

Uncertainty of predictions in absorption spectroscopy: Modelling with quantile regression forest

Alexandre M.J.-C. Wadoux^a,*, Leonardo Ramirez-Lopez^{b,c}

^a LISAH, Univ. Montpellier, AgroParisTech, INRAE, IRD, L'Institut Agro, Montpellier, France

^b Data Science Department, BUCHI Labortechnik AG, Switzerland

^c Imperial College London, Imperial Business School, South Kensington, Exhibition Rd, London SW7 2AZ, United Kingdom

ARTICLE INFO

Keywords:

Machine learning
Random forest
Chemometrics
Measurement error
Pedometrics

ABSTRACT

Machine learning modelling is becoming popular for estimating agricultural and environmental properties from their infrared spectra. Commonly in modelling with machine learning and in commercial software applications, however, uncertainty estimates of the prediction are seldom reported. Uncertainty quantification of variables predicted with infrared spectroscopy is yet highly relevant in a number of applications, such as in uncertainty propagation analyses studies or for drug exposure detection. In this paper, we report on the development and application of quantile regression forest to predict properties from infrared spectroscopic data along with a sample-specific estimate of the uncertainty. Quantile regression forest is a machine learning algorithm that builds on random forest and provides estimate of the mean but also of the full conditional distribution of the predicted variable. We illustrate the algorithm with two chemometric applications and evaluate the modelling approach for its ability for predict the variable of interest and quantify the uncertainty. Evaluation involved usual validation statistics but also the validation of the uncertainty with the prediction interval coverage probability calculated for various interval widths. We tested prediction and prediction uncertainty quantification of two soil properties (cation exchange capacity and total organic carbon) as well as the dry matter of mango. The results confirm the potential of quantile regression forests for prediction and uncertainty quantification of properties predicted from infrared spectroscopy data. In all cases, the predictions were accurate and sample-specific estimates of the uncertainty were obtained. Validation of the uncertainty showed that the interval width was too large, thus overestimating the uncertainty for most intervals. Nevertheless, we recommend its use for operational applications as well as in future software developments, in particular when the data inferred by the spectroscopic model are used in other applications.

1. Introduction

Applications of regression techniques in spectral chemometrics are concerned the calibration of models that can be used to accurately quantify meaningful properties of a sample from its spectral features [1]. In the past two decades, there has been growing interest in the combination of sensing techniques with advanced data analytics methods to analyze complex matrices or products in a time- and cost-effective way. In soil science, for example, chemometrics and absorption spectroscopy techniques are used to support the routine estimation of soil organic and inorganic components (e.g. N, P, K, cation exchange capacity) [2]. Near-infrared spectroscopy, in particular, has been widely used to determine chemical and physical properties of food and agricultural products since the 80s [3], and now spans across many fields with pharmaceutical, biological and cosmetics applications, among others. Absorption spectroscopy models are typically calibrated

from datasets that include hundreds to thousands of samples. In turn, the spectra also comprise hundreds to thousands of variables which are highly auto-correlated. To address this problem, multivariate data analysis techniques were developed in chemometrics for quantitative analysis of spectra.

In the literature, multivariate models were developed to relate the concentration or amount of samples to their measured spectra. Common prediction techniques are based on spectra dimensionality reduction, such as in principal component regression and partial least-squares regression [4]. These two techniques are the standard ones in spectral chemometrics. High-dimensional spectra can also be reduced into a smaller subspace of latent variables such as with principal component analysis and wavelet transform [5], and then be used as input to relatively simple model such as multivariate linear regression and decision trees. Data-driven and algorithmic tools from the field of

* Correspondence to: LISAH, 2 place Pierre Viala, 34090 Montpellier, France.
E-mail address: alexandre.wadoux@inrae.fr (Alexandre M.J.-C. Wadoux).

<https://doi.org/10.1016/j.chemolab.2025.105473>

Received 13 February 2024; Received in revised form 14 June 2025; Accepted 19 June 2025

Available online 4 July 2025

0169-7439/© 2025 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

machine learning were also used for this purpose and are becoming popular [6]. Example machine learning models are (deep) artificial neural networks, support vector regression and random forest. These models are particularly useful for finding pattern in large and complex spectra data and for their predictive capability [7].

When an empirical spectral model is used to predict a property, ideally such prediction should be accompanied with an estimate of its reliability. In practice, uncertainty quantification is of fundamental importance for setting reliable detection limits (e.g. for quality control or inspection of raw material) and for understanding how such uncertainty propagates when predicted values are used for calibrating other models. De Bièvre [8] argued in this sense: a result without estimates of uncertainty should not be published or communicated. It seems, however, that uncertainty analysis of multivariate calibration model prediction still lag behind, in particular in commercial software applications.

Statistical models are by essence able to provide estimate of the prediction uncertainty. In partial least squares regression uncertainty in prediction have been calculated based on a linear approximation of the vector of regression coefficients or by using bootstrapping and jackknifing to obtain prediction confidence intervals [9]. Alternatively, when assuming normality of the error the prediction intervals can be constructed from the standard error of prediction and *t*-test statistics [10]. These techniques are valid and useful but cannot always be applied to machine learning models. To the best of our knowledge, questions of uncertainty quantification of predictions made by machine learning have been largely disregarded in the chemometrics and applied chemistry literature.

Random forest (RF) is a form of machine learning based on decision trees, has been used in several applications in chemometrics [e.g. by 11–13]. RF has attractive properties that can be exploited for uncertainty quantification. Recent articles have shown that RF could theoretically provide reliable prediction intervals. Briefly, RF combines several trees, each built on a bootstrap sample of the original data and coupled with a randomness component in the fitting of the tree. Meinshausen [14] proposed quantile regression forest by modifying the output of a random forest so that the spread of the property is quantified by keeping the values of each node of each tree, from which prediction intervals are constructed. Although machine learning models are becoming popular in the chemometrics literature, quantile regression forests for estimating the prediction and prediction uncertainty has not yet been described and tested to date.

The purpose of this paper is to investigate the merit of quantile regression forest to quantify the uncertainty of prediction from absorption spectroscopy data. We report the prediction results for two soil properties predicted from their near-infrared spectra and one fruit variable predicted from their visible and near-infrared spectra. We evaluate the quality of the predictions and capacity of quantile regression forest to accurately predict the sample-specific uncertainty. The sample-specific uncertainty is represented by the prediction interval width and validated by indices that show the prediction interval coverage probability of various interval widths.

2. Material and methods

2.1. Random forest

Random forest is an ensemble machine learning model that combine multiple decision trees to make a prediction [15]. A single tree is built by recursive partitioning of the training data into sub-regions using a series of splitting rule. In the root node, the predictor part of the training data is grouped into a single partition called the root node. The predictor space composed of $X_1, X_2, \dots, X_p$ is then divided into two distinct and non-overlapping regions R_1, R_2 using a procedure called binary recursive partitioning. Possible partitions are tested and evaluated using a splitting metric and the best cut point t for the predictor

partition is determined. The residual sum of squares calculated by assigning the mean of response value for the training observations in each regions is often used as splitting metric [16, p. 306]. Once the two partitions are determined, each undergo the same procedure in order to split the training data and to minimize the splitting metric further. The steps are repeated to grow a tree until a user-defined stopping criterion is met. A common stopping criterion is the minimum number of observation in the partition. For regression (i.e. for continuous variables) the prediction is taken as the average of the observation values present in the final node of the decision tree.

Decision trees are unstable because a small change in the training data can result in a large change in the estimated tree. To reduce the prediction error variance of a single decision tree, Breiman [17] introduced bagging, which stands for bootstrap and aggregating. In this procedure, an ensemble of decision trees are built based on a bootstrap sample of the training dataset. Consider B bootstrap samples drawn with replacement from the original training set, grow a decision tree on each of the $1, \dots, B$ subsamples and obtain $B = K$ trees $\phi_k; k = 1, \dots, K$. Next the predictions of all individual trees are aggregated through averaging and this average is taken as the prediction:

$$\hat{f}_{bag}(x) = \frac{1}{B} \sum_{b=1}^B \hat{f}_{\phi_b}(x). \quad (1)$$

Breiman [15] proposed random forest as a further improvement to the bagging procedure. Consider a training dataset with a very strong predictor, predictions from the bagged trees will be correlated because most trees will use the same strong predictor in the top split. To address this, the bagged trees are decorrelated by means of an extra random perturbation during the splitting of the tree. For each partition, the algorithm is allowed to consider only one of subset m predictors from the total set of p predictors. Typically p is approximately as the rounded down square root of the total number of predictors (i.e. $p \approx \lfloor \sqrt{p} \rfloor$). The random forest estimate $\hat{f}(x)$ is taken as the average of the individual trees prediction.

2.2. Quantile regression forest (QRF)

The random forest estimate $\hat{f}(x)$ is an approximation of the conditional mean $E[Y|X = x]$ by averaging the observations of the response Y in the final node (i.e. the leaves). Meinshausen [14] introduced quantile regression forest (QRF) that does not average the observations of the response Y but instead keeps the distribution of all observed values of the leaf of each tree in the random forest. It thus reports the conditional probability function on which we can calculate the percentiles. The QRF procedure is summarized in Fig. 1.

Consider the conditional probability function F :

$$F(y|X = x) = P(Y \leq y|X = x). \quad (2)$$

Knowledge about this function enables calculating the α quantile defined as $Q_\alpha(x)$ so that the probability of Y being smaller than Q_α is approximately equal to α at x . Using α and F one can determine the quantile boundaries which results in $\alpha_{lower} = \frac{1-p}{2}$ and $\alpha_{upper} = \frac{1+p}{2}$ for the lower and upper boundaries, respectively. The prediction interval for width p (hereafter p -PI) can then be computed by:

$$p - \text{PI}(x) = [Q_{\alpha_{lower}}, Q_{\alpha_{upper}}]. \quad (3)$$

In Meinshausen [14] the function F is estimated in two steps. In the first step, the weight of a single tree is calculated as the proportion of the response observations in the terminal leaf, relative to the total number of observations that are present within the same leaf. In a second step, the weights of all trees are aggregated to construct the empirical conditional cumulative distribution function $\hat{F}$. Meinshausen [14] defines this function from the distribution of the training data within each leaf as:

$$\hat{F}(y|X = x) = \sum_{i=1}^n w_i(x) \mathbb{1}_{\{Y_i \leq y\}}, \quad (4)$$

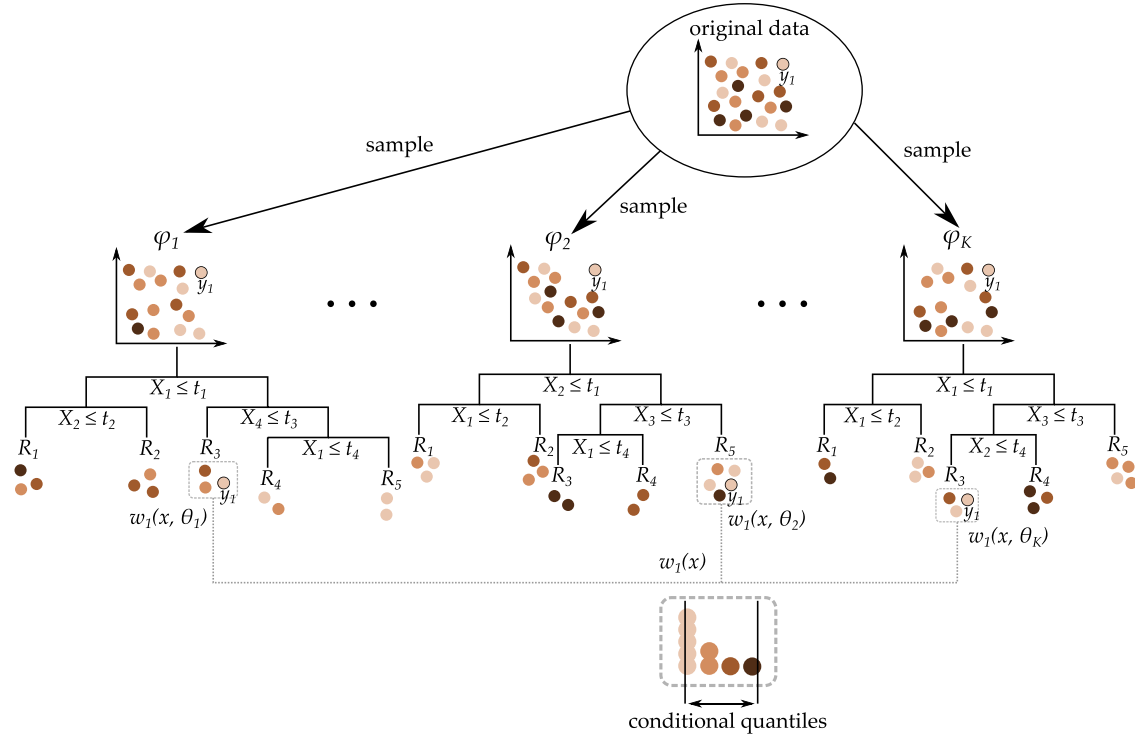


Fig. 1. Example visualization of a quantile regression forest model with three repetitions of the bootstrap from the original dataset. For an observation y_i the calculation of the weights at the leaves that the observation falls in is illustrated as well as the estimation of the conditional quantiles from which the prediction intervals and mean prediction can be derived.

Source: Adapted from Nussbaum [18].

where the indicator $\mathbb{1}$ determines whether the weights are counted or not, depending on $Y_i \leq y$. Thus, $E(\mathbb{1}_{\{Y \leq y\}} | X = x)$ is the weighted mean over the observations of $\mathbb{1}_{\{Y \leq y\}}$, and the weights are defined as the average of the weights $w_i(\theta)$ for a collection of K trees:

$$w_i(x) = \frac{1}{K} \sum_{t=1}^K w_i(x, \theta_t), \quad (5)$$

where θ is a parameter vector which determines how a tree is grown and $w_i(\theta)$ is the weight of a single tree:

$$w_i(x, \theta) = \frac{\mathbb{1}_{X_i \in R_{\ell(x, \theta)}}}{\#j : X_j \in R_{\ell(x, \theta)}}, \quad (6)$$

where the weight is a positive constant if X_i is part of the leaf $\ell(x, \theta)$ and zero otherwise and $\ell(x, \theta)$ is the leaf (i.e. the terminal node) for tree $\varphi(\theta_k)$ Meinshausen [14] summarized the algorithm as follows:

1. Grow K trees $\varphi(\theta_k)$ to create a random forest, but unlike random forest, for every leaf and every tree keep all observations in the leaf, not only the average of these observations.
2. For $X = x$ drop x down to all trees in the random forest and compute the weights for every tree (Eq. (6)) and compute the average of $w_i(x, \theta_k)$ over all trees (Eq. (5)).
3. Compute the estimate of the empirical conditional cumulative distribution function (Eq. (4)) and estimate the conditional quantile $\hat{Q}_\alpha(x)$ as $\inf \{y : \hat{F}(y|X=x) \geq \alpha\}$.

2.3. Validation of prediction

Validation of the prediction is made using common statistical indices using a 5-fold cross-validation strategy. The dataset was divided into 5 approximately equally sized folds. Four of these folds were used to fit the QRF model and predict the variable of interest on the remaining fold. This procedure was repeated 100 times, each time setting a different fold aside. For each validation fold the predicted

value by QRF was calculated and the error was derived by subtracting the predicted values from their corresponding observed ones. The predictions were plotted against the actual observations in a pair-wise validation scatterplot and statistical indices of the overall error and model performance were computed. We calculated the mean error (ME), the root mean square error, the squared Pearson's r correlation coefficient (r^2) and the Nash-Sutcliffe modelling efficiency (NSE). The latter is defined as [19]:

$$\text{NSE} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \quad (7)$$

where y denote the observed value of the property, $\hat{y}$ is its predicted value by QRF and $\bar{y}$ denotes the mean of the observed values. The NSE has an optimal value of 1 and it can be negative. It denotes the improvement made by the model over using the mean of the observations as predictor.

2.4. Validation of the uncertainty

To estimate the quality of the uncertainty quantification we used the prediction interval coverage probability. Consider a p -PI defined by the quantiles $(1-p)/2$ and $(1+p)/2$, the uncertainty is validated by counting the occurrence of the observed values falling within the p -PI. In statistical terms, let the fraction of occurrence within the quantile be expressed

$$\bar{\xi}(p) = \frac{1}{n} \sum_{i=1}^n \xi(y_i; p) \quad \forall p \in [0, 1], \quad (8)$$

where $\xi(s; p)$ is an indicator function given by:

$$\xi(y_i; p) = \begin{cases} 1 & \text{if } \hat{F}_i^{-1}\left(y_i; \frac{1-p}{2}\right) < y_i \leq \hat{F}_i^{-1}\left(y_i; \frac{1+p}{2}\right) \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

For all p in the interval $[0, 1]$ the value $\tilde{\xi}(p)$ is plotted against the nominal p value in a so-called accuracy plot [20,21]. The accuracy plot shows the fraction of cross-validation observations within a p interval, for various intervals. For example, for a 90% prediction interval (i.e. $p = 0.9$), lower quantile is $(1 - p)/2 = 0.05$ and upper quantile is $(1 + p)/2 = 0.95$, 90% of the cross-validated observations should fall within this interval. A perfect modelling of the uncertainty would therefore show the points on the 1:1 line in the accuracy plot (i.e. $\tilde{\xi}(p) = p$). Above this line the uncertainty is over-estimated whereas it is under-estimated if it falls under the line. Further, the values of $\tilde{\xi}(p)$ we averaged over 100 repetitions of the cross-validation, and these iterations were used to construct a 90% confidence interval.

Finally, we identified the specific observations that fall outside the interval width using the 100 repetitions of the cross-validation. For a 0.9 prediction interval, we counted the number of times each observation fall inside the prediction interval using an indicator function that returns 1 if the observation is inside and 0 if the observation is outside.

2.5. Absorbance spectroscopic data

Soil data. We used the near-infrared dataset from the “Chimimétrie 2006” challenge [22]. The dataset consist of 825 reflectance spectra of soil samples measured between 1100 and 2498 nm and recorded at a resolution of 2 nm. The spectra were obtained from dried and sieved soil samples collected within the cultivated areas of the Walloon region in Belgium. Two soil properties are available with the reflectance data: CEC (Cation Exchange Capacity in mmol(+) per 100 g of dry soil) and TOC (Carbon in percent mass). The reflectance spectra were resampled to a resolution of 5 nm. The spectra were then pre-processed with a standard normal variate (SNV) transformation followed by a Savitzky-Golay first derivative with a window of 25 nm and a first order polynomial. An example of 5 pre-processed spectra (out of the 825) is shown in Fig. 2.

Mango data. The mango dataset [23] consists of 4675 visible and near-infrared spectra. The apparent absorption spectra were collected between 300–1100 nm and recorded at a resolution of 3.3 nm following the protocol described in [23, section 2.1]. The dry matter content (DMC, % fresh weight) of mango fruit is provided in the dataset. The dry matter content was obtained through dehydration for 48 h at 65 °C. The near-infrared spectra were resampled to a resolution of 3 nm, and pre-processed with a SNV transformation followed by a Savitzky-Golay second derivative with a window of 51 nm and a second order polynomial. An example of 5 pre-processed spectra (out of the 4675) is shown in Fig. 2.

2.6. Outlier analysis

An outlier detection was applied by projecting the pre-processed spectra onto their principal components (PC) space using PC analysis. The components that explained more than 1% of the original spectral variance were retained. The Mahalanobis distance (MD) from every observation to the center of the data was computed on the PC scores space. Samples that showed MD values larger than 3 were considered as outliers and removed from the analysis. A total of 5 observations were removed for the soil dataset and no observations were removed for the mango dataset.

2.7. Practical implementation

The analysis was done in the R programming language [24]. Pre-processing of spectra was done with prospectr package [25] and outlier detection using the resemble package [26]. Fitting of the quantile regression forest was made with functions available in the ranger package. The number of trees was fixed at 150, whereas the two other user-defined QRF parameters, the minimum terminal node size and

Table 1

Validation statistics (mean $\pm$ standard deviation) obtained over 100 repetitions of the cross-validation procedure, for the two soil properties (Cation exchange capacity in mmol(+) per 100 g of dry soil and TOC in percent mass) and the dry matter of the mango in % fresh weight.

	ME	RMSE	r	NSE
Soil TOC	-0.02 ± 0.01	0.7 ± 0.03	0.92 ± 0.1	0.84 ± 0.01
Soil CEC	-0.06 ± 0.04	3.7 ± 0.1	0.81 ± 0.14	0.65 ± 0.02
Mango DMC	0 ± 0	0.77 ± 0	0.95 ± 0.03	0.9 ± 0

mtry (i.e. the number of covariates to randomly select at each split of the tree) were set to 5 and 80 respectively. All validation statistics were computed by averaging 100 repetitions of the cross-validation procedure.

3. Results and discussion

3.1. Modelling results

Table 1 shows the validation statistics of the QRF prediction for two soil properties (CEC and TOC) and the mango DMC. Note that we report the average of 100 repetitions of the cross-validation along with the standard deviation of the repetitions. Table 1 shows that the predictions have little bias (ME is close to zero in all cases). The Pearson's r correlation coefficient show that predictions have a strong linear correlation with the observed values in all cases. The best results are achieved for the mango data, where the NSE is 0.9, but good to excellent results are also obtained for the soil dataset (NSE is 0.65 and 0.84 for CEC and TOC, respectively).

These results are comparable or superior to previous studies using the same dataset and a different modelling strategy. For the soil dataset and the TOC property, Pierna and Dardenne [22] reported a RMSE values between 0.43 and 1.48 for different modelling strategy. The lowest RMSE was achieved in their case with a local modelling whereas the standard partial least square regression yielded a RMSE of 1.14. Minasny and McBratney [27] reported a RMSE of 0.35 and 0.76 for modelling with cubist and treenet, respectively. For the property CEC we obtained better or slightly better results than previous studies. For example, Pierna and Dardenne [22] reported RMSE values between 3.3 and 5.3 whereas Minasny and McBratney [27] obtained values between 5.4 and 10.5, which is to be compared with our RMSE value of 3.7. The modelling of the mango dataset yielded better results than all the tests conducted in Anderson et al. [23] for various modelling strategies (partial least squares, neural networks, and various validation tests). The RMSE value of 0.7 that we obtained is lower in 23 out of the 24 modelling strategies that they tested. We also achieved better results for the mango dataset than Diaz et al. [28] in which they found a RMSE of 0.93, 1.02, 0.87.

3.2. Uncertainty quantification

The scatterplot of predicted versus observed values along with the estimated uncertainty is shown in Fig. 3. The grey bars indicate the 90% prediction interval characterized by the 0.05 and 0.95 percentiles whereas the coloured bar indicates the 50% interval (i.e. the 0.25 and 0.75 percentiles). The colour represent the proportion of observation in the validation set falling within the 50% prediction interval, for 100 repetitions of the cross-validation. Fig. 3 confirms the statistics reported in Table 1. The predictions have no bias but a smoothing occurred where small values are slightly overestimated while large values seem underestimated.

Overall, Fig. 3 shows that some data points are better predicted than others, because the 50% and 90% prediction intervals vary greatly in length. This might be due to the hyperparameter combination, which we discuss later in this manuscript. The figure also further indicates

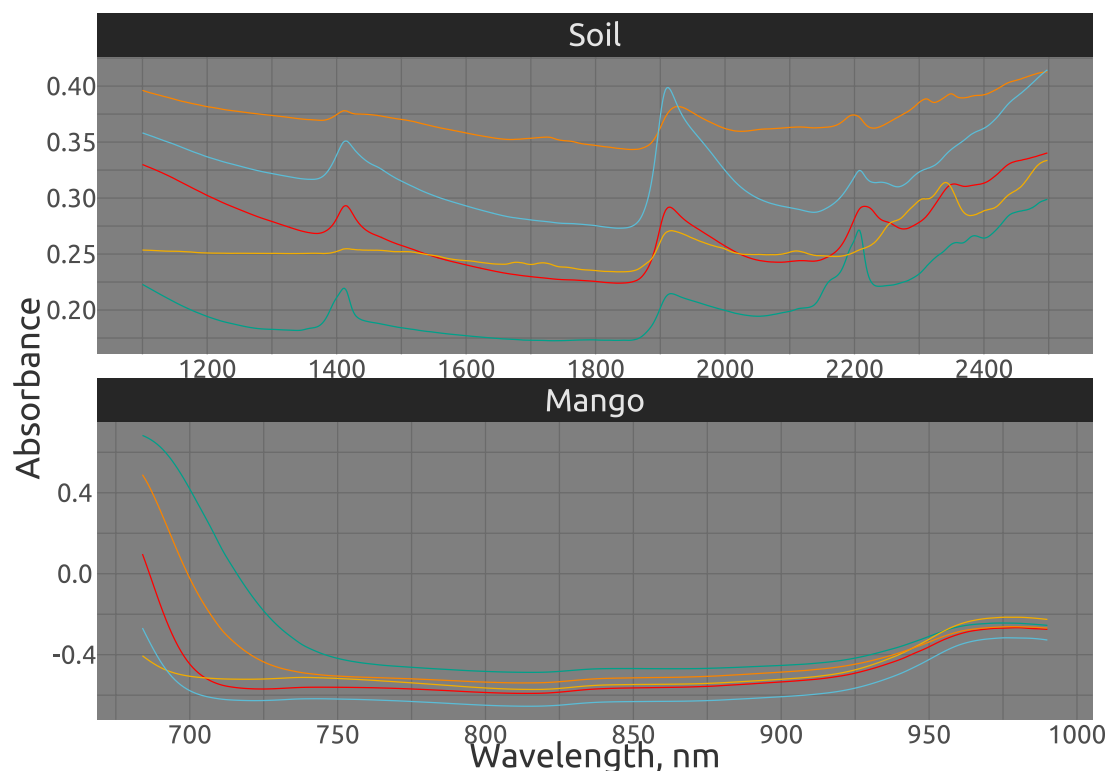


Fig. 2. Example set of five pre-processed spectra for the soil and mango datasets. The spectra were selected from the whole dataset using Kennard–Stone sampling.

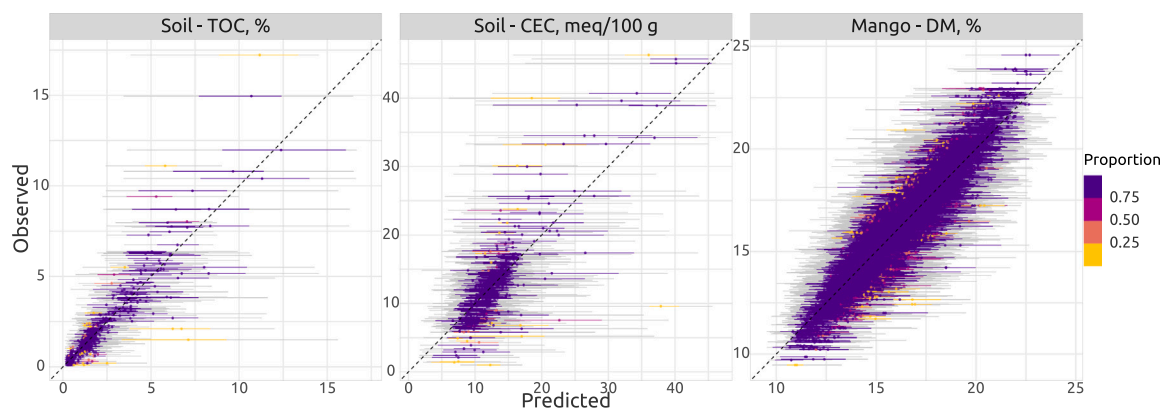


Fig. 3. Scatterplot of predicted versus observed values for the three properties. The grey bars indicate the 90% prediction interval whereas the coloured bar indicates the 50% interval. The colour represent the proportion of observation in the validation set falling within the 50% prediction interval over 100 repetitions of the cross-validation.

that most points systematically fall within the 50% interval (i.e. the proportion being 1). Few points are nearly always outside the prediction interval (yellow colour). Grey error bars that do not cross the 1:1 line are also outside the 90% prediction interval.

Fig. 4 shows the ordered predicted values along with the predicted value and the prediction intervals, for various prediction interval widths. The upper plots show that the interval width may vary greatly. For the soil properties CEC and TOC, the interval width changes, which indicates that some values are more accurate than others. There is no clear pattern of increasing the interval width for larger values of CEC or TOC. A different pattern is observed for the Mango DM, in which the interval width seem nearly constant. This indicates that most values are predicted with similar accuracy.

The bottom plots of Fig. 4 show the ordered predicted values against the predicted values and prediction intervals subtracted from the mean value of the dataset. Two observations can be made. The first is that low values of soil TOC have a much wider prediction intervals (e.g the 50%

interval in yellow) than the rest of the data, but that the interval widths increases for large values of TOC. The second is that, conversely, the soil CEC have an opposite pattern: large values of soil CEC have wide intervals and interval width tends to increase for small values of soil CEC. This results indicates that, for TOC, small values that are poorly predicted are below the detection limit of the spectroscopic model. The detection limit is the lowest detectable amount of a property of a sample given a method [29]. In our case, Fig. 4 shows that this detection limit is about 0.4% of organic carbon, which is much below the results reported in the literature [e.g. for organic matter³⁰]. Another explanation for the large prediction intervals at low TOC values is because of measurement error of the TOC data used to fit the model, in which case the TOC measurement are below 0.4% are too inaccurate to be used in model fitting. For CEC, the large prediction interval width for large values of CEC can be explained by the smoothing that is observed in Fig. 3: large values are strongly underestimated .

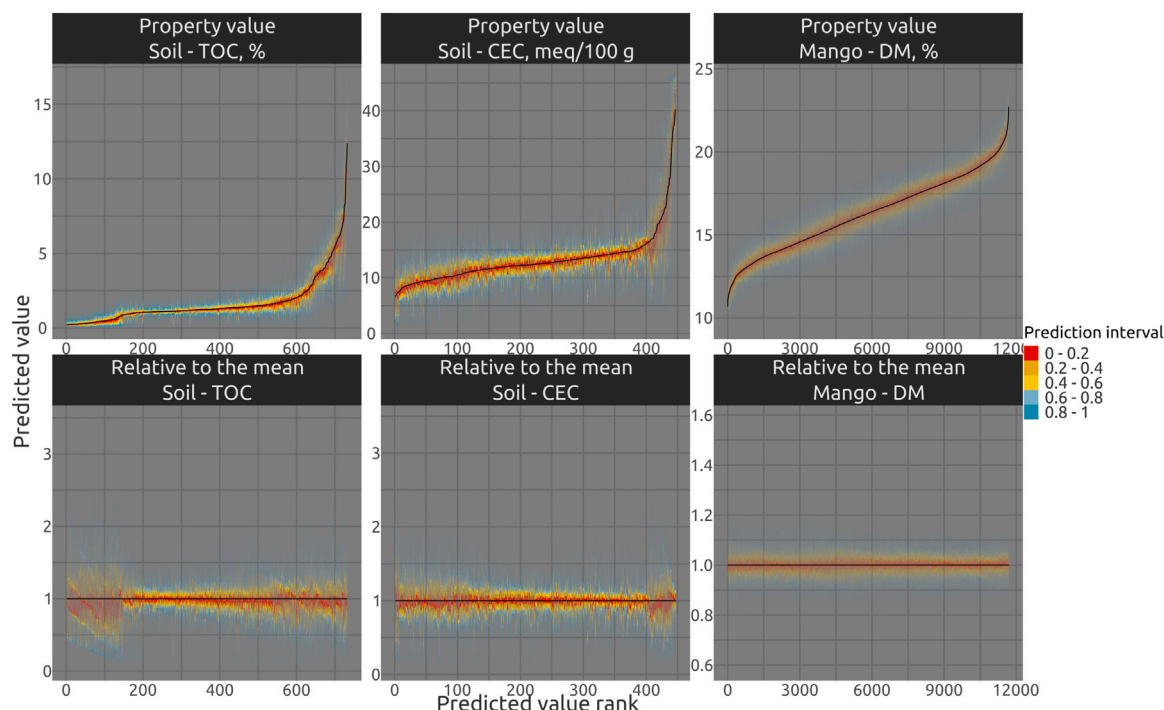


Fig. 4. Ordered predicted values against their nominal value. The black line is the mean predicted values and the colour indicates the prediction intervals, for 5 interval widths.

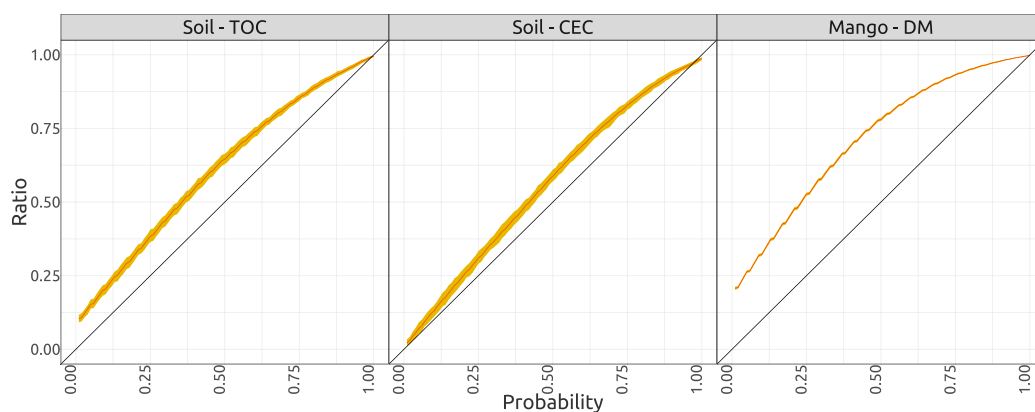


Fig. 5. Accuracy plot of the prediction interval versus the proportion of values falling within the interval, for various interval widths. The shaded area is the 90% confidence interval of the estimates obtained with the 100 repetitions of the cross-validation.

3.3. Validation of the uncertainty estimates

Fig. 5 shows that accuracy plots for the validation of the quantified uncertainty along with the 90% confidence interval obtained by the 100 repetitions of the cross-validation. In general, the uncertainty was overestimated. This was more pronounced for the mango dataset than for the soil dataset. Overestimation was considerable between the 50% and 80% prediction intervals. For example, in the 75% prediction interval of the soil TOC dataset, nearly 87% of the observations in the validation data would fall within this interval. Overestimation, however, decreases considerably around the 90% prediction interval, similar to what is usually reported in the literature [e.g. in]. For example, Fig. 5 reveals that about 94, 92 and 95% of the observations fall within this interval for TOC, CEC and mango DM, respectively.

Previous studies that quantify uncertainty for predictions from absorption spectroscopic models, usually disregard validation of the uncertainty estimates. For example, while Skou et al. [31] reported prediction intervals of urea concentrations obtained by partial least square and near infrared spectroscopy but did not provided information on

the validation of such estimates. To our knowledge, reporting a validation of the uncertainty is still uncommon practice in chemometrics. Validation of the uncertainty estimates is crucial so that end-users are not over-confident about the predictions obtained with spectroscopic models. The provision of an estimate of the prediction uncertainty is relevant for many downstream applications which take as input spectroscopically inferred data. In spatial modelling of soil properties, where spectroscopic predictions are commonly used, for example, it is common to propagate the uncertainty of the input data to the mapping product when it comprises both analytical and spectroscopically inferred data. Uncertainty estimates of the input data, however, are rarely available [32].

While QRF can provide prediction uncertainty estimates, we found random forest to be sensitive to the choice of parameters. We tested this by calculating the prediction NSE for various values of Mtry, minimum node size and number of trees, for the three properties investigated. The results for TOC (Fig. 6), CEC and DM (Supplementary Materials) show a similar pattern. The model predictions are strongly influenced by the minimum node size: the prediction benefits from a

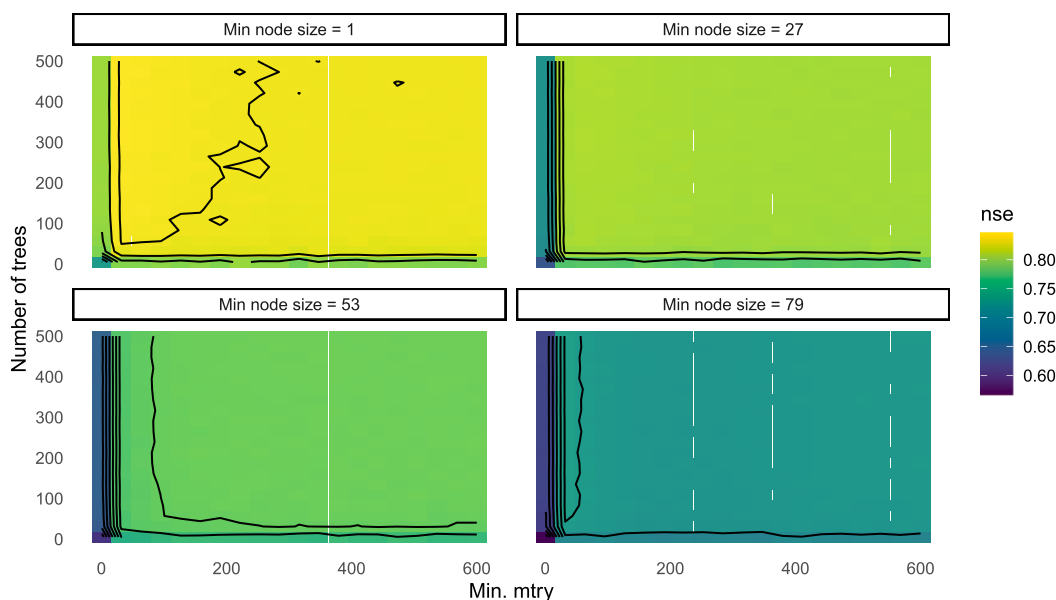


Fig. 6. Heat maps showing the sensitivity of the TOC prediction to the random forest parameters: the minimum node size, Mtry and the number of trees. nse: Nash–Sutcliffe model efficiency coefficient.

small number of observations in the final node. The number of tree and Mtry were less important but small value for these two parameters were preferred. We did not test QRF on synthetic datasets. This is because QRF has already been extensively evaluated in the literature, both on real-world and synthetic datasets. For instance, the seminal paper introducing the QRF [14] demonstrated the use of QRF through experiments on well-known datasets. Another example is Vaysse and Lagacherie [33] in which the performance of QRF for both prediction and uncertainty quantification is tested to address geospatial prediction of soil properties. These studies, and others, rendered further synthetic testing of the basic QRF behaviour in our study unnecessary.

4. Conclusion

The study evaluated the merit of quantile regression forest for predicting properties from infrared spectroscopic data along with a sample-specific estimate of the uncertainty. Quantile regression forest was evaluated both in terms of prediction accuracy and quality of the quantified uncertainty. Two public datasets comprising infrared spectroscopic data coupled with either soil properties and mango dry matter were used to test the algorithm.

The results showed that the prediction accuracy of the quantile regression forest algorithm was high and similar or better than existing results reported in the literature. More importantly, the algorithm estimated the uncertainty of the prediction by providing the full conditional distribution of the prediction from which quantiles can be calculated. We found that some values were predicted with more accuracy than others, and that large prediction intervals were obtained for values that fell below the detection limit of the model. Finally, validation of the estimated quantiles showed that the uncertainty was overestimated, but that this overestimation was negligible for the 90% prediction interval, which is the most commonly reported.

Machine learning is becoming popular for spectroscopic modelling, but to date there are limited options to obtain sample-specific uncertainty estimates of the predictions. Quantile regression forest can be used for this purpose, thus combining the high accuracy of machine learning algorithm and the capability of providing the prediction intervals. Quantifying the uncertainty of the prediction is still uncommon practice in spectroscopic modelling with machine learning, but it is necessary in a number of applications where spectroscopically inferred properties are used as input to other applications.

CRediT authorship contribution statement

Alexandre M.J.-C. Wadoux: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Leonardo Ramirez-Lopez:** Writing – review & editing, Writing – original draft, Validation, Methodology, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

All authors have participated in (a) conception and design, or analysis and interpretation of the data; (b) drafting the article or revising it critically for important intellectual content; and (c) approval of the final version.

This manuscript has not been submitted to, nor is under review at, another journal or other publishing venue.

The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript.

Acknowledgements

For the purpose of Open Access, a CC-BY public copyright licence has been applied by the authors to the present document and will be applied to all subsequent versions up to the Author Accepted Manuscript arising from this submission.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at <https://doi.org/10.1016/j.chemolab.2025.105473>.

Data availability

All data used in this study are publicly available.

References

- [1] R.G. Brereton, *Applied Chemometrics for Scientists*, John Wiley & Sons, Chichester, 2007.
- [2] M. Nocita, A. Stevens, B. van Wesemael, M. Aitkenhead, M. Bachmann, B. Barthès, E.B. Dor, D.J. Brown, M. Clairotte, A. Csorba, et al., Soil spectroscopy: An alternative to wet chemistry for soil monitoring, *Adv. Agron.* 132 (2015) 139–159.
- [3] Y. Ozaki, A.A. Christy, W.F. McClure, *Near-Infrared Spectroscopy in Food Science and Technology*, John Wiley & Sons, Chichester, 2006.
- [4] H. Wold, Partial least squares, in: S. Kotz, C.B. Read, N. Balakrishnan, B. Vidakovic, N.L. Johnson (Eds.), *Encyclopedia of Statistical Sciences*, Wiley, New York, 2006.
- [5] Z. Xiaobo, Z. Jiewen, M.J.W. Povey, M. Holmes, M. Hanpin, Variables selection methods in near-infrared spectroscopy, *Anal. Chim. Acta* 667 (1–2) (2010) 14–32.
- [6] C.A. Meza Ramirez, M. Greenop, L. Ashton, I.U. Rehman, Applications of machine learning in spectroscopy, *Appl. Spectrosc. Rev.* 56 (8–10) (2021) 733–763.
- [7] R. Houhou, T. Bocklitz, Trends in artificial intelligence, machine learning, and chemometrics applied to chemical data, *Anal. Sci. Adv.* 2 (3–4) (2021) 128–141.
- [8] P. De Bièvre, Measurement results without statements of reliability (uncertainty) should not be taken seriously, *Accredit. Qual. Assur.* 2 (1997) 269–269.
- [9] L. Zhang, S. Garcia-Munoz, A comparison of different methods to estimate prediction uncertainty using Partial Least Squares (PLS): a practitioner's perspective, *Chemometr. Intell. Lab. Syst.* 97 (2) (2009) 152–158.
- [10] A.C. Olivieri, N.M. Faber, J. Ferré, R. Boqué, J.H. Kalivas, H. Mark, Uncertainty estimation and figures of merit for multivariate calibration (IUPAC Technical Report), *Pure Appl. Chem.* 78 (3) (2006) 633–661.
- [11] D. Donald, D. Coomans, Y. Everingham, D. Cozzolino, M. Gishen, T. Hancock, Adaptive wavelet modelling of a nested 3 factor experimental design in NIR chemometrics, *Chemometr. Intell. Lab. Syst.* 82 (1–2) (2006) 122–129.
- [12] F.B. de Santana, A.M. de Souza, R.J. Poppi, Visible and near infrared spectroscopy coupled to random forest to quantify some soil quality parameters, *Spectrochim. Acta Part A: Mol. Biomol. Spectrosc.* 191 (2018) 454–462.
- [13] J. Liu, S. Sun, Z. Tan, Y. Liu, Nondestructive detection of sunset yellow in cream based on near-infrared spectroscopy and interval random forest, *Spectrochim. Acta Part A: Mol. Biomol. Spectrosc.* 242 (2020) 118718.
- [14] N. Meinshausen, Quantile regression forests, *J. Mach. Learn. Res.* 7 (6) (2006).
- [15] L. Breiman, Random forests, *Mach. Learn.* 45 (1) (2001) 5–32.
- [16] G. James, D. Witten, T. Hastie, R. Tibshirani, *An Introduction to Statistical Learning*, Springer, New York, 2013.
- [17] L. Breiman, Bagging predictors, *Mach. Learn.* 24 (2) (1996) 123–140.
- [18] M. Nussbaum, Machine learning and processing of large data, in: M.J. Goss, M. Oliver (Eds.), *Encyclopedia of Soils in the Environment*, second ed., Academic Press, Oxford, 2023, pp. 509–520.
- [19] P. Janssen, P. Heuberger, Calibration of process-oriented models, *Ecol. Model.* 83 (1–2) (1995) 55–66.
- [20] P. Goovaerts, Geostatistical modelling of uncertainty in soil science, *Geoderma* 103 (1–2) (2001) 3–26.
- [21] A.M.-C. Wadoux, D.J. Brus, G.B. Heuvelink, Accounting for non-stationary variance in geostatistical mapping of soil properties, *Geoderma* 324 (2018) 138–147.
- [22] J.A.F. Pierna, P. Dardenne, Soil parameter quantification by NIRS as a Chemometric challenge at “Chimio-métrie 2006”, *Chemometr. Intell. Lab. Syst.* 91 (1) (2008) 94–98.
- [23] N. Anderson, K. Walsh, P. Subedi, C. Hayes, Achieving robustness across season, location and cultivar for a NIRS model for intact mango fruit dry matter content, *Postharvest Biol. Technol.* 168 (2020) 111202.
- [24] R.C. Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2022, URL <https://www.R-project.org/>.
- [25] A. Stevens, L. Ramirez-Lopez, An introduction to the prospectr package, 2022, R package version 0.2.6. URL <https://CRAN.R-project.org/package=prospectr>. (Accessed 23 November 2022).
- [26] L. Ramirez-Lopez, A. Stevens, R. Viscarra Rossel, C. Lobse, A. Wadoux, T. Breure, Resemble: regression and similarity evaluation for memory-based learning in spectral chemometrics, 2022, R package version 2.2.1. URL <https://CRAN.R-project.org/package=resemble>. (Accessed 23 November 2022).
- [27] B. Minasny, A.B. McBratney, Regression rules as a tool for predicting soil properties from infrared reflectance spectroscopy, *Chemometr. Intell. Lab. Syst.* 94 (1) (2008) 72–79.
- [28] V.F. Diaz, P. Mishra, J.-M. Roger, W. Saeys, Domain invariant covariate selection (Di-CovSel) for selecting generalized features across domains, *Chemometr. Intell. Lab. Syst.* 222 (2022) 104499.
- [29] V. Thomsen, D. Schatzlein, D. Mercurio, Limits of detection in spectroscopy, *Spectroscopy* 18 (12) (2003) 112–114.
- [30] Y. Ogen, C. Neumann, S. Chabrilat, N. Goldshleger, E.B. Dor, Evaluating the detection limit of organic matter using point and imaging spectroscopy, *Geoderma* 321 (2018) 100–109.
- [31] P.B. Skou, T.A. Berg, S.D. Aunsbjerg, D. Thaysen, M.A. Rasmussen, F. van den Berg, Monitoring process water quality using near infrared spectroscopy and partial least squares regression with prediction uncertainty estimation, *Appl. Spectrosc.* 71 (3) (2017) 410–421.
- [32] B. Takoutsing, G.B.M. Heuvelink, J.J. Stoorvogel, K.D. Shepherd, E. Aynekulu, Accounting for analytical and proximal soil sensing errors in digital soil mapping, *Eur. J. Soil Sci.* 73 (2) (2022) e13226.
- [33] K. Vaysse, P. Lagacherie, Using quantile regression forest to estimate uncertainty of digital soil mapping products, *Geoderma* 291 (2017) 55–64.