



# Geographic quantile regression forest: a new method for spatial modelling of mineral commodities

**Kane Maxwell**

Matrix Geoscience  
Springfield Lakes, QLD  
k.maxwell@matrixgeoscience.com

**Mojtaba Rajabi**

University of Queensland  
St Lucia, QLD  
m.rajabi@hotmail.com

**Joan Esterle**

University of Queensland  
St Lucia, QLD  
j.esterle@uq.edu.au

## SUMMARY

Spatial modelling of analysis results such as grade, geochemical properties and density is required for resource estimation in almost all commodities. For spatial modelling of most commodity properties, geostatistical methods are the most popular because they are usually more accurate than deterministic methods, can quantify uncertainty and can use auxiliary information to improve predictive accuracy. However, geostatistical methods have the primary disadvantages that they have onerous pre-processing steps such as variogram modelling, rely on rigid statistical assumptions, and incorporation of numerous auxiliary variables which have non-linear relationship with the target variable is difficult. To address these disadvantages, a machine learning method based on quantile regression forest algorithm is proposed as an alternative approach for spatial modelling. This newly proposed method (termed geographic quantile regression forest) does not require variogram modelling, can quantify uncertainty and can easily incorporate numerous auxiliary information of differing data type. To evaluate the performance of the new method, the accuracy of predictions of coal relative density is compared to inverse distance weighting and regression kriging. Data from an active mine site in the Bowen Basin, Queensland Australia is used for the comparison. Using evaluation metrics from leave-one-out cross-validation, this paper demonstrates that the geographic quantile regression forest method has higher accuracy than inverse distance weighting and similar or higher accuracy than regression kriging accuracy across all geological domains. The high accuracy, similar performance to regression kriging and stated advantages over geostatistical methods makes it a candidate for future inclusion in geological model packages.

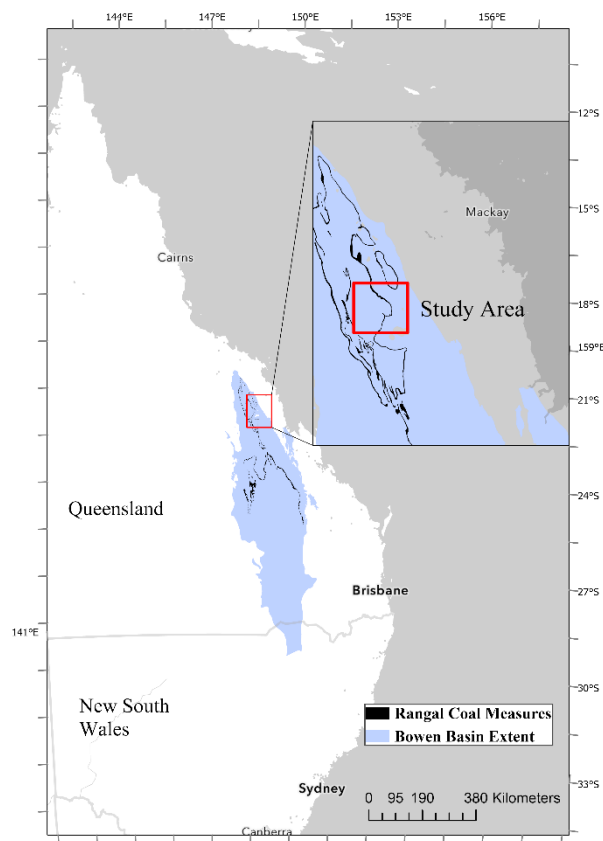
**Key words:** Spatial modelling, quantile regression forest, geostatistics, machine learning.

## INTRODUCTION

Spatial modelling is the process of predicting values at unknown locations based on existing observations (such as geochemical results obtained from drill core lab analysis). Spatial modelling is required for resource estimation in all commodities because it is impractical and costly to obtain information at high density. Numerous methods for spatial modelling exist, however geostatistical methods are most utilised for modelling mineral grade (Goovaerts and others 1997; Olea 2013; Srivastava 2013). For modelling of coal properties, geostatistical methods such as regression, universal and co-kriging have also shown to produce superior results to

deterministic methods, especially when auxiliary geophysical log data is available (Jeuken, Xu, and Dowd 2020; Webber, Costa, and Salvadoretti 2013). However, when there are numerous properties to model, and numerous geological domains are present, geostatistical methods may be considered onerous because variogram interpretation is required for all properties and domains. For example, in coal resource estimation, numerous properties including coal ash, volatile matter, moisture, density, yield, energy content, and coke properties are required to be modelled. In addition, in most coal deposits, numerous coal plies exist which must be modelled separately (domained). Due to this, inverse distance weighting (IDW), which does not require variogram modelling, remains the most popular method for spatial modelling of coal properties for resource estimation (Maxwell 2020). However, IDW has the primary disadvantages that it cannot quantify uncertainty and cannot use additional auxiliary data to improve estimates (Jeuken et al. 2020; Srivastava 2013). Therefore, ideally, a spatial model method that has the advantage of geostatistical methods, and does not require onerous pre-processing steps is preferred when modelling of numerous properties and domains is required.

Random forest, which is a machine learning method, has recently shown to produce similar or superior results to geostatistical methods when extended to account for spatial data (Georganos et al. 2019; Hengl et al. 2018). These extended 'spatial' random forest methods have numerous advantages over geostatistical methods including that variogram modelling and rigid statistical assumptions about the data are not required, and they can easily incorporate categorical variables (Hengl et al. 2018). In the context of this paper, the ability to incorporate categorical variables allows for a single modelling process for all coal seams and domains compared to IDW and geostatistical methods which require that each domain and coal seam is modelled separately. However, a primary disadvantage of traditional random forest is that it does not provide an estimate of error. To address this, the new method in this paper utilises quantile regression forest in place of random forest. This modification enables the calculation of uncertainty at any confidence interval (Meinshausen 2006). The new method is termed geographic quantile regression forest (QRF). To evaluate the performance of QRF, data from an active mine site in the Bowen Basin Queensland Australia is used (Figure 1.). The performance of the new method is evaluated against IDW and regression kriging (RK) which is a popular geostatistical method which can incorporate auxiliary data and is considered state of the art (Hengl, Heuvelink, and Rossiter 2007; Keskin and Grunwald 2018). To demonstrate the flexibility of QRF, results are provided for numerous geological domains and coal seams which are identified at the mine site. Relative density (Standards Australia International 2002), which is an important parameter for coal resource estimation is used as the variable to be spatially modelled.

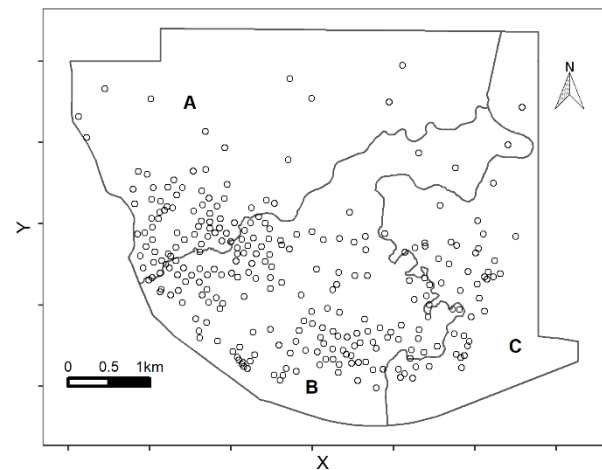


**Figure 1.** Approximate location of the study area, relative to the Bowen Basin (after (Sliwa et al. 2017)).

## METHOD AND RESULTS

### Database

Data provided by the mine site comprises of 272 boreholes with relative density and geophysical log parameters (density and natural gamma). The target coal seam at the mine site is the Leichardt coal seam which is part of the Upper Permian Rangal Coal Measures of the Bowen Basin (Sliwa et al. 2017). Locally, the coal seam at the mine is comprised of sub-plies named LCTU, LCL1, LCL2, LCL3 and LCL4 which are differentiated by their variable coal quality. Three geological domains are identified by the mine site, termed Domain A, B and C. Domain A is characterized by absence of LTCU, and Domain C is defined by the presence of Cretaceous age lamprophyres and dolerite intrusion (Ritchie 2010). Coal impacted by the intrusion in this domain has marked increase in density, marked decrease in volatile matter and slight increase in ash (Maxwell, Rajabi, and Esterle 2019). The spatial arrangement of the data in each domain is presented in Figure 2 and is further described in Table 1.



**Figure 2.** Locations of boreholes used in this study with respect to each geological domain.

Domain	No. Boreholes	Distance (m)		
		Min	Max	Mean
A	73	47	583	154
B	153	25	705	127
C	46	2	1001	159

**Table 1.** Number of boreholes used in this study, and statistics on the distance (in meters) between boreholes for each geological domain.

### Method

All data preparation, spatial modelling, evaluation, and plot generation was conducted in the R programming language (R Core Team 2017). The R library gstat (Pebesma 2014) was used to create the estimates for IDW and RK. The QRF method was also implemented in R and was later developed as an R package (Maxwell, Rajabi, and Esterle 2021). The QRF implementation relies on the R packages ranger (Wright and Ziegler 2017) and nabor (Elseberg et al. 2012). The settings for each spatial model method are supplied in Table 2 and were determined by random grid search optimisation (Bergstra and Bengio 2012).

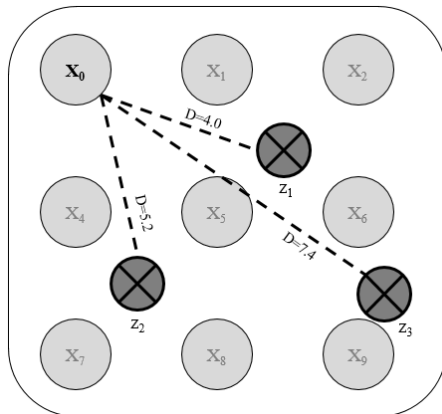
For background theory for IDW, the reader is referred to (Shepard 1968). For regression kriging, primary references are (Hengl et al. 2007; Keskin and Grunwald 2018). The QRF method comprises of the following steps, explained with an accompanying diagram:

For  $X_0$ :

1. Calculate the Euclidean distance between  $X_0$  and  $k$  (number of observed data). In Figure 3,  $k=3$  and the nearest observations are  $Z_1$ ,  $Z_2$  and  $Z_3$  with corresponding distances of 4, 5.2 and 7.4 respectively.
2. Scale the distances calculated in the previous step between 0 and 1. These will be used as weighting values in the following step. In this example, re-scaled distances would be 1, 0.64 and 0.0 respectively.
3. Train a quantile regression forest using the points from Step 1 and use the weights from Step 2 as

weighting values in the algorithm (e.g. equivalent to the case weights in R package ranger).

4. Use the trained quantile regression forest to predict the value at ( $x_0$ ).
5. Repeat this process for all locations e.g.  $X_{0-9}$



**Figure 3. Diagram supporting the step by step explanation of the QRF method where  $X_{0-9}$  are locations at which values are to be predicted and  $Z_{1-3}$  are observed (known values).**

Model Method	Settings	Auxiliary data
IDW	Power =2, k = 10	-
RK	Variogram= auto, k=10, repressor = random forest, mtry= 2, trees =500	Short spaced density, long spaced density, compensated density, natural gamma
QRF	k=50, mtry=2, trees= 500	Easting, northing, ply code, domain, short spaced density, long spaced density, compensated density, natural gamma

**Table 2. Spatial model parameter settings for the models used in this study.**

### Evaluation

All three methods are evaluated using “variance explained by predictive models based on cross-validation”, VEcv, (Li 2017). In this case leave-one-out cross validation is used (Efron 1982). In the leave-one-out case, VEcv is equivalent to calculating the Coefficient of Determination (R-squared) on the residuals. VEcv is used because VEcv is unit, scale, data mean and variance independent (Li 2017). Results of VEcv can be interpreted using Table 4.

VEcv	Interpretation
$\leq 10$	Very poor
$> 10 \leq 30$	Poor
$> 30 \leq 50$	Average
$> 50 \leq 80$	Above average
$> 80$	80 Excellent

**Table 3. Interpretation of model predictive performance based on VEcv result (Li 2016).**

### Results

VEcv results show that IDW is the worst performing method across all domains and plies with the exception of LCL2 and LCL3 in domain ‘A’ (Table 4). Across all domains, QRF produced better VEcv than RFK in Domain A, slightly worse results in domain B, and near identical results in domain C (Table 4).

In domain A, the standard deviation of density across all plies was much lower compared to domain B and C (Table 4). This indicates in areas of low variability IDW may perform similarly to RK and QRF. However, all models produced very poor to poor results (interpreting using Table 3) in Domain A. In domain B, across all plies, RK and QRF produced slightly above average models compared to IDW which produced poor results. Highest VEcv by all models was achieved in Domain C with RK and QRF producing near excellent results compared to IDW which produced average results. However, lack of data points (14-35) in this domain may cause overestimation in result when comparing directly against domain A and B.

Domain	Ply Code	N.	Sd	VEcv result for each model		
				QRF	IDW	RK
A	LCL1	50	0.05	23	19.93	21.63
	LCL2	55	0.05	9.81	12.3	9.18
	LCL3	53	0.05	10.1	11.56	17.71
	LCL4	58	0.08	29.49	-22.55	-17.95
	Total	216		24.12	-1.03	1.99
B	LCTU	123	0.07	23.58	10.24	39.9
	LCL1	122	0.04	18.18	7.85	32.52
	LCL2	109	0.06	32.39	-19.21	29.91
	LCL3	104	0.08	75.4	-6.49	68.89
	LCL4	128	0.14	37	12.84	45.34
	Total	586		52.1	24.98	57.22
C	LCTU	14	0.18	51	1.99	67.09
	LCL1	23	0.23	88.06	12.92	83.59
	LCL2	12	0.15	43.5	21.87	38.99
	LCL3	17	0.18	83.24	17.61	78.72
	LCL4	35	0.2	69.53	24.62	72.83
	Total	101		78.22	32.39	78.74

**Table 5. Relative density variance explained based on cross-validation (VEcv) for each model method broken down by geological domain and coal seam. Results can be interpreted using Table 4. The number of observations (N) and standard deviation (Sd) of relative density for each ply is also shown.**

### CONCLUSIONS

This paper demonstrated and evaluated a machine learning based method for spatial modelling. The method is based on the quantile regression forest (QRF) algorithm which was modified to account for spatial data by utilising distances between observations as case weights. Using coal relative density as the variable to be spatially modelled, the accuracy of the method

was compared to inverse distance weighting (IDW) and regression kriging (RK) across numerous coal plies and geological domains. Using evaluation based on leave-one-out cross validation, the results showed QRF and RK markedly outperformed IDW across all geological domains. Results also showed that QRF outperformed RK in a geological domain with comparatively low variance in ply relative density, and performed very similar in the other geological domains.

Major advantages of the QRF method over geostatistical methods are that it does not require variogram modelling and that a single modelling process can be used across numerous geological domains. These advantages greatly reduce the complexity of the modelling process when there are numerous geological domains and numerous variables to be spatially modelled.

Due to the similarity in performance to RK, and stated advantages over geostatistical methods, the QRF method should be considered as an alternative spatial model method for modelling of mineral commodity properties.

## ACKNOWLEDGMENTS

The authors acknowledge Peabody Energy for the supply of data for this paper.

## REFERENCES

- Bergstra, James, and Yoshua Bengio. 2012. "Random Search for Hyper-Parameter Optimization." *Journal of Machine Learning Research* 13(Feb):281–305.
- Efron, Bradley. 1982. *The Jackknife, the Bootstrap and Other Resampling Plans*. SIAM.
- Elseberg, J., S. Magnenat, R. Siegwart, and A. Nüchter. 2012. "Comparison of Nearest-Neighbor-Search Strategies and Implementations for Efficient Shape Registration." *Journal of Software Engineering for Robotics (JOSE)* 3(1):2–12.
- Georganos, Stefanos, Tais Grippa, Assane Niang Gadiaga, Catherine Linard, Moritz Lennert, Sabine Vanhuysse, Nicholas Mboga, Eléonore Wolff, and Stamatis Kalogirou. 2019. "Geographical Random Forests: A Spatial Extension of the Random Forest Algorithm to Address Spatial Heterogeneity in Remote Sensing and Population Modelling." *Geocarto International* 1(12). doi: 10.1080/10106049.2019.1595177.
- Goovaerts, Pierre, and others. 1997. *Geostatistics for Natural Resources Evaluation*. Oxford University Press on Demand.
- Hengl, Tomislav, Gerard B. M. Heuvelink, and David G. Rossiter. 2007. "About Regression-Kriging: From Equations to Case Studies." *Computers & Geosciences* 33(10):1301–15. doi: <https://doi.org/10.1016/j.cageo.2007.05.001>.
- Hengl, Tomislav, Madlene Nussbaum, Marvin N. Wright, Gerard B. M. Heuvelink, and Benedikt Gräler. 2018. "Random Forest as a Generic Framework for Predictive Modeling of Spatial and Spatio-Temporal Variables." *PeerJ* 2018(8). doi: 10.7717/peerj.5518.
- Jeuken, Rick, Chaoshui Xu, and Peter Dowd. 2020. "Improving Coal Quality Estimations with Geostatistics and Geophysical Logs." *Natural Resources Research* 1–18. doi: 10.1007/s11053-019-09609-y.
- Keskin, H., and S. Grunwald. 2018. "Regression Kriging as a Workhorse in the Digital Soil Mapper's Toolbox." *Geoderma* 326:22–41. doi: 10.1016/J.GEODERMA.2018.04.004.
- Li, Jin. 2016. "Assessing Spatial Predictive Models in the Environmental Sciences: Accuracy Measures, Data Variation and Variance Explained." *Environmental Modelling & Software* 80:1–8. doi: <https://doi.org/10.1016/j.envsoft.2016.02.004>.
- Li, Jin. 2017. "Assessing the Accuracy of Predictive Models for Numerical Data: Not r nor R2, Why Not? Then What?" *PloS One* 12(8):e0183250.
- Maxwell, Kane. 2020. "Review of the Spatial Interpolation Techniques Used for Estimating Economic Coal Properties." *AIG Journal* (1443–1017).
- Maxwell, Kane, Mojtaba Rajabi, and Joan Esterle. 2021. "Spatial Interpolation of Coal Geochemical Properties Using Geographic Quantile Regression Forest." *International Journal of Coal Geology* In Review.
- Maxwell, Kane, Mojtaba Rajabi, and Joan Esterle. 2019. "Automated Classification of Metamorphosed Coal from Geophysical Log Data Using Supervised Machine Learning Techniques." *International Journal of Coal Geology* 214:103284. doi: 10.1016/j.coal.2019.103284.
- Meinshausen, Nicolai. 2006. "Quantile Regression Forests." *J. Mach. Learn. Res.* 7:983–99.
- Olea, Ricardo A. 2013. "Special Issue on Geostatistical and Spatiotemporal Modeling of Coal Resources." *International Journal of Coal Geology* 112:1. doi: <https://doi.org/10.1016/j.coal.2013.01.010>.
- Pebesma, Edzer. 2014. "The Meuse Data Set : A Brief Tutorial for the Gstat R Package." 17.
- R Core Team. 2017. "R: A Language and Environment for Statistical Computing."
- Ritchie, Casandra. 2010. "Lamprophyric Intrusions in the Rangal Coal Measures, Bowen Basin Classification, Geochemistry and Tectonic Significance." University of Queensland.
- Shepard, Donald. 1968. "A Two-Dimensional Interpolation Function for Irregularly-Spaced Data." Pp. 517–24 in *Proceedings of the 1968 23rd ACM national conference*.
- Sliwa, Renate, Joan Esterle, Laura Phillips, and Steven Wilson. 2017. "Rangal Supermodel 2015: The Rangal-Baralaba-Bandanna Coal Measures in the Bowen and Galilee Basins. Final Report ACARP Project C22028."
- Srivastava, R. Mohan. 2013. "Geostatistics: A Toolkit for Data Analysis, Spatial Prediction and Risk Management in the Coal Industry." *International Journal of Coal Geology* 112:2–13. doi: 10.1016/j.coal.2013.01.011.
- Standards Australia International. 2002. "AS1038.21.1.1–2002, Coal and Coke – Analysis and Testing Part 21.1.1: Higher Rank Coal and Coke – Relative Density – Analysis Sample/Density Bottle Method."
- Webber, Tiago, João Felipe Coimbra Leite Costa, and Paulo Salvadoretti. 2013. "Using Borehole Geophysical Data as Soft Information in Indicator Kriging for Coal Quality Estimation." *International Journal of Coal Geology* 112:67–75. doi: 10.1016/j.coal.2012.11.005.
- Wright, Marvin N., and Andreas Ziegler. 2017. "Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R." *Journal of Statistical Software* 77(1):1–17. doi: 10.18637/jss.v077.i01.