated annealing: Pit limits, production schedules, multiple orebody scenarios and sensitivity analysis. Mining Technology, 118(2), 79–90. https://doi.org/10.1179/037178409X12541250836860
2. Dimitrakopoulos, R., & Lamghari, A. (2022). Simultaneous stochastic optimization of mining complexes/mineral value chains: An overview of concepts, examples and comparisons. International Journal of Mining, Reclamation and Environment. https://doi.org/10.1080/17480930.2022.2065730
3. Kaykov, D., Arsova, K., Kutsarov, K. (2023). A Design of Experiments Approach for Establishing Key Factors Determining the Profitability of Ultimate Pit Limits. 10.5281/zenodo.8334040.
4. Kumral, M. (2010) Robust stochastic mine production scheduling, Engineering Optimization, 42:6, 567-579, DOI: 10.1080/03052150903353336
5. Li, S., de Werk, M., St-Pierre, L., & Kumral, M. (2019a). Dimensioning a stockpile operation using principal component analysis. International Journal of Minerals, Metallurgy and Materials, 26(12), 1485–1494. https://doi.org/10.1007/s12613-019-1849-y
6. Li, S., Sari, Y. A., & Kumral, M. (2019b). New approaches to cognitive work analysis through latent variable modeling in mining operations. International Journal of Mining Science and Technology, 29(4). https://doi.org/10.1016/j.ijmst.2019.06.014
7. Montgomery, D. C. (2017). Design and analysis of experiments (9th ed.). John Wiley & Sons.
8. Morales Varela, N., Jélvez, E., Nancel-Penard, P., Marinho, A., & Guimaraes, O. (2015). A comparison of conventional and direct block scheduling methods for open-pit mine production scheduling.
9. Poblete, C., Smith, R., Das, R., Romero, J., Van, N., & Hout, D. (2021). Using design of experiments to improve strategic mine planning.
10. Poniewierski, J. (2017). Pseudoflow explained: A discussion of Deswik pseudoflow pit optimisation in comparison to Whittle LG pit optimisation. Accessed through: https://www.deswik.com/wp-content/uploads/2017/05/Pseudoflow-Explained.pdf
11. Whittle, G., Stange, W., & Hanson, N. (2007). Optimising project value and robustness. Project Evaluation Conference, Melbourne, VIC.
12. https://blog.minitab.com/en/bruno-scibilia/definitive-screening-designs-for-products-and-processes-optimization
13. https://matplotlib.org/
14. https://miningmath.com/2024/08/29/direct-block-scheduling-algorithms/
15. https://miningmath.com/docs/knowledgebase/formatting-data/datasets/
16. https://pandas.pydata.org/
17. https://scikit-learn.org/stable/index.html
18. https://umap-learn.readthedocs.io/en/latest/
19. https://www.cs.cornell.edu/courses/cs4780/2022fa/lectures/lecturenote02_kNN.html
20. https://www.minitab.com/en-us/
21. https://www.python.org/downloads/

All online sources were last accessed on 30th June 2025