AI-driven Spatial Data Augmentation for Geological Modelling and Resource Estimation

A. Gole[1] and S. Sullivan[2]

1. Machine Learning Engineer, Maptek Pty Ltd, Adelaide SA 5065 arpit.gole@maptek.com.au
2. Technical Lead, Maptek Pty Ltd, Adelaide SA 5065 steve.sullivan@maptek.com.au

ABSTRACT

In real-world data sets, missing values are unavoidable for various reasons. These missing values are typically represented by NaNs, default placeholders, or simply left as blank entries. Depending on the extent of missing data, this can significantly reduce the performance of statistical methods. Additionally, data sets with missing values are incompatible with many machine learning techniques, including random forests, regression models, and neural networks, which rely on the assumption that all features contain complete and relevant information related to the task at hand.

Geological data sets, which capture the 3D representation of a deposit using geological field observations, survey data, drill hole information, and assay grades, are no exception. In geological modelling, it is extremely rare to encounter complete data sets without any missing values. A common but simplistic approach is to exclude observations with missing values altogether. However, when a large portion of the data set contains missing entries, removing those records leads to substantial information loss. This highlights the importance of integrating effective missing data imputation techniques into the data preprocessing workflow—a process that presents several challenges (Rahm and Do, 2000).

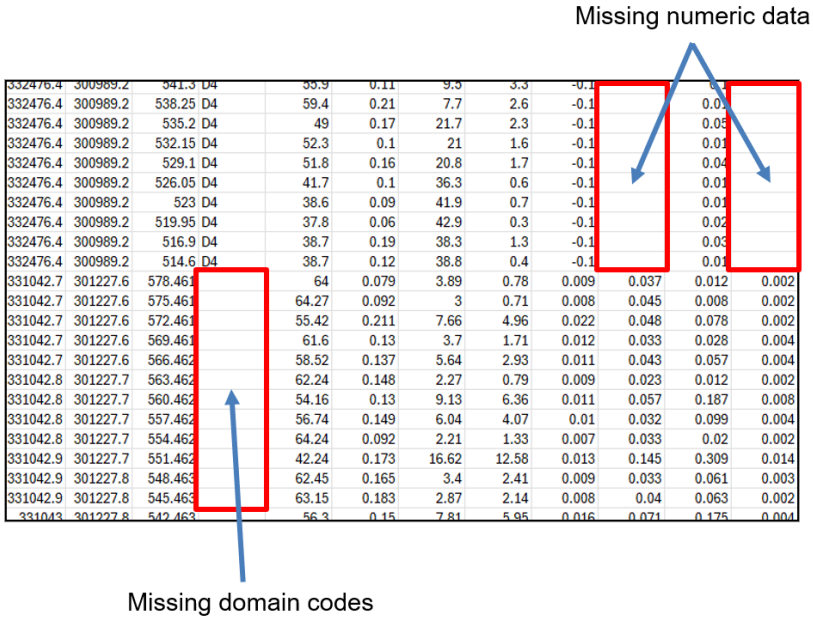
Effective imputation methods must account for naturally occurring geological patterns, such as the formation and spatial continuity of rock types, the proportions of different lithologies, and the uncertainty or potential misinterpretation introduced by incomplete data. A method has been developed that captures and reformulates a high level of correlation with the existing geological data. Performance gains using newly imputed data as input to machine learning processes are evaluated using several metrics to provide geological plausibility for the method.

Presented at APCOM 2025
held in Perth, Western Australia 10-13 August 2025

INTRODUCTION

For many reasons, real-world data sets always contain missing values. In diamond drilling, there is always some core loss; an underground drive may only expose part of the orebody for sampling (King, McMahon and Bujtor,1982). Generally, these missing values are represented in a data set by either NaNs, some placeholder value, or even a blank value. In other circumstances, the missing data are ignored completely. Depending on the volume of the missing data, it can degrade the performance of statistical methods. Furthermore, such data sets are incompatible with machine learning techniques like random forest, regression and more, which assume that the entire data set contains numerical values and all the features hold valuable information related to the task.

A geological data set contains a three-dimensional representation of the ore deposit based on geological field observations, survey data, drill hole data and chemical assay elemental grade data. In the case of geological modelling, fully observed data sets without any missing values are a rare occurrence, as shown in Figure 1.



332476.4	300989.2	541.3	D4	55.9	0.11	9.5	3.3	-0.1		0.1	
332476.4	300989.2	538.25	D4	59.4	0.21	7.7	2.6	-0.1		0.03	
332476.4	300989.2	535.2	D4	49	0.17	21.7	2.3	-0.1		0.05	
332476.4	300989.2	532.15	D4	52.3	0.1	21	1.6	-0.1		0.01	
332476.4	300989.2	529.1	D4	51.8	0.16	20.8	1.7	-0.1		0.04	
332476.4	300989.2	526.05	D4	41.7	0.1	36.3	0.6	-0.1		0.01	
332476.4	300989.2	523	D4	38.6	0.09	41.9	0.7	-0.1		0.01	
332476.4	300989.2	519.95	D4	37.8	0.06	42.9	0.3	-0.1		0.02	
332476.4	300989.2	516.9	D4	38.7	0.19	38.3	1.3	-0.1		0.03	
332476.4	300989.2	514.6	D4	38.7	0.12	38.8	0.4	-0.1		0.01	
331042.7	301227.6	578.461		64	0.079	3.89	0.78	0.009	0.037	0.012	0.002
331042.7	301227.6	575.461		64.27	0.092	3	0.71	0.008	0.045	0.008	0.002
331042.7	301227.6	572.461		55.42	0.211	7.66	4.96	0.022	0.048	0.078	0.002
331042.7	301227.6	569.461		61.6	0.13	3.7	1.71	0.012	0.033	0.028	0.004
331042.7	301227.6	566.462		58.52	0.137	5.64	2.93	0.011	0.043	0.057	0.004
331042.8	301227.7	563.462		62.24	0.148	2.27	0.79	0.009	0.023	0.012	0.002
331042.8	301227.7	560.462		54.16	0.13	9.13	6.36	0.011	0.057	0.187	0.008
331042.8	301227.7	557.462		56.74	0.149	6.04	4.07	0.01	0.032	0.099	0.004
331042.8	301227.7	554.462		64.24	0.092	2.21	1.33	0.007	0.033	0.02	0.002
331042.9	301227.7	551.462		42.24	0.173	16.62	12.58	0.013	0.145	0.309	0.014
331042.9	301227.8	548.463		62.45	0.165	3.4	2.41	0.009	0.033	0.061	0.003
331042.9	301227.8	545.463		63.15	0.183	2.87	2.14	0.008	0.04	0.063	0.002
331043	301227.8	542.463		56.3	0.15	7.81	5.95	0.016	0.071	0.175	0.004

FIG 1 – A table with analytical data from the case history with missing data attributes.

A naive solution is to ignore observations with missing values. However, dropping a significant proportion of the data set due to many observations having missing values is a primary source of information loss and hence poor modelling outcomes. It is consequential to keep in mind that geological data collection is a costly and time-consuming process. It becomes vital to impute the missing values in such data sets as a part of the data preprocessing workflow, which poses many challenges (Rahm and Do, 2000). While imputing the missing values, the need for a versatile approach that accounts for naturally occurring geological scenarios is required. Geologically acceptable models create rock shapes that exhibit spatial continuity and maintain proportions of rock type(s) comparable with that observed in source drilling data. Imputing additional data should aim to preserve this behaviour whilst improving certainty or lowering the probability of misinterpretation. This ability to efficiently handle, extract meaningful insights, and discover hidden knowledge from massive volumes of partially complete data becomes critical to any real-world artificial intelligence application in this sphere, as is the case with such applications generally (Lee and Yoon, 2017; Tsai *et al,* 2015).

The motivation and practical value of doing this research is to make faster and smarter uncertainty-based decisions in geological resource modelling—directly from the spatial geological data—especially in the presence of only partially complete information. This way of rapid modelling of geological resources bypasses many time-consuming manual processes that burden the resource geologist, allowing them to focus on interpretation and matters of economic interest rather than rote manual work. To accomplish this aim, the study revolves around the following research questions:

- RQ 1: Develop a method to impute the missing lithological domains (categorical rock type) when the chemical assay data (multi-variable continuous numerical data) is present.
- RQ 2: Develop a method to impute the missing assay data when the lithological domains (obtained from manual core logging) are present.
- RQ 3: Develop a method to impute partially missing assay data when the lithological domains and some assay values are present.

All three missing data scenarios occur with different data collection strategies. Exploration drilling data—logged by geologists for lithological domains and assayed in the lab for elemental composition—is only captured at low spatial density because it is expensive and occurs early on in the economic life of a mine. RC grade control drilling data – although captured at higher density – is not often logged by geologists for lithological domains because of time pressure and the absence of clean core samples – but is assayed in the lab for elemental composition. Blasthole drilling data also is often assayed for elemental composition just prior to mining and is at very high spatial density because these holes host the explosives used to fracture the rock for mining, but no domaining and lithological interpretation occurs on these data. The need for RQ 3 arises when during the initial period of exploration at a mining site the geologists/engineers are interested in certain elements and their assay grades but down the line, they become interested in an increased (or possibly reduced) set of elements and their assay grade values. A justifiable understanding of the above research questions in context with these different types of spatial distributed geological data will enhance the performance of downstream data-driven geological resource modelling methods. An example of this is Maptek's DomainMCF domain modelling solution (Sullivan *et al,* 2019).

In this work, we work towards developing a method to impute missing geological data to have a fully observed geological data set. Quantitatively, the imputed data exhibits a high level of statistical correlation with the existing geological data. Qualitatively, the veracity of the method is assessed using the downstream DomainMCF solution to compare models with and without imputed data in the hands of geologists. These approaches combine to build a sound understanding of the various assumptions and limitations of the technique with suitable geological explanations.

METHODOLOGY

In our context, geological data sets comprise both continuous numerical and categorical data. Additionally, the data set is spatially three-dimensional and is full of complex non-linear spatial interactions that are notably hard to analytically model. Missing data is a wreaking problem in the machine learning paradigm, and the current implementation of DomainMCF at Maptek is no exception. The present technique uses a neural network to implicitly define domain shapes by training on sample data consisting of logged lithological domains and assay grades – with the drawback that every sample needs to have the complete data (Sullivan *et al*, 2019; Kapageridis *et al,* 2021). Quality data is paramount as an input for training the neural network. Using machine learning techniques for imputing the missing values in the data, which then acts as an input to train another machine learning model could lead to a fundamental problem – the data incest problem (Krishnamurthy and Hamdi, 2013). To construct a statistically plausible full data set from a partial data set, let's first briefly discuss categories of missing data and then available and proposed methods for imputing them.

Three categories of missing data (Mislevy, 1991):

1. Missing at Random (MAR): Missing data are independent of the unobserved data, but systemically related to values of the observed data. For example, performing selective logging for a rock type based on experience and leaving certain rock types not logged.
2. Missing Not at Random (MNAR): Missing data are systemically related to the values of the unobserved data. For example, while drilling, some of the rock gets washed away due to its softness, preventing sampling of that part of the rock type.
3. Missing Completely at Random (MCAR): Missing data are independent of both observed and unobserved data. For example, while performing assay, 5 per cent of the samples go missing due to an event not related to anything geological.

The most prevalent imputation techniques include mean, mode, or median substitution in place of the missing value. Since these methods only consider a single feature for imputing the missing value, they do not consider the uncertainty of the imputed value or cross-correlations with other values. This introduces variance and covariance biases into the imputed data set (Arnab, 2017). Other prominent and much-advanced techniques like k-nearest neighbours (kNN Imputation) (Troyanskaya *et al,* 2001) are used for numerical data. Support vector machines (SVM Imputation) (Honghai *et al,* 2005) are used for classification with the ability to use different kernels to reduce computational cost and increase performance, such as Linear, Polynomial, Laplacian, and Radial Basis Function (RBF).

Kriging (Krige, 1953; Matheron, 1962) is a spatial interpolation technique in geostatistics. Kriging is also known as Gaussian process regression and is the primary technique used in the mining context. It has become a go-to method for optimally predicting the values at unknown points in space using nearby data. It uses a variogram to describe the spatial structures present in the data. It can be used to estimate unknown/missing values given nearby completely observed data. Given a set of observation points X and observation values at these locations $Z(x)$, it can be stated that the estimation at an unobserved location $Z^*(x_0)$ is a weighted mean:

$$Z^*(x_0) = \sum_{i=0}^{N} \lambda_i Z(x_i)$$

Here, N is the size of X and λ is the array of weights. To assure unbiasedness in the kriging system, we do this by constrained optimisation, we constrain the weights to sum to 1. A variogram explains how observed points become more dissimilar with distance. At its core, a variogram is only a function of *distance*, although parameters can vary with direction to model anisotropy. Depending upon the stochastic characteristics of the data distribution, different types of kriging apply: simple kriging to independently estimate unknown numerical variables, indicator kriging to estimate the transition probabilities between categories and co-kriging to model and estimate relationships between multiple co-dependent variables. For kriging in general, all weights must be calculated at each estimation location. For a large data set, this means a large matrix system and so solving can be extremely computationally intensive and hence slow.

To aid the machine learning algorithms, we developed a way to impute the missing values in data sets more effectively. The method assumes that the missing data follows specific patterns, such as when the likelihood of missingness is dependent on observed data rather than unobserved information, ie following the MAR principle. This method is highly adaptable and can be applied across various scenarios. It works by iteratively predicting missing values using models based on other available variables in the data set. Notably, the approach can handle different data types, including continuous, unordered categorical, and ordered categorical variables.

The core of this method is to artificially predict the missing data such that it has a similar data distribution as the observed data. It is a parametric multiple imputation method and the ability to tune the model parameters is a complex task that may depend on domain/prior knowledge. Tuning can have a drastic effect on the method's performance. Large data sets will naturally have huge numbers of variables to include in the imputation process. It may not be possible to identify all the meaningful variables to be used during the imputation process. Selection of such variables to include can be guided by various analyses and or based on domain knowledge, but this poses a risk of bias (Collins, Schafer and Kam, 2001). However, the chosen method works by automatically selecting the variables and imputing the values based on the most correlated variables.

In our context, geological data sets comprise both numerical and categorical data. It is 3D point cloud data, with: (a) x, y and z as the spatial coordinates, which are numerical; (b) lithology (also often called stratigraphy in sedimentary geology) as categorical data; and (c) assay data for elemental (or sometimes mineralogical) concentrations which are numerical data. We need evaluation metrics for both numerical and categorical data. *Geological plausibility of the imputed data set is much more than the quantitative statistical analysis of what has been imputed and involves a degree of subjective assessment by experienced geologists.*

The initial selection of the imputed data is based on quantitative analysis using the below metrics on the cross-validation data. Additionally, as part of this study, geologists are asked to visualise/verify the result for the geological plausibility of the imputed data set.

For categorical variables, the evaluation metric is the accuracy:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where *TP* is the number of *true positive* predictions, *TN* is the number of *true negative* predictions, *FP* is the number of *false positive* predictions, and *FN* is the number of false negative predictions.

For numerical variables, it is the root mean squared error (RMSE):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

Also, kernel density estimate (KDE) plots are plotted for each numerical variable. This helps in the visualisation of the distribution of the numerical values before and after the imputation method is applied to them.

CASE HISTORY

The following case study demonstrates the application of the study method and compares the results of machine learning with and without data imputation.

The Lisheen mine in County Tipperary, Ireland, was mined from 1999 until the mine closure in December 2015 as shown in Figure 2. It produced 22 Mt at 11.6 per cent Zn and 1.9 per cent Pb (Torremans *et al,* 2018). Mineralisation at Lisheen consists of several, largely stratiform, massive sulfide orebodies at or within 30 m of Lower Carboniferous marine carbonates and breccias. The mineralisation is manifest by semi-massive, disseminated, and vein-hosted sulfides in a complex of laterally discontinuous normal faults (Hitzman, 1999).

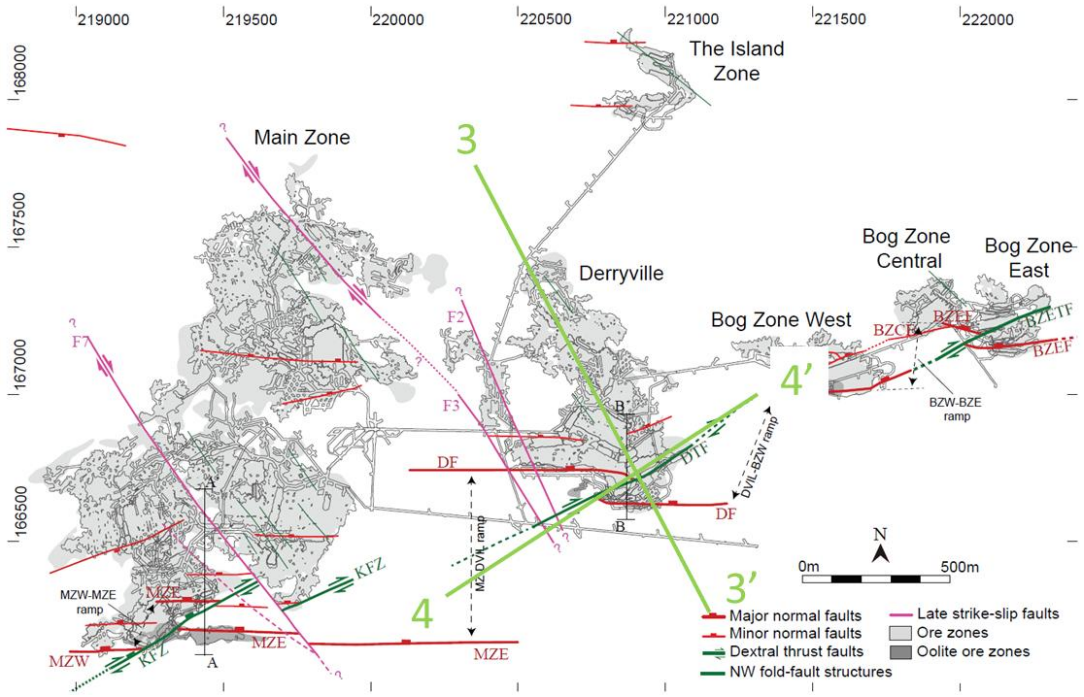


FIG 2 – Structural setting of the Lisheen base metal deposits in Ireland (from Kyne et al (2019) modified from Torremans et al, 2018).

Previous Geological Modelling

The Lisheen deposits were investigated using more than 7000 published holes, drilled from surface during exploration and subsequently from underground during mine production. Holes were geologically logged with more than 40 different domain codes being used to describe the host rocks and mineralised intervals. Chemical analysis of mineralised intervals and surrounding host rocks includes determinations for Zn, Pb and a wide suite of related elements. The data from the Lisheen mining operations was released by the government of Ireland as open source data on 27 June 2019 (Department of the Environment, Climate and Communications, 2019).

Modelling during the operation of the Lisheen mines used sectional interpretation of exploration and production drilling and underground channel sampling/mapping observations. Explicit wireframes were created for the footwall and hanging wall of each mineralised horizon. These two surfaces were then combined to make a solid model of each horizon, which was then used to flag an ore code into a 3D block model. Analytical data was used for estimating grades into the block model using inverse distance and ordinary kriging for the numeric attributes of economic interest.

Modelling using Machine Learning

The origin of the concept of data-driven modelling is discussed in detail by Solomatine *et al* (2009) and was further updated by Montánsa *et al* (2019). The rapid advance of computing power facilitates computer-based discovery and analysis. In the geological context this allows us to generate a documented workflow which updates the model when the input data changes. This could be a manual process but the most benefit is derived from an automated system which is triggered by a change in state in the underlying data.

This current study built a 3D geological model using the latest commercial machine learning engine, Maptek DomainMCF, which has its origins in collaborative research and development with the mining industry as described in Sullivan *et al* (2019). The details of the data preparation, data validation and the results of the geological modelling process using machine learning are presented in Sullivan, 2022.

Modelling using Data Imputation

A subset of the 7000-hole data set from Lisheen was used for testing the impact of the use of data imputation prior to running the machine learning process. This subset was taken from the SW of the orebody as shown in the study area indicated in Figure 3 and included 1618 exploration and underground production drill holes.

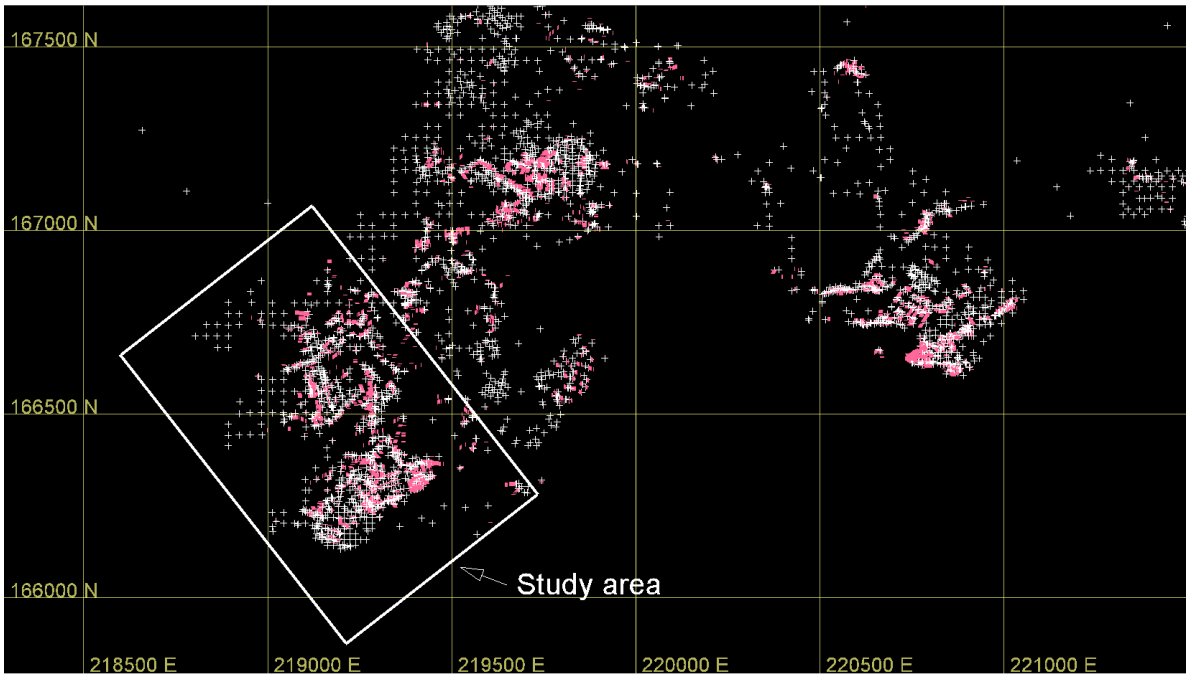


FIG 3 – Lisheen mine data drill hole sample locations with the study area highlighted.

Samples from the subset of drill holes were analysed for Zn, Pb and Fe and reported in per cent (%). Specific gravity (SG) determinations were made from a representative selection of samples and reported in grams per cubic centimetre (g/cm^3). As is the usual practice with semi-massive to massive sulfide orebodies, only drill core samples which have visibly mineralised material are sent for quantitative analysis, saving the cost of analysing barren or waste rock. This resulted in a significant portion of the spatial data points from the drill holes having missing attributes or null values for Zn, Pb, Fe and SG.

A total of 75 159 samples were included in the study area shown in Figure 3. Each data attribute, aside from the spatial coordinates, was well represented but not complete, with missing values of varying proportions as per Table 1.

TABLE 1

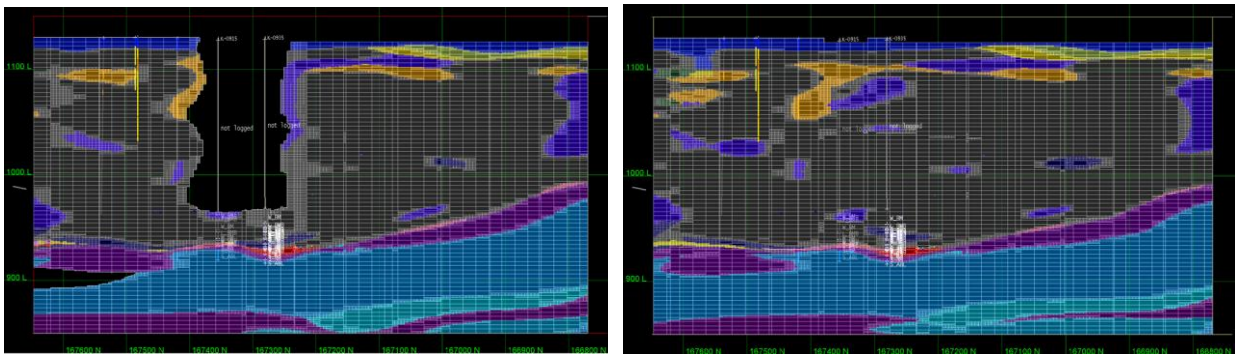
Showing the percent of missing data in the case study.

Attribute	Percentage missing
Fe	49.2
Pb	49.9
Zn	49.2
sg	68.8

As discussed earlier, machine learning requires data sets with a complete/full set of attributes for each spatial location. To achieve this, prior to the consideration of data imputation, rows of data with missing domain codes were deleted, as were rows with missing numeric attributes. As can be seen in Table 1, this would result in almost half the observed data being discarded. This scenario was run and used as a base case for comparison with the subsequent imputed scenario.

Before data is uploaded, it needs to be validated. There are hosts of simple rules that can be applied in this process: checking the validity of collar x, y, and z coordinates; checking *from* and *to* sample intervals are not missing or overlapping; and checking that analytical data uses the same units of measurement. Data validation rules are built into DomainMCF so that data going through to the modelling process is as clean as required. This process is critical as the output models will be directly impacted by erroneous input data.

In the Lisheen data, several drill holes were pre-collared through the overburden and were not geologically logged. When composited, these intervals are given the default code of NONE. If unlogged intervals remain in the data-driven process, a domain model will be generated for NONE as per Figure 4a.



(a) (b)

FIG 4 – (a) A cross-section through the overburden at Lisheen showing the impact of leaving unlogged intervals in the input data to machine learning. (b) The same cross-section shows the impact of the use of imputed data by imputing unlogged intervals.

To remedy the unlogged intervals, the geologist can either manually enter appropriate logging codes based on surrounding drill holes, or delete all unlogged intervals in the input data and allow the machine learning to predict and interpolate relevant domains into the unlogged space shown in Figure 4a. Neither of these options are ideal. Figure 4b shows the impact of applying data imputation to the data missing in Figure 4a. The cross-section through the 3D geology is now complete and a better approximation of the spatial distribution of the domain codes.

Empirically, the result of the domain prediction for the base case: a total of 30 geological domains were predicted from 37 576 samples. This resulted in an overall accuracy of 77 per cent. In the second scenario, data imputation was implemented, allowing all 75 159 samples to be used in the machine-learning process. This scenario was run, and the results of the domain prediction for the imputed data case are as follows: 33 geological domains were predicted, with a strong correlation showing that the method learned well for predictive capabilities. This resulted in an overall accuracy of 84 per cent. We can conclude that using data imputation with this real mine data in the study area allowed an additional three domains to be predicted over the base case. In addition, the accuracy of the domain code prediction was higher when data imputation was included in the modelling process.

Now to compare the distribution of the 37 576 samples before data imputation and for all 75 159 samples after the imputation process. Figure 5 shows the data distribution of the observed data distribution (original data) in blue and the complete data set imputed (imputed data) in red. Assay values are power transformed (Yeo and Johnson, 2000) to improve the interpretability of data distribution by reducing skewness and enhancing the effectiveness of graphical representations to detect the underlying patterns. We can right away see that the data imputation method models the underlying data distribution to a sufficient degree for the assay elements. The general acceptability of the method is taken into account given we only have about 50 per cent of the real data to start the imputation after initial preprocessing.

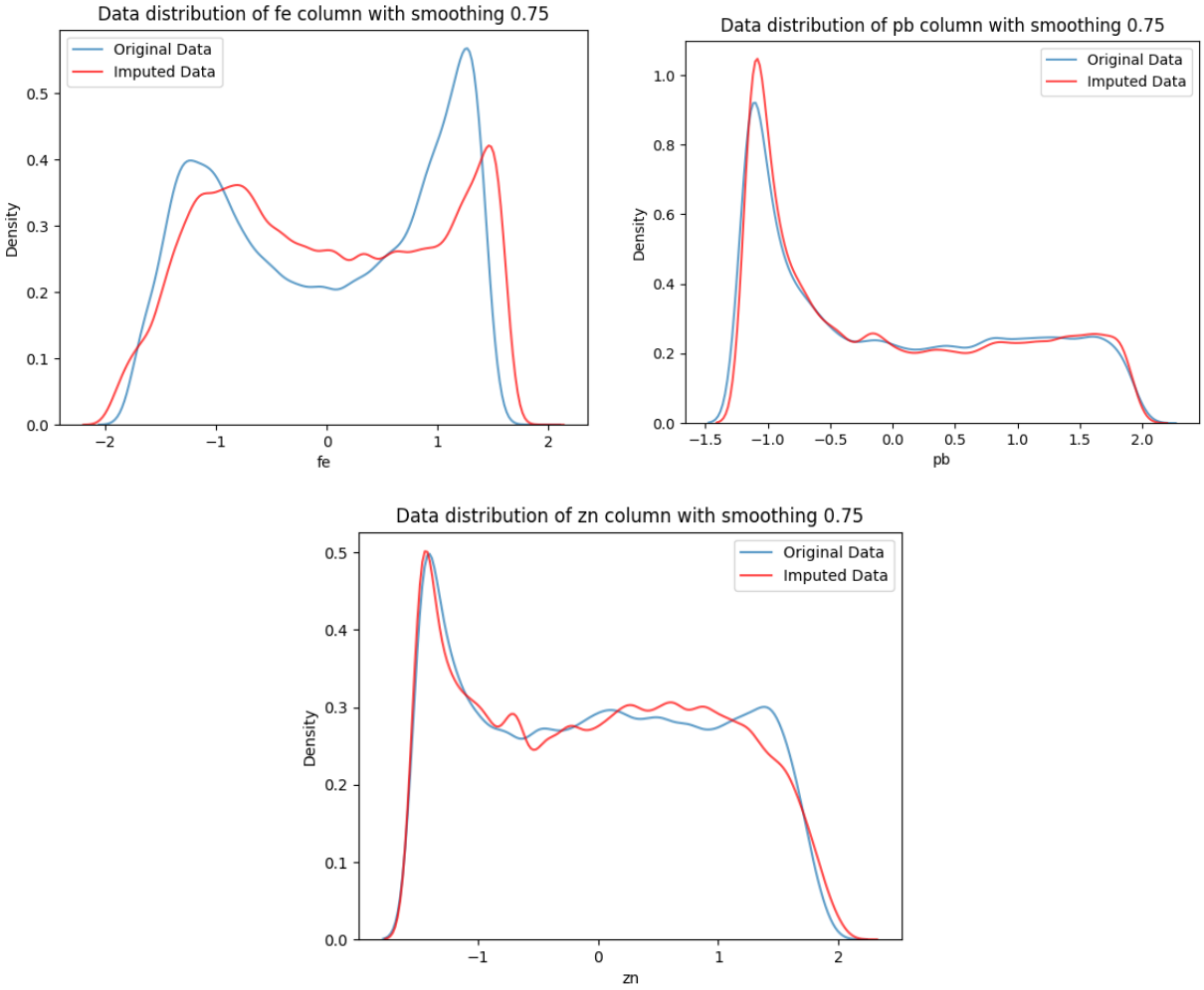


FIG 5 – Showing the data distributions before and after the imputation process by doing the power transformation of the assay values.

CONCLUSION AND FUTURE WORK

The data imputation method discussed is an extremely rapid and data-driven way to produce resource models with uniform and geologically plausible distributions. It requires less human supervision when compared to traditional methods solely relying on domain experts. Incorporating multi-element data allows correlations within these attributes to be analysed and missing domain and numeric data attributes to be imputed with cognisance of the underlying statistical and spatial relationships. Thus, data imputation will be an essential component for improving the outcomes from the implementation of machine learning processes in geological modelling and grade prediction. The documented case history above using real mine data, with all its complexities, shows a significant improvement in the prediction accuracies through the use of data imputation.

Immediate value can be gained in the modelling process by the use of the commercial machine learning engine, Maptek DomainMCF, integrated with the data imputation technique. Because geologists can now use more of the data they have available for rapid resource modelling:

1. Better validation of routine resource modelling can take place.
2. A better and more timely understanding of geological uncertainties can be derived.
3. Many more what-if scenarios can be done in a given amount of time providing a great deal of agility and ability to pivot future decisions.

ACKNOWLEDGEMENTS

The authors acknowledge Maptek for providing the time and approval to prepare and present this paper. They also appreciate the input from Maptek's internal peer review for proofreading and improving the clarity of the paper. Software developers from the DomainMCF team at Maptek are thanked for their ongoing commitment to creating a new environment for geological modelling.

REFERENCES

Arnab, R, 2017. Chapter 1 – Preliminaries and basics of probability sampling, in *Survey Sampling Theory and Applications* (ed: R Arnab), pp 1–21 (Academic Press). https://doi.org/10.1016/B978-0-12-811848-1.00001-7

Collins, L M, Schafer, J L and Kam, C M, 2001. A comparison of inclusive and restrictive strategies in modern missing data procedures, *Psychological Methods*, 6(4):330–351. https://doi.org/10.1037/1082-989X.6.4.330

Department of the Environment, Climate and Communications, 2019. Lisheen Mine Data Release. Available from: <https://www.gov.ie/en/publication/d187c-lisheen-mine-data-release>

Hitzman, M.W. (1999) 'Extensional faults that localize Irish syndiagenetic Zn-Pb deposits and their reactivation during Variscan compression', *Geological Society, London, Special Publications*, 155, pp. 233–245.

Honghai, F, Guoshun, C, Cheng, Y, Bingru, Y and Yumei, C, 2005. A SVM regression-based approach to filling in missing values, in *Knowledge-Based Intelligent Information and Engineering Systems, KES 2005* (eds: R Khosla, R J Howlett and L C Jain), Lecture Notes in Computer Science, 3683:581–587 (Springer: Berlin). https://doi.org/10.1007/11553939_83

Kapageridis, I, Albanopoulos, C, Sullivan, S, Buchanan, G and Gialamas, E, 2021. Application of machine learning to resource modelling of a marble quarry with DomainMCF, *Materials Proceedings*, 5:12. https://doi.org/10.3390/materproc2021005012

King, H F, McMahon, D W and Bujtor, G J, 1982. A guide to the understanding of ore reserve estimation, Supplement to Proceedings No. 281, March 1982 (The Australasian Institute of Mining and Metallurgy: Melbourne).

Krige, D G, 1953. A statistical approach to some basic mine valuation problems on the Witwatersrand, *Operational Research Quarterly*, 4(1):18–18.

Krishnamurthy, V and Hamdi, M, 2013. Mis-information removal in social networks: Constrained estimation on dynamic directed acyclic graphs, *IEEE Journal of Selected Topics in Signal Processing*, 7(2):333–346.

Kyne, R., Torremans, K., Guven, J., Doyle, R. & Walsh, J. (2019) '3-D modeling of the Lisheen and Silvermines deposits, County Tipperary, Ireland: Insights into structural controls on the formation of Irish Zn-Pb deposits', *Economic Geology*, 114(1), pp. 93–116.

Lee, C H and Yoon, H-J, 2017. Medical big data: Promise and challenges, *Kidney Research and Clinical Practice*, 36(1):3–11. Available from: <https://pubmed.ncbi.nlm.nih.gov/28392994>

Matheron, G, 1962. Treatise on Applied Geostatistics, Memoirs of the BRGM, BRGM. Available from: <https://books.google.com.au/books?id=Ej4gzQEACAAJ>

Mislevy, R J, 1991. Review: ‘Statistical Analysis with Missing Data’ by Roderick J A Little and Donald B Rubin, *Journal of Educational Statistics*, 16(2):150–155. http://www.jstor.org/stable/1165119

Montánsa, F, Chinesta, F, Gómez-Bombarelli, R and Kutz, J, 2019. Data-driven modeling and learning in science and engineering. Comptes Rendus Mecanique, vol 347, p 845–855.

Rahm, E and Do, H, 2000. Data cleaning: Problems and current approaches, IEEE Data Engineering Bulletin, 23:3–13.

Solomatine, D, See, L and Abrahart, R, 2009. Data-Driven Modelling: Concepts, Approaches and Experiences. In: R J Abrahart, L M See and D P Solomatine (eds), Practical Hydroinformatics Water Science and Technology Library, vol 68. Springer, Berlin, Heidelberg https://doi.org/10.1007/978–3-540–79881–1_2

Sullivan, S. (2022) 'Harnessing data complexity - how machine learning applies all project data for accurate resource modelling', *AusIMM 12th International Mining Geology Conference Proceedings*, Brisbane, Australia, 22-23 March.

Sullivan, S, Green, C, Carter, D, Sanderson, H and Batchelor, J, 2019. Deep learning – a new paradigm for ore body modelling, AusIMM Mining Geology Conference. Available from: <https://www.maptek.com/pdf/insight/Maptek_Deep_Learning_A_New_Paradigm_for_Orebody_Modelling.pdf>

Torremans, K., Kyne, R., Doyle, R., Güven, J.F. & Walsh, J.J. (2018) 'Controls on metal distributions at the Lisheen and Silvermines deposits: Insights into fluid-flow pathways in Irish-type Zn-Pb deposits', *Economic Geology*, 113(7), pp. 1455–1477.

Troyanskaya, O, Cantor, M, Sherlock, G, Brown, P, Hastie, T, Tibshirani, R, Botstein, D and Altman, R B, 2001. Missing value estimation methods for DNA microarrays, *Bioinformatics*, 17(6):520–525. https://doi.org/10.1093/bioinformatics/17.6.520

Tsai, C-W, Lai, C-F, Chao, H-C and Vasilakos, A V, 2015. Big data analytics: A survey, *Journal of Big Data*, 2(1):21. https://doi.org/10.1186/s40537-015-0030-3

Yeo, I K and Johnson, R A, 2000. A new family of power transformations to improve normality or symmetry, *Biometrika*, 87(4):954–959.