Evaluation of a depth to basement Bayesian model using cross-validation strategies

Hoël Seillé
CSIRO DEI FSP
CSIRO Mineral Resources
hoel.seille@csiro.au

Gerhard Visser
CSIRO Mineral Resources
gerhard.visser@csiro.au

Stephan Thiel
Geological Survey of South Australia
stephan.thiel@sa.gov.au

SUMMARY

Mapping the basement morphology undercover is crucial to identify structures significant for mineral exploration. Using geophysics, this mapping is done indirectly, and the solutions obtained present important uncertainty. Uncertainty can be estimated using Bayesian methods and can be reduced when introducing constraints such as direct drill hole observations of the depth to basement. Cross-validation techniques are commonly used to assess the predictive performance of machine learning or statistical methods. In our case, prediction applies to depth to basement and is observed using drill holes. However, a bias exists in the spatial distribution of the drill holes used for prediction. Certain exploration strategies, often based of geophysical observations, cause a dependency between adjacent observations that is incompatible with the statistical independence required by standard cross-validation techniques. This may cause overestimation of predictive performance. This paper compares different cross-validation strategies to assess their suitability to validating predictive performance of a workflow which maps depth to basement and its uncertainty in the Carrapateena province, South Australia, using Bayesian MT models, and drill holes measurements. Four validation strategies are considered:1) random k-folds cross-validation, 2) spatial block cross-validation, 3) depth block validation and 4) a random k-folds cross-validation with buffer. These results show that the spatial dependencies existing in the drill hole locations used for validation influences the estimation of the performance, and that a classic random k-folds cross-validation overestimates the performance that should be expected in application. We therefore recommend using random k-folds cross-validation with buffer or block validation strategies when using data derived from drill hole measurements for validation of predictive methods.

Key words: Bayesian model, cross-validation, prediction, depth to basement

INTRODUCTION

Most algorithms used in supervised machine learning applications focus on finding a function, or a model, capable of predicting an output given input data. For such applications, it is common to use an optimization algorithm to fit or train such a model on existing input-output pairs until a satisfactory accuracy is obtained. The same kind of optimization is used in geophysical inversion: an Earth model consisting of many model parameters (pixels, layers…) is obtained minimizing the residuals between input data (geophysical observations) and output data (calculated model responses). The end goal of the optimization in geophysical inversion is to obtain a model that can be used to describe and interpret the subsurface. In ML, the model itself has generally little interest, it is rather its capability to predict output data given data that were not used during the optimization that matters. In this paper, the model being used is a 3D depth to basement model, from which predictions of the depth to basement can be made at any given position. This model is derived using direct observations of the depth to basement from drill hole data and indirect observations derived from geophysical models.

It is common to assess the reliability of predictive methods using cross-validation techniques. It consists in training the model using an incomplete input dataset (known as the training set) and testing it on the remaining part of the dataset (known as the validation set) to evaluate its performance. For our application, the cross validation consists in generating a depth to basement model using a subset of the drill hole dataset and validating the model against the remainder of the dataset, predicting the depth of the basement in locations where we do not have drill holes. The input dataset consists of 2 different types of data, geophysically derived depth to basement estimations and drill holes depth to basement observations. Only the direct depth to basement observations in drill hole is considered as the truth, given the high ambiguity of geophysical models. We therefore use only drill hole data for validation.

When splitting the data into training and validation subsets for cross validations, the underlying assumption is that the distribution of validation items is representative of the distribution of items to be used in the application. Internal dependencies in datasets have been observed in many fields, such as ecology (Roberts et al., 2017) or psychology (de Rooij and Weeda, 2020). If the distribution of the locations in the application significantly differs from the distribution of validation item locations, these dependencies can lead to over optimistic estimates of predictive performance. In our

case, we can imagine that predicting the depth of the basement at the location of a drill hole given that a nearby drill hole was used in the training will give very good results, even if this doesn't mean our model will generalize equally well away from existing drill holes. In this study, we identify two types of biases that exists in the dataset and can cause dependency during the cross validation: a spatial bias and a depth bias. In general, the drill holes are concentrated in specific areas, where potential for locating a mineral deposit is higher, causing a spatial bias. For example, in early exploration projects, the drilling campaigns are generally carried out based on the results of regional geophysical results, such as aeromagnetic maps for instance. Magnetic anomalies or "bumps" are preferential targets that have higher chance to detect a mineral deposit. In Figure 1a we observe that many drill holes are located at or around magnetic anomalies. Then, regarding the depth of the drill holes. Because of the cost of exploiting a mineral resource increases with depth, shallower targets are generally preferred. Consequently, deep exploration drill holes are limited compared to the shallower ones. When using these drill holes to map the basement, we therefore have fewer information concerning the location of the basement when located deeper, causing a depth bias. To address these biases, we design cross validation tests splitting the data strategically, to account for the spatial and depth biases.

DEPTH TO BASEMENT BAYESIAN MODEL

The depth to basement Bayesian model we use in this study was previously derived by Seillé et al, (2022) using probabilistic MT models and drill holes. It is located in the Carrapateena Province in South Australia (Figure 1). The workflow consisted in 1) deriving depth to basement probabilities at the location of each MT sounding using Bayesian inversions and 2) interpolate laterally these estimations along with depth to basement drill hole estimations using a Bayesian Fusion algorithm (Visser and Markov, 2019) to produce a depth to basement probabilistic map. The output of this algorithm consists in an ensemble of 2D surfaces that fits the input estimates and approximate a depth to basement probability surface. The mean of the depth to basement probability distribution is shown in Figure 1b. In total 95 drill holes were used to constrain the interpolation.
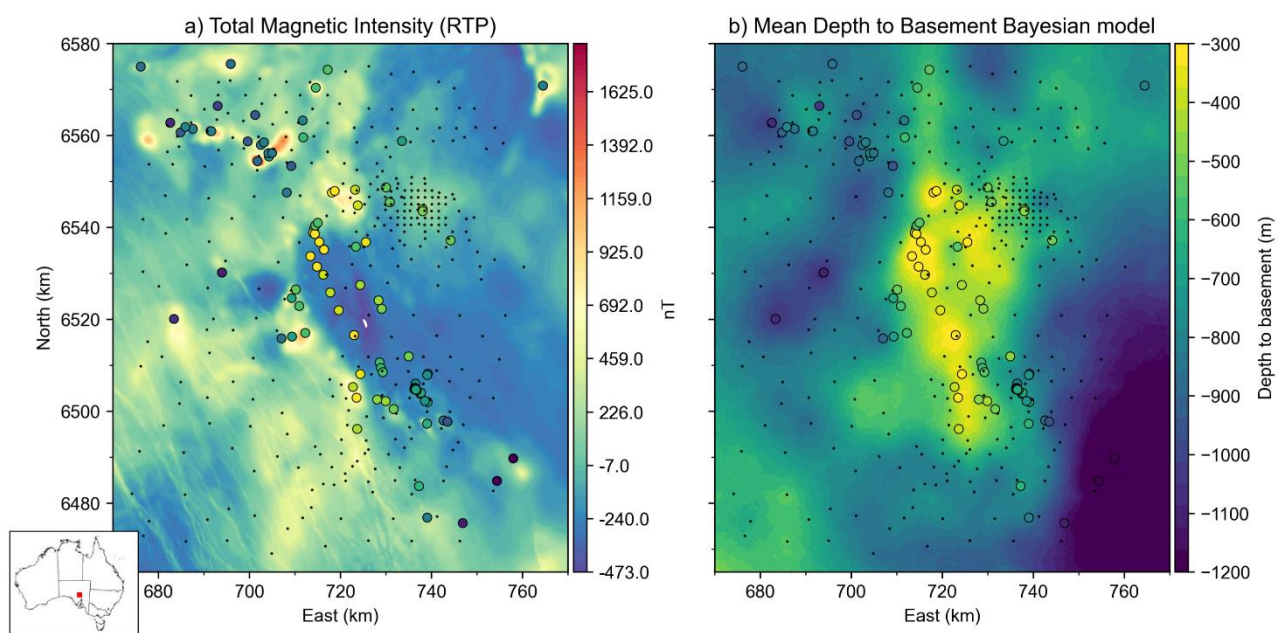

Figure 1. a) Magnetic map of the studied area. b) Mean of the depth to basement Bayesian model (after Seillé et al., 2022). The black dots are the location of the MT sites. The coloured circles are the drill holes, coloured by the depth at which they intersect the basement.

CROSS VALIDATION STRATEGIES

The following validation strategies were tested:

1) Random K-Folds (RKF): The dataset is split into 5 training/validation sets randomly, and each fold is used once for validation while the 4 remaining folds are used for training (Figure 2b). This random selection will frequently cause closely adjacent drillholes to be in training and validation sets.
2) Spatial blocking (SB): The area is divided in 9 blocks of equal sizes. Each block is consecutively used as a validation set, while the remaining 8 blocks are used for training (Figure 2c). This selection may cause uneven sizes of training/validation sets across the 9 folds. SB should give a better indicator of the depth to basement model's application performance in areas where no drill holes are available.

3) Depth blocking (DB): The drill holes are grouped into 5 depth intervals. Each interval is consecutively used as a validation set, while the remaining 4 intervals are used for the training set (Figure 2d and 2e). Because of the relative lateral continuity of the depth to basement, blocking by depth also has a spatial blocking effect: adjacent drill holes usually fall in the same depth interval.

4) Random K-Folds with buffer (RKF_B): Same as RKF, but this time a distance buffer of 3 km is applied to each drill hole of the validation set (Figure 2g). A distance buffer of 3 km is used, meaning that any drill hole located at a distance inferior of 3km from any of the validation drillholes will not be part of the training set, to ensure the spatial independence of the validation drill holes. This approach reduces the amount of data in the training set compared to a normal RKF.
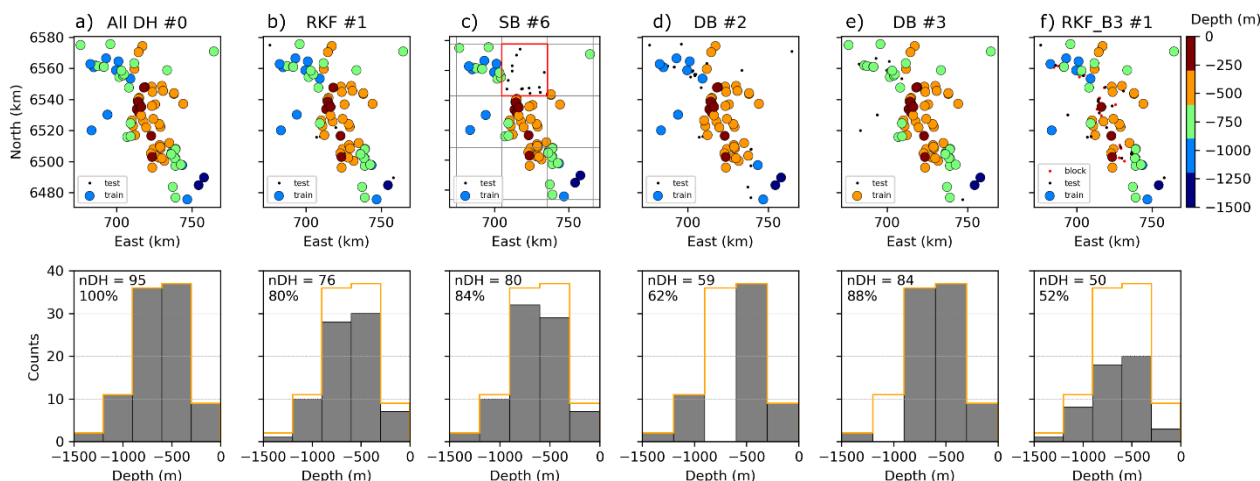
Figure 2. Distribution of the drill holes used in the different cross validation tests. Top row: spatial distribution of the training and validation datasets. Bottom row: Distribution of the training dataset (orange contour is the complete dataset). a) Complete dataset. b) Example of one fold of the Random K-Fold. c) Example of one fold of the Spatial blocking. d) and e) Examples of two folds of the Depth blocking. f) Example of one fold of the Random K-Fold with buffer = 3km. Red points are the drill holes that were excluded from the training set because of their vicinity to validation points (distance < 3km).

RESULTS

We evaluate the results of the prediction of each cross-validation test using a logarithmic scoring rule and a normalised root mean squared error (nRMSE). The logarithmic scoring rule measures how well the posterior probability distribution performs on real data, which in that case will be the unseen validation set. For each test, the scoring rule is defined as the average of the negative logarithm of the probability estimate for the observed depth to basement value at the location of the validation drill holes. Better predictions have lower values. The nRMSE measures the difference between the true drill hole depth to basement and the mean of the posterior distribution, normalized by its standard deviation. In Table 1 is shown the details of the tests. For comparison, we also compute the predictive performance of the test that used all the drill holes in the training set (Control 1 in Table 1) and the one that used none of them, being driven only by the MT models (Control 2 in Table 1).

CV strategy	Number of folds	Average training size (%)	nRMSE	Average -log lkh
Control 1 (MT+DH)	-	100 (Full dataset)	0.581	2.396
Control 2 (MT only)	-	0 (only MT models)	1.699	6.912
Random K-folds	5	80	1.289	3.745
Depth blocking	5	80	1.105	3.756
Spatial blocking	9	89	1.029	3.114
Random K-folds with buffer	5	56	1.583	5.746

Table 1. Number of validations folds, average percentage of data used for training across the folds, nRMSE and average negative log likelihood for each test. Better predictions have lower values.

As expected, the test using all the drill holes (Control 1 in Table 1) has the better predictive performance given that the training set is the same as the validation set. Similarly, the test with no drill holes (Control 2 in Table 1) shows the lowest predictive performance. Then, we observe that RKF, DB and SB have similar scores and that RKF_B has a lower predictive performance. Overall, RKF_B has less training examples that the other validation tests, which could explain

this score. However, because the items that were removed from the training sets in RKF_B are the ones located nearby validation items, they should not contribute to a generalization on new data. The applications in RKF_B therefore properly predicts the depth to basement in unseen areas, which is more realistic. Finally, given the wide variability observed in amount of data used in each fold for the DB and SB tests, a more detailed analysis of these results is required to assess their suitability for specific spatial and depth predictions.

CONCLUSIONS

This study shows that the choice of cross-validation strategy can have an impact on the validation of a method. When making predictions about new locations in application, it is important to consider if they are close to drill holes used to derive the models. If they are, RKF_B will be more appropriate thank RKF. In that case RKF_B can reduce the biases existing between training and validation using blocking strategies, to obtain more reliable prediction errors. Also, when considering predictions in deeper areas, DB might be more appropriate. The suitability of a validation method depends on where the predictions will be made.

ACKNOWLEDGMENTS

REFERENCES

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., ... & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography, 40(8), 913-929.

de Rooij, M., & Weeda, W. (2020). Cross-validation: A method every psychologist should know. Advances in Methods and Practices in Psychological Science, 3(2), 248-263.

Seillé, H., Thiel, S., Brand, K., Mulè, S., Visser, G., Fabris, A., & Munday, T. (2022a). Bayesian fusion of MT, AEM and geological data: Examples from the eastern Gawler Craton, South Australia. Submitted Exploration Geophysics.

Visser, G., & Markov, J. (2019). Cover thickness uncertainty mapping using Bayesian estimate fusion: leveraging domain knowledge. Geophysical Journal International, 219(3), 1474–1490.