Computational Statistics and Data Analysis 56 (2012) 1552–1565

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Recursive partitioning on incomplete data using surrogate decisions and multiple imputation

A. Hapfelmeier [a,*], T. Hothorn [b], K. Ulm [a]

[a] Institut für Medizinische Statistik und Epidemiologie, Technische Universität München, Ismaninger Str. 22, 81675 München, Germany

[b] Institut für Statistik, Ludwig-Maximilians-Universität, Ludwigstraße 33, 80539 München, Germany

ARTICLE INFO

Article history:
Received 21 October 2010
Received in revised form 19 September 2011
Accepted 22 September 2011
Available online 2 October 2011

Keywords:
Recursive partitioning
Classification and regression trees
Random Forests
Multiple imputation
MICE
Surrogates

ABSTRACT

The occurrence of missing data is a major problem in statistical data analysis. All scientific fields and data of all kinds and size are touched by this problem. There is a number of ad-hoc solutions which unfortunately lead to a loss of power, biased inference, underestimation of variability and distorted relationships between variables. A more promising approach of rising popularity is multiple imputation by chained equations (MICE) also known as imputation by full conditional specification (FCS). Alternatives to imputation are given by methods with built-in procedures. These include recursive partitioning by classification and regression trees as well as corresponding Random Forests. However there is only few literature comparing the two approaches. Existing evaluations often lack generalizability due to restrictions on data structure and simulation schemes. The application of both methods to several kinds of data and different simulation settings is meant to improve and extend the comparative analyses. Classification and regression studies are examined. Recursive partitioning is executed by two popular tree and one Random Forest implementation. Findings show that multiple imputation produces ambiguous performance results for both, simulated and real life data. Using surrogates instead is a fast and simple way to achieve performances which are only negligible worse and in many cases even superior.

1. Introduction

Recursive partitioning by classification and regression trees as well as Random Forests is a popular approach in applied statistics. It is used in many research fields such as econometrics, medical statistics and epidemiology. A detailed listing of further applications along with discussions about the current state of methodological research can be found in Strobl et al. (2009). The popularity of trees is rooted in its easy applicability and interpretability. Results are given by decision rules represented by binary split criteria which lead to conditional inferences about a response. Further strong advantages are its' ability to implicitly deal with missing values, correlation, interaction and high dimensional problems e.g. few observations for many variables. These positive aspects in combination with an improved prediction accuracy are already appreciated for Random Forests in the scientific society, e.g. by Lunetta et al. (2004). The special focus of this work is put on the ability of recursive partitioning to deal with missing data.

There are two approaches to handle missing values using recursive partitioning. The first one is the built-in methodology of surrogate splits. The second one is given by imputation methods as summarized by Schafer and Graham (2002) and Horton and Kleinman (2007). Ad hoc methods like available case and complete case analysis as well as single imputation

* Corresponding author. Tel.: +49 89 4140 4347; fax: +49 89 4140 4850.
E-mail address: Alexander.Hapfelmeier@tum.de (A. Hapfelmeier).

0167-9473/$ – see front matter © 2012 Published by Elsevier B.V.
doi:10.1016/j.csda.2011.09.024

by mean, hot-deck, conditional mean and predictive distribution substitution are known to produce a loss of power, biased inference and to underestimate the variance of estimators. In order to overcome these pitfalls and to allow for imputation of multivariate data without the need of specifying a joint distribution of all predictor variables Van Buuren et al. (2006) introduced multiple imputations by chained equations (MICE). Its superiority to ad hoc and single imputation methods is shown in many publications e.g. Janssen et al. (2009, 2010). The thereby multiply imputed datasets can subsequently be used to fit trees or Random Forests that are not depending on surrogate decisions.

Despite its widespread acceptance and application there is only little published knowledge about its performance in missing data situations comparing both approaches. Two reference publications investigating performance differences are given by Feelders (1999) and Farhangfar et al. (2008). However these works lack generalizability as modeling is restricted to classification tasks, categorical data and special simulation schemes. A third related paper is given by Rieger et al. (2010) focusing on Random Forests. Based on extensive simulation studies the authors conclude that *k*-nearest neighbor (kNN) imputation performs comparably to using surrogate decisions.

The motivation of this work is to expand the range of investigated methodology based on the recent state-of-art. The CART algorithm is compared to conditional inference trees and Random Forests. A simulation is set up in order to understand basic properties of using surrogates or multiple imputation. The habit of inducing missing data in real life data as followed by reference publications is retraced. These simulations are found to construct arbitrary schemes of missing values which are not comparable to real life data. Therefore further investigations of real life data already containing missing data are conducted. Both regression and classification tasks are explored.

2. Discussion of related publications

1. Feelders (1999) clearly favors the usage of imputation methods. This conclusion bases on classification tasks performed for two datasets. The rpart() routine implemented in *S* which closely resembles the CART algorithm proposed by Breiman et al. (1984) was applied. Procedures were compared by assessing the misclassification error rate (MER) which equals the fraction of wrong predictions.
   One of the datasets examined is the Pima Indians Diabetes Data Set (see Section 4.2). The performance of a tree using surrogate splits was 30.6% MER. Single imputations based on EM-estimates were repeated by 10 independent draws and achieved an averaged MER of 26.8%. Little and Rubin (2002) clearly show that the variability of estimates is likely to be underestimated by single imputation. Thus comparisons and tests within each of the repetitions might be invalid. In a second experiment a multiple imputation approach is applied ten times. The averaged MER equals 25.2%. To back up the observed differences an exact binomial tests is computed for each repetition. In the first experiment there are 6 of 10 and in the second experiment there are 9 of 10 *p*-values below 0.05. Nevertheless a test for the comparison of two proportions like the McNemar-Test would have been more appropriate. In addition only the training data contains incomplete observations.
   The second data is the waveform recognition data originally used by Breiman et al. (1984). Again missing values are introduced only in the training data completely at random in fractions between 10% and 45%. The imputation is done by a LDA model based on EM-estimates. The MER of trees is assessed in two experiments which differ by the application of single imputation and multiple imputation. In the former the MER of the tree build with imputed data is between 29.2% and 30.6% equally spread among all fractions of missing data. Trees using surrogate splits produce MER values that rise from 29.8% to 34.3%. Results are similar for the latter experiment. The MER of trees with imputation lies between 25.5% and 26.1%. Using surrogates the trees MER increases from 28.9% to 35.6%. Differences are becoming more and more pronounced with high fractions of missing values. However observing 45% missing values in each variable of a dataset is rather rare in real life. A data set containing only 5 variables would already include $1-(1-0.45)^5 = 95.0\%$ incomplete observations. Likewise an equal spread of missing data is rather artificial.
2. Farhangfar et al. (2008) published a profound comparison of classification methods applied to missing data with and without imputation methods. There are several single imputation methods and a multiple imputation approach by polytomous regression using the MICE algorithm. Classification models are support vector machines (SVM), *k*-nearest neighbors (kNN), C4.5 and further ones. The C4.5 method is a decision tree introduced by Quinlan (1993). In total 15 completely observed datasets are investigated by inducing missing values. These exclusively consisted of qualitative variables and responses. The results show that the application of MICE compared to the other imputation methods leads to superior results in most instances. For none of the data sets there was an improvement for the C4.5 method by using imputation. By contrast it often induced even worse MER values. The performance of C4.5 was also independent of the amount of missing data. Like Feelders (1999) the authors restrict the induction of missing values to the training data. The problem of too many missing values equally spread among the variables is present too. Up to 50% of observations per variable are set missing.
3. In an extensive simulation study Rieger et al. (2010) conclude that using a *k*-nearest neighbors (kNN) imputation approach does not improve the performance of conditional Random Forests. Combinations of classification and regression problems with three different correlation structures and seven schemes for missing values are investigated. These studies are repeated for high-dimensional settings with additional noise variables and for two differing scenarios introducing missing values in the training and test data or solely in the training data. The fraction of missing values is not varied and chosen to be two times 20% and one time 10% in three variables. The comparison of approaches is based on prediction

accuracy which is measured by binomial log-Likelihood and mean squared error (MSE). Despite varying results there is no clear advantage of using imputation. Although using elaborate simulation settings the authors point out that the results may not be generalizable due to particular choices of parameters. However this publication does not incorporate trees, uses a single imputation method and does not vary fractions of missing values.

The former two publications show that the MER increases with an increasing number of missing values for single trees using surrogate splits. Meanwhile the MER of trees with imputation almost does not change. Differences are rather weak for lower fractions of missing data which are more likely to be observed in real life data. Farhangfar et al. (2008) found no improvement for C4.5 Trees with imputed data. They even claim a harmful effect of imputation in this case. Pitfalls and drawbacks of the former two publications are surreal simulation schemes, invalid test procedures, the application of biased imputation and tree building methods and the limited generalizability due to the predominant examination of nonstandard polytomous data and classification tasks only. By contrast the work of Rieger et al. (2010) involves many of these issues by presenting an extensive simulation study for classification and regression tasks. The authors conclude that a *k*-nearest neighbor imputation approach is not able to improve the performance of Random Forests.

3. Methods

3.1. Recursive partitioning

In this work three different algorithms are investigated. These are the famous CART algorithm of Breiman et al. (1984), conditional inference trees by Hothorn et al. (2006) and the corresponding conditional Random Forests.

CART constructs trees by sequently splitting data usually into binary subsets (nodes). This growth of the tree is stopped when a certain criterion is reached e.g. a limiting number of observations in the final subsets (aka leaves). After a tree has reached its final size it is pruned again. Therefore the performance of trees of different size is evaluated via cross-validation. Finally the smallest tree whose performance is within the 95%-confidence interval of the best performing tree is chosen. Splits are conducted in single variables due to different criteria depending on the response type. For binary responses the Gini Index G is used. In a node j it is defined by

$$\widehat{G}_j = 2\frac{N_{1j}}{N_j}\frac{N_{2j}}{N_j}$$

with 1 and 2 indicating the response classes and N the number of observations. For example N_{2j} is the number of observations of class 2 in node j. The Gini-Index is used as a measure of node impurity. It takes values between 0 and 1/2 corresponding to pure (only one response class is represented in a node) and maximally impure nodes (both classes are equally represented in a node). An optimal split is found for the cutpoint of a variable maximizing the Gini gain of a node j to its left (Lj) and right (Rj) child nodes

$$\widehat{\Delta G}_j = \widehat{G}_j - \left(\frac{N_{Lj}}{N_j}\widehat{G}_{Lj} + \frac{N_{Rj}}{N_R}\widehat{G}_{Rj}\right).$$

N_{Lj} is the number of observations send to the left child node. For regression tasks the criterion can be chosen to be the maximization of the residual sum of squares (RSS) differences

$$\widehat{\Delta R}_j = \widehat{RSS}_j - \left(\widehat{RSS}_{Lj} + \widehat{RSS}_{Rj}\right).$$

As already stated by Breiman et al. (1984) the CART algorithm (as well as C4.5) tends to favor decisions for splits in continuous variables and qualitative variables with many categories. Many works like those of Lausen et al. (1994) and Hilsenbeck and Clark (1996) have proposed solutions to this problem of 'optimally selected cutpoints'. In an equal manner predictors with many missing values are preferred candidates for split rules based on an optimization of the Gini Index. Corresponding investigations of Strobl et al. (2007a) headed to an operable adaption.

Facing these pitfalls Hothorn et al. (2006) introduced conditional inference trees. Splits are performed in two steps. In the first step the relation of a variable to the response is assessed by permutation tests based on a theoretical conditional inference framework developed by Strasser and Weber (1999). This allows for a comparison independently of a predictors scale. Thus there is no bias in favor of continuous variables or variables with many categories any more. After the strongest relation is found by the minimal p-value of the permutation tests it is checked if significance to a certain level is still present after adjustment for multiple testing. One possibility is to use the Bonferroni method. Finally, in the second step the best cutpoint for a variable chosen in step one is determined. The growth of a tree stops as soon as there are no further significant relations found. Despite the advantage of unbiased variable selection Hothorn et al. (2006) showed that conditional inference trees do not overspend the alpha error and stick closer to the underlying data structure while producing equal performance results to CART.

Breiman (1996) enhanced the tree methodology by 'bagging' them. The performance of single trees benefits from using this ensemble method by reducing the variance of predictions. In this methodology several trees are fit to bootstrapped samples of the data to build a forest. Averaged values or majority votes of the outputs given by each single tree are used as predictions. Random Forests as introduced by Breiman (2001) are extending this approach. By contrast to bagging the node splits are performed in a random selection of covariates. Therefore even high correlated covariates are able to contribute

to the prediction accuracy of a forest. The trees are grown to a maximal size until terminal nodes are pure or reach a minimal size. Similar to bagging the benefit of forests is an improved prediction. Another property is the evaluation of a covariates importance in the light of importance measures. Strobl et al. (2007b, 2008) uncovered biases in their assessment but simultaneously presented according solutions. Anyhow in a following publication Nicodemus et al. (2010) state that this kind of 'bias' might even be beneficial in some instances. Thus it is appreciated in genome wide association studies for uncovering correlated regions of genetic markers. There is no advice how many trees should be used in a forest. Although Breiman (2001) proofs that with a rising number of trees the Random Forest does not overfit but '…produces a limiting value of the generalization error'.

Each method internally handles missing values by surrogate splits which try to mimic the primary split of a node. The latter is determined by the best split using only non-missing data. Surrogate splits are meant to resemble this split as closely as possible producing the same decisions i.e. sending observations to the same child node. The strength of similarity induces a natural order. If an observations is sent down a tree and faces a primary split for which its value is missing it is further processed along the order of surrogate splits until a decision is found.

3.2. Multivariate imputation by chained equations

Imputation is done by flexible specifications of predictive models per variable. There is no need to determine any joint distributions of the data. Cycling through incomplete variables iteratively updates imputations until convergence. Repeating the procedure several times leads to multiple imputed data sets. A short summary of theory and appealing properties is given in the following.

3.2.1. Multiple imputation

A simple and thus popular approach to handle missing data is the application of multiple imputation (MI) as introduced by Rubin (1987, 1996). Little and Rubin (2002) point out that an apparent advantage of this approach is its ability to make standard complete-data methods applicable to incomplete data. Therefore the user is able to stick to his preferred method of analysis. There is no necessity to switch to one he is not used to, he does not understand or is known to be less powerful.

In general any measure of interest Q (e.g. parameter estimates or response predictions $\widehat{y}$) is assessed by the average

$$\overline{Q}_M = \frac{1}{M}\sum_{i=1}^{M}\widehat{Q}_i$$

using M estimates $\widehat{Q}_i$ derived from the imputed complete data sets. The total variability of the estimate is given by

$$T_M = \overline{W}_M + \frac{M+1}{M}B_M$$

while

$$\overline{W}_M = \frac{1}{M}\sum_{i=1}^{M}\widehat{W}_i \quad \text{and} \quad B_M = \frac{1}{M-1}\sum_{i=1}^{M}(\widehat{Q}_i - \overline{Q}_M)^2$$

are the average of the within-imputation variances $\widehat{W}_i$ and the between-imputation variance, respectively. Of course the essential preceding step is the creation of M imputed data sets. If imputation was only done once, like in single imputation, the imputed values would be assumed to be the true values. This can lead to a severe underestimation of the variance, 'which affects confidence intervals and statistical tests' as stated by Harel and Zhou (2007). Still it is not sufficient to simply create more than 1 dataset by drawing from the conditional distribution $P(Y_{\text{mis}}|Y_{\text{obs}};\widehat{\theta})$. The uncertainty inherent to the estimate $\widehat{\theta}$ itself has to be incorporated too. The posterior predictive distribution of Y_{mis} is

$$P(Y_{\text{mis}}|Y_{\text{obs}}) = \int P(Y_{\text{mis}}|Y_{\text{obs}},\theta)P(\theta|Y_{\text{obs}})\mathrm{d}\theta$$

with

$$P(\theta|Y_{\text{obs}}) \propto P(\theta)\int P(Y_{\text{obs}}, Y_{\text{mis}}|\theta)\mathrm{d}Y_{\text{mis}}$$

denoting the observed-data posterior distribution of θ. A proper multiple imputation approach is supposed to first draw M estimates $\theta^{(1)},\ldots,\theta^{(M)}$ from $P(\theta|Y_{\text{obs}})$. These are subsequently used in the conditional distributions $P(Y_{\text{mis}}|Y_{\text{obs}};\widehat{\theta}^{(t)})$, $t = 1,\ldots,M$. An example of this procedure taken from Rubin (1987) for linear regression can be found in the electronic supplementary (S1).

3.2.2. MICE

Whenever there is more than one variable with missing values the imputation approach needs to be adopted. There are mainly two approaches for missing data imputation in this case. Joint modeling (JM) approaches, as presented by Schafer (1997), are not discussed in detail here. Still it is worth to mention that imputations are directly drawn from the parametric multivariate density $P(Y_{\text{mis}}, Y_{\text{obs}}, R|\theta)$ in this approach. Appropriate methods exist for the multivariate normal, log-linear and general location model. A more practical approach which makes it possible to bypass the specification of a joint distribution is MICE also called fully conditional specification (FCS) by van Buuren (2007). Although it lacks profound

theory Van Buuren et al. (2006) could show in simulation studies that it produces reasonable imputations and coverages of statistics of concern. FCS using linear regression even equals JM under the multivariate normal joint distribution given specific regularity conditions. The same holds for some special cases of the log linear model. According to van Buuren and Groothuis-Oudshoorn (in press) FCS is an attempt to obtain a posterior distribution of θ by chained equations. These authors state that starting with the imputation of missing values by random samples of the observed values the tth iteration of the chained equations is

$$\begin{aligned}
&\theta_1^t \sim P(\theta_1|Y_{1,\text{obs}}, Y_2^{t-1}, \ldots, Y_p^{t-1}),\\
&Y_{1,\text{mis}}^t \sim P(Y_1|Y_{1,\text{obs}}, Y_2^{t-1}, \ldots, Y_p^{t-1}, \theta_1^t),\\
&\vdots\\
&\theta_p^t \sim P(\theta_p|Y_{p,\text{obs}}, Y_1^t, \ldots, Y_{p-1}^t),\\
&Y_{p,\text{mis}}^t \sim P(Y_p|Y_{p,\text{obs}}, Y_1^t, \ldots, Y_{p-1}^t, \theta_p^t),
\end{aligned}$$

with Y_j^t being the jth imputed variable at iteration t. It is easy to see how turns are taken within the iterative steps to infer θ and Y_{mis}. After the convergence of the algorithm it is possible to draw $\widehat{\theta}$ from its posterior and to use it to obtain $\widehat{Y}_{\text{mis}}$. Several imputed datasets are produced by repeating the procedure with different starting values. A practical advantage of FCS are the many possibilities of modeling $P(Y_j|Y_{j,\text{obs}}, Y_{-j}, \theta_j)$. A profound discussion can be found in van Buuren and Groothuis-Oudshoorn (in press). MICE is especially suitable in MAR settings although Janssen et al. (2010) state that it should also be preferred to ad hoc methods like complete case analysis even in MNAR situations. In a review paper of epidemiologic literature Klebanoff and Cole (2008) conclude that MICE is still of minor popularity. Despite its outstanding benefits compared to simpler ad hoc methods it seems like researchers feel uncomfortable using it. Thus they give recommendations about the proper publication of methods using multiple imputation to increase popularity. These are followed in this paper and outlined in Section 3.4.

To date there are still proposals for further developments. For example Burgette and Reiter (2010) claim that complex distributions as well as interactions and nonlinear relations might better be fit using CART as imputation model within the MICE algorithm. They are able to present promising results in a simulation study and an application to real life data. Templ et al. (2011) emphasize that many imputation approaches '…assume that the data originate from a multivariate normal distribution…'. They suggest an iterative robust model-based imputation procedure to deal with data that deviates from this assumption (e.g. data that contains outliers or originates from skewed or multimodal distributions). Accordingly, future studies could as well take the diversity of MICE approaches into account.

3.3. Missing data

Rubin (1976) addresses the problem of correct statistical inference with missing values data. For this purpose he stresses the importance to determine and model the process that causes the missingness

- Missing completely at random (MCAR): $P(R|X_{\text{comp}}) = P(R)$.
- Missing at random (MAR): $P(R|X_{\text{comp}}) = P(R|X_{\text{obs}})$.
- Missing not at random (MNAR): $P(R|X_{\text{comp}}) = P(R|X_{\text{obs}}, X_{\text{mis}})$.

$P(R)$ denotes a distribution of the random variable R which is indicating the state of missingness. The letter R is a popular label because Rubin (1987) originally was dealing with '*R*'esponse rates in surveys. The complete variable set X_{comp} is made up by the observed values X_{obs} and the missing ones X_{mis}. $X_{\text{comp}} = \{X_{\text{obs}}, X_{\text{mis}}\}$. Therefore MCAR indicates that the probability of observing a missing value is independent of the observed and unobserved data. By contrast for MAR this probability is dependent on the observed values. Finally in MNAR it is dependent on unobserved or the missing values themselves.

Farhangfar et al. (2008) outline that the MCAR scheme is assumed for most imputation methods. Anyhow He et al. (2009) point out that the MICE algorithm is also capable to deal with MAR schemes as the imputation model becomes more general and includes more variables. In this situation it becomes more probable that missing values can be explained by observed data. The latter property is especially valuable for real life data already containing missing data. In such settings it is not clear which scheme really holds. Similar statements can be found in Janssen et al. (2010) which claim that even a false assumption of MAR under MNAR has minor impact on results in many realistic cases. The performance of Random Forests under several MAR schemes was investigated by Rieger et al. (2010). These authors compared the usage of surrogates against a single imputation method. In extensive simulation studies they were able to show competing results which did not differ between MCAR and MAR. For all these reasons the introduction of missing data is done in a MCAR scheme in the following studies.

3.4. Statistical software

All analyses were implemented by the R software for statistical computing (R Development Core Team, 2010). The CART algorithm is provided by the function `rpart()` which is part of the equally named package `rpart` (Therneau and Atkinson, 2009). It is opposed to conditional inference trees called by `ctree()` which is part of the `party` package

Table 1
Count of observations and independent variables listed for each complete real life data used in the simulation.

Data	Obs.	Ind. var.
H. Survival	306	3
Heart	270	12
Fertility	47	5
Birthweight	189	8

Table 2
Characteristics of real life data containing missing values. The number of independent variables and observations is given for each data set. In addition, the absolute and relative frequencies of missing observations and missing data in single variables are listed.

Data	Observations			Variables		
	#	≥1 Missing		#	Missing (per var.)	
Hepatitis	155	43	(27.7%)	18	14	(0.6%–18.7%)
Mammo	961	130	(13.5%)	4	4	(0.5%–7.9%)
Pima	768	376	(49.0%)	8	5	(0.7%–48.7%)
Ozone	2534	687	(27.1%)	73	73	(0.1%–11.8%)
Airquality	153	42	(27.5%)	4	2	(4.6%–24.2%)
El Nino	733	168	(22.9%)	4	2	(10.6%–12.4%)
CHAIN	508	173	(34.1%)	7	3	(2.8%–30.5%)
Sleep	62	20	(32.3%)	9	5	(6.5%–22.6%)

(Hothorn et al., 2008). This package also includes the function `cforest()` which is used for the implementation of Random Forests. Unfortunately the function `randomForest()` in the package `randomForest` (Liaw and Wiener, 2002) does not support the fitting of Random Forests to incomplete data. Thus this biased, CART based version could not be used for comparison matters. Multivariate Imputation by Chained Equations was done by the `mice()` function of the `mice` package (van Buuren and Groothuis-Oudshoorn, in press).

All functions were used with these settings:

- The number of trees in each forest equals ntree $= 500$.
- For each node a candidate set of mtry $= \min(5,$ variables available) variables is selected.
- Trees and Forests use maxsurrogate $= \min(3,$ variables available) surrogate splits.
- MICE produces $m = 5$ imputed datasets.
- A normal linear model was used for imputations in continuous variables, a logistic regression for binary variables and a polytomous regression for variables with more than two categories: defaultMethod $= c$("norm", "logreg", "polyreg").
- Concerning the training data each variable was used for the imputation. In the test data the response was excluded from the imputation.
- The fraction of imputed data and number of variables used for imputation can be read of Tables 1 and 2.

4. Studies

In order to observe results for several kinds of data which display a random sample of real life situations with a wide range of attributes there were no constrictive exclusion criteria applied for their selection. 12 datasets were used. Half of them were supposed to be capable for regression tasks and half of them for classification. Four datasets without any missing data were needed for the simulation. The remaining eight had to include missing values in advance. There were no restrictions about the number of observations or variables and amount of missing values. The datasets were included without any prior knowledge of these characteristics or any presumptions about potential outcomes. Therefore a broad set of real life data emerging from different scientific fields, including regression and classification tasks is used.

There are two kinds of studies. The first one is meant to retrace and extend the simulation schemes discussed in Section 2. Simulation settings for introducing missing values are varied for a deeper insight of effects caused by special patterns. The second study is based on the evaluation of real life data already including missing values. It is supposed to provide a more reliable image of potential benefits of imputation without the need of artificial specifications. To stick close to existing publications the performance assessment is given by the mean squared error (MSE). This measure equals the misclassification error rate (MER) when predicting binary responses. It has to be emphasized that an assessment of the error only allows for conclusions about the predictive strength of a modeling strategy. Further considerations about the appropriate estimation of tree structures, correctness of imputation and hypothesis testing cannot be made as a different kind of evaluation criterion as well as simulated data would be needed.

Evaluation is done by Monte-Carlo Cross-Validation (MCCV) as outlined by Boulesteix et al. (2008). Analyses are performed for two cases that differ by the decision of using imputation through MICE or surrogate decisions. It is well known that aiming for a valid evaluation the test data needs to be isolated from the training data in any aspect. This can

only be achieved if two separate imputation models are fit to each of these data sets. Otherwise it is believed that the MSE estimation would be positively biased as an imputation model fit to the training data and applied to the test data would transfer information. One could not call the observations of the test data 'unseen' to the predictive model any more. Consequently there are separate imputation models fit to the training and test dataset in the studies. Of course the response is also not allowed to be included in the imputation model for the test data. It is considered unknown until the comparison of predicted and real outcomes for evaluation purposes.

4.1. Simulation

The simulation study was processed in the MCAR pattern. Thus each observation has the same probability to be missing independent of any observed or unobserved data. From the pool of twelve datasets four were chosen for the simulation study as they were fully observed. Two of them are used for the classification of a binary response. Another two are used for regression tasks. All of them are completely observed for which reason missing values needed to be artificially introduced. This procedure is close to the one used by Feelders (1999) and Farhangfar et al. (2008). Although the induction of missing data was not restricted to the training set but extended to the test set too. Fractions of missing values are 0% (benchmark), 10%, 20%, 30% and 40%. The procedure was repeated 1000 times using Monte-Carlo Cross-Validation (MCCV). In each iteration a random sample of 80% was used for the training of the forest while the remaining 20% served as test set. This easily allows for a separate fitting of imputation models to the training data and the validation data.

Part of the criticism in Section 2 was the huge amount of missing values which is far beyond the fractions found in real life data. The number of observations containing at least one missing value is given by $1-(1-\%_{\text{missing}})^{n_{\text{variables}}}$. Thus a dataset that contains only 5 variables with 40% missing values already includes $1-(1-0.40)^5 = 92.2\%$ incomplete observations on average. To examine the effect introduced by the number of variables containing missing data and to stick closer to real life situations the simulation was repeated with a reduced number of variables containing missing data. Therefore instead of setting all variables to be partly missing in each iteration only a randomly chosen third was involved. Each of the predictors had the same probability to be chosen in each MCCV step.

A summary of the data is given by Table 1 and the following listing:

- *Haberman's Survival Data* contains data about the 5 year survival of patients after a breast cancer surgery. It was originally used for studies about log-linear models by Haberman (1976). The corresponding study was conducted between 1958 and 1970 at the University of Chicago's Billings Hospital. The data is provided by the open source UCI Machine Learning Repository (Asuncion and Newman, 2007). There are 306 observations in 3 independent variables namely the age, year of operation and number of positive axillary nodes. The classification task was to describe the survival status of a patient.
- The *Heart Disease Data* was collected at four clinical institutions. These are the Cleveland Clinic Foundation, the Hungarian Institute of Cardiology, Budapest, the V.A. Medical Center, Long Beach, CA and the University Hospital, Zurich, Switzerland. It is given by the open source UCI Machine Learning Repository (Asuncion and Newman, 2007), too. The data contains information about the incidence of a heart disease along with the assessment of a patients age, gender, chest pain type, resting blood pressure, serum cholesterol in mg/dl, a fasting blood sugar measurement (>120 mg/dl), resting electrocardiographic results, maximum heart rate achieved, exercise induced angina, ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, number of major vessels (0–3) colored by fluoroscopy and thallium scan status information. The data contains 270 observations. All risk factors were used for the classification task of a heart disease.
- The *Swiss Fertility and Socioeconomic Indicators Data* contains a standardized fertility measure and socio-economic indicators. It is provided by R and was used for regression analysis e.g. by Mosteller and Tukey (1977). Features are percentages of males involved in agriculture, draftees receiving highest mark on army examination, draftees with education beyond primary school, catholic population and the infant mortality within the first year of life. Data was gathered in 47 French-speaking provinces of Switzerland at about 1888. It contains 47 observations. The regression aims at the explanation of the standardized fertility measure in each province.
- The *Infant Birth Weight Data* is a collection of weights and risk factors of newborns assessed at the Baystate Medical Center, Springfield, Mass during the year 1986. It is part of the R package `MASS`. Venables and Ripley (2003) already used it for a binary classification of low and high birth weight. Here, by contrast, a regression task was performed. Thus a mother's age in years, mother's weight in pounds at last menstrual period, mother's race, mother's smoking status during pregnancy, number of previous premature labors, history of hypertension, presence of uterine irritability and the number of physician visits during the first trimester were used to predict a child's birth weight in grams. The data contains 189 observations.

4.2. Applications

Although simulation studies might be helpful to investigate theoretical properties there are some deficiencies in the simulation schemes that clearly limit generalizability and validity of the results. Corresponding statements have been discussed in Section 2. Thus the main focus of this work is put on the analysis of real life data that originally includes missing values. A total of eight datasets has been used in four classification and four regression tasks. Likewise to the simulation each

data was split in 1000 MCCV turns into 80% training and 20% test observations to estimate a methods performance in terms of MSE. The eight datasets summarized in Table 2 are:

- The *Hepatitis dataset* was obtained from the open source UCI Machine Learning Repository (Asuncion and Newman, 2007). It contains the data of 155 patients that suffered of hepatitis and of whom 32 died. A total of 19 independent variables is available for the classification task of predicting a patients survival. These variables include demographic data like sex and age, information about drugs intake like steroids and antivirals and further clinical factors. One missing value was observed in 4 variables, 5 missing values in 4 variables, and 4, 6, 10, 11, 16 and 29 missing values in one variable respectively. Therefore in 14 out of 18 variables missing values were present. The fraction of missing values per variable ranges from 0.6% to 18.7%. In total 43 (27.7%) of all observations contain at least one missing value.
- The *Mammographic Mass Data* was obtained from the open source UCI Machine Learning Repository (Asuncion and Newman, 2007) as well. Mammography is used for screening breast cancer. Physicians especially radiologists are able to determine the severity (benign or malign) of a suspicious lesion by the examination of these screenings. In the recent past efforts have been made to solve the classification problem by machine learning approaches. The resulting systems are called CAD (Computer Aided Decision/Detection) systems. The data was originally used by Elter et al. (2007) for the evaluation of such systems. Analyses were performed to describe the severity status of a leason. The data also contains information about the age and the shape, margin and density of mass lesions observed in 961 women. Age contains 5 missing values while shape, margin and density are missing 31, 48 and 76 times respectively. Thus the fractions of missing values are 0.5%, 3.2%, 5.0% and 7.9%. Overall 130 (13.5%) of the observations contain at least one missing value.
- The *Pima Indians Diabetes Data Set* was obtained from the open source UCI Machine Learning Repository (Asuncion and Newman, 2007). This data set was also used for the comparison of trees with and without imputation by Feelders (1999). It contains information about the diabetes disease of 768 pima Indian women which are at least 21 years old. In addition the number of pregnancies, plasma glucose concentration, diastolic blood pressure, triceps skin fold thickness, 2-Hour serum insulin, BMI, diabetes pedigree function and the age were recorded. The classification task was to determine whether a women showed signs of diabetes according to the WHO definition. At a first glance the data does not seem to contain any missing data. However these are indicated by many 0 values which are biologically implausible or impossible. Pearson (2006) denotes this instance as 'disguised missing data' and gives a profound discussion about their occurrence in the Pima Indians Diabetes Data Set. Finally there were five variables that contain missing data. The total numbers are 5, 35, 227, 374 and 11 which equals fractions of 0.7%, 4.6%, 29.6%, 48.7% and 1.4%. Overall 376 (49.0%) of all observations contain at least one missing value.
- The *Ozone Level Detection Data Set* is provided by the open source UCI Machine Learning Repository (Asuncion and Newman, 2007). It was collected from 1998 to 2004 at the Houston, Galveston and Brazoria area and contains information about geographic measures and ozone levels. The classification task is to distinguish days of high and low ozone concentration based on information about wind speed, temperature, solar radiation etc. In total there are 2534 observations in 73 measured features. Each of the variables contains between 2 (0.08%) and 300 (11.8%) missing values. For the whole data 687 (27.1%) observations contained at least one missing value.
- The *Airquality Data Set* contains daily measurements of the air quality in New York from May to September 1973. It is directly included in R. The ozone data was originally provided by the New York State Department of Conservation and the meteorological data by the National Weather Service. There are four variables that were of interest for the planed analyses. These are the ozone pollution in parts per billion (ppb), the solar radiation in Langleys (lang), the average wind speed in miles per hour (mph) and the maximum daily temperature in degrees Fahrenheit (degrees F). A more detailed explanation of the data can be found in Chambers (1983). In summary all variables are metric. Therefore the corresponding analyses are regression tasks. Having observed 153 days the ozone pollution contains 37 (24.2%) missing values while the solar radiation contains 7 (4.6%). The whole data contains 42 (27.5%) observations that have at least one missing value.
- The *El Nino Data Set* is provided by the open source UCI Machine Learning Repository (Asuncion and Newman, 2007). It was gathered with the Tropical Atmosphere Ocean (TAO) array of the international Tropical Ocean Global Atmosphere (TOGA) program. TAO is an assemblage of ca. 70 moored buoys recording oceanographic and surface meteorological variables in the equatorial Pacific. The present data contains information about the sea surface temperature, air temperature, humidity as well as zonal and meridional wind speeds. The regression task was to predict the sea surface temperature by the remaining variables. There are 733 observations of which 78 and 91 are missing for the air temperature and the humidity respectively. Thus the amount of missing values per variable is 10.6% and 12.4%. For the whole data 168 (22.9%) of the observations contained at least one missing value.
- The *CHAIN Project Data* contains information from a longitudinal cohort study of HIV infected people living in New York City by 1994. It was originally used by Messeri et al. (2003) for the assessment of HIV treatment effects on survival. The data is part of the R package MI. For the planed analyses there were 508 observations of seven variables. These are the log of self reported viral load level, age at time of interview, family annual income, a continuous scale of physical health, the CD4 count, a binary measure of poor mental health and an indicator for the intake of HAART. The regression task was to explain the continuous scale of physical health. There were 155 missing values in the self reported viral load level, 14 in the family annual income and 39 in the CD4 count. This equals 30.5%, 2.8% and 7.7%. At least one missing value in any variable was observed for 173 (34.1%) of the observations.

Table 3

Summary of mean MSE values and mean relative improvements $\left(\text{rel. imp.} = \frac{\text{MSE}_{\text{Sur.}} - \text{MSE}_{\text{MICE}}}{\text{MSE}_{\text{Sur.}}}\right)$ obtained using multiple imputation and surrogates. Please note that the mean relative improvement is given by the mean of improvements across simulation runs. It cannot simply be recomputed by using the mean MSE values in the formula given here (as the mean of ratios does not equal the ratio of means). Missing values were induced completely at random into real data that was originally fully observed (simulation). Two imputation schemes are distinguished. For one of them all variables and for another one, only one third of variables is partly set missing.

Type	Data	Tree	Missing variables	Missing values			rel. imp.
				0% benchmark	10%–40% surrogates	10%–40% MICE	
Classification	H. Survival	forest	3	0.27	0.28–0.29	0.27	0%–5%
			1		0.27	0.27	0%
		ctree	3	0.28	0.27–0.28	0.27–0.28	−2% to −1%
			1		0.27–0.28	0.28	−2% to −1%
		rpart	3	0.28	0.28	0.28	−2%–0%
			1		0.28	0.28	−1%–0%
	Heart	forest	12	0.17	0.19–0.26	0.18–0.23	0%–7%
			4		0.18–0.19	0.18–0.19	−5% to −2%
		ctree	12	0.24	0.27–0.35	0.24–0.27	7%–22%
			4		0.25	0.25	−2%–0%
		rpart	12	0.22	0.23–0.30	0.21–0.25	6%–13%
			4		0.22–0.23	0.21–0.22	−1%–2%
Regression	Fertility	forest	5	124	129–160	123–129	1%–17%
			2		123–131	122–129	−2%–0%
		ctree	5	126	144–164	116–126	12%–19%
			2		126–136	119–125	1%–4%
		rpart	5	128	129–143	113–121	5%–9%
			2		121–132	115–125	−1%–2%
	Birthweight	forest	8	46e+4	48e+4–52e+4	47e+4–51e+4	2%–3%
			3		46e+4–48e+4	46e+4–48e+4	0%–1%
		ctree	8	52e+4	54e+4–56e+4	51e+4–53e+4	4%–7%
			3		52e+4–54e+4	51e+4–52e+4	2%
		rpart	8	53e+4	54e+4–56e+4	51e+4–54e+4	3%–5%
			3		53e+4–55e+4	51e+4–52e+4	3%–4%

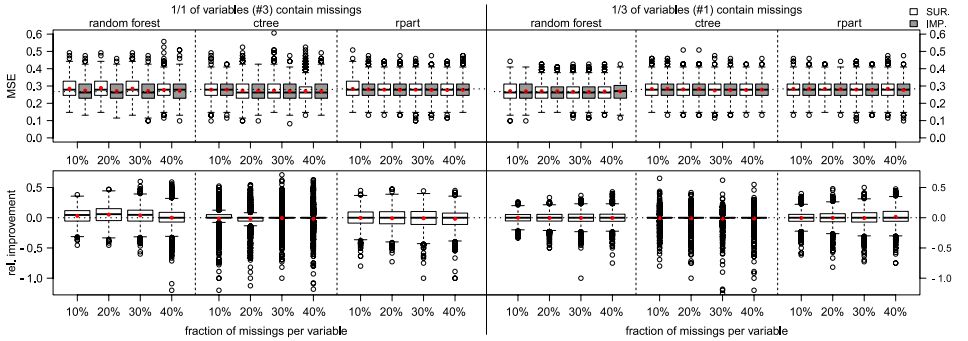
- The *Mammal Sleep Data* comprises features of 62 species ranging from mice over opposums and elephants to man. It can be obtained from the R package `VIM` and was originally used by Allison and Cicchetti (1976) to examine relations between sleep, ecological influences and constitutional characteristics. The observed sleep features include information about duration and depth of sleep phases as well as occurrence of dreams. Constitution is given by measures like body weight and brain weight. The safety of sleep is assessed by scaling for overall danger, danger for being hunted, sleep exposure as well as gestation time etc. One of the main findings in the original paper was a negative correlation between slow-wave sleep and body size. In alignment with these investigations the data was used for the prediction of body weight. There are 20 (32.3%) observations which are not completely observed for all 10 variables. It is interesting to note that Allison and Cicchetti (1976) had originally chosen a complete case analysis as they found the incomplete data to be '…not suitable for the multivariate analyses …'. There are five variables containing three times 4, 12 and 14 (6.5%, 19.4% and 22.6%) missing values.

5. Results

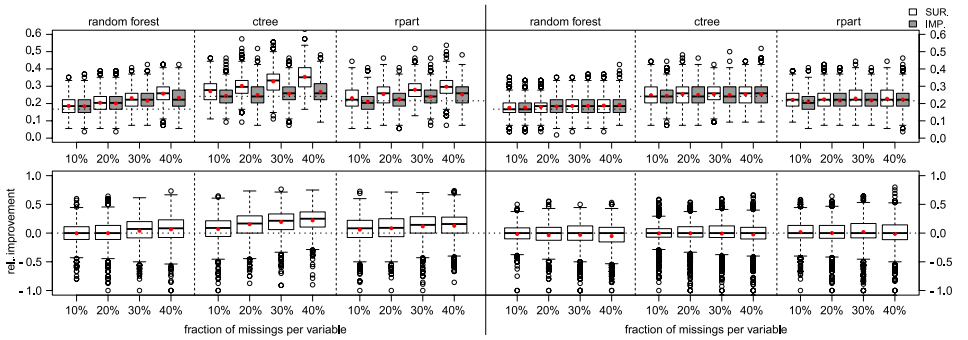
5.1. Simulation

The following contains a listing of discussions for each of the four investigated datasets. A corresponding graphical representation is given by Fig. 1. A summary of observed MSE values can be found in Table 3. An even more elaborate listing is given in the electronic supplementary (S2).

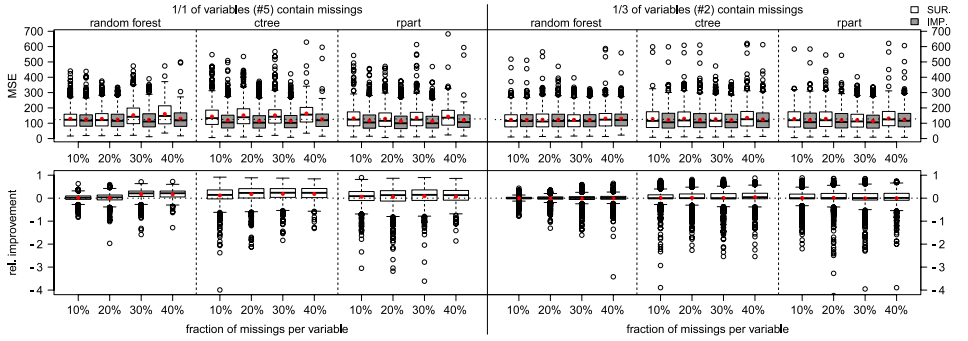
- *Haberman's Survival Data* is used to predict the 5 year survival of patients after a breast cancer surgery. Random Forests, `ctree` and `rpart` perform comparable. They are able to preserve the benchmark MSE (obtained for 0% missing values) independent of the procedure to handle missing values. The relative improvements by using MICE instead of surrogates reach from −2% to 5% and get even less pronounced (−2%–1%) when only one third of variables contains missing values.
- Via the *Heart Disease Data* it is assessed how well the presence of a heart disease can be predicted. In terms of prediction accuracy Random Forests outperforms both single tree methods. `rpart` produces slightly superior results to `ctree`. Increasing the number of missing values makes error rates rise especially when they are introduced in each variable. The



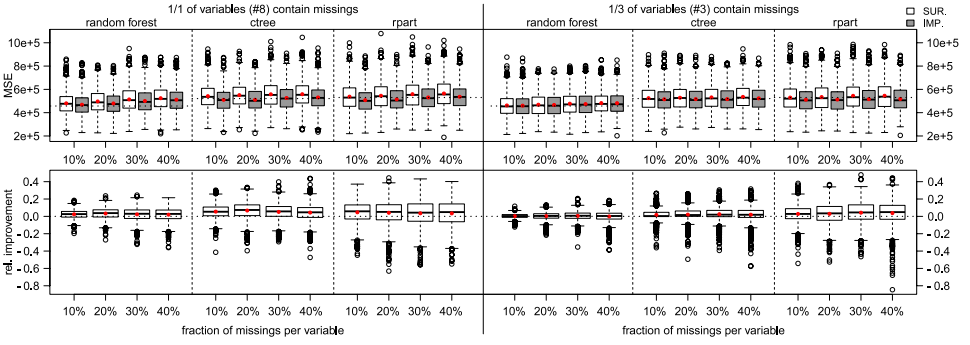
(a) H. Survival.



(b) Heart.



(c) Fertility.



(d) Birthweight.

Fig. 1. Boxplots of MSE values for the simulation setting. For each data the analysis is done twice. The left column shows results when all variables contain missing values. In the right column only one third of variables is partly missing. White boxes indicate the use of surrogates and gray boxes the use of imputation. Solid points represent corresponding means. Vertical dashed lines in the error plots represent the benchmark mean MSE obtained for data containing 0% missing values. In addition, the relative improvement by imputation opposed to surrogates is given below each error plot.

Table 4
Summary of the relative improvement (rel. imp. $= \frac{MSE_{Sur.} - MSE_{MICE}}{MSE_{Sur.}}$) obtained by using multiple imputation and surrogates within 1000 MCCV repetitions for the real life data originally including missing values (applications).

	Data	forest (%)	ctree (%)	rpart (%)
Classification	Hepatitis	−1	−4	1
	Mammo	4	3	0
	Pima	0	1	0
	Ozone	−39	10	0
Regression	Airquality	4	15	15
	El Nino	−41	−16	36
	CHAIN	1	2	4
	Sleep	1	0	−65

relative improvements by applying MICE shrink from a range of 0%–22% to a range of −5%–2% when only one third of variables contains missing values.

- Using the *Swiss Fertility and Socioeconomic Indicators Data* it is examined if a continuous fertility measure can be explained by socio-economic indicators. All three methods produce competitive results. Although in some instances one is able to produce results which are close to the benchmark using surrogates it is obvious that MICE even makes results exceed this level. A slight rise in differences between methods can be observed for an increased number of missing data. The mean relative improvement due to imputation is between 1% and 19%. When there are missing values in only one third of variables this relative improvement ranges from −2% to 4%.
- The *Infant Birth Weight Data* is used to predict a child's birth weight in grams. Random Forests clearly outperform its competitors while `ctree` and `rpart` perform comparable. The difference between using imputation and surrogates shows no clear rise with an increased number of missing values. In some instances MICE makes the performance exceed the benchmark. The improvements by imputation drop from 2%–7% to 0%–4% when the number of variables containing missing values is restricted to one third.

It has to be stressed that in terms of single trees a comparison between the application of MICE and surrogates is not quite fair. MICE produces multiple datasets that vary in the imputed values. To each of them a tree is fit which consequently differ from each other. Their average or majority decision for an observation is used for prediction. Several works e.g. Bühlmann and Yu (2002) and Breiman (1996) show that such an ensemble approach performs superior to single trees. This fact becomes apparent for the Swiss Fertility Data as MICE is able to exceed the benchmark obtained for 0% missing values. By contrast Random Forests are an ensemble method themselves which makes them less prone to this effect. One might find them even more suitable for a fair comparison. Anyhow as MICE is very popular this multiple imputation approach is still preferred to single imputation in order to reflect use-oriented results.

5.2. Applications

This section presents a short description of results obtained for the eight investigated real life datasets originally containing missing values. Graphical representations are given by Fig. 2. The relative improvement of using MICE instead of surrogates can be read off Table 4. An extensive listing of results can be found in the electronic supplementary (S3).

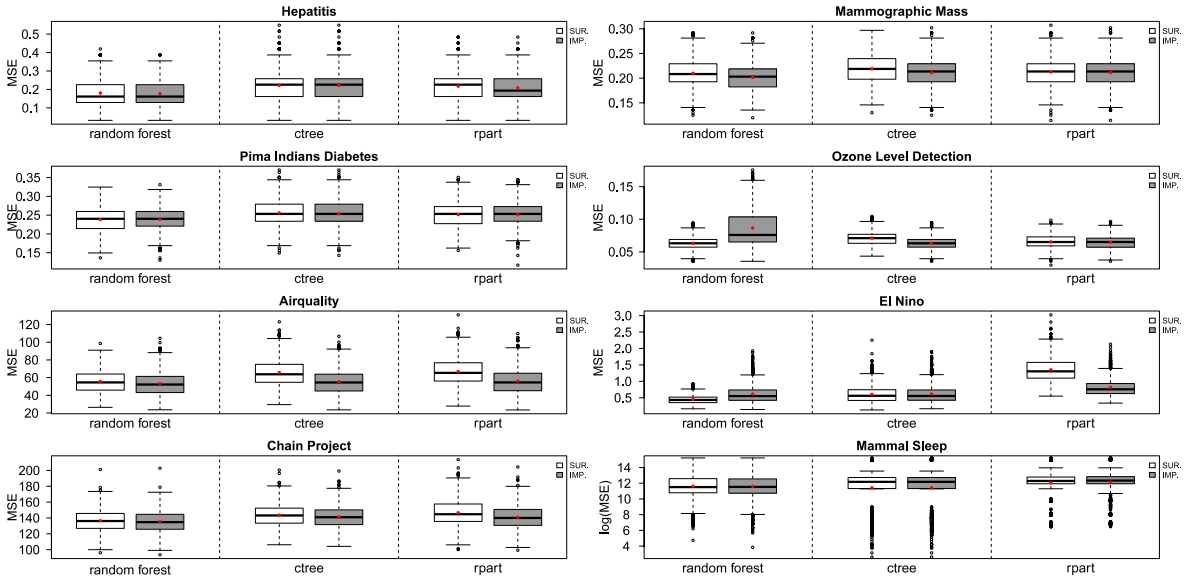
A close look at the observed MSE values reveals that Random Forest performs best while `ctree` and `rpart` are comparable. The relative improvement for Random Forests lies within −41% and 4%. For `ctree` these values are between −16% and 15%. Using `rpart` values range from −65% to 36%.

6. Discussion

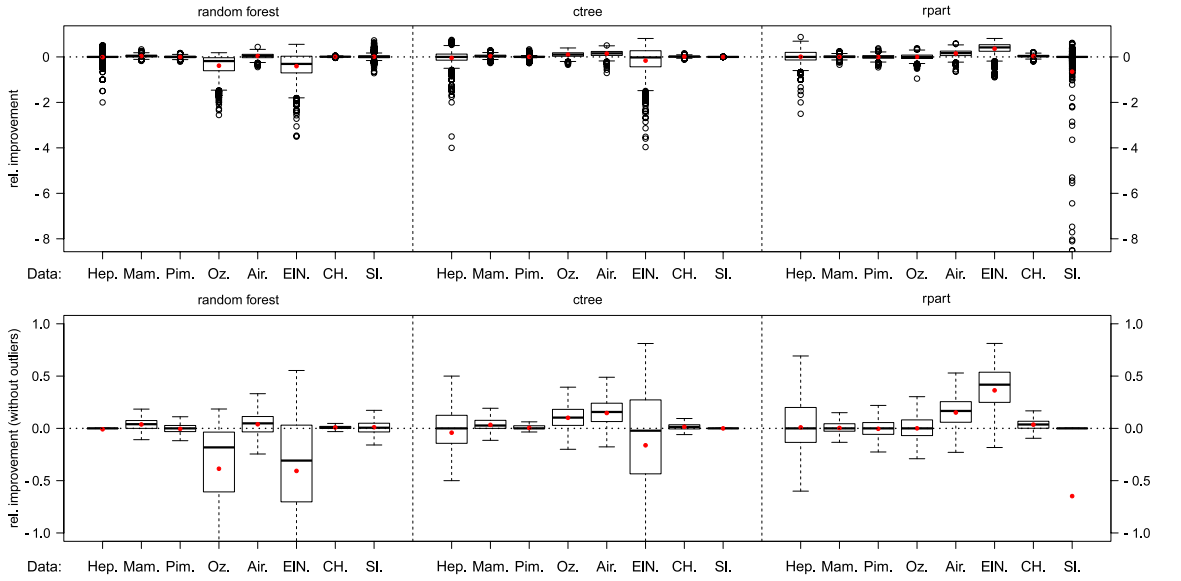
Summing up the findings of the simulation it has to be pointed out that `rpart` and `ctree` alternately beat each other in performance. Similar results were already found by Hothorn et al. (2006) in their work introducing `ctree`. Still one may tend to use `ctree` as `rpart` is known to be biased towards the selection of variables with many possible cutpoints and many missing values. In fulfillment of expectations Random Forests do not show inferior results compared to single tree approaches. Therefore it is recommendable for applications when the main interest is put on prediction strength.

6.1. Simulation

Independent of the statistical model, the underlying dataset and the fraction of missing values it is found that results are dependent on the amount of variables containing missing data. If there are missing values in all of them the relative improvement of using MICE instead of relying on surrogate decisions ranges from 0% to 17% for Random Forests. For `ctree` it is between −2% and 22%. And for `rpart` it lies within −2% and 13%. If only one third of variables contains missing values the improvement diminishes. Now it ranges from −5% to 1% for Random Forests, −2%–4% for `ctree` and −1%–4% for `rpart`. These results show that on one hand mice tends to be beneficial when there are many missing values in many variables. On the other hand it looses this ability when the number of missing values is limited and may even produce inferior results.



(a) Mean squared error.



(b) Relative improvement.

Fig. 2. Boxplots of MSE values for the applications setting are given in Fig. 2(a). White boxes indicate the use of surrogates and gray boxes the use of imputation. Fig. 2(b) shows the relative improvement of multiple imputation compared to surrogates. Solid points represent the corresponding means.

In combination with the considerations and findings about the simulation setting in Sections 2 and 5.1 this rises strong doubt about the usefulness of these comparisons for real life situations. The simulation pattern of equally spreading missing values among the whole data in much too high fractions is extremely artificial. There is a strong need to extend simulation to a wider range of patterns that are closer to those found in real life data. A first big step into this direction has already been taken by Rieger et al. (2010) varying the fractions of missing values and additionally taking MAR settings into account. However eligible structures are difficult to identify and it is easier to investigate real life data already including missing values.

6.2. Applications

The potential improvement by using imputation instead of surrogates lies within −41% and 4% for Random Forests. Results were equally ambiguous for tree methods although the gain by using MICE instead of surrogate decisions was slightly more pronounced. To some extend this might be affected by the property of MICE to implicitly produce ensembles of trees.

The relative improvement reaches from −16% to 15% for `ctree` and −65%–36% for `rpart`. Independent of the prediction method used MICE produced inferior results in some cases indicating that imputation may also decrease the prediction accuracy.

6.3. General

Recursive partitioning by trees is still the method of choice if one is interested in clear decision rules. Nevertheless the conducted studies confirmed the superiority of Random Forests in terms of prediction error. Thus it is advisable for applications focusing on prediction strength. There was no convincing improvement of using MICE in combination with Random Forests. In terms of prediction accuracy Random Forests seem to be capable to handle missing values by surrogates almost as well as by imputation. A slightly more distinct benefit was found for single tree procedures though it was negligible in many cases. For all methods and studies the application of MICE also produced inferior results. Furthermore the extra effort of using imputation should not be underestimated. For example one might decide to create five imputed datasets which results in a fit of five models. If these are subsequently applied to each of another five imputed datasets there are 25 predictions to be made. In total this makes 30 computational steps (5 times fitting + 25 times prediction). Using surrogates it takes only one fitting and one prediction step. Generally the number of computational steps is given by $n + n * m$ while n is the desired number of imputed datasets for the fitting and m for the application of the models. In addition, by using multiple imputation during the fitting process of single trees these loose their ability to provide simple decision rules which is often the main reason for their application.

Results for the habit of using imputation or surrogates in both, the training and test steps are presented in this work. Actually as soon as the fitting and the application of a statistical method is done by two different people these habits could also mix. Some researchers might not be used to imputation methods and therefore will not apply them. Others could have experienced good results using MICE which makes them use it whenever possible. Likewise it has often be claimed that one positive aspect of imputation is that the imputed data can be passed to third party analysts. Therefore all analyses have additionally been conducted by imputing the training set without touching the test set and vice versa. There were two reasons for not presenting these findings in this work. Firstly, the MSE values of both cases lay between those obtained for using imputation in both data sets or in none of them. Secondly, there was also no clear benefit or harm observed by imputing one set instead of the other one.

7. Conclusion and outline

All results indicate that the theoretical properties of the used recursive partitioning methods could be retraced in the simulation and real life data studies. Thus Random Forests showed the best or at least not inferior performances. The CART algorithm and conditional inference trees implemented by the functions `rpart` and `ctree` performed equally well.

The simulation based on four datasets showed no clear improvement of results by using multiple imputation. A potential benefit is highly dependent on the composition of the simulation setting. MICE may even produce inferior results when missing values are limited in number and are not arbitrarily spread across the entire data. Thus the generalizability of simulation results is limited. A broader application of differing simulation schemes is needed for further investigations.

In order to be close to practical applications another set of eight real life datasets was analyzed. The benefit of using imputation in terms of prediction accuracy was found to be ambiguous. Using Random Forests the relative reduction was rather negligible in six datasets ranging from −1% to 4%. In another two datasets it even showed extremely harmful effects (−41% and −39%). Similar results were found for single tree methods though the benefit was slightly more pronounced. Referring to `ctree` it reached from −16% to 15% while it was between −65% and 36% for `rpart`.

Due to reasons like lacking familiarity or additional work and time that needs to be spend for imputation a practitioner might not be willing to use multiple imputation in combination with recursive partitioning methods. Using surrogates instead is fast, simple, works in any data situation and leads to only negligible worse (and in some instances even superior) results. After all these statements are based on the analysis of as much as four simulation settings and eight real life datasets. Although they were chosen to cover a huge range of missing value producing schemes, variable scales, data dimensions and research fields the presented results need to proof generalizability in further studies.

Acknowledgments

The authors would like to thank the two anonymous reviewers and the journal's editor for their constructive criticism and useful suggestions, all of which have led to substantial improvements of this work.

Appendix. Supplementary data

Supplementary material related to this article can be found online at doi:10.1016/j.csda.2011.09.024.

References

Allison, T., Cicchetti, D.V., 1976. Sleep in mammals: ecological and constitutional correlates. Science 194, 732–734. http://www.sciencemag.org/cgi/reprint/194/4266/732.pdf.

Asuncion, A., Newman, D.J., 2007. UCI machine learning repository.

Boulesteix, A.L., Strobl, C., Augustin, T., Daumer, M., 2008. Evaluating microarray-based classifiers: an overview. Cancer Information 6, 77–97.

Breiman, L., 1996. Bagging predictors. Machine Learning 24, 123–140.

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

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and Regression Trees. Chapman & Hall, CRC.

Bühlmann, P., Yu, B., 2002. Analyzing bagging. Annals of Statistics 30, 927–961.

Burgette, L.F., Reiter, J.P., 2010. Multiple imputation for missing data via sequential regression trees. American Journal of Epidemiology 172, 1070–1076. http://aje.oxfordjournals.org/content/172/9/1070.full.pdf+html.

Chambers, J.M., 1983. Graphical Methods for Data Analysis (Statistics). Chapman & Hall, CRC.

Elter, M., Schulz-Wendtland, R., Wittenberg, T., 2007. The prediction of breast cancer biopsy outcomes using two cad approaches that both emphasize an intelligible decision process. Medical Physics 34, 4164–4172.

Farhangfar, A., Kurgan, L., Dy, J., 2008. Impact of imputation of missing values on classification error for discrete data. Pattern Recognition 41, 3692–3705.

Feelders, A.J., 1999. Handling missing data in trees: surrogate splits or statistical imputation. In: PKDD'99: Proceedings of the Third European Conference on Principles of Data Mining and Knowledge Discovery. Springer-Verlag, London, UK, pp. 329–334.

Haberman, S.J., 1976. Generalized residuals for log-linear models. In: Proceedings of the 9th International Biometrics Conference, pp. 104–122.

Harel, O., Zhou, X.H., 2007. Multiple imputation: review of theory, implementation and software. Statistics in Medicine 26, 3057–3077.

He, Y., Zaslavsky, A., Landrum, M., Harrington, D., Catalano, P., 2009. Multiple imputation in a large-scale complex survey: a practical guide. Statistical Methods in Medical Research.

Hilsenbeck, S.G., Clark, G.M., 1996. Practical *p*-value adjustment for optimally selected cutpoints. Statistics in Medicine 15, 103–112.

Horton, N.J., Kleinman, K.P., 2007. Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models. The American Statistician 61, 79–90.

Hothorn, T., Hornik, K., Strobl, C., Zeileis, A., 2008. Party: a laboratory for recursive part(y)itioning. R Package Version 0.9–9993.

Hothorn, T., Hornik, K., Zeileis, A., 2006. Unbiased recursive partitioning. Journal of Computational and Graphical Statistics 15, 651–674. http://pubs.amstat.org/doi/pdf/10.1198/106186006X133933.

Janssen, K.J., Donders, A.R., Harrell, F.E., Vergouwe, Y., Chen, Q., Grobbee, D.E., Moons, K.G., 2010. Missing covariate data in medical research: to impute is better than to ignore. Journal of Clinical Epidemiology 63, 721–727.

Janssen, K.J., Vergouwe, Y., Donders, A.R., Harrell, F.E., Chen, Q., Grobbee, D.E., Moons, K.G., 2009. Dealing with missing predictor values when applying clinical prediction models. Clinical Chemistry 55, 994–1001.

Klebanoff, M.A., Cole, S.R., 2008. Use of multiple imputation in the epidemiologic literature. American Journal of Epidemiology 168, 355–357. http://aje.oxfordjournals.org/content/168/4/355.full.pdf+html.

Lausen, B., Sauerbrei, W., Schumacher, M., 1994. Classification and regression trees (cart) used for the exploration of prognostic factors measured on different scales. In: Dirschedl, P., Ostermann, R. (Eds.), Computational Statistics. Physica-Verlag, Heidelberg, pp. 483–496.

Liaw, A., Wiener, M., 2002. Classification and regression by randomforest. R News 2, 18–22.

Little, R.J.A., Rubin, D.B., 2002. Statistical Analysis with Missing Data, Second Edition, 2nd ed. Wiley-Interscience.

Lunetta, K., Hayward, B.L., Segal, J., Van Eerdewegh, P., 2004. Screening large-scale association study data: exploiting interactions using random forests. BMC Genetics 5.

Messeri, P., Lee, G., Abramson, D.M., Aidala, A., Chiasson, M.A., Jessop, D.J., 2003. Antiretroviral therapy and declining aids mortality in New York city. Journal of Medical Care 4, 512–521.

Mosteller, F., Tukey, J.W., 1977. Data Analysis and Regression: A Second Course in Statistics. Addison-Wesley Pub. Co.

Nicodemus, K., Malley, J., Strobl, C., Ziegler, A., 2010. The behaviour of random forest permutation-based variable importance measures under predictor correlation. BMC Bioinformatics 11, 110+.

Pearson, R.K., 2006. The problem of disguised missing data. SIGKDD Explorations Newsletter 8, 83–92.

Quinlan, J.R., 1993. C4.5: Programs for Machine Learning (Morgan Kaufmann Series in Machine Learning), 1st ed. Morgan Kaufmann.

R Development Core Team, 2010. R: a language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. ISBN: 3-900051-07-0.

Rieger, A., Hothorn, T., Strobl, C., 2010. Random forests with missing values in the covariates.

Rubin, D.B., 1976. Inference and missing data. Biometrika 63, 581–592. http://biomet.oxfordjournals.org/cgi/reprint/63/3/581.pdf.

Rubin, D.B., 1987. Multiple Imputation for Nonresponse in Surveys. J. Wiley & Sons, New York.

Rubin, D.B., 1996. Multiple imputation after 18+ years. Journal of the American Statistical Association 91, 473–489.

Schafer, J.L., 1997. Analysis of Incomplete Multivariate Data. Chapman & Hall.

Schafer, J.L., Graham, J.W., 2002. Missing data: our view of the state of the art. Psychological Methods 7, 147–177.

Strasser, H., Weber, C., 1999. On the asymptotic theory of permutation statistics. Mathematical Methods of Statistics 2.

Strobl, C., Boulesteix, A.L., Augustin, T., 2007a. Unbiased split selection for classification trees based on the gini index. Computational Statistics & Data Analysis 52, 483–501.

Strobl, C., Boulesteix, A.L., Kneib, T., Augustin, T., Zeileis, A., 2008. Conditional variable importance for random forests. BMC Bioinformatics 9, 307+.

Strobl, C., Boulesteix, A.L., Zeileis, A., Hothorn, T., 2007b. Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics 8, 25+.

Strobl, C., Malley, J., Tutz, G., 2009. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychological Methods 14, 323–348.

Templ, M., Kowarik, A., Filzmoser, P., 2011. Iterative stepwise regression imputation using standard and robust methods. Computational Statistics & Data Analysis 55, 2793–2806.

Therneau, T.M., Atkinson, B., 2009. rpart: recursive partitioning. R Package Version 3.1–45; R Port by B. Ripley.

van Buuren, S., 2007. Multiple imputation of discrete and continuous data by fully conditional specification. Statistical Methods in Medical Research 16, 219–242. http://smm.sagepub.com/cgi/reprint/16/3/219.pdf.

Van Buuren, S., Brand, J.P.L., Groothuis-Oudshoorn, C.G.M., Rubin, D.B., 2006. Fully conditional specification in multivariate imputation. Journal of Statistical Computation and Simulation 76, 1049–1064.

van Buuren, S., Groothuis-Oudshoorn, K., 2011. Mice: multivariate imputation by chained equations in r. Journal of Statistical Software, pp. 1–68 (in press).

Venables, W.N., Ripley, B.D., 2003. Modern Applied Statistics with S, 4th ed. Springer, New York, USA.