A Robust Framework for Comparing Classification Models in Spatial Domains

Scott McManus[1], Azizur Rahman[2], Ana Horta[3] and Jacqueline Coombes[4]

[1] *Charles Sturt University, Port Macquarie NSW 2444, AUS*
[2] *Charles Sturt University, Wagga Wagga NSW 2650, AUS*
[3] *Charles Sturt University, Albury NSW 2640, AUS*
[4] *AMIRA International, Perth WA 6000, AUS*
[1]smcmanus@csu.edu.au

***ABSTRACT* – In the mining industry, a framework exists for quantitative assessment of interpretation uncertainty of spatial domains used to model a stationary spatial domain for mineral resource estimation. This framework will improve public reporting of mineral resource estimates, and improve the reliability of feasibility studies by ensuring successful communication of geological risk. In early-stage mineral projects, there is often not enough multielement laboratory data to enable the use of calculated geological methods for quantitative uncertainty assessment. Portable X-Ray Fluoresce (pXRF) is an accepted method of providing cost and time effective multielement measurements for early-stage projects. However, these measurements are of lower precision and accuracy, then laboratory-based measurements. Recent work has shown that quantitative uncertainty assessments using a Bayesian approximation method can successfully use both pXRF and laboratory data. Subjective visual assessment of uncertainty band graphs, drill hole plots, and confidence matrices suggest that models derived from the two types of data provide similar uncertainty assessments. This paper reviews recent advances in Null Hypothesis and Bayesian Hypothesis statistical methods for comparing models to propose a robust methodological framework for assessing the reliability and similarity of supervised classification models utilising confusion matrix model metrics for further research in the use of pXRF as a suitable measurement for geological spatial domain uncertainty.**

***Keywords*— Accuracy, classification model comparison, statistical significance, Bayesian approximation, pXRF measurements**.

INTRODUCTION

In early-stage minerals projects, there is often not enough data to produce robust variograms to allow geostatistical assessment of spatial uncertainty in spatial domains. Spatial domains are models that either define the three-dimensional geology or mineral estimation envelopes used for estimating volume and the average grade of economic material. As production has not started the industry practice of reconciliation against production actuals to assess model quality is also not possible. A workflow to assess the interpretation uncertainty of classified drill hole intercepts assigned to spatial domains as an alternative measure of quality in early-stage mineral projects has been developed [1].

Often the data used in the early stage to classify the drill hole samples to enable the construction of the spatial domain models is limited. Data that is available includes the laboratory analysed economic element, subjective visual logging by a geologist (expert), surface geological mapping, and geophysical surveys. A conceptual exploration model guides the collection of this data. So the exploration and mining company is often not willing to commit to the extra cost of laboratory multielement analysis until the conceptual model is confirmed. It is often the case that exploration and mining companies do not budget for more laboratory-based analysis than the economic element of interest.

Furthermore, one or two other elements that can be used as a pathfinder to show the sample is in or out of the mineralising system may also be analysed. Later in the project, when the project is economically viable, laboratory multielement geochemistry is analysed. Usually, as part of a metallurgical study and to identify toxic minerals.

The limited types of data in the early-stage can cause issues in regards to the spatial domain model creation and the resultant mineral resource estimate, which relies on geostatistical assumptions. Relying only on economic grade element based boundaries for the mineralisation spatial domains can be problematic. Boundaries based on current metal economics will affect geostatistical studies as they will not define stationary statistical principles ensuring similar geological, chemical and structural features within the model [2]. Additionally, geological logging may be inconsistent or of unknown quality due to industry employment demands, high turnover of staff, or work by successive minerals companies [3]. Insufficient subjective geological logging has led to the use of calculated geology to support and strengthen geological models.

Calculated geology has been successful using multielement data from Inductively Coupled Plasma (ICP) and portable X-Ray Fluorescence (pXRF) measurements. However, the pXRF measurements have lower confidence and have issues with heterogeneity, accuracy, and precision despite following best practice workflows [4]. Reference [5] showed that interpretation uncertainty assessment using calculated geology from pXRF multielement data was able to determine the quality of spatial domains created using only subjective geological data and economical grade element laboratory results.

Recently a company has sought to assess the interpretation uncertainty of spatial domains as they have four competing models. The first model used arbitrary high-grade modelling of the economic element. The second model used the geochemical alteration halo of the ground rock, which was generated by hot fluids that provided a chemical 'trap' for the emplacement of the economic material. The third model uses a pathfinder element's dispersal through the alteration halo. The final model combined economic element presence, structure, alteration, weathering and lithological data. The company has taken 1,045 samples from pulverised reject powders from previous laboratory

samples in a test area that includes six classes of the new spatial domains and collected ICP and pXRF measurements on the powders, which should reduce much of the heterogeneity issues of the pXRF method.

Separate Bayesian approximation models have been developed to assess the spatial domains' interpretation uncertainty using just the pXRF and just the ICP multielement data. As there are significant savings in time and cost of using pXRF measurements compared to ICP laboratory data, the company is interested in determining if the uncertainty assessment of the two models are similar, so that they can continue to use the cheaper pXRF method. Initial results suggest both the pXRF and ICP models are providing similar measures of uncertainty. Figures 1 and 2 show the graphed Bayesian uncertainty bands for the alteration spatial domain. A subjective visual comparison shows similar peaks and troughs in the prediction value for each sample point. Similar high uncertainty in the prediction at sample 72 is visible in both figures. Metrics from confusion matrixes also suggest the models are similar, with overall accuracy for both models being 0.94.

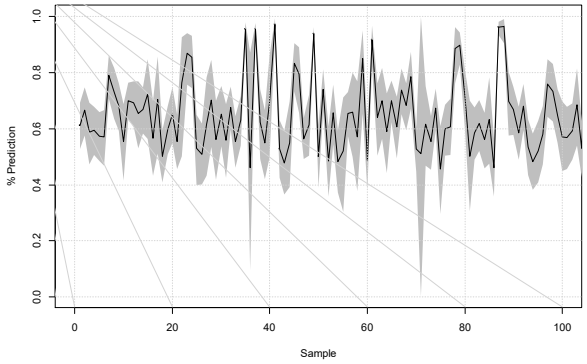

Fig. 1 Point prediction and uncertainty bands for alteration spatial domain, using ICP measurements for the first 100 samples.

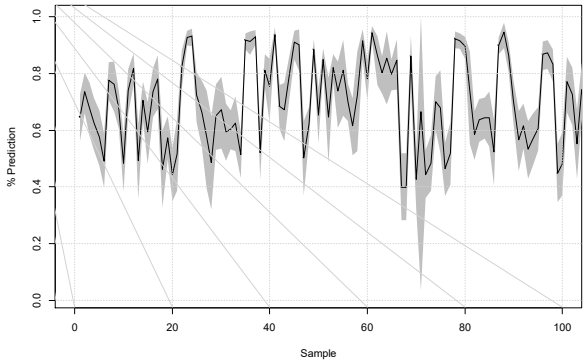

Fig. 2 Point prediction and uncertainty bands for alteration spatial domain, using pXRF measurements for the first 100 samples.

The main aim of the paper is to demonstrate a contemporary methodological framework for assessing the reliability of different supervised classification models. Following is a selective review of statistical significance tests in use in the spatial science and data science fields that might be appropriate for determining if the null hypothesis, 'That both models are equal', is true.

STATISTICAL MODEL COMPARISON METHODOLOGY

It is current practise to assess and compare spatial supervised classification models with the following criteria; these include accuracy, reproducibility, robustness, and suitability of the algorithms. The error matrix or confusion matrix [6] with resultant Overall Accuracy (OA), and Kappa (K) provides a base for determining these criteria. In its simplest form, comparison of models is achieved through a subjective comparison of a confusion matrix from each model presented as a table of results as well as a set of standard performance metrics computed from the confusion matrix.

A review of research papers that compared models found that the majority only provided a confusion matrix and resultant model performance metrics. The metrics used included, Overall Accuracy, Area under the receiver operating characteristic curve (AUC), sensitivity, specificity, F-score and Kappa. The highest metric determines the preferred modelling method. This methodology does not answer the question if two models provide different performance on the same data set using a test of statistical significance. The practice of just presenting the confusion matrix and model performance does not provide any evidence in regards to actual differences between model outcomes. However, it is useful in displaying the visual metrics of each model [7]. This simplistic method is also limited in that it does not provide practical information around predictive power, accuracy as well as ability and utility, especially when the data size is large for predictive modelling.

Modifications have been applied to the metrics to account for the limitations of subjective evaluation. These include; modification of the confusion matrix, the use of different Kappa formulas to account for discrepancies in the Cohen kappa formulation, and a normalised confusion matrix. These modifications assume the result is a 'hard' classification, one where a location or sample may only belong to one category [8]. The raw or modified metrics are then often used in a formal statistical test of significance using the t-test [9]. The use of the t-test in assessing Bayesian predictive modelling is also successfully demonstrated.[10]. However, a straightforward implementation of the frequentist t-test may be inappropriate as there are several issues in regards to assumptions.

The issues that affect model comparison when using a t-test include; an inflated Type I error rate, a low statistical power, or low reproducibility. We are not able to test multiple hypotheses, for example, when looking at the current problem, we can ask, does ICP perform similar to pXRF and if not, which one performs better? The t-test assumes a specific distribution, whereas our models may not conform to that assumption. There is no consideration for the data uncertainty. Finally, in regards to p-values, the p-value does not distinguish between the effect and sample size. The correlated t-test alleviates some of these issues when comparing confusion matrix metrics of different classifiers with the same data set [11] or the signed-rank test [12] when comparing different classifiers over multiple data sets.

To reduce the issues highlighted sampling from the distribution or resampling in some format from the confusion matrix is required. One such method is the Resampling method (without replacement) combined with the Student t-test. This method still produces a high Type I error rate as well as failing to meet one of the critical assumptions of the test, that of normality. This failure is due to the difference between the metrics of the two models not conforming to a normal

distribution. Both models use identical training and test data sets, which are highly correlated and during the resampling process, the resample sets may overlap. A similar method to Resampling is the bootstrap method but includes replacement. This method also produces a high Type I error, there is less overlap than in the previous method, but the models are still strongly correlated due to the identical training and test data sets. A third method, the K-Folds test, attempts to reduce the correlation but because there is an overlap between k-2 folds, also produces a high Type I error.

In order to reduce the inflated Type I error with the resampling methods [13] proposed to use the corrected resampled t-test using an expression in equation (1) which adjusted the estimated variance ($\hat{\sigma}^2$) of the t value to account for the inflated Type I Error:

$$\hat{\sigma}^2 = \frac{1}{J} + \frac{n_2}{n_1} S_{\mu_j}^2 \qquad (1)$$

where, $S_{\mu_j}^2$ is the sample variance of the estimates, n_1 is the number of samples from the training set, n_2 is the number of samples in the validation set and J is the number of samples, which is the number of k-folds x the number of repeats.

The corrected cross-validation test applies the same variance correction for the t-test. The corrected repeated k-fold Cross Validation test removes not only the Type I errors, through the adjusted t-value variance, but also improves the power and higher replicability issues. The method does this by repeating a 10 K Fold test 10 times, so there is a total of 100 samples. As there are 10 K Fold data sets (including training and test data sets), it also reduces the correlation seen in the previously discussed resampling methods.

The next two methods discussed, resample from the simulation of the metrics of the confusion matrix rather than actual data. Reference [14] proposed a Bayesian interpretation of the confusion matrix to allow assessment of uncertainty in the performance indicators of a classifier when the researcher only has the confusion matrix utilising the values of the matrix as a random vector following a multinomial distribution. The Bayesian method also allowed for the update of the model with prior information. The method draws on similar techniques used in the previously discussed method as well as in the Bayesian Correlated t-test discussed later. An alternative method that also worked just with the confusion matrix values utilises the discrete squared Hellinger distance and assumes a multinomial distribution. The difference between this method and the previous one is that this method uses individual cell values of the confusion matrix. This method considers the correct number of correctly classified and misclassified samples—the previous method averages and aggregates the individual matrix values to arrive at the Overall Accuracy and Kappa [15].

There is also an application of Corrected K-Fold cross-validation test through a Bayesian approximation methodology through two additional tests. The Bayesian hierarchical method for testing different models over multiple data sets [16] and the Bayesian correlated t-test for testing different models on the same data set [17]. Both tests use a data set of repeated k-fold cross-validation results. Which utilises the most robust resampling method, to improve the Type I error, power, and replicability issues. Moreover, when the prediction is our concern, the effects of prior distributions are significant for different models through [18], and the effect of prior distributions can be investigated by iteratively changing the prior.

Previously discussed tests used an averaging model which provides just the single value to assess in a t-test. Both of these Bayesian methods provide several outputs that are useful in examining the relationship between the two models. Firstly, a resultant posterior probability of the classifier confusion matrix metrics is available that can be used with the corrected variance t-test using the frequentist null hypothesis methodology. Secondly, the intersection of the posterior distribution function within a defined Region of Practical Equivalence (ROPE) shows the intersection of the models. Figure 3 shows the relationship of the ROPE with the posterior density function of the difference between the two models. In this case, it is the ICP and pXRF model for a single spatial domain using a single data set. With a ROPE of 1%, that is, if there is less than 1% difference in the Overall Accuracy, then both models are practically equivalent. The probability mass that is within the bounds of the ROPE is the probability that two models are practically equivalent. The areas outside the bounds of the ROPE provide the probability that one model is more accurate than the other model. So, for a ROPE of 1% the probability that ICP is better than pXRF is 23.3%, the probability that both models are practically equivalent is 75.9% and the probability that pXRF is better than ICP is 0.8%

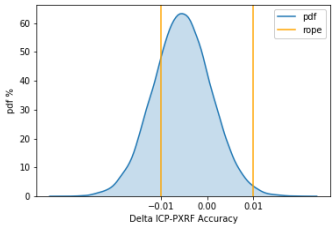

Fig. 3 Example of defined ROPE at 1% and posterior density of the difference between two models (ICP and pXRF) for the Accuracy metric.

DISCUSSION

In the context of the comparison of the ICP and pXRF models, the set of tests that assume one set of data is appropriate. That is the Correlated Bayesian t-test. The ICP and pXRF models use different variables, but the variables are from the same data set and the same sample points. So this will exclude the signed-rank test and the Bayesian hierarchical test, which test the same models over different data sets.

pXRF measurements are not possible for lighter elements, which may mean that geological features, such as alteration may not be as well defined compared to ICP measurements. An additional limitation may be that each metric might result in a different outcome for each of the five spatial domains—especially where the features of the spatial domains are sensitive to light elements. The pXRF model may be affected by low specificity in addition to the precision and accuracy of the measurements. It may be practicable to consider a series of ROPE from 1% to 5%. Iteratively looking for the percentage of ROPE where the most significant mass of probability exists will allow consideration of practical equivalence in models, where the pXRF model is not as accurate but still showing a slightly weaker equivalence. In the context of the study area, we may expect the set of spatial domains based on the alteration to have a weaker equivalence then say the spatial domain based

on the pathfinder element as the pathfinder element is a heavy element.

Even though the ROPE used for different metrics and different spatial domains, may differ, it is clear what level of practical equivalence and the mass of probability within that region of practical equivalence. The method reduces high Type-I errors, improving power and high replicability issues. Also, assessment of assumptions of the distribution and sensitivity of the prior is possible using the Bayesian method.

CONCLUSION

The use of a traditional frequentist Null Hypothesis t-test to compare model metrics has issues. Also, the subjective evaluation of the metrics and presentation of the confusion matrix for each model does not answer statistical significance with rigour. In reviewing current practise, resampling methodology and application of Bayesian techniques, it is clear that a method of resampling using k-fold cross-validation, combined with a t-test corrected for variance will remove these issues. The addition of Bayesian approximation allows the assessment of prior sensitivity, the use of different distributions and a statement of practical equivalence.

The following recommendation for the hypothesis, 'that pXRF and ICP models are practically equivalent for assessing interpretation uncertainty' is the following methodology framework;

- Presentation of the metrics and confusion matrix for each model for subjective evaluation,
- Correlated Bayesian t-test analysis of repeated cross-validation results (10x10) of the differences between the following model metrics; Overall Accuracy, AUC, Sensitivity, Kappa, precision, recall, and root mean square error,
- Analysis of prior sensitivities, and
- Selection of an appropriate ROPE for each spatial domain.

This methodology framework will reduce the Type I error, improve power and replicability as well as verify distribution assumptions.

ACKNOWLEDGEMENT

The authors would like to acknowledge the support provided by the Australian Government Research Training Program Scholarship, Gympie Gold Mines, Australian Bauxite Ltd, Qld DNRME's Exploration Data Centre, Zillmere, MMG Ltd, CSU Spatial Analysis Unit, the CSU Indigenous Support Centre, Port Macquarie, CSU School of Computing and Mathematics and the CSU School of Environmental Science. We would also like to thanks two reviewers for their valuable comments and stimulus.

REFERENCES

[1] S. McManus, J. Coombes, A. Horta, and A. Rahman, “A workflow for assessing interpretation uncertainty in spatial domains using bayesian approximation,” In International Future Mining Conference 2019: Incorporating the 11th Symposium on Green Mining;, 2019.

[2] [C. Stegman, “How domain envelopes impact on the resource estimate–case studies from the cobar gold field, nsw, australia,” Mineral resource and ore reserve estimation—the Aus IMM guide to good practice. The Australasian Institute of Mining and Metallurgy, Melbourne, pp. 221-236, 2001.

[3] F. Fouedjio, E. J. Hill, and C. Laukamp, “Geostatistical clustering as an aid for ore body domaining: Case study at the Rocklea dome channel iron ore deposit, western Australia,” Applied Earth Science, vol. 127, no. 1, pp. 15-29, 2018/01/02, 2018.

[4] M. F. Gazley, L. C. Bonnett, L. A. Fisher, W. Salama, and J. H. Price, “A workflow for exploration sampling in regolith-dominated terranes using portable x-ray fluorescence: Comparison with laboratory data and a case study,” Australian Journal of Earth Sciences, vol. 64, no. 7, pp. 903-917, 2017.

[5] S. McManus, A. Rahman, A. Horta, and J. Coombes, "Applied Bayesian modeling for assessment of interpretation uncertainty in spatial domains," Statistics for Data Science and Policy Analysis. pp. 3-13.

[6] A. Rahman, S. Nimmy, and G. Sarowar, "Developing an automated machine learning approach to test discontinuity in DNA for detecting tuberculosis." pp. 277-286.

[7] A. Rahman, “Statistics-based data preprocessing methods and machine learning algorithms for big data analysis,” International Journal of Artificial Intelligence, vol. 17, no. 2, pp. 44, 2019.

[8] D. Lu, and Q. Weng, “A survey of image classification methods and techniques for improving classification performance,” International Journal of Remote Sensing, vol. 28, no. 5, pp. 823-870, 2007/03/01, 2007.

[9] A. Rahman, Statistics for data science and policy analysis: Springer, 2020.

[10] [10] A. Rahman, Bayesian predictive inference for some linear models under student-t errors: VDM Publishing, 2008.

[11] C. Nadeau, and Y. Bengio, “Inference for the generalization error,” Machine Learning, vol. 52, no. 3, pp. 239-281, 2003/09/01, 2003.

[12] J. Demšar, “Statistical comparisons of classifiers over multiple data sets,” Journal of Machine learning research, vol. 7, no. Jan, pp. 1-30, 2006.

[13] [13] J. Gardner, and C. Brooks, “A statistical framework for predictive model evaluation in moocs,” in Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale, pp. 269-272, 2017.

[14] [14] O. Caelen, “A bayesian interpretation of the confusion matrix,” Annals of mathematics and artificial intelligence, vol. 81, no. 3-4, pp. 429-450, 2017.

[15] J. García-Balboa, M. Alba-Fernández, F. Ariza-López, and J. Rodríguez-Avi, “Analysis of thematic similarity using confusion matrices,” ISPRS international journal of geo-information, vol. 7, no. 6, pp. 233, 2018.

[16] G. Corani, A. Benavoli, J. Demšar, F. Mangili, and M. Zaffalon, “Statistical comparison of classifiers through bayesian hierarchical modelling,” Machine Learning, vol. 106, no. 11, pp. 1817-1837, 2017/11/01, 2017.

[17] A. Benavoli, G. Corani, J. Demšar, and M. Zaffalon, “Time for a change: A tutorial for comparing multiple classifiers through bayesian analysis,” The Journal of Machine Learning Research, vol. 18, no. 1, pp. 2653-2688, 2017.

[18] A. Rahman, J. Gao, C. D'Este, and S. E. Ahmed, “An assessment of the effects of prior distributions on the bayesian predictive inference,” International Journal of Statistics and Probability, vol. 5, no. 5, pp. 31, 2016.