

ePub^{WU} Institutional Repository

Carolin Strobl and Anne-Laure Boulesteix and Achim Zeileis and Torsten Hothorn

Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution

Paper

Original Citation:

Strobl, Carolin and Boulesteix, Anne-Laure and Zeileis, Achim and Hothorn, Torsten (2006)

Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution.

Research Report Series / Department of Statistics and Mathematics, 40. Department of Statistics and Mathematics, WU Vienna University of Economics and Business, Vienna.

This version is available at: <https://epub.wu.ac.at/1274/>

Available in ePub^{WU}: September 2006

ePub^{WU}, the institutional repository of the WU Vienna University of Economics and Business, is provided by the University Library and the IT-Services. The aim is to enable open access to the scholarly output of the WU.

Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution



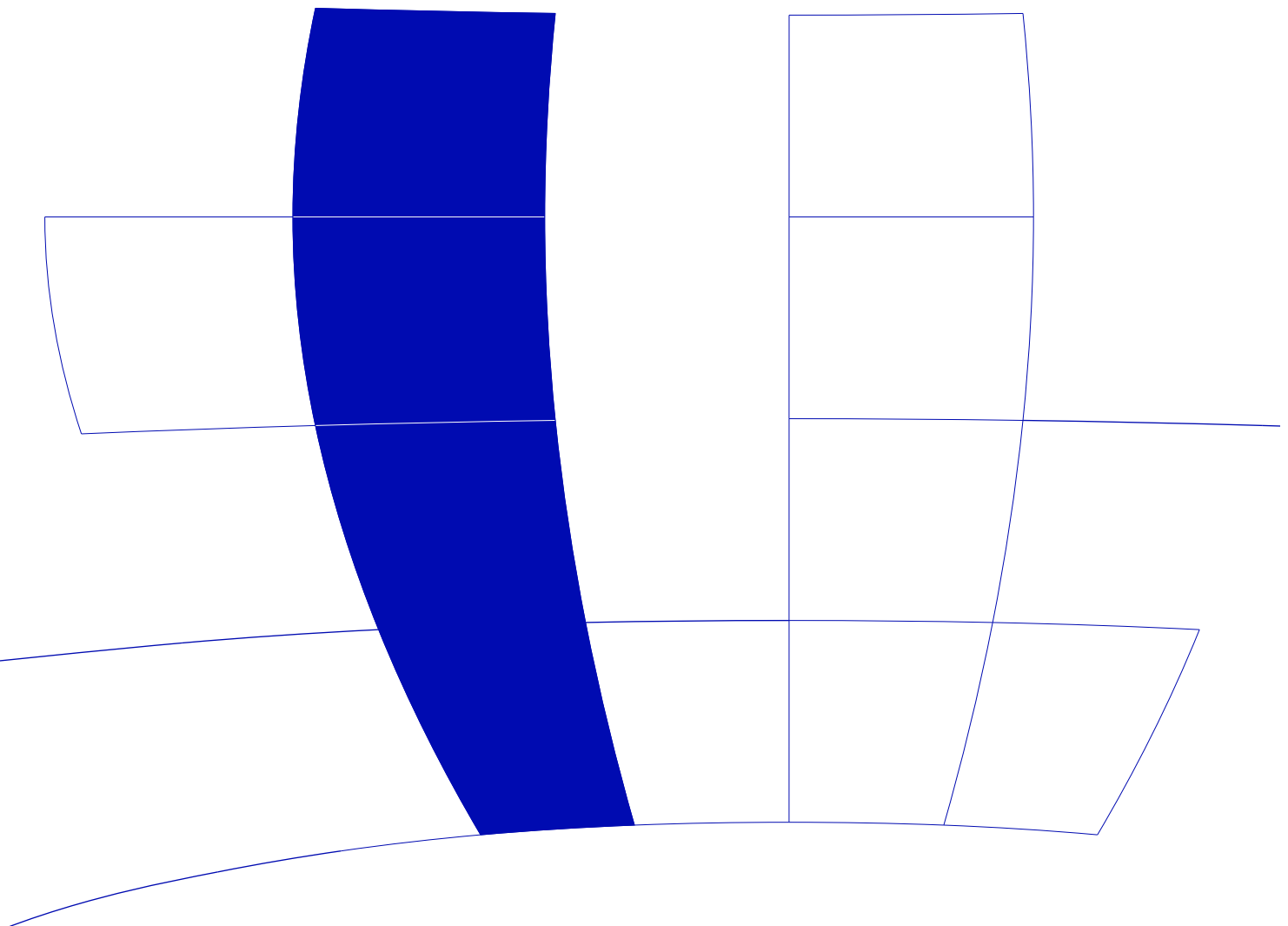
Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, Torsten Hothorn

Department of Statistics and Mathematics
Wirtschaftsuniversität Wien

Research Report Series

Report 40
September 2006

<http://statmath.wu-wien.ac.at/>



Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution

Carolin Strobl

Ludwig-Maximilians-Universität
München

Anne-Laure Boulesteix

Technische Universität München

Achim Zeileis

Wirtschaftsuniversität Wien

Torsten Hothorn

Friedrich-Alexander-Universität
Erlangen-Nürnberg

Abstract

Variable importance measures for random forests have been receiving increased attention as a means of variable selection in many classification tasks in bioinformatics and related scientific fields, for instance to select a subset of genetic markers relevant for the prediction of a certain disease. We show that random forest variable importance measures are a sensible means for variable selection in many applications, but are not reliable in situations where potential predictor variables vary in their scale level or their number of categories. This is particularly important in genomics and computational biology, where predictors often include variables of different types, for example when predictors include both sequence data and continuous variables such as folding energy, or when amino acid sequence data show different numbers of categories. Simulation studies are presented illustrating that, when random forest variable importance measures are used with data of varying types, the results are misleading because suboptimal predictor variables may be artificially preferred in variable selection. The two mechanisms underlying this deficiency are biased variable selection in the individual classification trees used to build the random forest on one hand, and effects induced by bootstrap sampling with replacement on the other hand. We propose to employ an alternative implementation of random forests, that provides unbiased variable selection in the individual classification trees. When this method is applied using subsampling without replacement, the resulting variable importance measures can be used reliably for variable selection even in situations where the potential predictor variables vary in their scale level or their number of categories. The usage of both random forest algorithms and their variable importance measures in the R system for statistical computing is illustrated and documented thoroughly in an application re-analysing data from a study on RNA editing. Therefore the suggested method can be applied straightforwardly by scientists in bioinformatics research.

Keywords: random forests, variable importance, Gini importance, variable selection bias.

1. Background

In bioinformatics and related scientific fields, such as statistical genomics and genetic epidemiology, an important task is the prediction of a categorical response variable (such as the disease status of a patient or the properties of a molecule) based on a large number of predictors. The aim of this research is on one hand to predict the value of the response variable from the values of the predictors, i.e. to create a diagnostic tool, and on the other hand to reliably identify relevant predictors from a large set of candidate variables.

From a statistical point of view, one of the challenges in identifying these relevant predictor

variables is the so-called “small n large p ” problem: Usual data sets in genomics often contain hundreds or thousands of genes or markers that serve as predictor variables X_1, \dots, X_p , but only for a comparatively small number n of subjects or tissue types.

Traditional statistical models used in clinical case control studies for predicting the disease status from selected predictor variables, such as logistic regression, are not suitable for “small n large p ” problems (Bureau, Dupuis, Falls, Lunetta, Hayward, Keith, and Eerdewegh 2005; Heidema, Boer, Nagelkerke, Mariman, van der A, and Feskens 2006). A more appropriate approach from machine learning, that has been proposed recently for prediction and variable selection in various fields related to bioinformatics and computational biology, is the nonlinear and nonparametric random forest method (Breiman 2001). It also provides variable importance measures for variable selection purposes.

Random forests have been successfully applied to various problems in, e.g., genetic epidemiology and microbiology in general within the last five years. Within a very short period of time, random forests have become a major data analysis tool, that performs well in comparison with many standard methods (Heidema *et al.* 2006; Díaz-Uriarte and de Andrés 2006). What has greatly contributed to the popularity of random forests is the fact that they can be applied to a wide range of prediction problems, even if they are non-linear and involve complex high-order interaction effects, and that random forests produce variable importance measures for each predictor variable.

Applications of random forests in bioinformatics include large-scale association studies for complex genetic diseases, as e.g. Lunetta, Hayward, Segal, and Eerdewegh (2004) and Bureau *et al.* (2005), who detect SNP-SNP interactions in the case-control context by means of computing a random forest variable importance measure for each polymorphism. An application to the analysis of gene expression data is presented by Díaz-Uriarte and de Andrés (2006), who recommend random forests as a gene selection method for sample classification with microarray data. (We refer to Gunther, Stone, Gerwien, Bento, and Heyes (2003); Huang, Pan, Grindle, Han, Chen, Park, Miller, and Hall (2005); Shih (2005) for other applications of the random forest methodology to microarray data.)

Prediction of phenotypes based on amino acid or DNA sequence is another important area of application of random forests, since possibly involving many interactions. For example, Segal, Barbour, and Grant (2004) use random forests to predict the replication capacity of viruses, such as HIV-1, based on amino acid sequence from reverse transcriptase and protease. Cummings and Segal (2004) link the rifampin resistance in *Mycobacterium tuberculosis* to a few amino acid positions in rpoB, whereas Cummings and Myers (2004) predict C-to-U edited sites in plant mitochondrial RNA based on sequence regions flanking edited sites and a few other (continuous) parameters.

The random forest approach was shown to outperform six other methods in the prediction of protein interactions based on various biological features such as gene expression, gene ontology (GO) features and sequence data (Qi, Bar-Joseph, and Klein-Seetharaman 2006). Other applications of random forests can be found in fields as different as quantitative structure-activity relationship (QSAR) modelling (Guha and Jurs 2003; Svetnik, Liaw, Tong, Culberson, Sheridan, and Feuston 2003), nuclear magnetic resonance spectroscopy (Arun and Langmead 2006), landscape epidemiology (Furlanello, Neteler, Merler, Menegon, Fontanari, Donini, Rizzoli, and Chemini 2003) and medicine in general (Ward, Pajevic, Dreyfuss, and Malley 2006).

The scope of this paper is to show that the variable importance measures of Breiman’s original random forest method (Breiman 2001), based on CART classification trees (Breiman, Friedman, Olshen, and Stone 1984), are a sensible means for variable selection in many of these applications, but are not reliable in situations where potential predictor variables vary in their scale level or their number of categories, as, e.g., when both genetic and environmental variables, individually and in interactions, are considered as potential predictors, or predictor variables of the same type vary in the number of categories present in a certain sample, as is often the case in genomics, bioinformatics and related disciplines.

Simulation studies are presented illustrating that variable selection with the variable importance

measure of the original random forest method bears the risk that suboptimal predictor variables are artificially preferred in such scenarios.

In an extra section, further details and explanations of the statistical sources underlying the deficiency of the variable importance measures of the original random forest method, namely biased variable selection in the individual classification trees used to build the random forest and effects induced by bootstrap sampling with replacement, are given.

We propose to employ an alternative random forest method, the variable importance measure of which can be employed to reliably select relevant predictor variables in any data set. The performance of this method is compared to that of the original random forest method in simulation studies, and is illustrated by an application to the prediction of C-to-U edited sites in plant mitochondrial RNA, re-analysing the data of [Cummings and Myers \(2004\)](#) that were previously analysed with the original random forest method.

2. Methods

Here we focus on the use of random forests for classification tasks, rather than regression tasks, for instance for predicting the disease status from a set of selected genetic and environmental risk factors, or for predicting whether a site of interest is edited by means of neighboring sites and other predictor variables as in our application example.

Random forests are an ensemble method that combines several individual classification trees in the following way: From the original sample several bootstrap samples are drawn, and an unpruned classification tree is fit to each bootstrap sample. The variable selection for each classification tree is conducted only from a small random subset of predictor variables, so that the “small n large p ” problem is avoided. From the complete forest the status of the response variable is predicted as an average or majority vote of the predictions of all trees.

Random forests can highly increase the prediction accuracy as compared to individual classification trees, because the ensemble adjusts for the instability of the individual trees induced by small changes in the learning sample, that impairs the prediction accuracy in test samples. However, the interpretability of a random forest is not as straightforward as that of an individual classification tree, where the influence of a predictor variable directly corresponds to its position in the tree. Thus, alternative measures for variable importance are required for the interpretation of random forests.

2.1. Random forest variable importance measures

A naive variable importance measure to use in tree-based ensemble methods is to merely count the number of times each variable is selected by all individual trees in the ensemble.

More elaborate variable importance measures incorporate a (weighted) mean of the individual trees’ improvement in the splitting criterion produced by each variable ([Friedman 2001](#)). An example for such a measure in classification is the “Gini importance” available in random forest implementations. The “Gini importance” describes the improvement in the “Gini gain” splitting criterion.

The most advanced variable importance measure available in random forests is the “permutation accuracy importance” measure. Its rationale is the following: By randomly permuting the predictor variable X_j , its original association with the response Y is broken. When the permuted variable is used to predict the response, the prediction accuracy (i.e. the number of observations classified correctly) decreases substantially as compared to the original variable, if the variable was associated with the response originally. Thus, a reasonable measure for variable importance is the difference in prediction accuracy before and after permuting X_j .

For variable selection purposes the advantage of the random forest permutation accuracy importance measure as compared to univariate screening methods is that it covers the impact of each predictor variable individually as well as in multivariate interactions with other predictor variables.

For example, Lunetta *et al.* (2004) find that genetic markers relevant in interactions with other markers or environmental variables can be detected more efficiently by means of random forests than by means of univariate screening methods like Fisher’s exact test.

The Gini importance and the permutation accuracy importance measures are employed as variable selection criteria in many recent studies in various disciplines related to bioinformatics, as outlined in the background section. Therefore we want to investigate their reliability as variable importance measures in different scenarios.

In the simulation studies presented in the next section, we compare the behavior of all three random forest variable importance measures, namely the number of times each variable is selected by all individual trees in the ensemble (termed “selection frequency” in the following), the “Gini importance” and the permutation accuracy importance measure (termed “permutation importance” in the following).

2.2. Simulation studies

The reference implementation of Breiman’s original random forest method (Breiman 2001) is available in the R system for statistical computing (R Development Core Team 2006) via the `randomForest` add-on package (Breiman, Cutler, Liaw, and Wiener 2006; Liaw and Wiener 2002). The behavior of the selection frequency, the Gini importance and the permutation importance of the `randomForest` function is explored in a simulation design where potential predictor variables vary in their scale level and number of categories.

As an alternative, we propose to use the new random forest function `cforest` available in the R add-on package `party` (Hothorn, Hornik, and Zeileis 2006a) in such scenarios. In contrast to `randomForest`, the `cforest` function creates random forests not from CART classification trees based on the Gini split criterion (Breiman *et al.* 1984), that are known to prefer variables with, e.g., more categories in variable selection (Breiman *et al.* 1984; Kononenko 1995; Kim and Loh 2001; Boulesteix 2006b,a; Strobl, Boulesteix, and Augustin 2005), but from unbiased classification trees based on a conditional inference framework (Hothorn, Hornik, and Zeileis 2006b). The problem of biased variable selection in classification trees is covered more thoroughly in a separate section below.

Predictor variables	
X_1	$\sim N(0, 1)$
X_2	$\sim M(2)$
X_3	$\sim M(4)$
X_4	$\sim M(10)$
X_5	$\sim M(20)$

Table 1: In the simulation studies the predictor variables are sampled independently from the following distributions. $N(0, 1)$ stands for the standard normal distribution, $M(k)$ stands for the multinomial distribution with values in $\{0, \dots, k - 1\}$ and equal probabilities, $B(p)$ stands for the binomial distribution with probability p , thus $M(2)$ equals $B(0.5)$.

Response variable		
null case	Y	$\sim B(0.5)$
power case	$Y X_2 = 1$	$\sim B(0.3)$
	$Y X_2 = 2$	$\sim B(0.7)$

Table 2: In the simulation studies the response variable is sampled from binomial distributions. The degree of dependence between the response Y and X_2 is regulated by the probability p of the binomial distribution $B(p)$ of Y conditional on X_2 .

Since the `cforest` function does not employ the Gini criterion, we investigate the behavior of the Gini importance for the `randomForest` function only. The selection frequency and the permutation importance is studied for both functions `randomForest` and `cforest` in two ways: Either the individual trees are built on samples drawn with replacement, as suggested in Breiman (2001), or on samples drawn without replacement.

The simulation design used throughout this paper represents a scenario where a binary response variable Y is supposed to be predicted from a set of potential predictor variables that vary in their scale level and number of categories. The first predictor variable X_1 is continuous, while the other predictor variables X_2, \dots, X_5 are categorical with their number of categories between two and up to twenty. The simulation designs of both studies are summarized in Tables 1 and 2. The sample size for all simulation studies was set to $n = 120$.

In the first simulation study, the so-called null case, none of the predictor variables is informative for the response, i.e. all predictor variables and the response are sampled independently. In this situation a sensible variable importance measure should not prefer any one predictor variable over any other.

In the second simulation study, the so-called power case, the predictor variable X_2 is informative for the response, i.e. the distribution of the response depends on the value of this predictor variable. In this situation, a sensible variable importance measure should be able to detect the informative predictor variable.

3. Results and discussion

Our simulation studies show that for the `randomForest` function all three variable importance measures are biased, the Gini importance even strongly so. For the `cforest` function reliable results can be achieved both with the selection frequency and the permutation importance if the function is used together with subsampling without replacement. Otherwise the measures are biased as well.

3.1. Results of the null case simulation study

In the null case, when all predictor variables are equally uninformative, the selection frequencies as well as the Gini importance and the permutation importance of all predictor variables are supposed to be equal.

However, as presented in Figure 1, the mean selection frequencies (over 1000 simulation runs) of the predictor variables differ substantially when the `randomForest` function (cf. top row in Figure 1) or the `cforest` function with bootstrap sampling (cf. bottom row, left plot in Figure 1) are used. Variables with more categories are obviously preferred. Only when the `cforest` function is used together with subsampling without replacement (cf. bottom row, right plot in Figure 1) are the variable selection frequencies for the uninformative predictor variables equally low as desired. (For sampling without replacement the subsample size is set to 0.632 times the original sample size, because in bootstrap sampling with replacement about 63.2% of the data end up in the bootstrap sample.)

It is obvious that variable importance cannot be represented reliably by the selection frequencies, that can be considered as very basic variable importance measures, if the potential predictor variables vary in their scale level or number of categories.

The mean Gini importance (over 1000 simulation runs), that is displayed in Figure 2 is biased even stronger. Like the selection frequencies for the `randomForest` function (cf. top row in Figure 1) the Gini importance shows a strong preference for variables with many categories and the continuous variable, the statistical sources of which are explained in the section on variable selection bias in classification trees below. We conclude that the Gini importance cannot be used to reliably measure variable importance in this situation either.

We now consider the more advanced permutation importance measure. We find that here an effect

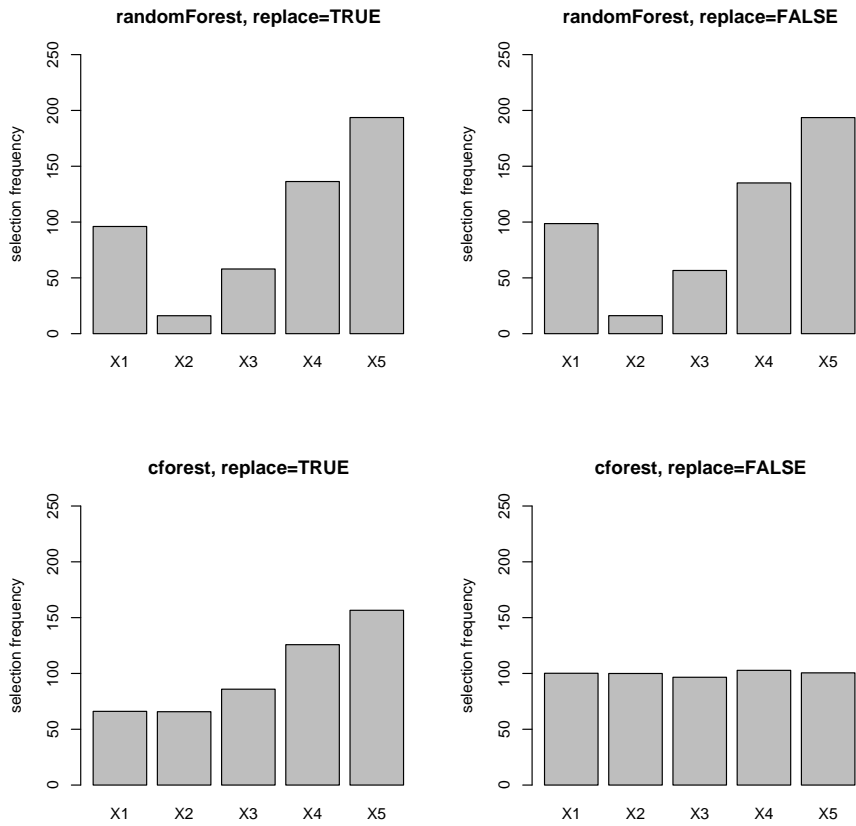


Figure 1: Mean variable selection frequencies for the null case study, where none of the predictor variables is informative. The plots in the top row display the frequencies when the `randomForest` function is used, the bottom row when the `cforest` function is used. The left column corresponds to bootstrap sampling with replacement, the right column to subsampling without replacement.

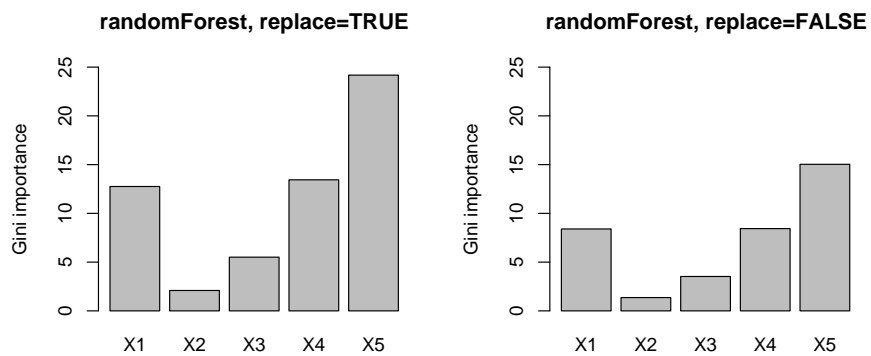


Figure 2: Mean Gini importance for the null case study, where none of the predictor variables is informative. The left plot corresponds to bootstrap sampling with replacement, the right plot to subsampling without replacement.

of the scale level or number of categories of the potential predictor variables is less obvious but still severely affects the reliability and interpretability of the variable importance measure.

Figure 3 shows boxplots of the distributions (over 1000 simulation runs) of the permutation importance measures of both functions for the null case. The plots in the top row again display the distribution when the `randomForest` function is used, the bottom row when the `cforest` function is used. The left column of plots displays the distributions when bootstrap sampling is conducted with replacement, while the right column displays the distributions when subsampling is conducted without replacement.

Figure 4 shows boxplots of the distributions of the scaled version of the permutation importance measures of both functions, incorporating the standard deviation of the measures.

The plots show that for the `randomForest` function (cf. top row in Figures 3 and 4) and, less pronounced, for the `cforest` function with bootstrap sampling (cf. bottom row, left plot in Figures 3 and 4), the deviation of the permutation importance measure over the simulation runs is highest for the variable X_5 with the highest number of categories, and decreases for the variables with less categories and the continuous variable. This effect is weakened but not substantially altered by scaling the measure (cf. Figure 3 vs. Figure 4).

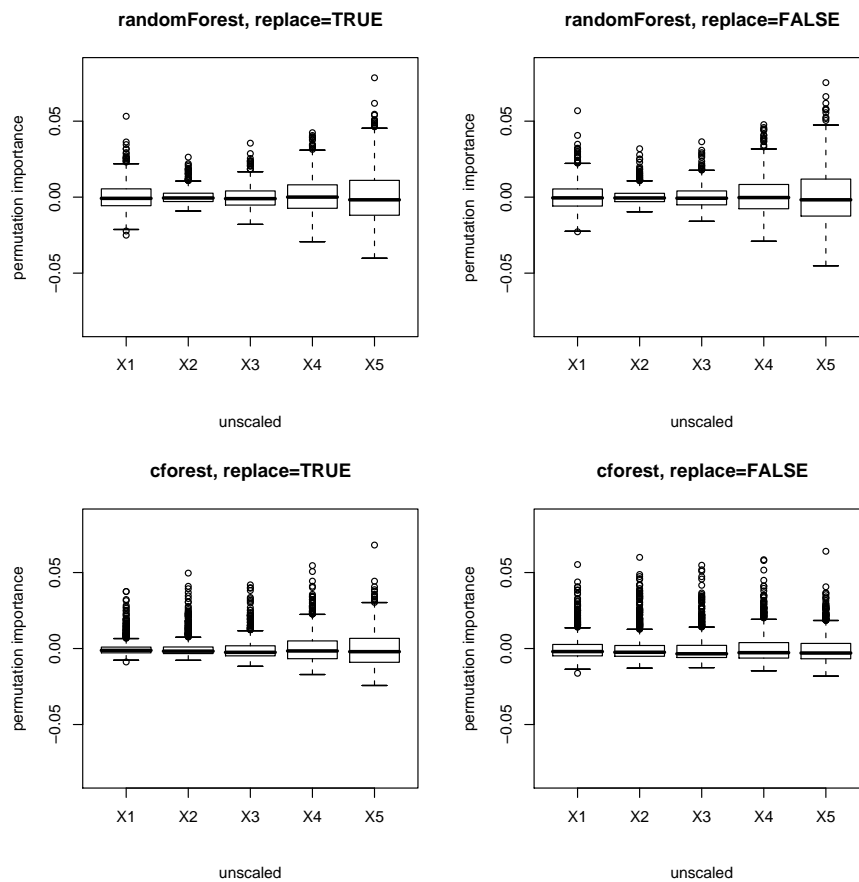


Figure 3: Distributions of the unscaled permutation importance measures for the null case study, where none of the predictor variables is informative. The plots in the top row display the distributions when the `randomForest` function is used, the bottom row when the `cforest` function is used. The left column corresponds to bootstrap sampling with replacement, the right column to subsampling without replacement.

As opposed to the obvious effect in the selection frequencies and the Gini importance, there is no effect in the mean values of the distributions of the permutation importance measures, which are in mean close to zero as expected for uninformative variables. However, the notable differences in the variance of the distributions for predictor variables with different scale level or number of categories seriously affect the expressiveness of the variable importance measure.

In a single trial this effect may lead to a severe under- or overestimation of the variable importance of variables that have more categories as an artefact of the method, even though they are no more informative than the other variables.

Only when the `cforest` function is used together with subsampling without replacement (cf. bottom row, right plot in Figures 3 and 4) does the deviation of the permutation importance measure over the simulation runs not depend on the number of categories or scale level of the predictor variables.

Thus, only the variable importance measure available in `cforest`, and only when used together with sampling without replacement, reliably reflects the true importance of potential predictor variables in a scenario where the potential predictor variables vary in their scale level or number of categories.

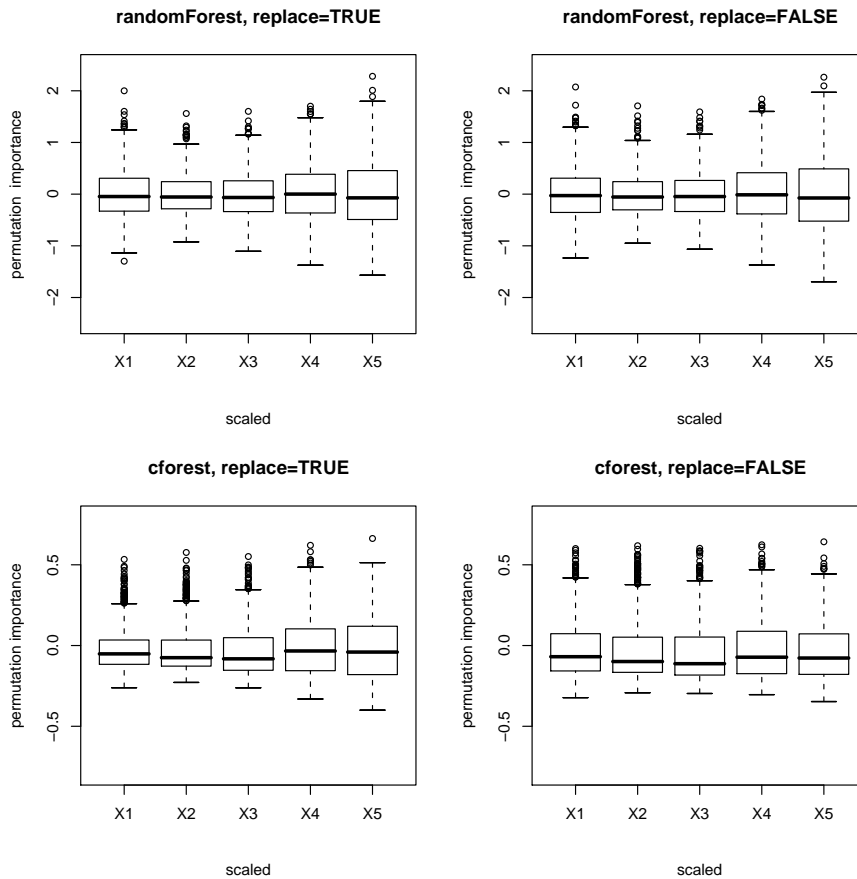


Figure 4: Distributions of the scaled permutation importance measures for the null case study, where none of the predictor variables is informative. The plots in the top row display the distributions when the `randomForest` function is used, the bottom row when the `cforest` function is used. The left column corresponds to bootstrap sampling with replacement, the right column to subsampling without replacement.

3.2. Results of the power case simulation study

In the power case, where only the predictor variable X_2 is informative, each of the variable importance measures should clearly prefer X_2 while the respective values for the remaining predictor variables should be equally low.

Figure 5 shows that the mean selection frequencies (again over 1000 simulation runs) of the predictor variables again differ substantially when the `randomForest` function (cf. top row in Figure 5) is used, and the relevant predictor variable X_2 cannot be identified. With the `cforest` function with bootstrap sampling (cf. bottom row, left plot in Figure 5) there is still bias obvious in the election frequencies of the categorical predictor variables with many categories. Only when the `cforest` function is used together with subsampling without replacement (cf. bottom row, right plot in Figure 5), are the variable selection frequencies for the uninformative predictor variables equally low as desired, and the value for the relevant predictor variable X_2 sticks out.

The mean Gini importance, that is displayed in Figure 6, again shows a strong bias towards variables with many categories and the continuous variable. It completely fails to identify the relevant predictor variable, with the mean value for the relevant variable X_2 only slightly higher than in the null case.

Figures 7 and 8 show boxplots of the distributions of the unscaled and scaled permutation importance measures of both functions. Again for the `randomForest` function (cf. top row in Figures 7 and 8) and, less pronounced, for the `cforest` function with bootstrap sampling (cf. bottom row, left plot in Figures 7 and 8), the deviation of the permutation importance measure over the simulation runs is highest for the variable X_5 with the highest number of categories, and decreases for the variables with less categories and the continuous variable. This effect is weakened but not substantially altered by scaling the measure (cf. Figure 7 vs. Figure 8).

As expected the mean value of the permutation importance measure for the informative predictor variable X_2 is higher than for the uninformative variables. However, the deviation of the variable importance measure for the uninformative variables with many categories X_4 and X_5 is so high that in a single trial these uninformative variables may outperform the informative variable as an artefact of the method.

Thus, only the variable importance measure computed with the `cforest` function, and only when used together with sampling without replacement, is able to reliably detect the informative variable out of a set of uninformative competitors.

We have seen that for the assessment of variable importance and variable selection purposes it is important to use a reliable method, that is not affected by other characteristics of the predictor variables, such as the scale level or number of categories. Statistical explanations of our findings are given in a later section.

In addition to its superiority in the assessment of variable importance the `cforest` method, especially when used together with sampling without replacement, can also be superior to the `randomForest` method with respect to classification accuracy in situations like that of the power case simulation study, where uninformative predictor variables with many categories “fool” the variable importance measure of the `randomForest` function.

Due to its artificial preference for uninformative predictor variables with many categories the `randomForest` method can produce a higher mean misclassification rate than the `cforest` method. For the simulation design chosen here, the mean misclassification rates (again over 1000 simulation runs) of the methods, applied with sampling with and without replacement, show the ranking displayed in Table 3. Each method was applied to the same simulated test set in each simulation run. The test sets were generated from the same data generating process as the learning sets. We find that the `cforest` method, especially with sampling without replacement, outperforms the other methods. A similar result is obtained in the application to C-to-U conversion data presented in the next section.

The differences in classification accuracy are moderate in the latter case, however one could think of more extreme situations that would produce even greater differences. This shows that the same

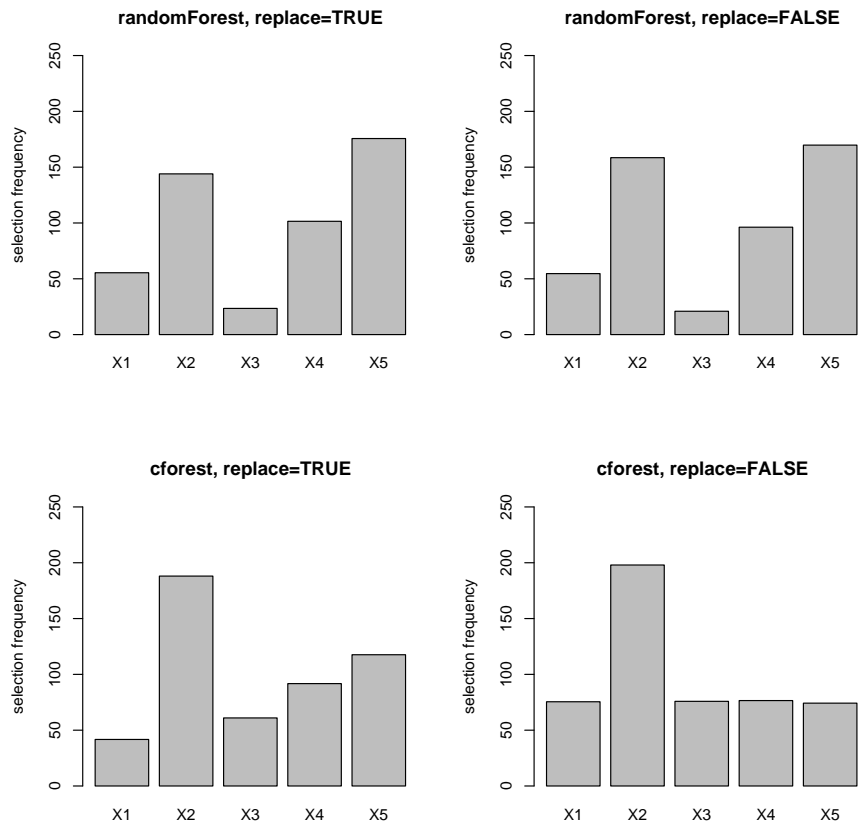


Figure 5: Mean variable selection frequencies for the power case study, where only the second predictor variable is informative. The arrangement of the plots corresponds to that in Figure 1.

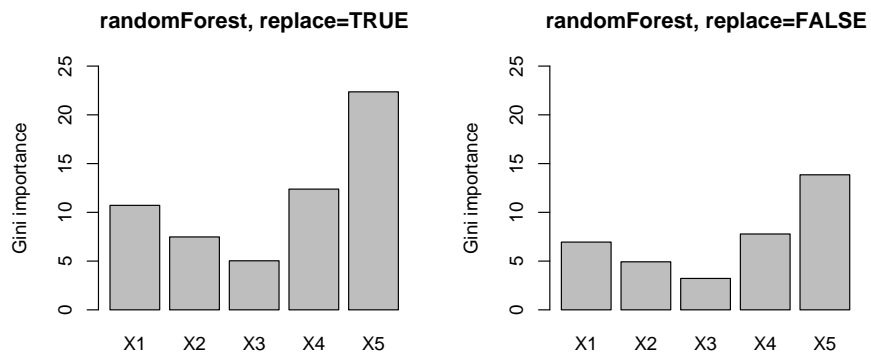


Figure 6: Mean Gini importance for the power case study, where only the second predictor variable is informative. The arrangement of the plots corresponds to that in Figure 2.

mechanisms underlying the variable importance bias can also affect the classification accuracy, e.g. when suboptimal predictor variables, that do not add to the classification accuracy, are artificially preferred in variable selection merely because they have more categories.

Method	Sampling with replacement	Mean misclassification rate	Std. error of mean
<code>randomForest</code>	true	0.40241	0.00161
	false	0.40241	0.00162
<code>cforest</code>	true	0.3519	0.00199
	false	0.34066	0.002

Table 3: Mean misclassification rates of the `randomForest` method applied with sampling with and without replacement as compared to those of the `cforest` method applied with sampling with and without replacement in the power case study.

3.3. Application to C-to-U conversion data

RNA editing is the process whereby RNA is modified from the sequence of the corresponding DNA template (Cummings and Myers 2004). For instance, cytidine-to-uridine conversion (abbreviated

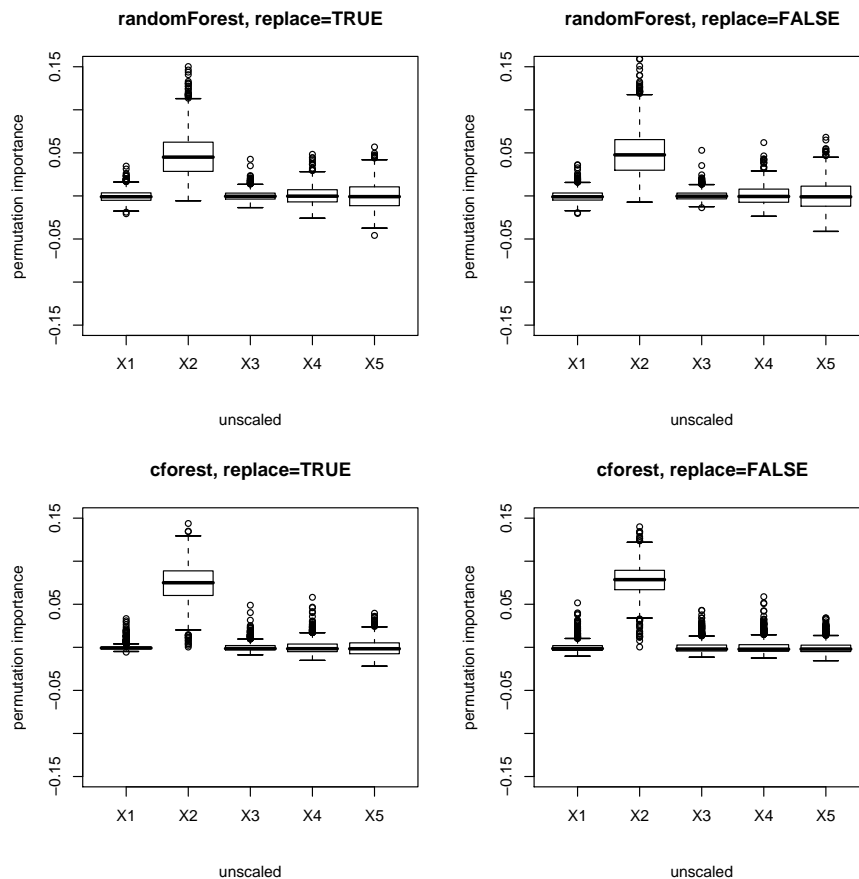


Figure 7: Distributions of the unscaled permutation importance measures for the power case study, where only the second predictor variable is informative. The arrangement of the plots corresponds to that in Figure 3.

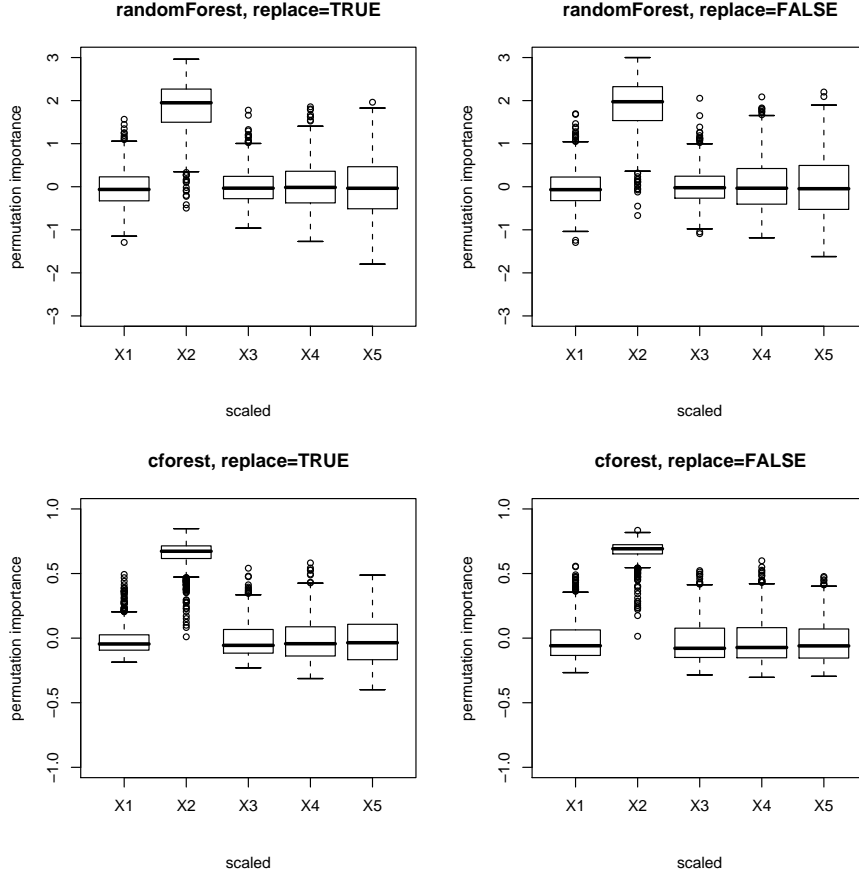


Figure 8: Distributions of the scaled permutation importance measures for the power case study, where only the second predictor variable is informative. The arrangement of the plots corresponds to that in Figure 4.

C-to-U conversion) is common in plant mitochondria. The mechanisms of this conversion remain largely unknown, although the role of neighboring nucleotides is emphasized. Cummings and Myers (2004) suggest to use information from sequence regions flanking the sites of interest to predict editing in *arabidopsis thaliana*, *brassica napus* and *oryza sativa* based on random forests. The *arabidopsis thaliana* data of Cummings and Myers (2004) can be loaded from the journal's homepage (<http://www.biomedcentral.com/content/supplementary/1471-2105-5-132-S1.txt>).

For each of the 876 observations, the data set gives

- the response at the site of interest (binary: edited/not edited)

and as potential predictor variables

- the 40 nucleotides at positions -20 to 20, relative to the edited site (4 categories),
- the codon position (4 categories),
- the estimated folding energy (continuous) and
- the difference in estimated folding energy between pre-edited and edited sequences (continuous).

We first derive the permutation importance measure for each of the 43 potential predictor variables with each method. As can be seen from the barplot in Figure 9, the (scaled) variable importance measures largely reflect the results of Cummings and Myers (2004) based on the Gini importance measure, but differ noticeably for the `randomForest` and `cforest` function and the different resampling schemes.

In particular, the variable importance measure of the `randomForest` function produces more “noise” than that of the `cforest` function: the contrast of amplitudes between irrelevant and relevant predictors is more pronounced when the `cforest` function is used. Moreover, the variable importance of the two continuous predictor variables “estimated folding energy” and “difference in estimated folding energy” (the two last bars on the right) is underestimated severely when bootstrap sampling with replacement is used.

Similarly to the simulation study, we also compared the prediction accuracy of the four approaches for this data set. To do so, we split the original data set into learning and test sets with size ratio 2:1 in a standard split-sample validation scheme. A random forest is grown based on the learning

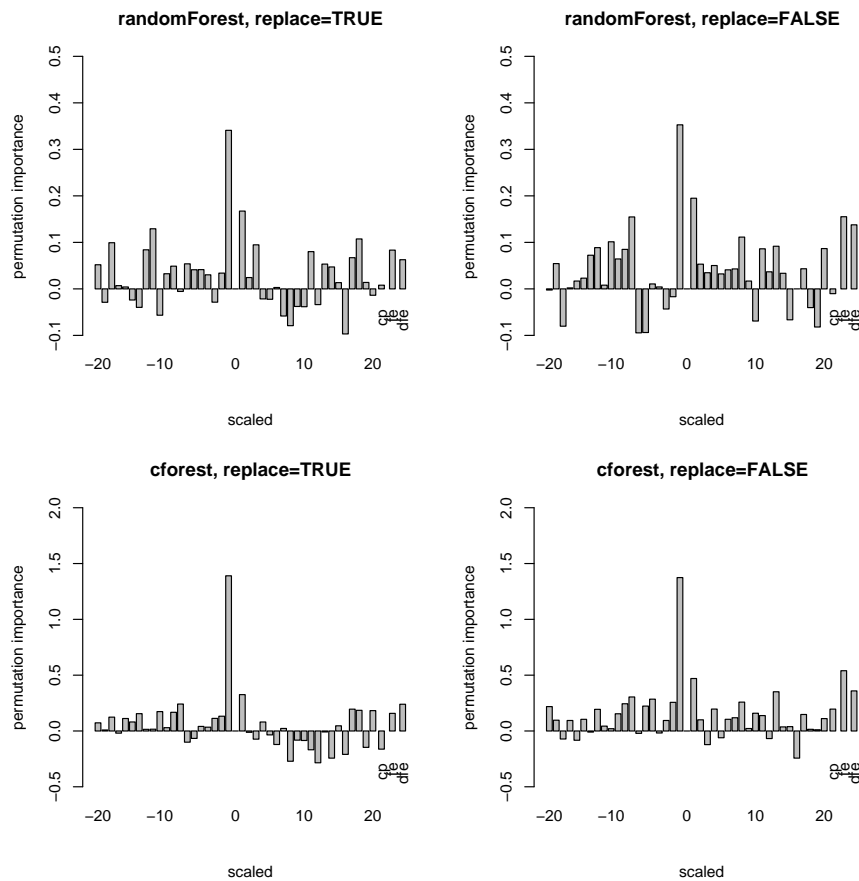


Figure 9: Scaled variable importance measures for the C-to-U conversion data. The plots in the top row display the measures when the `randomForest` function is used, the bottom row when the `cforest` function is used. The left column corresponds to bootstrap sampling with replacement, the right column to subsampling without replacement. In each plot the positions -20 through 20 indicate the nucleotides flanking the site of interest, and the last three bars on the right refer to the codon position (cp), the estimated folding energy (fe) and the difference in estimated folding energy (dfe).

Method	Sampling with replacement	Mean misclassification rate	Std. error of mean
<code>randomForest</code>	true	0.29682	0.00245
	false	0.29144	0.00244
<code>cforest</code>	true	0.28325	0.00253
	false	0.28055	0.00264

Table 4: Mean misclassification rates of the `randomForest` method applied with sampling with and without replacement as compared to those of the `cforest` method applied with sampling with and without replacement for the C-to-U conversion data.

set and subsequently used to predict the observations in the test set. This procedure is repeated 100 times, and the mean classification error rate over the 100 runs is reported in Table 4. Again we find a slight superiority of the `cforest` function, especially when sampling is conducted without replacement. (Differences to the accuracy values reported in Cummings and Myers (2004) are most likely due to their use of a different validation scheme, that is not reported in detail in Cummings and Myers (2004).)

All function calls and all important options of the `randomForest` and `cforest` functions used in the simulation studies and the application to C-to-U conversion data are documented in the supplementary file.

4. Sources of variable importance bias

The main difference between the `randomForest` function, based on CART trees (Breiman *et al.* 1984), and `cforest` function, based on conditional inference trees (Hothorn *et al.* 2006b), is that in `randomForest` the variable selection in the individual CART trees is biased, so that e.g. variables with more categories are preferred. This is illustrated in the next section on variable selection bias in individual classification trees.

However, even if the individual trees select variables in an unbiased way as in the `cforest` function, we find that the variable importance measures, as well as the selection frequencies of the variables, are affected by the bootstrap sampling with replacement. This is explained in the section on effects induced by bootstrapping.

4.1. Variable selection bias in the individual classification trees

Let us again consider the null case simulation study design, where none of the variables is informative, and thus should be selected with equally low probabilities in a classification tree.

In traditional classification tree algorithms, like CART, for each variable a split criterion like the “Gini index” is computed for all possible cutpoints within the range of that variable. The variable selected for the next split is the one that produced the highest criterion value overall, i.e. in its best cutpoint.

Obviously variables with more potential cutpoints are more likely to produce a good criterion value by chance, as in a multiple testing situation. Therefore, if we compare the highest criterion value of a variable with two categories, say, that provides only one cutpoint from which the criterion was computed, with a variable with four categories, that provides seven cutpoints from which the best criterion value is used, the latter is often preferred. Because the number of cutpoints grows exponentially with the number of categories of unordered categorical predictors we find a preference for variables with more categories in CART-like classification trees. (For further reading on variable selection bias in classification trees see, e.g., the corresponding sections in Kim and Loh (2001); Kononenko (1995); Strobl *et al.* (2005); Hothorn *et al.* (2006b); Dobra and Gehrke (2001); Strobl (2005a,b).)

Since the Gini importance measure in `randomForest` is directly derived from the Gini index split

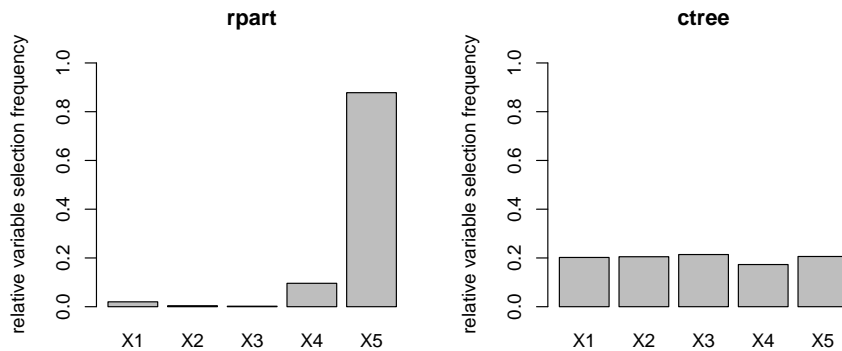


Figure 10: Relative selection frequencies for the `rpart` (left) and the `ctree` (right) classification tree methods. All variables are uninformative as in the null case simulation study.

criterion used in the underlying individual classification trees, it carries forward the same bias as was shown in Figures 2 and 6.

Conditional inference trees (Hothorn *et al.* 2006b), that are used to construct the classification trees in `cforest`, are unbiased in variable selection. Here, the variable selection is conducted by minimizing the p value of an independence test, like e.g. the χ^2 test, that incorporates the number of categories of each variable in the degrees of freedom.

The mean selection frequencies (again over 1000 simulation runs) of the five predictor variables of the null case simulation study design for both CART classification trees (as implemented in the `rpart` function (Therneau, Atkinson, and Ripley 2006)) and conditional inference trees (function `ctree` (Hothorn *et al.* 2006b)) are displayed in Figure 10. We find that the variable selection with the `rpart` function is highly biased, while for the `ctree` function it is unbiased.

The variable selection bias that occurs in every individual tree in the `randomForest` function also has a direct effect on the variable importance measures of this function. Predictor variables with more categories are artificially preferred in variable selection in each splitting decision. Thus, they are selected in more individual classification trees and tend to be situated closer to the root node in each tree.

The variable selection bias affects the variable importance measures in two respects. Firstly, the variable selection frequencies over all trees are directly affected by the variable selection bias in each individual tree. Secondly, the effect on the permutation importance is less obvious but just as severe.

When permuting the variables to compute their permutation importance measure, the variables closer to the root node can affect the prediction accuracy of a large set of observations represented by this node, while variables in the bottom nodes affect only small subsets of observations. Thus, the range of possible changes in prediction accuracy in the random forest, i.e. the deviation of the variable importance measure, is higher for variables that achieve positions closer to the root node due to variable selection bias in the individual trees.

We found in Figures 1 through 9, that the effects induced by the differences in scale level of the predictor variables were more pronounced for the `randomForest` function, where variable selection in the individual trees is biased, than for the `cforest` function, where the individual trees are unbiased. However, we also found that when the `cforest` function is used with bootstrap sampling, the variable selection frequencies of the categorical predictors still depend on their number of categories (cf., e.g., bottom row, left plot in Figure 1), and also the deviation of the permutation importance measure is still affected by the number of categories (cf., e.g., bottom

row, left plot in Figures 3 and 4).

Thus, there must be another source of bias, besides the variable selection bias in the individual trees, that affects the selection frequencies and the deviation of the permutation importance measure.

We show in the next section that this additional effect is due to bootstrap sampling with replacement, that is traditionally employed in random forests.

4.2. Effects induced by bootstrapping

From the comparison of left and right columns (representing sampling with and without replacement) in Figures 1 and 5 we learned that the variable selection frequencies in random forest functions are affected by the resampling scheme.

We found that, even when the `cforest` function based on unbiased classification trees is used, variables with more categories are preferred when bootstrap sampling is conducted with replacement, while no bias occurs when subsampling is conducted without replacement, as displayed in the bottom right plot in Figures 1 and 5. Thus, the bootstrap sampling induces an effect that is more pronounced for predictor variables with more categories.

For a better understanding of the underlying mechanism let us consider only the categorical predictor variables X_2 through X_5 with different numbers of categories from the null case simulation study design. Rather than trying to explain the effect of bootstrap sampling in the complex framework of random forests, we use a much simpler independence test for the explanation.

We consider the p values of χ^2 tests (computed from 1000 simulated data sets). In each simulation run, a χ^2 test is computed for each predictor variable and the binary response Y . Remember that the variables in the null case are not informative, i.e. the response is independent of all variables.

For independent variables the distribution of the p values of the χ^2 test is supposed to form a uniform distribution.

The left plot in Figure 11 displays the distribution of the p values of χ^2 tests from each predictor variable and the response Y as boxplots. We find that the boxplots range from 0 to 1 with median 0.5 as expected, because the p values of the χ^2 test form a uniform distribution when computed before bootstrapping.

However, if in each simulation run we draw a bootstrap sample from the original sample and then again compute the p values based on the bootstrap sample, we find that the distribution of the p values is shifted towards zero as displayed in the right plot in Figure 9.

Obviously, the bootstrap sampling artificially induces an association between the variables. This effect is more pronounced for the variables that have more categories.

The reason for this shift in the distribution of the p values is that each original sample, even if sampled from theoretically independent distributions, may show some minor variations from the null hypothesis of independence. These minor variations are aggravated by bootstrap sampling with replacement, because the cell counts in the contingency table are affected by observations that are either not included or are doubled or tripled in the bootstrap sample.

This effect is more pronounced for variables with more categories, because in larger tables (such as the 4×2 table from the cross-classification of X_3 and the binary response Y), the absolute cell counts are smaller than in smaller tables (such as the 2×2 table from the cross-classification of X_2 and the binary response Y). With respect to the smaller absolute cell counts, excluding or duplicating an observation produces more severe variations from the null hypothesis.

This effect is not eliminated if the sample size is increased, because in bootstrap sampling the size n of the original sample and the bootstrap sample size n increase simultaneously. However, if subsamples are drawn without replacement the effect disappears.

The apparent association that is induced by bootstrap sampling, and that is more pronounced for predictor variables with many categories, affects both variable importance measures: The selection frequency is again directly affected, and the permutation importance is affected because variables

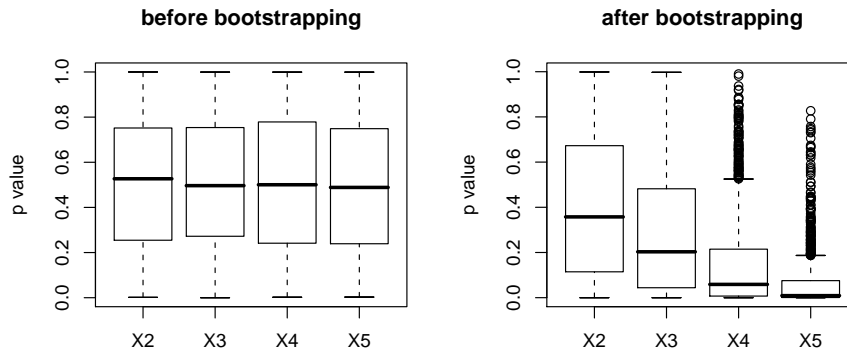


Figure 11: Distribution of the p values of χ^2 tests of each categorical variable X_2, \dots, X_5 and the binary response for the null case simulation study, where none of the predictor variables is informative. The left plots correspond to the distribution of the p values computed from the original sample before bootstrapping. The right plots correspond to the distribution of the p values computed for each variable from the bootstrap sample drawn with replacement.

with many categories gain positions closer to the root node in the individual trees. Together with the mechanisms described in the previous section, this explains all our findings.

From our simulation results we can also see that the effect of bootstrap sampling is mostly superposed by the much stronger effect of variable selection bias when comparing the conditions of sampling with and without replacement for the `randomForest` function only (cf. Figures 1 through 9, top row). Only when variable selection bias is removed by the `cforest` function the differences between the conditions of sampling with and without replacement are obvious (cf. Figures 1 through 9, bottom row).

We therefore conclude that in order to be able to reliably interpret the variable importance measures of a random forest, the forest must be built from unbiased classification trees, and sampling must be conducted without replacement.

5. Conclusions

Random forests are a powerful statistical tool, that has found many applicants in various scientific areas. It has been applied to such a wide variety of problems as large-scale association studies for complex genetic diseases, the prediction of phenotypes based on amino acid or DNA sequences, QSAR modelling and clinical medicine, to name just a few.

Features that have added to the popularity of random forests especially in bioinformatics and related fields, where identifying a subset of relevant predictor variables from very large sets of candidates is the major challenge, include its ability to deal with critical “small n large p ” data sets and the variable importance measures it provides for variable selection purposes.

However, when a method is used for variable selection, rather than prediction only, it is particularly important that the value and interpretation of the variable importance measure actually depict the importance of the variable, and are not affected by any other characteristics.

We found that for the original random forest method the variable importance measures are affected by the number of categories and scale level of the predictor variables, which are no indicators of the true importance of the variable. This finding is particularly relevant in studies where continuous variables like the folding energy are used in combination with categorical information like the

neighboring nucleotides, or when categorical predictors, as e.g. in amino acid sequence data, vary in their number of categories present in the sample.

Especially information on clinical and environmental variables is often gathered by means of questionnaires, where the number of categories can vary between questions. The number of categories is typically determined by many different influence factors and is not an indicator of variable importance. Similarly, the number of different categories of a predictor actually available in a certain sample is not necessarily an indicator of its relevance for predicting the response. Hence, the number of categories of a variable should obviously not influence its estimated importance—otherwise the results of a study could easily be affected if an irrelevant variable with many categories is included in the study design.

We showed that, due to variable selection bias in the individual classification trees and effects induced by bootstrap sampling, the variable importance measures of the `randomForest` function are not reliable in many scenarios relevant in applied research.

As an alternative random forest method we propose to use the `cforest` function, that provides unbiased variable selection in the individual classification trees. When this method is applied with subsampling without replacement the resulting variable importance measure can be used reliably for variable selection even in situations where the potential predictor variables vary in their scale level or their number of categories.

The aim of this paper is to explore the limits of the empirical measures of variable importance provided for random forests, to understand the underlying mechanisms and to use that understanding to guarantee unbiased and reliable variable selection in random forests.

In a more theoretical work [van der Laan \(2006\)](#) gives a fundamental definition of variable importance, as well as a statistical inference framework for estimating and testing variable importance. Inspired by this approach, future research on variable importance measures for variable selection with random forests aims at providing further means of statistical inference, that can be used to guide the decision on which and how many predictor variables to select in a certain problem.

Computational details

The results in this paper were obtained using R 2.3.1 ([R Development Core Team 2006](#)) and the packages `party` 0.9-3 ([Hothorn et al. 2006a](#)) `randomForest` 4.5-16 ([Liaw and Wiener 2002](#); [Breiman et al. 2006](#)) `rpart` 3.1-29 ([Therneau et al. 2006](#)).

Acknowledgements

CS was supported by the German Research Foundation (DFG), collaborative research center 386 “Statistical Analysis of Discrete Structures”. TH received financial support from DFG grant HO 3242/1-1. The authors would like to thank Thomas Augustin, Friedrich Leisch and Gerhard Tutz for fruitful discussions and for supporting our interest in this field of research.

References

- Arun K, Langmead CJ (2006). “Structure Based Chemical Shift Prediction Using Random Forests Non-linear Regression.” In “Proceedings of the Fourth Asia-Pacific Bioinformatics Conference (APBC),” pp. 317–326.
- Boulesteix AL (2006a). “Maximally Selected Chi-square Statistics and Binary Splits of Nominal Variables.” *Biometrical Journal*. Accepted for publication.
- Boulesteix AL (2006b). “Maximally Selected Chi-square Statistics for Ordinal Variables.” *Biometrical Journal*, **48**, 451–462.

- Breiman L (2001). “Random Forests.” *Machine Learning*, **45**(1), 5–32.
- Breiman L, Cutler A, Liaw A, Wiener M (2006). *Breiman and Cutler’s Random Forests for Classification and Regression*. R package version 4.5-16, URL <http://CRAN.R-project.org/>.
- Breiman L, Friedman JH, Olshen RA, Stone CJ (1984). *Classification and Regression Trees*. Chapman and Hall, New York.
- Bureau A, Dupuis J, Falls K, Lunetta KL, Hayward B, Keith TP, Eerdewegh PV (2005). “Identifying SNPs Predictive of Phenotype Using Random Forests.” *Genetic Epidemiology*, **28**, 171–182.
- Cummings MP, Myers DS (2004). “Simple Statistical Models Predict C-to-U Edited Sites in Plant Mitochondrial RNA.” *BMC Bioinformatics*, **5**, 132.
- Cummings MP, Segal MR (2004). “Few Amino Acid Positions in rpoB are Associated with Most of the Rifampin Resistance in Mycobacterium Tuberculosis.” *BMC Bioinformatics*, **5**, 137.
- Díaz-Uriarte R, de Andrés SA (2006). “Gene Selection and Classification of Microarray Data Using Random Forest.” *BMC Bioinformatics*, **7**, 3.
- Dobra A, Gehrke J (2001). “Bias Correction in Classification Tree Construction.” In CE Brodley, AP Danyluk (eds.), “Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA,” pp. 90–97. Morgan Kaufmann.
- Friedman J (2001). “Greedy Function Approximation: A Gradient Boosting Machine.” *The Annals of Statistics*, **29**(5), 1189–1232.
- Furlanello C, Neteler M, Merler S, Menegon S, Fontanari S, Donini D, Rizzoli A, Chemini C (2003). “GIS and the Random Forest Predictor: Integration in R for Tick-Borne Disease Risk Assessment.” In K Hornik, F Leisch, A Zeileis (eds.), “Proceedings of the 3rd International Workshop on Distributed Statistical Computing, Vienna, Austria,” ISSN 1609-395X, URL <http://www.ci.tuwien.ac.at/Conferences/DSC-2003/Proceedings/>.
- Guha R, Jurs PC (2003). “Development of Linear, Ensemble, and Nonlinear Models for the Prediction and Interpretation of the Biological Activity of a Set of PDGFR Inhibitors.” *Journal of Chemical Information and Computer Sciences*, **44**, 2179–2189.
- Gunther EC, Stone DJ, Gerwien RW, Bento P, Heyes MP (2003). “Prediction of Clinical Drug Efficacy by Classification of Drug-induced Genomic Expression Profiles *in vitro*.” *Proceedings of the National Academy of Sciences*, **100**, 9608–9613.
- Heidema AG, Boer JMA, Nagelkerke N, Mariman ECM, van der A DL, Feskens EJM (2006). “The Challenge for Genetic Epidemiologists: How to Analyze Large Numbers of SNPs in Relation to Complex Diseases.” *BMC Genetics*, **7**, 23.
- Hothorn T, Hornik K, Zeileis A (2006a). *party: A Laboratory for Recursive Part(y)itioning*. R package version 0.9-0, URL <http://CRAN.R-project.org/>.
- Hothorn T, Hornik K, Zeileis A (2006b). “Unbiased Recursive Partitioning: A Conditional Inference Framework.” *Journal of Computational and Graphical Statistics*, **15**(3), 651–674.
- Huang X, Pan W, Grindle S, Han X, Chen Y, Park SJ, Miller LW, Hall J (2005). “A Comparative Study of Discriminating Human Heart Failure Etiology Using Gene Expression Profiles.” *BMC Bioinformatics*, **6**, 205.
- Kim H, Loh W (2001). “Classification Trees with Unbiased Multiway Splits.” *Journal of the American Statistical Association*, **96**, 589–604.

- Kononenko I (1995). “On Biases in Estimating Multi-Valued Attributes.” In “Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence,” pp. 1034–1040.
- Liaw A, Wiener M (2002). “Classification and Regression by **randomForest**.” *R News*, **2**(3), 18–22. URL <http://CRAN.R-project.org/doc/Rnews/>.
- Lunetta KL, Hayward LB, Segal J, Eerdewegh PV (2004). “Screening Large-Scale Association Study Data: Exploiting Interactions Using Random Forests.” *BMC Genetics*, **5**, 32.
- Qi Y, Bar-Joseph Z, Klein-Seetharaman J (2006). “Evaluation of Different Biological Data and Computational Classification Methods for Use in Protein Interaction Prediction.” *Proteins*, **63**, 490–500.
- R Development Core Team (2006). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org/>.
- Segal MR, Barbour JD, Grant RM (2004). “Relating HIV-1 Sequence Variation to Replication Capacity via Trees and Forests.” *Statistical Applications in Genetics and Molecular Biology*, **3**, 2.
- Shih Y (2005). “Tumor Classification by Tissue Microarray Profiling: Random Forest Clustering Applied to Renal Cell Carcinoma.” *Modern Pathology*, **18**, 547–557.
- Strobl C (2005a). “Statistical Sources of Variable Selection Bias in Classification Tree Algorithms Based on the Gini Index.” *Discussion Paper 420*, SFB “Statistical Analysis of Discrete Structures”, Munich, Germany. URL <http://www.stat.uni-muenchen.de/sfb386/papers/dsp/paper420.ps>.
- Strobl C (2005b). “Variable Selection in Classification Trees Based on Imprecise Probabilities.” In F Cozman, R Nau, T Seidenfeld (eds.), “Proceedings of the Fourth International Symposium on Imprecise Probabilities and Their Applications, CMU Pittsburgh,” pp. 340–348. SIPTA, Manno.
- Strobl C, Boulesteix AL, Augustin T (2005). “Unbiased Split Selection for Classification Trees Based on the Gini Index.” *Discussion Paper 464*, SFB “Statistical Analysis of Discrete Structures”, Munich, Germany. URL <http://www.stat.uni-muenchen.de/sfb386/papers/dsp/paper464.pdf>.
- Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP (2003). “Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling.” *Journal of Chemical Information and Computer Sciences*, **43**, 1947–1958.
- Therneau TM, Atkinson B, Ripley BD (2006). **rpart**: *Recursive Partitioning*. R package version 3.1-30, URL <http://CRAN.R-project.org/>.
- van der Laan M (2006). “Statistical Inference for Variable Importance.” *International Journal of Biostatistics*, **2**(1), 1008–1008.
- Ward MM, Pajevic S, Dreyfuss J, Malley JD (2006). “Short-Term Prediction of Mortality in Patients with Systemic Lupus Erythematosus: Classification of Outcomes Using Random Forests.” *Arthritis and Rheumatism*, **55**, 74–80.