LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
INSTITUT FÜR STATISTIK

Silke Janitza

On the overestimation of random forest's out-of-bag error

Technical Report Number 204, 2017
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

On the overestimation of random forest's out-of-bag error

Silke Janitza

April 10, 2017

Department of Medical Informatics, Biometry and Epidemiology, University of Munich, Marchioninistr. 15, D-81377 Munich, Germany.

Abstract

Background The ensemble method random forests has become a popular classification tool in bioinformatics and related fields. The out-of-bag error is an error estimation technique which is often used to evaluate the accuracy of a random forest as well as for selecting appropriate values for tuning parameters, such as the number of candidate predictors that are randomly drawn for a split, referred to as mtry. However, for binary classification problems with metric predictors it was shown that the out-of-bag error overestimates the true prediction error. Based on simulated and real data this paper aims to identify settings for which the overestimation is likely. Moreover, the overestimation was shown to depend on the parameter mtry. Therefore, it is questionable if the out-of-bag error can be used in classification tasks for selecting tuning parameters like mtry.

Results The simulation-based and real-data based studies with metric predictor variables show that the overestimation is largest in balanced settings and in settings with few observations, a large number of predictor variables, small correlations between predictors and weak effects. There was hardly any impact of the overestimation on tuning parameter selection.

Conclusions Although the prediction performance of random forests was not substantially affected when using the out-of-bag error for tuning parameter selection in the present studies, one cannot be sure that this applies to all future data. For settings with metric predictor variables it is therefore recommended to always use stratified subsampling for both tuning parameter selection and error estimation in random forests. This yielded less biased estimates of the true prediction error.

Keywords: Random forests, OOB error, Out-of-bag, Parameter tuning, Error estimation.

1 Introduction

Random forests (RF) (Breiman; 2001) have become a popular classification tool in bioinformatics and related fields. They have shown excellent performance also in very complex data settings. Each tree in a RF is constructed based on a random sample of the observations, usually a bootstrap sample or a subsample of the original data. The observations that are not part of the bootstrap sample or subsample, respectively, are referred to as out-of-bag (OOB) observations. The OOB observations can be used for example for estimating the prediction error of RF, yielding the so-called OOB error. The OOB error is often used for assessing the prediction performance of RF. An advantage of the OOB error is that the complete original sample can be used for constructing the RF classifier. In contrast to cross-validation and related data splitting procedures, in which a subset of the samples are left out for RF construction, the OOB procedure enables using all information that is available in the data for classifier construction. This yields RF classifiers that have a higher accuracy than those obtained from cross-validation and related procedures. Another advantage of using the OOB error is its computational speed. In contrast to cross-validation or other data splitting approaches, only one RF has to be constructed, while for k-fold cross-validation k RF have to be constructed (Bylander; 2002; Zhang et al.; 2010). The use of the OOB error saves memory and computation time, especially when dealing with large data dimensions, where constructing a single RF might last several days or even weeks. These reasons might explain the frequent use of the OOB error for error estimation and tuning parameter selection in RF.

The OOB error is often claimed to be an unbiased estimator for the true error rate (Breiman; 2001; Goldstein et al.; 2011; Zhang et al.; 2010). However, for two-class classification problems it was reported that the OOB error overestimates the true prediction error (Bylander; 2002; Mitchell; 2011). The bias can be very substantial, as shown in these papers, and is also present when using classical cross-validation procedures for error estimation. It was thus recommended to use the OOB error only as an upper bound for the true prediction error (Mitchell; 2011). However, Mitchell (2011) considered only settings with completely balanced samples, sample sizes below 60 and two response classes, limiting the generality of the results.

Besides the fact that trees in RF are constructed on a random sample of the data, there is a second component which differs between standard classification and regression trees and the trees in RF. In the trees of a RF, not all variables but only a subset of the variables are considered for each split. This subset is randomly drawn from all candidate predictors at each split. The size of this subset is usually referred to as *mtry*. In practical applications, the most common approach for choosing appropriate values for *mtry* is to select the value over a grid of plausible values which yields the smallest OOB error (Oliveira et al.; 2012; Hassane et al.; 2008; Nicodemus et al.; 2010). Also in works on RF methodology, the OOB error has frequently been used to choose an appropriate value for *mtry* (Nicodemus and Malley; 2009; Kim et al.; 2006). In principle, other procedures like (repeated) cross-validation may be applied for selecting an optimal value for *mtry*, but the OOB error is usually the first choice for parameter tuning. This is due to the fact that, unlike many other approaches such as cross-validation, the whole data can be used to construct the RF and much computational effort is saved since only one RF has to be built for each candidate *mtry* value. Implementations exist that use the OOB error to select an appropriate value for *mtry*. In the statistical software R (R Core Team; 2013), for example, the function `tuneRF` (from the package randomForest; Liaw and Wiener; 2002) automatically searches over a grid of *mtry* values and selects the value for *mtry* for which the OOB error is smallest. However, the bias in the OOB

error has recently been shown to depend on the parameter *mtry* (Mitchell; 2011). This finding suggests that the *mtry* value minimizing the OOB error might possibly not minimize the true prediction error and thus may be suboptimal, making the OOB error based tuning approaches questionable. To date there are no studies investigating the reliability of the OOB error for tuning parameters like *mtry* in RF.

The contribution of this paper is three-fold: (i) the bias and its dependence on *mtry* in settings with metric predictor variables are quantitatively assessed through studies with different numbers of observations, predictors and response classes which helps to identify so-called "high-risk settings", (ii) the reasons for this bias and its dependence on *mtry* are studied in detail, and based on these findings, the use of alternatives, such as stratified sampling, are investigated, and (iii) the consequences of the bias for tuning parameter selection are explored.

This paper is structured as follows: In Section 2, simulation-based and real-data based studies are described after briefly introducing the RF method. The description includes an outline of the simulated data and real data, the considered settings and several different error estimation techniques that will be used. The results of the studies are subsequently reported in Section 3, and finally the findings are discussed and recommendations are given.

2 Methods

In this section, the RF method and the simulation-based and real data-based studies are described. Simulated data is used to study the behavior of the OOB error in simple settings, in which for example all predictors are uncorrelated. This shall give an insight into the mechanisms which lead to the bias in the OOB error. Based on these results, settings are identified, in which a bias in the OOB error is likely. To assess the extent of the bias in these settings in practice, complex data from the real world is used.

2.1 Random forests and its out-of-bag error

RF is an ensemble of classification or regression trees that was introduced by Breiman (2001). One of the two random components in RF concerns the choice of variables used for splitting. For each split in a tree, the best splitting variable from a random sample of *mtry* predictors is selected. If *mtry* is chosen too small, it might be that none of the variables contained in the subset is relevant and that irrelevant variables are often selected for a split. The resulting trees have poor predictive ability. If the subset contains a large number of predictors, in contrast, it is likely that the same variables, namely those with the largest effect, are often selected for a split, and that variables with smaller effects have hardly any chance of being selected. Therefore, *mtry* should be considered a tuning parameter.

The other random component in RF concerns the choice of training observations for a tree. Each tree in RF is built from a random sample of the data. This is usually a bootstrap sample or a subsample of size $0.632n$. Therefore not all observations are used to construct a specific tree. The observations that are not used to construct a tree are denoted by *out-of-bag (OOB) observations*. In a RF, each tree is built from a different sample of the original data, so each observation is "out-of-bag" for some of the trees. The prediction for an observation can then be obtained by using only those trees for which the observation was not used for the construction. A classification for each observation is obtained in this way and the error rate can be estimated

from these predictions. The resulting error rate is referred to as *OOB error*. This procedure was originally introduced by Breiman (1996b) and it has become an established method for error estimation in RF.

2.2 Simulation-based studies

The overestimation of the OOB error in different data settings with metric predictor variables was systematically investigated by means of simulation studies. Settings were considered with

- different associations between the predictors and the response. Either none of the predictors were associated with the response (the corresponding studies termed *null case*) or some of them were associated (*power case*);
- different numbers of predictors, $p \in \{10, 100, 1000\}$;
- different numbers of response classes, $k \in \{2, 4\}$. The studies are termed *binary* if $k = 2$ and *multiclass* if $k = 4$;
- different response class ratios. An equal number of observations of each response class was used (*balanced settings*) for $k \in \{2, 4\}$. For $k = 2$ two additional settings with unequal response class sizes were simulated (*binary unbalanced* and *binary extremely unbalanced*). In the first setting (*binary unbalanced*), the smaller class comprised 30% of the observations. In the second setting (*binary extremely unbalanced*), the smaller class comprised approximately 17% (ratio 1:5) of the observations.
- different numbers of observations, $n \in \{n_{small}, 100, 1000\}$, with $n_{small} = 20$ for *binary balanced* studies, $n_{small} = 30$ for *binary unbalanced* studies, $n_{small} = 60$ for *binary extremely unbalanced* studies and $n_{small} = 40$ for *multiclass balanced* studies.

Since one of the aims was to investigate the bias in dependence on *mtry*, several RFs with different *mtry* values were constructed for each setting. The grid of considered *mtry* values was $\{1, 2, 3, \ldots, 10\}$ for $p = 10$, $\{1, 10, 20, 30, \ldots, 100\}$ for $p = 100$ and $\{1, 5, 10, 50, 100, 200, 300, \ldots, 1000\}$ for $p = 1000$. Note that for $mtry = 1$ there is no selection of an optimal predictor variable for a split, while for $mtry = p$ the RF method coincides with the bagging procedure which selects the best predictor variable from all available predictors (Breiman; 1996a). The number of trees, usually referred to *ntree*, should be chosen very large, especially if the data comprises a large number of predictors. It is usually chosen as a compromise between accuracy and computational speed. The OOB error stabilized at around 250 trees in convergence studies of Goldstein et al. (2010), and they concluded that 1000 trees might be sufficiently large for their genome-wide data set. Also in the studies of Díaz-Uriarte and De Andres (2006) the results for RF with 1000 trees were almost the same as those for RF with 40000 trees, and in the high-dimensional settings of Genuer et al. (2008) RF with 500 trees and 1000 trees yielded very similar OOB errors. In accordance with these findings the number of trees was set to 1000 in all studies of this paper (including at most ~ 7000 predictors). Each setting was repeated 500 times to obtain stable results.

Only metric predictor variables were considered in the studies. In the null case study, the predictors $X_1, \ldots, X_p$ were independent and identically distributed (*i.i.d.*), each following a standard normal distribution (see Tables 1 and 2). In the power case study, both predictors associated with the response and predictors not associated with the response were considered. The predictors not

associated with the response followed a standard normal distribution. The distribution of predictors with association was different for the different response classes. The predictor values for observations from class 1 were always drawn from a standard normal distribution. The predictor values for observations from class 2 (or classes 2, 3, and 4 in settings with $k = 4$ response classes) were drawn from a normal distribution with variance 1 and a mean different from zero. Tables 1 and 2 give an overview over the distribution of predictors in the response classes for settings with $k = 2$ and $k = 4$ response classes, respectively. Let us consider the setting with $p = 10$ and $k = 4$ as an example (Table 2). The first two predictors X_1 and X_2 are associated with the response, while the other predictors $X_3, \ldots, X_{10}$ are not. Accordingly, $X_3, \ldots, X_{10}$ always follow a standard normal distribution, while the distribution of X_1 and X_2 depends on whether the observation comes from class 1 or a different class. If the observation comes from class 1 the distribution of X_1 and X_2 is $N(0, 1)$, and if it comes from class $r \in \{2, 3, 4\}$ the variables $X_j, j = 1, 2$ follow a normal distribution $N(\mu_{rj}, 1)$ with μ_{rj} drawn independently from $N(0.4, 1)$. Randomly drawing the mean separately for X_1 and X_2 and for each repetition of the study makes sure that predictors with different effect strengths are considered.

Study	No. predictors	Predictors	Class 1 $N(\mu_1, 1)$	Class 2 $N(\mu_2, 1)$
Null case	$p \in \{10, 100, 1000\}$	$X_1, \ldots, X_p$	$\mu_1 = 0$	$\mu_2 = 0$
Power case	$p = 10$	X_1	$\mu_1 = 0$	$\mu_2 \sim N(0.75, 1)$
		X_2	$\mu_1 = 0$	$\mu_2 \sim N(0.75, 1)$
		$X_3, \ldots, X_{10}$	$\mu_1 = 0$	$\mu_2 = 0$
	$p = 100$	X_1	$\mu_1 = 0$	$\mu_2 \sim N(0.75, 1)$
		$\vdots$	$\vdots$	$\vdots$
		X_{10}	$\mu_1 = 0$	$\mu_2 \sim N(0.75, 1)$
		$X_{11}, \ldots, X_{100}$	$\mu_1 = 0$	$\mu_2 = 0$
	$p = 1000$	X_1	$\mu_1 = 0$	$\mu_2 \sim N(0.1, 1)$
		$\vdots$	$\vdots$	$\vdots$
		X_{50}	$\mu_1 = 0$	$\mu_2 \sim N(0.1, 1)$
		$X_{51}, \ldots, X_{1000}$	$\mu_1 = 0$	$\mu_2 = 0$

Table 1: Distribution of predictors in class 1 and class 2 of the simulated data setting with $k = 2$ response classes.

Study	No. predictors	Predictors	Class 1 $N(\mu_1, 1)$	Class 2 $N(\mu_2, 1)$	Class 3 $N(\mu_3, 1)$	Class 4 $N(\mu_4, 1)$
Null case	$p \in \{10, 100, 1000\}$	$X_1, \ldots, X_p$	$\mu_1 = 0$	$\mu_2 = 0$	$\mu_3 = 0$	$\mu_4 = 0$
Power case	$p = 10$	X_1	$\mu_1 = 0$	$\mu_2 \sim N(0.4, 1)$	$\mu_3 \sim N(0.4, 1)$	$\mu_4 \sim N(0.4, 1)$
		X_2	$\mu_1 = 0$	$\mu_2 \sim N(0.4, 1)$	$\mu_3 \sim N(0.4, 1)$	$\mu_4 \sim N(0.4, 1)$
		$X_3, \ldots, X_{10}$	$\mu_1 = 0$	$\mu_2 = 0$	$\mu_3 = 0$	$\mu_4 = 0$
	$p = 100$	X_1	$\mu_1 = 0$	$\mu_2 \sim N(0.4, 1)$	$\mu_3 \sim N(0.4, 1)$	$\mu_4 \sim N(0.4, 1)$
		$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
		X_{10}	$\mu_1 = 0$	$\mu_2 \sim N(0.4, 1)$	$\mu_3 \sim N(0.4, 1)$	$\mu_4 \sim N(0.4, 1)$
		$X_{11}, \ldots, X_{100}$	$\mu_1 = 0$	$\mu_2 = 0$	$\mu_3 = 0$	$\mu_4 = 0$
	$p = 1000$	X_1	$\mu_1 = 0$	$\mu_2 \sim N(0.4, 1)$	$\mu_3 \sim N(0.4, 1)$	$\mu_4 \sim N(0.4, 1)$
		$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
		X_{50}	$\mu_1 = 0$	$\mu_2 \sim N(0.4, 1)$	$\mu_3 \sim N(0.4, 1)$	$\mu_4 \sim N(0.4, 1)$
		$X_{51}, \ldots, X_{1000}$	$\mu_1 = 0$	$\mu_2 = 0$	$\mu_3 = 0$	$\mu_4 = 0$

Table 2: Distribution of predictors in class 1, class 2, class 3 and class 4 of the simulated data setting with $k = 4$ response classes.

Despite considering metric predictors with different effects strength, the settings are simplistic because all predictors are uncorrelated. Although assuming no correlations between any of the predictors is not realistic, such settings are important to understand the mechanisms which lead

to a bias in the OOB error. The OOB error in more complex settings that include correlated predictors will be explored by means of real data.

2.3 Real data-based studies

Based on the results from simulated data, real data sets were to be considered in which the overestimation of the OOB error is expected to be most pronounced. As will be seen later, the OOB error is likely to occur in data settings with huge numbers of predictors, p, and small numbers of observations, n. Such settings are typically prevalent with genomic data. Therefore high-dimensional genomic data from the real world are considered for further investigations.

New data for evaluation can easily be generated in simulated data. In contrast to that, in real data applications, the original data has to be split up in order to obtain an independent test data set used for evaluation. Thus, six genomic datasets were selected that are large enough to randomly split the data into a training and a test set (Table 3). These datasets were often used by various authors for classification purposes (Díaz-Uriarte and De Andres; 2006; Dettling and Bühlmann; 2003; Tan and Gilbert; 2003) and are briefly described in the following. Note that no pre-selection of data sets based on the results obtained for this data was performed, and the results of all six datasets which were analyzed are reported (Boulesteix; 2015).

Dataset	No. response classes, k	No. predictors, p	Considered *mtry* values	Size of original data
Colon Cancer	2	2000	$\{1, 10, 100, 500, 1000, 2000\}$	62
Breast Cancer	3	4869	$\{1, 10, 100, 500, 1000, 2000, 3000, 4000, 4869\}$	95
Breast Cancer	2	4869	$\{1, 10, 100, 500, 1000, 2000, 3000, 4000, 4869\}$	77
Prostate Cancer	2	6033	$\{1, 10, 100, 500, 1000, 2000, 3000, 4000, 5000, 6033\}$	102
Embryonal Tumor	2	7129	$\{1, 10, 100, 500, 1000, 2000, 3000, 4000, 5000, 6000, 7129\}$	60
Leukemia	2	7129	$\{1, 10, 100, 500, 1000, 2000, 3000, 4000, 5000, 6000, 7129\}$	72

Table 3: Overview over high-dimensional genomic datasets.

2.3.1 Data

The first considered data is the *Colon Cancer data* of Alon et al. (1999). The expression levels of 40 tumor and 22 normal colon tissues for 6500 human genes were measured. The considered data set contains the expression of the 2000 genes with highest minimal intensity across the 62 tissues measured using the Affymetrix technology.

Two versions of the *Breast Cancer data* of van't Veer et al. (2002) were considered. The first version of this data was previously analyzed by Díaz-Uriarte and De Andres (2006) and contains $k = 3$ response classes: 33 patients developed distant metastases within 5 years, 44 remained disease-free for over 5 years and 18 patients had BRCA1 germline mutations. Missing data was imputed by using 5-nearest neighbor imputation. Further details on transformations of the original data are given in the supplement to the paper of Díaz-Uriarte and De Andres (2006). The second version which is considered in this paper is a subset of the dataset provided by Díaz-Uriarte and De Andres (2006). This subset does not contain the 18 patients with BRCA1 germline mutations. A differentiation is thus only made between the patients that developed distant metastases within 5 years ($n = 33$) and patients that remained disease-free for over 5 years ($n = 44$), that is the number of response classes is $k = 2$.

The fourth considered data set is the *Prostate Cancer data* of Singh et al. (2002). From 1995 to 1997 samples of prostate tumors and adjacent non-tumor prostate tissue were collected from

patients undergoing radical prostatectomy at the Brigham and Women's Hospital. High-quality expression profiles were obtained from 50 non-tumor prostate samples and 52 tumor specimens. The oligonucleotide microarrays contained probes for approximately 12600 genes.

The *Embryonal Tumor data* of Pomeroy et al. (2002) includes 60 patients with embryonal tumors of the central nervous system from whom biopsies were obtained before receiving treatment. The data was used to differentiate between patients who are alive after treatment ($n = 21$) and those who succumbed to their disease ($n = 39$) (dataset C in Pomeroy et al.; 2002). RNA was extracted from frozen specimens and was analyzed with oligonucleotide microarrays containing 7129 probes from 6817 genes.

The *Leukemia data* (Golub et al.; 1999) consists of 47 patients with acute lymphoblastic leukemia (ALL) and 25 patients with acute myeloid leukemia (AML). The considered dataset comprises both, training samples and test samples from Golub et al. (1999). The samples were assayed using Affymetrix Hgu6800 chips and data on the expression of 7129 genes are available.

The Colon Cancer data, the Prostate Cancer data and both Breast Cancer data sets were obtained from the website `http://ligarto.org/rdiaz/Papers/rfVS/randomForestVarSel.html`. The Embryonal Tumor data is available at `http://portals.broadinstitute.org/cgi-bin/cancer/datasets.cgi`, and the Leukemia data was retrieved from the Bioconductor package golubEsets.

2.3.2 Settings

Different settings were investigated which were created by modifying the original real data sets. The aims together with the modifications are outlined in the following:

Aim 1: To quantitatively assess the overestimation in the OOB error and its consequences for selecting an optimal value for *mtry* on real world datasets. For this purpose, the original data was used without making any modifications to the data. This study is referred to as "Real data study".

Aim 2: To investigate the behavior of the OOB error on datasets with realistic data structures but without any associations between the predictors and the response. To create a data set with realistic data structures, the matrix containing the values of the predictor variables of the real data sets was used and the response values of the original datasets were randomly permuted to break any associations between the predictors and the response. The studies with the permuted response are termed "Real data null case study with correlations", where the term correlation refers to the correlations between the predictor variables. Note that the datasets obtained in this way only differ to the original data in that none of the predictors is associated with the response, while in the original data some of the predictors are possibly associated.

Aim 3: To investigate the effect of correlations on the bias in the OOB error in realistic data settings. For this purpose, each predictor variable was permuted separately to create independence between them. This also breaks possible associations between the predictors and the response. This setting is called "Real data null case study without correlations". Note that, in order to assess the effect of correlations, the results for this study cannot be compared to the results for the *real data study* (described above) because in the *real data study* some of the predictors are possibly associated with the response, while in the *real data null case*

study without correlations this is not the case. This makes it impossible to decide whether differences are due to the correlations between predictors or are due to the fact that some of the predictors are associated in one study but not in the other. However, the results of the *real data null case study without correlations* can be compared to those of the *real data null case study with correlations*, in which there are correlations between predictors but none of the predictors is associated with the response.

Only a part of the observations was used to construct the RF (training set) while the other part was used for assessing the performance of the RF (test set). The number of trees was always set to 1000. For the datasets with $k = 2$ response classes, the training set consists of $n = 20$ observations that were randomly drawn, and for the Breast Cancer data (i.e. the only dataset with $k = 3$), the training set consists of $n = 30$ observations. In contrast to the simulation studies, the response class ratio in the training set was not fixed. However, a minimum of 8 observations were required from each response class to prevent that there are too few observations from a response class. With only $n = 20$ observations, this means that the response class distribution is nearly balanced and that only slight class imbalances can occur in the considered settings. Note that we chose to use only 20 and 30 observations, respectively, to train RF since these are settings in which a bias in the OOB error is most likely, as will be shown in the rest of this paper. Although modern studies include far more observations, such small sample sizes are still encountered in practice (Floares et al.; 2016).

For all settings RF with different *mtry* values were constructed. The grid of *mtry* values was $\{1, 10, 100, 500, 1000, 2000, \ldots, p\}$, with p denoting the total number of predictors. Table 3 shows the grids for the considered datasets. Each setting was repeated 1000 times.

2.4 Alternative strategies for error estimation

The following strategies for error estimation were considered as possible alternatives to the OOB error:

- *Test error:* This error rate was computed for observations that are not part of the set of n observations that were used to construct a RF. Since these observations are usually referred to as test observations, the resulting error rate is referred to as test error. In the simulation studies, data for 10000 additional observations (test observations) was generated in order to estimate the prediction error of the RF. The response class distributions were the same in the two samples of size 10000 and n. In the real data studies, the n observations that were used to construct the RF ($n = 20$ for $k = 2$, $n = 30$ for $k = 3$) were randomly sampled from all available observations, while making sure that at least 8 observations from each response class were sampled. In order to have the same response class distribution in the two sets, as test set the largest subset of the remaining observations was used in which the response class distribution equals that in the sample of n observations.

- *Stratified OOB error:* In this paper, the OOB error was also computed for a RF which is based on a stratified sampling scheme. This strategy was also investigated in the studies of Mitchell (2011). In this stratified sampling scheme, trees were grown on subsamples of size $\lfloor 0.632n \rfloor$, in which the response class distribution of the original data of n observations is preserved in each subsample. The OOB error was computed based on the OOB observations as usual. In this paper, it is referred to as the stratified OOB error. Note that, in contrast

to the test error, the (stratified) OOB error uses the n observations for both constructing the RF and estimating its prediction error.

- *Cross-validation (CV) error:* In contrast to the OOB error, CV is a strategy for estimating the error rate of an arbitrary classification method and is not specific to RF. To estimate the CV error of a RF, the data of size n was partitioned into l sets of equal size. Each of the l sets was used for computing the error rate of a RF, while the other $l-1$ sets were used for creating the RF. The CV error was computed as the average of the l error rates. Ten-fold cross-validation was used in all studies (i.e., $l = 10$). While the test and OOB error (stratified and unstratified) estimation strategies use all of the n observations to construct the RF, in cross-validation the n observations are split into a training and a test set and only the $n(l-1)/l$ training observations are used to construct a RF. This means that the CV error is computed from l models that are fit based on only a subset of the data. Thus, the CV error slightly overestimates the true prediction error that would be obtained for a model that was fit based on all n observations (Kohavi; 1995).

- *Stratified cross-validation (CV) error:* For computing the stratified CV error the data of size n was randomly split into $l = 10$ sets in a way, that within each set the distribution of response classes is the same as in the original data. The error estimation was then done in exactly the same way as was described for the CV error.

Since the test error is an accurate estimate for the generalization error, it is treated as a "gold standard" in this paper against which the OOB and CV errors (stratified and unstratified) are compared. In simulation studies, one should therefore prefer estimating the error rate by means of an additional large independent test sample. In real data settings, in contrast, the number of observations is limited and is usually not sufficient to enable splitting the data into a training set and a large test set. Moreover, sample sizes are rather small and it is often desired to use all available information for building a model which has high predictive ability. Thus, in real data applications it is rarely the case that there is a large test set available from which the error rate can be computed, and different approaches to estimating the error rate, such as cross-validation procedures, have to be applied.

2.5 Random forest implementation and computational issues

Several implementations of the original RF version (Breiman; 2001) are available for different softwares (see Boulesteix et al.; 2012, for an overview). However, it was shown that split selection is biased in the original RF version (Kim and Loh; 2001; Strobl et al.; 2007; Nicodemus; 2011; Boulesteix et al.; 2012a). Certain types of predictors, such as predictors that offer many possible cutpoints, are preferentially selected for a split. For example, categorical variables with many categories are preferentially selected over metric variables or categorical variables with fewer categories. The RF version of Hothorn et al. (2006) implements an unbiased split selection that is based on conditional inference tests. In contrast to the original version, the selection of the variable for splitting and the selection of the cutpoint are performed in two different steps. In the first step, each variable is globally tested for its association with the response, yielding a p-value for each variable. The variable with the smallest p-value is selected for splitting. The optimal cutpoint within the variable is then chosen based on a special two-sample-test for all possible binary split points within the variable. This procedure avoids a preferential selection of variables

that offer many cutpoints (Strobl et al.; 2007). Although this version implements an unbiased split selection, the original version of Breiman is far more often used in practice and there is an ongoing development of even faster implementations of this version (see Wright and Ziegler; 2017, and references therein). In the context of computing time, the RF implementation of Breiman by far outperforms the (unbiased) RF implementation of Hothorn et al. Due to its frequent use in practice and its technical advantages the original RF version of Breiman and its implementation in the statistical software R (Liaw and Wiener; 2002) was used for all the studies. For some simulation settings, however, the RF version of Hothorn et al. implemented in the R package party (Hothorn et al.; 2012) was also used to make sure that the results do not depend on the choice of the RF version.

Note that the studies shown in this paper include only metric predictor variables. The inclusion of only metric predictor variables avoids affecting results by the biased split selection in the classical RF implementation. Moreover, subsampling (i.e., sampling from the original data without replacement) was used instead of bootstrapping in order to avoid possible biases induced by the bootstrap (Strobl et al.; 2007). As was suggested subsamples of size $\lfloor 0.632n \rfloor$ were used, where n denotes the number of observations (Strobl et al.; 2007). No pruning was applied and trees were always grown to full size.

3 Results

Figures 1 - 5 show the estimated error rates over a grid of *mtry* values for the five different error estimates (test error, OOB error, stratified OOB error, CV error, stratified CV error). In the following the bias in the OOB error is quantified based on these results. Further the sources of the bias and the dependence of this bias on RF parameters and data characteristics are investigated, and finally the consequences of using the OOB error for tuning *mtry* are assessed.

3.1 Quantitative assessment of the bias

For the *binary null case study (balanced)* the true error rate for new observations is 0.5, given that new observations come from both response classes equally often. Figure 1 shows the estimated error rates for the *binary null case study (balanced)*. The test error approximates 0.5 very well in all balanced settings and for all considered *mtry* values. For small sample sizes ($n = 20$), the OOB error is larger than the test error which is considered to be a good estimate of the true prediction error. For larger sample sizes ($n = 100$), the difference between the test error and the OOB error is smaller but still present. Finally, if the sample size is increased to $n = 1000$, the OOB error seems to approximate the test error well. When comparing the results for different parameter settings, it can be seen that the overestimation does not only depend on the number of observations but also on the number of predictors, or rather the ratio of the number of observations and predictors. In settings with both, large predictor numbers and small sample sizes (here $n = 20, p = 1000$), the overestimation is most extreme. Depending on the chosen value for *mtry*, the difference between the OOB error and the test error lies between 10% and 30% in this setting. In contrast to that, there is no overestimation in settings with large sample sizes and small predictor numbers ($n = 1000, p = 10$). Exactly the same results are obtained for the CV error. This suggests that CV is not a reasonable alternative to the OOB error and, moreover, that there is a common source of the overestimation. In contrast to the OOB error and the CV error, the stratified OOB error

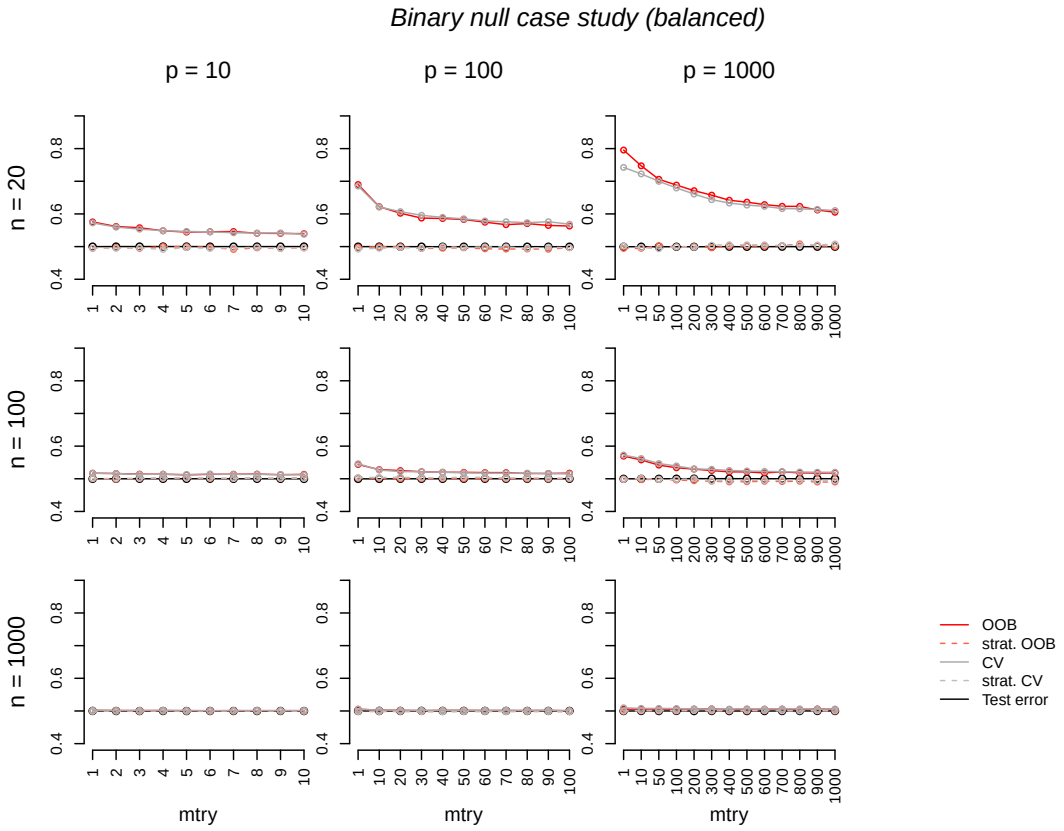

Figure 1: Error rate estimates for the *binary null case study (balanced)*. Shown are different error rate estimates for the setting with two response classes of equal size and without any predictors with effect. The error rate was estimated through the test error, the OOB error, the stratified OOB error, the CV error, and the stratified CV error for settings with different sample sizes, n, and numbers of predictors, p. The mean error rate over 500 repetitions was obtained for a range of *mtry* values.

and the stratified CV error approximate the test error very well and are a reasonable alternative to the unstratified sampling procedures in the considered study.

Comparable results were obtained for the *binary power case study (balanced)*. However, the difference between the OOB error (CV error) and the test error is smaller than in the study without any associations. In particular, there is only a small overestimation for large *mtry* values. Moreover, in contrast to the *binary null case study (balanced)*, there is no overestimation in the settings with a moderate sample size of $n = 100$. Similar results were also obtained for balanced settings with four response classes (Figures A1 and A2). This shows that the overestimation also occurs in settings with more than two response classes.

The findings of the *binary null case study (balanced)* and the *binary power case study (balanced)* do not transfer to the settings with unbalanced response classes. In the *binary null case study (unbalanced)*, the OOB error and the CV error are far closer to the test error (Figure 3). For the study with more extreme class imbalance (ratio 1:5) there are hardly any differences between the error rates estimated by the different strategies (Figure A3). Overall, this suggests a good performance of these two error estimation techniques in unbalanced data settings. The fact that the prediction error is much lower than 0.5 in the unbalanced data settings is not surprising. If for example all observations are classified into the larger class, one achieves an error rate which

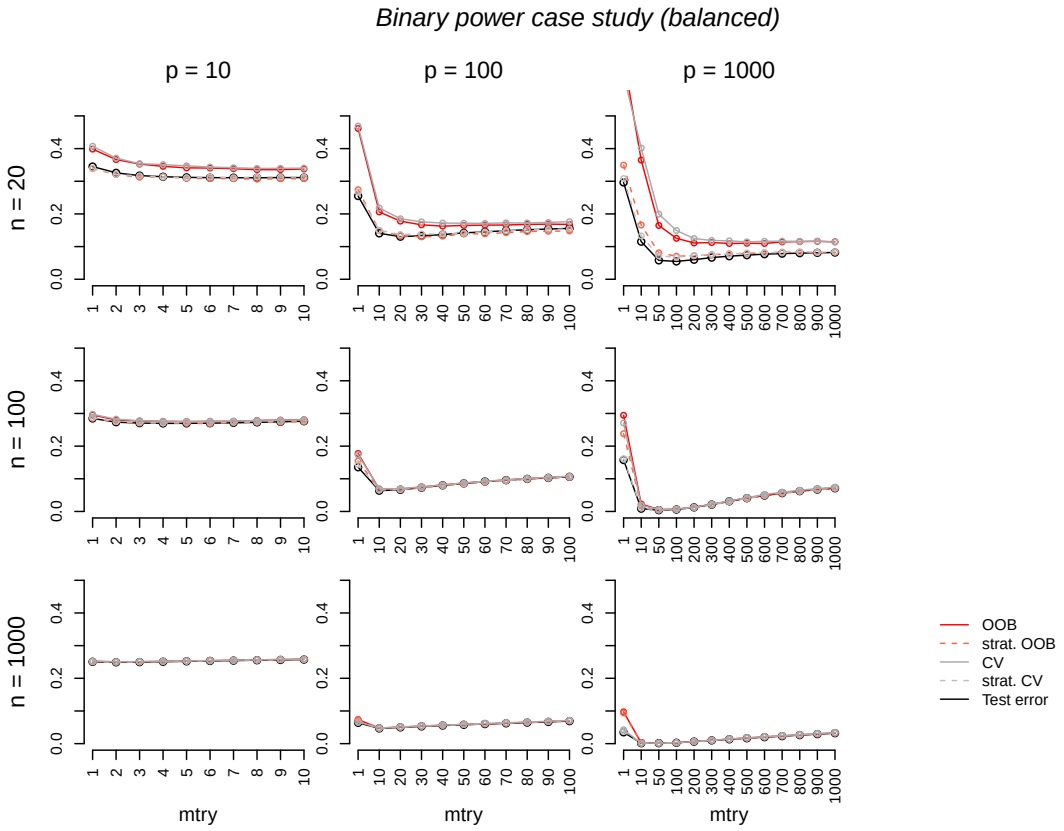

Figure 2: Error rate estimates for the *binary power case study (balanced)*. Shown are different error rate estimates for the setting with two response classes of equal size and with both predictors with effect and without effect. The error rate was estimated through the test error, the OOB error, the stratified OOB error, the CV error, and the stratified CV error for settings with different sample sizes, n, and numbers of predictors, p. The mean error rate over 500 repetitions was obtained for a range of *mtry* values.

equals the proportion of the smaller class. With 30% observations belonging to the smaller class, the proportion of misclassified observations in a null case study could therefore be expected to be about 30%. These expectations are in line with the test error in Figure 3.

Some differences between the stratified OOB error and the test error can be observed in some of the power case settings (right column of Figure 2, right upper plot in Figure 4). In some balanced settings, the stratified OOB error is larger than the test error especially for *mtry* values close to one. However, in general such small *mtry* values are not recommended because, in the presence of many variables without any effect, small *mtry* values prevent the selection of relevant variables yielding RF that have poor performance (Genuer et al.; 2008; Boulesteix et al.; 2012).

Additional simulation studies with many predictor variables with effect show that if many predictors are associated with the response, there is a larger difference between the stratified procedures and the test error (Figure 6; see the Appendix for details on the design). However, in all considered settings the difference between the stratified procedures and the test error is (substantially) smaller than that between the unstratified procedures and the test error.

To conclude, based on these results we have identified settings with (i) (nearly) balanced response classes, (ii) large predictor numbers, (iii) small sample sizes and (iv) a high signal-to-noise ratio as "high-risk settings" in which a large overestimation in the OOB error can be

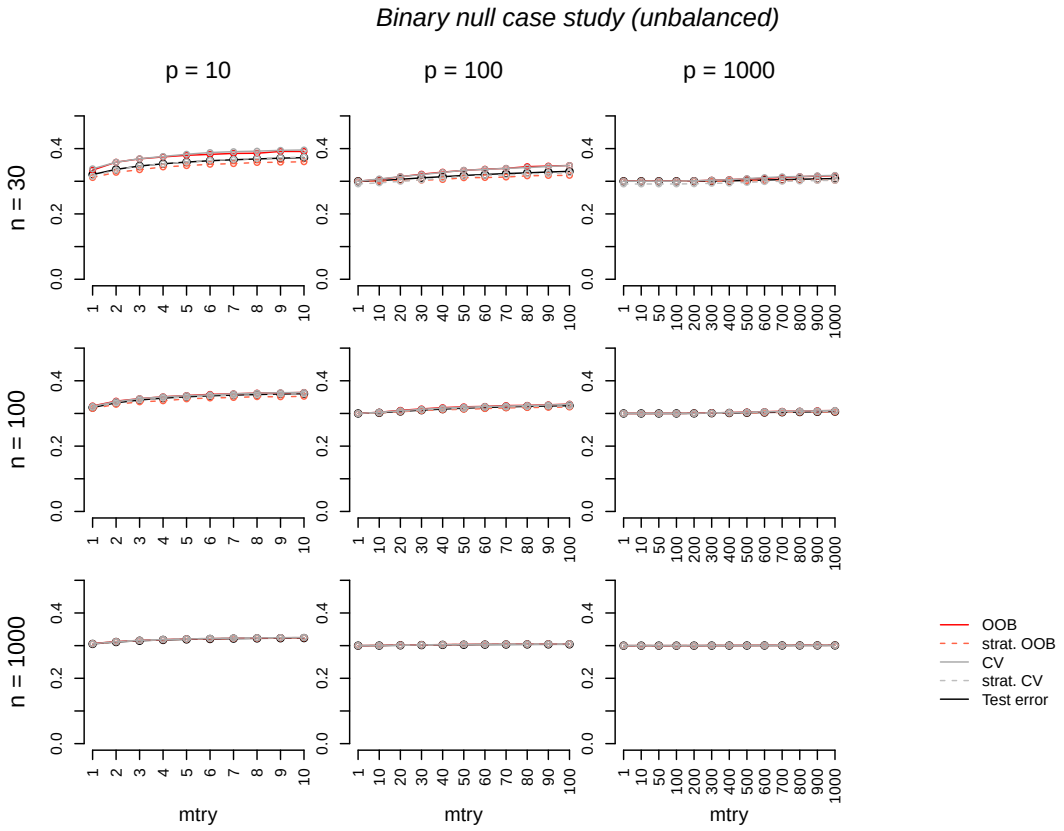

Figure 3: Error rate estimates for the *binary null case study (unbalanced)*. Shown are different error rate estimates for the setting with two response classes of unequal size (smaller class containing 30% of the observations) and without any predictors with effect. The error rate was estimated through the test error, the OOB error, the stratified OOB error, the CV error, and the stratified CV error for settings with different sample sizes, n, and numbers of predictors, p. The mean error rate over 500 repetitions was obtained for a range of *mtry* values.

expected. By now we have quantified the bias for rather simplistic settings which might not be realistic. The results for the real world high-dimensional genomic datasets in which (i)–(iv) apply, are shown in Figure 5. They are in line with the results obtained for the simulation studies: the OOB error and the CV error substantially overestimate the true prediction error for all datasets. The difference between the test error and the error estimated by the OOB procedure or CV is about 5%. CV performs worse than the OOB procedure for the Colon Cancer data, the Prostate Cancer data and the Leukemia data. This might be related to the fact that the CV error is computed from models that are fit based on only a subset of the data, yielding only an upper bound of the prediction error (Kohavi; 1995). Both CV and OOB error are very similar for the three remaining data sets. The stratified OOB error and the stratified CV error, in contrast, have a good performance and approximate the test error very well. An overestimation can, however, be also seen for the stratified CV error and the stratified OOB error for two of the datasets (Colon Cancer data, Prostate Cancer data). However, it is only marginal.

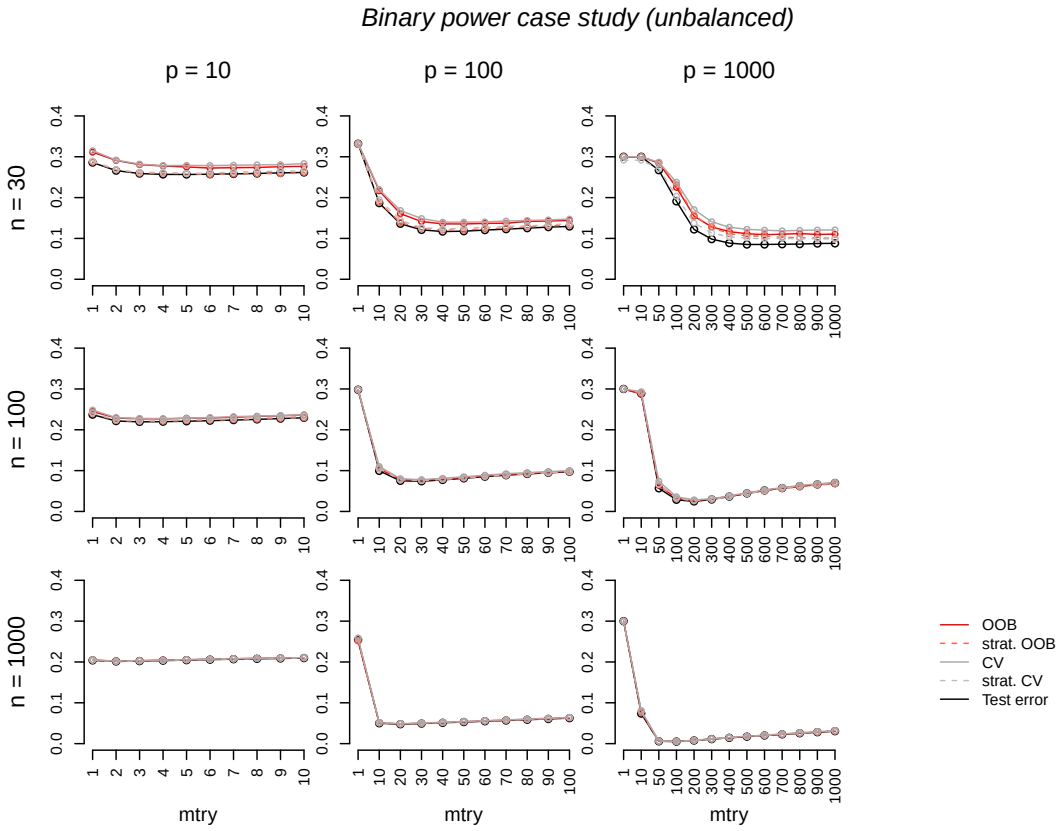

Figure 4: Error rate estimates for the *binary power case study (unbalanced)*. Shown are different error rate estimates for the setting with two response classes of unequal size (smaller class containing 30% of the observations) and with both predictors with effect and without effect. The error rate was estimated through the test error, the OOB error, the stratified OOB error, the CV error, and the stratified CV error for settings with different sample sizes, n, and numbers of predictors, p. The mean error rate over 500 repetitions was obtained for a range of *mtry* values.

3.2 Sources of the bias

The main source of the systematic deviation between the OOB error and the test error was already described in the literature (Mitchell; 2011). In the following, it is described before we explain the bias and its dependence on specific parameters.

In a nutshell, the bias is attributable to the trees' sensitivity to class imbalance. It is well known that classification trees are greatly affected by class imbalance in the sense that trees that were trained on unbalanced samples preferentially classify new observations into the class from which most training observations come. This is also relevant to settings in which there is an equal number of observations from both classes; later it will be shown that for balanced samples the problem is even more severe than for unbalanced settings.

Let us assume in the following that we have a sample with an equal number of observations from both response classes. When constructing trees for a RF we randomly draw subsamples (or bootstrap samples) of observations from the original balanced sample. The subsample may comprise for example, 63.2% of the observations contained in the original sample. In contrast to the original sample, the resulting subsamples generally do not include exactly the same number of observations from each class, that is, the subsamples are often not exactly balanced or may

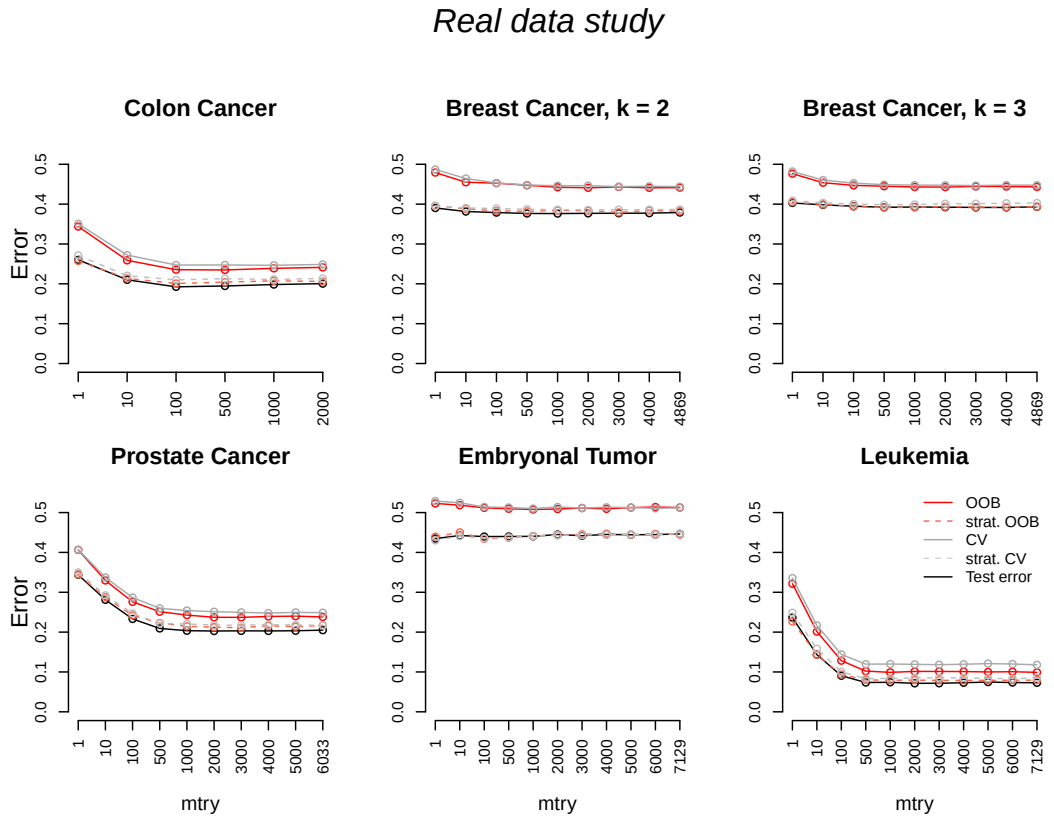

Figure 5: Error rate estimates for the *real data study*. Shown are different error rate estimates for six real data sets with two or three response classes, respectively, of nearly the same size. The error rate was estimated through the test error, the OOB error, the stratified OOB error, the CV error, and the stratified CV error for settings with different sample sizes, n, and numbers of predictors, p. The mean error rate over 1000 repetitions was obtained for a range of *mtry* values.

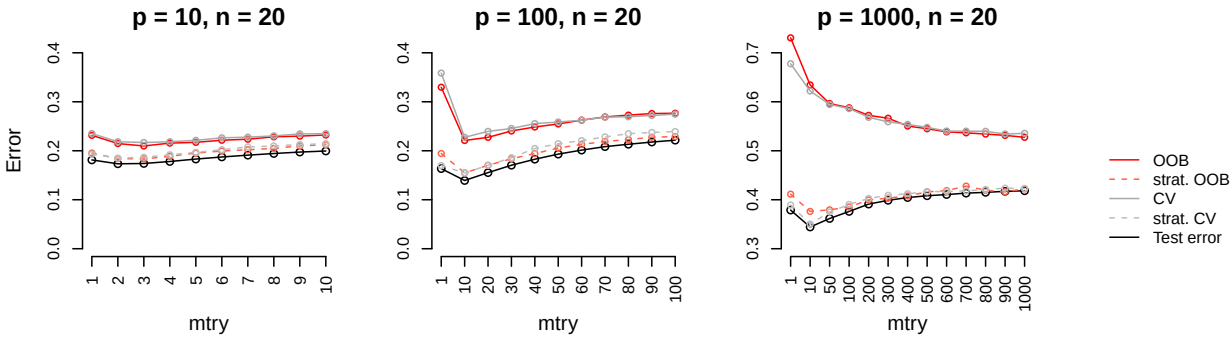

Figure 6: Error rate estimates for additional simulation studies with many predictors with effect and $n = 20$. Shown are different error rate estimates for an additional simulation study with two response classes of equal size and many predictor variables with effect. The error rate was estimated through the test error, the OOB error, the stratified OOB error, the CV error, and the stratified CV error for settings with sample size $n = 20$ and different numbers of predictors, p. The mean error rate over 500 repetitions was obtained for a range of *mtry* values.

even be extremely unbalanced if much more observations from one class are drawn by chance. The degree of class imbalance in the subsample is directly dependent on the sample size of the original sample, n. If n is large the chance for a moderate to extreme class imbalance in the subsample will be rather small, while for small n, the chance will be large. As an example, Figure 7 shows the degree of class imbalance in subsamples of size 63.2% that are drawn from balanced samples of sizes $n = 1000, n = 100$ and $n = 20$. The distributions showing the frequency of class 1 observations in the subsamples were determined based on the hypergeometric distribution. As can be seen, there is a high chance of an extreme class imbalance for small samples. For large samples ($n = 1000$), in contrast, there is only a small degree of class imbalance. The class imbalance in the subsamples yields trees that preferentially predict the class most often represented in the subsample and the more extreme the class imbalance the more extreme the preferential prediction. Thus, the preferential prediction for a class is more pronounced for smaller samples than for larger samples.

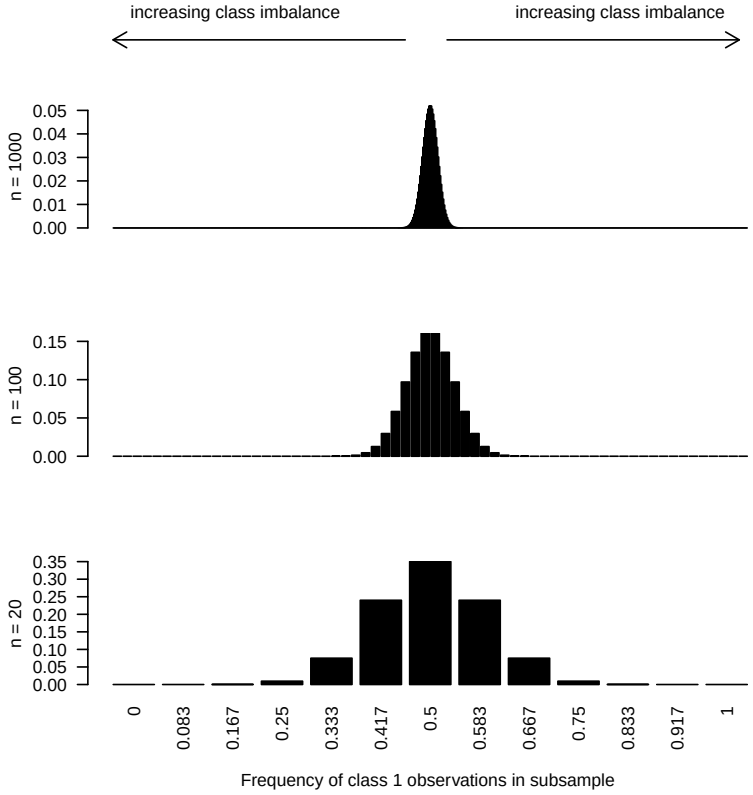

Figure 7: Class imbalance in subsamples drawn from a balanced original sample. Distribution of the frequency of class 1 observations in subsamples of size $\lfloor 0.632n \rfloor$, randomly drawn from a balanced sample with a total of $n = 1000$ (upper), $n = 100$ (middle), and $n = 20$ (lower), observations from classes 1 and 2.

If a prediction shall be obtained for a new observation, all trees in the RF are used to derive a prediction. Then we expect approximately the same number of trees preferentially predicting class 1 and trees preferentially predicting class 2. Overall, there is no preferential prediction for a new observation. In contrast to that, for OOB observations not all trees but only those trees for which the observation was not part of the subsample, are used to derive the prediction. If assuming that an observation i comes from class 1, for example, there are more subsamples

without i that contain more observations from class 2 than subsamples without i that contain more observations from class 1. Accordingly, more of these trees (i.e., trees for which observation i is "out-of-bag") preferentially predict class 2, that is the wrong class. Again, the sample sizes plays an important role. If it is large, there are not substantially more subsamples without i that contain more observations from class 2. Then there is hardly any preferential prediction for the wrong class. In contrast to that, if sample size is small, say $n = 10$, there are substantially more subsamples without i that contain more observations from class 2, yielding substantially more trees preferentially predicting the wrong class. This illustrates that the OOB predictions are worse than predictions that are obtained from the RF if the observation was not used for the construction of the RF. This mechanism finally leads to an OOB error that is too pessimistic, that is, it overestimates the error which is computed for new data.

In line with results from the literature, our studies suggest that a large amount of the overestimation can be solved by drawing subsamples in which the class distribution of the original data set is preserved (Mitchell; 2011). All trees in the RF will then have the same preference for a class, and this preference depends on the class distribution of the original sample. Thus, all trees in the RF and the subset of the trees that is used to derive a prediction for an OOB observation have exactly the same preference for a class which leads to similar test errors and OOB errors. Note that computing the OOB error from an RF based on stratified subsamples yields the stratified OOB error introduced in Section 2.4. The results shown in this paper support the findings of Mitchell (2011) who claims that most of the bias can be eliminated by this alternative OOB error estimation.

In the following subsections, the reason for the dependence of the overestimation on data characteristics and RF parameters are investigated.

3.2.1 Role of the number of observations

The role of the sample size has already been described in detail in 3.2. It was motivated that there is a larger overestimation for smaller sample sizes. In a nutshell, large class imbalance in subsamples is especially a problem for smaller samples. The class imbalance results in trees that tend to more often predict the class that is more represented in the corresponding in-bag sample, or equivalently, that is less often represented in the corresponding OOB sample, leading to higher OOB errors. The dependence of the overestimation on the sample size is seen in the simulation results shown in Section 3.1. These show that the bias is almost negligible for $n = 1000$, while it is large for $n = 20$.

3.2.2 Role of mtry

Figures 1 and 2 show that, in particular for balanced data, the difference between the OOB error and the test error may strongly depend on the parameter *mtry*. The difference is larger for smaller *mtry* values. For unbalanced data in contrast, the difference is smaller for smaller *mtry* values (Figures 3 and 4). The reasons for this are investigated separately for unbalanced and balanced settings in the following.

Unbalanced settings: Let us first consider the setting with unbalanced data and no associations between the predictors and the response (*null case study*). Although there is no association between the predictors and the response in truth, some of the predictors may discriminate in-bag

observations from different classes well by chance. If a large *mtry* value is used, these predictors are chosen for a split and the in-bag observations can be separated well. This yields trees that predict both classes and not only one of the classes (e.g. the most frequent class). In contrast to that, the well-discriminating predictors are not frequently selected as splitting variables in a tree if *mtry* is small. The resulting trees cannot discriminate between in-bag observations from different classes well and tend to predict the larger class more often. Then the RF, which uses the majority vote of the trees, predicts the larger class for almost all observations. This can also be seen by inspecting class predictions that are obtained from RFs with different *mtry* values in empirical studies. The inspection of class predictions was done using simulation studies and is outlined next.

Class predictions were obtained from RFs constructed using 10 observations from class 1 and 20 observations from class 2. The number of predictors, p, was 100. A null case scenario was simulated in which all predictors $X_1, \ldots, X_{100}$ were drawn from a standard normal distribution. Predictions by the RFs were obtained for $n = 10000$ test observations, with an equal number of observations from class 1 and class 2. The proportion of class 1 (minority class) predictions for the test observations was finally computed. This process was repeated 500 times. Figure 8 shows the frequency of class 1 predictions over the 500 repetitions for different values of *mtry*. A clear trend can be seen that the larger class (class 2 in this simulation study) is more often predicted if *mtry* is small. For *mtry* values close to one, class 2 is almost always predicted.

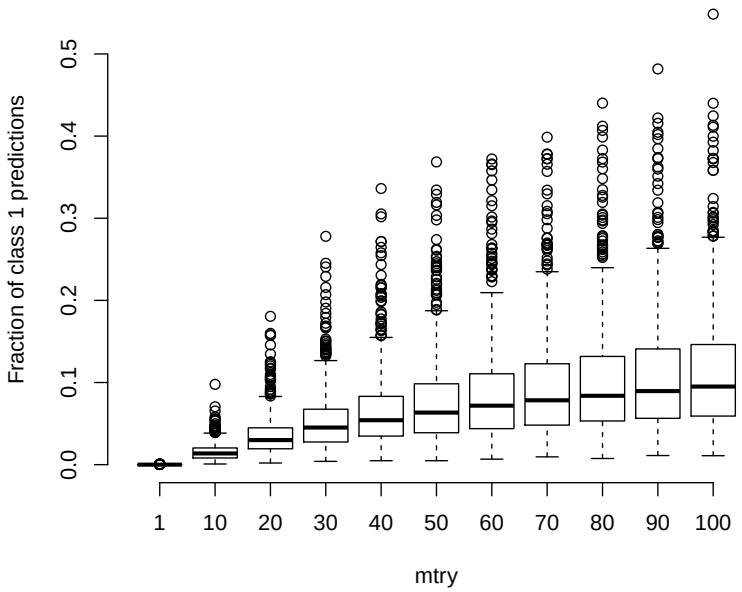

Figure 8: The trees' preference for predicting the larger class in dependence on *mtry*. Fraction of class 1 (minority class in training sample) predictions obtained for balanced test samples with 5000 observations, each from class 1 and 2, and $p = 100$ (null case setting). Predictions were obtained by RFs with specific *mtry* (x-axis). RFs were trained on $n = 30$ observations (10 from class 1 and 20 from class 2) with $p = 100$. Results are shown for 500 repetitions.

The OOB error and the test error are almost the same if *mtry* is very small because most of the trees in the RF predict the same class. In contrast to that, the trees do not always predict the same class if *mtry* is large. For large *mtry* the phenomenon that for the OOB observations the trees tend to predict the opposite class becomes relevant again. This explains the finding that

there is a larger difference between the test error and the OOB error for large *mtry* than for small *mtry*. However, in contrast to balanced settings in which the trees tend to predict the opposite class for an OOB observation, in unbalanced settings most of the trees have the preference for the same class, namely the largest class in the original sample. This reduces the risk that the trees tend to predict the opposite class for an OOB observation. Thus, the difference between the test error and the OOB error is far smaller in the unbalanced simulation settings than in the balanced simulation settings and is smallest in settings with very extreme class imbalance (Figure A4).

Also note that, if *mtry* is set to 1 the prediction of only one class may yield low error rates in specific settings. These are settings in which most of the observations, for which the predictions shall be obtained, are from the class that is always predicted by the RF. For example, if the test data includes 30% of observations from class 1, and the RF always predicts class 2, then the test error is 30%. The same applies to the OOB error. In the simulated data, for example, the OOB error is estimated based on observations, in which approx. 70% of the observations come from class 2 and 30% come from class 1. In the case of small *mtry* values, the RF very frequently predicts class 2 (cf. Figure 8), yielding an OOB error close to 30%. This is also the reason why smaller test and OOB errors were obtained for smaller *mtry* values than for larger *mtry* values in the unbalanced null case scenarios, seen in Figure 3 and A3. The other error estimation strategies are similarly affected.

Balanced settings: Let us now consider the *balanced null case study*, in which there is an equal number of observations from all classes. When drawing samples for tree construction, it is usually the case that not exactly the same number of observations is drawn from each class. When drawing subsamples of size $0.632n$ from $n = 20$ observations (10 from each response class), for example, there is a 50% chance of obtaining subsamples with a different number of observations from each class (cf. Figure 7). When drawing from $n = 100$ observations, the chance for an unbalanced subsample is about 84%. The trees grown on unbalanced samples tend to predict the larger class more often, especially if *mtry* is small. However, in contrast to the settings with an unbalanced original data, in the case of a balanced original sample there are approximately as many trees preferentially predicting class 1 as trees preferentially predicting class 2. In the absence of any associations between the predictors and the response, a new observation would then be classified to class 1 by 50% of the trees, while the other 50% of the trees classify the observation to class 2. This is independent of which value for *mtry* is chosen. Thus, there is no preferential prediction by the RF for new observations in balanced data settings. The test error computed from new observations is therefore not affected by different values for *mtry* if the original sample is balanced.

The OOB error, in contrast, is affected by the choice of *mtry* (cf. Figure 1). When obtaining predictions for an OOB observation i that comes from, say class 2, not all trees of a RF are used but only the trees that are constructed based on samples in which the observation was out-of-bag. Most importantly, even if the original sample is completely balanced, in the samples that do not contain the observation i, the proportion of observations from class 1 is higher on average than the proportion of observations from class 2. Thus, by construction, an OOB observation is out-of-bag for trees that tend to more often predict a class different than the true class the OOB observation belongs to. As explained before, this leads to the high OOB error rates observed in Figure 1. The OOB errors even exceed 0.5, which is the error rate of a random prediction in the absence of any associations between predictors and the response. As was outlined in the previous paragraph, the

trees' preference for the larger class in the subsample (i.e., most often the "wrong" class for the OOB observation) is stronger when small *mtry* values are used. This explains the finding that the OOB error is larger for RFs in which a small *mtry* value is used.

So far we focused on the case in which neither of the predictors are associated with the response. The mechanism described for the *null case study* may also play a role for the *power case study*, especially if there are only few predictors with effect and if the effects are small. In settings with only few influential predictors and many noise predictors, very small *mtry* values lead to trees that frequently select irrelevant variables for a split. Similar to the *null case study*, the trees then preferentially predict the class from which most training observations come. This explains the finding that in the simulation study (including only few relevant variables with rather small effects) the bias in the OOB error is larger for smaller *mtry* values in balanced settings and the opposite is true for unbalanced settings.

3.2.3 Role of the predictors

The simulation results have shown that the bias in the OOB error also greatly depends on the total number of predictors. This is again attributable to the trees' preference for the larger class. It can be shown that the presence of more predictors leads to a more extreme preference for the majority class. We repeated the null case studies presented in Section 3.2.2 and Figure 8 for settings with $p = 10$ and $p = 1000$. Figure 9 shows the fraction of class 1 predictions (average of 500 repetitions) for $p = 10$, $p = 100$ and $p = 1000$. It shows that the preference for predicting class 1 by RF is more pronounced for settings with a larger number of predictors.

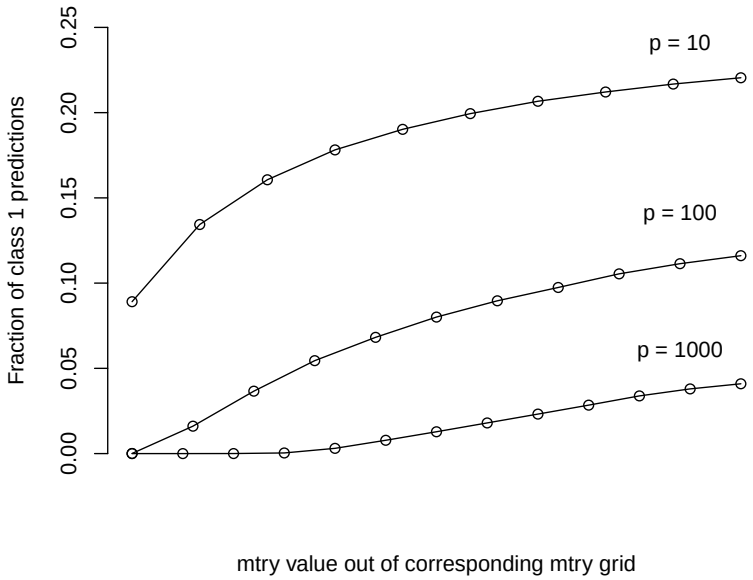

Figure 9: The trees' preference for predicting the larger class in dependence on *mtry* and the predictor number. Fraction of class 1 (minority class in training sample) predictions obtained for balanced test samples with 5000 observations from class 1 and 2, each (null case setting). Predictions were obtained by RFs with specific *mtry* from a corresponding grid of *mtry* values ($\{1, 2, \ldots, 10\}$ for $p = 10$, $\{1, 10, 20, \ldots, 100\}$ for $p = 100$, $\{1, 100, 200, \ldots, 1000\}$ for $p = 1000$). RFs were trained on $n = 30$ observations (10 from class 1 and 20 from class 2) with $p \in \{10, 100, 1000\}$. The mean fractions over 500 repetitions are shown.

Again, depending on the class imbalance in the data used to construct the RF, a preference for the larger class can be of advantage or disadvantage for the bias in the OOB error. With unbalanced training data, a preference for the majority class will lead to a smaller bias in the OOB error; see Section 3.2.2. A larger bias in the OOB error will be obtained in contrast if the training data is balanced.

Correlations between predictors also play a role, as can be seen when comparing the results of the *real data null case studies* with and without any correlations, respectively (Figures 10 and 11). We observe that the bias of the OOB error and the CV error is larger if predictors are uncorrelated. Intuitively, if predictors are correlated, they contain more or less the same (or at least similar) information. Thus, there is less information contained in correlated predictors than in uncorrelated predictors. A similar mechanism occurs that has been described for the number of predictors: the less information that is contained in the data (e.g. due to a small number of predictors or high correlations), the less extreme the trees' preference for one of the classes and the smaller the bias in the OOB error.

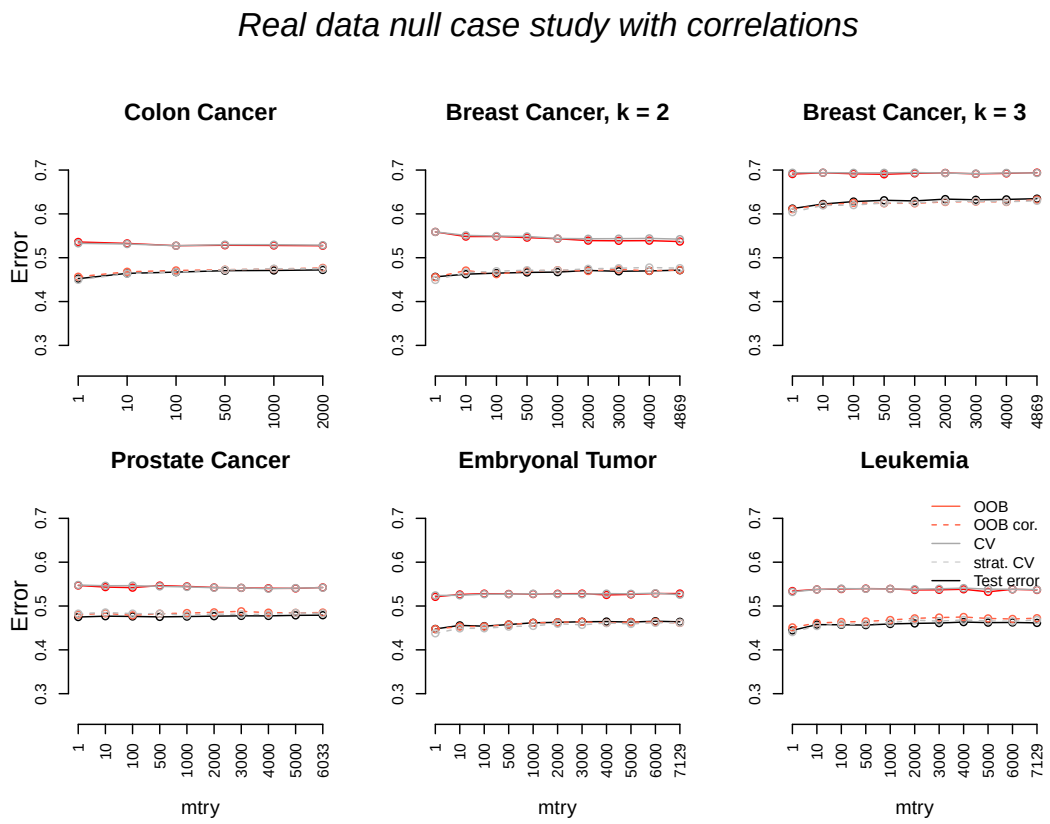

Figure 10: Error rate estimates for the *real data null case study with correlations*. Shown are different error rate estimates for studies based on six real data sets with correlated predictors and two or three response classes, respectively, of nearly the same size. The error rate was estimated through the test error, the OOB error, the stratified OOB error, the CV error, and the stratified CV error for settings with different sample sizes, n, and numbers of predictors, p. The mean error rate over 1000 repetitions was obtained for a range of *mtry* values.

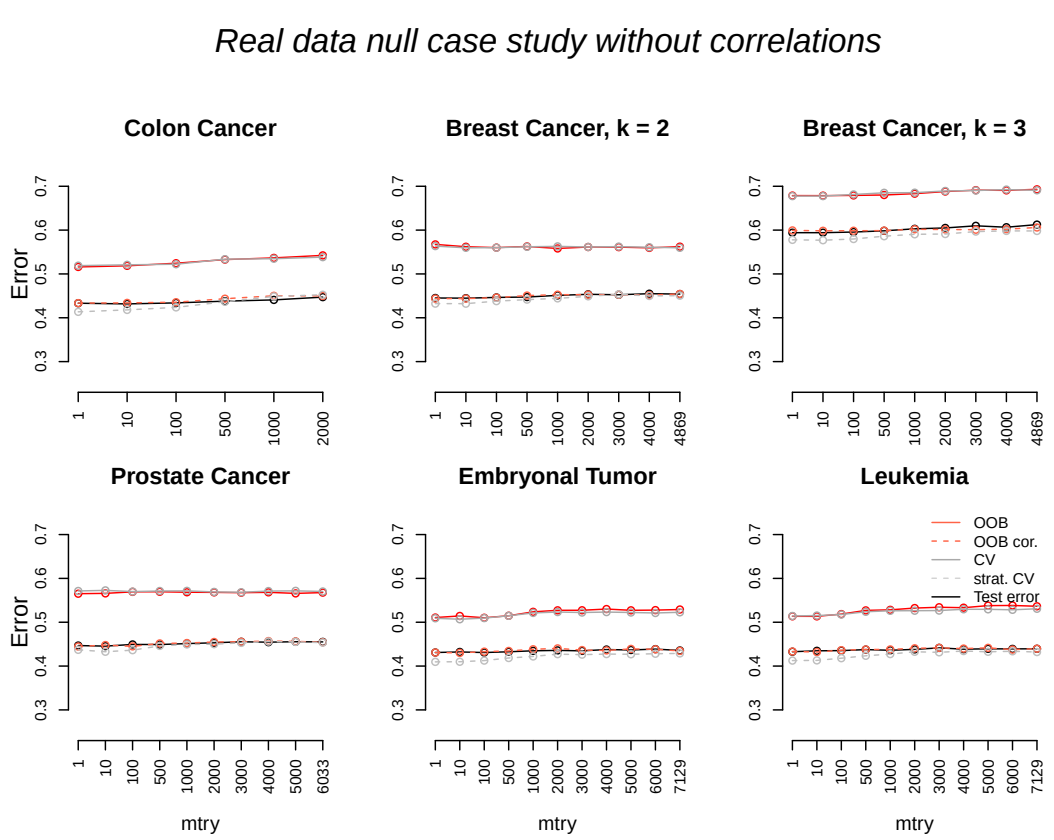

Figure 11: Error rate estimates for the *real data null case study without correlations*. Shown are different error rate estimates for studies based on six real data sets with uncorrelated predictors and two or three response classes, respectively, of nearly the same size. The error rate was estimated through the test error, the OOB error, the stratified OOB error, the CV error, and the stratified CV error for settings with different sample sizes, n, and numbers of predictors, p. The mean error rate over 1000 repetitions was obtained for a range of *mtry* values.

3.3 Consequences for tuning mtry

The OOB error is frequently used to tune parameters like *mtry*. From the studies in Section 3.1, we have seen that the unstratified OOB error and the unstratified CV error often overestimate the true prediction error. Further, it was seen in some settings that the overestimation depends on *mtry*. This was not the case for the unstratified procedures, which were almost unbiased. In the following, we compare the performance of RF when the *mtry* value is chosen based on the OOB error, the stratified OOB error, the CV error and the stratified CV error. The performance was measured by the error rate which was computed based on the independent test data set. A different performance between RFs selected based on the stratified and the unstratified error estimation procedures would suggest that the bias affects tuning parameter selection, or in other words, that a suboptimal model might be chosen when the OOB error (or unstratified cross-validation) is used for parameter tuning.

There were no systematic differences between the four methods in the considered simulation studies and in the real data studies (not shown). However, for the additional simulation studies with many variables with effect, there are differences in the settings with $p = 1000$ and $n = 20$. Figure 6 (right) shows that a small *mtry* of 10 yields the RF with the best performance since the

test error is smallest when using this *mtry* value. The OOB error, however, steadily decreases with larger values for *mtry*, suggesting that large values of *mtry*, such as 1000, should be used instead. Figure 12 shows the performance of the resulting RFs for 500 repetitions of the studies. For the setting with $p = 1000$ and $n = 20$ (Figure 12, right) the mean difference in performance between the OOB error and the stratified OOB error is 1.5%, and the mean difference between the unstratified CV error and the stratified CV error is 1.9%. The bias in the OOB error thus impacts tuning parameter selection and leads to the selection of suboptimal classifiers in this case. However, the impact of the bias is very small and probably of no relevance in practice. For the two settings with smaller predictor numbers ($p = 10, p = 100$), there is again no difference between the four methods (first and second plot in Figure 12).

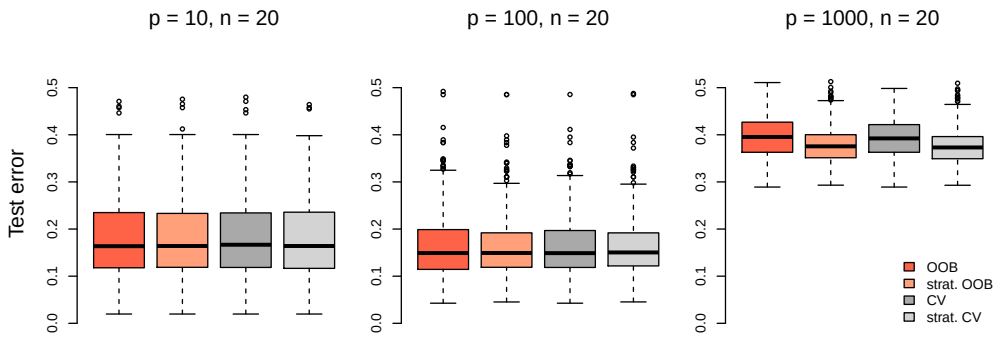

Figure 12: The effect of the bias of the OOB error on RF's performance when used for *mtry* selection. Performance of RF classifiers when *mtry* was selected based on the OOB error, the stratified OOB error, the unstratified CV error and the stratified CV error for the additional simulation studies with many variables with effect. The performance of RF was measured using a large independent test data set.

4 Discussion

Although it was shown that the OOB error may overestimate the true prediction error (Bylander; 2002; Mitchell; 2011), the OOB error is still often used in practice as an estimate of the true prediction error in classification tasks (e.g., DeVries et al.; 2016; Kim et al.; 2014; Marston et al.; 2014). The overestimation is due to the fact that the OOB observations that are used to derive predictions from the trees, might not be representative, in the sense that the response class distribution in the OOB sample might be very different from that of the in-bag sample. A classification tree that is trained on an in-bag sample in which the majority of the observations, say 90%, come from class 1, will have poor predictive ability for an OOB sample, in which only 10% of the observations belong to class 1. Due to random variations different response class distributions in the in-bag and the OOB samples are more likely when the original sample is small. This is the reason why in all the studies shown in this paper, the overestimation in the OOB error was large in small samples. This was also seen in the studies of Mitchell (2011) who considered only a few very specific settings with small sample sizes. The current studies also show that there is hardly any overestimation in large samples, and that the OOB error can be regarded as a good estimate of the true prediction error in very large samples. However, it is difficult to foresee in which settings

the OOB error will be a good estimate of the true prediction error because there are many factors that affect the bias in the OOB error and there is an interplay between the factors. These factors are related to both the data and the parameters of RF.

Concerning parameters in RF, *mtry* was identified as the parameter with the greatest influence on the bias of the OOB error (Mitchell; 2011). We performed additional studies (not shown) that suggest that the parameters controlling the size of trees, in contrast, do not depend on the bias of the OOB error. Depending on the response class distribution in the original sample, larger values for *mtry* might increase (unbalanced settings) or decrease (balanced settings) the bias.

The dependency between *mtry* and the bias in the OOB error might be problematic in the context of parameter tuning if the OOB error is used for selecting an appropriate value for *mtry*. Although there was a clear dependence between the bias and *mtry* in some of the studies, in only one of them this has lead to the selection of suboptimal RF classifiers. This can be explained by the fact that in nearly all studies, it seemed as if the specific choice of *mtry* was not crucial. There was a wide range of *mtry* values that yielded optimal performance, especially for the high-dimensional genomic data sets with values for *mtry* larger than 100 yielding very similar performance. However, one cannot be sure that this applies to all future data sets. Among our studies there was one study that had a clear performance peak at a specific *mtry* value. In this setting the tuning parameter selection based on the stratified OOB error yielded slightly more accurate RF models than that based on the classical, that is unstratified, OOB error.

With respect to data-dependent factors, the present studies identified the response class distribution of the original sample, the predictor number, the correlation between predictors as well as their predictive ability as relevant factors that have an effect on the bias. The studies reported in the literature consider only settings in which there is an equal number of observations from all response classes (Mitchell; 2011). The results in this paper show that the effect of *mtry* on the bias depends on the response class distribution of the original sample. For completely balanced samples, we observed a more extreme overestimation of the true error rate for smaller values of *mtry*. For unbalanced samples the opposite was true. This again underlines that it is difficult to assess whether there will be any bias in future real data applications and how severe this bias is because it depends on several different factors acting together.

Of note, the problem that leads to the overestimation in the error rate is not specific to OOB estimation in RF, but is relevant to any data splitting procedure, such as cross-validation, applied to classification methods that are sensitive towards class-imbalance. This was also seen in the present studies, in which 10-fold cross-validation also yielded too pessimistic error rates. Therefore, cross-validation and related procedures are no alternatives for preventing the overestimation. Instead stratified procedures, such as stratified cross-validation, have been recommended to bypass this problem (Witten and Frank; 2005). The use of stratified cross-validation for error estimation in the context of RF has not been systematically investigated so far. In the present studies, stratified cross-validation resulted in good approximations of the true prediction error of RF in the considered settings.

In benchmarking studies, cross-validation is often applied to compare the performance of different statistical methods. If it is applied in a non-stratified manner, it might happen that the performance for RF might look worse than it actually is. If RF (or a different method that is sensitive towards class imbalance) is considered as a competing method in a benchmark study, it is recommended to use stratified cross-validation to avoid misinterpretations on the performance of RF or other methods that are similarly affected. Note that this problem is relevant especially

to settings in which the original data contains (almost) exactly the same number of observations from the response classes, that is, it is *not* a problem that is encountered especially in unbalanced data settings.

In the original RF version of Breiman (2001), the trees are constructed based on bootstrap samples. In the studies of Mitchell (2011), the use of bootstrap sampling was shown to further increase the bias. Irrespective of this, bootstrap sampling has been shown to induce a preferential selection of certain types of predictors for a split (Strobl et al.; 2007). Therefore, the use of bootstrapping in RF is strictly disapproved to avoid misleading conclusions, and the R package party, for example, draws subsamples by default for this reason. Accordingly, the results in this paper are shown for RF that are always constructed based on subsampling – either unstratified or stratified, the latter leading to the correction addressed above.

The studies shown in this paper are mainly based on the original RF version of Breiman (2001). Some of the simulation settings were also performed with the RF version based on conditional inference trees (Hothorn et al.; 2006) implemented in the R package party to assess if there are any differences (results not shown). The results obtained for this RF version were very similar suggesting that the conclusions drawn from the studies are not specific to the RF version used. Moreover, the problem is not specific to the use of the error rate as performance measure. Any different measure is affected in the same manner. The area under the curve (AUC), for example, represents the probability that for an observation from the diseased class the probability of being diseased is higher than for an observation from the class of healthy subjects (Pepe; 2004). It is often used as an alternative to the error rate for assessing the prediction accuracy in unbalanced binary classification settings. However, the AUC computed from OOB observations similarly underestimates the true AUC, and one cannot circumvent the problem of the biased OOB error by using a performance measure different than the error rate.

Both the stratified OOB error and the error rate computed from stratified cross-validation also overestimated the true prediction error in some of our studies with metric predictor variables. The overestimation was larger if many variables were associated with the response and only marginal if only few variables were associated. Overall, the overestimation through the stratified procedures was considerably smaller than that obtained through the unstratified procedures, supporting the use of stratified procedures. Future studies might aim at developing alternative error estimation strategies that are both unbiased and computationally tractable.

5 Conclusions

To date, very little is known about the bias of the OOB error, and the OOB error is still frequently used for error estimation in classification settings. Simulation-based and real-data based studies with metric predictor variables show that the overestimation is not restricted to binary classification settings and that it is largest in settings with

- an equal number of observations from all response classes (i.e., balanced),
- small sample sizes,
- a large number of predictor variables,
- small correlations between predictors and
- weak effects.

These factors act together which makes it difficult to foresee in which settings the OOB error will greatly overestimate the true prediction error.

The overestimation encountered in settings with metric predictor variables might depend on the parameter *mtry*. This might be a problem when the OOB error is used for selecting an appropriate value for *mtry*, a procedure frequently done in practice. Overall, the prediction performance of RF was not substantially affected when using the OOB error for selecting an appropriate value for *mtry* in the studies shown in this paper. However, one cannot be sure that this applies to all future data.

In line with results reported in the literature (Mitchell; 2011), the use of stratified subsampling yielded almost unbiased error rates in most settings with metric predictors. It might therefore be a solution that is easy to apply in order to reduce the bias in the OOB error. This "correction of the OOB error" consists in using stratified subsampling in place of an unstratified sampling (bootstrap or subsampling) which is usually used when drawing samples on which trees are constructed. Thus, it comes at no additional costs and is easy to apply. For any settings that include only metric predictor variables it is therefore recommended preferring stratified subsampling over unstratified sampling that is, by default, used in RF. This reduces the risk for misinterpretations regarding the predictive accuracy of RF, and might avoid choosing a value for *mtry* that possibly leads to suboptimal performance when using the OOB error for parameter tuning.

Acknowledgements

The author thanks Anne-Laure Boulesteix for fruitful discussions and Sarah Tegenfeldt for language corrections. SJ was supported by grants BO3139/6-1 and BO3139/2-2 from the German Science Foundation.

References

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D. and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, *Proceedings of the National Academy of Sciences* **96**(12): 6745–6750.

Boulesteix, A.-L. (2015). Ten simple rules for reducing overoptimistic reporting in methodological computational research, *PLOS Computational Biology* **11**(4): e1004191.

Boulesteix, A. L., Bender, A., Bermejo, J. L. and Strobl, C. (2012a). Random forest Gini importance favours SNPs with large minor allele frequency: assessment, sources and recommendations, *Briefings in Bioinformatics* **13**: 292–304.

Boulesteix, A.-L., Janitza, S., Kruppa, J. and König, I. R. (2012). Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **2**(6): 493–507.

Breiman, L. (1996a). Bagging predictors, *Machine Learning* **24**(2): 123–140.

Breiman, L. (1996b). Out-of-bag estimation, *Technical report*, Citeseer.

Breiman, L. (2001). Random forests, *Machine Learning* **45**(1): 5–32.

Bylander, T. (2002). Estimating generalization error on two-class datasets using out-of-bag estimates, *Machine Learning* **48**(1-3): 287–297.

Dettling, M. and Bühlmann, P. (2003). Boosting for tumor classification with gene expression data, *Bioinformatics* **19**(9): 1061–1069.

DeVries, B., Pratihast, A. K., Verbesselt, J., Kooistra, L. and Herold, M. (2016). Characterizing forest change using community-based monitoring data and landsat time series, *PLOS ONE* **11**(3): e0147121.

Díaz-Uriarte, R. and De Andres, S. A. (2006). Gene selection and classification of microarray data using random forest, *BMC Bioinformatics* **7**: 3.

Floares, A. G., Calin, G. A. and Manolache, F. B. (2016). *Bigger Data Is Better for Molecular Diagnosis Tests Based on Decision Trees*, Springer, Cham, pp. 288–295.

Genuer, R., Poggi, J.-M. and Tuleau, C. (2008). Random forests: some methodological insights, *arXiv preprint arXiv:0811.3619* .

Goldstein, B. A., Hubbard, A. E., Cutler, A. and Barcellos, L. F. (2010). An application of random forests to a genome-wide association dataset: methodological considerations & new findings, *BMC Genetics* **11**: 1.

Goldstein, B. A., Polley, E. C. and Briggs, F. (2011). Random forests for genetic association studies, *Statistical Applications in Genetics and Molecular Biology* **10**(1): 1–34.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A. et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science* **286**(5439): 531–537.

Hassane, D. C., Guzman, M. L., Corbett, C., Li, X., Abboud, R., Young, F., Liesveld, J. L., Carroll, M. and Jordan, C. T. (2008). Discovery of agents that eradicate leukemia stem cells using an in silico screen of public gene expression data, *Blood* **111**(12): 5654–5662.

Hothorn, T., Hornik, K. and Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework, *Journal of Computational and Graphical Statistics* **15**(3): 651–674.

Hothorn, T., Hornik, K. and Zeileis, A. (2012). Party: a laboratory for recursive partytioning, *R package version 1.0-3, URL http://cran.r-project.org/package=party* .

Kim, D. S., Lee, S. M. and Park, J. S. (2006). *Building Lightweight Intrusion Detection System Based on Random Forest*, Springer, Berlin, Heidelberg, pp. 224–230.

Kim, H. and Loh, W.-Y. (2001). Classification trees with unbiased multiway splits, *Journal of the American Statistical Association* **96**(454): 589–604.

Kim, K.-Y., Zhang, X. and Cha, I.-H. (2014). Combined genomic expressions as a diagnostic factor for oral squamous cell carcinoma, *Genomics* **103**(5): 317–322.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection, *International Joint Conference on Artificial Intelligence*, IJCAI'95, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 1137–1143.

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

Marston, C. G., Danson, F. M., Armitage, R. P., Giraudoux, P., Pleydell, D. R., Wang, Q., Qui, J. and Craig, P. S. (2014). A random forest approach for predicting the presence of Echinococcus multilocularis intermediate host Ochotona spp. presence in relation to landscape characteristics in western China, *Applied Geography* **55**: 176–183.

Mitchell, M. W. (2011). Bias of the random forest out-of-bag (OOB) error for certain input parameters, *Open Journal of Statistics* **1**(3): 205–211.

Nicodemus, K. (2011). Letter to the editor: On the stability and ranking of predictors from random forest variable importance measures, *Briefings in Bioinformatics* **12**(4): 369–373.

Nicodemus, K. K., Callicott, J. H., Higier, R. G., Luna, A., Nixon, D. C., Lipska, B. K., Vakkalanka, R., Giegling, I., Rujescu, D., Clair, D. S. et al. (2010). Evidence of statistical epistasis between DISC1, CIT and NDEL1 impacting risk for schizophrenia: biological validation with functional neuroimaging, *Human Genetics* **127**(4): 441–452.

Nicodemus, K. K. and Malley, J. D. (2009). Predictor correlation impacts machine learning algorithms: implications for genomic studies, *Bioinformatics* **25**(15): 1884–1890.

Oliveira, S., Oehler, F., San-Miguel-Ayanz, J., Camia, A. and Pereira, J. (2012). Modeling spatial patterns of fire occurrence in Mediterranean Europe using multiple regression and random forest, *Forest Ecology and Management* **275**: 117–129.

Pepe, M. (2004). *The statistical evaluation of medical tests for classification and prediction*, Oxford University Press, USA.

Pomeroy, S. L., Tamayo, P., Gaasenbeek, M., Sturla, L. M., Angelo, M., McLaughlin, M. E., Kim, J. Y., Goumnerova, L. C., Black, P. M., Lau, C. et al. (2002). Prediction of central nervous system embryonal tumour outcome based on gene expression, *Nature* **415**(6870): 436–442.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *http://www.R-project.org/*

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P. et al. (2002). Gene expression correlates of clinical prostate cancer behavior, *Cancer Cell* **1**(2): 203–209.

Strobl, C., Boulesteix, A.-L., Zeileis, A. and Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution, *BMC Bioinformatics* **8**: 25.

Tan, A. C. and Gilbert, D. (2003). Ensemble machine learning on gene expression data for cancer classification, *Applied Bioinformatics* **2**(3 Suppl): S75 – S83.

van't Veer, L. J., Dai, H., Van De Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T. et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer, *Nature* **415**(6871): 530–536.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*, Morgan Kaufmann, San Francisco.

Wright, M. N. and Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R, *Journal of Statistical Software* **77**: 1–17.

Zhang, G.-Y., Zhang, C.-X. and Zhang, J.-S. (2010). Out-of-bag estimation of the optimal hyperparameter in subbag ensemble method, *Communications in Statistics – Simulation and Computation* **39**(10): 1877–1892.

Appendix

A.1 Results of multiclass power and null case studies

The results of the *multiclass null case study* and the *multiclass power case study* are shown in Figures A1 and A2.

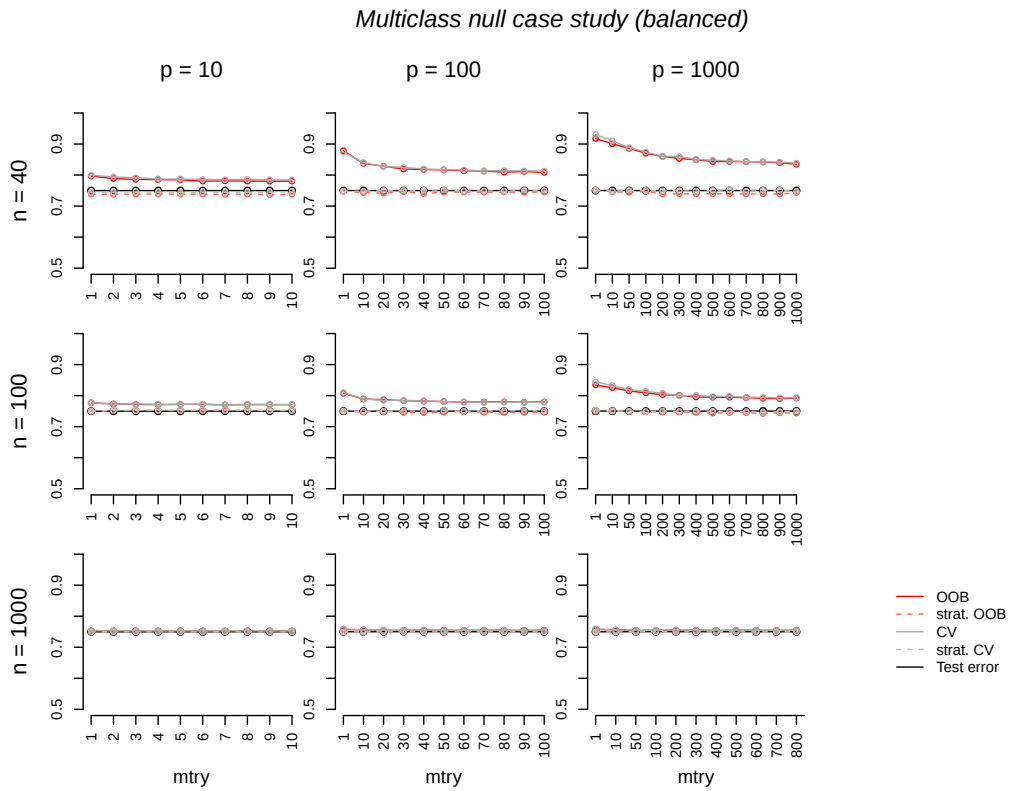

Figure A1: Shown are different error rate estimates for the setting with four response classes of equal size and predictors without any effect. The error rate was estimated through the test error, the OOB error, the stratified OOB error, the CV error, and the stratified CV error for settings with different sample sizes, n, and numbers of predictors, p. The mean error rate over 500 repetitions was obtained for a range of *mtry* values.

A.2 Results of binary power and null case studies with extreme class imbalance (ratio 1:5)

The results of the study with binary response and class imbalance ratio 1:5 are shown in Figures A3 and A4.

A.3 Additional simulation studies: Binary power case study with many predictors with effect

A study with many predictors with effect was simulated. The predictors not associated with the response followed a standard normal distribution. The distribution of predictors with association was different for the two response classes. The predictor values for observations from class 1 were

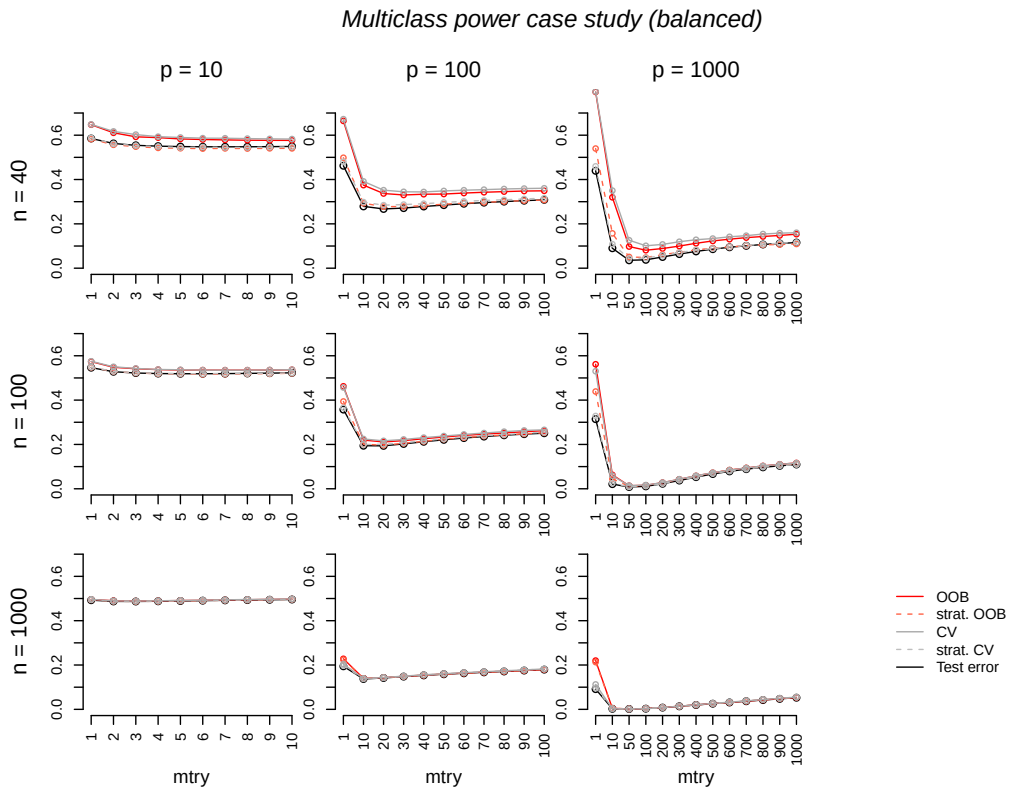

Figure A2: Shown are different error rate estimates for the setting with four response classes of equal size and both predictors with effect and without any effect. The error rate was estimated through the test error, the OOB error, the stratified OOB error, the CV error, and the stratified CV error for settings with different sample sizes, n, and numbers of predictors, p. The mean error rate over 500 repetitions was obtained for a range of *mtry* values.

always drawn from a standard normal distribution. The predictor values for observations from class 2 were drawn from a normal distribution with variance 1 and a mean that was in turn drawn from a normal distribution specified in Table A1.

No. predictors	Predictors	Class 1 $N(\mu_1, 1)$	Class 2 $N(\mu_2, 1)$
$p = 10$	X_1	$\mu_1 = 0$	$\mu_2 \sim N(0, 1)$
	$\vdots$	$\vdots$	$\vdots$
	X_8	$\mu_1 = 0$	$\mu_2 \sim N(0, 1)$
	X_9, X_{10}	$\mu_1 = 0$	$\mu_2 = 0$
$p = 100$	X_1	$\mu_1 = 0$	$\mu_2 \sim N(0, 0.6^2)$
	$\vdots$	$\vdots$	$\vdots$
	X_{50}	$\mu_1 = 0$	$\mu_2 \sim N(0, 0.6^2)$
	$X_{51}, \ldots, X_{100}$	$\mu_1 = 0$	$\mu_2 = 0$
$p = 1000$	X_1	$\mu_1 = 0$	$\mu_2 \sim N(0, 0.2^2)$
	$\vdots$	$\vdots$	$\vdots$
	X_{500}	$\mu_1 = 0$	$\mu_2 \sim N(0, 0.2^2)$
	$X_{501}, \ldots, X_{1000}$	$\mu_1 = 0$	$\mu_2 = 0$

Table A1: Distribution of predictors in class 1 and class 2 in the binary power case study with many predictors with effect.

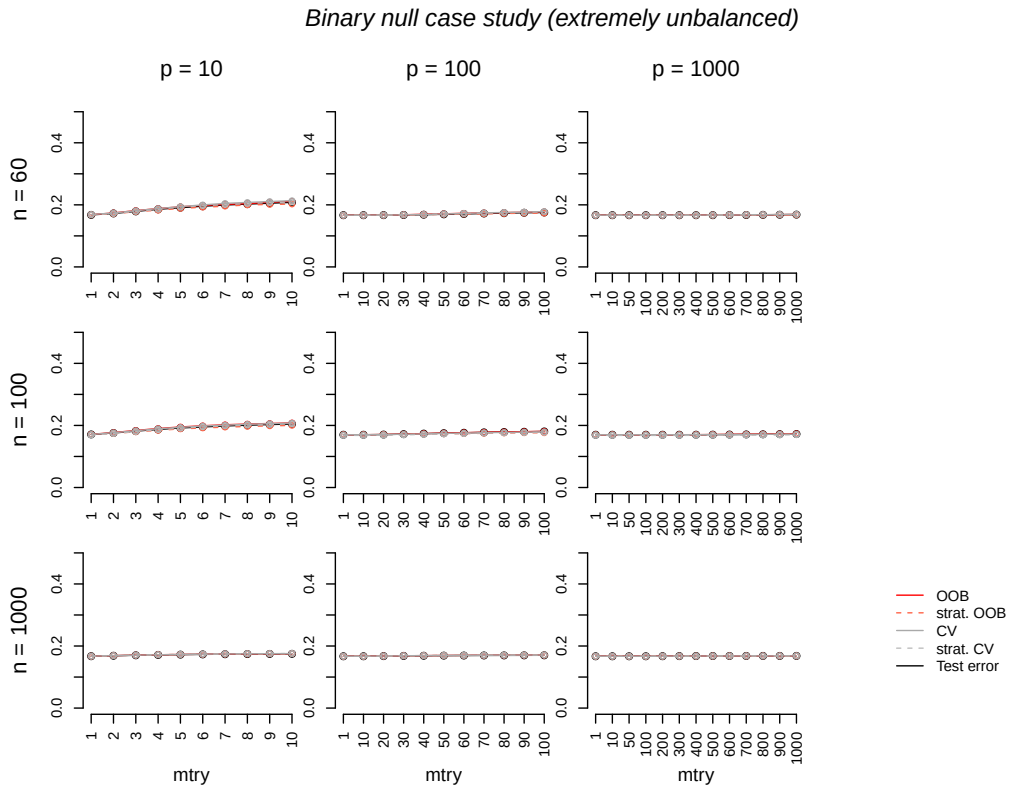

Figure A3: Shown are different error rate estimates for the setting with two extremely unbalanced response classes (ratio 5:1) and predictors without any effect. The error rate was estimated through the test error, the OOB error, the stratified OOB error, the CV error, and the stratified CV error for settings with different sample sizes, n, and numbers of predictors, p. The mean error rate over 500 repetitions was obtained for a range of *mtry* values.

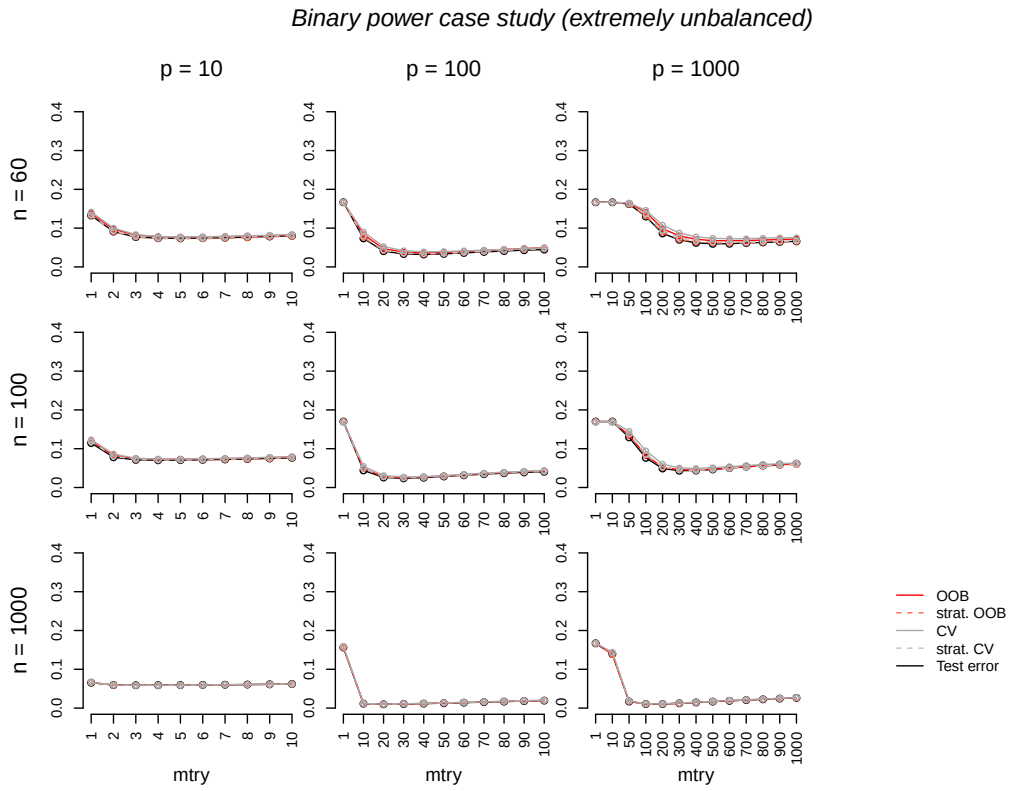

Figure A4: Shown are different error rate estimates for the setting with two extremely unbalanced response classes (ratio 1:5) and both predictors with effect and without any effect. The error rate was estimated through the test error, the OOB error, the stratified OOB error, the CV error, and the stratified CV error for settings with different sample sizes, n, and numbers of predictors, p. The mean error rate over 500 repetitions was obtained for a range of *mtry* values.