A machine learning approach to geochemical mapping

Charlie Kirkwood [a,*], Mark Cave [a], David Beamish [a],
Stephen Grebby [a], Antonio Ferreira [a]

[a] *British Geological Survey, Environmental Science Centre, Keyworth, Nottingham, NG12 5GG, UK*

* Corresponding author. Tel.: +44 1159363344
Email address: cwk@bgs.ac.uk (C.W.Kirkwood)

Abstract

Geochemical maps provide invaluable evidence to guide decisions on issues of mineral exploration, agriculture, and environmental health. However, the high cost of chemical analysis means that the ground sampling density will always be limited. Traditionally, geochemical maps have been produced through the interpolation of measured element concentrations between sample sites using models based on the spatial autocorrelation of data (e.g semivariogram models for ordinary kriging). In their simplest form such models fail to consider potentially useful auxiliary information about the region and the accuracy of the maps may suffer as a result. In contrast, this study uses quantile regression forests (an elaboration of random forest) to investigate the potential of high resolution auxiliary information alone to support the generation of accurate and interpretable geochemical maps. This paper presents a summary of the performance of quantile regression forests in predicting element concentrations, loss on ignition and pH in the soils of south west England using high resolution remote sensing and geophysical survey data.

Through stratified 10-fold cross validation we find the accuracy of quantile regression forests in predicting soil geochemistry in south west England to be a general improvement over that offered by ordinary kriging. Concentrations of immobile elements whose distributions are most tightly controlled by bedrock lithology are predicted with the greatest accuracy (e.g. Al with a cross-validated R^2 of 0.79), while concentrations of more mobile elements prove harder to predict. In addition to providing a high level of prediction accuracy, models built on high resolution auxiliary variables allow for informative, process based, interpretations to be made. In conclusion, this study has highlighted the ability to map and understand the surface environment with greater accuracy and detail than previously possible by combining information from multiple datasets. As the quality and coverage of remote sensing and geophysical surveys continue to improve, machine learning methods will provide a means to interpret the otherwise-uninterpretable.

Keywords:
Uncertainty
Modelling
Soil geochemistry
Quantile regression
Random forest
South west England

1. Introduction

The value of geochemical maps to mineral exploration (e.g. Hawkes and Webb, 1962; Levinson, 1974; Beus and Grigorian, 1977; Xuejing and Xueqiu, 1991; Xu and Cheng, 2001; Johnson et al., 2005), agriculture (e.g. Webb et al., 1971; Jordan et al., 1975; Reid and Horvath, 1980; Lewis et al., 1986; White and Zasoski, 1999; Reimann et al., 2003), and studies of environmental and human health (e.g. Thornton and Plant, 1980; Bowie and Thornton, 1985; Alloway, 1990; Appleton and Ridgway, 1993; Thornton, 1993; Fordyce, 2013) is well established. Surficial geochemistry should be considered an essential component of any comprehensive description of the natural environment (Darnley, 1990). In these times of increasing environmental concern, there is a need for increasingly effective geochemical mapping techniques to support the making of good evidence-based decisions about our interactions with the natural environment.

Geochemical maps are produced by the regional interpolation of element concentration data obtained from samples of surface media such as stream sediments, soil or water (e.g. Salminen et al., 1998). The sampling density is often limited by the relatively high cost of sample collection and chemical analysis, resulting in large expanses between sample sites in which there is much uncertainty about concentrations of elements. Traditionally, the interpolation of element concentrations has been based on the spatial autocorrelation of the data, as in ordinary kriging

(Cressie, 1988) which uses semivariogram models. While these spatial models are considered optimal for univariate interpolation in regions where no other information is present, their ignorance of auxiliary information makes them suboptimal for use in regions for which auxiliary variables have been measured. For geochemical mapping auxiliary variables might include anything that provides insight into surface-subsurface conditions, for example airborne gamma spectrometry and magnetic survey data.

Spatial autocorrelation based models such as ordinary kriging can be adapted to make use of auxiliary information, either by combination with regression models, as in regression-kriging or kriging with external drift approaches (e.g. Hengl et al., 2003), or by co-kriging (e.g. Knotters et al., 1995). However, the importance of considering spatial autocorrelation in predictive models decreases as the explanatory power of the auxiliary variables increases: eventually the spatial autocorrelation of the target variable is entirely captured within the auxiliary variables. Models which do not rely on spatial autocorrelation information are desirable as they greatly improve the interpretability of the resultant maps. The predicted element concentrations are no longer the product of a crude distance-weighted blend of geographically neighbouring measurements, but instead can be explained by the context of the prediction point within the more informative, process related, feature space of the auxiliary variables. The residuals of such models are useful as they indicate the degree to which samples have been subject to atypical processes.

Thanks in part to the Tellus South West airborne geophysical survey (Beamish et al., 2014), south west England is now one of the most thoroughly surveyed areas of Great Britain, and possesses a wealth of quantitative high resolution geoscientific data. It is therefore an ideal study area in which to investigate the ability of the available high resolution data to explain the variations of measured element concentrations in soils. There are many possible regression techniques with which to model soil element concentrations from auxiliary geoscientific data, however, to account for the lack of independence and normality in both predictor and target variables, nonparametric ‘machine learning’ techniques are advantageous. Interpretability is also a priority; in order to have impact, the resultant models and maps must be explainable to policy makers. Random forest (Breiman, 2001) is a machine learning technique which has been demonstrated to be highly accurate, adaptable and interpretable. The technique uses an ensemble of decision trees, and is capable of both classification and regression. It is gaining popularity for use in predictive mapping in various fields; for example species distribution mapping (e.g. Lawrence et al., 2006; Cutler et al., 2007; Evans et al., 2011), land-cover classification (e.g. Gislason et al., 2006; Rodriguez-Galiano et al., 2012), geological mapping (Cracknell and Reading, 2014) , digital soil mapping (e.g. Henderson et al., 2005; Wiesmeier et al., 2011) and mineral prospectivity mapping (e.g. Carranza and Laborte, 2015; Harris et al., 2015; Rodriguez-Galiano et al., 2015).

In this study quantile regression forests (Meinshausen, 2006) – an uncertainty-conscious elaboration of random forest (Breiman, 2001) – are utilised to model the concentrations of elements in the soils of south west England using high resolution geophysical and remote sensed data. The ability of quantile regression forests to use these auxiliary variables to produce high resolution, interpretable geochemical maps with quantified prediction intervals is demonstrated. This approach has important implications for future geochemical survey planning procedure. Additionally, interrogation of the underlying models facilitates improved understanding of the geochemical environment of south west England and has implications for decisions about our interaction with the natural environment.

2. Materials

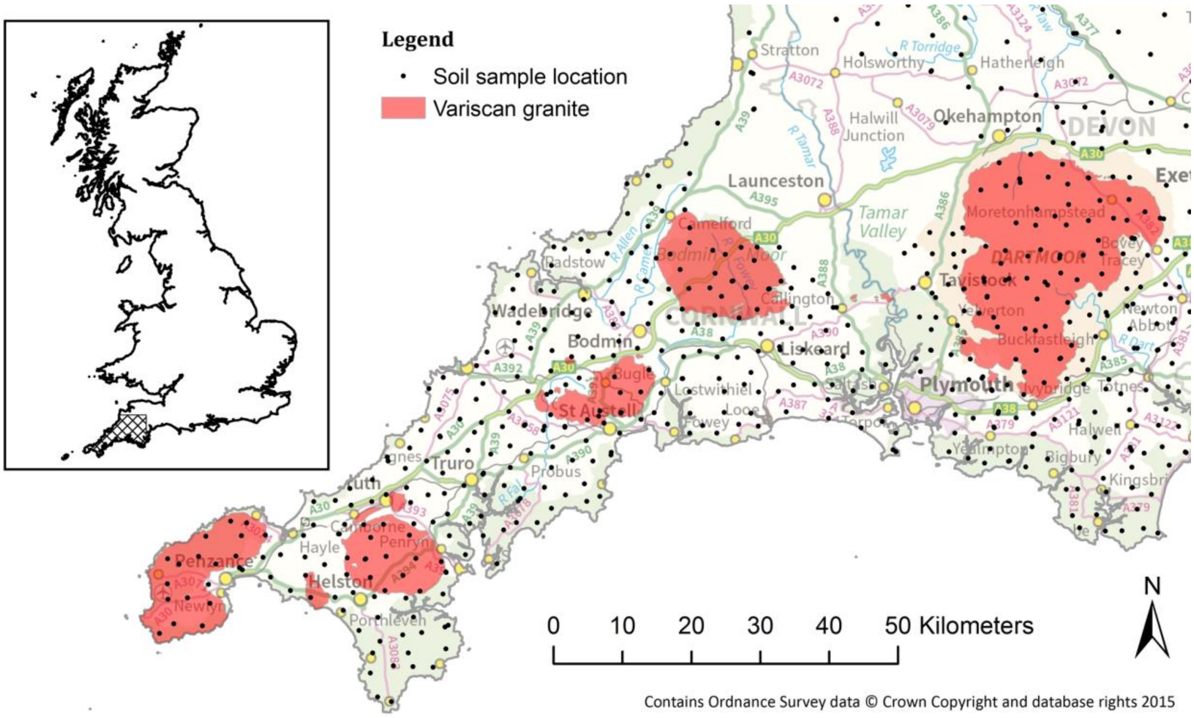
2.1 Study area

The study area, south west England, is located at the southwestern tip of the British Isles (Fig. 1). A wealth of high resolution geoscientific data has been collected across south west England owing to complex and economically significant geology. In brief summary, the geology of the region consists of a suite of metasedimentary facies originally deposited in a series of Devonian-Carboniferous east-west trending basins (Shail and Leveridge, 2009). The granites of the Cornubian Batholith were then emplaced following basin inversion during the late Carboniferous to early Permian Variscan Orogeny (Charoy, 1986; Floyd et al., 1993), and have provided a heat source for extensive hydrothermal activity. The result of this hydrothermal activity is that the region is both rich in polymetallic mineralisation (Dines, 1956; Willis-Richards and Jackson, 1989) and complex in terms of mapping and understanding element distributions (e.g. Colbourn et al., 1975; Alderton et al., 1980; Smedley, 1991; Kirkwood et al., 2016).

2.2 Target variables - soil geochemical data

The soil geochemical data used in this study is derived from samples collected across south west England during the summer field campaign of 2012 by the British Geological Survey following standard Geochemical Baseline Survey of the Environment (G-BASE) methods (Johnson et al., 2005). A total of 568 samples were collected within the study area at an average sampling density of one sample per 12.2 km^2 (Fig 1). Samples were collected at random, but exclude coverage of the Tamar Valley area which was sampled in 2004. The Tamar Valley data is not used in this study due to inferior lower limits of detection as a result of advancements in analytical procedure between the years of 2004 and 2012. The soil samples were collected from a depth of 5-20cm and sieved to <2mm grain size before being dried, ground and pelletised prior to analysis by XRF for 48 major and trace elements according to standard G-BASE procedures (Johnson et al., 2005). The 5-20cm sampling depth is intended to target the A horizon of typical soils, with material from the O horizon being excluded with the topmost 5cm. However, soil horizon representation within each sample varies according to local soil profiles. The pH and loss on ignition (LOI) of each sample was also measured. Data quality was assured by the inclusion of duplicate samples, replicate samples, and certified reference materials within the analytical runs.

Total concentrations of the following elements were determined along with pH and LOI: Ag, Al, As, Ba, Bi, Br, Ca, Cd, Ce, Co, Cr, Cs, Cu, Fe, Ga, Ge, Hf, I, K, La, Mg, Mn, Mo, Na, Nb, Nd, Ni, P, Pb, Rb, Sb, Sc, Se, Si, Sm, Sn, Sr, Ta, Te, Th, Ti, Tl, U, V, W, Y, Zn and Zr. The major elements (Al, Ca, Fe, K, Mg, Mn, Na, P, Si, Ti, Zr) were assumed to exist as their common oxides, and were each appended with the appropriate additional mass of oxygen so that the sum of all element concentrations for each sample approached 100%, or in the units of the study, 1 million milligrams per kilogram. For most samples though, the chemical analyses do not sum to 100%. This 'remainder' (referred to as 'R') is included in the study, to see if it too could be modelled and explained.



Contains Ordnance Survey data © Crown Copyright and database rights 2015

Fig. 1. Locations of 2012 field season G-BASE soil samples within the study area in south west England. The inset map shows the study area (cross-hatched) in reference to the rest of Great Britain. The granites of the Cornubian Batholith are shown as they form prominent geological and geochemical landmarks within the region.

2.3 Auxiliary variables – high resolution geophysics and remote sensed data

In order to provide the quantile regression forest models with as much information as possible from which to make predictions, all available regional geophysics and remote sensed data sets were utilised. The available data sets comprise airborne magnetic and radiometric surveys from the Tellus South West project (Beamish et al., 2014), aerial elevation survey from NEXTMap (Intermap Technologies, 2007), land gravity survey from the British Geological Survey et al. (1968), and Landsat 8 satellite imagery (Roy et al., 2014). All these auxiliary variables and their derivatives (Table 1) were resampled from their original data grids to a regular 100 m grid covering the study area using bilinear interpolation.

The 61,000 line-km of airborne geophysical data collected for the Tellus South West project, and the processing undertaken to produce the original magnetics and radiometrics data grids, is described by Beamish and White (2014). The survey used a N-S line separation of 200 m and a magnetic data sampling of 20 Hz providing a mean along-line sampling of 3.6 m. Radiometric data were sampled at 1 Hz intervals providing a sampling of 71 m. Data grids were generated using bicubic spline interpolation (magnetic) and minimum curvature (radiometric). The land gravity survey data were gridded using minimum curvature.

Table 1
Explanations of the geophysical and remote sensed variables used in the modelling.

Variable name	Explanation
Elevation	NEXTMap Britain Digital Terrain Model
Slope	Terrain slope angle
Wetness_index	Terrain wetness index
Topographic_position_index	Terrain topographic position index
Plan_curvature	Terrain plan curvature
Profile_curvature	Terrain profile curvature
Landsat_B1	Landsat 8 band 1 – Coastal Aerosol (0.43-0.45 μm)
Landsat_B2	Landsat 8 band 2 – Blue (0.45-0.51 μm)
Landsat_B3	Landsat 8 band 3 – Green (0.53-0.59 μm)
Landsat_B4	Landsat 8 band 4 – Red (0.64-0.67 μm)
Landsat_B5	Landsat 8 band 5 – Near Infrared (0.85-0.88 μm)
Landsat_B6	Landsat 8 band 6 – Short Wave Infrared 1 (1.57-1.65 μm)
Landsat_B7	Landsat 8 band 7 – Short Wave Infrared 2 (2.11-2.29 μm)
Landsat_B8	Landsat 8 band 8 – Panchromatic (0.50-0.68 μm)
Landsat_B10	Landsat 8 band 10 – Thermal Infrared 1 (10.60-11.19 μm)
Landsat_B11	Landsat 8 band 11 – Thermal Infrared 2 (11.50-12.51 μm)
Regional_bouguer_anomaly	Gravity survey bouguer anomaly

Residual_bouguer_anomaly	Gravity survey high pass filtered bouguer anomaly
TMI_IGRF	International Geomagnetic Reference Field corrected TMI
TMI_IGRF_1VD	1st vertical derivative of TMI_IGRF
TMI_IGRF_AS	Analytical signal of TMI_IGRF
TMI_IGRF_REDP	Reduction to the pole of TMI
Radiometrics_uranium	Uranium counts from gamma ray spectrometry
Radiometrics_thorium	Thorium counts from gamma ray spectrometry
Radiometrics_potassium	Potassium counts from gamma ray spectrometry
Radiometrics_total_count	Total count of unmixed gamma ray signal

3. Methods

3.1 Quantile regression forests

Quantile regression forests (Meinshausen, 2006) are an elaboration of random forest (Breiman, 2001); an ensemble model based on the averaged outputs of multiple decision trees (Breiman et al., 1984). Where random forest takes the mean of the outputs of the ensemble of decision trees as the final prediction, quantile regression forests also take specified quantiles from the outputs of the ensemble of decision trees, providing a quantification of the uncertainty associated with each prediction.

The decision trees themselves are constructed through recursive partitioning starting with a root node which contains all the data provided to the tree. The root node is split by defining an optimal threshold in whichever auxiliary variable works best to provide two resulting data partitions each with the greatest purity (the least variation in the target variable). This process is then repeated successively on child partitions until the terminal nodes ('leaves') are reached, at which point each partition contains just a single sample (or specified small number of samples) whose target variable value (or mean value) is explained by a series of increasingly precise "if-then" conditional statements referring to the context of the sample in the auxiliary variable feature space.

If all of the decision trees were grown from the same training data there would be no point in using an ensemble – the trees would all grow identically and the resultant model would be highly liable to overfit the data. Breiman's (2001) random forest overcomes the problem of overfitting decision trees by using bootstrap aggregation, or bagging (Breiman, 1996), to grow each tree from a separate subsample (roughly two thirds) of the full training dataset, thus reducing the chance of fitting to noise when the outputs of the multiple trees are averaged. In addition to bagging, random forest also provides only a random subset of the auxiliary variables on which to make each split in each tree, which reduces the chance of the same very strong predictors being chosen at every split, and therefore prevents trees from becoming overly correlated. The resulting algorithm is recognised as a highly competitive machine learning technique (e.g. Liu et al., 2013; Rodriguez-Galiano et al., 2015).

One drawback of the random forest method is that, as a consequence of each prediction being equivalent to a weighted average of the target variable values in the training data set (Lin and Jeon, 2006), predictions towards the limits of the training data values are increasingly biased towards the mean. This results in a tendency for low value predictions to exhibit positive bias, and high value predictions to exhibit negative bias (Zhang and Lu, 2012). To correct for this all random forest models were appended with a linear transformation defined by a robust linear model (iterative reweighted least squares; Venables and Ripley, 2013) of observations against random forest predictions during their training phase. This process effectively stretches the predictive range of the random forest in order to correct for central tendency bias.

All modelling was conducted in R (R Core Team, 2014) with a framework developed around the randomForest package (Liaw and Wiener, 2002). The models each used 1001 decision trees - a sufficient number to allow convergence of error to a stable minimum. The odd number of trees prevents possible ties in variable importance. Each tree was grown until the terminal nodes contained 8 samples in order to reduce overfitting to outliers. The default number of variables to try at each split – one third of the number of features – was used. The mean of the outputs of the ensemble of decision trees was used as the predicted value, and for each prediction the 2.5th and 97.5th percentiles of the ensemble were used as the lower and upper limits of a 95% prediction interval.

3.2 Model validation

The training dataset was constructed by joining the auxiliary variable data at each soil sample site to the geochemical data for each soil sample, using bilinear interpolation, in order to form a single table of both geochemical and auxiliary variable values for each sample site. A stratified 10-fold cross validation process was then used, in which the training data was randomly split into 10 equal folds of approximately equal mean (Kohavi, 1995). Then, for each element, a quantile regression forest model was constructed using the data in 9 of the folds before being tested by predicting the measured element concentrations in the remaining fold. The folds were cycled through and the modelling process repeated so that, in the course of the full 10-fold cross validation, every sample was used as test data. This process allows the accuracy of the model's predictions and prediction intervals (uncertainty estimates) to be assessed for each element, which is visualised in this study using scatter plots of the predicted against observed values. The prediction interval accuracies are assessed for each model on the basis of how closely the percentage of samples that are observed to fall within the prediction interval match the expected percentage (according to the specified prediction interval). In the case of this study we use a 95% prediction interval and therefore expect that 95% of samples will fall within it during cross-validation.

To allow the quality of each element's model to be compared, cross-validated R^2 values, root-mean-square error (RMSE) and range-normalised RMSE values were derived according to the relationship between each model's predictions and the actual measurements. In addition, Moran's I (Moran, 1950) was also calculated on each element's residuals to provide a measure of residual spatial autocorrelation. The Moran's I scale runs from -1 (perfect dispersion) to 1 (perfect correlation), with values close to zero indicating spatially random phenomena and suggesting that model performance would not be increased by directly taking spatial autocorrelation into account.

In order to provide some context to the prediction accuracy of the quantile regression forest models, ordinary kriging (using the R package 'automap'; Hiemstra et al., 2009) was run in parallel to the quantile regression forest modelling during the 10-fold cross validation, from which cross-validated R^2 values were derived.

3.3 Regional geochemical map production

The geochemical maps for each element were produced using a quantile regression forest model constructed on the full 568 sample training dataset. For each element, both concentration and uncertainty maps were produced. The value assigned to each grid cell in the concentration map is a prediction based on the measured values of the auxiliary variables. The value assigned to each grid cell in the uncertainty map is the width of the 95% prediction interval associated with each concentration prediction. No further measurements of soil geochemistry are used to test the map, but the results of the 10-fold cross validation form an acceptable approximation of the performance of each element's model (and therefore the quality of each element's map)(Kohavi, 1995; Vanwinckelen and Blockeel, 2012). For further assessment of model quality, the residuals of the quantile regression forests were mapped using inverse distance weighted interpolation. This allows for any spatial patterns within the residuals to be assessed (a more involved alternative to the Moran's I metric). Concentration maps were also produced by ordinary kriging to allow visual comparison with the quantile regression forest maps. However, caution is advised against making critical comparisons between methods based on the appearance of the maps alone – the image format encourages far more subjective (and potentially misleading) interpretations than objective model quality measures such as cross-validated R^2. All maps were symbolised using a CubeHelix continuous colour scale to prevent loss of information when viewing in greyscale (Green, 2011).

3.4 Model interpretation

With the help of the R package forestFloor (Welling, 2015) partial dependence scatter plots were produced to visualise the contribution of a given variable to the predicted element concentration (Palczewska et al., 2013). Additionally, each quantile regression forest model provides a measure of the average ability of each auxiliary variable to increase node purity in child partitions; thus providing a measure of the importance of each auxiliary variable to the predictions of each element.

The combination of these outputs provides insight into the controls behind each element's distribution.

4. Results and discussion

4.1 Model performance

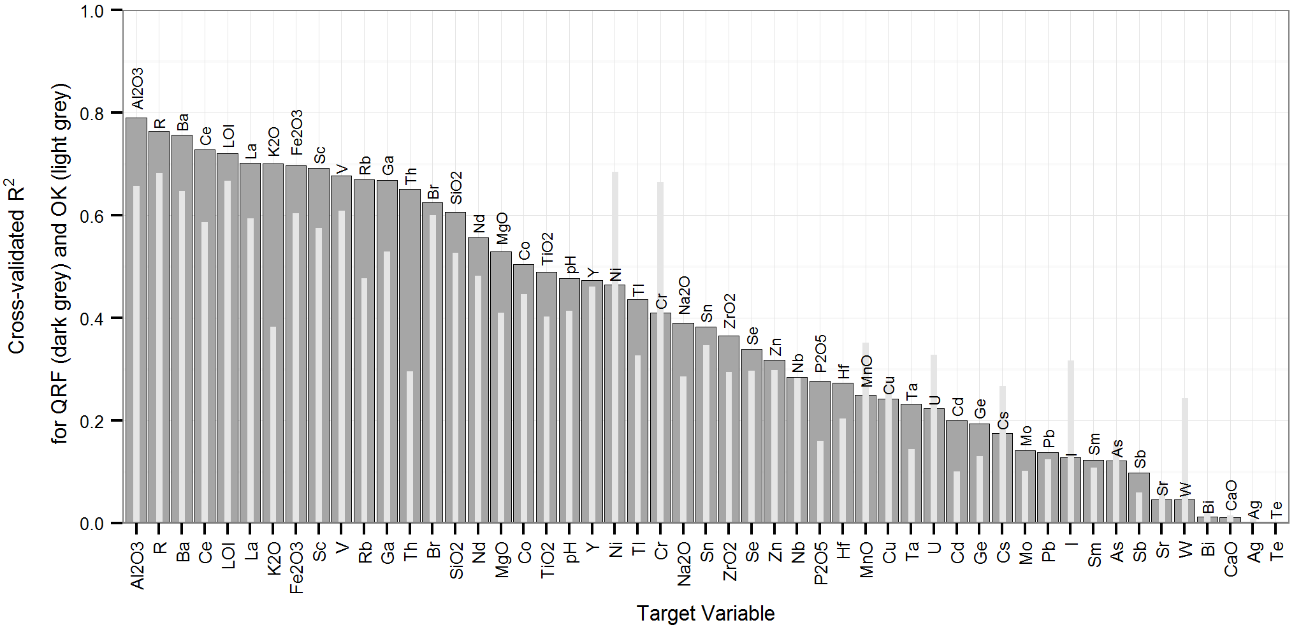


Fig. 2. Cross-validated R^2 values for comparison of quantile regression forest (QRF) model quality between each element (and R, LOI and pH). The corresponding cross-validated R^2 values achieved by ordinary kriging (OK) are overlain to provide some context to the overall quality of predictions.

Comparison of cross-validated R^2 values between quantile regression forests and ordinary kriging reveals that quantile regression forests provide overall improved prediction accuracy for 37 of the 51 target variables modelled (Fig. 2). Aside from Ni and Cr, which are unique in the strength of their association with the Lizard Ophiolite Complex (the region's southernmost pensinsula; Kirby, 1979; Kirkwood et al., 2016), the majority of the 14 elements for which ordinary kriging provided better predictions were minor or trace elements, and poorly predicted by either method. This is an encouraging result for the validity of geochemical maps produced by quantile regression forests using this data in south west England.

Cross-validated R^2 values for the quantile regression forest models vary greatly across the range of elements from 0.79 (Al) to 0 (Te). There appears to be a general inverse relationship between prediction accuracy and element mobility: elements which are known to be relatively immobile (and thus reflect the underlying lithology), such as Al, La and Ce are predicted with little error, while hydrothermally mobile elements such as W,Bi,Te,Ag and As are predicted with higher error. This discrepancy suggests a relative lack of explanation of hydrothermal processes within the suite of auxiliary variables. However, the Moran's I values for the residuals of all quantile regression forest models (Table 2) only deviate from zero by 0.011 in the worst case (Ge). This suggests that the auxiliary variables used have successfully captured the spatial dependence of all target variables at the scale of the predictor grid. Any residual variation in element concentrations which has not been captured by the models can therefore be attributed to processes which essentially appear to be spatially random at the scale of the geochemical survey, but which additional high resolution auxiliary variables may be capable of explaining. This is supported by inspection of variograms of the residuals of each element (not shown), which appeared to exhibit pure nugget effect.

The limited ability of the auxiliary variables used here to explain the distributions of the more mobile elements could perhaps be improved by the inclusion of additional variables which provide more information on spatial context. For example, a measure such as 'distance to nearest fault' could provide valuable context in relation to fluid flow pathways. However, a strength of the modelling approach in its current state is the consistency, transparency, and fully quantitative nature of the auxiliary variable datasets; each collected by sensing equipment, thus avoiding the potential inconsistencies of observations made by multiple geologists in the field. Currently any 'distance to nearest fault' or similar variables would need to be derived from traditional geological maps and consistency would suffer. However, with sufficient spatial resolution there is no reason why structural features such as faults would not be recognisable within the data. To make the best use of such structural information it would become beneficial to use an approach which is capable of learning higher order context (learning textures and spatial patterns, rather than just point properties), perhaps based on artificial neural networks. Such models could potentially learn processes of soil erosion and accumulation (and hydrothermal mobilisation) from spatial context without explicitly being provided with contextual derivatives as input variables. However, such deep learning would increase the effective degrees of freedom within each model, and would require more training data (perhaps more than would ever be financially viable) in order to produce reliable results. The combination of quantile regression forests and the auxiliary variables used in this study therefore represent a promising first step forward given the currently available data and the requirement for transparent and interpretable models.

Plots of predicted concentrations against measured concentrations from the 10-fold cross validation of the quantile regression forests allow for more detailed visualisation of model quality. The examples of La and Sn (Fig. 3), chosen as they provide insight into the models of both immobile (La) and mobile (Sn) elements, show how the prediction interval (2.5th to 97.5th forest quantiles) is unique for each prediction. The cross validation has shown these prediction intervals to be a remarkably accurate (if slightly conservative) probabilistic estimate for all elements (see Table. 2). This is very useful; even for elements with relatively low prediction accuracies the prediction intervals still provide reasonable upper and lower limits on predictions, which could be used to drive further geochemical sampling of areas that are of interest as a result of their probable geochemical properties.

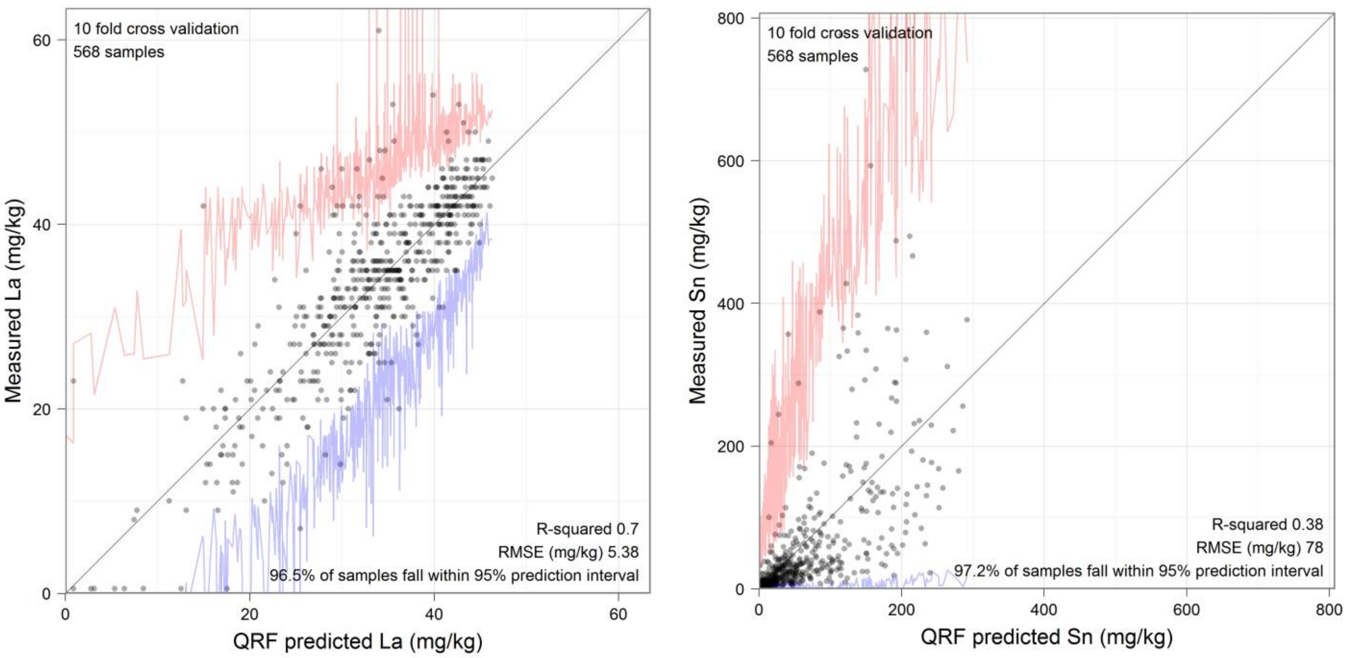


Fig. 3. Quantile regression forest predicted concentration vs measured concentration scatter plots for La and Sn. For each quantile regression forest prediction the 2.5th percentile is shown in blue and the 97.5th percentile shown in red; these are percentiles of the distribution of the outputs of the individual decision trees in the forest. The range between the 2.5th and 97.5th percentiles forms the 95% prediction interval; a measure of the uncertainty associated with each prediction.

A comparison of the fit of the predicted values between La and Sn reveals how the fit is deteriorated for the more mobile, highly-skewed, elements; prediction accuracy (and certainty) decreases in the long tail of the data. This is not explicitly due to the data having a skewed distribution, as random forest techniques are scale and transformation invariant. Rather, it is the inevitable result of having fewer data points on which to base the learning of the most ‘extreme’ situations within the context of the auxiliary variables. In this case, these situations are likely to represent relatively rare spikes of localised mineralisation. A geochemical sampling strategy designed around the auxiliary variable feature-space rather than the geographic space would take more samples from the locations of these ‘extreme’ situations and should improve the learning of the distributions of mobile elements (or any highly skewed target variable).

Table 2
Cross-validated measures of quantile regression forest model quality.

Target variable	Cross-validated R^2	RMSE (mg/kg)	Range-normalised RMSE	Moran's I of residuals	Samples in 95% prediction interval (%)
Ag	0.00	0.24	0.27	0.000	96.3
Al2O3	0.79	21552	0.10	-0.002	98.2
As	0.12	87.12	0.25	-0.006	97.7
Ba	0.76	52.57	0.13	0.001	96.1
Bi	0.01	4.46	0.11	-0.001	96.8
Br	0.62	26.58	0.10	0.001	96.7
CaO	0.01	14932	0.08	-0.003	97.5
Cd	0.20	0.28	0.24	-0.005	98.4
Ce	0.73	8.85	0.12	0.000	96
Co	0.50	7.15	0.14	-0.006	96.5
Cr	0.41	86.93	0.15	0.001	97.2
Cs	0.17	15.87	0.23	-0.008	96.7
Cu	0.24	34.54	0.22	-0.007	97.7
Fe2O3	0.70	12962	0.14	-0.001	96.7
Ga	0.67	3.57	0.12	-0.003	97.9
Ge	0.19	0.49	0.21	0.011	98.1
Hf	0.27	1.46	0.17	-0.008	97.7
I	0.13	7.93	0.18	-0.002	97.4
K2O	0.70	3771	0.11	-0.004	96
La	0.70	5.38	0.12	-0.004	96.5
LOI	0.72	71562	0.08	-0.006	97
MgO	0.53	3610	0.13	-0.006	97.9
MnO	0.25	1233	0.19	0.000	96.3
Mo	0.14	0.92	0.19	-0.004	97
Na2O	0.39	2082	0.17	0.001	98.6
Nb	0.28	4.22	0.17	-0.004	97.9
Nd	0.56	6.60	0.17	-0.005	96.1
Ni	0.46	32.67	0.13	-0.001	97.5
P2O5	0.28	1091	0.21	0.011	98.2
Pb	0.14	41.74	0.24	0.003	98.1
pH	0.48	0.65	0.18	-0.011	97.4
R	0.76	79204	0.09	-0.005	96
Rb	0.67	42.57	0.12	-0.002	96
Sb	0.10	4.86	0.13	0.003	96.3
Sc	0.69	2.85	0.15	-0.002	97.4
Se	0.34	0.49	0.16	0.001	96.8
SiO2	0.61	71748	0.10	-0.005	97.5
Sm	0.12	1.82	0.23	-0.005	98.8
Sn	0.38	77.97	0.26	-0.007	97.2
Sr	0.05	73.40	0.09	-0.002	98.1
Ta	0.23	1.19	0.16	-0.001	97
Te	0.00	0.07	0.32	0.001	98.2
Th	0.65	1.69	0.09	0.002	96.7
TiO2	0.49	2153	0.14	-0.005	95.4
Tl	0.44	0.37	0.17	0.002	95.6
U	0.22	2.49	0.13	0.000	96
V	0.68	27.58	0.15	-0.005	97
W	0.05	19.25	0.23	0.001	96.7
Y	0.47	5.26	0.18	-0.001	97.2
Zn	0.32	63.29	0.24	-0.001	97.9
ZrO2	0.37	68.71	0.14	-0.010	98.1

4.2 Geochemical maps

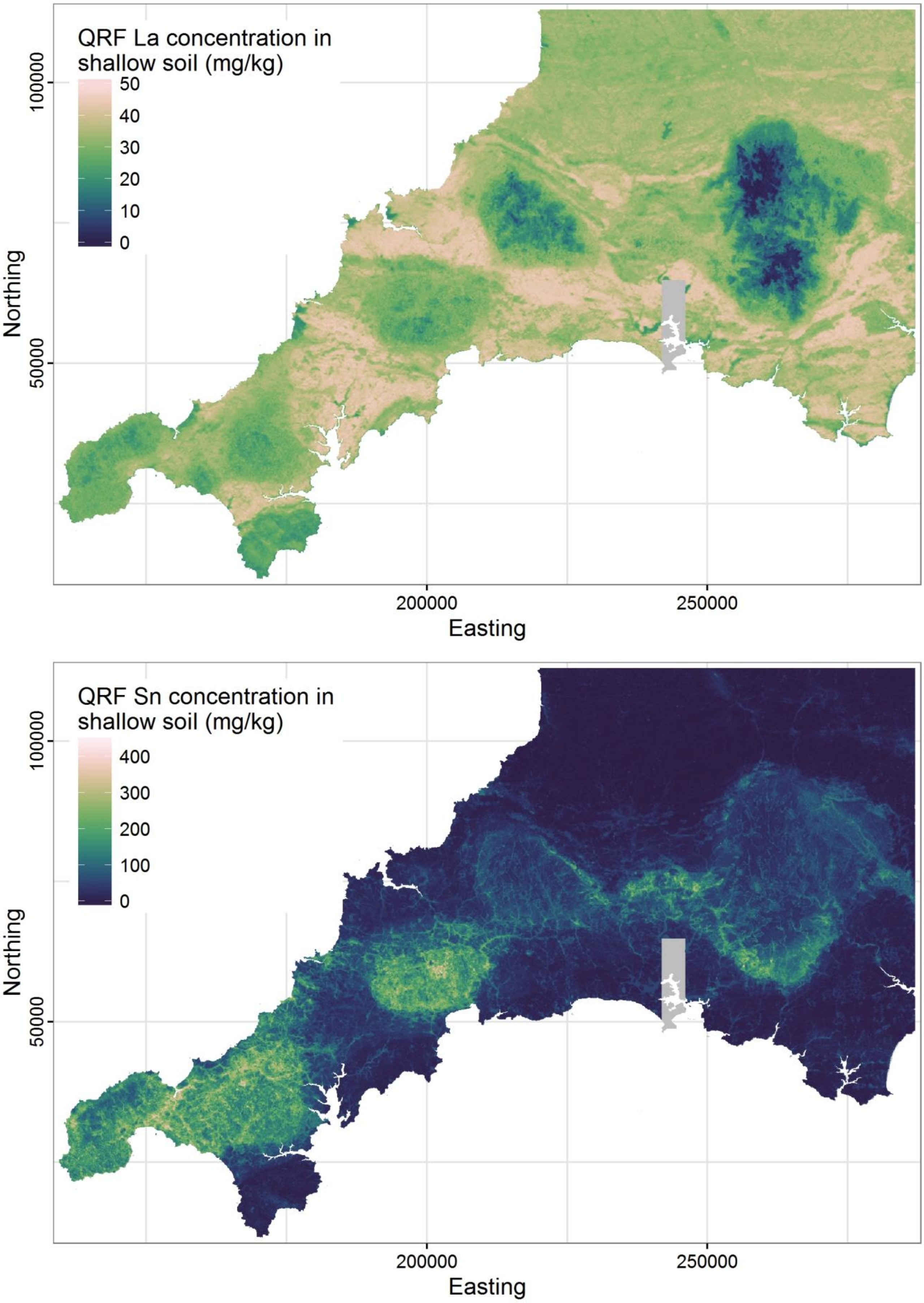


Fig. 4. Quantile regression forest predicted concentration maps for La and Sn in shallow soils.

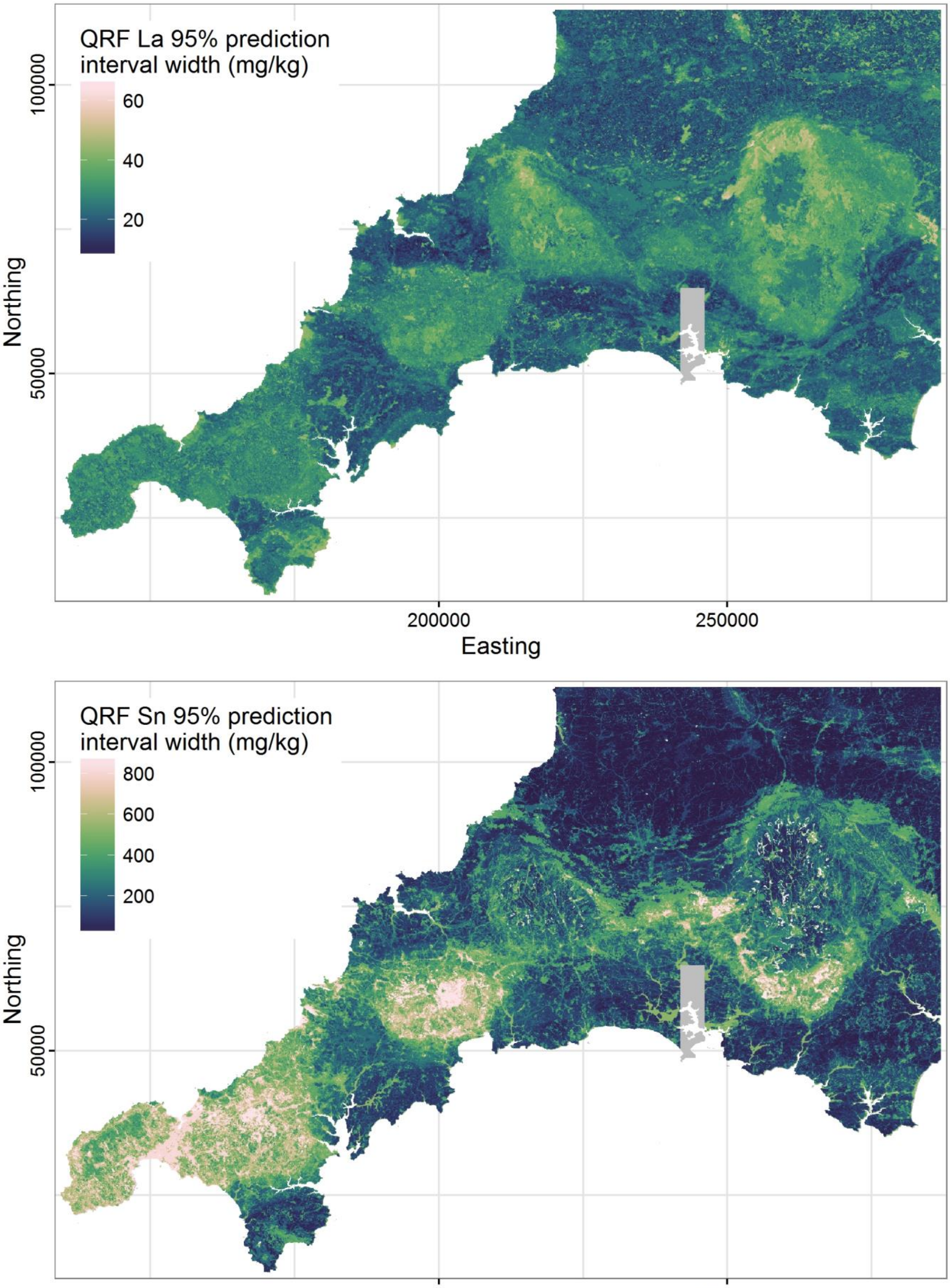
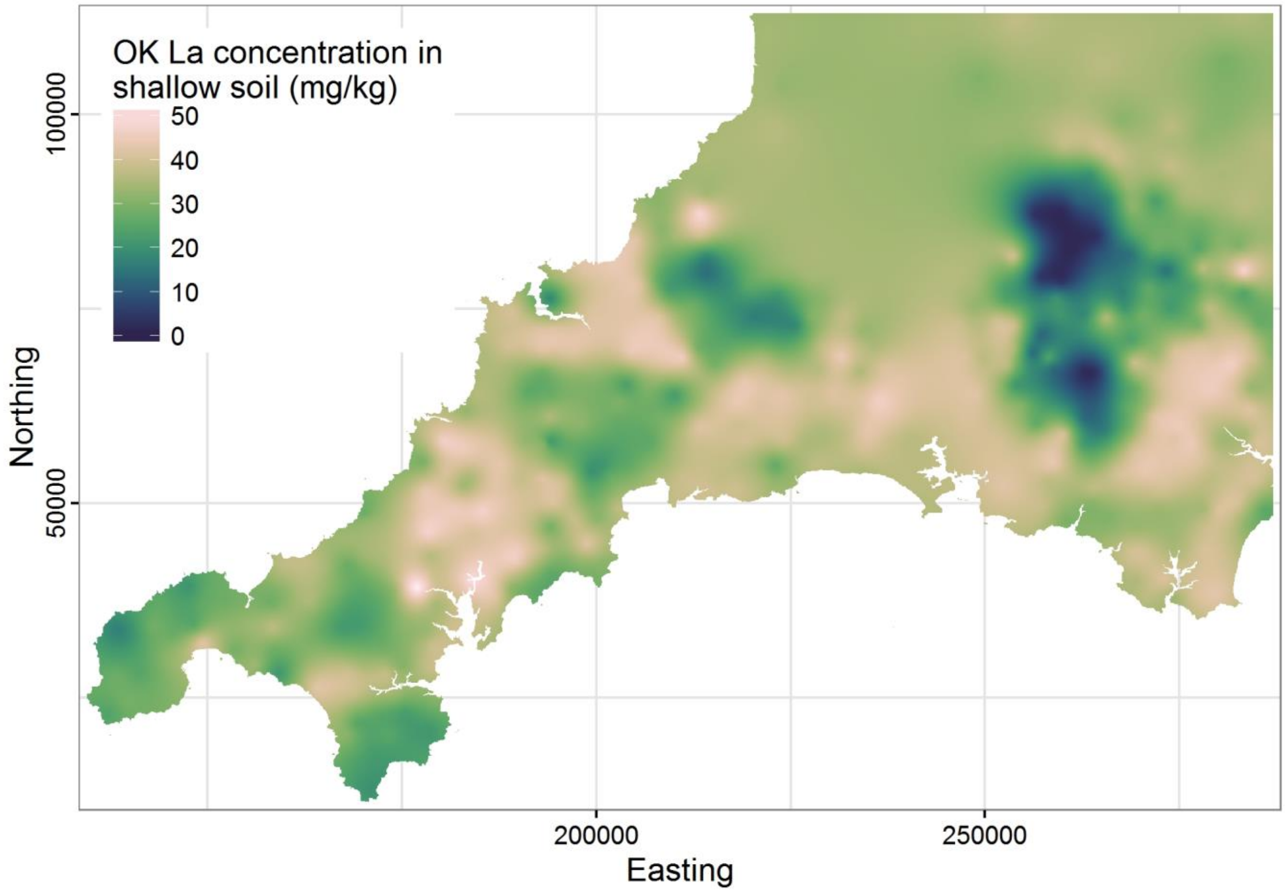


Fig. 5. Quantile regression forest prediction interval maps for La and Sn in shallow soils.



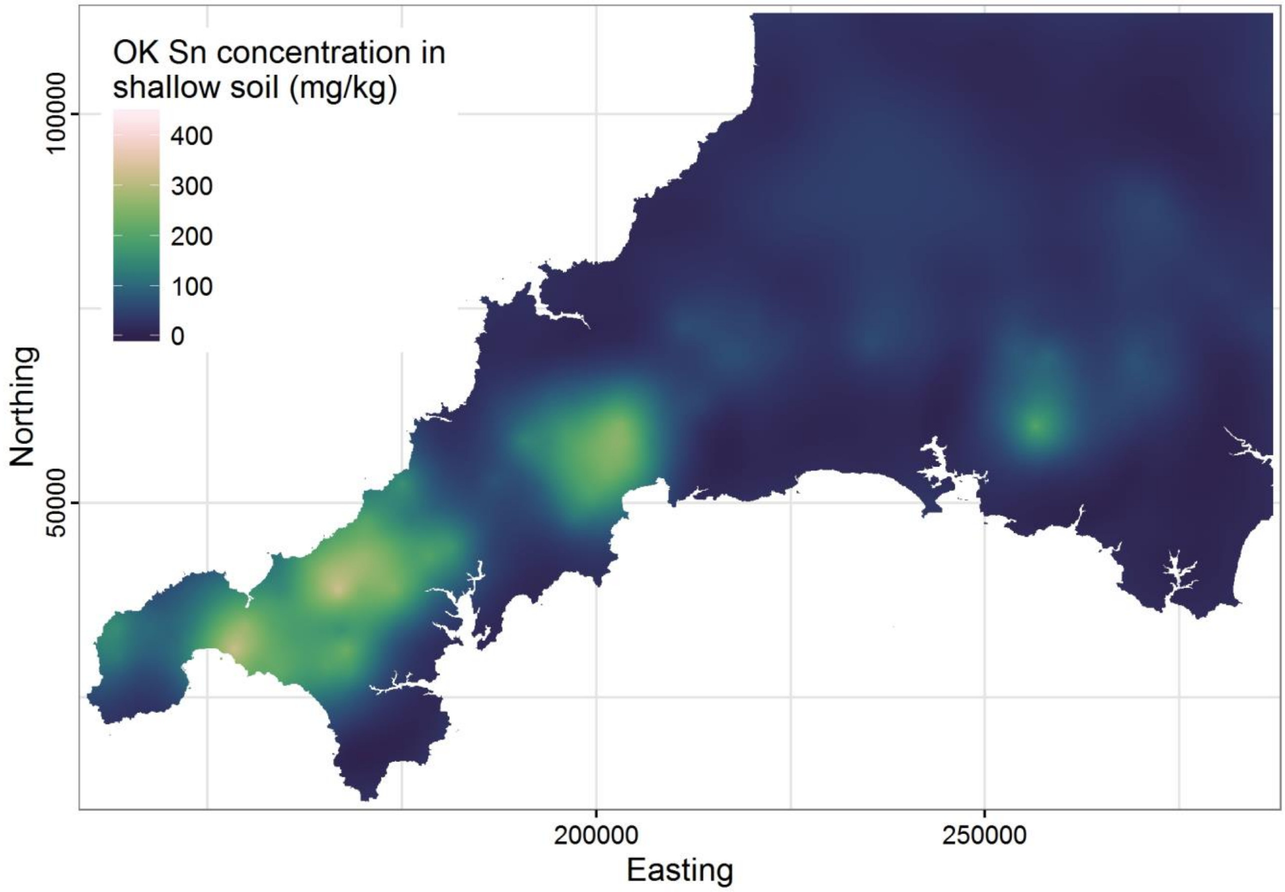
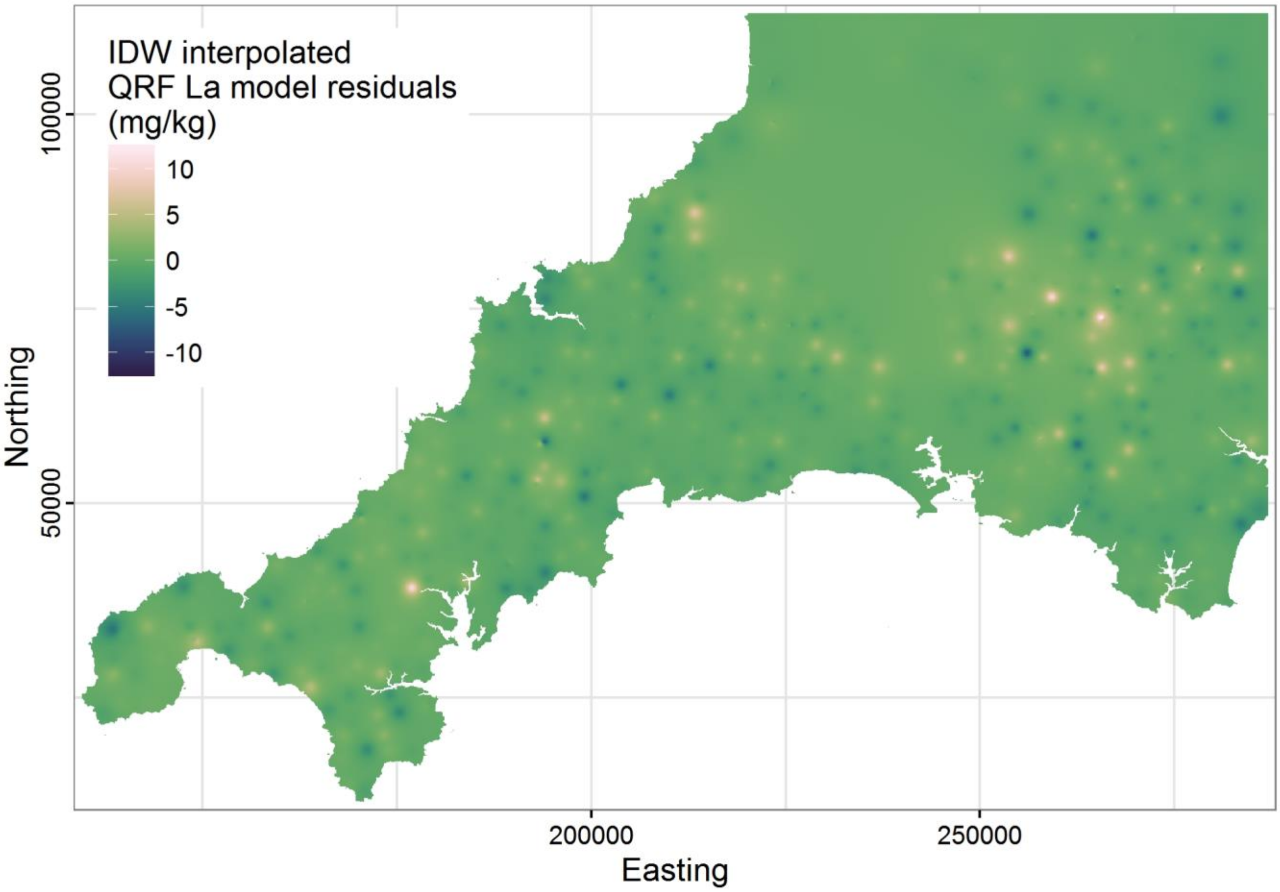


Fig. 6. Ordinary kriging predicted concentration maps for La and Sn in shallow soils, for comparison.



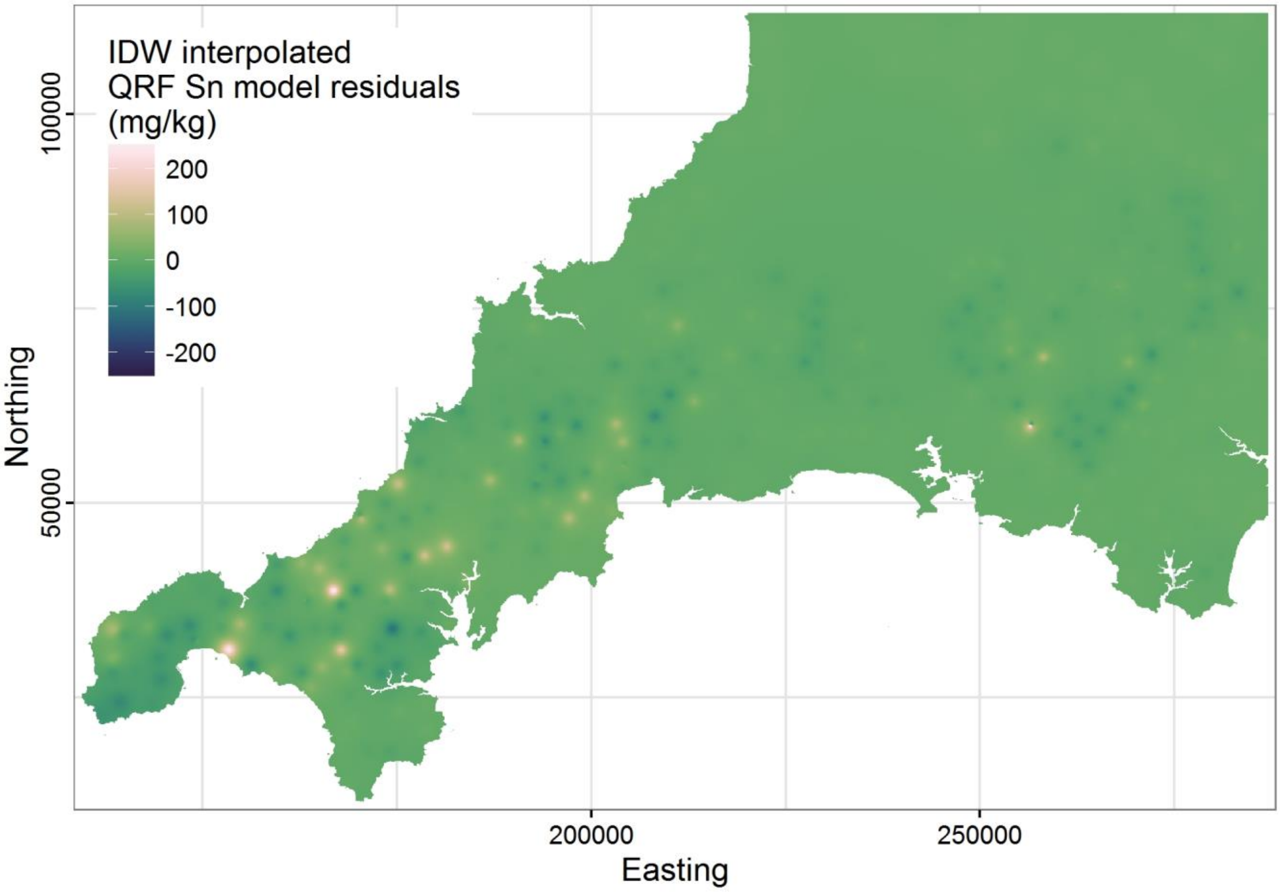


Fig. 7. Quantile regression forest residuals for La and Sn in shallow soils, interpolated using inverse distance weighting.

The geochemical maps produced using the quantile regression forest method have a spatial resolution governed by that of the auxiliary variables. Accordingly, with a resolution of 100 m, these maps are capable of resolving the spatial distribution of the elements in much more detail than traditional inverse distance weighted or ordinary kriged interpolated geochemical maps, which are limited by the spatial density of the geochemical sampling. The increased detail is evident when comparing concentration maps produced by quantile regression forests (Fig. 4) and ordinary kriging (Fig. 6). In addition, all quantile regression forest concentration maps are accompanied by uncertainty maps (Fig. 5) in the form of mapped prediction intervals – 95% in the case of this study, but it is possible to map any chosen quantile or interval for each of the quantile regression forest predictions. The quantile regression forest model residual maps (Fig. 7) display the lack of spatial autocorrelation within the residuals in agreement with the Moran's I results (Table 2). Inverse distance weighted interpolation, rather than kriging, was used to visualise the residuals as their variograms exhibited pure nugget, and kriging would therefore have produced maps of flat zero values. This reinforces the assertion that the quantile regression forest models are accounting for the spatial autocorrelation of the element concentrations at the scale of the auxiliary variable grid. The quantile regression forest maps for both example elements – La and Sn (Fig. 4) provide insight into the geochemistry of the region at a level of detail never before seen.

A traditional geochemical map interpretation would involve qualitative comparison of trends seen in the map with trends seen in other datasets. For example, geochemical maps might be compared with geological maps to try to understand the relationships between bedrock geology and surface geochemistry. The details of south west England's geology are beyond the scope of this paper, but it is well summarised by Shail and Leveridge (2009). A traditional interpretation of the quantile regression forest La map (Fig. 4) might conclude that the concentration of La in soil is strongly constrained by the underlying lithology, a relationship which the high resolution quantile regression forest map reveals in detail. Similarly, a traditional interpretation of the quantile regression forest Sn map (Fig. 6) might conclude that the concentration of Sn in soil is strongly controlled by hydrothermal mineralisation and as a result has become concentrated in close proximity to the granite intrusions, though the relationship is not consistent for all intrusions. However, interpretation of the quantile regression forest models themselves, rather than just the geochemical maps, allows the quality of interpretations of the controls on element distributions to be improved over traditional methods.

4.2 Controls on element distributions

Considering the relative importance of each auxiliary variable to the prediction of each element is a simple means by which to gain insight into the controls on the distributions of each element. In addition to this, partial dependence plots provide insight into the nature of the relationship between each predictor and the target variable. The end user can use this information to devise better informed interpretations and hypotheses of the controls on an element's distribution.

For example, the quantile regression forest model for La concentration finds elevation to be the most important predictor, followed by regional bouguer anomaly, residual bouguer anomaly and radiometric thorium concentration (Fig. 8). The negative correlation between La and elevation at elevations above 200 m indicates a close associated with the granites – which are found outcropping as elevated plateaus at ≤200 m. Furthermore, the association between La and the presence of granites is also evident in the regional bouguer anomaly – whose signal is dominated by the granites – as a sharp transition at around -11 mGal, which represents the granite-country rock contact. As can be expected, the same granite contact is less imposing in the residual bouguer anomaly, which captures fine scale (shallow depth) gravitational variations that are more influenced by other less deep-rooted lithologies in the region. More subtle lithological information in the La map appear to be revealed by the radiometrics data, in particular the relationship between La and Th. The multimodal appearance of this and other partial relationships is an effect of interaction between predictor variables. For example the La–Th relationship appears to fork into two probable trends upwards of 10 ppm of Th. Colouring the points according to elevation reveals that it is an interaction of Th with elevation (and the inversely correlated regional bouguer anomaly) which separates the upper trend from the lower trend. The lower trend, formed of samples of high elevation and low bouguer anomaly, represents the distinct relationship between La and Th over granites compared to the steeper and more linear relationship between La and Th on the surrounding rocks of lower elevation.

In contrast, the quantile regression forest model for Sn concentration finds regional bouguer anomaly, total magnetic intensity (TMI), radiometrics uranium and elevation to be the most important predictors (Fig. 9). The negative correlation between Sn and regional bouguer anomaly can be taken as proxy for the relationship between Sn and granite; generally, Sn values are elevated on and around granite bodies. The gradual transition to the Sn plateau upwards of 10 mGal gives some indication of the mobility of Sn, whose concentrations at the regional scale form gradational rather than sharp boundaries. The relationship between Sn and TMI is complex, but there is a strong negative relationship between Sn concentration and TMI values between -50 and 0 nT, particularly over granite (low regional bouguer anomaly), although it does not extend beyond this range. Similarly, there is a strong positive relationship between Sn and radiometric U between 1.9 and 2.1 ppm U which presumably represents the transition onto granite. The broadly negative relationship between Sn and elevation is heavily influenced by interactions. With the help of a regional bouguer anomaly based colour scheme it is apparent that this relationship is relatively weak over the granites, but indicates increased Sn concentrations at lower granite elevations. This may represent the fact that, on average, the interiors of the granites have lower Sn concentrations than the perimeters due to differentiation between granite phases, and the influence of hydrothermal processes. The off-granite relationship is stronger, and shows an almost exponential increase in Sn concentrations descending towards sea level from an elevation of about 100 m, above which the influence of elevation on Sn is fairly negligible. This may relate to Sn enrichment of floodplains as a result of sediment transport from mineralised areas.

4.3 A note on compositions, LOI and the unmeasured 'remainder', R.

Despite not implementing compositional data analysis methods (Aitchison, 1986; Egozcue et al., 2003; Pawlowsky-Glahn and Buccianti, 2011) to intrinsically ensure that modelled element concentrations sum to 100% at every prediction point (at the cost of computational expense and additional complexity to interpretations), we find that the sum of predicted concentrations of measured elements, and the unmeasured 'remainder' (R), fall very close to 100% in the vast majority of situations (Fig. 10). The 95% interval of summed predictions (predicted element concentrations plus predicted remainder concentration) spans from 96.0% to 105.4%. In addition, we find that R has a very close relationship with loss on ignition (LOI): their quadratic relationship could be explained by a discrepancy in calibration between the two measurement methods, but it appears that they are essentially two separate measures of the same thing (Fig. 11). The models of LOI and R achieved some of the highest prediction accuracies in the study according to the cross-validated R^2 and normalised RMSE metrics (Table 2).

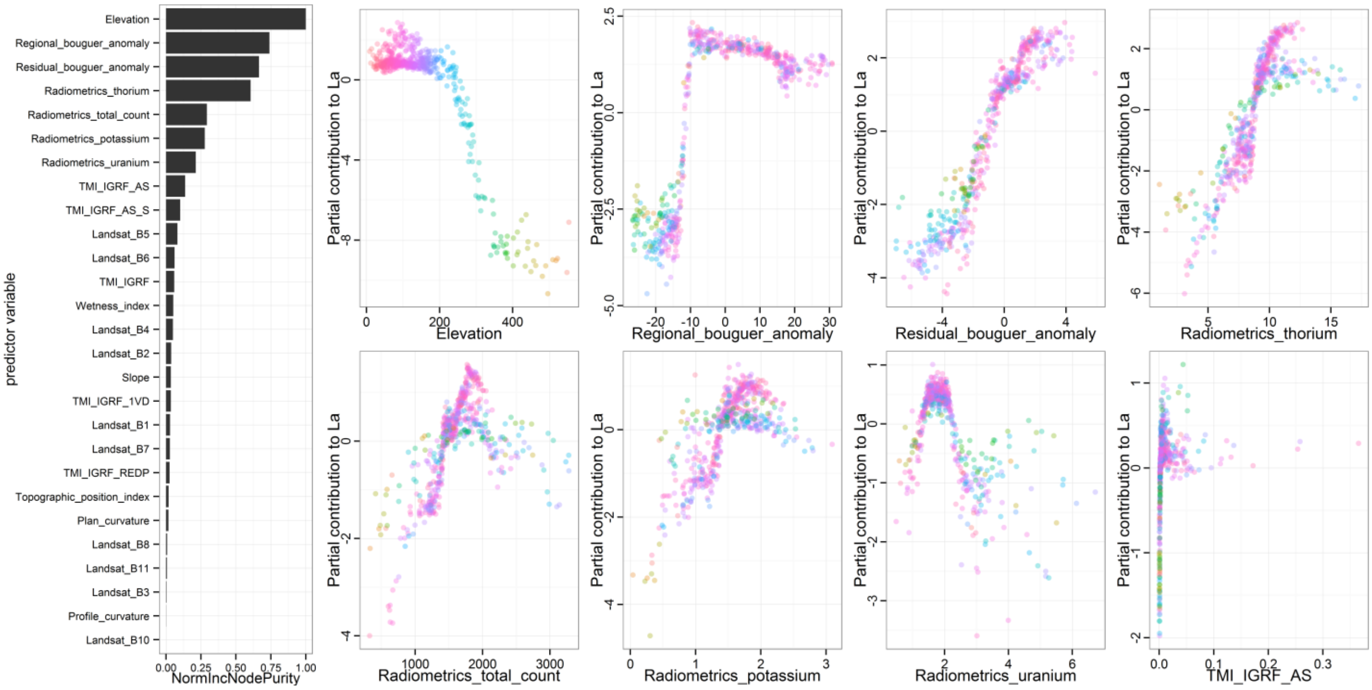


Fig. 8. Variable importance plot and top eight most important partial dependence plots for La, with points coloured according to elevation (the most important predictor).

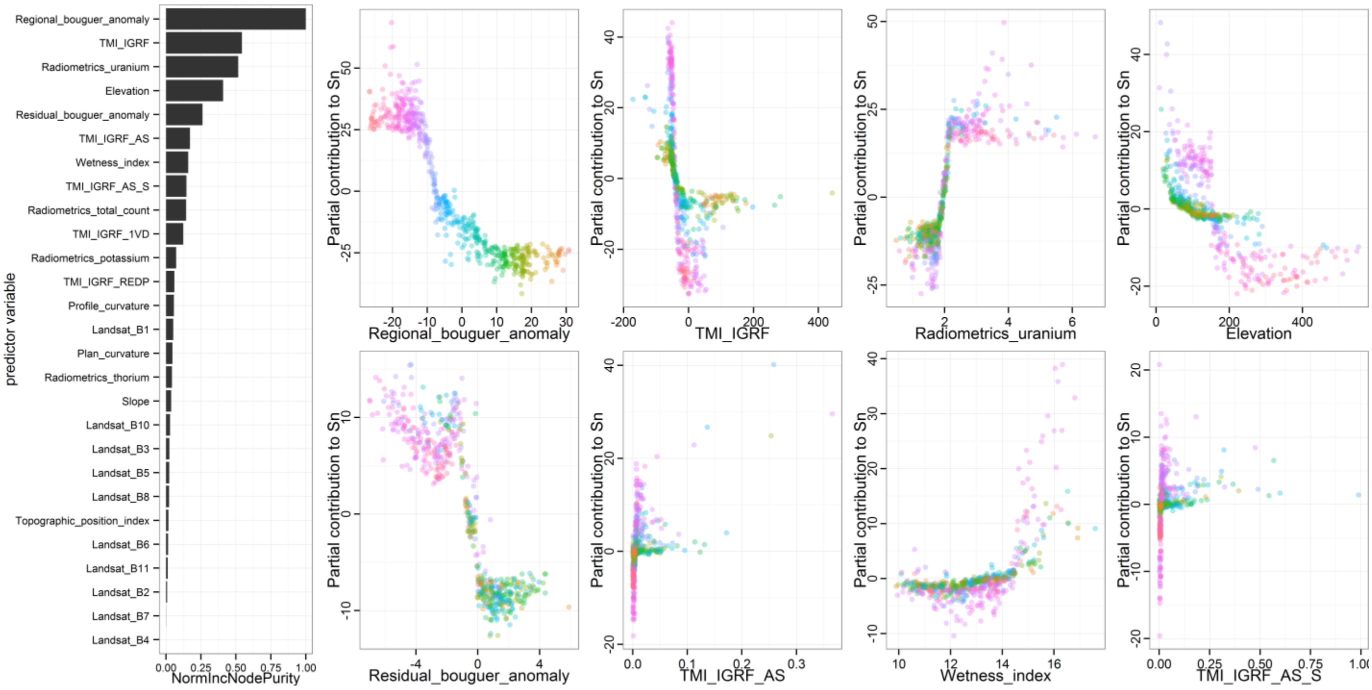


Fig. 9. Variable importance plot and top eight most important partial dependence plots for Sn, with points coloured according to regional bouguer anomaly (the most important predictor).

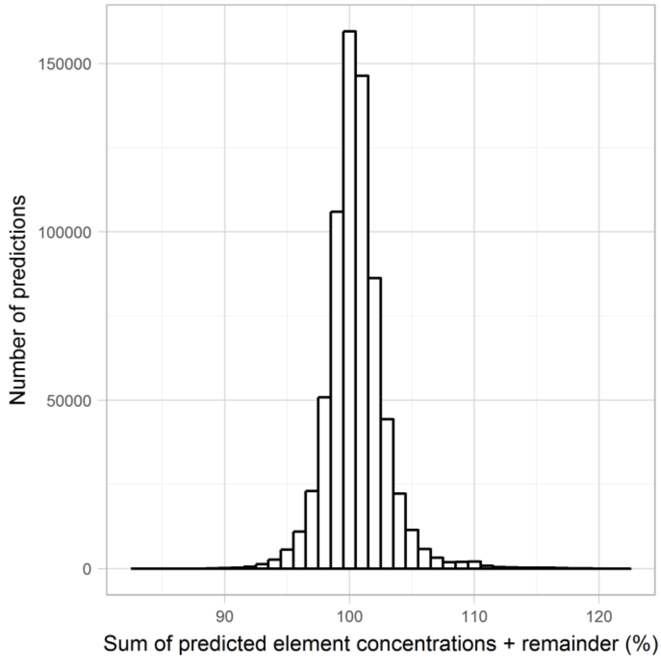


Fig. 10. Sum of predicted element concentrations + R.

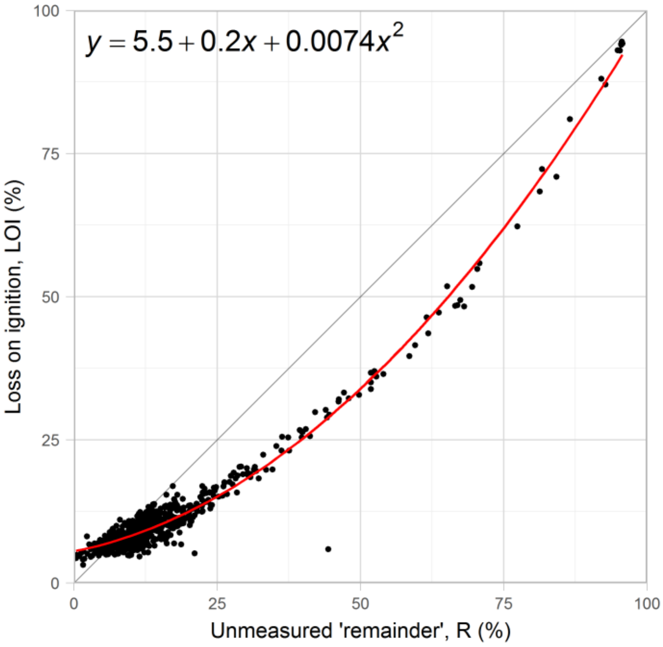


Fig. 11. Relationship between LOI and R in training data. The equation describes a quadratic curve (red line) which fits the data with an R^2 of 0.98.

5. Conclusions

The implementation of quantile regression forests to map regional soil geochemistry at high resolution (100 m) using only information from auxiliary variables has produced very encouraging results. The major, immobile, elements are modelled with sufficient accuracy to promote the development of fully quantitative geological mapping using remotely sensed data such as those used in this study. Immobile elements are modelled with a lesser degree of accuracy due to a combination of the relative under-sampling of their ‘extreme’ events (which could be improved with a change in sampling design to target anomalous locations in the context of the available auxiliary variables) and perhaps a lack of relevant information in existing auxiliary variables. Further developments to sampling design strategies, sensing technologies, and auxiliary variable derivatives (or the use of more advanced learners) should be capable of improving the modelling of mobile elements in the future.

For now, these models are capable of making an interpretable and uncertainty-aware prediction of the geochemical properties of the soil at any point on the basis of magnetic, gravity, radiometric, spectral and topographic information. The prediction process is similar to the decision making process which might be made by a human, but with the objectivity and accuracy of an optimally self-training algorithm. Allowing the model to consider the spatial dependence of the target variables might gain improvements in some situations, but the Moran's I results of the residuals suggest that the processes controlling the residuals appear to be operating randomly at the scale of the geochemical survey, and so it is the case that we currently do not have sufficient information to explain them.

The maps produced by the quantile regression forests are more useful than their spatially interpolated equivalents, providing increased detail, accuracy, interpretability and uncertainty awareness. Accordingly, the use of machine learning methods in conjunction with geophysical, radiometric, spectral and topographic information seems very capable of bringing significant improvements to geological mapping, agriculture, environmental survey and mineral exploration practices, and all the policies that surround them.

Acknowledgements

This research was funded by the British Geological Survey. Thanks to all colleagues and reviewers who have helped to guide this study. Thanks also to all G-BASE volunteers for their hard work in collecting a valuable geochemical dataset.

Aitchison, J., 1986. The statistical analysis of compositional data. Chapman & Hall, London.

Alderton, D., Pearce, J.A., Potts, P., 1980. Rare earth element mobility during granite alteration: evidence from southwest England. Earth and Planetary Science Letters 49, 149-165.

Alloway, B.J., 1990. Heavy metals in soils. Blackie & Son Ltd.

Appleton, J., Ridgway, J., 1993. Regional geochemical mapping in developing countries and its application to environmental studies. Applied geochemistry 8, 103-110.

Beamish, D., Howard, A.S., Ward, E.K., White, J., Young, M.E., 2014. Tellus South West airborne geophysical data. Natural Environment Research Council, British Geological Survey.

Beus, A.A., Grigorian, S.V., 1977. Geochemical exploration methods for mineral deposits.

Bowie, S.H.U., Thornton, I., 1985. Environmental geochemistry and health. Springer Science & Business Media.

Breiman, L., 1996. Bagging predictors. Machine learning 24, 123-140.

Breiman, L., 2001. Random forests. Machine learning 45, 5-32.

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and regression trees. CRC press.

British Geological Survey et al., 1968. GB Land Gravity Survey. British Geological Survey.

Carranza, E.J.M., Laborte, A.G., 2015. Random forest predictive modeling of mineral prospectivity with small number of prospects and data with missing values in Abra (Philippines). Computers & Geosciences 74, 60-70.

Colbourn, P., Alloway, B., Thornton, I., 1975. Arsenic and heavy metals in soils associated with regional geochemical anomalies in south-west England. Science of the Total Environment 4, 359-363.

Cracknell, M.J., Reading, A.M., 2014. Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information. Computers & Geosciences 63, 22-33.

Cressie, N., 1988. Spatial prediction and ordinary kriging. Mathematical Geology 20, 405-421.

Cutler, D.R., Edwards Jr, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., Lawler, J.J., 2007. Random forests for classification in ecology. Ecology 88, 2783-2792.

Darnley, A.G., 1990. International geochemical mapping: a new global project. Journal of Geochemical Exploration 39, 1-13.

Dines, H.G., 1956. The metalliferous mining region of south-west England. HM Stationery Office.

Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barcelo-Vidal, C., 2003. Isometric logratio transformations for compositional data analysis. Mathematical Geology 35, 279-300.

Evans, J.S., Murphy, M.A., Holden, Z.A., Cushman, S.A., 2011. Modeling species distribution and change using random forest, Predictive Species and Habitat Modeling in Landscape Ecology. Springer, pp. 139-159.

Fordyce, F.M., 2013. Selenium deficiency and toxicity in the environment. Springer.

Gislason, P.O., Benediktsson, J.A., Sveinsson, J.R., 2006. Random forests for land cover classification. Pattern Recognition Letters 27, 294-300.

Green, D., 2011. A colour scheme for the display of astronomical intensity images. arXiv preprint arXiv:1108.5083.

Harris, J., Grunsky, E., Behnia, P., Corrigan, D., 2015. Data-and knowledge-driven mineral prospectivity maps for Canada's North. Ore Geology Reviews.

Hawkes, H.E., Webb, J.S., 1962. Geochemistry in mineral exploration.

Henderson, B.L., Bui, E.N., Moran, C.J., Simon, D., 2005. Australia-wide predictions of soil properties using decision trees. Geoderma 124, 383-398.

Hengl, T., Heuvelink, G.B., Stein, A., 2003. Comparison of kriging with external drift and regression-kriging. Technical note, ITC 51.

Hiemstra, P.H., Pebesma, E.J., Twenhöfel, C.J., Heuvelink, G.B., 2009. Real-time automatic interpolation of ambient gamma dose rates from the Dutch radioactivity monitoring network. Computers & Geosciences 35, 1711-1721.

Intermap Technologies, 2007. NEXTMap British Digital Terrain Model Dataset Produced by Intermap, NERC Earth Observation Data Centre.

Johnson, C., Breward, N., Ander, E., Ault, L., 2005. G-BASE: baseline geochemical mapping of Great Britain and Northern Ireland. Geochemistry: Exploration, Environment, Analysis 5, 347-357.

Jordan, W.J., Alloway, B.J., Thornton, I., 1975. The application of regional geochemical reconnaissance data in areas of arable cropping. Journal of the Science of Food and Agriculture 26, 1413-1423.

Kirby, G., 1979. The Lizard complex as an ophiolite.

Kirkwood, C., Everett, P., Ferreira, A., Lister, B., 2016. Stream sediment geochemistry as a tool for enhancing geological understanding: An overview of new data from south west England. Journal of Geochemical Exploration 163, 28-40.

Knotters, M., Brus, D., Voshaar, J.O., 1995. A comparison of kriging, co-kriging and kriging combined with regression for spatial interpolation of horizon depth with censored observations. Geoderma 67, 227-246.

Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection, Ijcai, pp. 1137-1145.

Lawrence, R.L., Wood, S.D., Sheley, R.L., 2006. Mapping invasive plants using hyperspectral imagery and Breiman Cutler classifications (RandomForest). Remote Sensing of Environment 100, 356-362.

Levinson, A.A., 1974. Introduction to exploration geochemistry.[Textbook].

Lewis, G., Thornton, I., Howarth, R., 1986. Geochemistry and animal health, Applied geochemistry in the 1980s: proceedings of a meeting to honour the contribution of professor John S. Webb to applied geochemistry, held on 29 April 1983 at Imperial College, London. John Wiley & Sons, p. 260.

Liaw, A., Wiener, M., 2002. Classification and regression by randomforest. R News 2 (3): 18–22. URL: http://CRAN.R-project.org/doc/Rnews.

Lin, Y., Jeon, Y., 2006. Random forests and adaptive nearest neighbors. Journal of the American Statistical Association 101, 578-590.

Liu, M., Wang, M., Wang, J., Li, D., 2013. Comparison of random forest, support vector machine and back propagation neural network for electronic tongue data classification: Application to the recognition of orange beverage and Chinese vinegar. Sensors and Actuators B: Chemical 177, 970-980.

Meinshausen, N., 2006. Quantile regression forests. The Journal of Machine Learning Research 7, 983-999.

Moran, P.A., 1950. Notes on continuous stochastic phenomena. Biometrika, 17-23.

Palczewska, A., Palczewski, J., Robinson, R.M., Neagu, D., 2013. Interpreting random forest models using a feature contribution method, Information Reuse and Integration (IRI), 2013 IEEE 14th International Conference on. IEEE, pp. 112-119.

Pawlowsky-Glahn, V., Buccianti, A., 2011. Compositional data analysis: Theory and applications. John Wiley & Sons.

R Core Team, 2014. R: A Language and Environment for Statistical Computing, R version 3.1.1 (2014-07-10) ed. R Foundation for Statistical Computing, Vienna, Austria.

Reid, R., Horvath, D., 1980. Soil chemistry and mineral problems in farm livestock. A review. Animal Feed Science and Technology 5, 95-167.

Reimann, C., Siewers, U., Tarvainen, T., Bityukova, L., Eriksson, J., Gilucis, A., Gregorauskiene, V., Lukashev, V., Matinian, N., Pasieczna, A., 2003. Agricultural soils in Northern Europe: a geochemical atlas. E. Schweizerbart'sche Verlagsbuchhandlung.

Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., Chica-Rivas, M., 2015. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geology Reviews.

Rodriguez-Galiano, V.F., Ghimire, B., Rogan, J., Chica-Olmo, M., Rigol-Sanchez, J.P., 2012. An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS Journal of Photogrammetry and Remote Sensing 67, 93-104.

Roy, D.P., Wulder, M., Loveland, T., Woodcock, C., Allen, R., Anderson, M., Helder, D., Irons, J., Johnson, D., Kennedy, R., 2014. Landsat-8: Science and product vision for terrestrial global change research. Remote Sensing of Environment 145, 154-172.

Salminen, R., Tarvainen, T., Demetriades, A., Duris, M., Fordyce, F., Gregorauskiene, V., Kahelin, H., Kivisilla, J., Klaver, G., Klein, H., 1998. FOREGS geochemical mapping field manual.

Shail, R.K., Leveridge, B.E., 2009. The Rhenohercynian passive margin of SW England: Development, inversion and extensional reactivation. Comptes Rendus Geoscience 341, 140-155.

Smedley, P.L., 1991. The geochemistry of rare earth elements in groundwater from the Carnmenellis area, southwest England. Geochimica et Cosmochimica Acta 55, 2767-2779.

Thornton, I., 1993. Environmental geochemistry and health in the 1990s: a global perspective. Applied geochemistry 8, 203-210.

Thornton, I., Plant, J., 1980. Regional geochemical mapping and health in the United Kingdom. Journal of the Geological Society 137, 575-586.

Vanwinckelen, G., Blockeel, H., 2012. On estimating model accuracy with repeated cross-validation, BeneLearn 2012: Proceedings of the 21st Belgian-Dutch Conference on Machine Learning, pp. 39-44.

Venables, W.N., Ripley, B.D., 2013. Modern applied statistics with S-PLUS. Springer Science & Business Media.

Webb, J., Thornton, I., Nichol, I., 1971. The agricultural significance of regional geochemical reconnaissance in the United Kingdom. Trace Elements in Soils and Crops, Min. Agr. Fish. Food Tech. Bull 21, 1-7.

Welling, S.H., 2015. forestFloor: Visualizes Random Forests with Feature Contributions. URL: http://CRAN.R-project.org/package=forestFloor.

White, J.G., Zasoski, R.J., 1999. Mapping soil micronutrients. Field Crops Research 60, 11-26.

Wiesmeier, M., Barthold, F., Blank, B., Kögel-Knabner, I., 2011. Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem. Plant and soil 340, 7-24.

Willis-Richards, J., Jackson, N.J., 1989. Evolution of the Cornubian ore field, Southwest England; Part I, Batholith modeling and ore distribution. Economic Geology 84, 1078-1100.

Xu, Y., Cheng, Q., 2001. A fractal filtering technique for processing regional geochemical maps for mineral exploration. Geochemistry: Exploration, environment, analysis 1, 147-156.

Xuejing, X., Xueqiu, W., 1991. Geochemical exploration for gold: a new approach to an old problem. Journal of Geochemical Exploration 40, 25-48.

Zhang, G., Lu, Y., 2012. Bias-corrected random forests in regression. Journal of Applied Statistics 39, 151-160.