Ludwig-Maximilians-Universität

Intitute for Statistics

Master Thesis

with the title:

Imputation and Prediction of HIV using NHS Survey Data from Nigeria

Author: Benedikt Baus

Advisor: Prof. Dr. Christian Heumann

Date: July 12, 2018

Abstract

The human immunodeficiency virus (HIV) is one of the biggest pandemic of our time. As there exists a therapy that can suppress the viral load, it is important to identify as many HIV positive people as possible since the majority does not know about their condition. To gather information, national health services (NHS) conduct surveys which include a voluntary HIV test. It would be desirable to be able to predict the result of that test for people who did not attend it.

To achieve this, first multiple imputation is used to accommodate the missing data in co-variables. Then, machine learning methods are applied. Five models are deployed to construct classifiers. The models are a logistic regression model, a mixed effects logistic regression model, random forests, boosted trees and naive Bayes. Additionally, sampling techniques are used to accommodate the highly imbalanced data of the HIV test result.

With none of the techniques was it possible to construct a satisfactory classifier. All classifiers predicted all missing test results as negative. Though it is possible to classify some cases as positive, this comes at a high cost of many false predicted positive test results.

Contents

1 Introduction

Since the first description of the human immunodeficiency virus (HIV) in 1983, it has become a worldwide epidemic. Until today, it was not possible to develop a cure or vaccination. Although, globally, the epidemic reached its highest rate of new infections in 1997 and has been falling ever since, as of 2010 there are still 2.4 million to 2.9 million new infections and 1.6 million to 1.9 million deaths per year. [1]

In Nigeria, 150 000 to 310 000 new infections and 110 000 to 230 000 deaths, related to the acquired immune deficiency syndrome (AIDS), occurred in 2016. HIV prevalence rates, the proportion of a population being infected by HIV, fell from 5.8% in 2001 to 2.9% (2.1%-4.0%) in 2016. Nigeria has the second largest HIV epidemic in the world. While there exist countries with much higher HIV prevalence rates, especially in Sub-Saharan Africa, the size of the Nigerian population means that there are between 2.3 and 4.3 million people living with HIV. Only South Africa has a higher population that is HIV positive. [2]

As there exists an antiretroviral therapy (ART) that can suppress the viral load, it is important to identify the people who have HIV to decrease further spread of the disease. A suppressed viral load means that a person's viral load is reduced to an undetectable level. In Nigeria, 1.1 million people live with HIV and know their status, which equals a rate of 34% (25%-46%) of the total estimated population of HIV positives. Of this 1.1 million, 970 000 are on ART and 780 000 have suppressed viral loads. [2]

The goal of this work is to predict HIV for respondents of the National HIV & AIDS and Reproductive Health Survey (NARHS), who refused to take part in the HIV test. To achieve this, missing values among the data set will be imputed. The structure of this thesis is as follows: First the theoretical aspects of this work are highlighted and then the practical aspects. More precisely, theory about missing data and concepts to handle missing data are introduced, followed by models and measures to validate the predictive properties of these models. Then the data is described and results are presented.

2 Missing Data

2.1 Missing Mechanisms

Ignoring missing data as well as an inappropriate handling of it may lead to biased estimates, incorrect standard errors and incorrect inferences and results. Therefore, an appropriate handling of missing data is quite important. As all missing data handling methods require a certain missing data mechanism, it is crucial to know as

much as possible about the reasons for missing data. Therefore, three missing data mechanisms were introduced by Rubin (1987). [20]
The first one is the missing not at random (MNAR) mechanism. The assumption of MNAR is the existence of a systematic relationship between the probability of missing data on a variable Y and the values of Y, even after controlling for other variables. It is not possible to test for MNAR. This is due to the fact that there is no way to confirm the MNAR mechanism without knowing the missing values themselves. [8, p.8]
If the underlying missing data mechanism is MNAR, then the data mechanism is said to be non-ignorable, as it is required to model the missing data mechanism as part of the estimation process. The denotation of the MNAR mechanism looks as follows:

$$p(R|Y^{obs}, Y^{mis}, \phi)$$

with Y^{obs} being the observed part of the data and Y^{mis} the missing part of the data. Further $R \in \{0, 1\}$ is the missing data indicator, where $R = 1$ indicates that the data is available and $R = 0$ indicates that the data is missing. ϕ is a set of parameters describing the relationship between R and the data. [8, p.11]
Another missing data mechanism is the missing at random (MAR) mechanism. MAR assumes a systematic relationship between the probability of missing data and one or more measured variables. Furthermore, the probability of missing data on a variable Y is not related to the values of Y itself. As for MNAR, it is not possible to test the MAR assumption due to the fact that it is not possible to confirm that the probability of missing data on Y are solely a function of other measured values. [8, pp.6,11] The MAR mechanism is denoted as follows:

$$p(R|Y^{obs}, \phi)$$

The last missing data mechanism is called missing completely at random (MCAR). MCAR is more restrictive than MAR as it assumes that there is no (systematic) relationship between the probability of missing data on a variable Y and the variable itself or other variables in the dataset. MCAR is the only missing data mechanism that can be tested for. [8, pp.7f.,12]
The MCAR mechanism is denoted as follows:

$$p(R|\phi)$$

If the underlying missing data mechanism is MAR or MCAR, then the mechanism is said to be ignorable. For both of these missing mechanisms, it is not needed to model the missing data mechanism as part of the estimation process.

Generally, with given data, it is not possible to determine whether the missing data mechanism is MAR or MNAR. Using missing data methods that require MAR and/or MCAR (like the methods used in this work) might cause bias if the missing data mechanism is in reality MNAR. If performing such missing data methods is problematic depends on the kind of MNAR. A confounder, an unmeasured variable that correlates with outcome and missingness, is not as severe as long as the correlation between the missing outcome and the unmeasured variable is not relatively strong (i.e. below 0.4). However, if the correlation is relatively strong or if there exists a direct relationship between missingness and outcome, then using MAR missing data methods is problematic. According to some researchers, serious violations of MAR are relatively rare. If there exists a confounder and this variable would be observed, then this MNAR scenario would become a MAR scenario. [8, pp.14ff.]

2.2 Tests on MCAR

There exist some tests on the MCAR assumption. One possibility to test for missing completely at random are univariate t-tests or chi-squared tests. The latter have the advantage that they are as well usable with only categorical data. The idea behind the t-tests is to separate missing and complete cases of a variable and use a t-test to examine group mean differences on other variables. In case of the chi-squared test it is tested if there are differences in the frequency in other variables between the missing and complete cases. As the t-test, it always tests one variable against one of the variables with missing data. If the test statistic is significant, then this proves that the underlying mechanism is not MCAR, but MAR or MNAR. If the test statistic is insignificant, then the underlying mechanism is MCAR. This is valid for both t-test and chi-squared test. [8, pp.18f]
Another possible test is Little's MCAR Test, which is a multivariate extension of the t-test. It is a global test of MCAR. If the statistic is significant, then this is evidence against MCAR. [8, pp.19ff.] A problem with Little's MCAR Test is that it cannot identify specific variables that violate MCAR.

2.3 Overview of Missing Data Handling Methods

Complete Case Analysis and Available Case Analysis
There exists the list-wise deletion or complete case analysis and the pairwise deletion or available case analysis. The difference between both is that list-wise deletion eliminates all cases with missing data while pairwise deletion only eliminates all cases with missing data in a variable that is important for the desired statistic. These approaches are the default in many statistical programs for missing data handling.

These approaches require a MCAR-mechanism and produce biased parameter estimates if data is not MCAR. [8, pp.37-42]

Single Imputation

Imputation replaces missing values with possible values and therefore does not throw data away as above approaches do. Generally said, single imputation tends to underestimate standard errors. There exist many imputation methods of which some are introduced below.

Arithmetic Mean Imputation

Missing values are replaced by the mean of their variables. This distorts resulting parameter estimates and underestimates variance and correlations.

Regression Imputation (Conditional Mean Imputation)

Missing values are imputed using regression on the missing variables with complete case analysis. No missing values are allowed in the predictors. With multiple variables with missing values, one has to estimate a model for each missing data pattern. This imputation method is superior to mean imputation, but has also bias. There is perfect correlation on imputed values and it tends to overestimate R^2 and correlations. Further it can underestimate (co-)variances, but less severe than mean imputation. [8, pp.44ff]

Stochastic Regression Imputation

This approach is like regression imputation, but augments each predicted score with a normally distributed residual term. With the addition of residuals to the imputed values, the lost variability can be restored and therefore the bias of the regression imputation approach can be eliminated. Studies show that stochastic regression imputation gives unbiased parameter estimates when the missing mechanism is MAR. It tend to attenuates standard errors.[8, pp.46ff]

Hot-Deck Imputation

This method takes 'similar' scores from other observations that share the same background variables. This means they are a random draw of a sub sample of respondents that have similar scores on a set of matching variables. Hot-deck imputation underestimates standard errors. Further it is bad for estimating measures of association and can result in biased regression coefficients and correlations. [8, p.49]

Predictive Mean Matching

This approach is kind of a combination of regression and hot-deck imputation. It is similar to the regression method except that for each missing value, it imputes a value randomly from a set of observed values whose predicted values are closest to the predicted value for the missing value from the simulated regression model. Or shortly said it takes the observed value from someone with a similar predicted value. The advantage of this approach is that imputed values are possible even in the case of bounds. [16]

Multiple Imputation

Multiple imputation generates several copies of the data and for each copy it imputes the estimates of the missing data values. Imputation itself in each of the copies is carried out using single imputation methods. Multiple imputation requires a MAR or MCAR mechanism. This method will be used in this work. Further details about this method can be found in chapter 3. [8]

Maximum Likelihood

Unlike most above mentioned methods, maximum likelihood does not impute the missing values. It does not fill in missing values in the data. Moreover it is used to calculate a statistic in a missing value case but is only returning estimates of the statistics like the mean, variance or correlation. Like multiple imputation it requires a MAR or MCAR mechanism. Additional information on maximum likelihood imputation can be found in Enders (2010). [8, p.113]

Random Forest

Random forest missing data approaches use random forest techniques for imputation. They can give unbiased results for MAR and MCAR mechanisms. Further, they are able to accommodate for e.g. interactions. One of the random forest missing data approach strategies is for example as follows. At the begin the data should be preimputed, then for each variable with missing values a forest is grown and further used to predict the missing values. The missing values get updated with the predicted values and finally this procedure is iterated for improved results. Further information on random forests is available in chapter 4.5. Additional information specifically on random forest imputation techniques can be found in Tang (2017) [21].

Algorithms for MNAR Data

Examples of algorithms for MNAR data are the selection models and the pattern mixture model. These algorithms attempt to describe the probability of missingness and the joint distribution of the data. The problem is that selection models rely on distributional assumptions that cannot be tested on. Pattern mixture models require users to specify assumed values for at least one inestimable parameter. For both cases, it is impossible to test on these assumptions. Violations or wrong specifications of these assumptions can introduce more bias than a MAR-based analysis such as multiple imputation. Specifying wrong values can produce considerable bias even with MAR-data. [8, pp.326ff]

For more information on the pattern mixture model see Little (1993) [15]. For more information on the selection models see Heckman (1976) [12].

3 Multiple Imputation

Multiple Imputation (MI) was developed by Rubin (1987) [20] within the Bayesian framework. It assumes at least a missing at random (MAR) mechanism [8, p.187]. MI generates several copies of the data and for each copy it imputes the estimates of the missing data values. Imputation itself in each of the copies is done by imputing values for missing data using single imputation methods. An advantage is that multiple imputation can reflect uncertainty about the values to impute. It reflects sampling variability which would also exist if there were no missing data and it can also reflect variability that exists due to the uncertainty about the reasons of non-response. [19, p.38]
There are three phases of the multiple imputation analysis. The first phase is the imputation phase. It is an iterative procedure, relying on Bayesian estimation principles, to create m copies of the data set. The second phase is the analysis phase. Here, complete data methods are used to perform the desired analysis for each copy of the data set. The last phase is the pooling phase, where the m estimates of the analysis phase are combined to a single set of results. [8, p.187]

3.1 Introduction to Bayesian Estimation

As already mentioned, multiple imputation is a Bayesian estimation approach. In a Bayesian framework a parameter is a random variable with its own distribution. This is the difference to many disciplines, where a parameter is a fixed characteristic of the population. This changes the interpretation of for example the confidence interval, or Bayesian credible interval. The interpretation of such an interval is that the parameter falls between the values of the lower boundary and the upper boundary. A credible interval attaches the probability to the parameter itself instead of the data. [8, p.165]
The three steps of a Bayesian analysis are the following. First, a prior distribution for the parameter of interest is specified. Second, a likelihood function summarizes the data's information for the parameter of interest. Third, the information of the likelihood and the prior are combined to construct the posterior distribution that describes the relative probability of different parameter values. [8, p.165]

$$Posterior \propto Prior \times Likelihood \tag{1}$$

For the specification of the prior distribution three hyperparameters are needed. They are the location of the distribution (e.g. the mean), the spread of the distribution (e.g. the variance) and the number of hypothetical data points. [8, p.169]

Often a non-informative prior distribution is used. This non-informative prior is also called Jeffrey's prior. Jeffrey's prior changes for different likelihoods. Using conjugate distributions has the advantage that the posterior distribution also belongs to the same distribution family as the likelihood and the prior. [8, p.173]
The basic idea behind the posterior distribution is to weight each point on the likelihood function by the magnitude of the prior [8, p.167]. The underlying Bayes theorem is

$$P(\theta|Y) = \frac{P(\theta)P(Y|\theta)}{P(Y)} \Rightarrow Posterior = \frac{Prior \times Likelihood}{ScalingFactor} \tag{2}$$

where θ is the parameter of interest, Y is the sample data, $P(\theta)$ is the prior distribution of the parameter, $P(Y|\theta)$ is the likelihood, $P(Y)$ is the marginal distribution of the data and $P(\theta|Y)$ is the posterior distribution. [8, p.170]
Note that equation (1) equals equation (2), only that the denominator is left out.

3.2 Imputation Phase

There exist a number of algorithms for the imputation phase. The algorithm that will be used in this work is called fully conditional specification (FCS). FCS is also referred to as sequential regression (multivariate) imputation (SRMI) or chained equations. It is a semi-parametric approach that specifies the multivariate imputation model by a series of conditional models. [22, p.219]
Its big advantage is that every variable with missing data gets its own model. This ensures that it is quite easy to handle non-normal data such as categorical variables. Therefore it is an approach that imputes the data on a variable-by-variable basis. It can produce unbiased parameter estimates and standard errors.
Generally said, it is a Bayesian approach that specifies an explicit model for each variable with missing values in a manner that they are conditional on the fully observed variables and their prior distribution. The result is a posterior predictive distribution of the missing values conditional on the observed values for each variable. The imputations are drawn from the posterior distribution. Thus, this approach is fully conditional on all the observed information. In many cases a non-informative prior will be used. [18]
Let $X = (X_1, ..., X_l)$ be a vector of l complete variables and let $Y = (Y_1, ..., Y_k)$ be a set of k incomplete variables. The matrix x with dimension $n \times l$ is an i.i.d. sample of the vector X and the matrix y with dimension $n \times k$ is an i.i.d. sample of the vector Y. The matrix y can also be illustrated by the vectors $y = (y_1, ..., y_k)$ with $y_i = (y_{i1}, ..., y_{ik})$. The part of the missing data in y is denoted y^{mis} and the observed part y^{obs}. Let $y_{-j} = (y_1, ..., y_{j-1}, y_{j+1}, ..., y_k)$ be the $k-1$ variables in y

except y_j. Furthermore, let $R = (R_1, ..., R_k)$ be a set of response indicators with $R_{-j} = (R_1, ..., R_{j-1}, R_{j+1}, ..., R_k)$. For the response indicator R_j holds

$$R_j = \begin{cases} 1, Y_j \text{ is observed} \\ 0, Y_j \text{ is missing} \end{cases}$$

The imputation of y_j^{mis} is based on the relation between the predictors $y_{-j} = (y_1, ..., y_{j-1}, y_{j+1}, ..., y_k)$ and x and the incomplete variable y_j. This is the MAR missing data scenario.

To create multiple imputations y^* of y^{mis} the following procedure is applied.

1. Calculate the posterior distribution $P(\theta|x, y^{obs}, R)$ of θ based on the observed data y^{obs}, the complete variables x and the response indicator R.
2. Draw a value θ^* from $P(\theta|x, y^{obs}, R)$.
3. Draw a value y^* from the conditional posterior distribution of y^{mis} given $\theta = \theta^*$, x and R, $P(y^{mis}|x, y^{obs}, R, \theta = \theta^*)$.

For multiple imputations steps two to three are repeated M times.

In the case of multivariate y, explicitly or implicitly getting the multivariate distribution of θ is the main problem. To obtain a posterior distribution of θ, FCS samples iteratively from separate conditional distributions of the form

$$P(Y_j|X, Y_{-j}, R, \theta_j) \text{ for each variable } Y_j, j = 1, ..., k. \tag{3}$$

The parameters θ_j are taken as specific to the respective conditional densities. They are not necessarily the product of some factorization of the true joint distribution $P(Y, X, R|\theta)$. It is possible to draw values from the conditional distributions in equation (3) through a Gibbs sampler. A Gibbs sampler is sampling from a conditional distribution, because that is simpler than marginalizing by integrating over a joint distribution. In this scenario, a Gibbs sampler is used to sample values θ^* and y^*. The initial values can be determined randomly. From there on, it samples each θ_j^* and y_j^* from distributions conditioned on all other components,

$$\begin{aligned}
\theta_1^{*(t)} &\backsim P\left(\theta_1|x, R_1, y_1^{obs}, y_2^{(t-1)}, ..., y_k^{(t-1)}\right) \\
y_1^{*(t)} &\backsim P\left(y_1^{mis}|x, R_1, y_1^{obs}, y_2^{(t-1)}, ..., y_k^{(t-1)}, \theta_1^{*(t)}\right) \\
&\vdots \qquad\qquad , \\
\theta_k^{*(t)} &\backsim P\left(\theta_k|x, R_k, y_k^{obs}, y_1^{(t)} y_2^{(t)}, ..., y_{k-1}^{(t)}\right) \\
y_k^{*(t)} &\backsim P\left(y_k^{mis}|x, R_k, y_k^{obs}, y_1^{(t)}, ..., y_{k-1}^{(t)}, \theta_k^{*(t)}\right)
\end{aligned} \tag{4}$$

where t resembles the t-th iteration of the Gibbs sampler. Samples of variable j in the t-th iteration are based on its observed part, the complete variables x, the

parameter $\theta_j^{*(t)}$ of the t-th iteration and the completed variables y_{-j} of the t-th iteration, if they were sampled before variable j, or of iteration $t-1$ if they will be sampled after variable j. The Gibbs sampler returns the final imputation value y^* from step 3 in above described procedure.
To generate M multiple imputations, the iterations of equation (4) are executed M times in parallel. An assumption of this approach is that the joint distribution is specified by equation (3) and the Gibbs sampler in equation (4) draws from this joint distribution. Note that the algorithm includes already imputed data as complete data. This means that for the first imputation of Y_2 in iteration one Y^{obs} is updated by Y_1. Starting with the last variable to impute at iteration one, the model for imputation of missing values at that variable uses all available data. In iteration two and later, every model for missing values at any Y_j, $j = 1, ..., k$, will use all available data. This also means that in iteration one for the first variables only a subset of cases with complete data in all predictor variables is used. [23]

Compatibility

There exists the possibility that a set of conditional distributions has no multivariate density. This so called incompatibility of conditionals is a theoretical weakness of FCS. This is due to the fact that in this scenario the multivariate distribution, the implicit joint distribution to which the algorithm converges, is unknown. This could make the assessment of convergence ambiguous. Compatibility in data can be destroyed e.g. by rounding errors. However van Buuren et al. (2006) [23] show that FCS is quite robust against violations of compatibility in a set of simulations. If the conditionals are compatible, FCS is guaranteed to work. [22]

Convergence

Monitoring convergence is achieved by plotting the draws in each of the M sampling streams against the iterations. The paths should be inspected for any absence of trend and be freely intermingled with each other. [24, p.37]

Ignorability

The imputation model for variable j is $P(Y_j|X, Y_{-j}, R)$. It utilizes relations between X, Y and R. It is only possible to fit models for $P(Y_j|X, Y_{-j}, R = 1)$ and not for $P(Y_j|X, Y_{-j}, R = 0)$. However, imputations have to be drawn from
$P(Y_j|X, Y_{-j}, R = 0)$. Under MAR or MCAR it is possible to simply set
$P(Y_j|X, Y_{-j}, R = 0) = P(Y_j|X, Y_{-j}, R = 1)$. If this is not possible, because the data is MNAR and therefore the assumption of ignorability does not hold, MI will still work if it is possible to specify $P(Y_j|X, Y_{-j}, R = 0)$ so that it reflects the missing mechanism. Errors in the specification of $P(Y_j|X, Y_{-j}, R = 0)$ would introduce bias to the imputations. [22]

Advantages and Disadvantages

One advantage of FCS is its flexibility in creating multivariate models. Specialized imputation methods that are difficult to formulate as part of a multivariate density $P(X, Y, R|\theta)$ can be used. FCS gives the possibility to preserve unique features in the data such as bounds, interactions or skip patterns. Furthermore, constraints between variables to avoid logical inconsistencies in the imputed data can be maintained. [22]

A disadvantage is that FCS requires some modeling effort, as each model needs to be specified. Another disadvantage is its lack of a satisfactory theory. [22]

Number of Covariates

Rubin said that "the advice has always been to include as many variables as possible when doing multiple imputation." [8, p.133]

It is of importance that especially all variables that will be taken as predictors in the analysis phase are already included in the imputation phase. The same applies for any interactions or transformations like e.g. quadratic predictors.

Number of Imputations

Originally five was considered to be enough from an efficiency point of view. However, as multiple imputation standard errors decrease as the number of imputations M increase, it is favourable to create more imputed data sets. A large M can also improve power. [8, p.212]

3.3 Imputation Methods

For this chapter, y will be the vector of the dependent variable and x will be a matrix of the independent variables.

Logistic Regression Imputation

For binary variables a Bayesian logistic regression model is used in this work, as proposed in Rubin (1987) [20]. Let y be the dependent binary variable whose missing values should be imputed and let $x_1, ..., x_p$ be the set of numerical predictor variables. For these predictors, possible categorical variables are replaced by their corresponding dummies. The general logistical regression model is

$$ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$$

where $\pi = P(y = 1|x_1, ..., x_p)$ is the conditional density that $y = 1$ given the values of the predictor variables $x_1, ..., x_p$ and $\beta = (\beta_0, ..., \beta_p)$ is a vector of regression coefficients. More on the general logistic regression model is presented in chapter 4.1. The regression model is calculated only for observed data. If all missing data in the predictor variables is already in imputed, then the regression model is calculated on all available data. An imputation y^* is generated according to the following scheme: First fit a logit model and calculate $\hat{\beta} = (\hat{\beta}_0, ..., \hat{\beta}_p)$, the maximum likelihood estimator of $\beta = (\beta_0, ..., \beta_p)$, by an iterative least square algorithm and estimate the posterior covariance matrix of β, $V(\hat{\beta})$.
Second draw $\hat{\beta}^* = (\hat{\beta}_0^*, \hat{\beta}_1^*, ..., \hat{\beta}_p^*)$ from the approximate posterior distribution $N\left(\hat{\beta}, V(\hat{\beta})\right)$. Third calculate

$$\pi_i^* = \frac{1}{1 + exp(-(\hat{\beta}_0^* + \hat{\beta}_1^* X_{i1} + ... + \hat{\beta}_p^* X_{ip}))}$$

for $i = 1, ..., n_{mis}$, where n_{mis} represents the number of missing observations in y. Finally draw $u_i \sim \text{unif}(0, 1)$, $i = 1, ..., n_{mis}$. If $u_i > \pi_i$, impute $y_i^* = 0$, otherwise $y_i^* = 1$. [6, pp.93f.]

Polytomous Regression Imputation

For categorical variables with more than two levels polytomous (multinomial) logistic regression was applied. Let y be the dependent categorical variable with unordered categories $0, ..., s - 1$ whose missing values should be imputed and let $x_1, ..., x_p$ be the set of numerical predictor variables. For this x, possible categorical variables are replaced by their corresponding dummies. Polytomous regression is modeled as a set of $s - 1$ separate logistic regression models against a baseline category 0 according to

$$ln\left(\frac{P(y = j|x)}{P(y = 0|x)}\right) = \beta_{j0} + \beta_{j1} x_1 + ... + \beta_{jp} x_p, \text{ for } j = 1, ..., s - 1.$$

An imputation Y^* is generated according to the following scheme:
First draw $\hat{\beta}^*$ from the approximate posterior distribution $N\left(\hat{\beta}, V(\hat{\beta})\right)$, where $V(\hat{\beta})$ is the estimated covariance matrix of $\hat{\beta}$ and $\hat{\beta}$ is the maximum likelihood estimator of β. Note, that in this case β is a vector of regression coefficient vectors $\beta_j = (\beta_1, ..., \beta_{s-1})$ where each $\beta_j^T = (\beta_{j0}, ..., \beta_{jp})$ corresponds to the regression coefficients of each separate logistic regression model. Thus, β is a matrix with dimension $p \times n - 1$. Second, let

$$\pi_{ij}^{mis} = \frac{exp(-(\hat{\beta}_{j0}^* + \hat{\beta}_{j1}^* x_{i1}^{mis} + ... + \hat{\beta}_{jp}^* x_{ip}^{mis}))}{1 + \sum_{\nu=1}^{s-1} exp(-(\hat{\beta}_{\nu 0}^* + \hat{\beta}_{\nu 1}^* x_{i1}^{mis} + ... + \hat{\beta}_{\nu p}^* x_{ip}^{mis}))},$$

where $i = 1, ..., n_{mis}$ and $j = 0, ..., s-1$, with $\hat{\beta}_0^T = (\hat{\beta}_{00}, ..., \hat{\beta}_{0p}) = 0$. This means that π_{ij} is the probability that the i-th missing data entry is equal to the j-th category of y corresponding to the drawn regression coefficients $\hat{\beta}_j^*$.
Third generate imputations y_i^* for each missing data entry y_i, such that $y_i^* = j$ with probability π_{ij}^{mis} for $i = 1, ..., n^{mis}$ and $j = 0, ..., s-1$. [6, pp.94f.]

Proportional Odds Regression Imputation
For ordered categorical variables proportional odds logistic regression was applied in this work. The algorithm follows basically the one for unordered categorical variables with the biggest difference being the use of a proportional odds logistic regression model instead of a polytomous logistic regression model.
In all three imputation methods data augmentation according to the method of White, Daniel and Royston (2010) [26] is used in order to avoid bias due to perfect prediction.

3.4 Analysis Phase

In the analysis phase complete data-methods are used to analyze the filled-in data sets from the preceding step. The statistics of interest are calculated M times, once for each filled-in data set from the imputation phase. The results are M statistics of interest. The M statistics of interest differ only because the imputations differ. [8, pp.218f.]

3.5 Pooling Phase

The pooling phase is returning a single estimate of the statistics of interest. This is achieved by combining the M statistics of interest from the analysis phase. The pooling parameter estimate or multiple imputation point estimate for the estimates from the regression phase is often the arithmetic average of these estimates. [20]
Its formula is

$$\bar{\theta} = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_m \tag{5}$$

where $\bar{\theta}$ is the pooled estimate and $\hat{\theta}_m$ is the parameter estimate from data set m [8, p.219]. In the multivariate case, $\hat{\theta}_m$ and $\bar{\theta}$ are column vectors [8, p.234].
The pooled sampling variance consists of two parts. The within-imputation variance and the between-imputation variance.

The equation of the within-imputation variance V_W is

$$V_W = \frac{1}{M} \sum_{m=1}^{M} SE_m^2 \tag{6}$$

where SE_m^2 denotes the square of the sampling variance from data set m. It is the mean of the M estimated sampling variances from the analysis phase. So the within-imputation variance estimates the variance that would have resulted if there were no missing data. Its multivariate analogous, the within-imputation covariance matrix looks like this:

$$V_W = \frac{1}{M} \sum_{m=1}^{M} var(\hat{\theta}_m)$$

where V_W is the average within-imputation covariance matrix and $var(\hat{\theta}_m)$ is the parameter covariance matrix from data set m. [8, p.234]
The between-imputation variance is the part of the variance that results from the fact that there is missing data. It resembles the variability of the M parameter estimates. Its equation is

$$V_B = \frac{1}{M-1} \sum_{m=1}^{M} (\hat{\theta}_m - \bar{\theta})^2 \tag{7}$$

with V_B being the between-imputation variance, $\bar{\theta}$ being the point estimate from equation (5) and $\hat{\theta}_m$ being the parameter estimate from data set m. The multivariate analogous is the between-imputation covariance matrix:

$$V_B = \frac{1}{M-1} \sum_{m=1}^{M} (\hat{\theta}_m - \bar{\theta})(\hat{\theta}_m - \bar{\theta})^T \tag{8}$$

where $\hat{\theta}_m$ and $\bar{\theta}$ are vectors and V_B is a covariance matrix. The diagonal elements of V_B contain the between-imputation variance estimates and the off-diagonal elements qualify the relationship between two between-imputation fluctuations in two parameters. [8, pp.234f.]
The total sampling variance is a combination of these two variances, in detail:

$$V_T = V_W + V_B + \frac{V_B}{M} \tag{9}$$

The V_B/M from (9) is due to the sampling variance of the mean. The mean or average parameter estimate $\bar{\theta}$ from (5) also has a sampling error. This term serves as a correction factor for using a finite number of imputations. [8, pp.222f.]
In the multivariate case V_T becomes the total parameter covariance matrix which

reflects the total sampling fluctuation in a set of parameter estimates. Again, the diagonal elements contain sampling variances and the off-diagonal elements contain covariances between pairs of estimates. [8, p.235]

3.6 Fraction of missing information

The fraction of missing information describes the influence of missing data on the sampling variance of a parameter estimate. More precisely it is the proportion of the total sampling variance that exists due to missing data. Its equation for an infinite (very large) number of imputations is as follows:

$$FMI = \frac{V_B + \frac{V_B}{M}}{V_T}$$

If the number of imputations is finite, the equation changes to:

$$FMI = \frac{V_B + \frac{V_B}{M} + \frac{2}{\nu+3}}{V_T}$$

with ν being the number of degrees of freedom. The degrees of freedom can be calculated as follows:

$$\nu = (M-1)\left(1 + \frac{V_W}{V_B + \frac{V_B}{M}}\right)^2 = (M-1)\left(\frac{1}{FMI^2}\right) \tag{10}$$

The degrees of freedom ν increase as the number of imputations M increase or the fraction of missing information FMI decreases.
As the multiple imputation degrees of freedom can substantially exceed the complete data degrees of freedom (in small and moderate samples), an adjusted version of the MI degrees of freedom can be used to correct this problem:

$$\nu_1 = \left(\frac{1}{\nu} + \frac{1}{\tilde{\nu}}\right)^{-1}$$

where

$$\tilde{\nu} = (1 - FMI)\left(\frac{df_{com} + 1}{df_{com} + 3}\right) df_{com}$$

and df_{com} is the number of degrees of freedom of the complete data case. [4]
The adjusted degrees of freedom increase as the sample size increases and never exceeds the complete data degrees of freedom. Typically the missing data rate is higher than the fraction of missing information. This is valid especially if the variables in the imputation model are predictive of the missing values because then the correlations mitigate the information loss.

It is a useful diagnostics tool, as it influences the convergence of the data augmentation algorithm. Parameters which have high rates of missing information normally converge more slowly.
The fraction of missing information tends to be noisy and untrustworthy, especially for less than 100 imputations. [8, pp.225f.]

4 Models

In the following, classification models that can be used in the analysis phase will be presented. Classification is a fundamental issue in machine learning and data mining. The goal is to construct a classifier, given a set of training examples, to predict the class of future cases. The outcome in this work is of binary nature, meaning it has two classes. This is a typical framework of machine learning, where first a model is trained and its accuracy of predictions is tested on some test set, to find the best model. Subsequently predictions are made on new data that was not used in training or testing.
All the following models can produce classifiers that return a discrete class label but also a real valued prediction.

4.1 Logistic Regression

Logistic regression is a special case of the generalized linear model. In binomial logistic regression a single binary outcome variable y_i, $i = 1, ..., n$ follows a Bernoulli probability function, $y_i \sim Bernoulli(y_i|\pi_i)$ where

$$P(y_i = 1) = \pi_i = \frac{1}{1 + exp(-x_i^T \beta)}. \tag{11}$$

In equation (11), x_i^T is a vector of independent variables and β is a vector of regression coefficients including an intercept. Together they form the linear predictor which is referred to as $\eta_i = x_i^T \beta$. The matrix equivalent of x_i^T is denoted X with dimension $n \times p$. The inverse of equation (11) is the log odds or logit link function

$$g(\pi_i) = ln\left(\frac{\pi_i}{1 - \pi_i}\right) = \eta_i = x_i^T \beta.$$

Logically, the equation of the odds is as follows:

$$odds = \frac{\pi_i}{1 - \pi_i} = exp(x_i^T \beta).$$

This implies that, if x_k is increased by one, then the chance for y is changed by $exp(\beta_k)$.
The parameters are estimated by maximum likelihood which assumes independence of the observations. The equation of the likelihood is denoted as follows:

$$L(\beta) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}.$$

The log-likelihood function becomes

$$l(\beta) = \ln(L(\beta)) = \sum_{i=1}^{n} y_i \ln(\pi_i) + (1-y_i)\ln(1-\pi_i). \tag{12}$$

The maximum likelihood estimate, $\hat{\beta}_{ML}$, is then calculated by deriving the log-likelihood from equation (12) and equaling it to zero. [14]

4.2 Logistic Regression with Mixed Effects

Mixed effects models are models that include random effects and fixed effects. Let X be the model matrix for p fixed effects with dimension $n \times p$ and let Z denote the model matrix for the q random effects with dimension $n \times q$. The main difference to the logistic regression model is that a stochastic component is included in the linear predictor:

$$\eta = X\beta + Zb, \tag{13}$$

where b is a vector of unknown random effects that usually is assumed to be normally distributed, $b \sim N(0, \Sigma(\theta))$, where Σ resembles the variance-covariance matrix with dimension $q \times q$ and θ resembles a parameter vector determining $\Sigma(\theta)$. In same notation as above equation (13) becomes

$$\eta_i = x_i^T\beta + z_i^T b, \tag{14}$$

where x_i is the i-row of X and z_i is the i-th row of Z. Logit link function and odds are calculated in the same way as for logistic regression in chapter 4.1, the only difference being the linear predictor. However, the likelihood changes as now there are two parameters β and θ that need to be maximized given the observed data y. The likelihood is the numerically equivalent to the marginal density of y given β and θ. It is denoted as

$$L(y|\beta,\theta) = \int p(y|\beta,b) f(b|\Sigma(\theta))db, \tag{15}$$

where $f(b|\Sigma)$ is the probability density at b and $p(y|\beta, b)$ is the probability mass function of y, given β and θ. If $p(y|\beta, b)$ is binomial, the integral in equation (15) has no closed-form solution and as a result must be approximated. One of the possibilities for this is the Laplace approximation. The conditional modes of the random effects are determined for given values of β and θ by

$$\tilde{b}(\beta, \theta) = \arg\max_b p(y|\beta, b) f(b|\Sigma(\theta)). \tag{16}$$

These are the values of the random effects that maximize the integrand of equation (15). [5]

A penalized iteratively reweighted least squares algorithm (PIRLS) can be used to determine the conditional modes of equation (16). In this algorithm, an offset, $X\beta$, is applied to incorporate the contribution of the fixed effects parameters β. To incorporate the contribution of the variance components, θ, a penalty term in the weighted least squares fit is used. For a detailed description of the PIRLS algorithm see Bates (2011) [5].

To get approximate values of the maximum likelihood estimate for the parameters and the corresponding conditional modes of the random effects Laplace approximation is used. The Laplace approximation to the likelihood in equation (15) is carried out by replacing the integrand of that likelihood. It is replaced by the second order Taylor series approximation to the log of the integrand at the conditional modes of equation (16). The approximation on the scale of the deviance (negative twice the log-likelihood) is

$$\begin{aligned} -2l(\beta, \theta|Y) &= -2\log\left\{\int p(y|\beta, b) f(b|\Sigma(\theta)) db\right\} \\ &\approx 2\log\left\{\int exp\left\{-\frac{1}{2}\left[d(\beta, \tilde{b}, y) + \tilde{b}^T\tilde{b}^* + \tilde{b}^T D^{-1} b\right]\right\} db\right\} \\ &= d(\beta, \tilde{b}, y) + \tilde{b}^{*T}\tilde{b}^* + \log|D|, \end{aligned}$$

where $\tilde{b}^*$ are the conditional modes from the PIRLS algorithm at convergence and D is an approximation of the variance-covariance matrix of these conditional modes. Furthermore, $d(\beta, \tilde{b}, y) = -2\log(p(y|\beta, b))$ is the deviance function from the linear predictor only. The sum of the deviance residuals can be used for the evaluation of this quantity. [5, pp.27-31]

4.3 Decision Trees

A decision tree can be seen as a representation of a decision procedure with the aim of determining the class of a given instance. It consists of nodes and links (branches).

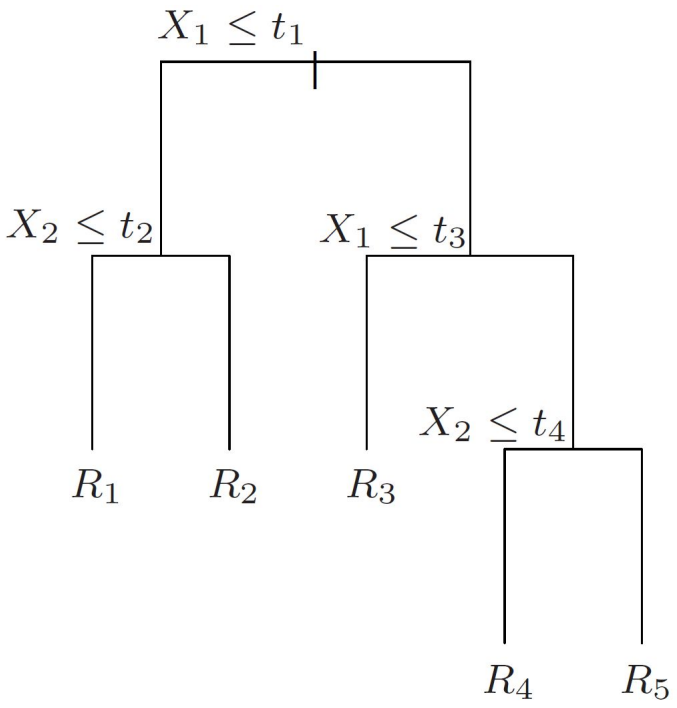

Figure 1: Exemplary decision tree, source: [11, p.306]

Hereby, nodes represent a feature or attribute and each branch represents a decision or rule. The leafs represent the outcome. In best case scenarios leafs are pure and contain only a single outcome.

Building a tree starts with finding the single variable that best splits the data in two groups. This variable should then serve as the root node. In the simple example of Figure 1 with two continuous input variables X_1 and X_2 and a continuous output, the first split is taken at $X_1 = t_1$ where t_1 is a threshold of some sort that splits the data in two. This serves as the root node.

The data gets separated by this criterion and then separately the single variable that best splits each subgroup is used to split the data further. Here this means that the region $X_1 > t_1$ is split at $X_1 = t_3$ and the region $X_1 \leqslant t_1$ is split at $X_2 = t_2$. This process is applied until no improvement can be made or the subgroups reach a minimum size. Here, on the left side such criteria are already met and the results are the leaves or terminal nodes that correspond to regions R_1 and R_2. On the right hand side, the region where $X_1 > t_3$ is split further at $X_2 = t_4$ which results in the terminal nodes corresponding to R_4 and R_5. The region where $X_1 \leqslant t_3$ already reached its terminal node this node corresponds to R_3. [11, pp.305f.]

Let the outcome be a categorical variable with K levels. For each of the n observations, the data consists of p inputs and a response, (x_i, y_i), with $x_i = (x_{i1}, x_{i2}, ..., x_{ip})$ for $i = 1, 2, ..., n$. The partition consists of M regions, $R_1, R_2, ..., R_M$.

The proportion of observations corresponding to class k, $k = 1, ..., K$ in node m,

$m = 1, ..., M$ is

$$\hat{p}_{mk} = \frac{1}{n_m} \sum_{x_i \in R_m} I(y_i = k), \tag{17}$$

where n_m is the number of observations in region R_m. The majority class in node m is classified by the observations in this node to class $k(m) = \arg\max_k \hat{p}_{mk}$. Node impurity measures $Q_m(T)$, where T represents the tree, include the Gini index,

$$Q_m^{Gini}(T) = \sum_{k \neq k'} \hat{p}_{mk}\hat{p}_{mk'} = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) = 1 - \sum_{k=1}^{K} \hat{p}_{mk}^2,$$

and the cross entropy or deviance,

$$Q_m^{CE}(T) = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}.$$

In the case of $K = 2$ with p being the proportion in the second class, the measures are $2p(1 - p)$ for the gini index and $-p\ \log(p) - (1 - p)\log(1 - p)$ for the cross entropy.

The minimum value of the Gini index is zero. This equals the case of perfect separation where all data belongs to the same class. The maximum value would be at $1 - 1/k$. The maximum resembles the case when all target classes are equally distributed. Perfect classification is achieved at an entropy of zero. A higher entropy has higher potential for improvement of the classification.

The node impurity measures must be weighted by the number of observations in the child nodes that were created when splitting node m, namely n_{m_L} and n_{m_R}. They replace n_m in equation (17).

To build the tree, the impurity measure is calculated for the whole data set, using n_m in equation (17). Later this will be calculated for the existing branch before the consequent split and not any more for the whole data. To find the best split, the impurity measure is calculated for each input, using the weights of the child nodes. The input with the lowest impurity measure is chosen for the split if its impurity measure value is lower than the one of the whole data set. This procedure is repeated until the final, large tree T_0 is built. The splitting process is stopped as soon as some minimum node size is reached.

The size of the tree is important, as a small tree could be insufficient to capture the structure of the data. On the other hand, a large tree might overfit the data. The optimal tree size should be chosen from the data. One approach to find the optimal tree size is pruning. [11, pp.307-311]

Once the large tree T_0 is created, it gets pruned. One pruning approach is cost-

complexity pruning. This approach defines a sub-tree $T \subset T_0$. This can be any tree that can be obtained by pruning the large tree T_0. Pruning can be achieved by collapsing any number of its internal nodes. The terminal nodes itself are indexed by m representing the corresponding region R_m with N_m observations. A common way to define the cost complexity criterion is:

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha|T|,$$

where $|T|$ is the number of terminal nodes in the sub-tree T and $Q_m(T)$ is either the Gini index or cross entropy. Further, $\alpha \geq 0$ is a tuning parameter governing the trade-off between tree size and the corresponding goodness of fit to the data. For each α, the idea is to find the sub-tree $T_\alpha \subseteq T_0$ that minimizes $C_\alpha(T)$. Small values of α result in bigger trees T_α with the full tree T_0 as solution if $\alpha = 0$. Large values of α result in smaller trees T_α. For each α there exists a unique smallest sub-tree T_α that minimizes the cost complexity criterion $C_\alpha(T)$. To find T_α, weakest link pruning is carried out. It works by collapsing the internal node with the smallest per-node increase in $\sum_m N_m Q_m(T)$. This procedure is continued until the single root node tree is produced, resulting in a finite sequence of sub-trees. This sequence must contain T_α. The estimation of α is achieved by cross validation. More on cross validation can be found in chapter 5.2. The final tree is $T_{\hat{\alpha}}$, where $\hat{\alpha}$ is the value to minimize the cross validated impurity measure.

Splitting a categorical predictor with q unordered levels into two groups, gives $2^{q-1} - 1$ possible partitions of these values which could lead to prohibitive computational time for large q. With a binary outcome this can be simplified. The predictor classes are ordered according to the proportion falling in outcome class one. Then the predictor gets split as if it were an ordered predictor. Among the $2^{q-1} - 1$ splits, this results in the optimal split.

Categorical predictors with large q tend to be favoured by partitioning algorithms. The reason for this behaviour is that the number of partitions grows exponentially in q, increasing the chance to find a good one for the data at hand. As a consequence, severe overfitting can occur if a predictor has many levels, making such variables sub-optimal.

So far, splits were supposed to be binary, but it is also possible to have splits with more links. A problem with such multi-way splits is that they divide the data too quickly, which could lead to insufficient data at the next level down the tree. However, multi-way splitting can be achieved by a series of binary splits like X_1 in Figure 1. [11, pp.307-311]

Above two node impurity measures were described. This is due to the fact that there exist two major algorithms to build trees. CART (Classification and Regres-

sion Trees) and C5.0. The latter uses information-based criteria like the entropy as metrics to find the best split while the first uses the Gini index as a metric. [27, p.4] C5.0 has a unique feature to derive rule sets. It is sometimes possible to simplify the set of rules that define a terminal node. This can be executed if one or more condition can be dropped, but the subset of observations that fall into the same node is not changed. If the set of rules can be simplified, they no longer follow a tree structure.
Advantages of decision trees are that it is possible to understand how decisions are derived at each node. A disadvantage of decision trees is their high variance. Small changes in the data can make huge differences. This is due to the hierarchical structure of trees. Changes in the split defining the root node or one of the early nodes will propagate down to all the following splits. A solution could be to build multiple trees and average their results. One such approach is boosting which will be described in the following chapter. [11, p.312]

4.4 Boosted Trees

Single decision trees are prone to changes in the data and seldom provide the best possible achievable predictive accuracy. The idea behind boosting is to build many trees and take the weighted average over all trees which should be a lot more robust to changes in the data.
Boosting trees can dramatically improve their accuracy while maintaining many desirable properties. However, boosted trees sacrifice speed and interpretability. [11, p.352]
One of the most popular boosting approaches is called "AdaBoost.M1". This is the boosting approach taken in the C5.0 algorithm. In short words, misclassified events are re-weighted and the new tree is built with the re-weighted events. Furthermore, each tree is assigned to a score which will be used as weight when averaging over all trees.
AdaBoost starts with initializing observation weights ω_i by $\omega_i = 1/n$, $i = 1, 2, ..., n$. The algorithm will iterate over the number of the (weak) classifiers $G_m(x)$, $m = 1, 2, ..., M$. In each iteration round a classifier $G_m(x)$ using weights ω_i is fitted to the training data. In this work boosting is carried out with trees as classifiers. This classifier is then used to compute the corresponding weighted error rate,

$$err_m = \frac{\sum_{i=1}^{N} \omega_i I(y_i \neq G_m(x_i))}{\sum_{i=1}^{N} \omega_i}.$$

Then a weight

$$\alpha_m = \log((1 - err_m)/err_m)$$

is computed that will be used to weigh $G_m(x)$ at the calculation of the final classifier $G(x)$. Now, the individual weights get updated for the next iteration:

$$\omega_i \leftarrow \omega_i \exp(\alpha_m I(y_i \neq G_m(x_i))).$$

The idea is to increase the relative influence of observations that were misclassified by $G_m(x)$ when inducing the next classifier $G_{m+1}(x)$. This is achieved by scaling their weights by the factor $\exp(\alpha_m)$. As iterations proceed, the influence of observations that are difficult to classify correctly will increase. This forces each successive classifier to concentrate on the training observations that were misclassified by previous ones. The final classifier,

$$G(x) = \operatorname{sgn}\left(\sum_{m=1}^{M} \alpha_m G_m(x)\right),$$

is the weighted average over all classifiers $G_m(x)$ with sgn being the signum function. [11, pp.337ff.]

4.5 Random Forest

Random forests construct many decision trees at training and return the class that is the majority vote of the classes at testing. A difference to boosted trees from the previous chapter is the random selection of features at the nodes as only a subset of all possible features is available at each node. Decision trees in random forests are independent, whereas in boosted trees they depend on each other. Unlike boosting, where bias is reduced because trees are grown in an adaptive way, the bias in random forests is the same as that of any of the individual trees. Therefore, improvements in prediction are solely a result of variance reduction. Variance reduction is achieved by averaging over many trees and further by reducing the correlation between trees. The latter is achieved through random selection of input variables at the nodes in the progress of tree-growing. Before each split, $l \leqslant p$ of the input variables are chosen at random as candidates for splitting. In classification, typical values for l are $\sqrt{p}$ and the minimum value is one. [11, pp.587ff.]
In detail, classifications in random forests are achieved as follows. For the number of random forest trees, B, first a bootstrap sample Z^* of size N is drawn from the training data. Then a random forest tree T_b, $b = 1, ..., B$ is grown to the bootstrapped data. To grow the tree, for each terminal node of it the following steps are recursively repeated until the minimum node size is reached.
Out of the l selected variables, the one which produces the best split is picked and the node is split into two daughter nodes.

The output is the ensemble of trees $\{T_b\}_1^B$. To make predictions at a new point x, random forest obtains a class vote from each tree. The prediction is the majority vote,

$$\hat{C}_{rf}^B(x) = majority\ vote\{\hat{C}_b(x)\}_1^B,$$

where $\hat{C}_b(x)$ is the class prediction of the b-th random forest tree. [11, p.588]
Random forests should further correct the habit of overfitting to the training set that occurs with decision trees. Hereby, an important aspect is that increasing B does not cause the random forest sequence to overfit. [11, p.596]

4.6 Naive Bayes

The idea behind naive Bayes classification is to calculate the conditional probabilities for every factor given an event occurred. The outcome with the highest probability is then selected.
In this chapter, a case E is represented by a tuple of attribute values (variables) $(x_1, x_2, ..., x_p)$. The classification variable or outcome is represented by C and c are the values of C. As the outcome in this work is binary, C is binary with the two classes positive and negative. Recall the Bayes theory in chapter 3.1. Equation (2) stated the Bayes theorem,

$$P(c|E) = \frac{P(E|c)P(c)}{P(E)},$$

describing the frequency of class c given case $E = (x_1, ..., x_p)$. This is the probability that case E will be in class c. All attributes are assumed independent which leads to

$$P(E|c) = P(x_1, x_2, ..., x_p|c) = \prod_{j=1}^{p} P(x_j|c). \tag{18}$$

The shape of $P(x_j|c)$ in equation (18) depends on the type of the data x_j. If x_j is binary, than $P(x_j|c)$ is assumed to have the shape of a Bernoulli probability mass function which would lead to $P(x_j|C_k) = \pi_{kj}^{x_j}(1 - \pi_{kj})^{(1-x_j)}$ for class k. If x_j is categorical with more than two categories, $P(x_j|c)$ could have the shape of a multinomial probability mass function. If x_j is continuous, it might typically be assumed to be Gaussian distributed.
A case will be classified positive, if $P(C = positive|E) \geq P(C = negative|E)$. The

resulting naive Bayes classifier is:

$$f_{nb}(E) = \frac{P(C = positive)}{P(C = negative)} \prod_{j=1}^{p} \frac{P(x_j|C = positive)}{P(x_j|C = negative)}.$$

If $f_{nb}(E) \geq 1$, a case will be classified positive.
The assumption of independence between attributes means that the algorithm cannot learn the relationships between them. This is a disadvantage of naive Bayes as this assumption often does not hold in real-world applications. Its advantage is speed as it is a fast, highly scalable algorithm which is at the same time quite simple. [28]

5 Measures for Validation

As the classification problem is dichotomous, in the following mostly special cases for dichotomous outcome are regarded. Note that many of these techniques function as well on and are easy to adapt to categorical data with more than two classes.

5.1 Sampling Techniques

Most statistical classification models assume both classes to appear with more or less equal frequencies. If this is not the case, then the data is called imbalanced. The problem is that in imbalanced situations the minority class is typically of primary interest. Models induced over imbalanced data sets tend to have poor predictive accuracy with respect to the minority class. Therefore, sampling techniques are used to achieve a (almost) balanced data set. There are different sampling techniques available.

Oversampling
One approach is called oversampling. This approach randomly duplicates cases from the minority class to increase their population. The problem is that the increase of the total size of the data set is achieved by duplicates of existing data which may lead to over-fitting. As a result variables may appear to have lower variances than in reality.

Under-sampling
Another approach is under-sampling. The majority class is randomly down-sampled until the data set is balanced. A disadvantage is that valuable data gets thrown away which may lead to bias. Furthermore, independent variables can appear to have a higher variance than in reality. According to literature, under-sampling the majority class leads to better classifiers than oversampling the majority class. [7, p.326]

Synthetic Sampling

Synthetic sampling synthesizes new samples instead of resampling existing ones to achieve class equality. One such approach is the SMOTE (Synthetic Minority Oversampling TEchnique) algorithm. Basically the SMOTE approach works as follows. Among the minority class, find the l nearest neighbors of a data sample, ignoring cases from the majority class. Take the difference between the feature vector (sample) and its nearest neighbors and multiply it with a random number between zero and one. The result will be added to the feature vector under consideration and define a sample along the line segment between the two data samples. Repeat this procedure for each data sample in the minority class. If more synthetic samples are desired, the above process will be repeated, resulting in more than one new sample between two existing ones. [7]

So far this resembles an oversampling approach. But this approach can be combined with an under-sampling one, by first down-sampling the majority class to a certain amount and then oversampling using SMOTE. [7]

Important to remember about SMOTE is that it can only generate new samples within the body of existing minority samples.

In the case of nominal features, the above mentioned difference between nearest neighbors is not as self-explanatory as in the continuous case. In the nominal case a modified version of Value Difference Metric (VDM) is used. This metric looks at the overlap of feature values over all feature vectors. It creates a matrix that defines the distance between feature values. Certain distance elements of this matrix, δ are defined as follows.

$$\delta(V_1, V_2) = \sum_{i=1}^{n} |\frac{C_{1i}}{C_1} - \frac{C_{2i}}{C_2}|^k, \tag{19}$$

where k is a constant, usually set to one, V_1 and V_2 are the two corresponding feature values, C_{ji} is the number of occurrences of feature value V_j for class i and C_j is the total number of occurrences of feature value V_j, with $j = 1, 2$. Equation (19) equals a geometric distance on a fixed, finite set of values. Between two feature vectors X and Y the distance Δ is defined as follows:

$$\Delta(X, Y) = \omega_x \omega_y \sum_{i=1}^{N} \delta(x_i, y_i)^r, \tag{20}$$

where ω_x and ω_y are the exemplar weights in the modified VDM. $r = 1$ yields the Manhattan distance and $r = 2$ yields the Euclidean distance. For a new feature vector, the weight is $\omega_y = 1$ and the weight ω_x is the bias towards more reliable feature vectors with $\omega_x \approx 1$ for more accurate feature vectors. [7, pp.349ff.]

The weights in equation (20) are mostly ignored as SMOTE is not directly used for classification purposes. They can be redefined if the weight of the minority class feature vectors falling closer to the majority class feature vectors should be increased to make them appear further away. [7, pp.349ff.]
Another approach for imbalanced data would be to change the cost function in a way that increases the cost for misclassifications of minority instances in comparison to misclassifications of majority instances.

5.2 Cross Validation

Cross validation is a model evaluation method that can be used to compare different models or a set of different parameters for one model. In the latter case it can be used for getting the best parameter set of a model. Cross validation further gives an evaluation of the ability for predictions of the model.
The problem in the normal model fitting approach is that it is not possible to evaluate the performance of the model on new data, as all data was used to train the model. Therefore, the idea is to split the data and hold a (small) part of it back to test the fitted model. Cross validation partitions the data into a training set and a test set. The model is first fitted on the training set and then it is used to predict the unseen data of the test set. As the data of the test set is known, the predicted values can be evaluated.
One approach to cross validation is called k-fold cross validation. For this approach the original data is partitioned into k folds or sub-samples of equal size. One of the k sub-samples is then retained for the test set and the remaining $k-1$ folds are used as the training data. This process is then repeated k times such that each of the k sub-samples is once used as the test set. The k results are finally averaged to produce a single estimation. The measures used for validation of the methods will be introduced in the following sub-sections. [10]
The parameter k must be chosen carefully as poorly chosen values for it may result in high bias or high variance. Generally, the choice of k is a bias-variance trade-off, as both cannot be minimized at the same time. High variance occurs if too little data is left in the test set. This could lead to over-fitting and an untrustworthy estimate of error. High bias could appear if too much data was hold-out for the test set, leaving not enough information for a solid model in the training set. Therefore, increasing k generally reduces this bias as the test set gets smaller. Typically $k = 5$ or $k = 10$ is chosen. These values have been shown to yield estimates that suffer neither from very high variance nor high bias. [13, p.184]
The sub-samples can be partitioned randomly or in a stratified manner. For the latter it means that the mean response value is approximately the same in all folds. In the case of categorical data this means that the proportion of each class is roughly

the same in each fold.
For repeated k-fold cross validation the above procedure is repeated multiple times. The number of repetitions depends on the data itself but also on the computational power at hand as repeated k-fold cross validation can be very time-consuming, especially with larger data sets. In general, more repetitions reduce the probability that the results are the outcome of a certain partitioning of the data. This improves the validity of the results.
Another possible approach to cross validation is called leave-one-out cross validation. In this scenario k is set to the number of data points, n, and in each run one data point is assigned to the test set while the training set consists of $n - 1$ data points. Thus leave-one-out cross validation is run n times. [10]
All steps of the algorithm of the model must be repeated in each cross validation loop. Prior specifying implies significant bias. [25]
As an example take a combination of cross validation with oversampling. It is important that the sampling technique is carried out in each run of cross validation and not once before it. The reason is that in oversampling, cases from the minority class are duplicated to achieve equal proportions between both classes in the data. If this is carried out before the folds are defined, then it is very likely that the fold with the test data consists partly of duplicated data which was already used to calculate the method. This clearly counteracts with the concept of testing the calculated method on unseen data.

5.3 Confusion Matrix

A confusion matrix is a special kind of contingency table in the field of machine learning and especially supervised learning in classification problems. Each column represents the actual classes while each row represents a predicted class. It is an easy measure to see how good predictions actually are. Table 1 is an exemplary confusion matrix for a dichotomous outcome. Note that false positives (FP) resemble type I error while false negatives (FN) are equivalent to type II error. For the following equations, TP denotes the true positives and TN denotes the true negatives. Further, P is the number of real positive cases in the data and N will be the number of real negative cases in the data. [9]

	Actual Positive	Actual Negative
Predicted Positive	True Positives	False Positives
Predicted Negative	False Negatives	True Negatives

Table 1: Confusion Matrix

5.4 Measures from a Confusion Matrix

The commonly used measure for validation in dichotomous classification problems is accuracy which simply describes the percentage of correctly predicted cases. Its calculation is:

$$accuracy = \frac{TP + TN}{P + N} = \frac{TP + TN}{TP + TN + FP + FN}.$$

In general this is a good measure as it describes exactly the desired properties but in a case where the (dichotomous) data is highly imbalanced the accuracy may not be the right measure. This is due to the fact that in the case of unbalanced data accuracy is high if all data is predicted to be in the majority class. If the proportion between the two classes is 99 to one, then accuracy would be as high as 99 percent as only one percent would have a wrong prediction since it is originally in the minority class. Predicting cases to the minority class often comes at the cost of misclassifications of cases from the majority class which can easily reduce accuracy. In the following, measures will be introduced that may be superior for imbalanced data. One of these measures is the sensitivity or true positive rate (TPR),

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN}. \tag{21}$$

Another name for this measure is recall. Another measure called specificity is calculated as:

$$specificity = \frac{TN}{N} = \frac{TN}{TN + FP}. \tag{22}$$

Precision is the proportion of true positives among all positive predicted cases:

$$precision = \frac{TP}{TP + FP}. \tag{23}$$

Another often seen measure is the F_1 score which is the harmonic mean of precision and sensitivity,

$$F_1 = \frac{2}{1/precision + 1/TPR} = \frac{2TP}{2TP + FP + FN}.$$

[9]

5.5 ROC Curve

The receiver operating characteristics (roc) curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various thresholds. The

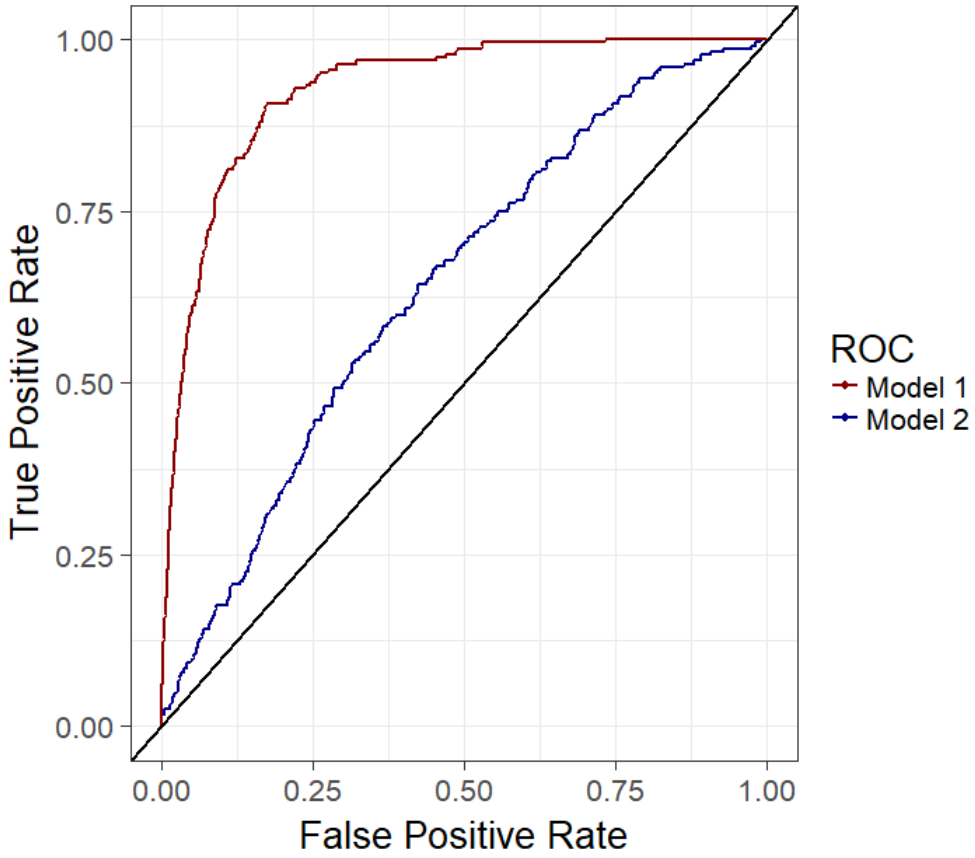

Figure 2: Exemplary roc curves, model 1 has an auc of 0.927 and model 2 has an auc of 0.641

equation for the false positive rate is

$$FPR = \frac{FP}{N} = \frac{FP}{FP + TN} = 1 - specificity.$$

The definition of the specificity can be found in equation (22) and TPR is defined in equation (21). The underlying idea behind roc curves is that distributions for positives and negatives are not equal. When choosing the rounding threshold, it is important to decide what is worse, increasing the false positives (false alarms) or the false negatives (misses).

Figure 2 depicts exemplary roc curves. Best prediction would go through the point (0,1) indicating that there are no false positives while all positives are depicted correctly. Generally said, a steep curve is desired. The angle bisector in the plot is the line for chance accuracy. If the curve follows this line, the underlying model has the predictive precision of coin flips. The graph shows that model one is clearly a better classifier than model two. It is not possible to directly read the threshold from a roc curve, but the TPR and FPR can be read directly. From there the best threshold can directly be read from the data.

Figure 3 depicts an example of the distributions of two classes. On the x-axis is the

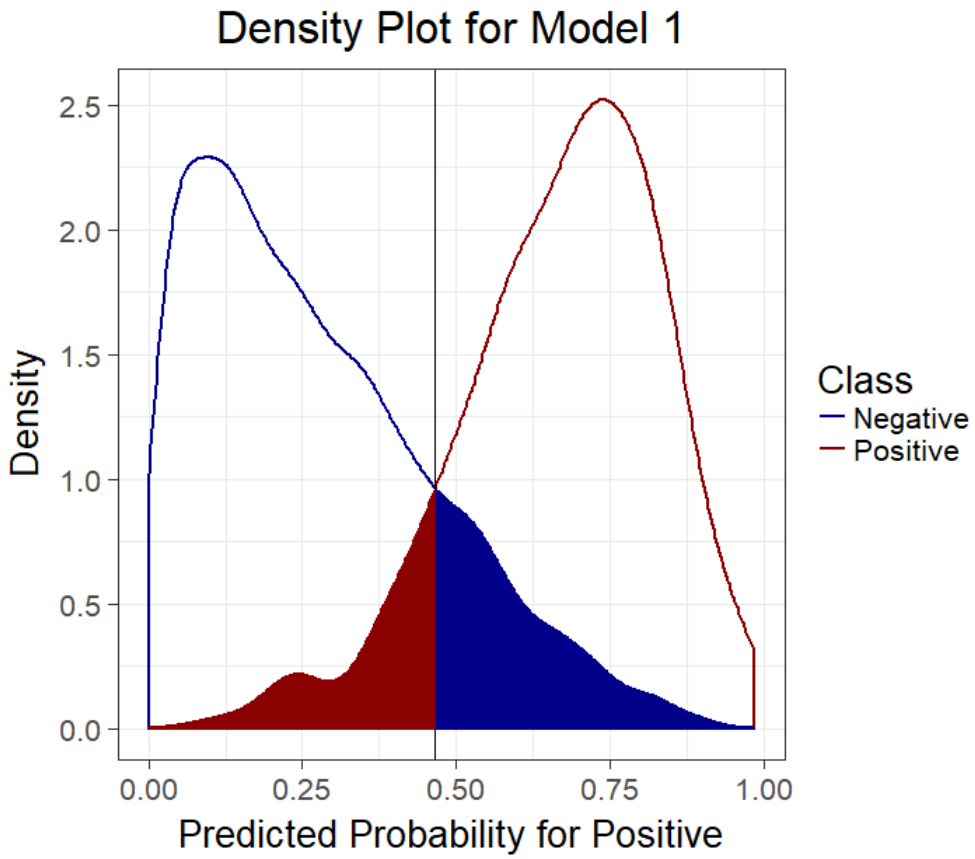

Figure 3: Density plot for model 1 from Figure 2, the blue filled area is the area for the false positives and the red filled area is the area for the false negatives at a threshold of 0.465

predicted probability for a positive test result. Depicted is a case that minimizes the total number of false classifications. If the cost for false negatives would be higher, then the threshold would be lowered and the red area (false negatives) be decreased. At the same time the blue area (false positives) would increase. The area for the true positives is the whole area under the red curve to the right of the threshold and the area for the true negatives is the whole area under the blue curve to the left of the threshold. The area under the curve or *auc* is a summary statistic of the roc curve which should depict in one number the quality of the classifier. It equals the probability that a random positive example will be ranked above a random negative example. Like the name implies the *auc* is the area under the curve and is calculated as follows:

$$auc = \int ROC(t)dt,$$

where $ROC(t)$ is the roc curve of a classifier t. The *auc* ranges between 0.5 and 1, where 0.5 is equivalent to random predictions and one to perfect predictions. Note that the underlying model from Figure 3 has a high auc of 0.927. [9]

For some specific cost and class distributions the classifier with the highest auc may not be the best. [7]
For the predicted cases it is best to get the probabilities instead of the classifications themselves. Classifications are just the rounded probabilities which were assigned to the class labels. Normally rounding is done with a 0.5 rounding threshold which may not be the best threshold for the available data. The optimal case would be to set the threshold such that the precision from equation (23) is one, because in that case there exist no false positives. At the same time the sensitivity from equation (21) should be high.

6 Description of Study and Descriptive Analysis

6.1 Description of Study

The National HIV & AIDS and Reproductive Health Survey (NARHS) is a nationally representative survey on human immunodeficiency virus (HIV) and acquired immunodeficiency syndrome (AIDS) in Nigeria [17]. There, it was the first such survey on HIV and AIDS [3, p.621].
The major objectives of the NARHS studies are to obtain HIV prevalence estimates and information on risk factors related to HIV infection. Other objectives are to monitor trends and changes in behavior which influence HIV & AIDS intervention strategies and to identify information gaps which need further exploration. Such knowledge determines Nigeria's response to the HIV & AIDS epidemic as it guides the development of appropriate HIV & AIDS intervention strategies. [17, pp.48f.]
The NARHS study consists of two parts. A survey on knowledge and behavior in fields that are relevant for HIV and a voluntary HIV test. Data was collected on sexual and reproductive health indicators. Respondents were females aged 15-49 years and males aged 15-64 years in Nigeria. Selection of respondents was based on a probability multi-stage sampling method, ensuring that respondents come from all over the country. [17, p.49]
HIV testing was carried out using five finger prick blood samples stored as dried blood spots (DBS) on the same filter paper. For identification a unique random identification number was assigned to each DBS and questionnaire. Testing itself was carried out using the enzyme-linked immunosorbent assay (ELISA) test of DBS of 10% of non-reactive, all reactive and all discordant specimens. [17, pp.53f.]

6.2 Description and Preparation of Data

In this work data of 2007 NARHS Plus and 2012 NARHS Plus II is used containing a total of 42756 respondents and 24 variables. Only data from respondents who were

successfully interviewed is included in the data set. This means that they answered to at least a subset of questions and did not refuse the interview totally.
There are 11521 respondents from the 2007 study and 31235 from the 2012 study. Respondents in both studies could attend a voluntary and free HIV test which 9610 or 22.48% of them refused to take. The percentage of deniers was quite similar in both studies. Item non-response does not only occur at the variable for the HIV test result but also at many other of the given variables. The given subset of variables from the two original studies includes variables representing the state, zone and location of living, the wealth, education, religion, age, gender and marital status of the respondent, the year of the study, his or her sexual behavior, the age at first sex as well as his or her knowledge about condoms and HIV, the result of the HIV test and if the respondent had contact with sexual transmittable infections in the last 12 months.
The data contains more variables but they do not contain additional information as they are mostly binary variables of categorical variables listed above. The variable `Religion`, which is mentioned above, was created using two variables with information if respondents are Muslims, Christians or have another religion.
There is also a categorical version of the age and a continuous version of the variable `AgeSexcat`, the age at first sex. For the age, the continuous version was used as it contains more information than the categorical version of the age. For the age at first sex, the categorical version was used, as the continuous version does not contain information about people who never had sex. At variable `AgeSexcat`, the level "No response" was removed and treated as missing values.
Further details about the given variables can be found in Table 2. Unit non-response does not exist in the available data set as every respondent answered the questionnaire at least partly.
In the data some illogical combinations could be found. Therefore, it was possible to do logical imputation in some cases. In detail, this was possible for eight respondents, who claimed to not have heard of condoms or did not answer to this question, `CDHeard`, but answered one of the sub-questions, `CD-AIDS`, `CD-STD`, `CD-Obtain` or `CD-Afford`, which were only possible to answer if a respondent had in fact heard of condoms. Respondents who had never heard of condoms, had originally missing values at sub-questions. To solve this, they were assigned to the level "Don't Know". For variable `CD-Afford`, this level did not exist and had first to be created. They were not assigned to a new level, for example "Not heard of CD", because this would have caused problems in some models later. The reason for these problems is that the new levels would contain the exact same observations as the level "No" in variable `CDHeard` and therefore would not add any new information but repeat already existing information. Further, a new variable, `CDagree`, was created as the overlap of data between the levels of the questions on condoms is rather big. This

new variable simply counts how often a respondent agreed to the statement in the sub-variables. The new variable has a continuous scale and no missing values. If there was a missing value at one of the variables used to create `CDagree`, it simply got ignored. The sub-variables are the same as above, namely `CD-AIDS`, `CD-STD`, `CD-Obtain` and `CD-Afford`. If respondents had ever heard of condoms, i.e. the variable `CDHeard` is not included in the variable `CDagree`.
For variables `HeardHIV` and the corresponding sub-variable `CompknoHIV` the same procedure as above was applied. The sub-variable had one level "Not comprehensive". This level got renamed "Not comprehensive/No knowledge" and all respondents who did not know about HIV were assigned to this new level.
Questions about sexual behavior had originally little less than 20% missing cases. This could be reduced dramatically to around 0.6% per variable because everybody who claimed that he or she never had had sex in his or her life at variable `AgeSexcat` had not responded to any of the questions about sexual behavior. Their missing values could be set to "No" at all four questions about sexual behavior, as a "Yes" on any of these would imply that they in fact had sex at least once in their life. Questions about sexual behavior include the variables `Sexgift`, `MultSex`, `Sex12m` and `NonmarSex1`. One of the variables, `NonmarSex1`, changed all the missing cases to a denial of non-marital sex and therefore this variable became complete.
For subjects who were 15 years old (`Respage`=15), claimed to not have had sex in the last year (`Sex12m`=No) and that they were 15 years or older when they first had sex (`AgeSexcat`=15 years and above), their age at first sex was set to the level "Below 15 years". This was valid for 12 cases. The column with the missing values in Table 2 has already incorporated the above changes. The portion of missing cases is calculated after logical imputation.

Variable	Description	Scale of Measurement	Detail of Categorical Variables	Missing Values	Method of Imputation
wealthq	Wealth Quintile	Ordinal	Poorest to Wealthiest	0.17%	polr
zone	Geopolitical Zone	Categorical	South West (reference), South South, South East, North West, North East, North Central		
location	Locality of residence	Dichotomous	Rural or Urban		
State	State/district of residence	Categorical	36 States and the Federal Capital Territory (FCT)		
Sexgift	Had sex in exchange for gift	Dichotomous	Yes or No	0.61%	logreg
MultSex	Had multiple sex partners	Dichotomous	Yes or No	0.62%	logreg
Sex12m	Had sexual intercourse in the last 12 months	Dichotomous	Yes or No	0.63%	logreg
NonmarSex1	Had sex with non-marital partner	Dichotomous	Yes or No		
CDHeard	Ever heard of condom	Dichotomous	Yes or No	0.32%	logreg
CD-AIDS	CD protects against AIDS	Categorical	Don't Know (reference), Agree, Disagree	0.33%	polyreg
CD-STD	CD protects against STDs	Categorical	Don't Know (reference), Agree, Disagree	0.33%	polyreg
CD-Obtain	CDs are easy to obtain	Categorical	Don't Know (reference), Agree, Disagree	0.35%	polyreg
CD-Afford	CDs are affordable	Categorical	Don't Know (reference), Not Affordable, Affordable	0.39%	polyreg
CDAgree	Number of Agree at CD variables	Continuous			
AgeSexcat	Age at first sex	Categorical	Below 15 years (reference), Can't Remember, 15 years and above, Never	1.31%	polyreg
educ-cat	Educational attainment	Categorical	No formal education (reference), Quranic, Primary, Secondary, Higher	0.16%	polyreg
ExpSTIs	Experienced STIs in the last 12 months	Dichotomous	Yes or No	0.58%	logreg
RespAge	Respondent's age	Continuous			
HIVTest-res	HIV test result	Dichotomous	Positive or Negative	22.48%	logreg
HeardHIV	Heard of HIV	Dichotomous	Yes or No	0.58%	logreg
CompknoHIV	Comprehensive knowledge of HIV	Dichotomous	Comprehensive knowledge (reference), Not comprehensive/No knowledge	0.64%	logreg
Yearstud	Year of study	Dichotomous	2012 or 2007		
Marital-cat	Marital status	Categorical	Currently Married (reference), Formerly Married, Never Married	1.03%	polyreg
Male	Gender	Dichotomous	Male (reference), Female		
Religion	Religion	Categorical	Others (reference), Christian, Muslim		

Table 2: Description of variables, STI: Sexual Transmittable Infections and STD: Sexual Transmittable Diseases, CD: Condom, polr: proportional odds regression, logreg: logistic regression, polyreg: polytomous regression

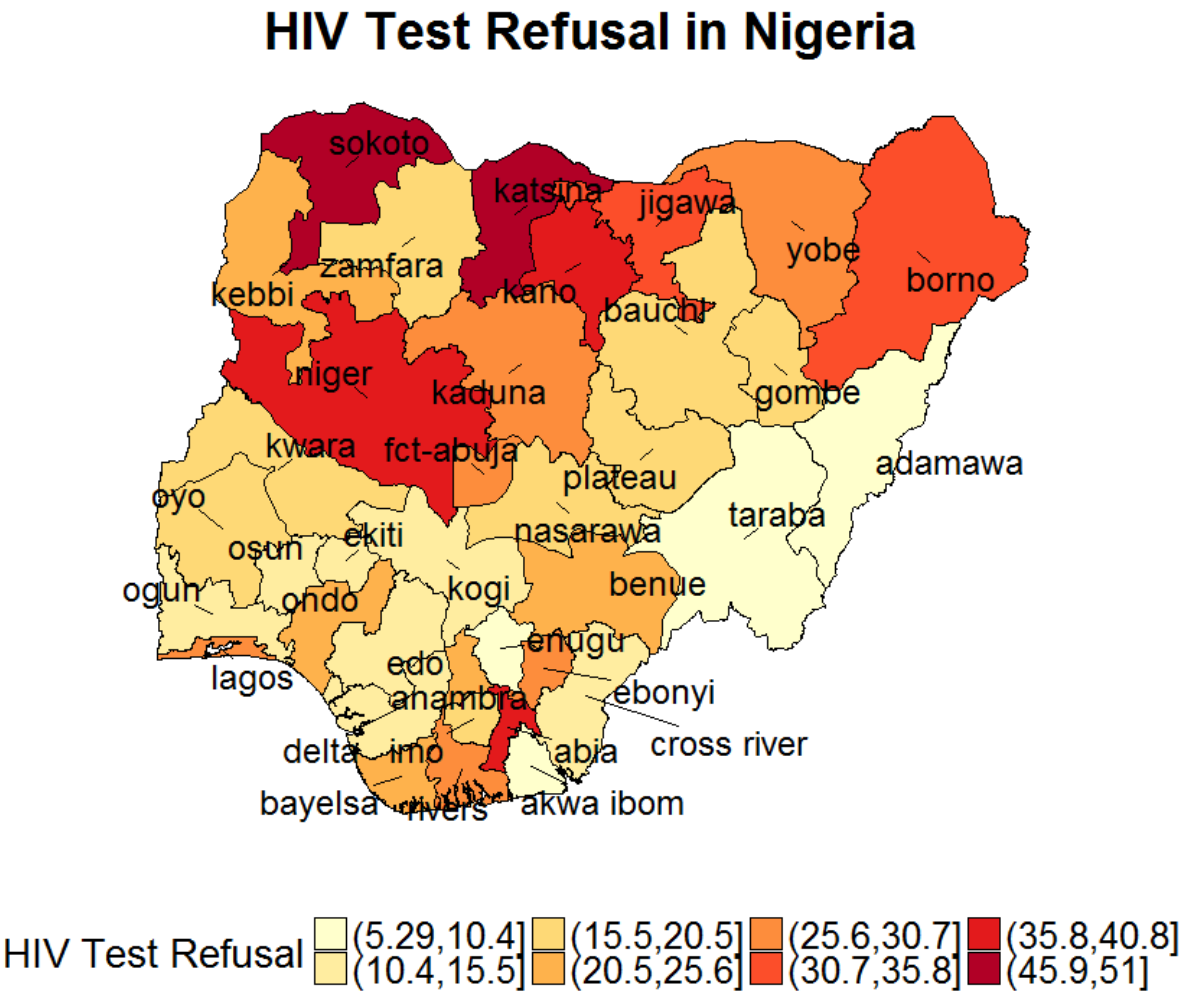

Figure 4: HIV test refusal in Nigeria by states in percent

6.3 Descriptive Analysis

According to the study design the number of females and males is almost equal, with the percentage of males at 50.89%. Two thirds of the respondents live in a rural environment. The number of respondents in each state ranges between 796 in Ondo to 1599 in Kano. In Figure 4 the percentage of respondents who refused the HIV test is plotted for each state. It shows that the proportion of respondents refusing the HIV test varies strongly between the different states of Nigeria. It can be seen, that the percentage of HIV test deniers is higher in many northern states and some southern ones. In Sokoto less than 50% of the respondents participated in the HIV test. More than one third of the respondents refused testing in the following states: Katsina, Niger, Abia and Kano. The lowest rates of denial can be found in mostly eastern and southern states. In Akwa Ibom, Adamawa and Enugu less than 10% of the respondents denied HIV testing. On the national level, 22.47% of respondents in Nigeria refused to take the HIV test. The corresponding plot of HIV test deniers on the zone level can be found in Appendix A.

Among the respondents who attended the HIV test, 3.4% were tested positive. In

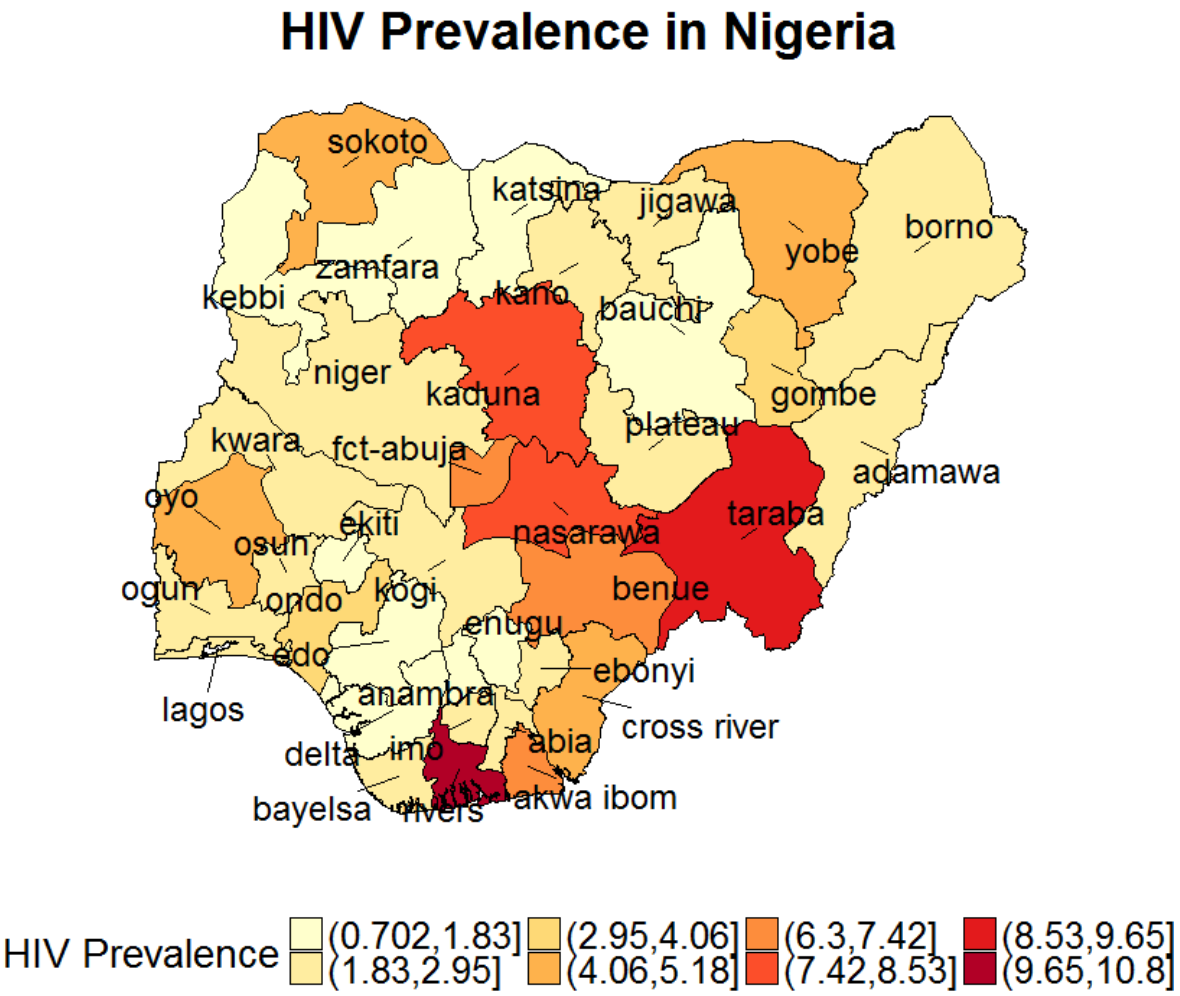

Figure 5: HIV prevalence rate in Nigeria by states in percent, no HIV test refusals were regarded

Figure 5 the HIV prevalence rate among respondents who participated in the HIV test is plotted on the state level. The highest prevalence rates can be found in Rivers, Taraba, Kaduna and Nasarawa and central and southern-eastern states, while many northern and southern states have lower prevalence rates, with the lowest rate in Zamfara.

In Figure 6 the HIV prevalence rate for the different zones of Nigeria is plotted. It can be seen, that differences in HIV prevalence rates are lower between the zones than between states. The prevalence rate in the north central region is more as double the prevalence rate in the south east region. In south east, south west and north west zones the prevalence rates are lower than the national prevalence rates. Further, a gap between the three regions with the lowest HIV prevalence rates and the three zones with the highest prevalence rates can be observed. Note, that only those who participated in the HIV test are regarded.

Figure 7 shows the distribution of the respondent's age for the possible HIV test results, positive and negative and for the deniers of that test. While the lines for

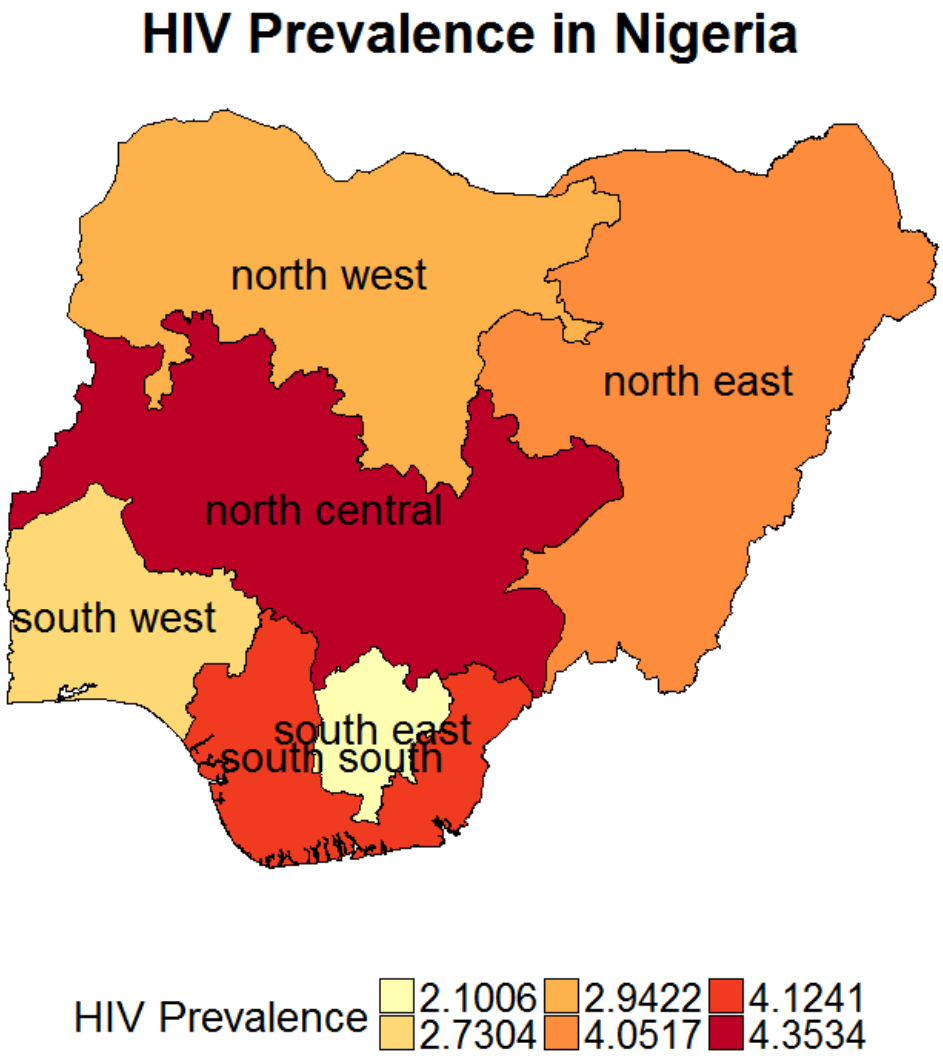

Figure 6: HIV prevalence rate in Nigeria by zones in percent, no HIV test refusals were regarded

negative and refusal are overlapping with the ticks in negative being larger, the line for positive has its maximum value at a higher age and is smoother.

Figure 8 shows a mosaic plot giving information about the HIV test result and if respondents had sex in exchange for gifts. On the x-axis the proportion of the variable `Sexgift` is drawn and on the y-axis the corresponding proportion of the HIV test result is drawn. Only a small portion of the respondents had sex in exchange for gifts, but their probability for having a positive test result is higher than for those who did not have sex in exchange for gifts. Respondents who had sex in exchange for gifts had a higher probability to participate in the HIV test. For the variables `MultSex` and `NonmarSex1` the mosaic plots look quite similar. The portion of respondents ticking "Yes" at one of these variables is higher than for variable `Sexgift`, but people who ticked "Yes" have higher probability for being tested positive and also a higher probability to participate in the test. The corresponding mosaic plots can be found in Appendix A.

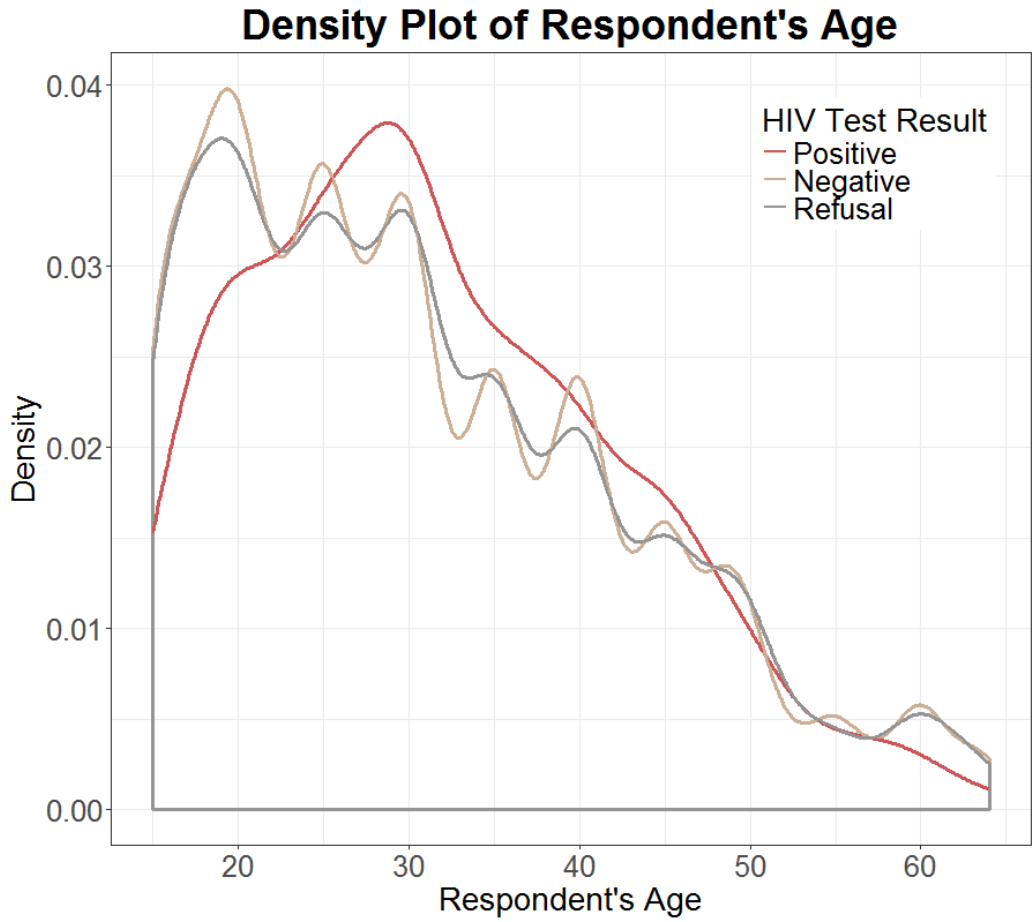

Figure 7: Density plot for HIV test result and respondents age

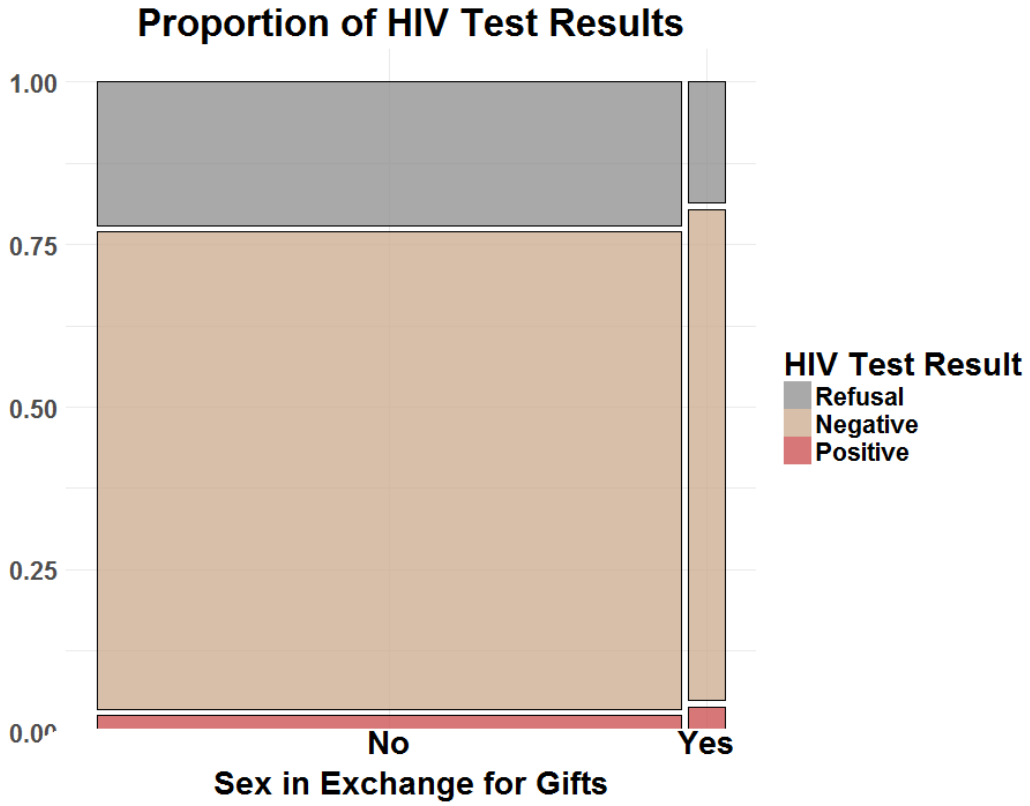

Figure 8: Mosaic plot for HIV test result and sex in exchange for gifts

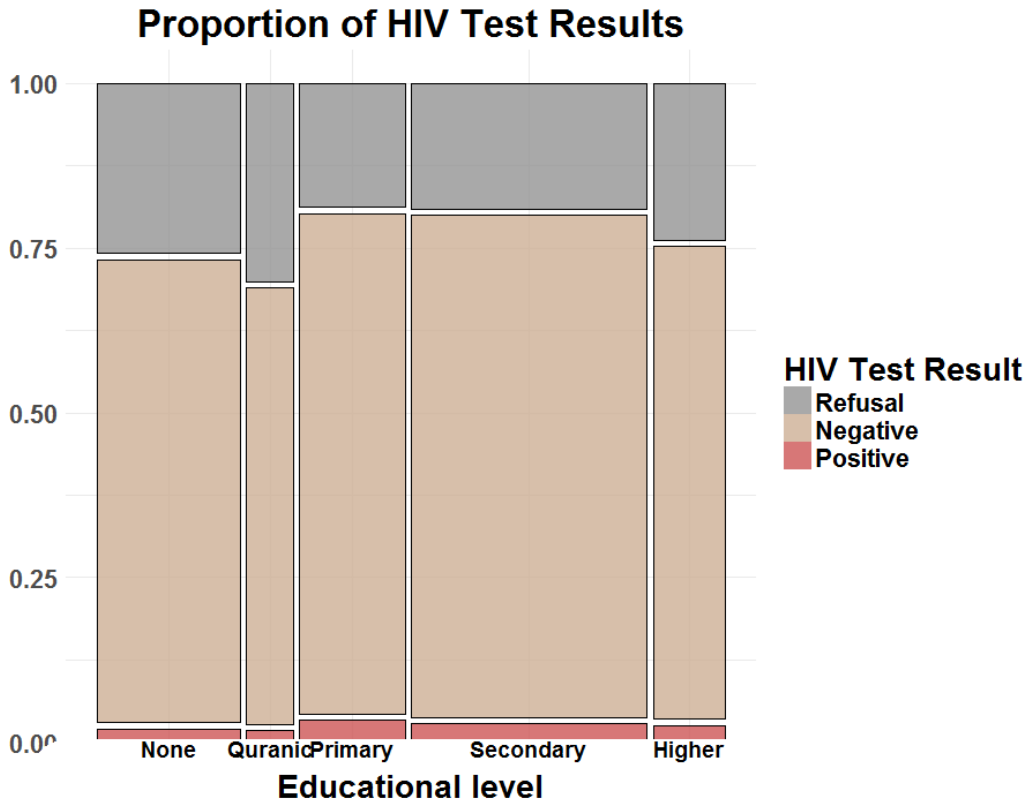

Figure 9: Mosaic plot for HIV test result and the highest educational level

Figure 9 is the mosaic plot for the highest educational level and the result of the HIV test. The educational levels are none, quranic, primary, secondary and higher with none being the short for no formal education from variable `educ-cat` in Table 2. The biggest group for the highest educational level is secondary, followed by none and primary. Quranic education as the highest level is the smallest group. Regarding the outcome of the HIV test, differences between the groups are rather small, but respondents in the primary group seem to have the highest rate of positive test results, while respondents from the quranic group have the highest proportion of test refusals.

In Figure 10 a stacked bar plot for the result of the HIV test result according to the groups in the variable `ExpSTIs` can be seen. The y-axis gives the proportion of each possible outcome of the HIV test result, refusal, negative and positive, in each of the levels for the variable `ExpSTIs` on the x-axis. There exist big differences between the levels of `ExpSTIs`. Respondents who experienced sexual transmittable infections (STIs) in the last year have a probability of 12.61% to be tested positive. Note that this value does not correspond to the one in Figure 10, as the 12.61% correspond to the portion of respondents with positive test result among all who attended the test in the group of people who experienced STIs. This is the HIV prevalence rate for this group, as respondents refusing HIV testing have to be seen as missing values. The portion drawn in Figure 10 for positive test result and expe-

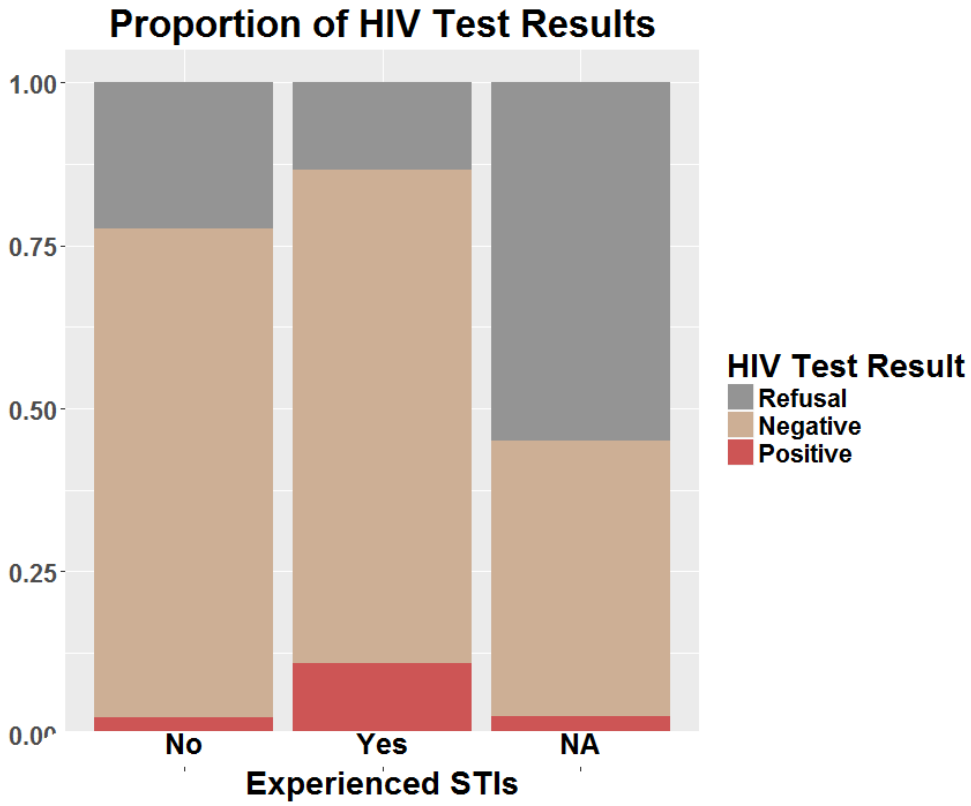

Figure 10: Bar plot for HIV test result and experienced sexual transmittable infections in the last 12 months

rienced STIs is 10.94%. This value corresponds to the percentage of all respondents who experienced STIs to be tested positive, whether or not they attended the HIV test. The value of 10.94% itself is not important, however the general expression that respondents who experienced STIs have a higher probability for being tested positive is valid. Further, one can read the percentage of people refusing the HIV test from Figure 10, which is 22.31% for those who did not experience STIs, 13.28% for those who did and 55.02% for those who did not answer to this question. This shows that those who experienced STIs have a higher probability to attend the HIV test and those who did not respond to the variable `ExpSTIs` have a high probability to not attend the HIV test. The variable `ExpSTIs` is extreme, as almost all data can be found in the level "No". The proportion of respondents who experienced sexual transmittable infections in the last year is as low as 0.3% and the proportion of missing values is as low as 0.58%.

Figure 11 is the mosaic plot for the marital status and result of the HIV test including the refusal of the test. More than half of the respondents are currently married, a big part of the rest was never married and few are formerly married. Respondents who were formerly married seem to have a higher probability for being tested positive than those who were never or are currently married. Overall, differences in the HIV test results are rather small between the levels in the marital status.

Figure 12 depicts the mosaic plot for religion and the result of the HIV test. Most

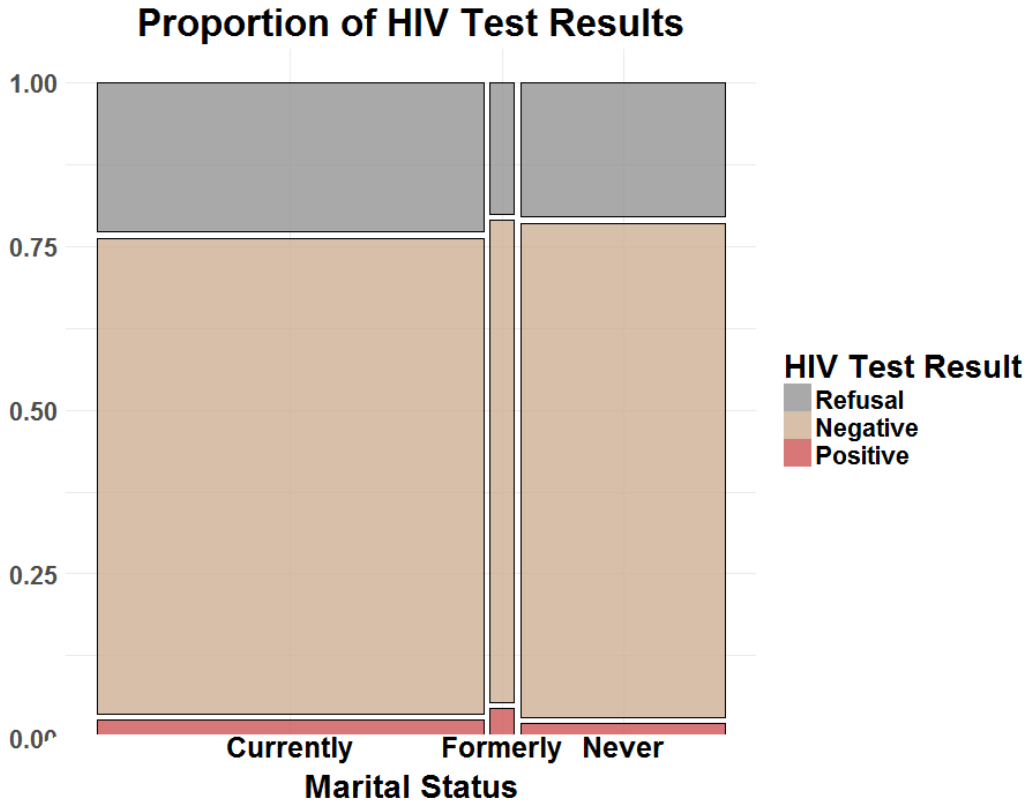

Figure 11: Mosaic plot for HIV test result and marital status of respondent

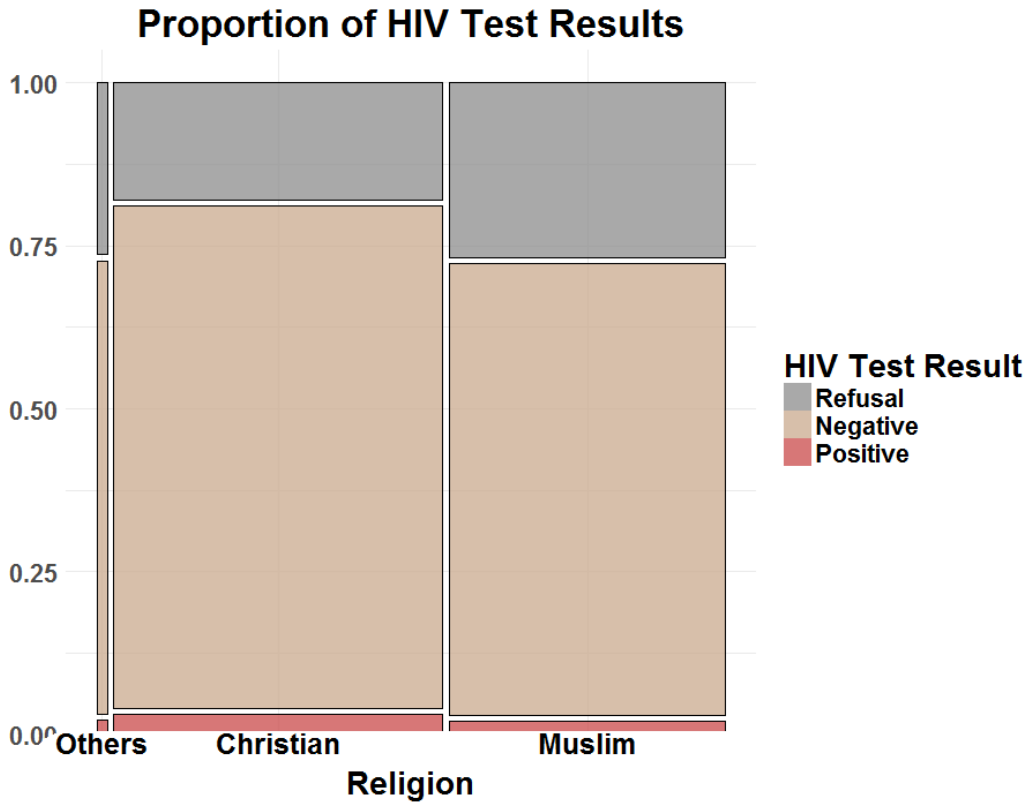

Figure 12: Mosaic plot for HIV test result and religion

respondents are either Christian or Muslim and only a few have another religion. It can be seen that Christians refused the HIV test with a lower probability than Muslims or those with other beliefs. Christians also seem to have a higher probability for being tested positive. Overall, the differences between the different religious groups with regard to the result of the HIV test including denial of the test, are rather small.
Mosaic plots like the ones above can be found in Appendix A for all variables. As in the plots shown above there are not big differences between the groups, for many plots the differences are smaller than for the figures shown above. Worth mentioning is that respondents who claim to never having sex in their life have only a slightly smaller probability for being tested positive on HIV.

6.4 Missing data

After the data preparation steps of chapter 6.2, 24.33% of observations had missing data at at least one variable. Regarding only the amount of missing data in all variables except the one of the HIV test leaves 2.82% of cases with missing data. As can be seen in Table 2 in chapter 6.2, missing values occurred in the following variables: `educ-cat`, `wealthq`, `CDHeard`, `CD-AIDS`, `CD-STD`, `CD-Obtain`, `CD-Afford`, `ExpSTIs`, `HeardHIV`, `Sexgift`, `MultSex`, `Sex12m`, `CompknoHIV`, `Marital-cat`, `AgeSexcat` and `HIVTest-res`. The highest missing data rate exists in variable `HIVTest-res`, the variable of the result of the HIV test, at 22.48%. Besides `HIVTest-res`, the variables `AgeSexcat` and `Marital-cat` are the only ones with more than 1% missing values. All other variables with missing values have less than 1% missing cases.
For the respondents who have missing values at one or more items, except the HIV test result, the probability for being tested positive is at 3.28%. This is a little lower than the 3.44% among the respondents without missing values. The probability to refuse the HIV test for respondents with missing values is 34.38%. This is higher than the probability for respondents without missing values, whose probability is 22.13%.

Figure 13 gives a graphical overview over the existing missing data patterns. On the x-axis the variables which have missing values are drawn and on the y-axis respondents who have at least one missing value are drawn. This means that only a subset of the data was used. More precisely, 24.33% of the observations are used and those without missing values are excluded for this plot. Black color indicates missing values and gray color indicates observed values. It can be seen that variable `HIVTest-res` has the biggest portion of missing values. If a respondent has missing values at one variable that is not the HIV test result, there are often missing values at other variables as well. In total, there exist 94 missing data patterns, which can be studied in detail in Table 9 in appendix B.

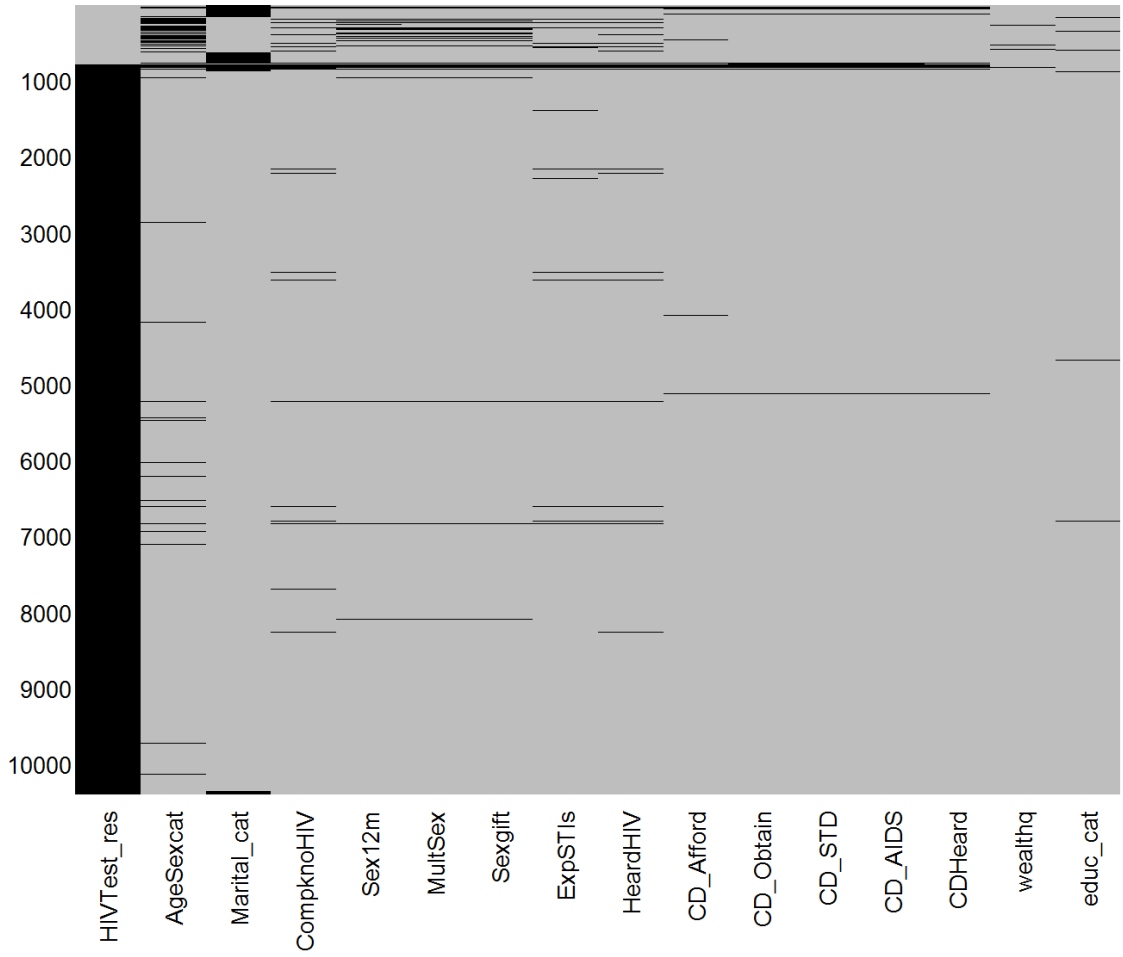

Figure 13: Missingness Map, where black color indicates missing values, grey color indicates observed values, only cases and variables with missing values were used for this plot

7 Results

R was used for every particular step and result in this and the previous chapter.

7.1 Test on MCAR

Whether or not the data is MAR cannot be ascertained as there exist no tests to distinguish between MAR and MNAR. However, it is possible to test on MCAR, thus Little's MCAR Test was applied to the data and showed a significant test result. This means that there is at least one variable that is not MCAR which was expectable considering the variables.

Next it was tested for dependencies between certain variables. Therefore, chi-squared tests were applied as all of the variables with missing values were categorical. For every variable with missing values the chi-squared test was significant with at least one other variable, meaning that there exists some dependency between the variables. This is further proof that the underlying missing data mechanism for all variables is either MAR or MNAR. There is not enough information to specify

a MNAR mechanism and therefore, the data is assumed to be missing at random (MAR).

7.2 Imputation

Multiple imputation as described in chapter 3 was applied to impute the missing values. In the imputation phase, the fully conditional specification (FCS) algorithm was applied through its implementation in the mice package in R. The assumption of MAR is assumed to hold, although it has to be remembered that there is no proof for this. The sub-variables of the variable about the knowledge of existence of condoms were excluded for the multiple imputation process. That is, the variables `CD-AIDS`, `CD-STD`, `CD-Obtain` and `CD-Afford` were excluded from the imputation process and variable `CDagree` was used instead. All other variables, from Table 2, except the ones mentioned before were used in the imputation phase. Zones are a grouping of states in Nigeria which means that each zone would be regarded as a combination of states. When applying to a (imputation) model, some coefficients would not be available. The result would be NAs (Not Available), which should be avoided. Due to this reason only one of both can be used in imputation. As every variable included in any model in the analysis phase should also be included in the imputation phase, multiple imputation was applied twice. One time the zones are excluded from imputation and the other time states are excluded.
The variable `CompknoHIV` was excluded in the imputation of the variable `HeardHIV`, because `HeardHIV` serves as a screening variable for `CompknoHIV`. Excluding it should prevent to simply impute the same values as in the previous iteration. It is possible to post-process imputations in each iteration if they take on an implausible value. This is achieved by checking directly after the imputation in each iteration if they match the criterion and correct them if not. After this check the imputation model of the next variable starts. For the variable about comprehensive knowledge about HIV this means that it is checked if someone who has never heard of HIV got imputed into the group "Not comprehensive/No knowledge". If not and this observation got imputed to be in the "Comprehensive knowledge" category, then this observation got post-processed to the "Not comprehensive/No knowledge" category. For variables `Sexgift`, `MultSex` and `Sex12m`, post-process checked if an observation that never had sex before was set to "Yes" at any of these variables. If that is the case, this observation was set to "No" at the corresponding variable about sexual behavior. The same applies vice versa for the age at the first sex-variable. If an observation was imputed to "Never", but in any of the four questions about sexual behavior ticked at least once "Yes", then it was set to "Can't Remember" at variable `AgeSexcat`.
The order of the imputations is monotone from the lowest amount of missing values

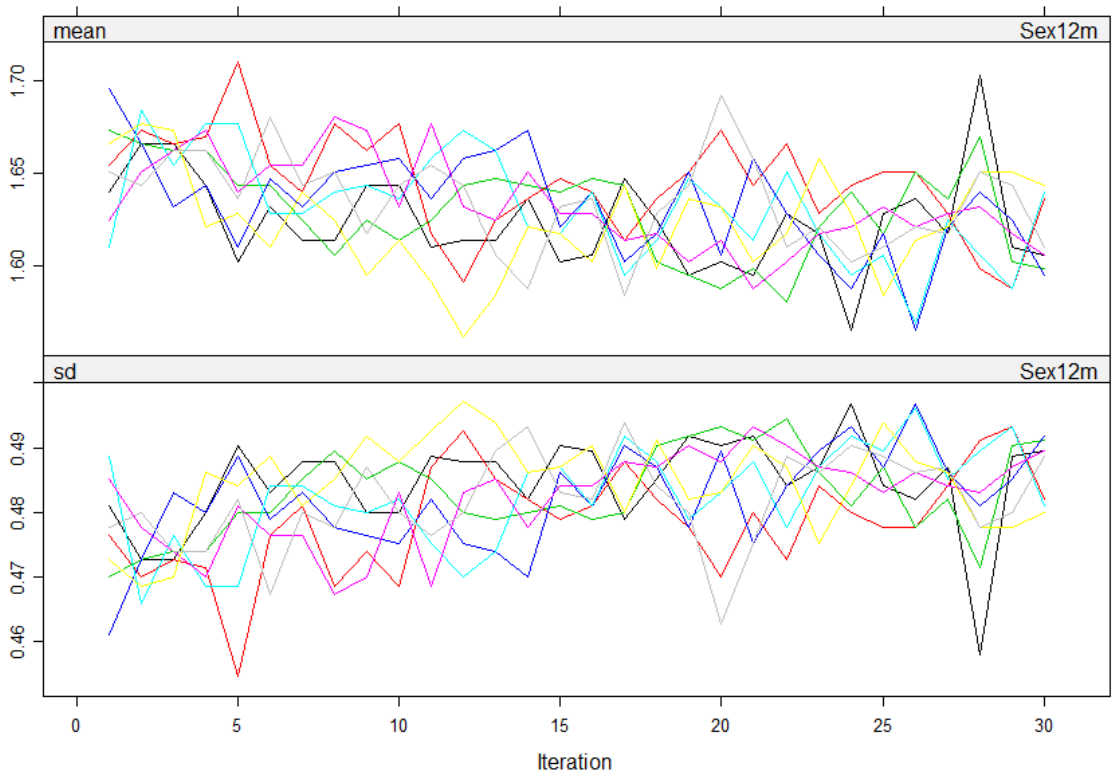

Figure 14: Trace line plot for imputations of variable `Sex12m` with covariable `State` each line represents one of the M imputations of this variable

to the biggest. It was made sure that the variable `HeardHIV` got imputed before the variable `CompknoHIV`.

An advantage of the FCS-algorithm is that it can incorporate individual models for each variable. For variables with a dichotomous scale of measurement, logistic regression was the imputation method used. For categorical variables polytomous regression was utilized and for ordinal variables proportional odds regression. Continuous variables did not have missing values. See Table 2 in chapter 6.2 for the imputation method of each variable.

The number of imputations is determined by the number of cores of the computer as it is possible to parallel the imputation sets. The computer used has eight cores which led to $M = 8$ imputation sets. 30 iterations were used which proved to be sufficient.

The reasoning is demonstrated exemplary for variable `Sex12m`. Figure 14 portrays a trace line plot of variable `Sex12m`. This is the imputation process, were `State` was a covariable. In a trace line plot for multiple imputation, the imputed values of a variable are plotted against the iteration number. Each line represents one of the M replications. Plotted is the mean and the standard deviation of the variable `Sex12m`. To calculate the mean and standard deviation, the variable is assumed continuous. The levels of `Sex12m` are represented by a one for "No" and a two for

"Yes". On convergence, the streams should be free of any trend and intermingle. Here, the streams intermingle pretty soon, but until iteration 20 a negative trend for the mean and a positive trend for the standard deviation can be observed. For 30 iterations, convergence is achieved. The trace line plots for all variables with imputed values for both imputation runs can be found in appendix C.
The result of the imputation phase were eight complete data sets. Once for the imputation with `State` and once for imputations using variable `zone`. The outcome variable, `HIVTest-res`, was imputed along with every other variable with missing data during the imputation phase of multiple imputation. For the following models the imputations on the outcome variable were removed.

7.3 Models

As mentioned above, the imputations of variable `HIVTest-res` were removed in all eight imputed data sets. The following models were applied to the eight imputed data sets.
All observations with missing values in `HIVTest-res` were excluded from the data set and saved for later prediction. The data set for model training and testing steps consisted of 33146 observations.
Models were trained using repeated 10-fold cross validation (cv). The number of repetitions was between 15 and 50, depending on the computational effort of the model. Increasing the number of repetitions did not seem to improve the predictive power of the models. Some models could be tuned in cross validation, for others cross validation was only used to have an idea of the predictive power. At the end of the cross validation process, the final model was calculated using all available data. An algorithm was constructed to be able to execute these steps in parallel. The final model was used to predict the result of the HIV test. Predictions were given in the form of probabilities, not classes. This ensures that the rounding threshold can be set manually to optimize the predictions. For each of the eight models, the optimal threshold was determined. Then each model predicted the probability for the missing cases of variable `HIVTest-res`. Next, their eight results were averaged, as well as the eight thresholds. If the probability for a positive test result was higher as the threshold, then it was set to "Positive". If it was lower, it was set to "Negative".
The outcome variable for all following models was `HIVTest-res`, the HIV test result with levels "Positive" and "Negative". The task was to get a high precision. This is the proportion of true positives among all positive predicted cases.
If SMOTE was used, the minority class was over-sampled by a factor of five and the majority class was down-sampled such that the ratio between both was almost 50:50. Under-sampling and over-sampling also reached this ratio.

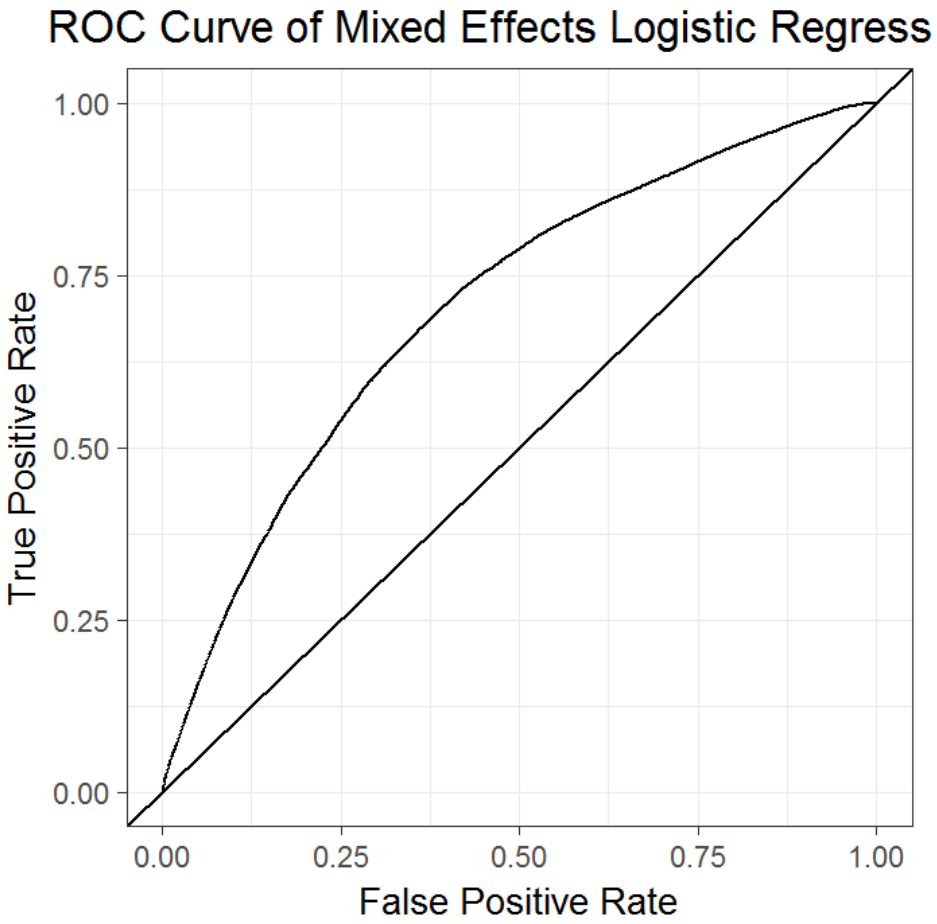

Figure 15: ROC curve of logistic regression with mixed effects model

The variable `State` has a total of 37 levels which lead to the idea on testing mixed effects models with `State` serving as the random variable. Since variable `State` had to be included in the predictors, this model was not applied with the variable `zone`. The model was calculated using the glmer-function from package lme4 in R. To eliminate errors and warnings in the code, the continuous variables got rescaled and the optimizer was chosen to be "bobyqa" with the maximum number of function evaluation set to 10000 to prevent convergence failures. 25 replications on 10-fold cross validation were used resulting in 825168 observations on which the predictive power of the model was tested.

The summary of the pooled model can be found in appendix D. The summary contains information about the parameter estimates, their standard error, degrees of freedom, t-test and the corresponding p value, the lower and upper bound of a confidence interval and the fraction of missing information.

Figure 15 plots the corresponding roc curve. In this case only one roc curve is plotted, because the roc curves for all $M = 8$ are indistinguishable. Table 3 gives the area under the curve (auc) of all eight roc curves. The *auc* ranges between 0.7027 and 0.7034, which is quite similar. It is not exactly similar, due to the different imputations in each data set, but the differences are rather small, as the portion

ROC Curve of Imputation Set								
	1	2	3	4	5	6	7	8
auc	0.7032	0.7034	0.7030	0.7034	0.7027	0.7032	0.7030	0.7029

Table 3: Area under the curve of all eight roc curves for the mixed effects logistic regression model

of missing data was as well quite small. An *auc* of 0.703 is not very good, which can also be seen when regarding the possible thresholds for rounding. The maximum probability for a positive test result among all test cases was 0.4132 and was in reality a respondent who had a negative test result. Table 4 depicts the real test result for the cases that had highest predicted probability to be tested positive. The two respondents with highest probability are in fact HIV negative and would be classified wrong. For a rounding threshold of 0.41, four respondents would be classified positive, two correctly and two falsely. The precision would be maximized at this point at a value of 0.5. Reducing the threshold further would decrease the precision, as can be seen in Table 4 for ten cases. If the threshold would be set such that 100 cases are predicted positive, precision would be a lot lower at 0.18. For a threshold of 0.41 the sensitivity or true positive rate would be at 0.00007. The F_1 score for this case is 0.00014. The measures prove the point that this result is far from good. As the precision is maximized at a threshold of 0.41, this will be the threshold chosen for the prediction of the respondents who refused HIV testing.

Among those who refused to take the HIV test, the maximum probability to be

	Real Class	Positive
1	Negative	0.4180
2	Negative	0.4172
3	Positive	0.4151
4	Positive	0.4104
5	Negative	0.4079
6	Negative	0.4053
7	Negative	0.3973
8	Positive	0.3958
9	Negative	0.3941
10	Positive	0.3913

Table 4: The ten cases in testing with highest probability for positive HIV test and their real result

HIV positive was as low as 0.1313 with a mean of 0.0175. At a threshold of 0.41, everybody got classified negative. The above results are represented at one of the eight imputations. The results of the other imputation sets were pretty similar. The maximum probability in testing was slightly different, between 0.4132 and 0.4254. Seven of the eight chose the same cut point resulting in a precision of 0.5. One maximized the precision at a cut point were five respondents were classified positive,

		ROC Curve of Imputation Set							
	Sampling	1	2	3	4	5	6	7	8
State		0.7029	0.7032	0.7028	0.7032	0.7024	0.703	0.7027	0.7026
	down	0.6957	0.696	0.6955	0.696	0.6952	0.6957	0.6955	0.6954
	SMOTE	0.6964	0.6967	0.6963	0.6967	0.696	0.6965	0.6962	0.6961
	up	0.7032	0.7035	0.7030	0.7035	0.7027	0.7032	0.7030	0.703
Zone	up	0.6226	0.6224	0.6217	0.6217	0.623	0.622	0.6218	0.6222
		0.6239	0.6236	0.6229	0.6228	0.6242	0.6231	0.623	0.6234

Table 5: Area under the curve of all eight roc curves for the logistic regression model

	Actual Positive	Actual Negative
Predicted Positive	39	226
Predicted Negative	22741	639914

Table 6: Confusion Matrix for up-sampled logistic regression model with rounding threshold 0.881

two of them correctly.

One of the models to be tested was logistic regression. The *auc* values of the different sampling approaches to logistic regression can be seen in Table 5. If no sampling method is given, then no sampling method was applied. The table shows the *auc* for all eight models that derive from the use of multiple imputation. It can be seen that the usage of the data set that included the variable `State` results in higher *auc* values than those with variable `zone`. Figure 16 depicts the corresponding roc curves. Only two are plotted. This is due to the fact that roc curves between the eight models are indistinguishable and in this case also roc curves between different sampling approaches but with the same co-variable are indistinguishable.

The problem is that the predictions are quite inaccurate. For all models, it was maximal possible to get some true positive predictions for the first 4 to 15 highest predicted probabilities. If a threshold was chosen that high to achieve a high precision according to testing in cross validation then the threshold was too high for any predicted probability for the unknown test results. The result was that all are classified "Negative".

The Confusion Matrix in Table 6 depicts this. To construct it, the threshold was chosen to be the highest predicted value for the unknown test results. The given probability was 0.881. The underlying model showed one of the best results as it depicted 12 of the highest 15 probabilities in cv-testing correctly. The corresponding table can be found in appendixD along with the pooled model coefficients. The model used the variable `zone` and up-sampling. The threshold derived was 0.94 and thus higher than the highest probability among the respondents with unknown test result. Choosing the threshold at 0.881 resulted in 39 true positive classifications and 226 false positive classifications. This yields a precision of 0.1472 which is low.

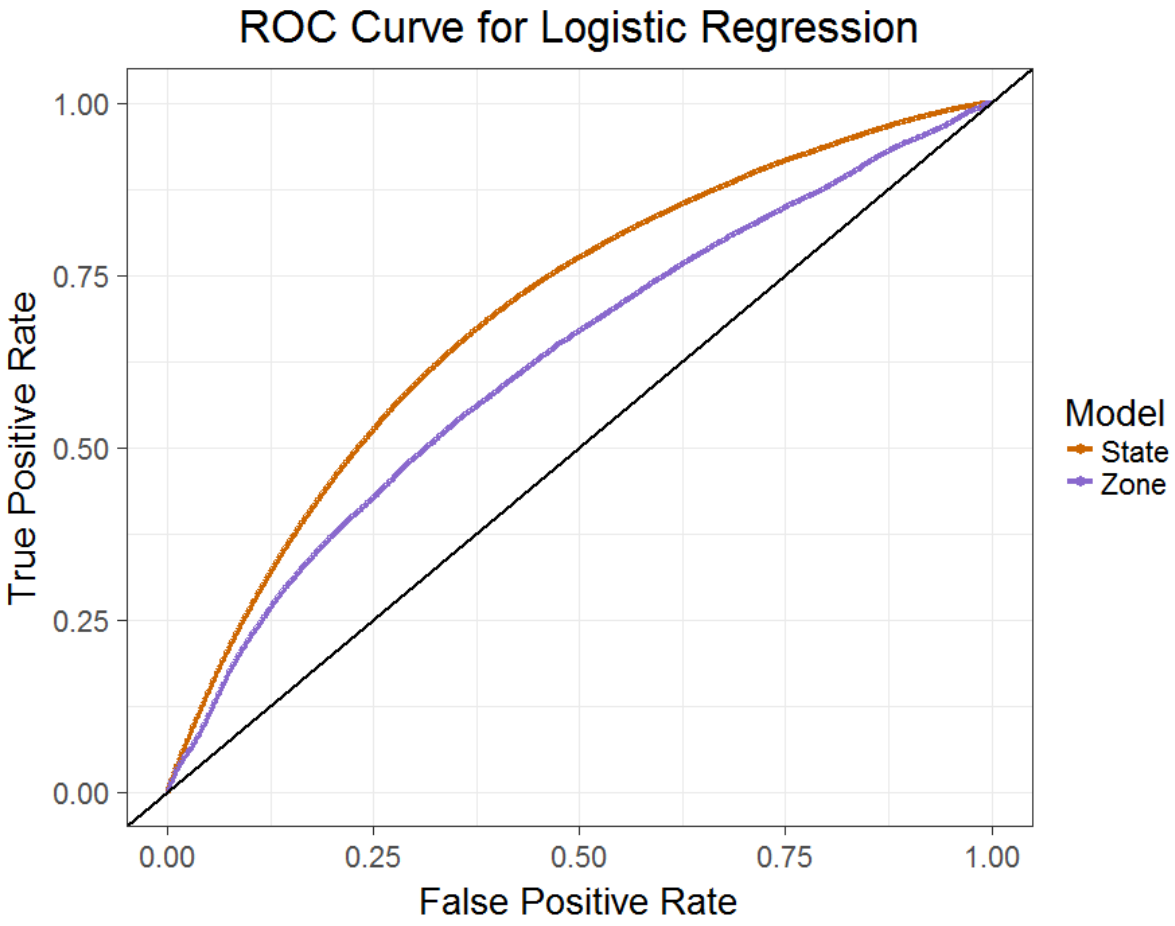

Figure 16: ROC curve of logistic regression

Figure 17 shows the densities of the two classes of variable `HIVTest-res` for above model. It can be seen that both densities overlap and are not as distinguishable as desired.

Fore boosted decision trees the algorithm "C5.0" was applied. Predictions with this algorithm were very inaccurate. For SMOTE sampling and in the case without any sampling many cases were predicted to a probability of one, making it impossible to distinguish between them. For the models including the variable `zone`, all models were totally inaccurate. The corresponding table with the *auc* values and the figure with the roc curves can be found in appendix D.

For random forests, the number of variables that get randomly selected at each node were tuned with cross validation. The number of trees was set to 2000. The measure for the best model was the *auc*. For seven out of eight models on the imputed data sets, the best *auc* was achieved with 12 selected variables at each node. In one model 13 variables at each node showed best *auc*. However, differences were rather small, as all *auc* values were between 0.65 and 0.67. The roc curve in Figure 18 resembles the *auc* values from Table 7. The curve is rather close to the angle bisector. Predictions from testing in cross validation are quite underwhelming, as the cases who got assigned the highest probability to be positive, are in fact negative. To predict

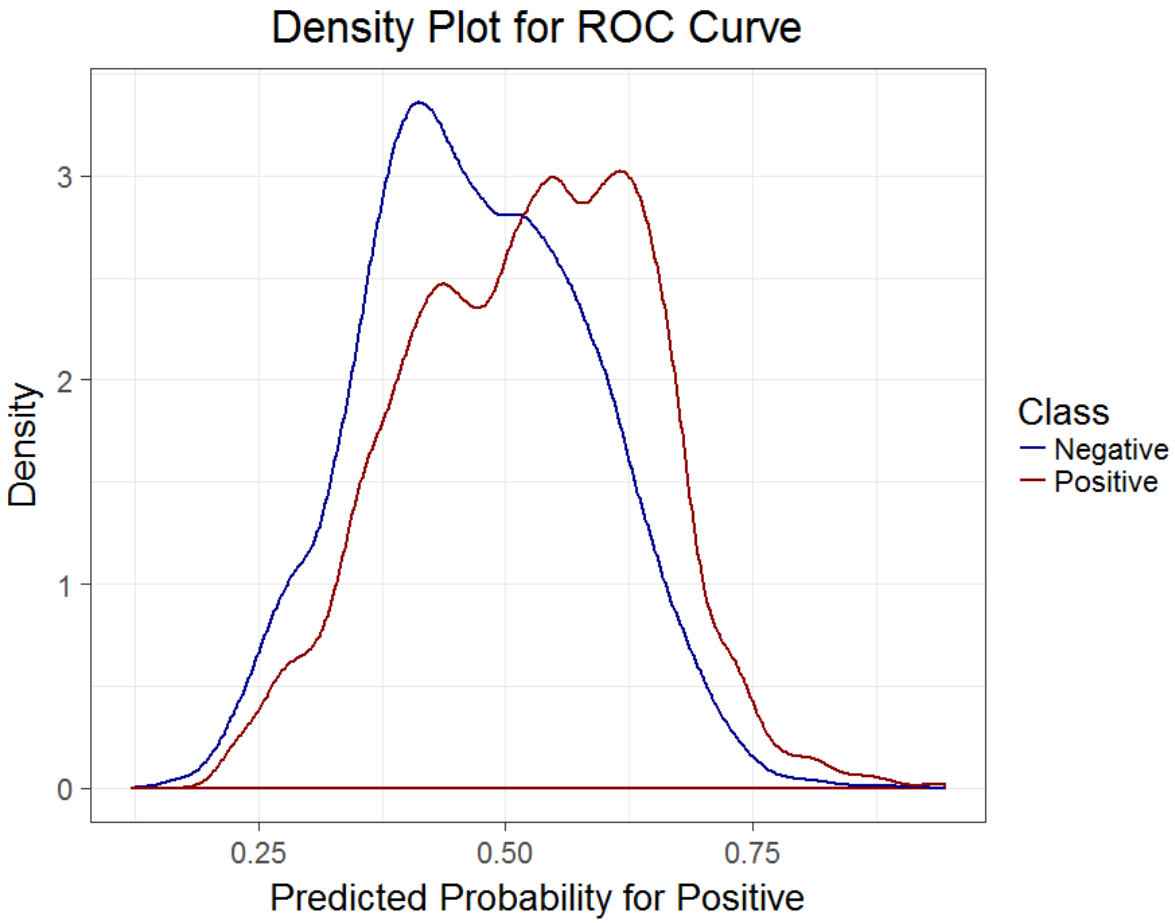

Figure 17: Density plot for logistic regression with up-sampling and co-variable `zone`

at least one true positive, on average 62 have already been predicted false positive. This results in a low precision and makes predictions on unknown data rather untrustworthy. The first to be predicted true positive had an probability of 0.725 with very little variation between the eight models. As precision was the measure to be maximized, one could argue that infinity would be the best threshold as the real maximum in precision holds many misclassifications. For example, if the threshold would be chosen to be at 0.6, over 17% of the positive classified cases would be true positives. This result is far from the desired case to correctly predict HIV positive cases with a small to no false positive rate. For the predictions of the HIV test result for respondents who refused testing, the maximum probability was at 0.68 and the

		ROC Curve of Imputation Set							
	Sampling	1	2	3	4	5	6	7	8
		0.664	0.6644	0.664	0.6644	0.6633	0.6647	0.6639	0.6631
`State`	SMOTE	0.6618	0.6628	0.6617	0.6619	0.6622	0.6626	0.6613	0.6611
	down	0.6882	0.6884	0.6882	0.6886	0.6881	0.6879	0.6879	0.6877
`zone`	down	0.615	0.6153	0.6153	0.6149	0.6162	0.6157	0.6144	0.6156

Table 7: *auc* values for all eight random forest models

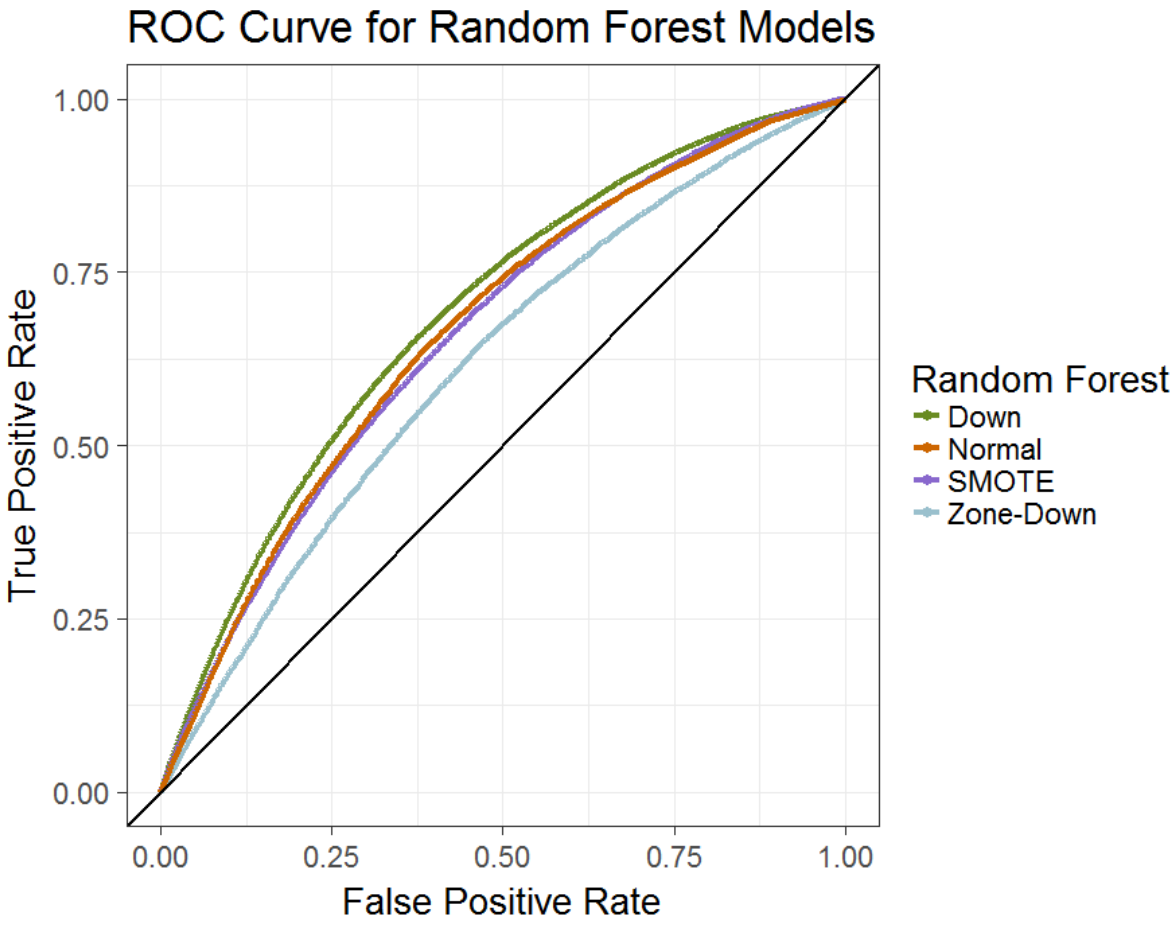

Figure 18: ROC curves for random forests

mean was at 0.04.

Random forests were also executed using the sampling method SMOTE. The number of trees was the same as above, 2000. In seven out of eight models, five was found to be the best number of variables that were chosen at random as candidates for splitting the node. In one case two variables were found to be best. As above, differences in the *auc* between different numbers of candidates for splitting were small. Although the *auc* values were quite close to the ones above, prediction was improved. In average, it took 19 false positives to classify the first true positive. In one of the eight models, it was possible to predict one true positive before predicting false positives. However, this would hold a threshold of 0.961. This threshold would yield only false positive classifications in all other model runs. Predicting the probability of a positive test result for respondents with missing values at variable `HIVTest-res` resulted in a maximum estimate of 0.9545 which would be lower than above threshold and thus classify all as "Negative".

Applying under-sampling resulted six times in choosing two variables and two times in choosing nine variables as candidates for splitting. The number of trees was constant at 2500 trees. Down-sampling yielded the highest *auc* values, as can be seen in Table 7. In average it took 19 false positives to classify one true positive. This

would result in a low precision which would would make predictions untrustworthy. So far all random forests were made using the variable `State` as a possible split variable and `zone` being excluded. Now random forests are applied with the possible split variable `zone` instead of `State`. The data was down-sampled before model tuning. Three variables were considered as candidates at each split in all eight models and the number of trees was set to 2500. In average it took 14 false positives to classify one true positive. This approach has the lowest *auc*, but is in comparison to the other random forest approaches the most promising in terms of predictive power. The mean of the predicted probabilities on HIV test deniers was 0.43 which is rather high. At the same time the maximum was 0.9 and no sensible threshold was found as cross validation showed that respondents with high probability have mostly a negative HIV test result. In all approaches to random forest models, the *auc* was rather small and the predictive power was poor.

The results of the naive Bayes classifier matched the ones from the other models. The roc curve, the *auc* values and the density plots to naive Bayes models with down-sampling and SMOTE can be found in appendix D. Table 8 shows the probability, observation and ID of the cases that got highest probability in cross validation testing. Assume that all of them would be classified "Positive". Out of the ten cases with highest probability, four of the five true positives come from the same observation in different resamples. Three of the six false positives are also from only one observation. The problem is that these cases may be outliers which have very high probabilities and could possibly pull the threshold up, making it harder that the threshold is reached by the unknown cases.

Generally said, predicted probabilities were higher in combination with sampling

	obs	Positive	ID
1	Positive	0.9964	33111
2	Positive	0.9960	33111
3	Negative	0.9960	31361
4	Positive	0.9957	32876
5	Negative	0.9955	31361
6	Negative	0.9951	4472
7	Positive	0.9951	33111
8	Negative	0.9947	20503
9	Positive	0.9946	33111
10	Negative	0.9946	31361

Table 8: The ten cases in testing with highest probability for positive HIV test and their real result and case number for naive Bayes with down-sampling

approaches. However, this did not improve anything as all probabilities were increased. Results between the models for the eight imputed data sets were rather small. ROC curves were indistinguishable and as a result *auc* values were almost

identical. The predicted values for variables were also quite close. The respondents yielding the high probabilities to be HIV positive in each prediction were often the same.
Besides the above mentioned models, the following models were tested: Logistic regression with step AIC, ridge regression, lasso and elastic net, gradient boosting models and elastic net. All of these models did not perform superior to the ones mentioned above and are quite CPU-intensive.
Not all combinations between sampling methods and models were tried as especially oversampling was very CPU-intensive. As literature implies that under-sampling the majority class leads to better classifiers than oversampling the majority class, focus was more on under-sampling. The combination of cross validation and model tuning is also very time-consuming.

8 Conclusion

In this work, first missing data and missing data handling methods, particularly multiple imputation was described. Multiple imputation consists of three phases: the imputation phase, the analysis phase and the pooling phase. Subsequently models that can be used in the analysis phase were described along with techniques to improve models and measures to validate them. Finally these concepts were applied to HIV data with the goal to predict the result of a HIV test for respondents who refused to participate.
Unfortunately, this was not possible as the predictive power of all tested models was far from optimal. In all models there were many false positives at the highest probabilities and in many the highest probability for a positive result was among observations that had in fact a negative result. In cases were the highest probability for a positive result was by an observation with an observed positive result, the rounding threshold was too high for predictions on data with unknown result.
Improved predictive power may be achieved with more parameters. The given variables lacked indications to risk groups. HIV statistics indicate that homosexuality among men and/or drug abuse increase the probability for having HIV. However, these variables were not included in the data set.
Another possible problem is weather or not the data can be trusted. The given data set contains variables to sensitive information like sexual behavior. In general, in topics of sensible data, the chance of incorrect data is increased. This issue is valid for HIV risk-groups like homosexuals and drug addicts, especially since both are not legal and criminalized in Nigeria.

List of Tables

List of Figures

Bibliography

[1] How to get to zero: Faster. smarter. better., 2011. UNAIDS World AIDS Day Report.

[2] Country factsheet nigeria, 2016. UNAIDS data.

[3] Samson B. Adebayo, Ludwig Fahrmeir, Christian Seiler, and Christian Heumann. Geoadditive latent variable modeling of count data on multiple sexual partnering in nigeria. *BIOMETRICS*, 67:620–628, June 2011.

[4] J. Barnard and Donald B. Rubin. Small-sample degrees of freedom with multiple imputation. *Biometrika*, 86:948–955, 1999.

[5] Douglas Bates. Linear mixed model implementation in lme4. manuscript, 2011. University of Wisconsin.

[6] Jaap P.L. Brand. *Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets.* PhD thesis, Erasmus University Rotterdam, 1999.

[7] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 2002.

[8] Craig K. Enders. *Applied Missing Data Analysis.* The Guilford Press, 2010.

[9] Tom Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27:861–874, 2006.

[10] Seymour Geisser. The predictive sample reuse method with application. *Journal of the American Statistical Association*, 70:320–328, 06 1975.

[11] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning.* Springer, 2 edition, 2016.

[12] J.T. Heckman. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *The Annals of Economic and Social Measurement*, 5:475–492, 1976.

[13] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning.* Springer, 2013.

[14] Helmut Küchenhoff. Lineare modelle-das logistische regressionsmodell. manuscript, 2014. Ludwig-Maximilians-Universität München.

[15] R.J.A. Little. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88:125–134, 1993.

[16] Tim P Morris, Ian R White, and Patrick Royston. Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Medical Research Methodology*, 14(75), 2014.

[17] Federal Ministry of Health [Nigeria]. National hiv & aids and reproductive health survey, 2012 (narhs plus), 2013. Federal Ministry of Health Abuja, Nigeria.

[18] Trivellore E. Raghunathan, James M. Lepkowski, John Van Hoewyk, and Peter Solenberger. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 27(1):85–95, June 2001.

[19] Donald B. Rubin. Basic ideas of multiple imputation for nonresponse. *Survey Methodology*, 12(1):37–47, June 1986.

[20] Donald B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. 2. John Wiley & Sons, 1987.

[21] Fei Tang. *Random Forest Missing Data Approaches*. Open access dissertations, University of Miami, 2017.

[22] Stef van Buuren. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, 16:219–242, 2007.

[23] Stef van Buuren, J.P.L. Brand, C.G.M. Groothuis-Oudshoorn, and Donald B. Rubin. Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76(12):1049–1064, December 2006.

[24] Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67, December 2011.

[25] Sudhir Varma and Richard Simon. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7(91), 2006.

[26] Ian R. White, Rhian Daniel, and Patrick Royston. Avoiding bias due to perfect prediction in multiple imputation of incomplete categorical variables. *Computational Statistics and Data Analysis*, 54(10):2267–2275, October 2010.

[27] Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, and Dan Steinberg. Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1):1–37, 2007.

[28] Harry Zhang. The otimality of naive bayes, 2004. FLAIRS conference.

A Appendix to Chapter 6.3

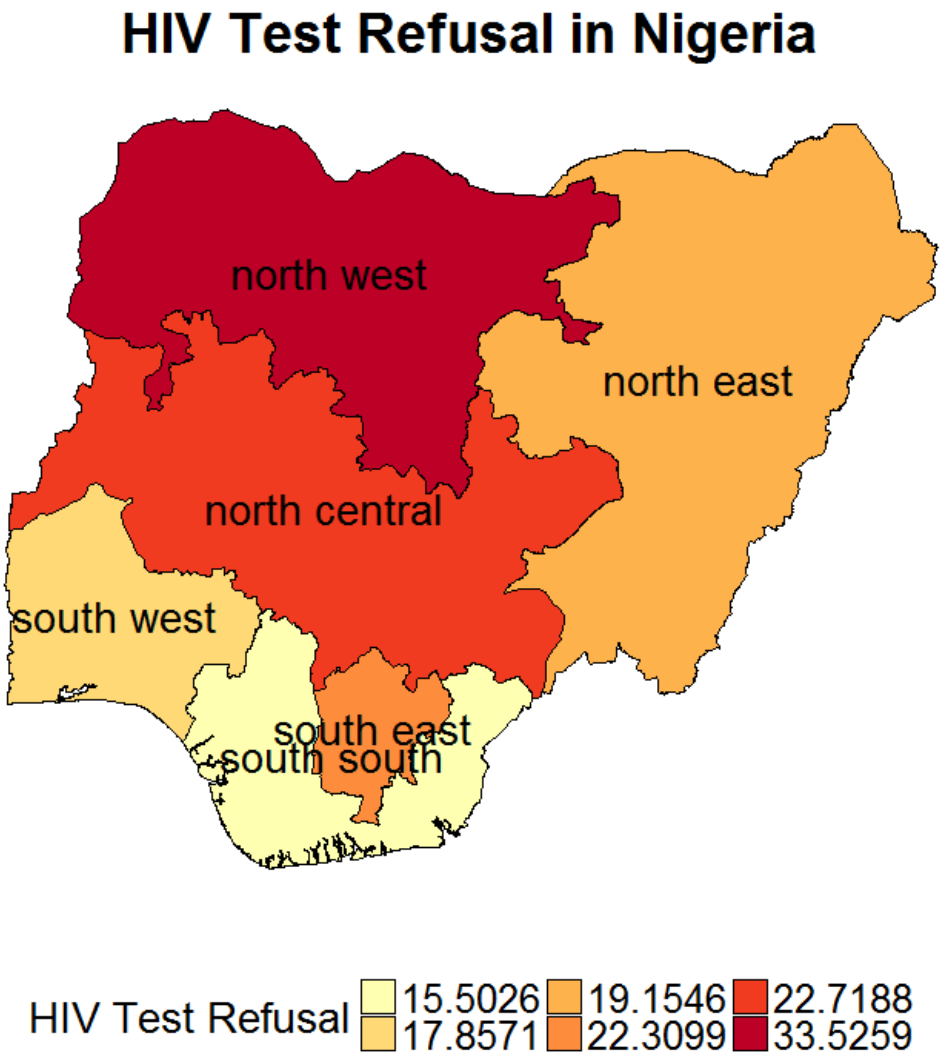

Figure 19: HIV test refusal in Nigeria by zones

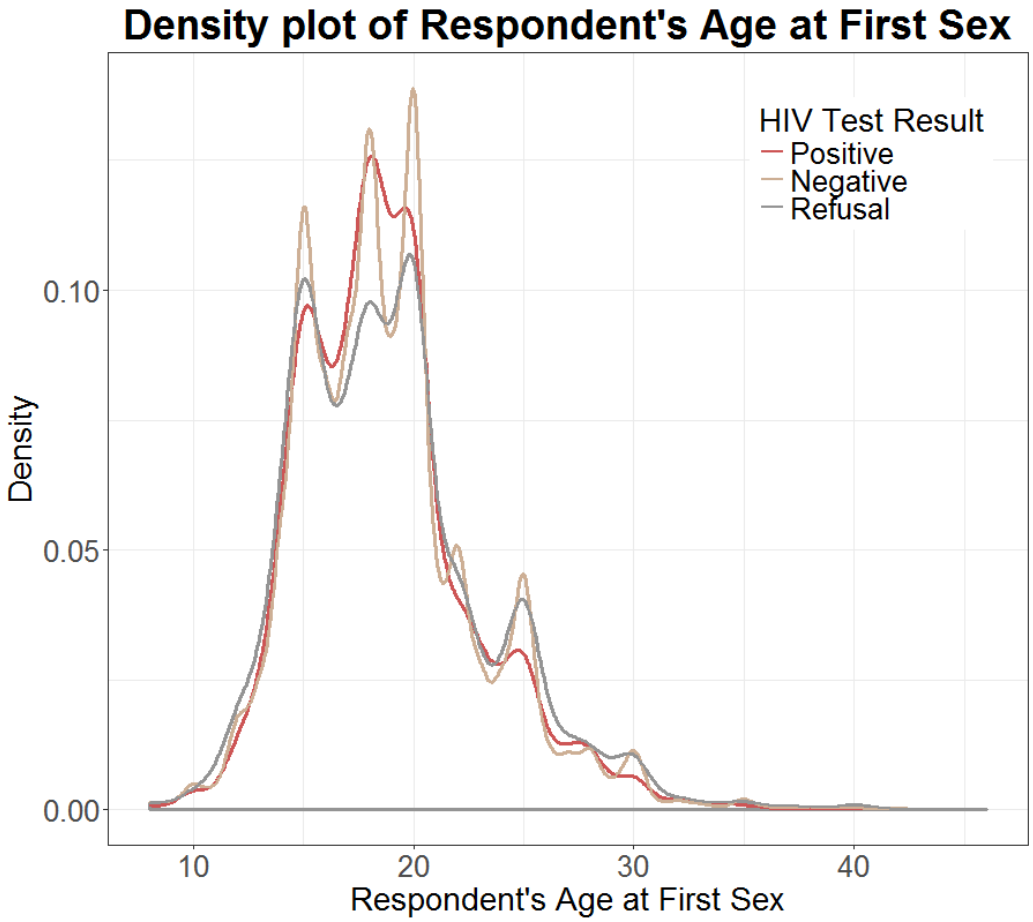

Figure 20: Density plot for HIV test result and respondents age at first sex

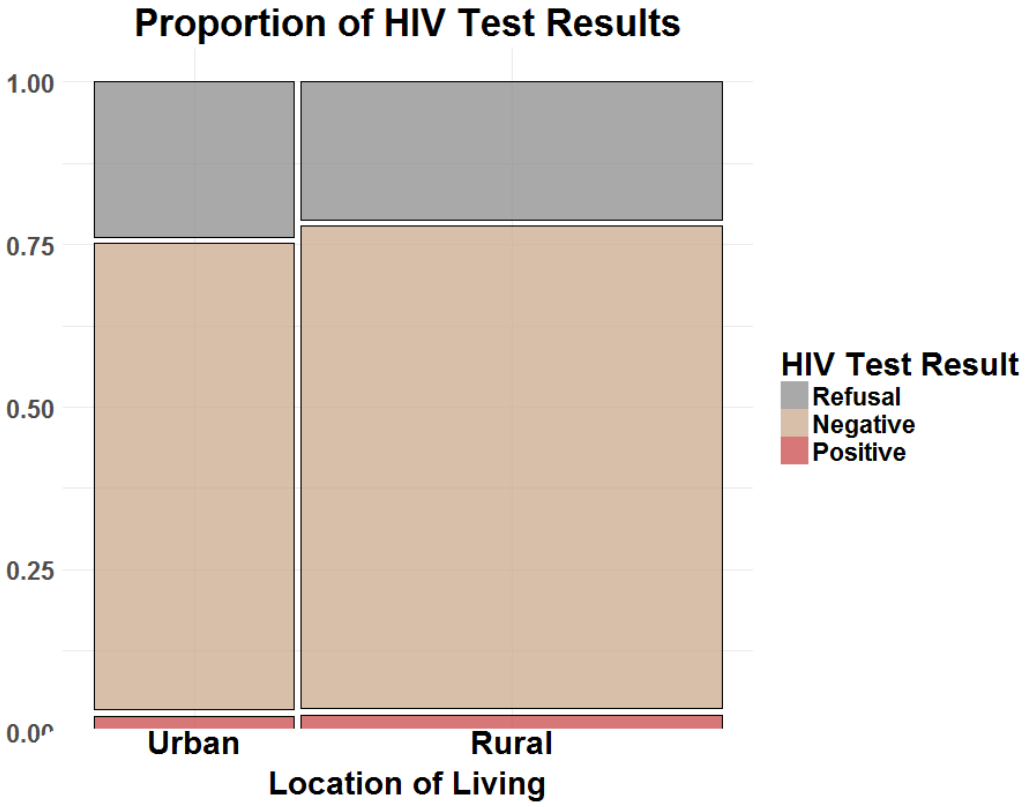

Figure 21: Mosaic plot for HIV test result and location of living

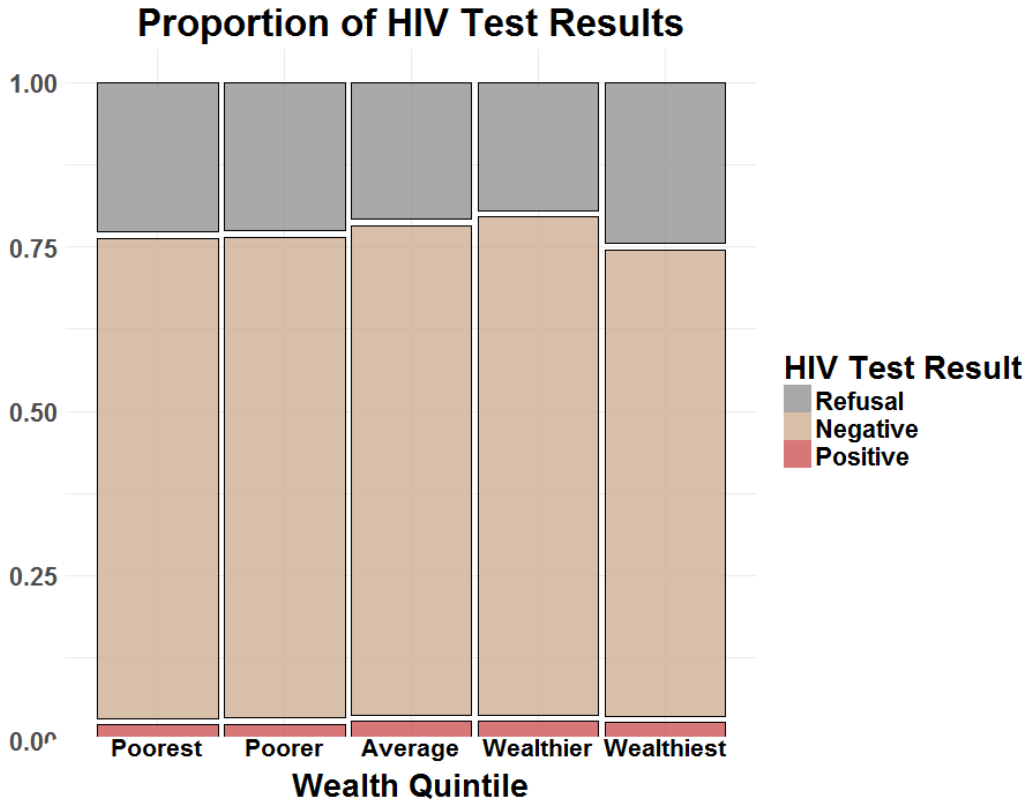

Figure 22: Mosaic plot for HIV test result and wealth quintile

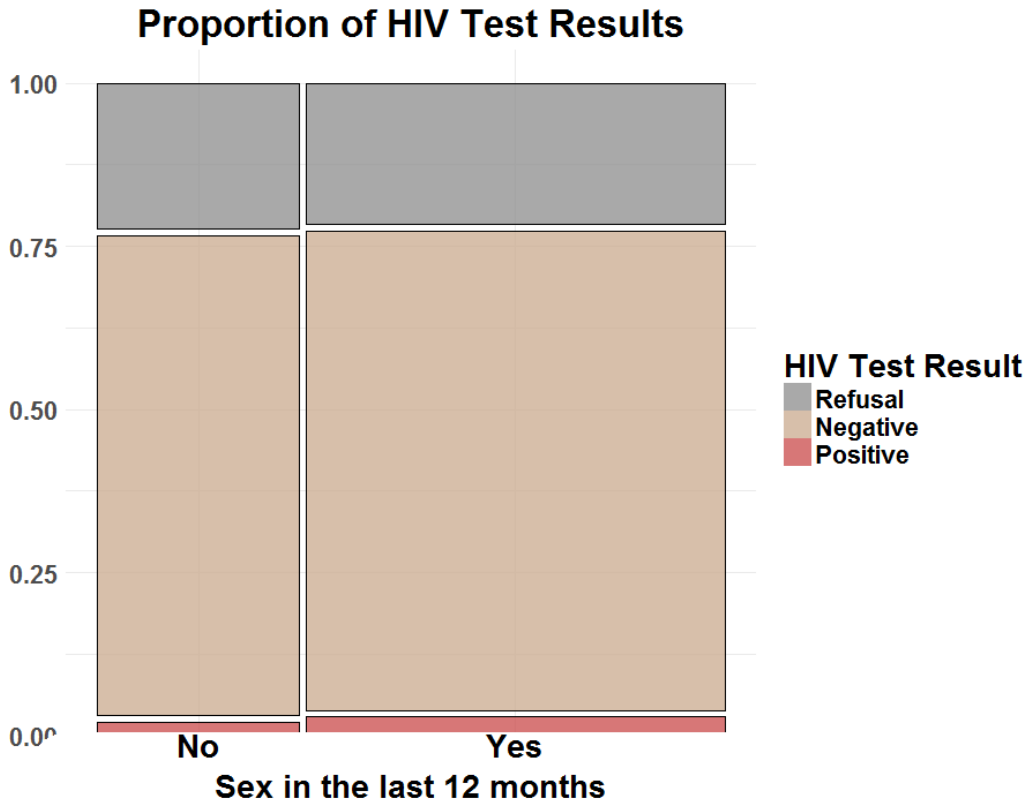

Figure 23: Mosaic plot for HIV test result and sex in the last 12 months

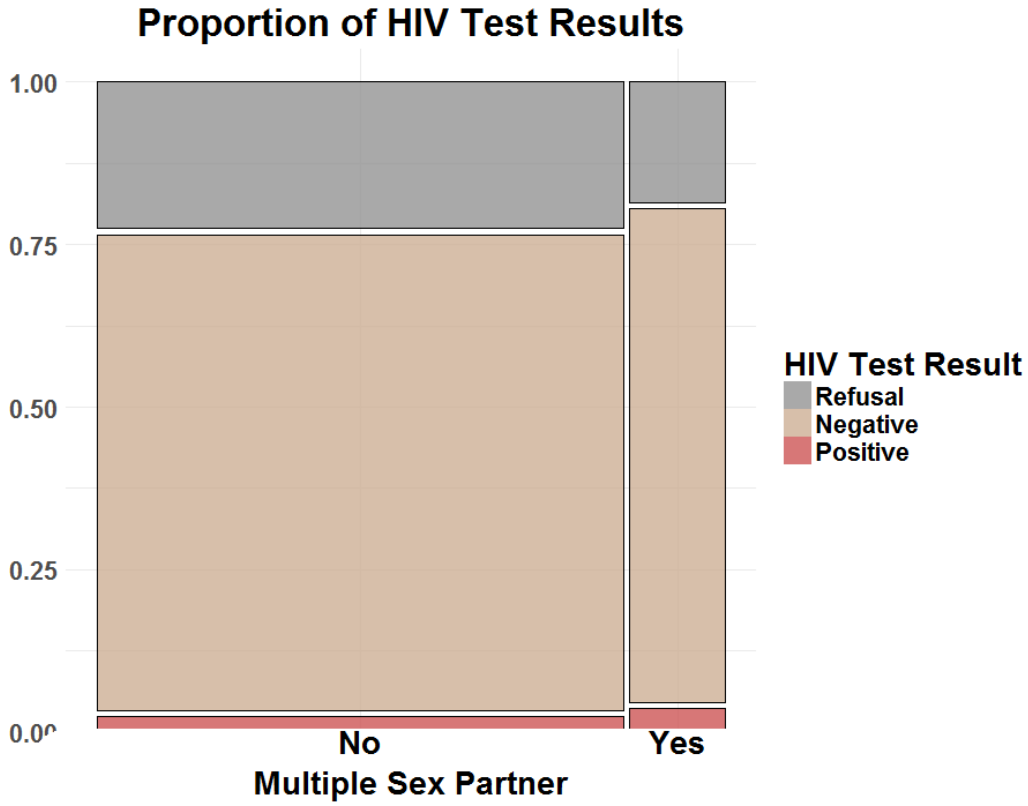

Figure 24: Mosaic plot for HIV test result and multiple sex partners

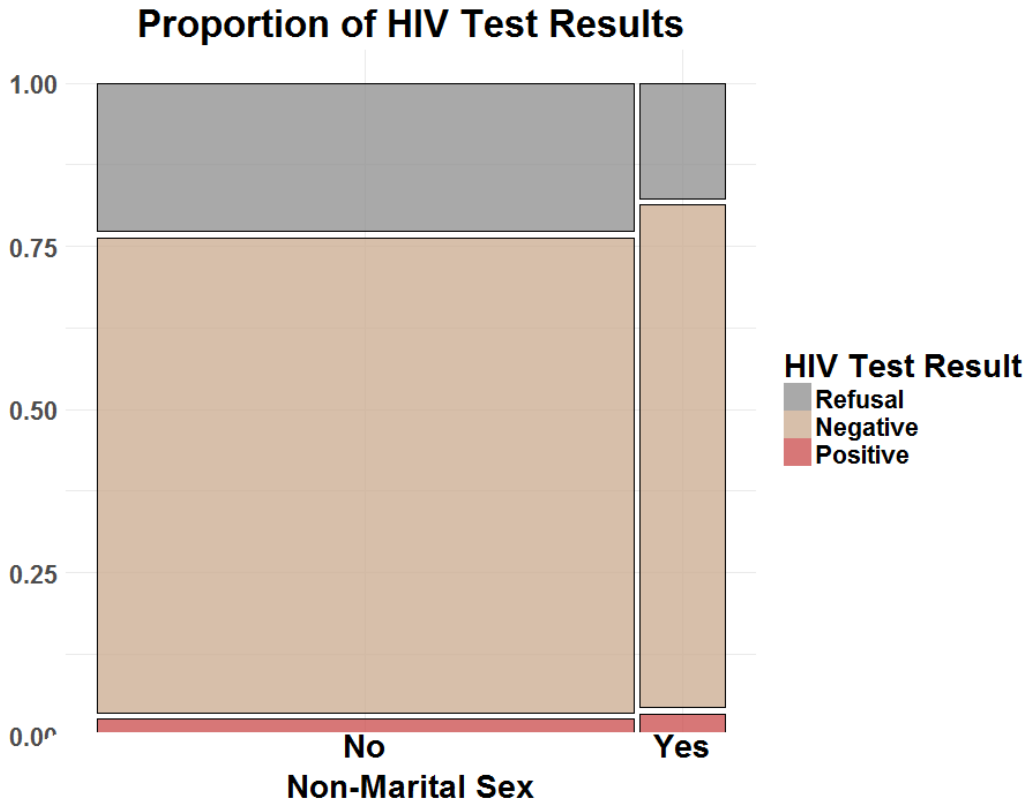

Figure 25: Mosaic plot for HIV test result and non-marital sex

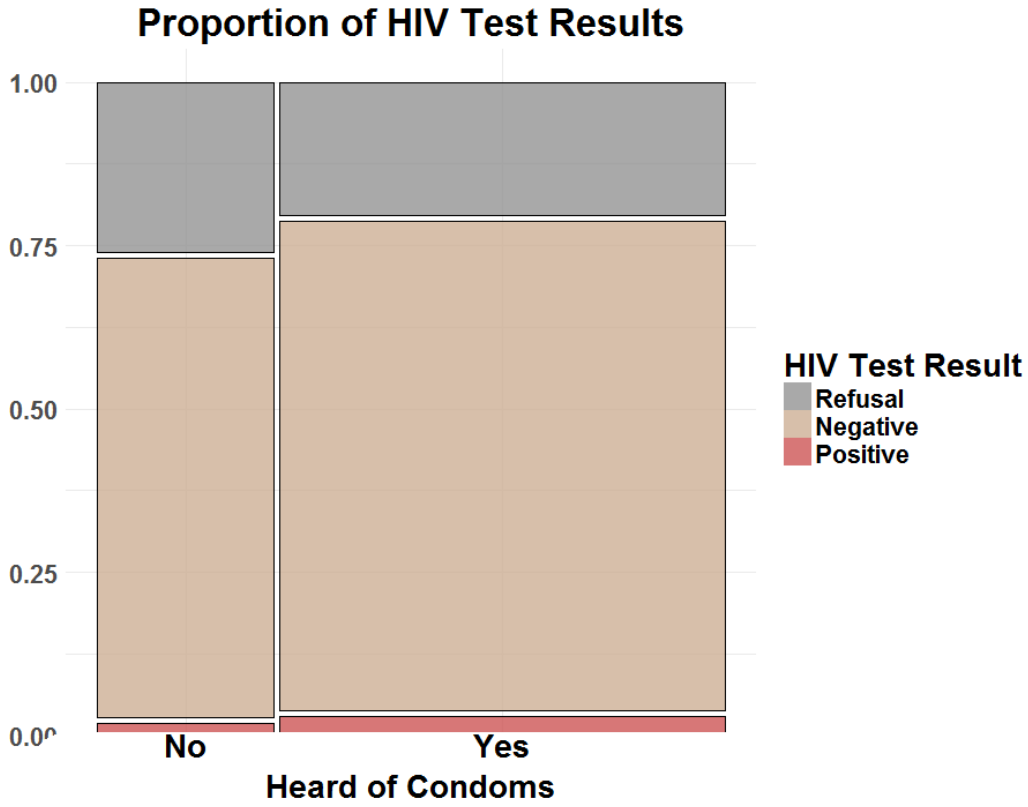

Figure 26: Mosaic plot for HIV test result and `CDHeard`

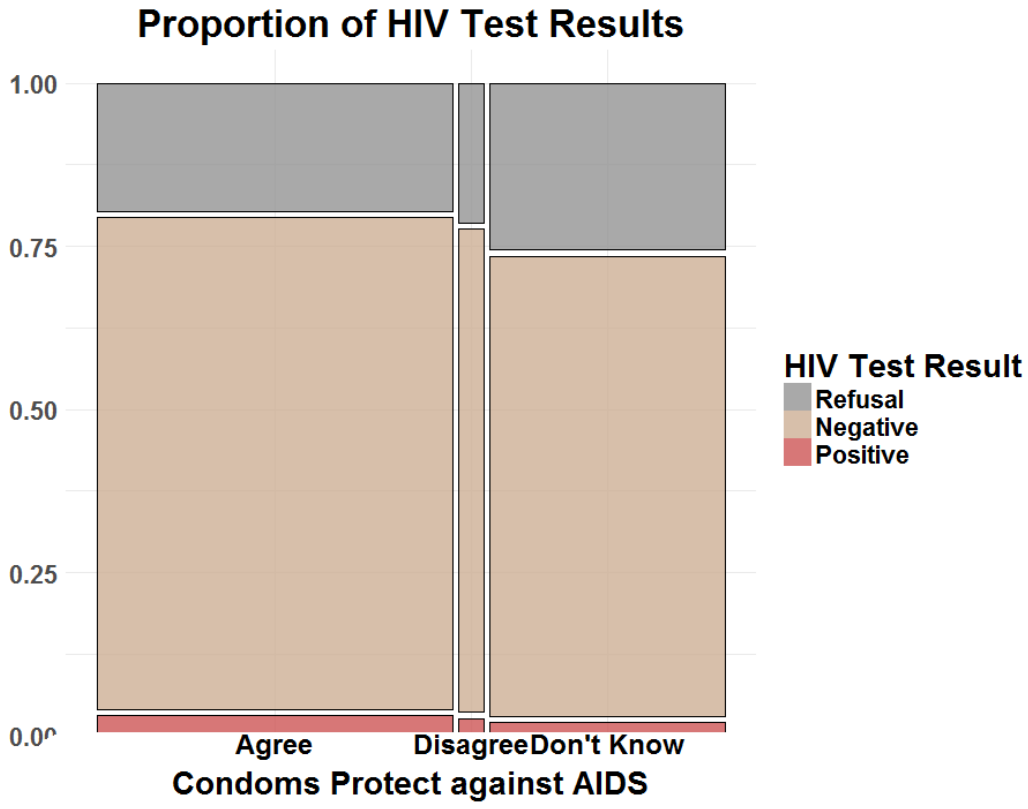

Figure 27: Mosaic plot for HIV test result and `CD-AIDS`

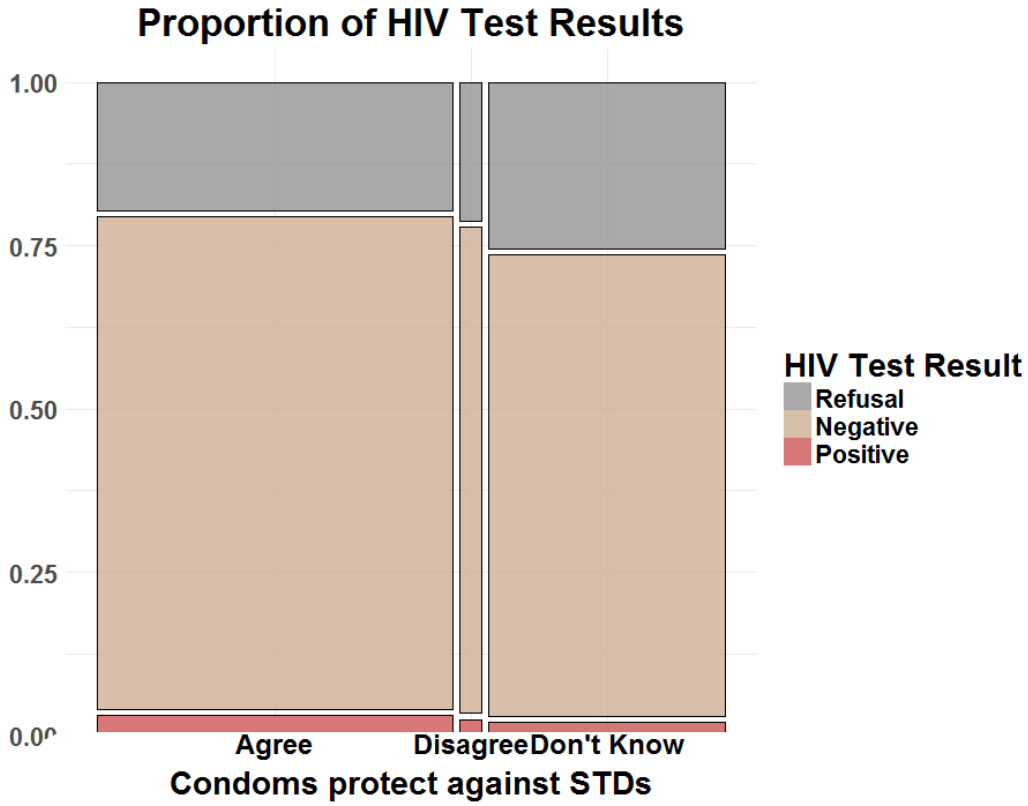

Figure 28: Mosaic plot for HIV test result and `CD-STD`

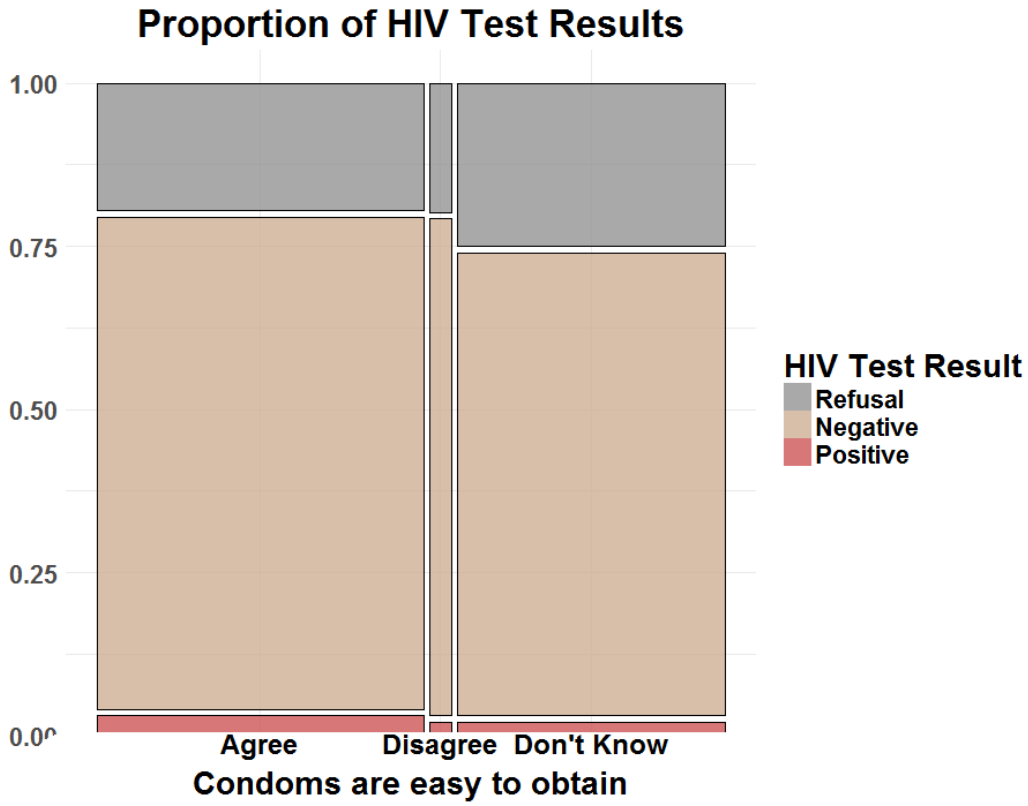

Figure 29: Mosaic plot for HIV test result and `CD-Obtain`

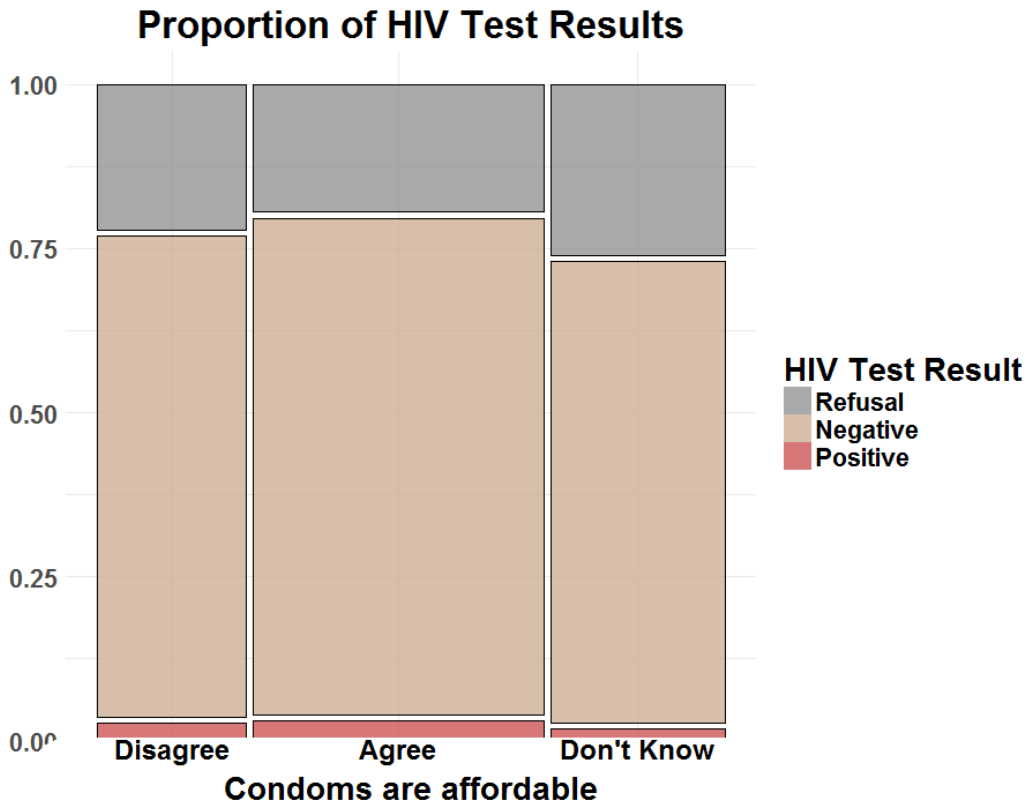

Figure 30: Mosaic plot for HIV test result and `CD-Afford`

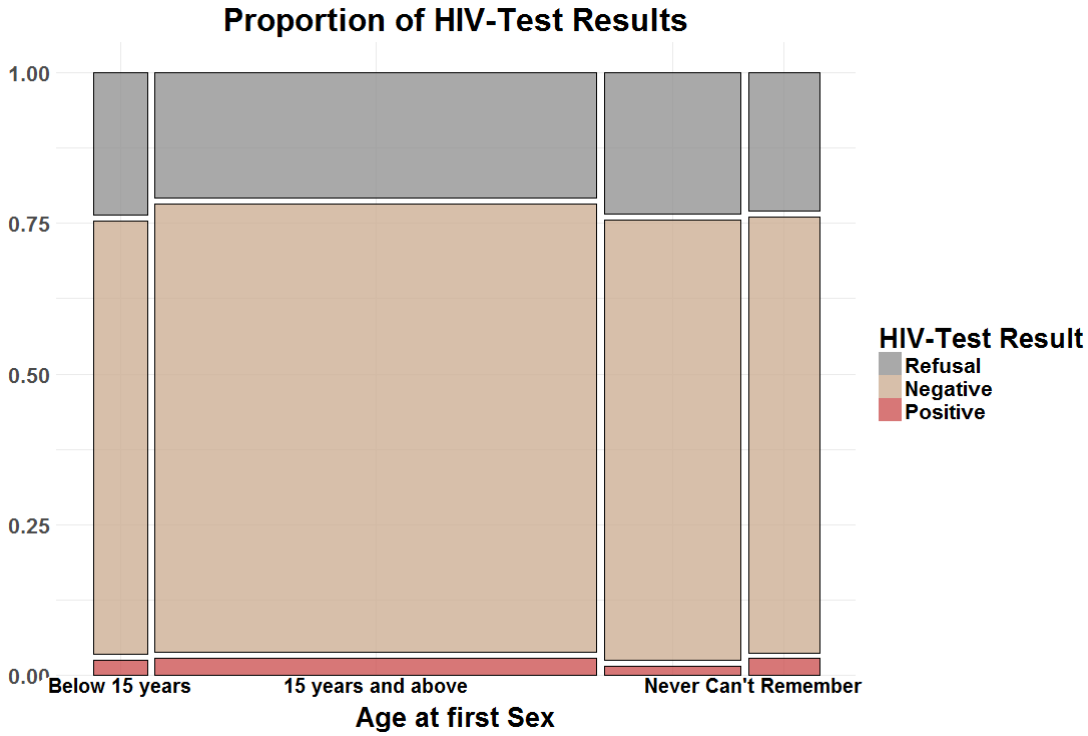

Figure 31: Mosaic plot for HIV test result and age at first sex

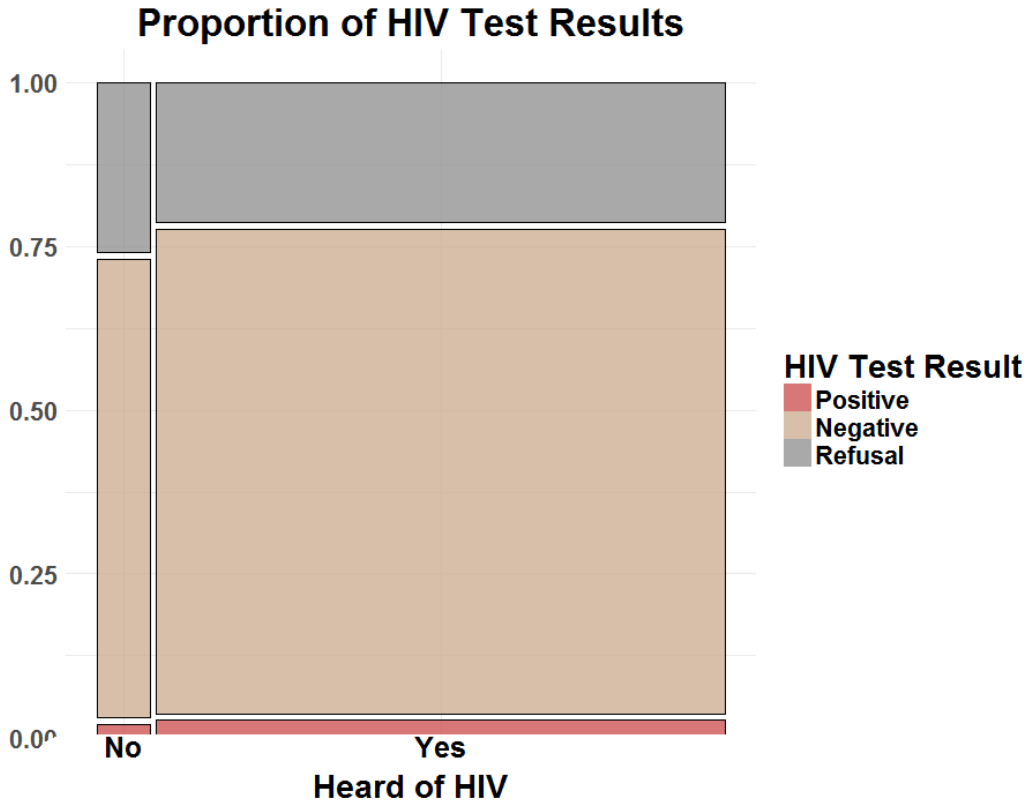

Figure 32: Mosaic plot for HIV test result and `HeardHIV`

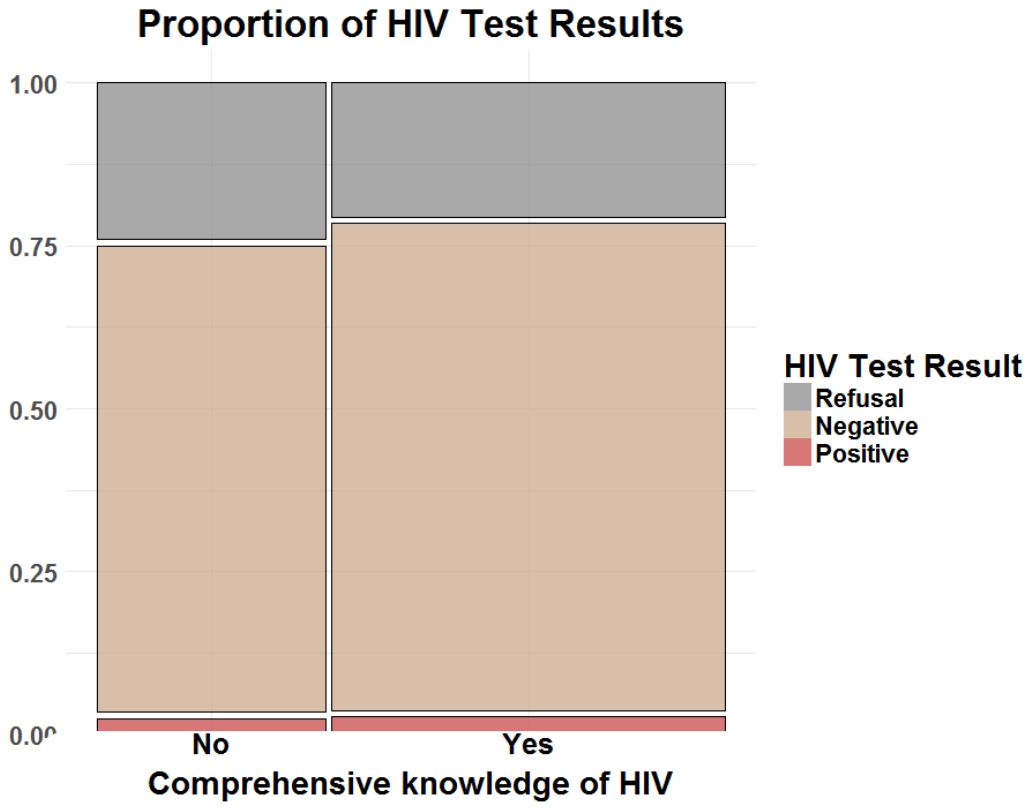

Figure 33: Mosaic plot for HIV test result and `CompknoHIV`

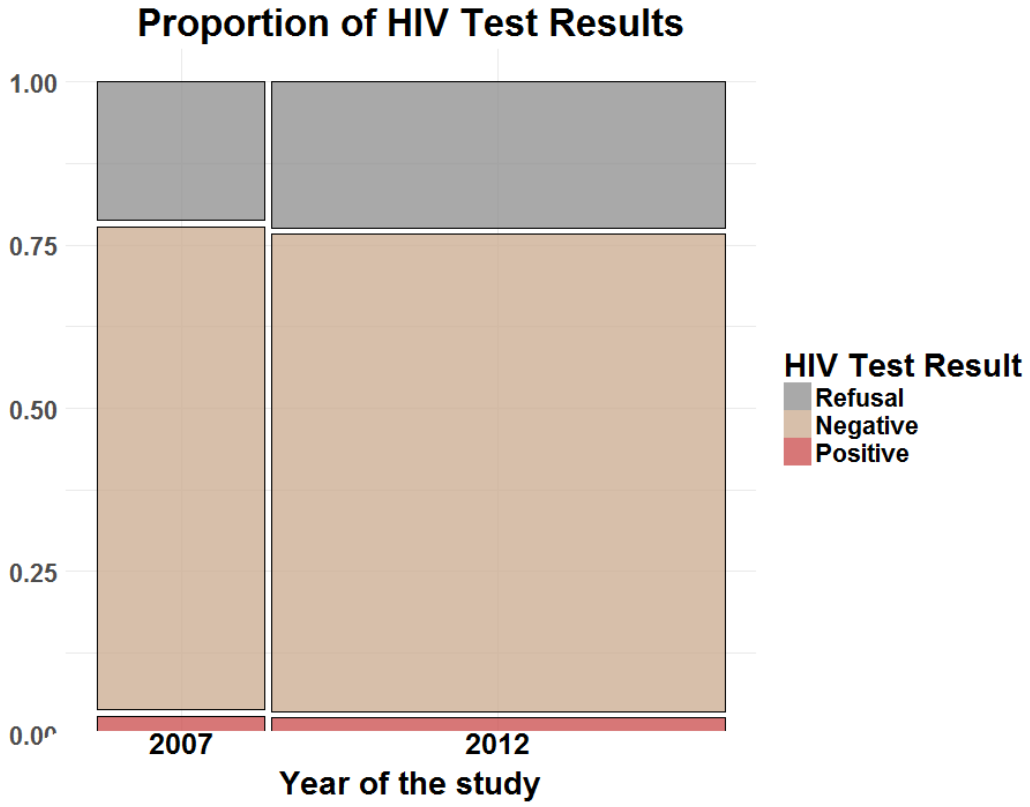

Figure 34: Mosaic plot for HIV test result and the year of the study

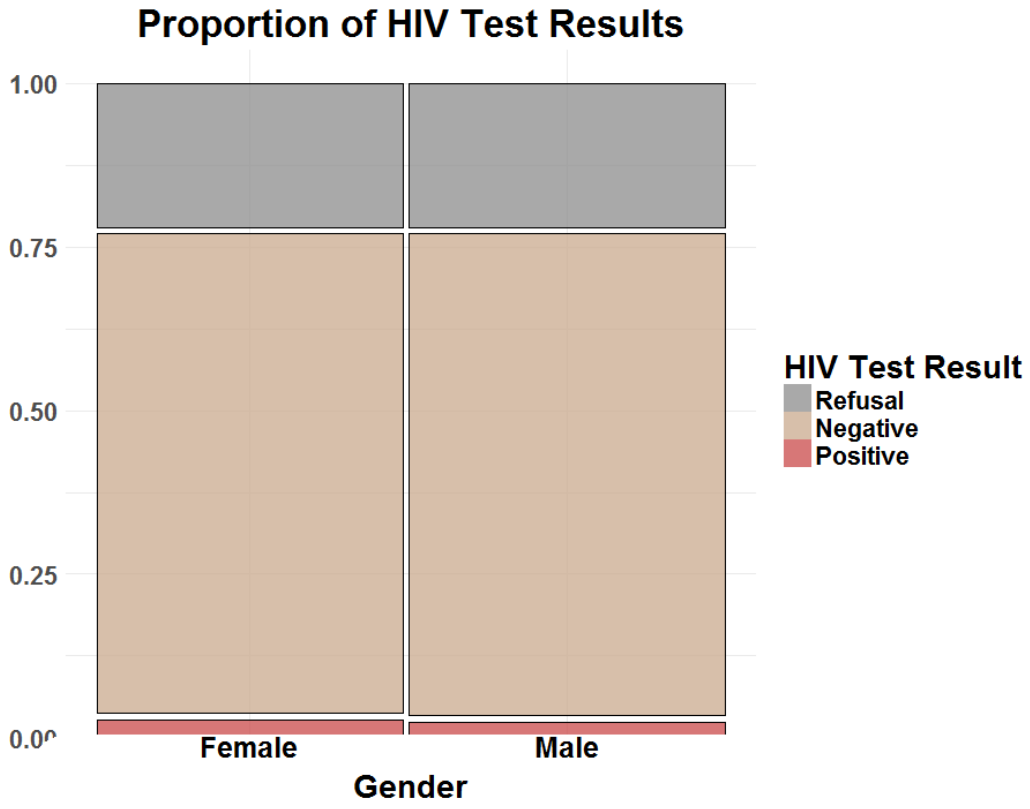

Figure 35: Mosaic plot for HIV test result and gender

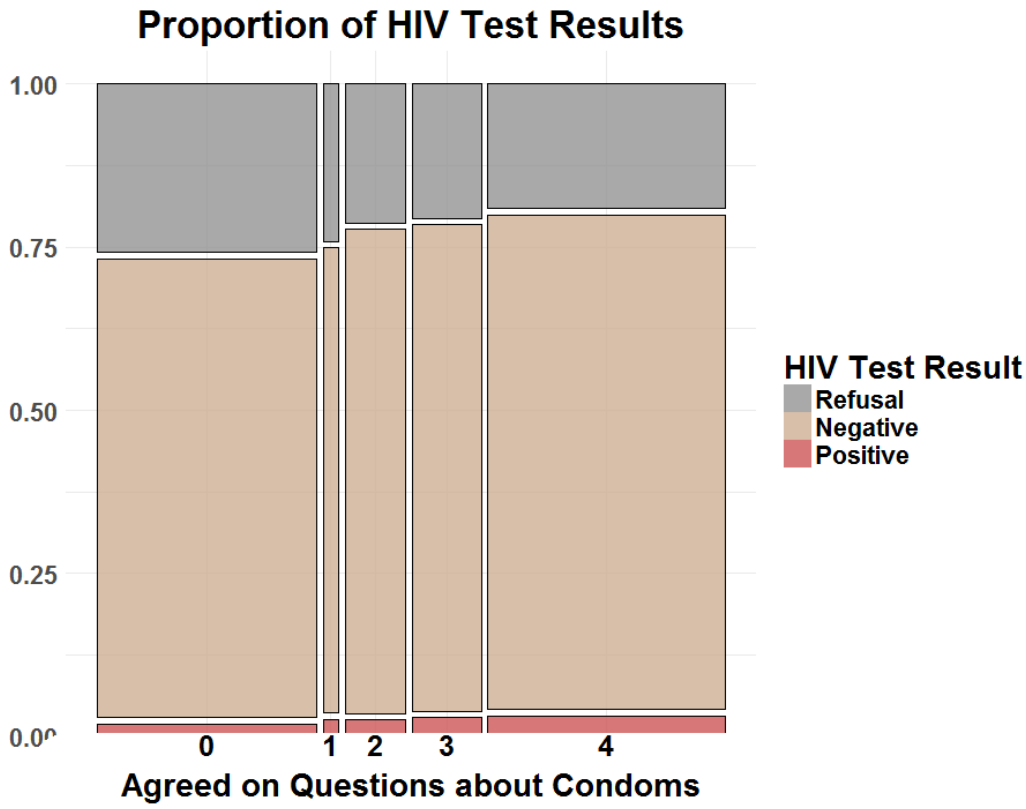

Figure 36: Mosaic plot for HIV test result and `CDagree`

B Appendix to Chapter 6.4

	educ_cat	wealthq	CDHeard	CD_AIDS	CD_STD	CD_Obtain	CD_Afford	HeardHIV	ExpSTIs	Sexgift	MultSex	Sex12m	CompknoHIV	Marital_cat	AgeSexcat	HIVTest_res	Missing Variables	frequency
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	32354
2	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	34
3	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	4
4	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	4
5	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	8
6	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	3
7	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	14
8	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	188
9	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	37
10	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	26
11	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	9195
12	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	13
13	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	225
14	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	2	3
15	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	2	1
16	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	2	2
17	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	2	1
18	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	2	1
19	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	2	6
20	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	2	1
21	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	2	1
22	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	2	1
23	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	2	5
24	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	2	112
25	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	2	17
26	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	2	5
27	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	2	1
28	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	2	1
29	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	2	9
30	1	1	1	1	1	1	1	0	1	1	1	1	0	1	1	1	2	17
31	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	2	2
32	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	1	2	1
33	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	2	3
34	1	1	1	1	1	1	1	1	0	1	1	1	1	0	1	1	2	2
35	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	2	62
36	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1	1	3	3
37	1	1	1	0	0	0	1	1	1	1	1	1	1	1	1	1	3	1
38	1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	0	3	1
39	1	1	1	1	1	1	1	1	1	0	1	0	1	1	1	0	3	1
40	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	0	3	1
41	1	1	1	1	1	1	1	0	1	1	1	1	0	1	0	1	3	1
42	1	1	1	1	1	1	1	0	0	1	1	1	0	1	1	1	3	22
43	1	1	1	1	1	1	1	0	1	1	1	1	0	1	1	0	3	12
44	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	3	2
45	1	1	1	1	1	1	1	1	0	1	1	1	1	0	1	0	3	1
46	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	0	3	1
47	1	1	1	1	1	1	1	0	1	1	1	1	0	0	1	1	3	1
48	1	1	1	0	0	0	0	1	1	1	1	1	1	1	1	1	4	2
49	1	1	1	1	1	1	1	1	1	0	0	0	1	1	0	1	4	71
50	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1	0	4	2
51	1	1	1	1	1	1	1	0	0	1	1	1	0	1	0	1	4	2
52	1	1	1	1	1	1	1	0	0	1	1	1	0	1	1	0	4	44

53	1	1	1	0	0	0	1	1	1	1	1	1	1	0	1	1	4	1
54	1	1	1	0	0	1	0	1	1	1	1	1	1	0	1	1	4	1
55	1	1	1	1	1	1	1	0	0	1	1	1	0	0	1	1	4	1
56	1	1	1	1	1	1	1	0	1	1	1	1	0	0	1	0	4	2
57	1	1	0	0	0	0	0	1	1	1	1	1	1	1	1	1	5	4
58	1	1	1	1	1	1	1	1	0	0	0	0	1	1	0	1	5	1
59	1	1	1	0	0	0	0	1	1	1	1	1	1	1	1	0	5	1
60	1	1	1	1	1	1	1	1	1	0	0	0	1	1	0	0	5	29
61	1	0	1	1	1	1	1	0	0	1	1	1	0	1	1	0	5	1
62	1	1	1	1	1	1	1	0	0	0	1	1	0	1	1	0	5	1
63	1	1	1	1	1	1	1	0	0	1	1	1	0	1	0	0	5	2
64	0	1	1	1	1	1	1	0	0	1	1	1	0	1	1	0	5	1
65	1	1	1	1	1	1	1	1	1	0	0	0	1	0	0	1	5	6
66	1	1	1	1	1	1	1	0	0	1	1	1	0	0	1	0	5	6
67	1	1	0	0	0	0	0	1	1	1	1	1	1	1	1	0	6	1
68	1	1	1	1	1	1	1	1	1	0	0	0	0	1	0	0	6	2
69	1	1	0	0	0	0	0	1	1	1	1	1	1	0	1	1	6	21
70	1	1	1	1	1	1	1	1	1	0	0	0	1	0	0	0	6	2
71	1	1	0	0	0	0	0	1	1	1	1	1	1	1	0	0	7	1
72	1	1	1	1	1	1	1	0	0	0	0	0	0	1	0	1	7	10
73	1	1	1	1	1	1	1	0	0	0	0	0	0	1	1	0	7	1
74	1	1	0	0	0	0	0	1	1	1	1	1	1	0	0	1	7	2
75	1	1	0	0	0	0	0	1	1	1	1	1	1	0	1	0	7	5
76	1	1	1	1	1	1	0	0	0	0	0	0	0	1	0	1	8	3
77	1	1	1	1	1	1	1	0	0	0	0	0	0	1	0	0	8	19
78	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	1	8	1
79	1	1	1	1	1	0	0	0	0	0	0	0	0	1	0	1	9	1
80	1	1	1	1	1	1	0	0	0	0	0	0	0	1	0	0	9	1
81	1	1	0	0	0	0	0	0	0	1	1	1	0	0	1	1	9	1
82	1	1	0	0	0	0	0	1	1	0	0	0	1	0	0	1	10	2
83	1	1	0	0	0	0	0	1	1	0	0	0	1	0	1	0	10	1
84	1	0	0	0	0	0	0	1	1	0	0	0	1	0	0	1	11	1
85	1	1	0	0	0	0	0	1	0	0	0	0	1	0	0	1	11	1
86	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	12	6
87	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	12	1
88	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	13	2
89	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	13	27
90	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	14	2
91	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14	32
92	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	15	5
93	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	15	13
94	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	8
NAs	69	73	136	143	143	148	167	246	249	259	264	269	273	442	560	9610		

Table 9: Table of the missing data patterns

C Appendix to Chapter 7.2

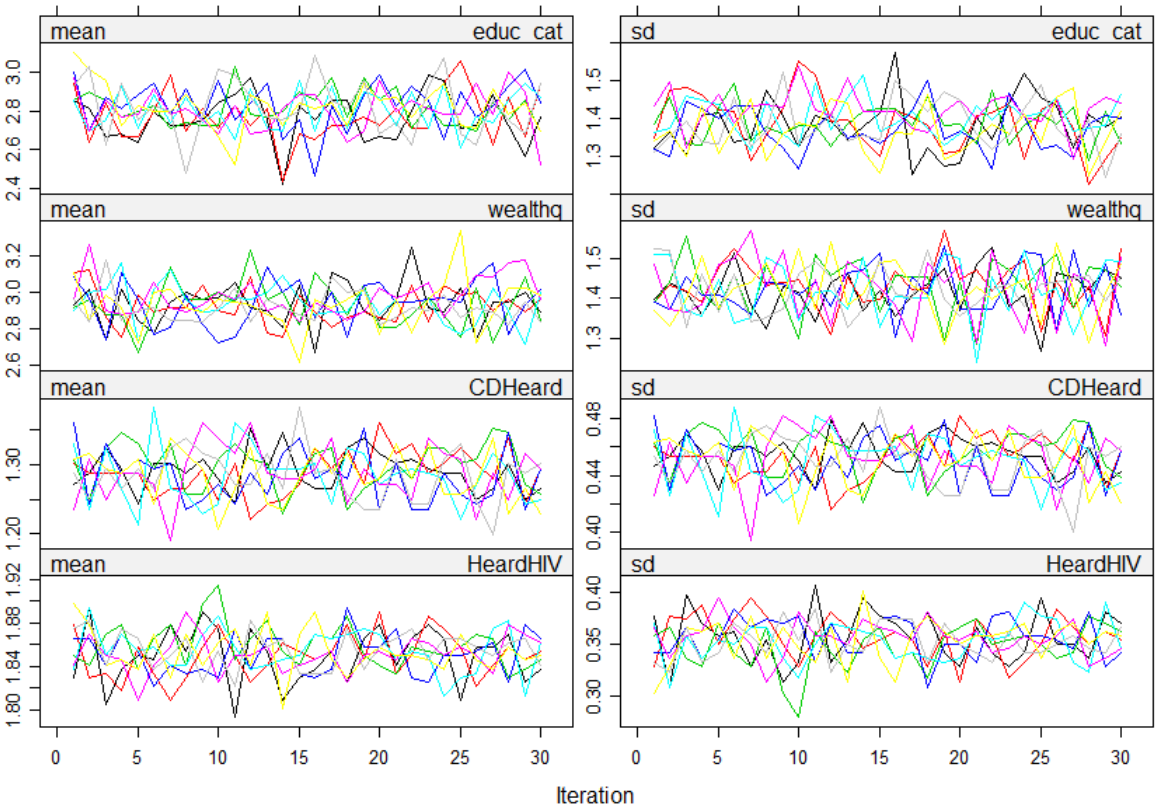

Figure 37: Trace line plot for imputations of variables `educ-cat`, `wealthq`, `CDHeard` and `HeardHIV` with covariable `State`

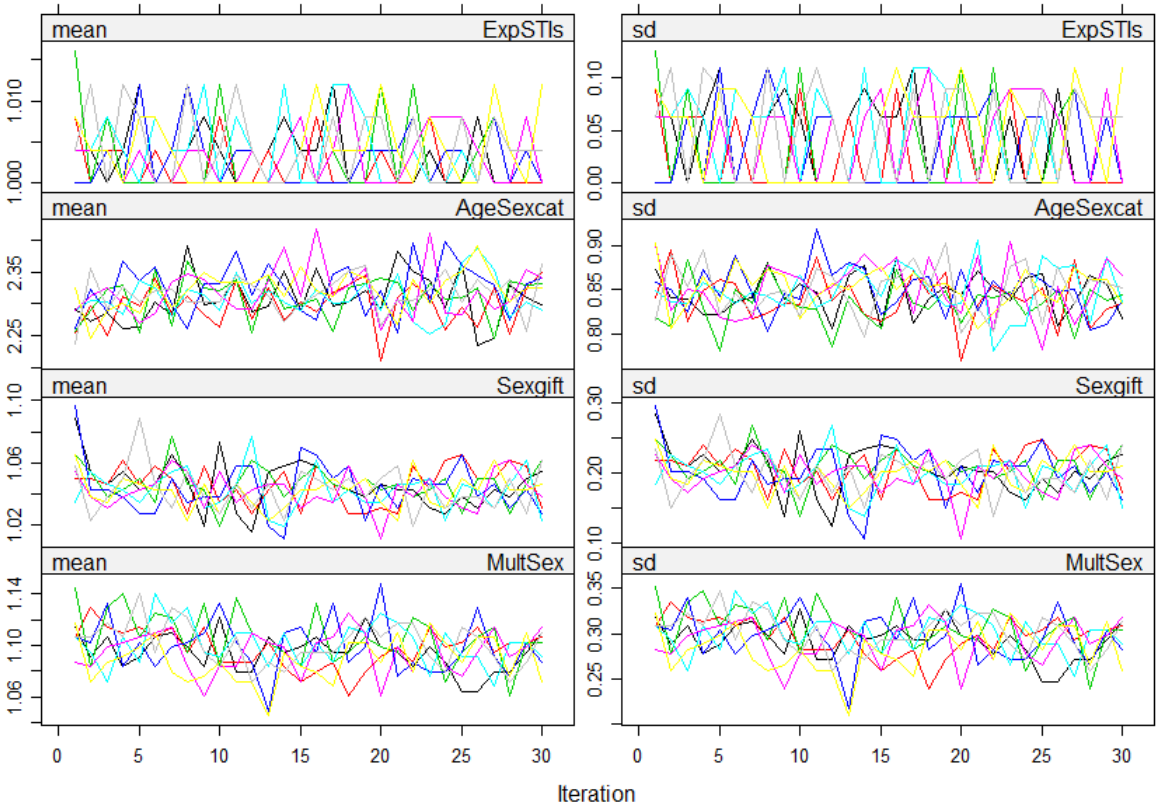

Figure 38: Trace line plot for imputations of variables `ExpSTIs`, `AgeSexcat`, `Sexgift` and `MultSex` with covariable `State`

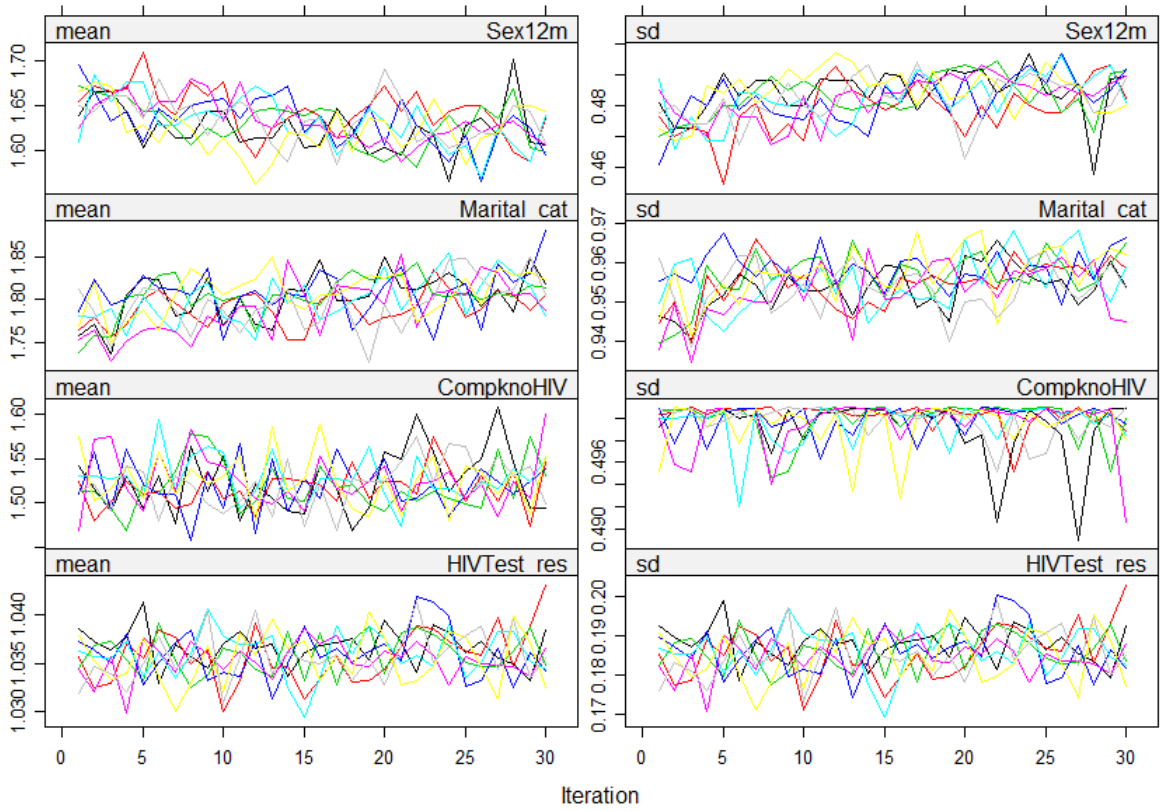

Figure 39: Trace line plot for imputations of variables `Sex12m`, `Marital-cat`, `CompknoHIV` and `HIVTest-res` with covariable `State`

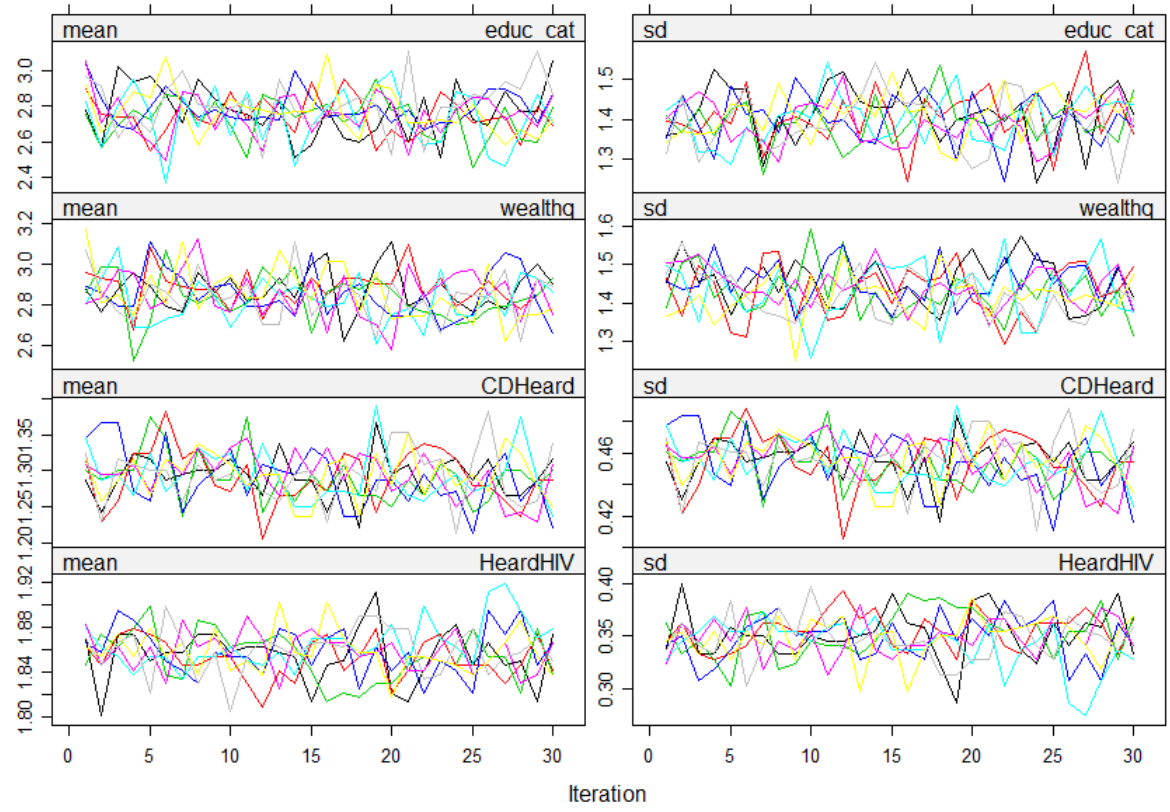

Figure 40: Trace line plot for imputations of variables `educ-cat`, `wealthq`, `CDHeard` and `HeardHIV` with covariable `zone`

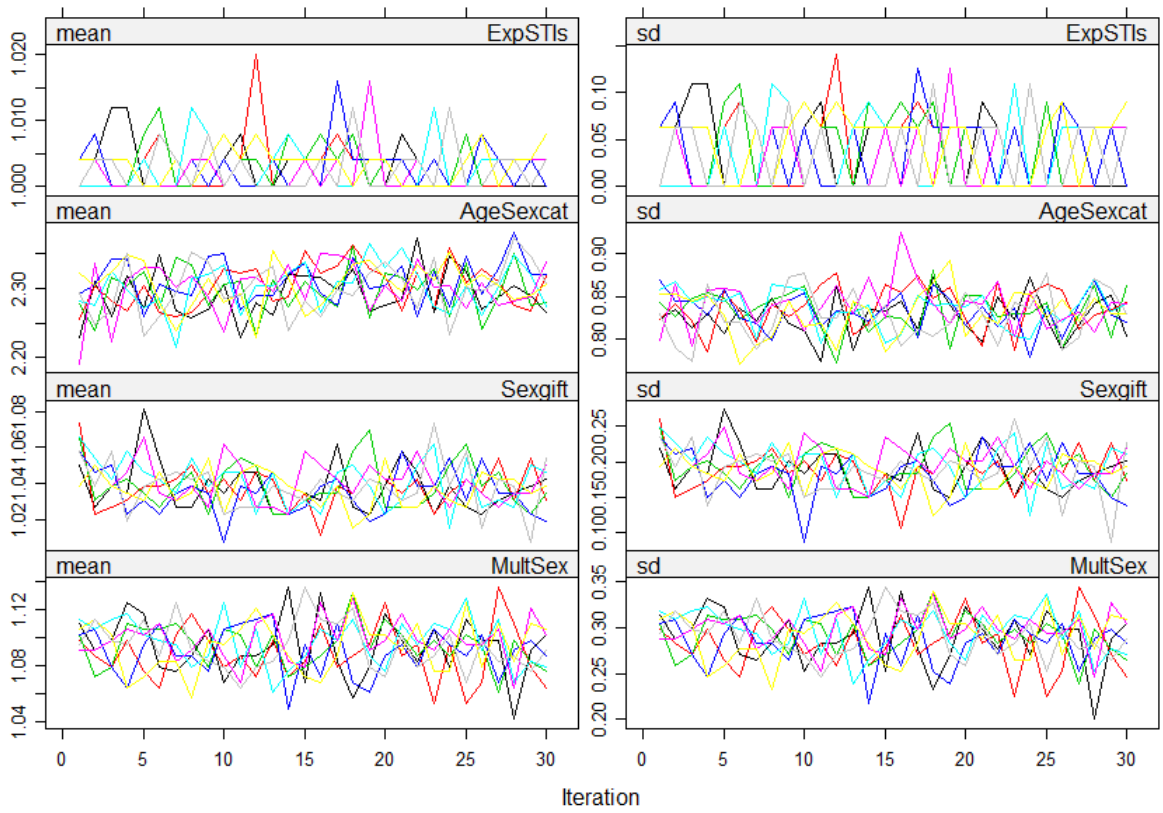

Figure 41: Trace line plot for imputations of variables `ExpSTIs`, `AgeSexcat`, `Sexgift` and `MultSex` with covariable `zone`

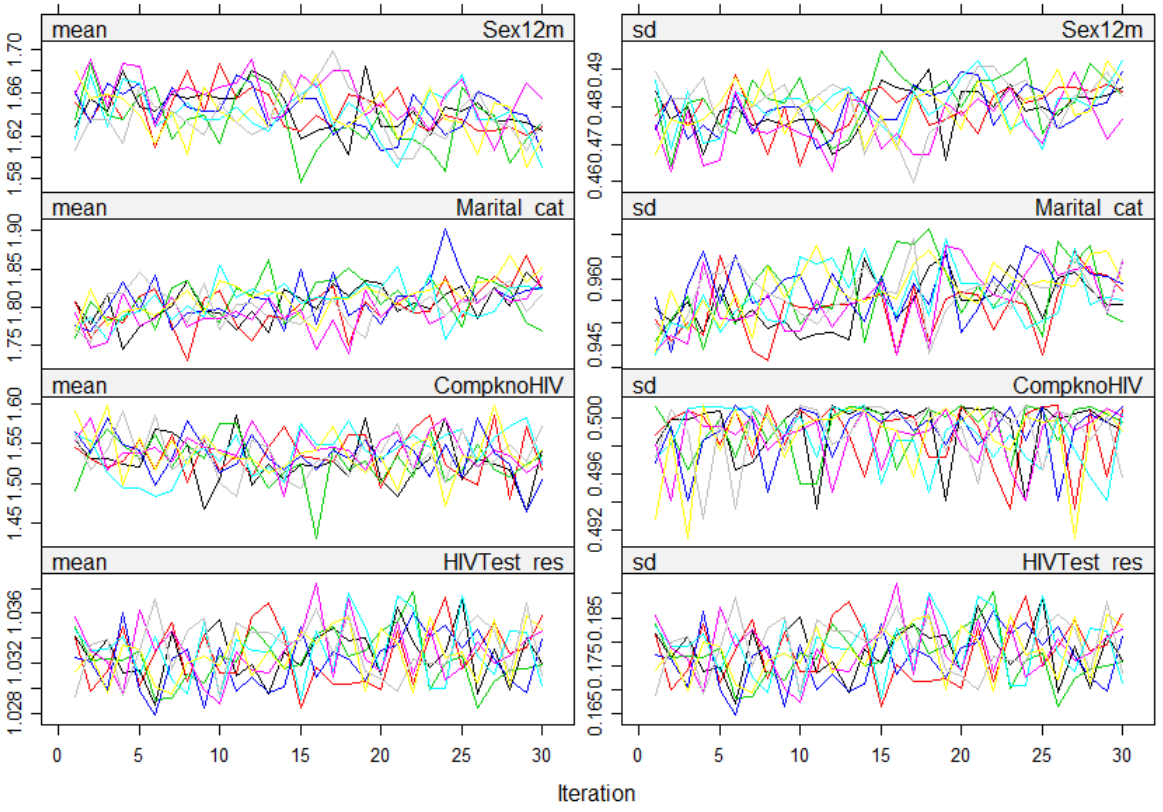

Figure 42: Trace line plot for imputations of variables `Sex12m`, `Marital-cat`, `CompknoHIV` and `HIVTest-res` with covariable `zone`

D Appendix to Chapter 7.3

	est	se	t	df	Pr(> |t|)	lo 95	hi 95	fmi
(Intercept)	-3.6245	0.3635	-9.9706	31478	0	-4.3370	-2.9120	0.0033
wealthq2	0.0285	0.1041	0.2738	32959	0.7842	-0.1755	0.2325	0.0010
wealthq3	0.2719	0.1084	2.5095	31340	0.0121	0.0595	0.4843	0.0034
wealthq4	0.3262	0.1189	2.7434	31849	0.0061	0.0931	0.5592	0.0029
wealthq5	0.3613	0.1317	2.7424	31223	0.0061	0.1031	0.6195	0.0035
location2	-0.0258	0.0824	-0.3132	33064	0.7541	-0.1873	0.1357	0.0005
Sexgift2	0.0831	0.1150	0.7227	31541	0.4699	-0.1423	0.3086	0.0032
MultSex2	0.1525	0.0845	1.8050	31851	0.0711	-0.0131	0.3181	0.0029
Sex12m2	-0.0757	0.1060	-0.7139	30284	0.4753	-0.2835	0.1321	0.0044
NonmarSex12	0.0791	0.1147	0.6902	32817	0.4901	-0.1456	0.3039	0.0013
CDHeard2	0.2211	0.1199	1.8438	19410	0.0652	-0.0139	0.4560	0.0122
AgeSexcat2	0.0522	0.1194	0.4373	30092	0.6619	-0.1818	0.2862	0.0046
AgeSexcat3	-0.3299	0.1822	-1.8109	32244	0.0702	-0.6869	0.0272	0.0023
AgeSexcat4	0.1178	0.1510	0.7802	9728	0.4353	-0.1782	0.4139	0.0226
educ_cat2	0.0846	0.1567	0.5400	33056	0.5892	-0.2225	0.3918	0.0006
educ_cat3	0.3503	0.1074	3.2633	31985	0.0011	0.1399	0.5608	0.0027
educ_cat4	0.2315	0.1110	2.0847	31842	0.0371	0.0138	0.4492	0.0029
educ_cat5	0.0387	0.1397	0.2771	31679	0.7817	-0.2351	0.3125	0.0031
ExpSTIs2	0.9772	0.2996	3.2620	33083	0.0011	0.3900	1.5643	0.0004
RespAge	-0.0237	0.0429	-0.5526	32618	0.5805	-0.1078	0.0604	0.0018
HeardHIV2	-0.0555	0.1406	-0.3945	14720	0.6932	-0.3311	0.2202	0.0163
CompknoHIV2	-0.1221	0.0719	-1.6981	21688	0.0895	-0.2630	0.0188	0.0105
Yearstud2	-0.1352	0.0693	-1.9500	33097	0.0512	-0.2711	0.0007	0.0003
Marital_cat2	0.4503	0.1429	3.1500	8321	0.0016	0.1701	0.7305	0.0252
Marital_cat3	-0.2542	0.1212	-2.0978	32865	0.0359	-0.4917	-0.0167	0.0012
Male2	-0.1899	0.0710	-2.6739	33076	0.0075	-0.3291	-0.0507	0.0005
Religion2	0.0081	0.2699	0.0299	33097	0.9761	-0.5210	0.5371	0.0003
Religion3	-0.1919	0.2759	-0.6955	33096	0.4868	-0.7325	0.3488	0.0003
CDagree	0.0182	0.0509	0.3579	30389	0.7204	-0.0816	0.1180	0.0043

Table 10: Pooled mixed effects logistic regression model

	obs	Positive	ID
1	Positive	0.9948	33111
2	Negative	0.9932	8881
3	Positive	0.9905	33111
4	Negative	0.9905	24281
5	Negative	0.9885	4472
6	Negative	0.9863	23105
7	Negative	0.9854	4472
8	Positive	0.9853	33111
9	Positive	0.9844	32184
10	Negative	0.9841	8881
11	Negative	0.9835	23105
12	Positive	0.9824	33111
13	Negative	0.9819	3853
14	Negative	0.9803	11362
15	Negative	0.9795	684
16	Negative	0.9792	3853
17	Negative	0.9792	30612
18	Negative	0.9789	4348
19	Positive	0.9779	32876
20	Negative	0.9776	30612
21	Negative	0.9774	3853
22	Positive	0.9773	32184
23	Negative	0.9759	8881
24	Positive	0.9753	32876
25	Negative	0.9753	7236

Table 11: The cases in testing with highest probability for positive HIV test and their real result and case number for naive Bayes with down-sampling and covariable `zone`

	obs	Positive
1	Positive	0.9450
2	Positive	0.9428
3	Positive	0.9428
4	Positive	0.9425
5	Positive	0.9425
6	Positive	0.9424
7	Positive	0.9421
8	Positive	0.9417
9	Positive	0.9414
10	Positive	0.9411
11	Negative	0.9406
12	Positive	0.9385
13	Positive	0.9373
14	Negative	0.9370
15	Negative	0.9361
16	Negative	0.9358
17	Negative	0.9345
18	Positive	0.9343
19	Positive	0.9341
20	Negative	0.9339

Table 12: The cases in testing with highest probability for positive HIV test and their real result and case number for logistic regression with up-sampling and covariable `zone`

	est	se	t	df	Pr(> |t|)	lo 95	hi 95	fmi
(Intercept)	-0.2564	0.0977	-2.6240	1447.0000	0.0088	-0.4482	-0.0647	0.0700
wealthq2	0.0314	0.0277	1.1330	8518.0000	0.2572	-0.0229	0.0857	0.0269
wealthq3	0.2425	0.0290	8.3708	13691.0000	0.0000	0.1857	0.2992	0.0201
wealthq4	0.2878	0.0317	9.0690	9373.0000	0.0000	0.2256	0.3500	0.0254
wealthq5	0.3317	0.0347	9.5657	10446.0000	0.0000	0.2638	0.3997	0.0238
zone2	0.2243	0.0280	8.0194	40547.0000	0.0000	0.1695	0.2792	0.0079
zone3	-0.0179	0.0299	-0.5980	31726.0000	0.5499	-0.0764	0.0407	0.0106
zone4	-1.0229	0.0332	-30.7960	26180.0000	0.0000	-1.0880	-0.9578	0.0126
zone5	-0.3537	0.0279	-12.6603	53401.0000	0.0000	-0.4084	-0.2989	0.0046
zone6	-0.5615	0.0286	-19.6286	63682.0000	0.0000	-0.6176	-0.5054	0.0007
location2	0.1535	0.0216	7.0913	24093.0000	0.0000	0.1111	0.1960	0.0135
Sexgift2	0.1361	0.0342	3.9814	2216.0000	0.0001	0.0691	0.2031	0.0560
MultSex2	0.1695	0.0242	7.0159	2601.0000	0.0000	0.1221	0.2169	0.0515
Sex12m2	-0.0685	0.0326	-2.0987	124.0000	0.0379	-0.1330	-0.0039	0.2496
NonmarSex12	0.0569	0.0321	1.7703	1294.0000	0.0769	-0.0062	0.1199	0.0742
CDHeard2	0.2235	0.0313	7.1491	5300.0000	0.0000	0.1622	0.2848	0.0351
AgeSexcat2	0.0757	0.0331	2.2864	978.0000	0.0224	0.0107	0.1407	0.0858
AgeSexcat3	-0.3519	0.0489	-7.1889	2506.0000	0.0000	-0.4478	-0.2559	0.0525
AgeSexcat4	0.1523	0.0471	3.2366	91.0000	0.0017	0.0588	0.2458	0.2921
educ_cat2	0.0770	0.0394	1.9539	42608.0000	0.0507	-0.0002	0.1542	0.0074
educ_cat3	0.4239	0.0287	14.7821	9254.0000	0.0000	0.3677	0.4802	0.0256
educ_cat4	0.2937	0.0291	10.0845	11016.0000	0.0000	0.2366	0.3508	0.0231
educ_cat5	0.1158	0.0370	3.1328	16584.0000	0.0017	0.0433	0.1882	0.0177
ExpSTIs2	1.0749	0.1179	9.1135	54580.0000	0.0000	0.8437	1.3060	0.0043
RespAge	-0.0005	0.0010	-0.5118	16471.0000	0.6088	-0.0024	0.0014	0.0178
HeardHIV2	-0.0658	0.0374	-1.7588	1550.0000	0.0788	-0.1393	0.0076	0.0675
CompknoHIV2	-0.0948	0.0213	-4.4537	199.0000	0.0000	-0.1367	-0.0528	0.1954
Yearstud2	-0.1374	0.0185	-7.4298	62104.0000	0.0000	-0.1737	-0.1012	0.0018
Marital_cat2	0.4925	0.0534	9.2270	45.0000	0.0000	0.3850	0.6000	0.4199
Marital_cat3	-0.1442	0.0333	-4.3317	1206.0000	0.0000	-0.2094	-0.0789	0.0769
Male2	-0.1777	0.0189	-9.4013	52222.0000	0.0000	-0.2148	-0.1407	0.0049
Religion2	0.1836	0.0698	2.6307	43112.0000	0.0085	0.0468	0.3203	0.0073
Religion3	-0.2033	0.0710	-2.8644	39162.0000	0.0042	-0.3425	-0.0642	0.0083
CDagree	0.0357	0.0078	4.5744	9994.0000	0.0000	0.0204	0.0510	0.0244

Table 13: Pooled logistic regression model with up-sampling and `zone` as covariable

	obs	Positive
1	Positive	0.3818
2	Positive	0.3760
3	Positive	0.3687
4	Positive	0.3603
5	Negative	0.3498
6	Positive	0.3495
7	Positive	0.3487
8	Positive	0.3466
9	Positive	0.3454
10	Positive	0.3451
11	Positive	0.3441
12	Positive	0.3430
13	Positive	0.3411
14	Negative	0.3292
15	Positive	0.3288
16	Positive	0.3278
17	Negative	0.3260
18	Negative	0.3228
19	Negative	0.3214
20	Positive	0.3194
21	Positive	0.3163
22	Positive	0.3161
23	Positive	0.3140
24	Negative	0.3138
25	Negative	0.3132
26	Negative	0.3117
27	Negative	0.3115
28	Negative	0.3098
29	Negative	0.3094
30	Negative	0.3089

Table 14: The cases in testing with highest probability for positive HIV test and their real result and case number for logistic regression with covariable `zone`

	est	se	t	df	Pr(> |t|)	lo 95	hi 95	fmi
(Intercept)	-3.4632	0.3553	-9.7475	32606.0000	0.0000	-4.1595	-2.7668	0.0018
wealthq2	0.0045	0.1035	0.0438	32920.0000	0.9651	-0.1983	0.2073	0.0011
wealthq3	0.2194	0.1070	2.0516	32855.0000	0.0402	0.0098	0.4291	0.0012
wealthq4	0.2690	0.1168	2.3033	32783.0000	0.0213	0.0401	0.4979	0.0014
wealthq5	0.3424	0.1285	2.6646	32763.0000	0.0077	0.0905	0.5943	0.0015
zone2	0.2205	0.0990	2.2261	33100.0000	0.0260	0.0263	0.4146	0.0002
zone3	0.0573	0.1105	0.5189	33083.0000	0.6038	-0.1592	0.2738	0.0004
zone4	-1.0359	0.1293	-8.0103	33098.0000	0.0000	-1.2894	-0.7824	0.0003
zone5	-0.3882	0.0979	-3.9638	33087.0000	0.0001	-0.5802	-0.1962	0.0004
zone6	-0.6090	0.1069	-5.6953	33105.0000	0.0000	-0.8186	-0.3994	0.0001
location2	0.1396	0.0813	1.7183	33051.0000	0.0858	-0.0196	0.2989	0.0006
Sexgift2	0.1411	0.1137	1.2404	30625.0000	0.2149	-0.0819	0.3640	0.0041
MultSex2	0.1843	0.0842	2.1883	31676.0000	0.0287	0.0192	0.3494	0.0031
Sex12m2	-0.0575	0.1060	-0.5425	13996.0000	0.5875	-0.2652	0.1502	0.0170
NonmarSex12	0.0780	0.1139	0.6844	29950.0000	0.4937	-0.1454	0.3013	0.0047
CDHeard2	0.2562	0.1193	2.1483	31447.0000	0.0317	0.0225	0.4900	0.0033
AgeSexcat2	0.0621	0.1182	0.5253	30982.0000	0.5994	-0.1696	0.2938	0.0038
AgeSexcat3	-0.3473	0.1809	-1.9194	31970.0000	0.0549	-0.7020	0.0074	0.0027
AgeSexcat4	0.1870	0.1489	1.2565	12083.0000	0.2090	-0.1047	0.4788	0.0192
educ_cat2	0.0337	0.1544	0.2180	33083.0000	0.8274	-0.2690	0.3363	0.0004
educ_cat3	0.3684	0.1072	3.4374	32578.0000	0.0006	0.1584	0.5785	0.0018
educ_cat4	0.2605	0.1110	2.3470	32412.0000	0.0189	0.0430	0.4781	0.0021
educ_cat5	0.0647	0.1400	0.4618	32766.0000	0.6442	-0.2098	0.3391	0.0014
ExpSTIs2	1.1015	0.2939	3.7472	32975.0000	0.0002	0.5253	1.6776	0.0009
RespAge	-0.0023	0.0036	-0.6377	32740.0000	0.5236	-0.0094	0.0048	0.0015
HeardHIV2	-0.0583	0.1384	-0.4208	27731.0000	0.6739	-0.3296	0.2131	0.0064
CompknoHIV2	-0.1225	0.0711	-1.7243	25253.0000	0.0847	-0.2618	0.0168	0.0081
Yearstud2	-0.1657	0.0684	-2.4228	33105.0000	0.0154	-0.2997	-0.0316	0.0001
Marital_cat2	0.4699	0.1431	3.2840	3471.0000	0.0010	0.1893	0.7504	0.0429
Marital_cat3	-0.2084	0.1207	-1.7262	29985.0000	0.0843	-0.4449	0.0282	0.0047
Male2	-0.2006	0.0711	-2.8226	33068.0000	0.0048	-0.3400	-0.0613	0.0005
Religion2	0.2039	0.2683	0.7602	33105.0000	0.4472	-0.3219	0.7298	0.0001
Religion3	-0.2185	0.2732	-0.8000	33105.0000	0.4237	-0.7540	0.3169	0.0001
CDagree	0.0392	0.0289	1.3551	32786.0000	0.1754	-0.0175	0.0958	0.0014

Table 15: Pooled logistic regression model with `zone` as covariable

		ROC Curve of Imputation Set							
	Sampling	1	2	3	4	5	6	7	8
		0.6895	0.6904	0.6881	0.689	0.6879	0.6876	0.6889	0.6879
`State`	SMOTE	0.6789	0.6796	0.6792	0.6792	0.6801	0.6797	0.6794	0.6792
	down	0.6882	0.6881	0.6882	0.6861	0.6858	0.6862	0.6866	0.688

Table 16: Area under the curve of all eight roc curves for boosted trees

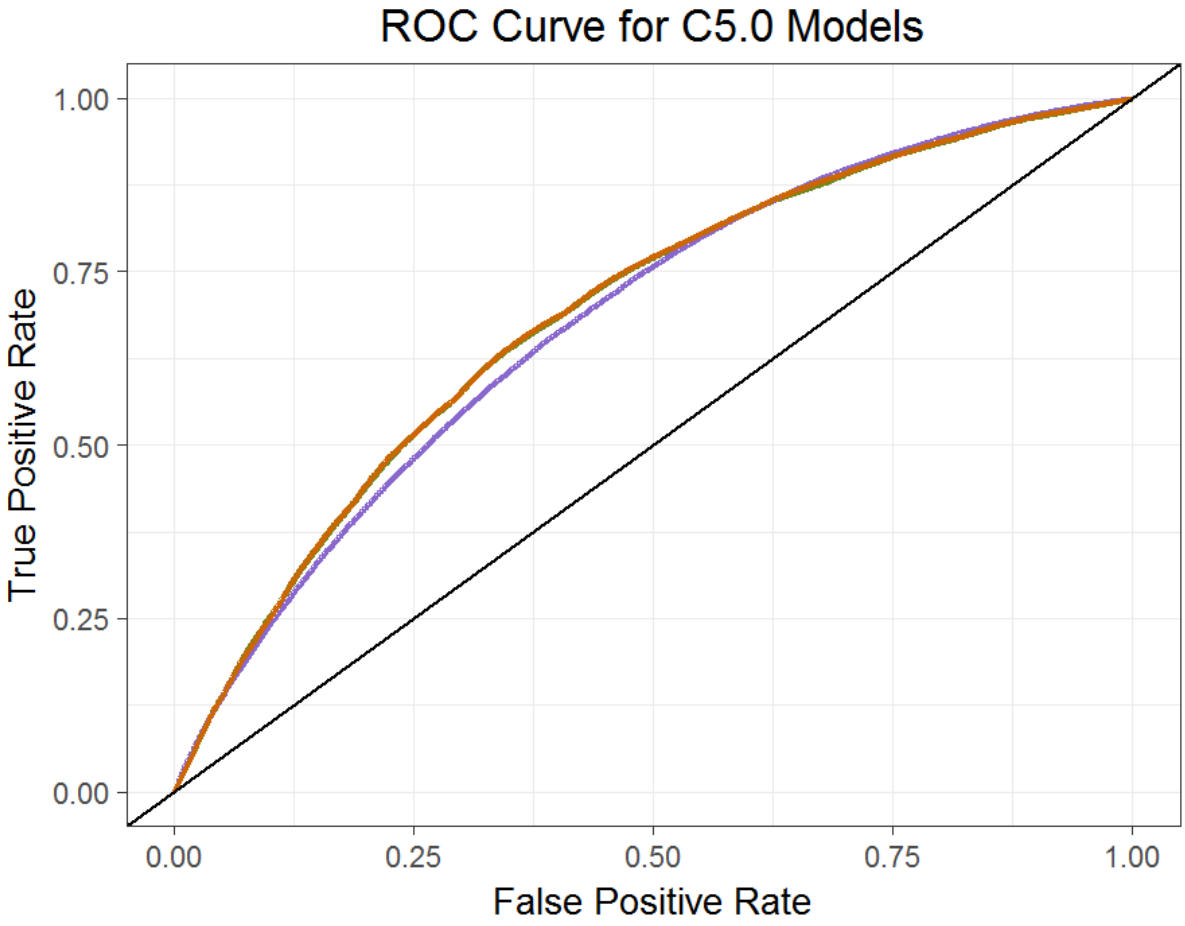

Figure 43: ROC curves for Boosted Tree Model

		ROC Curve of Imputation Set							
	Sampling	1	2	3	4	5	6	7	8
`State`	SMOTE	0.6457	0.6467	0.646	0.6465	0.6466	0.6469	0.6462	0.6453
	down	0.6748	0.6751	0.6746	0.675	0.6744	0.6748	0.6745	0.6744
`zone`	down	0.6057	0.6056	0.6053	0.6051	0.6062	0.6056	0.6054	0.6056

Table 17: Area under the curve of all eight roc curves for naive bayes with covariable `State`

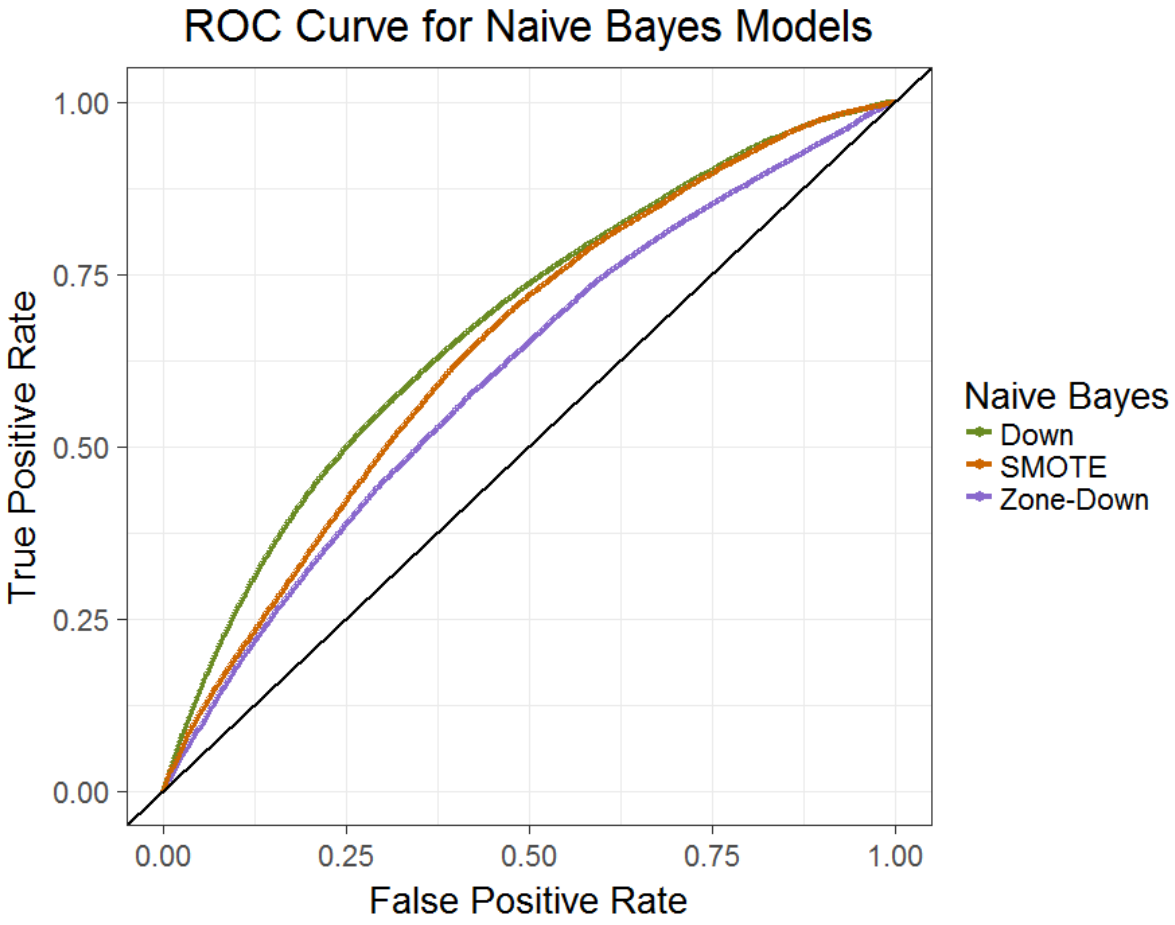

Figure 44: ROC curves for naive Bayes

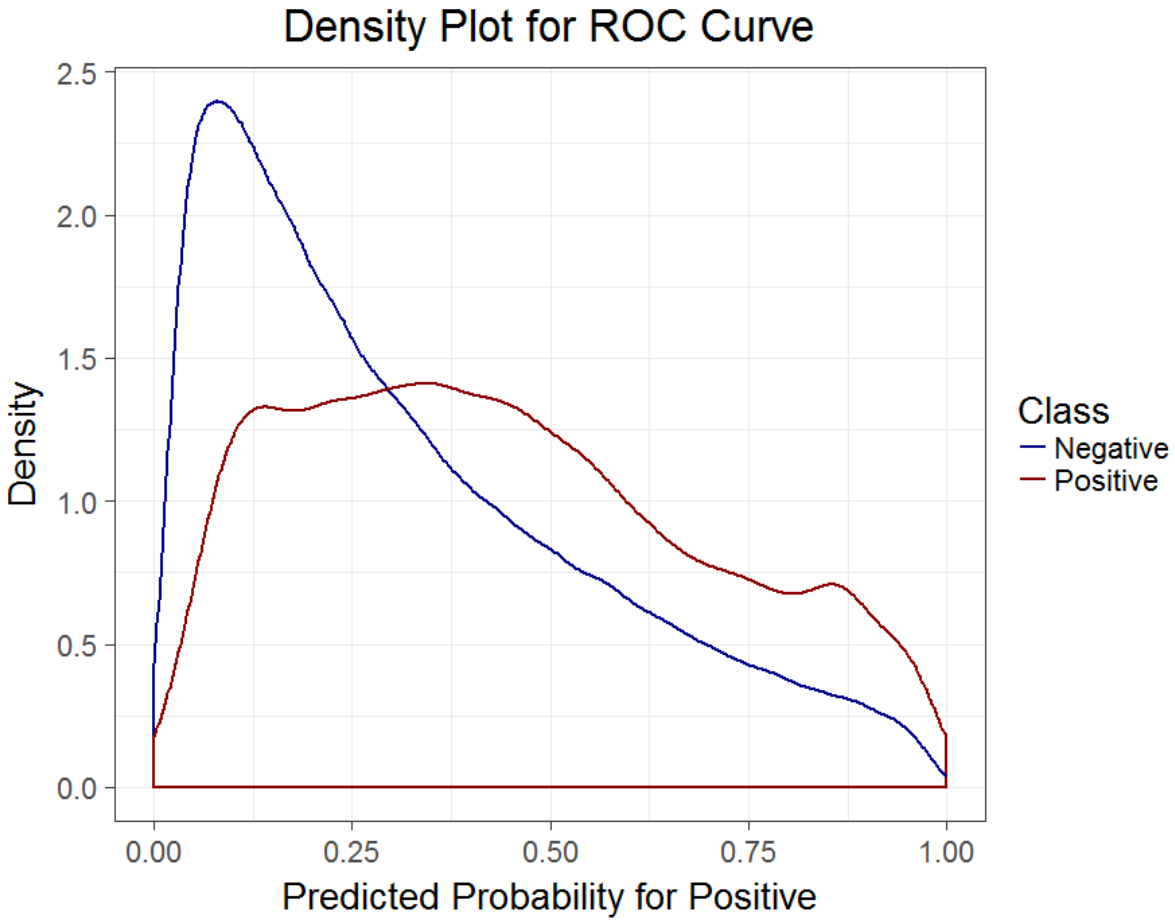

Figure 45: Density plot for naive Bayes with SMOTE and co-variable `State`

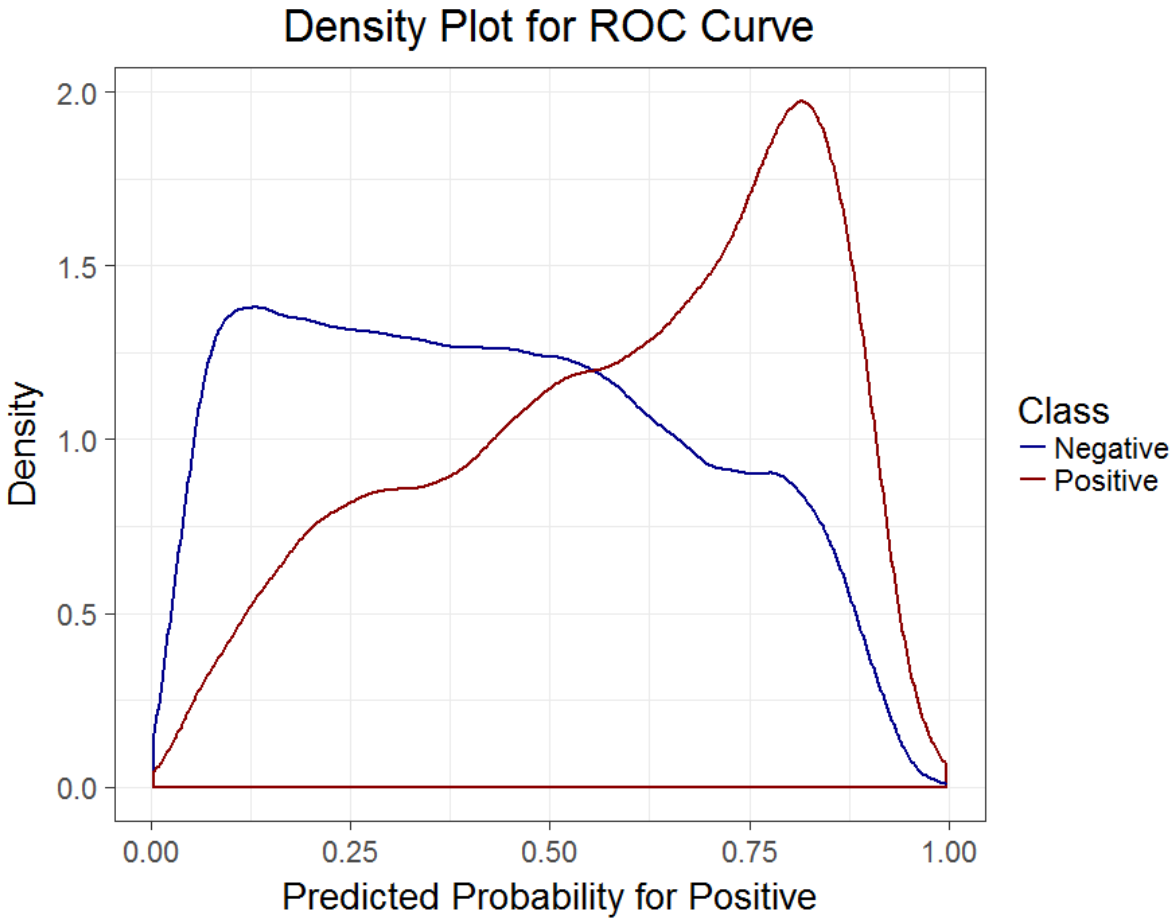

Figure 46: Density plot for naive Bayes with down-sampling and co-variable `State`

E R-Code

The 'R' code used for the calculations can be found on a CD attached to the last page.

Declaration

Herewith I declare that I completed this work on my own and that all information that has been directly or indirectly taken from other sources has been noted as such. Neither this, nor a similar work, has been published or presented to an examination committee.

Munich, July 12, 2018

Benedikt Baus