Erasmus University Rotterdam
Erasmus School of Economics
Bachelor Thesis International Bachelor of Econometrics and Operations Research

Determining Variable Importance Across Quantiles in Quantile Regression Forests

Quinty Okhuijsen (563743)

Supervisor:	Dr. PC Schoonees
Second assessor:	D.J.W. (Daniël) Touw.
Date final version:	1st July 2024

Abstract

This study replicates the results of the original Quantile Regression Forests (QRF) paper (Meinshausen, 2006) and extends on it by examining variable importance across quantiles. Using datasets—Boston Housing, Ozone, Abalone, Big Mac, and Fuel—this replication compares QRF to other quantile regression methods, including Linear Quantile Regression with interaction terms (LQR), and without interaction terms (QQR), and tree-based regression models (TRC, TRM, TRP). The findings show QRF generally achieves the lowest average loss and is robust to noise. Differences in some datasets and models suggest potential updates or data processing differences. The extension introduces two methods to assess predictor importance at different quantiles using standard and conditional permutation schemes. Applied to NHANES and Boston Housing datasets, results indicate predictor importance varies across quantiles. For NHANES for instance, BMI more important at lower quantiles and Age at higher quantiles. This information can help tailor prediction models for specific quantiles in order to make their predictions more reliable.

1 Introduction

"If your head is in the oven and your feet are in the freezer, on average, you feel just fine." This common saying in statistics highlights how misleading averages can be, as they do not account for the spread of the data. Despite this, many prominent regression and classification techniques rely on these averages. To instead gain a deeper understanding of a variable's full distribution, one can consider quantiles. Particularly conditional quantiles are a great tool, as they show how the various quantiles of a response variable can differ given a change in a predictor's value. When making predictions, these conditional quantiles can then serve as confidence bounds that present the best and worst-case scenarios, which provides information on the reliability of the forecast. This information is crucial for many fields, such as finance or healthcare, where outcomes can change significantly based on dynamic conditions and where the stakes of making a dependable prediction are high.

Meinshausen (2006) presents a way of viewing these conditional quantiles by incorporating quantile regression (QR) into random forests (RF), known accordingly as quantile regression forests (QRF). QR (Koenker, 1978) builds on linear regression as it is able to predict the conditional quantiles of the response variable rather than just the mean. RF (Breiman, 2001), on the other hand, is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of their predictions for classification or the mean for regression. The authors of QRF realised that the nature of the RF regression model allows for determining the full conditional distribution of a response variable. Moreover, they found that QRF generally performs better than linear quantile regression (LQR) and other tree-based quantile regression methods.

In this paper, an attempt is made to replicate the results of the original QRF paper (Meinshausen, 2006). To do this, the same data sets as in the original paper are used, namely the *Boston Housing*, *Ozone*, *Abalone*, *BigMac*, and *Fuel* datasets, found from the *mlbench* (Leisch, 2024) (Blake, 1998) and *alr4* (Weisberg, 2014) packages from RStudio (R Core Team, 2023), as well as the UCI machine learning repository. It is noteworthy that some of these datasets have been updated since the original paper was published, which can cause the results of this paper to differ from the original paper. Moreover, as in the original paper, this paper compares the QRF method to several other quantile regression methods, starting with LQR and a method called QQR, which is like LQR but then with added interaction terms. Then, QRF is compared to different variations of regression trees, consisting of quantile regression trees with with piecewise constant (TRC), piecewise multiple linear (TRM), and piecewise second-degree polynomial form (TRP). Each of these models and datasets are also evaluated under the addition of noise, as in the original QRF paper. Moreover, prediction intervals are constructed for the datasets using QRF to replicate the original QRF paper.

Afterwards, as an extension on the ideas of the QRF paper, this paper explores the concept of variable importance at different quantiles of the response variable. Variable importance refers to the extent to which the inclusion of a predictor variable in the model improves the accuracy of the prediction. This paper investigates whether, for instance, a variable can be a better predictor for a response variable at a lower conditional quantile as opposed to a higher conditional quantile. As mentioned before, the conditional quantiles act as confidence bounds, and the smaller these

confidence bounds are, the more reliable a prediction is. When we know which predictors are most important at a certain quantile, we can tailor the model to include those predictors, leading to more accurate predictions. This variability of predictor importance at different quantiles is also relevant to know in fields that need to deal carefully with their predictions. In finance, this can involve risk management, where an understanding of what causes low losses or high profits can help to develop strategies that take these factors into account. Likewise, in healthcare, analysing variable importance at different quantiles can allow for more personalised medicine, as it can help identify for example which factors are critical at different stages, or quantiles, of a disease. For standard random forests, a well-known technique for determining variable importance in regression tasks is known as the mean decrease in accuracy (MDA), which is a permutation-based scheme. However, this paper prefers to use the term Mean Increase in Error (MIE) to describe the technique, as it is a more accurate description of what is being calculated. This paper proposes an adapted approach to MIE to include the variable importance at different quantiles, which is explained in more in detail in Sec. 4. Since variable importance techniques can introduce bias (Strobl et al., 2007), as is detailed further in Sec. 2, this paper proposes another method that incorporates the so-called conditional variable importance devised by Strobl et al. (2008) into calculating the importance of variables across quantiles. This concept of variable importance across quantiles has to my current knowledge not been explored in other literature. This paper therefore intends to fill this gap in knowledge. This leads to the main research question of this paper:

> *How does the importance of predictors vary across different quantiles in the distribution of a response variable when analysed with QRF?*

This research question comes with various different subquestions too. First of all, *what method can be used to effectively determine the importance of predictors at various quantiles?* Then, due to the potential relevance in the context of healthcare, a second subquestion arises: *"What are the implications of predictor importance variability for practical applications in healthcare?"* To explore this second subquestion, the well known medical *NHANES* dataset from the R package NHANES (Pruim, 2015) is used to explore the effect of several predictors on blood pressure at different quantiles. This is to test the hypothesis that for a given set of predictors, the importance of those predictors differs at various quantiles of blood pressure. For instance, Body Mass Index (BMI) might be a more important predictor for higher quantiles of blood pressure compared to lower quantiles of blood pressure, as is highlighted in a study by Linderman et al. (2018) that involved 1.7 million Chinese adults and found that the association of blood pressure and BMI was stronger in groups with higher BMI. Lastly, to investigate the generalisability of this approach, the third subquestion is in kind: *How robust are the findings across different datasets?* To do this, the MIE across quantiles approach is also implemented on the *Boston Housing* dataset.

In the replication part of this paper, the results indicate that the QRF model generally exhibits the lowest average loss across most quantiles compared to other methods. However, unlike in the original paper, the TRM model sometimes outperforms other models, though with more variability. Additionally, QRF is robust to noise, showing minimal change in loss across datasets, unlike LQR and QQR models, which display more variable results when noise is added.

However, this paper finds that some datasets and models do not produce the same output when an attempt is made to replicate them. It is suspected that this is due to updated datasets or software since the original paper was published, or a failure of the original paper to mention any data pre-processing steps.

In terms of determining variable importance across quantiles, this paper found some intriguing results. When evaluated on the *NHANES* and *Boston Housing* datasets, the standard variable importance and conditional variable importance showed different results. Interestingly, for the conditional variable importance of *NHANES*, the variable BMI seems to be the most important at lower quantiles, but is overtaken by the variable Age at higher quantiles, showing that the importance of variables can change across quantiles. However, the importance scores show an interesting trend across these quantiles. Both standard and conditional importance scores tend to peak around the middle quantiles and decrease for the more extreme values, giving an almost parabolic shape. It is theorised that this could be due to heteroscedasticity and the tails of a variable likely having more outliers, which could deflate the importance of predictors. This makes it more difficult to get a true sense of variable importance at extremer quantiles. This paper proposes several future research directions that could allow for a more truthful view of the importance of predictors at extremer quantiles. These involve using a weighted quantile loss function, considering extremal random forests, or using a localised permutation scheme. However, the proposed method is a good tool for when you are interested in a broader range of quantiles, not just the extreme values.

The github code of this paper is https://github.com/QuintyOk/BachelorThesisEconometrics.git. The rest of this proposal is structured as follows. In Sec. 2, a brief literature overview of variable importance measures is given, followed by Sec. 3, in which the datasets are discussed in greater detail. Then, Sec. 4 thoroughly explains the used methodologies, after which a discussion of the results follows in Sec. 6, and ending with a conclusion in Sec. 7.

2 Theory and Relevant Work

To my current knowledge, variable importance across quantiles in QRF has not been explored in other literature. Therefore, this study first investigates how similar models determine variable importance.

2.1 QR

In QR, for instance, variable importance is measured and interpreted similarly to that of standard linear regression. The QR model looks as follows

$$Q_Y(\alpha \mid \mathbf{X}) = \mathbf{X}^\top \boldsymbol{\beta}(\alpha) \tag{1}$$

where $Q_Y(\alpha \mid \mathbf{X})$ is the α-th conditional quantile of Y given $\mathbf{X}$, $\mathbf{X}$ is the matrix of predictors including an intercept term, and $\boldsymbol{\beta}(\alpha)$ is the vector of coefficients corresponding to the α-th quantile. We can obtain an estimate for the α-th quantile regression estimator, namely $\hat{\boldsymbol{\beta}}(\alpha)$, by solving the following optimisation problem

$$\hat{\boldsymbol{\beta}}(\alpha) = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} L_\alpha(y_i - \mathbf{x}_i^\top \beta) \tag{2}$$

where $\mathbf{x}_i$ is the i-th observed vector of predictors with their corresponding coefficients β, n is the number of observations in the dataset and

$$L_\alpha = \begin{cases} \alpha|y_i - q_i^\alpha| & \text{if } y_i > q_i^\alpha \\ (1-\alpha)|y_i - q_i^\alpha| & \text{if } y_i \leq q_i^\alpha \end{cases} \tag{3}$$

is a type of loss function called the 'pinball' or quantile loss function, which penalises the over predictions proportionally to the corresponding quantile α. Here, y_i is the i-th observed value of the response variable, and q_i^α is the predicted value of the quantile α in question for y_i.

The intercept coefficient $\beta_0(\alpha)$ is then interpreted as the α-th quantile of the response variable when all other predictors are zero. Furthermore, each slope coefficient $\beta_j(\alpha)$ represents the change in the α-th quantile of the response variable for a one-unit increase in the predictor X_j, holding all other predictors constant.

2.2 RF and QRF

The theory behind RF is discussed in detail in Sec. 4, but in simple words, it performs regression as follows. Starting of with a set of predictor variables and a certain response variable, decision trees are built using a technique called bootstrapping. Bootstrapping involves randomly selecting data points with replacement from the original data to create a new dataset of the same size as the original dataset, meaning that some data points may appear more than once in a dataset. Each of these new bootstrapped datasets is then used to grow a decision tree. When growing each tree, the model makes splits in the data in order to create branches, which are based on a random subset of predictor variables. The tree will continue to recursively split the data until a stopping criterion is reached, such as the maximum tree depth or minimum amount of data points in leaf node. After all decision trees are built, predictions can be made for the response variable based on the features of a new data point by dropping it down each individual decision tree and storing which leaf node it ends up in. Since the trees were built on different samples and considered different predictor variables at each split, the leaf nodes the new data point ends up in can be different for each decision tree. Then, the final prediction of the RF model is the average of all the individual decision tree predictions.

QRF builds on the RF model for regression by not only predicting the average outcome but also estimating the conditional distribution of the response variable. It achieves this by using the same collection of decision trees as in the RF model, but instead of just averaging the predictions, it keeps track of all the response variable values in the leaf nodes to estimate the quantiles of the response variable. When considering the prediction error of predictions made via QRF, a special loss function needs to be considered, which happens to be the same quantile loss function as for QR, presented in Eq. 3.

2.3 Variable Importance

The original RF paper (Breiman, 2001) was the one to introduce MDA, which still appears to be most prevalent variable importance technique. This technique starts of by using the out-of-bag samples (OOB) for each tree, which are the remaining data points that were not used in the bootstrapped sample to train that tree, to calculate the prediction error. In RF, this prediction error is usually calculated using Mean Square Error (MSE). Hence, let us denote this error by MSE_{OOB}. Then, for each predictor X_j for which we want to determine the importance, we permute the values of X_j in the OOB samples in order to break the relationship between it and the response variable. We then calculate the new prediction error using the permuted OOB values, and denote its error by $MSE_{OOB-Perm}$. The MDA for predictor X_j can then be calculated as the relative increase in error due to the permutation of X_j of all $k = 1, ..., K$ trees

$$\text{MDA}(X_j) = \frac{1}{K} \sum_{k=1}^{K} \left(\frac{\text{MSE}_{\text{OOB-Perm}}(X_j) - \text{MSE}_{\text{OOB}}}{\text{MSE}_{\text{OOB}}} \right) \tag{4}$$

It is important to note that the interpretation of these variable importance scores are relative. The importance score of one variable can for instance only be evaluated in the context of the importance scores of other variables. This also means that the importance of a variable at a certain quantile is also relative to the importance of that variable at other quantiles. To therefore interpret the results, this paper looks at multiple quantiles at once in order to see whether the importance at these quantiles changes.

There are some issues regarding bias and interpretability with this technique that are crucial to consider. According to Strobl et al. (2007), particularly when predictors are correlated with each other, bias in the variable importance measure can arise, leading to unreliable results. This bias has two main causes when the MDA technique is used.The first issue occurs during the construction of decision trees in RF. When predictors are correlated with other important predictors, they are more likely to be used for splitting nodes in the decision tree. This correlation can make them appear more informative than they actually are. As a result, these predictors might receive higher importance scores than they truly deserve, despite them not being as crucial in predicting the response variable. The second cause of bias occurs during the permutation process of MDA. The standard permutation importance calculation does not take the correlation structure of predictors into account either. When a predictor is permuted, and before the permutation it was correlated with other predictors, the correlation with this predictor is also broken during the permutation. This leads to a higher increase in error as the predictors no longer give coherent information together, leading to an overestimation of the importance of the predictor of interest.

To mitigate these biases, another paper by Strobl et al. (2008) introduces a framework called conditional variable importance. This procedure permutes the predictor variable whilst preserving the correlation with other predictors. In simple words, it uses the splits created by the fitted RF model to condition the permutations, better reflecting the true impact of each predictor variable. This technique is discussed more in detail in Sec. 4, as it is used for my extension.

3 Data

3.1 Replication

As mentioned in Sec. 1, the datasets used in the replication part are obtained from the following R software packages. *mlbench* (Leisch, 2024), *alr4* (Weisberg, 2014) and the UCI machine learning repository. These datasets include the following: the *Boston Housing* data set, which contains information on housing in Boston; the *Ozone* data set, which involves different variables related to air quality, with missing values removed as in the original paper; the *Abalone* data set, which includes measurements of a type of snail called abalones, limited to 500 randomly chosen observations as in the original paper; the *Big Mac* data set, which contains economic indicators related to the price of a Big Mac in various cities around the world; and the *Fuel* data set, which contains data about average gas consumption for all states in the U.S. The original paper does not mention the response variables for all datasets, but in this paper they are assumed to be *medv* for the Boston Housing dataset, *V4* for the Ozone dataset, *Rings* for the Abalone dataset, and *BigMac* for the BigMac dataset. The response variable for the Fuel dataset is said to be the average gas-mileage in the original paper, constructed by the ratio of total gallons of gasoline sold and the approximate number of miles driven. This paper assumes this is done by dividing the variable *FuelC*, which is the Gasoline sold for road use (1000s of gal.), divided by the variable *Miles*, which represents the miles of Federal-aid highway miles in the state. This new response variable was named *GasM*, and *Miles* and *FuelC* were removed from the list of predictor variables. Another pre-processing step that this paper takes is to remove the first three variables of the Ozone dataset, namely *V1, V2* and *V3*. These variables represent the month, day of month, and day of week respectively, and were causing issues in numerous models due to mutli-collinearity, in particular in the QQR and tree-based methods. This paper is unsure whether the original paper also took these pre-processing steps. A summary of all observations and predictor variables is shown in Tab. 1.

Table 1: Summary of datasets used in the study, including the number of observations n and number of predictor variables p.

Property	**Boston Housing**	**Ozone**	**Abalone**	**Big Mac**	**Fuel**
n	506	203	500	69	51
p	13	9	8	9	5

3.2 Extension

For the extension part of this paper, the National Health and Nutrition Examination Survey (NHANES) dataset is used, which is a collection of extensive health information from a nationally representative sample. This dataset was part of the a package in R called *NHANES* (Pruim, 2015). A subset of variables is taking from this dataset, namely our response variable - average blood pressure (BPSysAve), and some well-known predictors of blood pressure - BMI, Age, Total Cholesterol, Alcohol per Year, and the Number of Physical Active Days. All of the missing values were removed, leaving us with $n = 4161$ observations and $p = 5$ predictor variables. Then

as an additional dataset to test the extension on the same Boston Housing dataset as in the replication part is used. To see how the variables in the datasets are correlated with each other, two correlation matrices are computed, as shown in Fig. 1 and 2. In the *NHANES* dataset, none of the variables are highly correlated with each other. In the *Boston Housing* dataset, on the other hand, we can see that some variables are indeed highly correlated with each other. The conditional variable importance permutation scheme that this paper applies hopes to reduce the effect of these correlations in the importance scoring process.

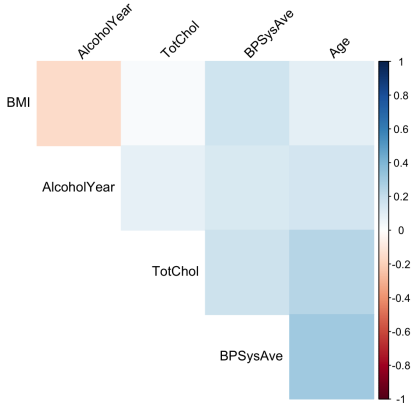

Figure 1: The correlation matrix of the variables in the NHANES dataset.

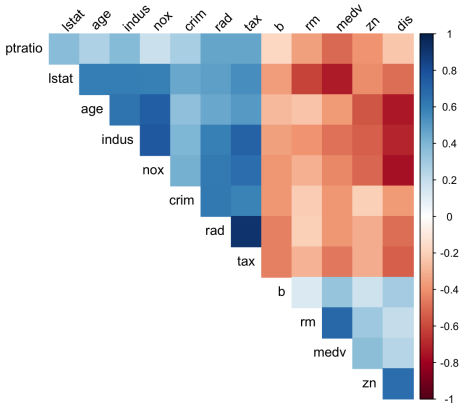

Figure 2: The correlation matrix of the variables in the Boston Housing dataset.

4 Methodology

4.1 Replication

As outlined in Sec. 1, this paper will compare QRF to other quantile regression methods. In this section, these methods are described in greater detail. As in the original QRF paper, all the models are evaluated with the following quantile loss function

$$\mathbf{L}_\alpha = \frac{1}{n}\sum_{i=1}^{n} L_\alpha \tag{5}$$

which takes the average of the individual quantile loss functions L_α of all n observations in the dataset, where L_α is

$$L_\alpha = \begin{cases} \alpha|y_i - q_i^\alpha| & \text{if } y_i > q_i^\alpha \\ (1-\alpha)|y_i - q_i^\alpha| & \text{if } y_i \leq q_i^\alpha \end{cases} \tag{6}$$

where y_i is the observed value, q_i^α is the quantile estimate at level α, and α is the quantile level, and which calculates the weighted absolute differences between the observed values and the estimated quantiles, where . This loss function is more robust when dealing with quantile estimations as it appropriately weights overestimation and underestimation.

Furthermore, all of the models are evaluated within a 5-fold cross validation loop. This means that the data is divided into five equal parts, known as folds. In each iteration of the cross-validation loop, the model of interest is trained using four of these folds, while the remaining

fold is used to evaluate the model. This process is repeated five times, each time with a different fold as the test set.

To evaluate how QRF performs compared to the other methods, 95% bootstrap confidence bounds are constructed for those methods. The original QRF paper was not clear in the methodology of creating those confidence bounds, but this paper interprets it as follows. For each quantile, the average losses of all fold are resampled 1000 times, after which the mean loss is computed again. Then, the 2.5th and 97.5th percentiles of these bootstrap means are computed to obtain the confidence bound.

4.1.1 RF

As mentioned before, QRF is an extension of the RF model, which was explained briefly in Sec. 2. It grows trees in the same manner as RF, but instead of only keeping the mean of the observations in each leaf, QRF keeps the full set of observations in order to estimate the conditional distribution of Y given $\mathbf{X} = \mathbf{x}$. Using the same notation as in Meinshausen (2006) and Breiman (2001), let $\boldsymbol{\theta}$ represent the random parameter vector that determines tree growth, such as the m out of p predictors to consider for splits. The tree corresponding to $\boldsymbol{\theta}$ is denoted by $T(\boldsymbol{\theta})$. The predictor space $B \subseteq \mathbb{R}^p$ represents the range of possible values for the predictor variables $\mathbf{X}$. Each leaf l of a tree $T(\boldsymbol{\theta})$ corresponds to a rectangular region $R_l \subseteq B$. For any $\mathbf{x} \in B$, there is a unique leaf $l(\mathbf{x}, \boldsymbol{\theta})$ such that $\mathbf{x} \in R_l$.

A tree's prediction for an input $\mathbf{X} = \mathbf{x}$ is the average of the values in the leaf that ends up containing the input $\mathbf{x}$. The weight vector $w_i(\mathbf{x}, \boldsymbol{\theta})$ is defined as a positive constant if observation $\mathbf{x}_i$ from the original dataset is in the same leaf as $\mathbf{x}$, and is otherwise set to zero. If $\mathbf{x}_i$ is in the same leaf as $\mathbf{x}$, the weight of observation i becomes one divided by the amount of times an observation $\mathbf{x}_j$ is in the same leaf node as $\mathbf{x}$, as shown in

$$w_i(\mathbf{x}, \boldsymbol{\theta}) = \frac{1\{\mathbf{x}_i \in R_{l(\mathbf{x},\boldsymbol{\theta})}\}}{\#\{j : \mathbf{x}_j \in R_{l(\mathbf{x},\boldsymbol{\theta})}\}} \tag{7}$$

The prediction for input $\mathbf{x}$ using a single tree is then

$$\hat{\mu}(\mathbf{x}) = \sum_{i=1}^{n} w_i(\mathbf{x}, \boldsymbol{\theta}) Y_i \tag{8}$$

RF models aim to approximate the conditional mean $E(Y \mid \mathbf{X} = \mathbf{x})$ by averaging the predictions of k trees, where each tree is constructed with an independent parameter vector $\boldsymbol{\theta}_t$. The average weight $w_i(\mathbf{x})$ across all trees is

$$w_i(\mathbf{x}) = \frac{1}{k} \sum_{t=1}^{k} w_i(\mathbf{x}, \boldsymbol{\theta}_t) \tag{9}$$

The overall prediction for the random forest is then given by

$$\hat{\mu}(\mathbf{x}) = \sum_{i=1}^{n} w_i(\mathbf{x}) Y_i \tag{10}$$

This approach gives weights to the observations based on how similar the conditional distribution

of Y given $\mathbf{X} = \mathbf{x}_i$ is to the conditional distribution of Y given $\mathbf{X} = \mathbf{x}$.

4.1.2 QRF

QRF extends this idea to estimate the full conditional distribution of Y given $\mathbf{X} = \mathbf{x}$. The conditional distribution function $F(y \mid \mathbf{X} = \mathbf{x})$ is defined as

$$F(y \mid \mathbf{X} = \mathbf{x}) = P(Y \leq y \mid \mathbf{X} = \mathbf{x}) = E(1\{Y \leq y\} \mid \mathbf{X} = \mathbf{x}) \tag{11}$$

QRF uses the same weights $w_i(\mathbf{x})$ as random forests, but applies them to indicator functions instead of the response variable directly. Thus, the estimate for the conditional distribution function is

$$\hat{F}(y \mid \mathbf{X} = \mathbf{x}) = \sum_{i=1}^{n} w_i(\mathbf{x}) 1\{Y_i \leq y\} \tag{12}$$

The procedure for computing $\hat{F}(y \mid \mathbf{X} = \mathbf{x})$ can be summarised as follows: First, grow k trees $T(\boldsymbol{\theta}_t)$, $t = 1, \ldots, k$, recording all observations in each leaf. For a given $\mathbf{X} = \mathbf{x}$, drop $\mathbf{x}$ down all trees to compute $w_i(\mathbf{x}, \boldsymbol{\theta}_t)$, and average these weights to get $w_i(\mathbf{x})$

$$w_i(\mathbf{x}) = \frac{1}{k} \sum_{t=1}^{k} w_i(\mathbf{x}, \boldsymbol{\theta}_t) \tag{13}$$

Then, compute the estimated distribution function for all y

$$\hat{F}(y \mid \mathbf{X} = \mathbf{x}) = \sum_{i=1}^{n} w_i(\mathbf{x}) 1\{Y_i \leq y\} \tag{14}$$

Finally, conditional quantiles $\hat{Q}_\alpha(\mathbf{x})$ are derived from the estimated distribution $\hat{F}(y \mid \mathbf{X} = \mathbf{x})$.

As in the original QRF paper, this paper uses the R package *quantregForest* to implement QRF (N. Meinshausen, 2017). The same parameters as in the original paper are used too, which includes building $k = 1000$ trees, setting the number of variables to use for splitting a node, or *mtry*, to one-third of all variables, and restricting each node to have more than 10 observations.

4.1.3 LQR and QQR

The theory of linear quantile regression is discussed in Sec. 2, and its main goal is to optimise the problem in Equation 2. The author of the original QRF paper does not elaborate in detail on how this LQR model is built, but this paper uses the R package *quantreg* R. Koenker (2024).

For QQR, the original papers instructs to start from the LQR model and then add interaction terms between variables by forward selection until the 5-fold cross validation error attains a minimum. This paper interprets this as iteratively adding a certain interaction term to the model when the quantile loss in Eq. 5 of the model with that interaction term is lower than the quantile loss of the model without that interaction term, as well as lower than adding any other interaction term, until this quantile loss function has reached a minimum.

In summary, the LQR and QQR are generated as follows for each fold and quantile α in the cross validation:

Algorithm 1 LQR and QQR for each fold

Initialize:
$\mathcal{M}_0 \leftarrow$ LQR model
$E_0 \leftarrow \frac{1}{n}\sum_{i=1}^{n} \text{Loss}_{LQR}$
Define interaction terms:
Interaction terms $\leftarrow \{x_h \cdot x_j \mid h \neq j\}$
Iteratively improve the model:
Improved $\leftarrow$ True
while Improved **do**
  Improved $\leftarrow$ False
  for all $x_h \cdot x_j$ **where** $h \neq j$ **do**
    $\mathcal{M}_{\text{current}} \leftarrow \mathcal{M}_{\text{best}} + x_h \cdot x_j$
    for all $\alpha \in$ quantiles **do**
      Calculate the loss for $\mathcal{M}_{\text{current}}$ for quantile α using Eq. 5:
      $\hat{\beta}_{\text{current}}(\alpha) \leftarrow \arg\min_{\beta} \sum_{i=1}^{n} L_\alpha(y_i - \mathbf{x}_i^\top \beta)$
      $\hat{y}_i(\alpha) \leftarrow \mathbf{x}_i^\top \hat{\beta}_{\text{current}}(\alpha)$
      $\text{Loss}_{\text{current}} \leftarrow \frac{1}{n}\sum_{i=1}^{n} L_\alpha(y_i - \hat{y}_i(\alpha))$
      if $\text{Loss}_{\text{current}} < E_{\text{best}}$ **then**
        $\mathcal{M}_{\text{best}} \leftarrow \mathcal{M}_{\text{current}}$
        $E_{\text{best}} \leftarrow \text{Loss}_{\text{current}}$
        Improved $\leftarrow$ True
      end if
    end for
  end for
end while
Final model:
$\mathcal{M}_{\text{final}} \leftarrow \mathcal{M}_{\text{best}}$

4.1.4 TRC, TRM, and TRP

As in the original QRF paper, the tree-based methods TRC, TRM and TRP are also compared with QRF. In TRC, each leaf node of the tree predicts a constant value for each quantile of the response variable. TRM, on the other hand, has each leaf node fit a linear model to predict the quantile of the response variable. TRP takes this one step further as it has each leaf node fit a second-degree polynomial model to predict the quantile of the response variable. Hence, TRC is the simplest and most interpretable, whilst the other two allow for more flexibility and complexity. To replicate these results, we use the same software package as in the original QRF paper called GUIDE, which is now currently available from **https://pages.stat.wisc.edu/ loh/guide.html** - a different page than the original QRF paper. We use this package for all three tree-based methods, and use the default settings for each method, as in the original QRF paper.

4.1.5 Adding additional noise to models

The original QRF paper tests the models under the addition of noise as well. This paper replicates this too. The methodology used for this is as follows. For each dataset, every predictor variable is taken and randomly permuted. Then all permuted predictor variables are added as additional variables to the model, meaning there are now double the amount of predictor variables. Then, all of the models are retrained on the new set of predictor variables, and the average quantile loss is recalculated.

4.1.6 Prediction Intervals for QRF

This paper also replicates the 95% prediction intervals created by the original QRF paper. For each data point in the test fold, the conditional quantiles are estimated with QRF using 5-fold cross-validation. Then, for better visualisation, the observations are ordered according to the length of their prediction interval.

4.2 Extension

In this sub section, the methodology of this paper's extension is presented. Firstly, this paper illustrates a proposed standard permutation technique based on MDA for measuring variable importance across quantiles in QRF is illustrated. This method is referred to as the standard variable importance across quantiles. Next, this paper proposes an augmented version of that technique, referred to as the conditional variable importance across quantiles, which aims to produce more unbiased and accurate results. These methods are based on the following proposed equation to measure the variable importance, namely MIE

$$\text{MIE}_{j,\alpha} = \frac{\mathbf{L}_{\alpha,\text{permuted}} - \mathbf{L}_{\alpha}}{\mathbf{L}_{\alpha}} \tag{15}$$

where $\mathbf{L}_{\alpha}$ is the quantile loss function in Eq. 5 and how $\mathbf{L}_{\alpha,\text{permuted}}$ is computed depends on the type of method used, which are explained in the next two subsections.

As in the replication part if this paper, the importance values, or MIE, are calculated within a 5-fold cross validation loop. As explained before, this means the data is split into 5 equal folds and in each iteration one fold is used as a test set, whilst the remaining folds are used as the training set. When training the QRF models, an *mtree* value of 2, a minimum nodesize of 10, and a total of 1000 trees are used.

4.2.1 Standard Variable Importance across Quantiles

This proposed method is inspired by a common way to determine the importance of a predictor variable in random forests — MDA, which works as follows. First, the QRF model, as described in sub Sec. 4.1.1, is trained on the in-bag sample of the dataset. Then, the baseline accuracy, or prediction error, of the QRF model is calculated by using each OOB sample as a test set for the trees that did not use that sample during training. Next, the OOB values of the predictor variable for which we wish to know the importance are permuted to break its relationship with the other variables, and the prediction error is recalculated. The importance of that variable

can now be computed as the relative increase in prediction error of the original and permuted model. To adapt this for QRF, we use the quantile loss function shown in Eq. 5 to determine the accuracy per quantile instead of the MSE. Furthermore, instead of using the OOB sample as a test set, we use a 5-fold cross validation mechanism to make the predictions.

The steps to calculate the standard variable importance across quantiles in QRF are outlined as follows:

Algorithm 2 Standard Variable Importance using MIE in QRF with 5-Fold Cross Validation

Step 1: Initialize
Split the dataset into 5 folds.
Step 2: Cross-validation loop
for $k = 1$ to 5 **do**
  Use fold k as the test fold and the remaining folds as the training folds.
  Step 3: Train QRF model using the training folds
  Step 4: Predict the conditional quantiles for observations in the test fold
  Step 5: Calculate prediction error
  for all quantiles α **do**
    Calculate the prediction error using the quantile loss function $\mathbf{L}_\alpha$ (see Eq. 5).
  end for
  Step 6: Permute variable j and calculate loss
  Permute the values of the variable of interest j in the test fold data.
  for all quantiles α **do**
    Predict quantiles using the permuted data and recalculate the quantile loss $\mathbf{L}_{\alpha,\text{permuted}}$ (see Eq. 5).
  end for
  Step 7: Calculate variable importance
  for all quantiles α **do**
    Calculate the importance of variable j (see Eq. 15).
  end for
end for

The larger the relative increase in error or quantile loss, the more important variable j is at predicting the response variable at quantile α. Hence, this approach adapts the traditional MDA method to work with quantile regression forests, allowing for the assessment of variable importance across different quantiles of the response distribution.

4.2.2 Conditional Variable Importance across Quantiles

To address the limitations of standard variable importance measures in the presence of correlated predictors as mentioned in Sec. 2, we employ a conditional permutation scheme to compute the importance of each variable while accounting for the correlations with other predictors. This methodology is inspired by Strobl et al. (2008) and works as follows.

As mentioned before, the variable importance across quantiles, or MIE, will be calculated within a 5-fold cross validation loop. After training the QRF model on the training folds, the

model predicts the conditional quantiles for the observations in the test fold for all desired quantiles. The prediction error for each quantile is then again calculated using the quantile loss function as defined in Eq. 5. Next, to assess the importance for a certain predictor variable j, the most important conditioning variables for that variable j are identified. Conditioning variables are those that have the highest correlation with the predictor variable j. Then, a grid of data points is constructed by extracting the cut points from the trained QRF model. These cut points are specific values at which the data is split by the conditioning variables during the tree-building process. The grid then effectively holds the partitions of the data created by the cut points from the splits of the conditioning variables. Following this, the values of variable j are permuted within each partition of the grid. Hence, the values are randomly shuffled in each separate partition to break the association with the response variable. The quantile loss is then recalculated with the permuted data using the same quantile loss function in Eq. 5. As in the standard variable importance in the above sub section, the conditional variable importance can also be calculated with the MIE in Eq. 15. In this way, the effect of correlated predictors is reduced, as only the impact of each variable within context-specific data partitions is assessed according to Strobl et al. (2008).

Algorithm 3 Conditional Variable Importance using MIE in QRF with 5-Fold Cross Validation

Step 1: Initialize
Split the dataset into 5 folds.
Step 2: Cross-validation loop
for $k = 1$ to 5 **do**
  Use fold k as the test fold and the remaining folds as the training folds.
  Step 3: Train QRF model using the training folds
  Step 4: Predict the conditional quantiles for observations in the test fold
  Step 5: Calculate prediction error
  for all quantiles α **do**
    Calculate the prediction error using the quantile loss function $L_\alpha(\mathbf{y}, \mathbf{q})$ (see Eq. 5).
  end for
  Step 6: Identify conditioning variables for each predictor j
  Step 7: Construct grid by extracting cut points from the trained QRF model.
  Step 8: Permute variable j and calculate loss
  for all quantiles α **do**
    Permute the values of variable j within each partition of the grid and recalculate the quantile loss $L_{\alpha,\text{permuted}}(\mathbf{y}, \mathbf{q})$ (see Eq. 5).
  end for
  Step 9: Calculate variable importance
  for all quantiles α **do**
    Calculate the importance of variable j (see Eq. 15).
  end for
end for

Again, the larger the difference in quantile loss, or relative increase in prediction error, the

more important variable j is at predicting the response variable at quantile α. Therefore, this method adapts the standard MDA method to work with QRF, allowing for the assessment of variable importance across different quantiles of the response distribution, while accounting for the conditional dependencies among predictor variables.

To determine the conditioning variables for a target variable j, the importance scores of all predictors excluding the response variable are calculated for each j using a random forest model, where j is now the response variable. Then, the two most important predictors are selected as conditioning variables for that target predictor j. This paper choose two conditioning variables to reduce computation time, but more conditioning variables can be selected if need be. Furthermore in the permutation step of the variable importance measure, this paper randomly permutes the variables 100 times. For each permutation, it calculates the quantile loss and then aggregates these results to achieve a more robust outcome. A summary of the procedure is given below:

5 Results

5.1 Replication

First, the average loss for the several quantiles and methods explained in Sec. 4 are replicated which. The results of this replication are shown in Fig. 3, which corresponds to Fig. 1 in the original paper. Then, the replication results for the average loss under additional noise are displayed in Fig. 9 in the appendix, whic correspond to Fig. 2 in the original paper. The vertical bars indicate the 95% bootstrap confidence intervals. According to the original QRF paper, if these intervals do not cross the horizontal striped line, which represents the average loss of QRF, the difference in average loss is statistically significant. Note that these graphs are do not have the same scale as the original paper, which can make the results appear to be different. First, Fig. 3 is discussed. In the Fuel dataset, QRF generally performs well, with its average loss being lower than or comparable to other methods across most quantiles. For the 0.5 quantile, however, the TRM method shows a significantly higher loss, as its confidence interval does not overlap with the QRF benchmark. In the Boston Housing dataset, QRF consistently shows the lowest average loss, particularly at the extreme quantiles (0.005, 0.025, 0.975, 0.995), where TRM and other methods show higher losses with non-overlapping confidence intervals, indicating statistical significance. In the Ozone dataset, QRF also outperforms other methods across most quantiles, with only minor exceptions at the extreme quantiles where TRC and TRP sometimes show comparable performance. Again, TRM is has a significantly lower average loss at the 0.5th quantile. For the Abalone dataset, QRF maintains the lowest average loss across nearly all quantiles, with the other methods showing statistically significant higher losses at several points. In the BigMac dataset, QRF generally performs better, though TRM shows better performance at the higher quantiles (0.95, 0.975, 0.995), where its lower average loss and non-overlapping confidence intervals suggest statistical significance. Most of the results are similar to the original QRF paper, with the exception of the *Fuel* dataset, which shows completely different loss values, as well as the tree-based methods TRC, TRM and TRP. Theories on why these differences occur are discussed in Sec. 6.

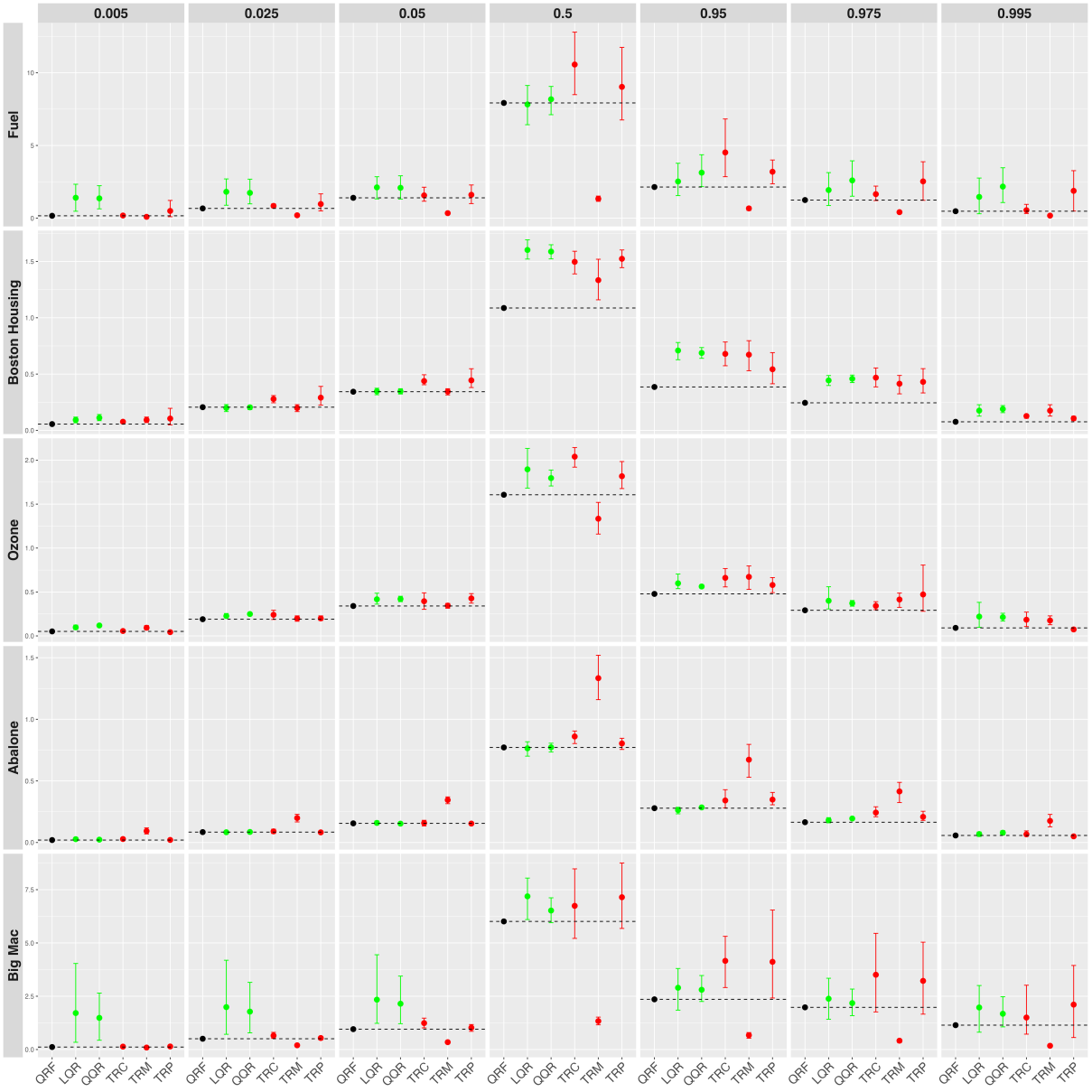

Figure 3: For each dataset, method and quantile, the average loss is depicted. From left to right, the dots present QRF, LQR, QQR, TRC, TRM and TRP respectively in each graph. The columns represent the seven quantiles, whilst the rows illustrate the five different datasets. The vertical bars represent the 95% bootstrap confidence bounds.

When noise is added to the model, which is shown in Fig. 9 in the appendix, the loss of QRF does not change much in all datasets, showing that it is robust to noise. LQR and QQR, on the other hand, show more varying results compared to their counterparts without noise. Sometimes their loss is lower, and sometimes their loss is higher than the models without noise. The average loss of the tree-based models also do not differ significantly from their non-noise equivalents.

In Fig. 4, the prediction intervals explained in sub Sec. 4.1.6 are illustrated for the Boston Housing dataset, which correspond to Fig. 3 in the original QRF paper. These results are similar to the original paper, such as the length of the prediction intervals varying greatly, meaning some observations can be predicted with more accuracy than others. The ordered prediction intervals

for the other datasets are shown in Fig. 10 in the appendix, which again look similar to the original QRF paper's Fig. 4, except for the Fuel dataset. The theory for that is again discussed in Sec. 6.

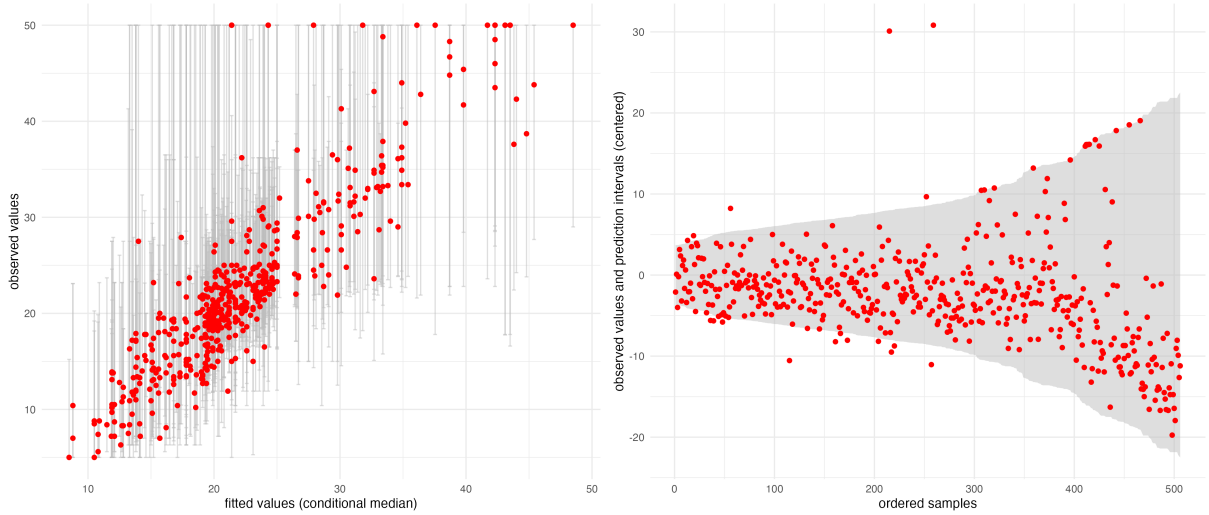

Figure 4: On the right, the observed values of the Boston Housing dataset are plotted in red against the predicted median values. Prediction intervals are displayed for each $i = 1, \ldots, n$ as transparent grey bars, with vertical black lines marking the lower and upper bounds. In the left graph, the observations $i = 1, \ldots, n$ are sorted by the length of their respective prediction intervals and the average of the upper and lower bounds of the prediction intervals is subtracted from all observations and prediction intervals.

5.2 Extension

Figure 5 shows the standard variable importance calculation for the sampled *NHANES* dataset, as described in sub Sec. 4.2.1. The way this graph is interpreted is as follows. Since the MIE calculates the relative increase in prediction error, shown in Eq. 15, we can interpret a predictor's importance score of zero as the prediction error not changing when that predictor is permuted, and hence that predictor not creating a better predictive accuracy and thus not being important. A positive MIE indicates that the prediction error increases when the predictor is permuted, indicating that when that predictor is included in the model, it has allows for a better predictve accuracy. Vice versa, when the MIE is positive, the inclusion of the predictor is actually worse for the model. Since the MIE is a relative measure indicating how much the error has increased relative to the initial error, it doesn't have units or a specific scale, making it a pure number. Notably, Age emerges as the most important predictor throughout the quantiles, reaching its peak around the median quantile before decreasing again towards the extremes. BMI follows a similar pattern with slightly lower importance values. Total Cholesterol (TotChol) shows moderate importance compared to Age and BMI, and is followed by Alcohol Consumption per Year and Physical Activity (PhysActive). Note that the importance of these variables are all relative to one another, meaning Age and BMI are more important predictors relative to the others. The apparent parabola shape of the importance values is interesting, and theories of why this arises are also discussed in Sec. 6.

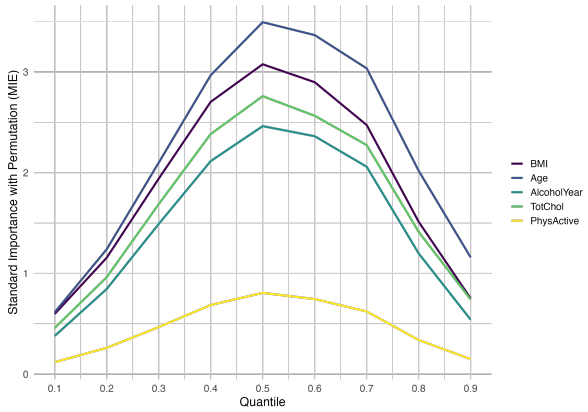

Figure 5: The standard variable importance by permutation (MIE) for the NHANES dataset across quantiles

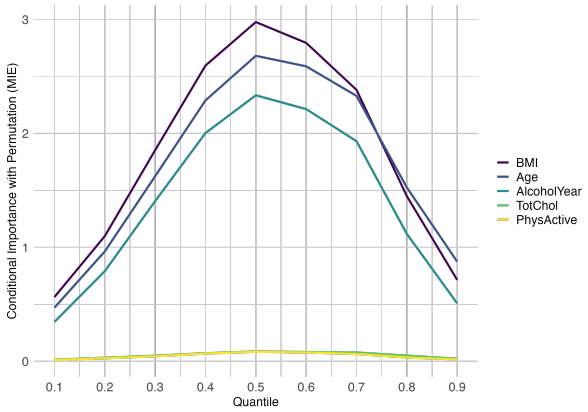

Figure 6: The conditional variable importance by permutation (MIE) for the NHANES dataset across quantiles

In Fig. 6 the conditional variable importance of the *NHANES* dataset is calculated, which was described in sub Sec. 4.2.2. It is immediately evident that this graphs looks different to the standard variable importance graph on the left. Particularly, the importance scores are a lot lower compared to the standard variable importance values, suggesting that the standard values were indeed inflated due to the correlation between the predictors. Again, BMI and Age are the most important predictors relative to the other predictors, but interestingly BMI appears to be a the most important predictor for blood pressure at lower quantiles, whilst Age is the most important at higher quantiles. This shows that variable importance can indeed differ across quantiles. It is also noteworthy that the variable Alcohol per Year now seems to be much more important than Total Cholesterol, and that Total Cholesterol and Physically Active Days are now significantly less important than all of the other predictors. This suggests they were perhaps highly correlated with the other predictors. Again, the graph of the conditional variable importance values follow this parabola shape, where the variables appear to be less important at extreme quantiles. Theories for this are discussed in Sec. 6 as well.

To test the outcome of the proposed variable importance across quantiles procedure on a different dataset, the Boston Housing dataset is used. The standard variable importance is for this dataset is shown in Fig. 7, where it is clear that the variables *rm* and *lstat* appear most important. When applying the conditional permutation scheme shown in Fig. 8, we see that the importance value dynamic shifts completely, with variable *crim* proving to be more important. The importance values are almost ten times lower as well, meaning the correlations were inflating the importance scores quite a bit. When looking at the correlations matrix of the variables in Fig. 2, we see that there are quite a few significant correlations between several variables, which can explain this more drastic change as compared to the *NHANES* dataset. The shape the variable *crim*, which stands for per capita crime rate, is also interesting. It peaks around the 0.5 quantile, dips down to the 0.8, but then rises up again after. This dip implies that for homes in the higher value range closer to the 0.8 quantile, crime rate becomes less critical as a predictor. One possible explanation is that high-value neighborhoods might have better security measures in place, making the crime rate less variable and hence less significant in predicting home values. However, the rise in importance of *crim* after the 0.9 quantile indicates that at the very high

end of the housing market, the crime rate once again becomes a significant predictor. This could be due to the increased attractiveness of high-value homes to potential criminals.

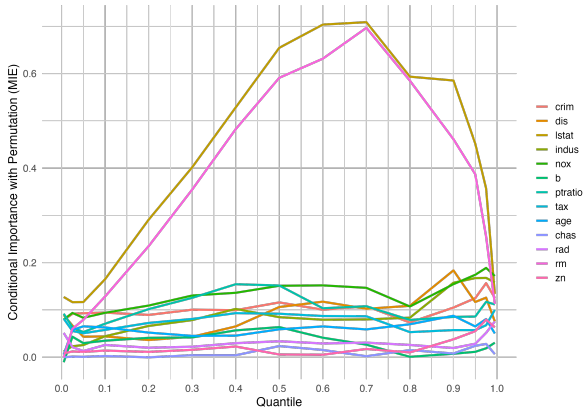

Figure 7: The standard variable importance by permutation (MIE) for the whole Boston Housing dataset.

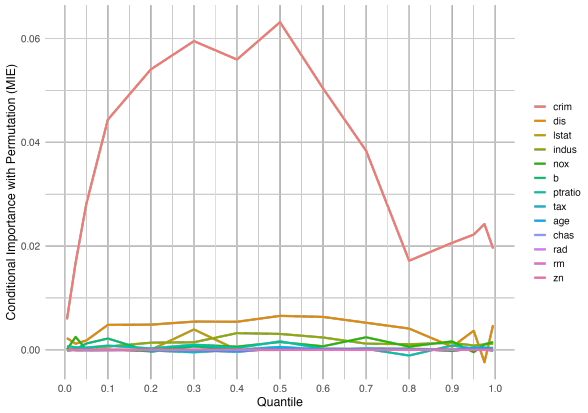

Figure 8: The conditional variable importance by permutation (MIE) for the whole Boston Housing dataset.

6 Discussion

6.1 Replication

Starting with our replication results, some significant differences are observed compared to the original QRF paper, despite using the same methods. The QRF results for all datasets, except the Fuel dataset, match those in the original QRF paper, which suggests that there may be an issue with the Fuel dataset. The original QRF paper was published in 2006, and since then the package *alr3* that hosts the Fuel dataset is no longer available and replaced by the package *alr4*. It is unsure whether this means that the dataset was updated, but it could provide a reason for why the replication results of this paper are so different to the original paper. Another reason why the Fuel dataset has such different results could be to do with the response variable. To my understanding and as described in Sec. 3, the response variable is calculated by dividing the FuelC by Miles in order to obtain the average gas-mileage. Perhaps the response variable in the original paper was calculated in a different way but not clearly described, leading to such different results.

The replicated LQR and QQR methods have values that are close to the original QRF paper, but not completely. This could be due to a difference in implementation. In the original paper, it is not described how LQR and QQR are implemented. This paper used the R package *quantreg* to perform both. This difference in implementation could cause the results to differ. Furthermore, for QQR, the original paper is somewhat vague on the method of adding interaction terms. It could be that they use a slightly different methodology of adding the interaction terms, or evaluate the cross validation error with another loss function than the quantile loss function, which could explain the differences.

For the replicated tree-based methods, the only method that appears to produce the same results as in the original paper is the TRC method. The fact that TRM shows loss values so much lower than in the original QRF paper, even lower than the QRF method sometimes, is

intriguing, as just like in the original paper, the default settings were used. It appears as though the software package *GUIDE* that is used to run these regression trees is regularly updated, which could perhaps explain the difference. The latest update even appears to be April 27th 2024. Apart from that reason, this author has no other arguments as to where the difference could come from, as *GUIDE* is a very straight-forward program to use.

6.2 Extension

This part intends to discuss the results of the extension of this paper, which was to determine the variable importance across quantiles for the NHANES dataset. In particular, the reason for the parabolic shape in Fig. 5 and Fig. 6, which suggests that the predictor variables are less important at extreme quantiles. This result seems less intuitive, but let us discuss theories as to why this may happen and how we should then interpret the results.

One possible explanation is the presence of heteroscedasticity, where in the context of QRF at extreme quantiles, the spread of the data points can be larger. This increased variance makes it more challenging for the model to capture the relationship between predictors and the response variable. This can result in the predictors appearing less important at these extreme quantiles because the model's predictions are more influenced by this noise. Furthermore, the tails of the distribution representing the extreme quantiles are often more likely to contain outliers. This can blur the true relationship between the predictors and the response variable and thereby decrease the importance of the predictors.

The effects of predictors at extreme quantiles are being underestimated with this approach, which can have negative effects in fields such as healthcare or finance, where understanding the behavior of predictors at extreme quantiles is crucial. This leads to a few future research directions that can potentially help with the true interpretation of the importance scores at extreme quantiles. First of all, a weighted quantile loss function can make the model more sensitive to extreme values by giving more weight to observations in the tails of the distribution. This may help reduce the parabolic shape of the importance values across quantiles. Extremal random forests are another interesting future research direction to consider. This type of random forest focuses specifically on capturing the behavior of extreme values, which can help interpret the importance values of those extreme values better. Moreover, perhaps a localised permutation importance measure could improve interpretation of extreme quantiles by restricting permutations to data subsets at specific quantiles. Lastly, when it comes to selecting the conditioning variables for the conditional variable importance measure, instead of using standard random forests, an interesting area to look into are conditional inference forests, as they help reduce bias in variable selection.

7 Conclusion

This study aims to replicate the results of the original QRF paper (Meinshausen, 2006) and extend the idea of QRF by examining variable importance across quantiles. This replication involves using the same datasets - Boston Housing, Ozone, Abalone, BigMac, and Fuel - to see how QRF compares to other quantile regressions methods in terms of predicting conditional

quantiles. These comparison methods include LQR, QQR, TRC, TRM and TRP. The results indicate that most of the time QRF achieves the lowest average loss, which is consistent with the results of the original QRF paper. Additionally, QRF is robust to noise, showing minimal change in loss across datasets, unlike LQR and QQR models, which display more variable results when noise is added. However, for the Fuel dataset, the results of this paper are very different from the original paper, which could be due to updates in the dataset or differences in response variable calculation. Moreover, the some of tree-based methods - TRM and TRP - produce results dissimilar to the original paper, with TRM occasionally outperforming the other models. This could be attributed to the software program used for the tree-based models receiving regular updates since the original paper was published, or to pre-processing steps that are taken in the original paper but not explicitly mentioned. Regardless, the findings of this paper reinforce the effectiveness of QRF.

The extension of this paper focuses on developing a method to assess the importance of a predictor at several different conditional quantiles in QRF. The proposed methods are all based on a permutation scheme and the MIE, which represents the relative increase in prediction error when a model uses a permuted variable compared to the original variable. The first proposed method, termed standard variable importance, uses the permutation scheme without any modifications. In contrast, the second method, called conditional variable importance, adjusts for correlations between variables by permuting values only within partitions defined by the cut points of conditioning variables in the QRF model. These concepts have not been explored in other literature until now, so this paper also investigates whether the importance of predictors changes across quantiles. To examine this, the proposed methods are applied to the *NHANES* dataset, where the importance of several predictors on the response variable blood pressure is analysed, as well as the *Boston Housing* dataset in order to test the generalisability of the results. The findings reveal that the standard and conditional permutation schemes produce different results, with the conditional permutation results being significantly lower, suggesting that the standard importance values are indeed being inflated due to correlations between predictors. For the *NHANES* dataset, the results indicate that Age and BMI are the most important predictors of blood pressure overall. Interestingly, the conditional variable importance results reveal that BMI is more important at lower conditional quantiles, whereas Age is more important at higher conditional quantiles, signifying that the importance of predictors can indeed vary across quantiles. For the *Boston Housing dataset*, there are some interesting patterns in the importance scores, with for instance the variable *crim* peaking around the 0.5th quantile, then dipping down to the 0.8th quantile, and then rising again afterwards, which again implies that the importance of variables is dependent on conditional quantiles. The importance scores do tend to drop at extremer quantiles, which can be due to heteroscedasticy or outliers in the tails, which can make those results difficult to interpret truthfully. This paper suggests that future research could explore methods to improve this interpretation of importance scores at extreme quantiles, such as weighted quantile loss functions, extremal random forests, and localised permutation importance measures. Overall, the findings from the extension indicate that predictor importance is indeed quantile-dependent. Consequently, it is possible to adjust the model to include predictors specific to the conditional quantiles of interest to achieve more reliable predictions.

References

Blake, C. L. (1998). Uci repository of machine learning databases [Computer software manual]. Irvine, CA. (Formerly available from `http://www.ics.uci.edu/~mlearn/MLRepository.html`)

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Koenker. (1978). Regression quantiles. *Econometrica: Journal of the Econometric Society*, 33–50.

Koenker, R. (2024). quantreg: Quantile regression [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=quantreg` (R package version 5.98)

Leisch, F. (2024). mlbench: Machine learning benchmark problems [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=mlbench` (R package version 2.1-5)

Linderman, G. C., Lu, J., Lu, Y., Sun, X., Xu, W., Nasir, K., ... Krumholz, H. M. (2018). Association of body mass index with blood pressure among 1.7 million chinese adults. *JAMA Network Open*, *1*(4), e181271–e181271.

Meinshausen. (2006). Quantile regression forests. *Journal of Machine Learning Research*, *7*(6).

Meinshausen, N. (2017). quantregforest: Quantile regression forests [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=quantregForest` (R package version 1.3-7)

Pruim, R. (2015). Nhanes: Data from the us national health and nutrition examination study [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=NHANES` (R package version 2.1.0)

R Core Team. (2023). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T. & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*, 1–11.

Strobl, C., Boulesteix, A.-L., Zeileis, A. & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *8*, 1–21.

Weisberg, S. (2014). *Applied linear regression* (Fourth ed.). Hoboken NJ: Wiley. Retrieved from `http://z.umn.edu/alr4ed`

A Programming code

The code of this paper can be found on the following github page:
https://github.com/QuintyOk/BachelorThesisEconometrics.git.

B Extension

B.1 Replication

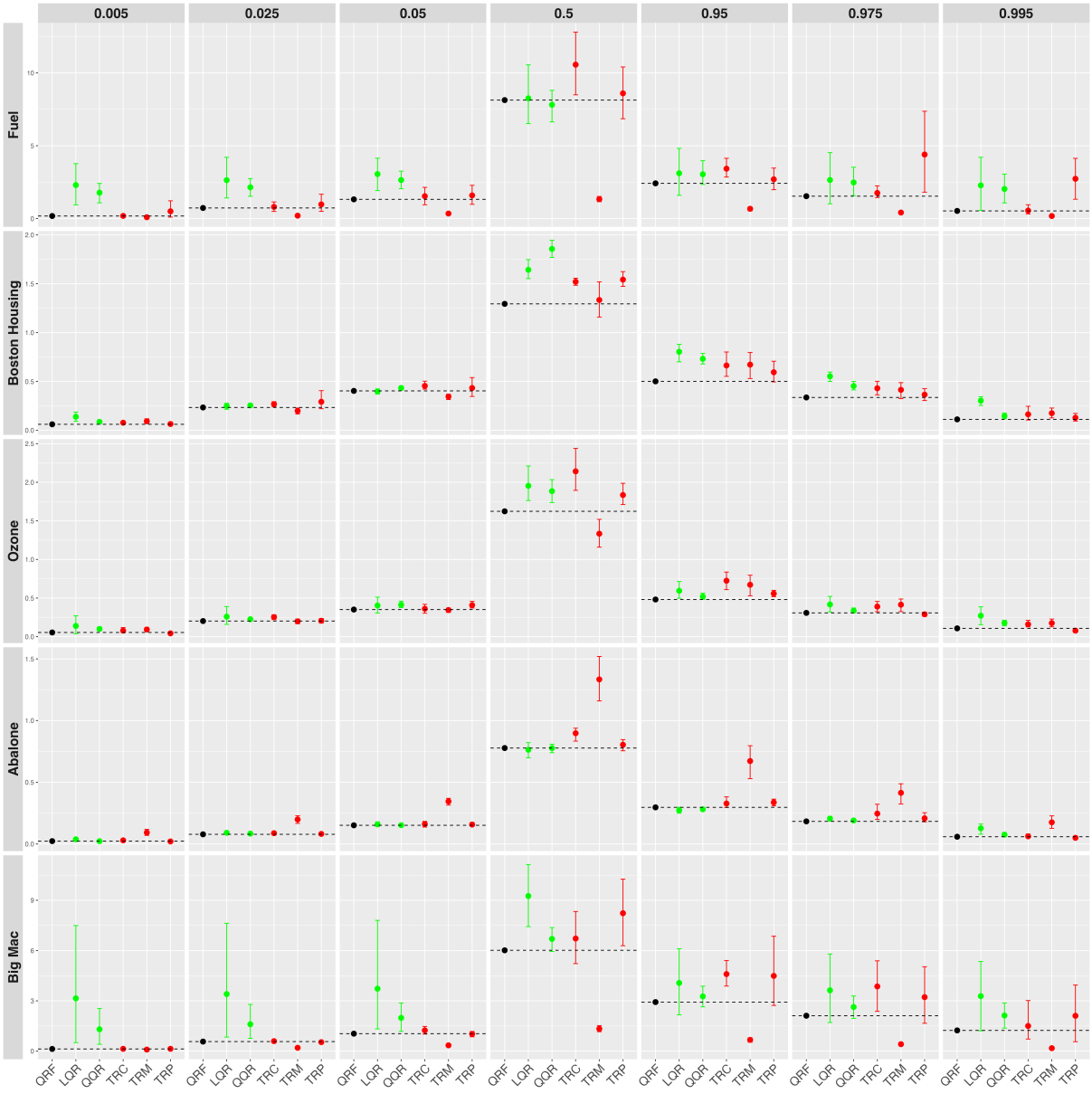

Figure 9: For each dataset, method and quantile, the average loss under additional noise is depicted. From left to right, the dots present QRF, LQR, QQR, TRC, TRM and TRP respectively in each graph. The columns represent the seven quantiles, whilst the rows illustrate the five different datasets. The vertical bars represent the 95% bootstrap confidence bounds.

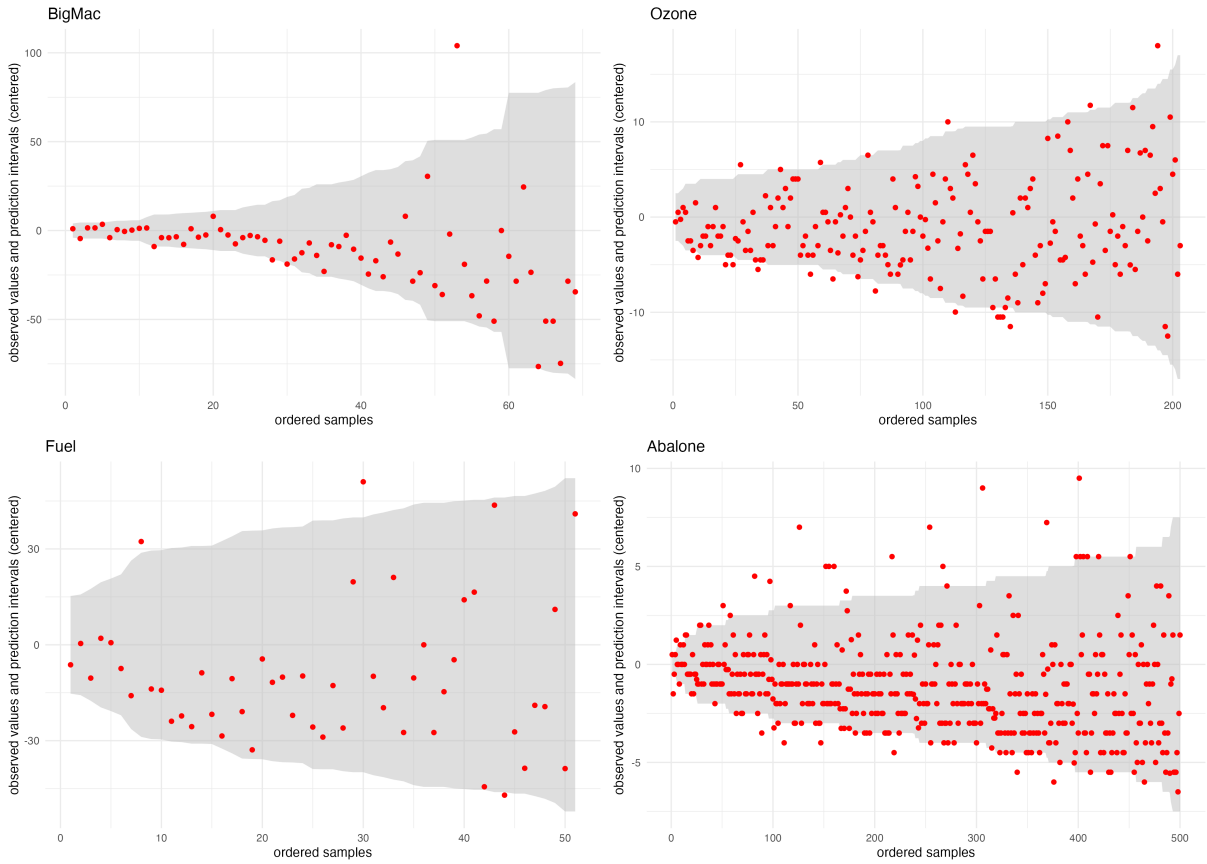

Figure 10: The sorted prediction intervals, as in Fig. 4, for the remaining datasets.

B.2 Extension

Since the Boston dataset yields such variable results, the five most important predictors according to Fig. 8 are selected and both the standard variable importance and conditional variable importance are calculated across quantiles, which are respectively illustrated in Fig. 11 and Fig. 12. Interestingly, instead of *crim* being the top predictor, *lstat*, which represents the lower status of the population, is the most important predictor, and shows an upwards trend across quantiles. Again, the importance values are different when computed with the conditional permutation scheme, with *crim* being the second most important predictor, whilst *indus* and *nox* appear to have no importance whatsoever, whilst *dis* fluctuates between being a positive and negative predictor, meaning that for some quantiles it is better to include it for predictions and for others it is worse.

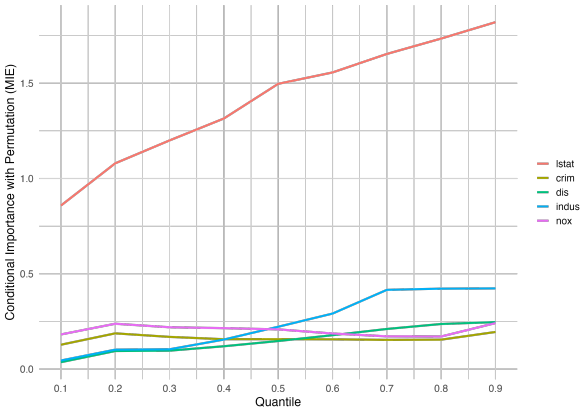

Figure 11: The standard variable importance by permutation (MIE) for the Boston Housing dataset with a subset of its variables.

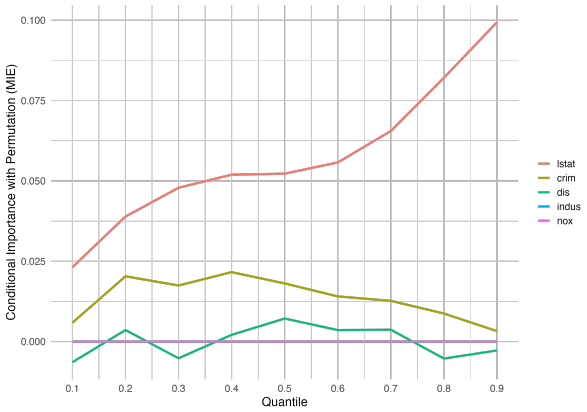

Figure 12: The standard variable importance by permutation (MIE) for the Boston Housing dataset with a subset of its variables.