Comparing Non-Response Adjustment Methods in the Panel on Household Finances

Julian Sengewald* Panagiota Tzamourani†

Abstract

This paper computes different non-response adjustment methods in the Panel on Household Finances and compares the resulting weights with respect to their effect on reducing non-response bias of key survey estimates. The Panel on Household Finances is a household survey conducted by the Deutsche Bundesbank and it forms the German component of the euro area Household Finance and Consumption Survey. The non-response adjustment methods are based on logistic response propensity models and a random forest algorithm. The estimated response propensities are used either directly to form non-response adjustments or are stratified to form response homogeneity groups. Within the latter, different non-response adjustment factors are used. The choice of the auxiliary variables used in the estimation of response propensities and the use of weights are further discussed. The study focuses on the non-response adjustment for the panel component, for which a rich set of variables from the previous wave is known for both respondents and non-respondents. These are also used to measure the reduction in bias between the resulting non-response adjusted weights.

Key Words: Survey weights, non-response adjustment, logistic models, random forests, bias.

1. Introduction

This paper examines different non-response adjustment methods in the Panel on Household Finances and compares them with respect to their effect on the mean and variance of the resulting weights and the bias reduction on key survey estimates. The Panel on Household Finances is a panel household survey conducted by the Deutsche Bundesbank and it forms the German component of the Household Finance and Consumption Survey, which is conducted across euro area countries. Its first wave took place in 2010/11 and the following two in 2014 and 2017. In both 2014 and 2017, random refreshment samples were added to the sample in order to address panel attrition and to include new important population subgroups, such as migrants. The panel structure mimics that of the Panel Study of Income Dynamics (PSID) (Gouskova et al., 2008). All households are re-contacted, and all individuals are tracked. Original sample members breaking off the original households are followed, and the new households they belong to are added to the panel. [1]

Population surveys are subject to non-response, which can result in bias in the survey estimates. If non-respondents differ from respondents with respect to the surveyed characteristics, the estimates based only on respondents will be biased, that is they will differ from the estimates that would have been obtained under full response. Adjustments for non-response can reduce, or ideally eliminate such bias.

There is an enormous literature on weighting for non-response. A comprehensive review of the rationale and methods can be found, for example, in Lynn (1996), Rizzo et al.

*University of Bamberg, Kapuzinerstrasse 16, 96047 Bamberg, Germany

†Deutsche Bundesbank, Wilhelm-Epstein-Strasse 14, 60431 Frankfurt am Main, Germany. The findings reported in this article represent the authors' views and do not necessarily represent the views of the Deutsche Bundesbank or its staff.

[1] An overview of the survey is given in von Kalckreuth et al. (2012). Further information can be found in `www.bundesbank.de/phf-research`.

(1996), Buskirk and Kolenikov (2015). There are usually three steps in the computation of survey weights: first, design weights are produced to adjust for the different sample selection probabilities. Second, the design weights are adjusted for non-response. Third, the non-response adjusted design weights are calibrated to match population totals, producing the final survey weights. The design of the survey may require additional steps. For example, in panel surveys, the non-response adjustment is carried out separately for the components of the households that have participated in different waves. Multiple frames require separate treatment, as for example in the Survey of Consumer Finances (see Kennickell and Woodburn (1999)).

In the PHF, in both the first and second waves of the survey, logistic response propensity models were used to compute the non-response adjustments. In the second wave of the survey, which is the first wave in which some households, the wave 1 respondents or "panel" households are re-contacted, there are two different sources of non-response. Besides non-response in the refreshment sample, some of the panel households also failed to respond. The non-response adjustment stage is carried out separately for the panel, split[2] and the refreshment sample components. Whereas non-response adjustments are typically applied to design weights, for the panel (and split) households, non-response adjustments are applied to the previous wave final weights. After the non-response adjustment in each sample component, the three components are merged and calibrated together to match official population statistics.[3]

The paper compares different methodologies to produce panel non-response adjustments and discusses the choice of variables involved in these methodologies. It focuses on the comparison between logistic regression modelling and a random forest algorithm. It compares the non-response adjusted weights arising from inverse probability weighting and weighting class estimators, based on the estimated response probabilities from the two methods. Within weighting classes, different adjustments factors are compared: (the inverses of) the observed response rate, the weighted observed response rate and the mean estimated response probability. For both the random forest and the logistic regression models, variables from wave 2 and the previous wave are used. Whereas the random forest algorithm uses all available variables, subsets of variables are used to fit different logistic regression models. We discuss the choice of variables and the use of weights. The resulting weighting adjustments are compared with respect to the bias reduction in the resulting variable estimates. In particular, the weighting adjustments are applied to the responses of the respondents of the second wave for the variables of the first wave, and the difference between these responses and the responses from the whole sample (of the previous wave) are computed.

The results of this study serve to inform of best practice methods at the phase of non-response adjustment, though this is not the final phase in the production of the survey weights. After non-response adjustment in each sample component, the three sample components, panel, split and refresher, are merged and the non-response adjusted weights are calibrated together to known population totals. Whether the choice of different adjustment methods still matters after calibration is beyond the scope of this study and is the subject of further research.

We find that overall the non-response adjustments estimated from the logistic regression model overall performed better in terms of bias reduction for the variables we studied.

[2]Split households are the new households that original sample members breaking off the original sample households belong to.

[3]The methodology of the second wave of the survey is described in Knerr et al. (2015). Specific features of the computation of the weights, such as the weight sharing adjustments for panel households, are further elaborated in Tzamourani (2015).

Moreover, a more parsimonious logistic regression model that includes the survey variables of interest performs better in terms of bias reduction for the particular variables. The use of weights in the estimation of the logistic regression model plays a significant role and reduces bias in the resulting estimates. Furthermore, different adjustment factors need to be applied within each framework. Whereas we find small differences in terms of the resulting bias between the observed response rates and the mean estimated response propensities, for the adjustments based on the logistic models, the observed response rates perform clearly better for the adjustments based on the random forest algorithm. Our results indicate, that the random forest algorithm can be useful in indicating variables related to the non-response process that might be overlooked when building a response propensity model. The choice of the non-response model may also depend on the overall purpose of the non-response adjustment, that is whether it should serve as many variables as possible, or whether some variables are of particular interest and improving their accuracy is of foremost importance. In that case a more parsimonious carefully defined logistic response model would be preferable.

2. Non-response Weighting Methods

Weighting adjustments methods are now usually based on estimated response propensities from a model or an algorithm which predicts the binary unit response outcome using a set of variables available from the sample for both respondents and non-respondents. Response propensity models attempt to "balance" respondents and non-respondents using a single propensity score that is a function of both continuous and categorical response predictors, as well as predictors that are also associated with the survey outcomes of interest (Kalton and Flores-Cervantes (2003), Little and Vartivarian (2005), Brick (2013)).

Logistic or probit models have been the most common method to estimate response propensities (Chen et al., 2015). Alternative methods, such as local polynomial regression (da Silva and Opsomer, 2009) or classification trees (Lee et al., 2010) have also been explored. These new approaches have been shown in simulation studies to be more flexible regarding misspecification of the functional form of the response process. More recently, random forests, a nonparametric ensemble tree-based method are being used for generating non-response adjustments, see for example Buskirk and Kolenikov (2015).

The inverse of the estimated response propensity can be used as a direct non-response adjustment for each respondent. A common problem encountered with this method is that some cases have very small response propensities, which lead to increased variance (Little, 1986). One way to reduce the impact of increased variance would be to employ trimming of the weights, see for example Kalton and Flores-Cervantes (2003).

An alternative method is to form weighting classes, or adjustment cells. The weighting classes can be formed by grouping the estimated response propensities (Little, 1986), or by cross-classifying observed categorical variables, available for respondents and non-respondents. In general they should be defined by factors that are believed to be associated with differences between respondents are non-respondents, so that within each adjustment cell it can be assumed that non-respondents and respondents are comparable. Estimates derived from the respondents' data should then have little bias that could be attributed to differences between respondents and non-respondents. If weighting cells can be defined so that within each cell respondents (households in our case) are missing at random (MAR), then the weighted class estimator of the mean is unbiased (Rubin, 1976; Little and Rubin, 2014).

Within the weighting classes, or adjustment cells, different adjustment factors can be considered (see for example Rizzo et al. (1996)). The adjustment factor, which will typically be applied to the design weight to produce the non-response weight within a weighting class c, is of the form $w_c = (\phi_c)^{-1}$.

Depending on the definition of ϕ_c, we can have the following adjustment factors:

- **Response rate** $\phi_c = \frac{n_{1c}}{n_{0c}+n_{1c}}$
- **Design weighted response rate** $\phi_c = \frac{\sum_{i \in c \cap R} d_i}{\sum_{i \in c} d_i}$
- **Mean propensity score** $\phi_c = \frac{1}{n_{1c}} \sum_{i \in c \cap R} \phi_i$, where c is the weighting class, n_{1c} is the number of respondents in class c, n_{0c} is the number of non-respondents, R is the set of respondents and d_i is the design weight of observation i.

  Assuming a size of the weighting classes equal to 1, then the last case becomes $\phi_c = \phi_i$, the inverse of the propensity score.

In our paper we estimate the response probabilities for the above weighting methods with logistic regression models and with a random forest algorithm. In forming weighting classes, we use the deciles of the estimated response propensities and form ten classes.

3. Logistic Regression Modeling

In this class of models the response propensity is modelled in terms of several variables that are assumed to affect non-response. Auxiliary variables that are related to the survey outcome should also be included in order to reduce bias (Little and Vartivarian, 2003, 2005; Kalton and Flores-Cervantes, 2003; Andy Peytchev and Kristen Olson, 2007; Kreuter et al., 2010).

Since our analysis focuses on the panel component of the sample, it is possible to use in our models variables from the first wave of the survey, as these are available for both respondents and non-respondents. We therefore include in our models, besides variables that are assumed to be related to non-response, the key survey variables, namely household net wealth and income, and variables that are related to these, such as homeownership status, from the first wave of the survey. To test the effect of particular variables we build models sequentially. After preselection of several variables we included the following sets of variables:

1) Sociodemographic factors: person level data, such as age, gender, nationality and education. These refer to the household reference person, that is the person that answered the household questionnaire, as the most knowledgeable for the household's finances. Such characteristics are in general considered to be related to survey non-response. In addition, some of these characteristics, particularly age, nationality, education, are related to the key survey variables, wealth and income.

2) Information on sampling design: this includes stratum and municipality size class. In addition the state ("Bundesland") of the household residence is also included. Besides the possible relation of sampling design variables to the response mechanism, for example, urban cities displaying higher non-response, the stratification variables are by design related to household wealth, a key survey outcome.

3) Dwelling characteristics: these are part of the paradata collected by interviewers for the all households in the drawn sample. We include two variables: (i) rating of the outward appearance of the dwelling (ii) comparison of the dwelling with respect to the other dwellings in the neighbourhood. These measures could be related to non-response, since, for example the condition of the dwelling or the neighbourhood may be related to social inclusion or cohesion, and hence readiness to participate in the survey. These variables should also be related to net wealth, since higher net wealth households would be more likely to live in dwellings of better condition, or better looking neighbourhoods. For a limited number of households paradata was not collected. We therefore included a paradata dummy variable, indicating whether paradada collected or not, so as to be able to include these households in the model. This variable is also related to non-response, as it indicates limited interviewer effort.

4) Fieldwork related data: In the final models we include information about the contact effort, which is measured in days between the first contact with the household until the final survey outcome. We also include a dummy variable indicating whether there was a change in the interviewer for that household within the fieldwork period. These variables are clearly related to non-response, as they indicate a difficulty in contacting the household or even persuading it to take part in the survey.

*5) Economic Situation:*These include income quantiles, net wealth quantiles and the homeownership status, since homeownership is correlated with household net wealth.

Table 1 shows results from selected logistic regression models. Model 1 is our base model and includes all the variables to be also included in later models, namely the household sociodemographic characteristics, the variables on dwelling appearance, the municipality size and the Bundesland indicators. Model 2 is as Model 1, but estimated using the survey weights [4] In Model 3 we also include the strata indicators. Besides being a main design variable, which is typically included in such models, the stratum indicator is related to household wealth, the main survey variable, since stratification in the survey was explicitly made on the basis of variables related to it. Model 5 includes in addition the wealth quantiles. Model 6 we exclude net wealth but include household income and home-ownership status. Besides income being correlated with wealth, as homeownership is an important determinant of net wealth, it partially controls for the latter. To examine the effect of weights using also this richer model, we estimate this model without using survey weights. Model 7 is the same model as Model 6, estimated using weights.

The estimated models indicate a hump shaped pattern of age in association with non-response, with smaller response propensities for the youngest and oldest household groups, compared to the middle category. There is no gender effect on response propensity. Wealthier (non-urban) municipalities display higher response propensities than other (non-urban) municipalities (base category), though wealthy or other street sections (both in urban municipalities) have lower response propensities than the latter (the base category). Better maintained and better rated dwellings are associated with higher response propensity. We further see increasing response propensity with increasing income and wealth, though the coefficients are not significant, as we have already included in the model the other variables also indicative of household income and wealth described above. Homeowners also have a higher response propensity relative to the households which do not own the dwelling they live in. Last, difficulty in contacting again the household (more than 90 days elapsed between first and last contact) and having a change in the interviewer in the middle of fieldwork, are associated with lower response propensity. The first may capture respondents who are evasive as well as ones who are very busy. Having an interviewer change during the field period may reflect the attempt by the survey managers to find an interviewer would have a better "fit" with a difficult respondent.

The (adjusted) R squared generally increases as we progressively add variables. Adding wealth in model 4 increases R squared, though it remains the same if we replace that with income and homeownership status. R squared increases by a larger degree going from model 5 to 7, compared to the changes between the previous models, when we add the last

[4]Since we are using the panel component of the sample, we use as survey weights the final weights of the previous wave, wave 1.

two variables relating to the fieldwork process, indicating the importance of these variables in relation to non-response.

Table 1: Logistic response propensity models

	(1) (Without weights)	(2)	(3)	(4)	(5)	(6) (Without weights)	(7)
Age							
(Base = 36-56)							
19-35	0.67**	0.58**	0.58**	0.63**	0.68*	0.80	0.69*
	(0.08)	(0.10)	(0.10)	(0.11)	(0.12)	(0.11)	(0.13)
57-63	1.32**	1.23	1.23	1.20	1.17	1.12	1.05
	(0.14)	(0.18)	(0.18)	(0.18)	(0.17)	(0.12)	(0.16)
63+	0.95	0.76	0.77	0.75	0.75	0.70**	0.60**
	(0.11)	(0.12)	(0.12)	(0.12)	(0.13)	(0.09)	(0.10)
Education							
(Base = Primary)							
Secondary	1.48**	1.40	1.41	1.36	1.28	1.37*	1.24
	(0.20)	(0.25)	(0.26)	(0.25)	(0.23)	(0.20)	(0.23)
Tertiary	2.23***	2.21***	2.23***	2.02***	1.89**	2.14***	1.87**
	(0.33)	(0.45)	(0.46)	(0.42)	(0.40)	(0.35)	(0.41)
Gender							
Female	0.92	0.95	0.95	0.96	0.97	0.90	0.94
	(0.07)	(0.11)	(0.11)	(0.11)	(0.11)	(0.08)	(0.12)
Dwelling appearance							
(Base = Moderate/A few small cracks)							
Well maintained	1.25*	1.22	1.24	1.22	1.21	1.10	1.09
	(0.13)	(0.18)	(0.18)	(0.18)	(0.18)	(0.12)	(0.17)
Badly maintained	0.64*	0.66	0.66	0.66	0.67	0.68	0.71
	(0.13)	(0.18)	(0.18)	(0.18)	(0.18)	(0.15)	(0.20)
Dwelling rating							
(Base = Satisf.)							
Exclusive/very good	1.25*	1.29	1.29	1.23	1.22	1.21	1.28
	(0.12)	(0.19)	(0.19)	(0.18)	(0.18)	(0.13)	(0.19)
Modestv.modest	1.34	1.44	1.46	1.47	1.47*	1.27	1.33
	(0.20)	(0.28)	(0.29)	(0.29)	(0.28)	(0.20)	(0.27)
Paradata avail.	0.13***	0.18**	0.19**	0.18**	0.17**	0.13***	0.18**
	(0.05)	(0.11)	(0.11)	(0.11)	(0.10)	(0.05)	(0.11)
Municipality size							
(Base = <5,000)							
5,000 - 20,000	0.87	0.95	0.97	1.00	1.06	0.87	0.88
	(0.12)	(0.20)	(0.21)	(0.21)	(0.22)	(0.13)	(0.19)
20,000 - 100,000	0.75*	0.73	0.75	0.79	0.85	0.83	0.78
	(0.11)	(0.16)	(0.16)	(0.17)	(0.18)	(0.13)	(0.17)
100,000 and above	0.71*	0.77	1.25	1.30	1.38	1.29	1.17
	(0.10)	(0.16)	(0.57)	(0.57)	(0.60)	(0.39)	(0.54)
Strata							
(Base = other municipality)							
Rich municipality			1.13	1.12	1.14	1.28*	1.43*
			(0.15)	(0.15)	(0.15)	(0.15)	(0.20)
Rich street section			0.57	0.54	0.57	0.73	0.74
			(0.25)	(0.23)	(0.24)	(0.22)	(0.34)
Other street section			0.62	0.65	0.73	0.65	0.89
			(0.25)	(0.26)	(0.29)	(0.19)	(0.39)
Wealth Quant.							
(Base = Up to 20%)							
20% - 40%				1.00			
				(0.20)			
40% -60%				1.19			

continued on next page

Table 1 – continued from previous page

	(1) (Without weights)	(2)	(3)	(4)	(5)	(6) (Without weights)	(7)
				(0.24)			
60% -80%				1.37			
				(0.29)			
80% -100%				1.47			
				(0.31)			
Income Quant.							
(Base = Up to 20%)							
20% -40%					1.17	1.15	1.23
					(0.23)	(0.19)	(0.25)
40% -60%					1.31	1.37^{*}	1.37
					(0.27)	(0.22)	(0.29)
60% -80%					1.36	1.37^{*}	1.46
					(0.28)	(0.22)	(0.31)
80% -100%					1.23	1.26	1.23
					(0.26)	(0.21)	(0.27)
Homeowner					1.73^{***}	1.29^{*}	1.70^{***}
					(0.24)	(0.13)	(0.24)
Intervw. Change						0.57^{***}	0.57^{***}
						(0.05)	(0.07)
Days first to last contact (>90)						0.20^{***}	0.21^{***}
						(0.02)	(0.03)
Bundesland	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	3207	3207	3207	3207	3207	3207	3207
McFadden's R2	0.07	0.07	0.08	0.08	0.09	0.15	0.15

Exponentiated coefficients; Standard errors in brackets

4. Random Forest

Random forests are a combination of classification trees and as such aim to construct prediction models from data. The classification algorithm is based by recursively splitting the predictor variables in nonoverlapping parts. The partitioning of the predictors variables, at each node, is chosen so that a measure of "impurity" is minimized, which is based on the distribution of the observed values of the class variable at the node. The random forest algorithm grows an ensemble of such trees and lets each tree cast a vote for the most popular class. The aggregated votes form the predictor. In order to grow these ensembles, often random vectors are generated that govern the growth of each tree in the ensemble (Breiman, 2001).

Random Forest expands the algorithm of classification trees in that it tries to reduce variance and bias. Variance is reduced by averaging several trees grown on bootstrapped training data. As one usually has only one training dataset, more training data can be obtained by resampling from the original training data set Breiman (1996). If resampling is done by bootstrap then this is known as bagging and was originally proposed in (Breiman, 1996). Furthermore within each tree at every splitting step only a randomly chosen set of predictors is considered (Friedman et al., 2009). The latter leads to a further reduction in variance of the random forest prediction, through decorrelation of trees, otherwise all trees would resemble each other as predictors might stay the same on every bootstrapped sample.[5] Using this additional step of decorrelation and averaging gives the random forest

[5]To see this, note that the variance of the average of the predictions, e.g. obtained by bagging, is given by $\rho\sigma^2 + \frac{1-\rho}{\#trees}\sigma^2$, where $\#trees$ is the size of the ensemble of trees and ρ is the average correlation between the trees (Friedman et al., 2009).

estimator. A classification prediction for an observation is then produced by the majority vote of the predictions of all trees together.

Tree based methods and in particular random forests have been adapted several times in the context of non-response adjustment, see Schouten and de Nooij (2005); Lee et al. (2010); Phipps and Toth (2012); Earp et al. (2014); Buskirk and Kolenikov (2015); Lohr et al. (2015). Random forest also seems to be generally superior in terms of accuracy when compared to the performance of several other classifiers on a range of different datasets in a benchmarking study by Fernández-Delgado et al. (2014).

In our study we will focus on the comparative effect of non-response weights stemming from a random forest algorithm to bias reduction.

Implementation aspects We used the package `party` in R for the implementation of our random forest algorithm. This algorithm has the feature `cforest`, which is an implementation of the random forest and bagging ensemble algorithms utilizing conditional inference trees as base learners and differs from the reference implementation in `randomForest` with respect to the base learners used and the aggregation scheme applied.[6] A particular feature of this algorithm is that splitting of nodes is performed on the basis of permutation testing, which is more appropriate when comparing variables of different scales (Torsten Hothorn et al., 2006). This algorithm further uses subsampling instead of bootstrapping, that is, a smaller set of observations than the full set is sampled. This smaller set is the out-of-bag observations. Subsampling without replacement should allow for unbiased variable selection (Strobl et al., 2007).

As the algorithm employs an automated variable selection scheme of all the variables in the file, it is important to ensure that the file contains "eligible" variables to enter the algorithm and so the variables that do not fulfill certain conditions must be removed. "Eligible" are variables, for example, that have been observed for both respondents and no respondents or variables that do not contain a large amount of missing data. Survey outcome variables need to be excluded, as the latter perfectly discriminate between respondents and non-respondents, but do not inform of the formation of non-response. Furthermore, observation identifiers and variables containing only text should also be excluded. We have kept flag variables[7], as these are indicators of item non-response. The final dataset included 216 variables.

Results The random forest has increased predictive power, compared to a single tree. However, since the output is an ensemble of trees, it is more difficult to interpret it. The "variable importance" measures, produce by the algorithm, can help to understand the results. We investigate two measures of variable importance. The first one is the permutation variable importance. Here the values of a variable are randomly permuted. Then the difference in accuracy before and after permutation is calculated. This step is repeated over the whole ensemble of trees and then averaged. This gives the permutation variable importance measure. Missing data are handled by a method proposed by Hapfelmeier et al. (2014), with which instead of permuting the predictor variable, cases are just randomly assigned to the left or to the right child node.

For unbalanced data, an alternative variance importance measure, "VarimpAUC", using the Area Under the Curve as an error function to compute accuracy (Janitza et al., 2013) can perform better in determining important variables in the classification procedure.

The first ten important variables with the two importance measures are listed in Table 2. Some of these variables have already been included in our logistic response propen-

[6] See: https://cran.r-project.org/web/packages/party/party.pdf.

[7] Each variable in the survey has a corresponding "flag variable", which indicates the outcome of the data collection process.

Table 2: Variable importance measures, Varimp and AUC, for the highest scoring predictors on Varimp

Variable	**Varimp**	**VarimpAUC**
Days between first and last contact	0.108	0.150
Education	0.003	0.014
Rating of dwelling	0.003	0.007
Being suspicious after interview	0.003	0.003
Ease in responding	0.003	0.007
Net wealth (quantiles)	0.002	0.001
Rating of residential area	0.002	0.003
Non visible/no security measure	0.001	0.001
Age of sampled person	0.001	0.005
Homeowner	0.001	0.002

Exponentiated coefficients; Standard errors in brackets

sity models, whereas the algorithm indicated some further variables, mainly paradata, as important in discriminating respondents from non-respondents. In the first set are the net wealth variable, homeownership status, education, days between first and last contact, as also the rating (by the interviewers) of the dwelling, which can be thought of similar to the rating of the area, and the age of the reference person, which often coincides to the age of the sampled person (age of person selected from the population register). The algorithm further indicated as important the variable reporting whether the respondent was suspicious after the interview, whether they had ease in responding the questions, and the indication of no security measure in the dwelling (the variable records they types of security measures, if available). These three variables belong to the paradata collected by the interviewer.

To better compare the random forest with the logistic regression model, we fitted another logistic response propensity model using as covariates the first ten variables that came out as most important from our random forest implementation. The model is given in Table 3. Since no weights were used in the random forest, we also estimated this model without the use of weights.

Table 3: Logistic response model using the most important variables from the random forest algorithm

Predictor	b/(se)
Days untill conversion	0.99^{***}
	(0.00)
Education	
(Base = Primary)	
Secondary	1.26
	(0.18)
Tertiary	1.78^{***}
	(0.28)

Predictor	b/(se)
Dwelling Rating	
(Base = Satisf.)	
Excl./very good	1.06
	(0.11)
(Very) Simple	1.46*
	(0.22)
Suspicious after interview	
(Base = Absolutely not))	
A little	0.65***
	(0.08)
Very much	0.32*
	(0.18)
Easiness in responding	
(Base = Fairly difficult & Difficult)	
Very easy	1.74**
	(0.32)
Fairly easy	1.68**
	(0.30)
Easy	1.03
	(0.20)
Income Quant.	
(Base = Up to 20%)	
20% - 40%	0.71*
	(0.11)
40% - 60%	1.04
	(0.17)
60% - 80%	0.96
	(0.17)
80% - 100%	0.81
	(0.15)
Rating of residential area	
(Base = Unsatisfactory & Poor)	
Very Good	2.85**
	(1.07)
Good	1.87
	(0.66)
Satisfactory	1.30
	(0.45)
Adequate	0.75
	(0.27)
Non-visible security measure	1.03
	(0.09)
Age	
(Base = Ages 40-49)	
18-29 years	0.71
	(0.14)
30-39 years	0.86
	(0.14)
50-65 years	1.08
	(0.14)
66+ years	0.78
	(0.10)
Noparadata	0.08***
	(0.04)
No data for easiness/suspicious	16.76***
	(9.85)
Homeowner	1.34*
	(0.18)

N	3207
McFadden's R2	0.14

Exponentiated coefficients; Standard errors in brackets
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

5. Evaluation

Our evaluation of the non-response adjustments resulting from the procedures described above is mainly focused on the reduction of bias achieved using these non-response adjustments, relative to unadjusted estimates.

Since we are comparing methods for a panel sample, we can use the data available from the previous wave to measure bias. In particular, we define the relative bias (RB) of an estimate, as the difference between the estimate obtained from both respondents and non-respondents from the first wave, minus the estimate obtained from the respondents of the second wave, relative to the former.

So, the relative bias is given by

$$RB = \frac{1/n_1 \sum_{k \in r_1} w_{wave1,k} y_{wave1,k} - 1/n_2 \sum_{k \in r_2} w_{wave2,k} y_{wave1,k}}{1/n_1 \sum_{k \in r_1} w_{wave1,k} y_{wave1,k}},$$

where y is the variable of interest, r_1 is the full set of respondents in wave 1, w_{wave1_k} are the final weights of wave 1, r_2 the respondents in wave 2 and $w_{wave2,k}$ are the non-response adjusted weights of wave 2, which are produced by applying the non-response adjustments described above to the final weights of wave 1.

Using the above measure, we will attempt to assess and disentangle the effects of the following on the final non-response weights: the different adjustment factors, the choice of variables in the different logistic regression models, the use of weights in estimating a logistic response propensity model, the use of such a model with a restricted, but carefully chosen set of covariates vs. a random forest algorithm.

5.1 Adjustment factors

Tables 2, 4 and 6 show the means, relative biases and the standard errors of our key survey variables (all obtained from the respondents of wave 2) after applying different adjustment factors, based on the response propensities estimated with either logistic regression models or the random forest algorithm. The column "unadjusted" indicates that the estimates were obtained using the survey weights of wave 1, without any non-response adjustments. Columns "RHG:rr", "RHG:dw" and "RGH:mp", indicate that the response adjustments were obtained by forming Response Homogeneity Groups (RHG), and then applying the response rate, the weighted response rate and the mean estimated response propensity as adjustment factors. The last column, "ipw", indicates that the response adjustment is the inverse of the estimated response propensity. Table 2 shows these quantities, where the response propensities are estimated with the logistic regression model 1. Overall, we see a marked reduction in relative bias, for all except one variable, for the non-response adjusted estimates, compared to the unadjusted estimate. The variation in relative bias between the different non-response adjusted methods is very small. Table 4 shows the same quantities, but now the estimated response propensity is derived from model 7, which is our richest model, and weights were used in its estimation. Again there is little variation between the different adjustment methods though the adjustments based on the propensity score, seem to perform slightly better in terms of relative bias than the adjustments based on the observed response rates (the response rate or the design weighted response rate).

Table 6 shows the same quantities coming from the random forest algorithm, that is, the formation of the response homogeneity classes was formed on the basis of the estimated response propensity from the random forest, and this is also used in “RHG:mp”, “ipw”. Here we observe the non-response adjustment based on the observed response rates perform much better than the response adjustment using the mean of the estimated response propensity within a class, or using directly the estimated response propensity of a household.

5.2 Choice of variables in the logistic response propensity model and the use of weights

Now, to compare the effect of the different models, i.e. the inclusion of the various auxiliary variables in the non-response models, and also the effect of using weights in the estimation of these models, we compare for all models the relative bias and the standard error of the estimates of our key variables, using a single adjustment method at a time.

Table 7 shows the estimate, relative bias and the standard error of our key survey variables from all our estimated logistic models, using the average response propensity within a group as the weighting adjustment and Table 8 shows the same quantities based the (individual) response propensity adjustment.

Use of weights. In all models weights were used for their estimation, except in models 1 and 6. Model 1 and model 6 are the same as models 2 and 7, respectively, except that weights were not used in the estimation of the former. We see that the relative bias of most variables, after being adjusted with a weighting adjustment coming from a model where weights are used, is smaller than the relative bias of the variable adjusted with a weighting adjustment coming from a model in which no weights are used. So, the use of weights in the logistic response propensity models, seems to produce non-response weights that lead to more accurate survey estimates.

Choice of predictor variables. Models 2,3,4,5,and 7 all use weights but differ in terms of the included variables. The “richer” models 4,5, and 7, display for most variables smaller biases than models 2 and 3.

We see a notable reduction in the relative bias of the net wealth estimate when we include net wealth as an auxiliary variable in Model 4. Adding income and homeownership status instead of net wealth, resulted in the same size relative reduction of bias for the net wealth variable. Adding the two fieldwork variables in the model, resulted in a small increase in the bias of the net wealth and the total assets variable, though the bias of the other variables remained the same.

Comparing model 6 with the model using the “important” variables from the random forest, which are both estimated without the use of weights, we note that the latter performs better for most variables, in particular for net wealth, compared to the relatively richer model 6 (which includes income and homeownership status but not net wealth). This indicates that the random forest can be a useful tool in determining the variables to be included in a logistic response propensity model, perhaps in addition to other variables that the researcher thinks important.

5.3 Logistic regression models vs. random forest

The estimates based on the random forest adjustments are worse in terms of relative bias, though the adjustment factor, individual or mean response propensity, is not yielding the

Table 4: Mean, relative bias and standard error of selected PHF variables, using non-response adjusted weights based on the logistic response propensity model 1

	Full	**Unadjusted**	**RHG.rr**	**RHG.dw**	**RHG.mp**	**ipw**
Net wealth	198,856	236,178	213,691	213,444	213,190	212,274
RB	.	0.19	0.07	0.07	0.07	0.07
SE	15,660	23,060	20,216	20,104	20,171	20,588
Real assets	219,354	252,127	234,610	234,267	234,400	235,719
RB	.	0.15	0.07	0.07	0.06	0.07
SE	18,565	26,640	23,915	23,760	23,917	23,954
Fin. assets	48,312	56,500	51,546	51,449	51,499	50,901
RB	.	0.17	0.07	0.06	0.06	0.05
SE	2,640	3,375	2,975	2,959	2,983	3,139
Total debt	53,824	60,448	56,137	55,969	56,007	56,665
RB	.	0.12	0.04	0.04	0.04	0.05
SE	3,160	4,110	4,011	3,985	3,997	3,984
HomeOwnRate	0.44	0.50	0.48	0.48	0.47	0.48
RB	.	0.14	0.08	0.08	0.07	0.09
SE	0.015	0.017	0.017	0.017	0.017	0.021
Value of home	206,992	205,865	201,658	201,623	201,473	199,701
RB	.	-0.01	-0.03	-0.03	-0.03	-0.04
SE	7,975	8,182	7,825	7,787	7,800	7,810
Income	43,747	47,387	44,908	44,869	44,809	44,626
RB	.	0.08	0.03	0.03	0.02	0.02
SE	1,174	1,437	1,334	1,335	1,328	1,522

Full: Estimation based on full sample, that is respondents and non-respondents of wave 1
Unadjusted: No non-response adjustment was used in the estimation (weighting only with final weights of wave 1)
RHG.rr: Weighting using the inverse of response rate within Response Homogeneity Groups
RHG.dw: Weighting using the inverse of the design weighted response rate within Response Homogeneity Groups
RHG.mp: Weighting using the inverse average estimated response propensity within Response Homogeneity Groups
ipw: Weighting using the inverse response propensity

Table 5: Mean, relative bias and standard error of selected PHF variables, using non-response adjusted weights based on the logistic response propensity model 7

	Full	**Unadjusted**	**RHG.rr**	**RHG.dw**	**RHG.mp**	**ipw**
Net wealth	198,856	236,178	208,104	205,605	205,213	202,517
RB	.	0.19	0.05	0.03	0.03	0.02
SE	15,660	23,060	19,584	19,320	19,252	19,996
Real assets	219,354	252,127	228,995	227,202	226,633	228,569
RB	.	0.15	0.04	0.04	0.03	0.04
SE	18,565	26,640	23,277	23,061	23,001	23,966
Fin. assets	48,312	56,500	50,939	50,298	50,280	49,369
RB	.	0.17	0.05	0.04	0.04	0.02
SE	2,640	3,375	2,997	2,971	2,943	3,091
Total debt	53,824	60,448	53,758	53,000	53,093	53,312
RB	.	0.12	-0.00	-0.02	-0.01	-0.01
SE	3,160	4,110	4,101	4,116	4,121	4,286
HomeOwnRate	0.44	0.50	0.45	0.44	0.44	0.43
RB	.	0.14	0.02	0.00	0.01	-0.02
SE	0.015	0.017	0.018	0.018	0.018	0.020
Value of home	206,992	205,865	203,189	203,192	202,805	202,863
RB	.	-0.01	-0.02	-0.02	-0.02	-0.02
SE	7,975	8,182	7,893	7,851	7,852	7,949
Income	43,747	47,387	44,750	44,525	44,501	43,808
RB	.	0.08	0.02	0.02	0.02	0.00
SE	1,174	1,437	1,360	1,363	1,362	1,554

Full: Estimation based on full sample, that is respondents and non-respondents of wave 1
Unadjusted: No non-response adjustment was used in the estimation (weighting only with final weights of wave 1)
RHG.rr: Weighting using the inverse of response rate within Response Homogeneity Groups
RHG.dw: Weighting using the inverse of the design weighted response rate within Response Homogeneity Groups
RHG.mp: Weighting using the inverse average estimated response propensity within Response Homogeneity Groups
ipw: Weighting using the inverse response propensity

Table 6: Mean, relative bias and standard error of selected PHF variables, using non-response adjusted weights based on a conditional random forest algorithm

	Full	Unadjusted	RHG.rr	RHG.dw	RHG.mp	ipw
Net wealth	198,856	236,178	207,994	206,887	214,152	213,669
RB	.	0.19	0.05	0.04	0.07	0.07
SE	15,660	23,060	19,582	19,197	20,551	20,428
Real assets	219,354	252,127	229,734	229,347	233,006	232,196
RB	.	0.15	0.05	0.05	0.06	0.06
SE	18,565	26,640	23,189	22,798	24,076	23,897
Fin. assets	48,312	56,500	49,897	49,820	51,606	51,642
RB	.	0.17	0.03	0.03	0.07	0.07
SE	2,640	3,375	2,995	3,033	3,067	3,064
Total debt	53,824	60,448	55,442	55,611	56,511	56,325
RB	.	0.12	0.03	0.03	0.05	0.05
SE	3,160	4,110	4,383	4,498	4,221	4,148
HomeOwnRate	0.44	0.50	0.46	0.46	0.48	0.48
RB	.	0.14	0.05	0.05	0.07	0.08
SE	0.015	0.017	0.019	0.019	0.018	0.018
Value of home	206,992	205,865	200,442	200,761	201,286	201,110
RB	.	-0.01	-0.03	-0.03	-0.03	-0.03
SE	7,975	8,182	8,359	8,385	8,303	8,301
Income	43,747	47,387	44,428	44,349	45,373	45,323
RB	.	0.08	0.02	0.01	0.04	0.04
SE	1,174	1,437	1,355	1,372	1,362	1,358

Full: Estimation based on full sample, that is respondents and non-respondents of wave 1
Unadjusted: No non-response adjustment was used in the estimation (weighting only with final weights of wave 1)
RHG.rr: Weighting using the inverse of response rate within Response Homogeneity Groups
RHG.dw: Weighting using the inverse of the design weighted response rate within Response Homogeneity Groups
RHG.mp: Weighting using the inverse average estimated response propensity within Response Homogeneity Groups
ipw: Weighting using the inverse response propensity

best non-response adjusted weights derived from the random forest algorithm. Using the observed or design weighted response rate yields smaller biases (Table 6), though still larger than the ones derived from the logistic regression models 4,5, and 7, that use net wealth or income, which relate to the variables under study (and were estimated using survey weights). Using a logistic response model with the most "important" variables from the random forest algorithm (results reported in last column in Tables 7 and 8), results in non-response adjustments that yield smaller bias than the random forest that uses all available variables. It seems that a more parsimonious model, may result in smaller bias than using all available variables in the prediction.

Table 7: Mean, relative bias and standard error of selected PHF variables, using non-response adjusted weights based on ten weighting classes and using the inverse of the mean propensity as the adjustment factor

	Full	**LM 1** (Without weights)	**LM 2**	**LM 3**	**LM 4**	**LM 5**	**LM 6** (Without weights)	**LM 7**	**RF**	**LM using VarImp** (Without weights
Net wealth	198,856	213,190	211,953	211,176	203,849	201,311	215,906	205,213	214,152	206,151
RB	0.00	0.07	0.07	0.06	0.03	0.01	0.09	0.03	0.08	0.04
SE	15,660	20,171	19,956	19,964	19,240	18,970	21,448	19,252	20,551	20,427
Real assets	219,354	234,400	233,624	233,009	227,407	223,980	238,329	226,633	233,006	224,605
RB	0.00	0.07	0.07	0.06	0.04	0.02	0.09	0.03	0.06	0.02
SE	18,565	23,917	23,690	23,751	23,012	22,867	26,174	23,001	24,076	23,911
Fin. assets	48,312	51,499	51,257	51,045	49,743	50,213	51,388	50,280	51,606	51,289
RB	0.00	0.07	0.06	0.06	0.03	0.04	0.06	0.04	0.07	0.06
SE	2,640	2,983	2,937	2,930	2,893	3,005	3,054	2,943	3,067	3,289
Total debt	53,824	56,007	55,885	55,799	54,855	53,684	55,474	53,093	56,511	55,585
RB	0.00	0.04	0.04	0.04	0.02	-0.00	0.03	-0.01	0.05	0.03
SE	3,160	3,997	3,996	4,040	3,918	3,895	4,675	4,121	4,221	4,228
HomeOwnRate	0.44	0.47	0.47	0.47	0.46	0.44	0.46	0.44	0.48	0.46
RB	0.00	0.07	0.07	0.07	0.04	-0.01	0.05	0.01	0.08	0.04
SE	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
Value of home	206,992	201,473	200,215	199,609	199,392	201,872	202,216	202,805	201,286	206,866
RB	0.00	-0.03	-0.03	-0.04	-0.04	-0.02	-0.02	-0.02	-0.03	-0.00
SE	7,975	7,800	7,791	7,827	7,904	7,898	7,936	7,852	8,303	9,038
Income	43,747	44,809	44,790	44,754	44,268	43,861	44,930	44,501	45,373	44,859
RB	0.00	0.02	0.02	0.02	0.01	0.00	0.03	0.02	0.04	0.03
SE	1,174	1,328	1,339	1,342	1,331	1,323	1,381	1,362	1,362	1,404

Full: Estimation based on full sample, that is respondents and non-respondents of wave 1
LM 1 - LM7: Response weighting classes and adjustment based on the estimated response propensity logistic regression models 1 to 7.
RF: Response weighting classes and adjustment based on the estimated response propensity from the Random Forest algorithm.
LM using VarImp: Response weighting classes and adjustment based on the estimated response propensity from the logistic regression model with the most "important" variables from the Random Forest algorithm.

Table 8: Mean, relative bias and standard error of selected PHF variables, using non-response adjusted weights based on ten weighting classes and using the inverse of the direct response propensity as the adjustment factor

	Full	**LM 1** (Without weights)	**LM 2**	**LM 3**	**LM 4**	**LM 5**	**LM 6** (Without weights)	**LM 7**	**RF**	**LM using VarImp** (Without weights
Net wealth	198,856	212,274	208,711	208,845	201,475	197,660	211,114	202,517	213,669	202,815
RB	0.00	0.07	0.05	0.05	0.01	-0.01	0.06	0.02	0.07	0.02
SE	15,660	20,588	20,391	20,375	19,690	20,004	21,028	19,996	20,428	20,000
Real assets	219,354	235,719	233,620	234,043	228,382	226,184	236,399	228,569	232,196	220,545
RB	0.00	0.07	0.07	0.07	0.04	0.03	0.08	0.04	0.06	0.01
SE	18,565	23,954	23,601	23,656	22,869	23,029	25,547	23,966	23,897	23,336
Fin. assets	48,312	50,901	50,249	50,266	48,986	49,062	50,434	49,369	51,642	50,853
RB	0.00	0.05	0.04	0.04	0.01	0.02	0.04	0.02	0.07	0.05
SE	2,640	3,139	3,140	3,137	3,130	3,385	3,050	3,091	3,064	3,275
Total debt	53,824	56,665	56,177	56,297	55,459	54,166	55,019	53,312	56,325	54,947
RB	0.00	0.05	0.04	0.05	0.03	0.01	0.02	-0.01	0.05	0.02
SE	3,160	3,984	3,935	3,969	3,927	3,873	4,475	4,286	4,148	4,183
HomeOwnRate	0.44	0.48	0.47	0.47	0.46	0.43	0.46	0.43	0.48	0.45
RB	0.00	0.09	0.07	0.07	0.04	-0.02	0.04	-0.02	0.08	0.02
SE	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
Value of home	206,992	199,701	198,955	199,052	198,560	201,561	201,226	202,863	201,110	207,161
RB	0.00	-0.04	-0.04	-0.04	-0.04	-0.03	-0.03	-0.02	-0.03	0.00
SE	7,975	7,810	7,737	7,760	7,840	7,848	8,022	7,949	8,301	8,985
Income	43,747	44,626	44,190	44,209	43,689	42,948	44,353	43,808	45,323	44,824
RB	0.00	0.02	0.01	0.01	-0.00	-0.02	0.01	0.00	0.04	0.02
SE	1,174	1,522	1,611	1,597	1,636	1,803	1,444	1,554	1,358	1,395

Full: Estimation based on full sample, that is respondents and non-respondents of wave 1
LM 1 - LM7: Response weighting classes and adjustment based on the estimated response propensity logistic regression models 1 to 7.
RF: Response weighting classes and adjustment based on the estimated response propensity from the Random Forest algorithm.
LM using VarImp: Response weighting classes and adjustment based on the estimated response propensity from the logistic regression model with the most “important” variables from the Random Forest algorithm.

Distribution of weights We last compare the distributions of the non-response adjusted weights, and provide some statistics on the mean and standard deviation of the main substantive variables of the survey, income and wealth.

Table 9 shows basic statistics of selected non-response adjusted weights. We see that adjusting directly with the inverse of the response propensity, rather than its mean, increases the variance of the weights, though this is not the case for the adjustment with the response propensity derived from the random forest algorithm. Adjusting directly in that case yields much smaller variance than adjusting within a weighting class, as for example with the weighted response rate. We note, that the latter adjustment had yielded a smaller bias than adjusting directly with the estimated response propensity, again in contrast with the logistic response propensity models. The larger variances in the weights have a modest effect on the standard deviations of net wealth, as for example for the direct response propensity adjustments based on models 6 and (Table 10), but not on the standard deviation of income (as the values of the variables for the observations with the large weights are small).

Table 9: Statistics of selected non-response adjusted weights

Weight	min	max	p50	sd	cv
Final wave 1 weight, unadj.	2.62	89,586	5,851	11,599	1.13
rpm6 + ipw	3.19	562,227	8,430	24,496	1.5
rpm7 + RHG:dw	3.54	262,447	8,640	21,778	1.33
rpm7 + ipw	3.48	809,570	8,601	27,345	1.65
rpm7 + RHG:mp	3.53	257,741	8,680	21,555	1.32
cRF + ipw	3.22	131,341	8,363	17,531	1.16
cRF + RHG:dw	2.64	210,513	8,303	22,859	1.4

Table 10: Mean and standard deviation of household income and wealth of wave 2, using selected non-response adjusted weights

Weight	Income	sd	Wealth	sd
Final wave 1 weight, unadj.	54,508	1,917	254,261	18,453
rpm6 + ipw	52,131	1,836	226,279	16,592
rpm7 + cells + dw	52,256	1,910	222,906	15,794
rpm7 + cells + mp	52,240	1,906	222,730	15,637
rpm7 + ipw	51,802	1,922	218,046	16,510
cRF + ipw	53,179	1,920	228,372	14,000
cRF + dw	51,502	2,193	208,998	14,963

6. Conclusions

This paper examined different non-response adjustment methods based on logistic regression models and a random forest algorithm for the panel component of the Panel on Household Finances. Besides the functional form of the response process, logistic regression or random forest, different adjustments were compared: the individual estimated response propensity vs. response homogeneity groups, and within the latter different adjustment factors were tested, such as the observed response rate, the design weighted response rate and the mean estimated propensity within the group. Since our study focused on the panel component of the second wave of the survey, it was possible to use variables from the previous wave of the survey in the estimation of the response propensities. The latter also served to evaluate the resulting weighting adjustments. In particular, we compared the weighting adjustments in terms of the reduction in the relative bias of the resulting estimates of selected variables of the previous wave of the survey, wave 1.

For the logistic regression models, the choice of variables was hand-picked, and included sociodemographic variables, variables related to the survey design, paradata recorded by the interviewers, data related to the contact effort and variables related to key survey variables, namely household wealth and income, or variables related to them. The random forest has a rather flexible overall function form (ensemble of trees). It allows a far larger number to be included as predictors in the classification process and, since the search of covariates is automatic, it can prevent the omission of important predictors.

However, there was no reduction of bias using the random forest algorithm. For the estimates of income and wealth, more parsimonious logistic regression models that include these variables result in non-response adjusted weights that lead to smaller bias estimators than the random forest algorithms. In fact, augmenting the models results in non-response adjustments that increase the bias for these variables, though it is beneficial for other survey variables. Our results indicate, that the random forest algorithm can be useful in indicating variables related to the non-response process that could later feed in a logistic regression model. The choice of model and the breadth of variables included may also depend on the overall purpose of overall adjustment, that is whether it should serve as many variables as possible, or whether some variables are of particular interest and improving their accuracy is of foremost importance. In that case a more parsimonious carefully defined logistic response model would serve better this purpose.

As regards the non-response adjustment factors within weighting classes, results are different between the logistic regression models and the random forest. In the adjustments based on the logistic regression models, adjusting with the inverse of the mean response propensity within a weighting class, rather than the observed response rates, seemed to lead to a reduction in bias for some variables, though differences were small. In contrast to that, for the adjustments based on the random forest, adjusting with the observed response rates, either unweighted or weighted, always resulted in a more marked reduction in bias in the resulting non-response adjusted estimates of the selected variables. In both functional forms, logistic models or random forest, the relative bias of the estimates derived using either the inverse of the response propensity or the inverse of the mean propensity of a class is very similar. Since the latter provides weights with smaller variance, it should be the preferred method of adjustment when a logistic regression model is used.

7. Acknowledgements

We thank Arthur B. Kennickell for organising the session on "Measuring Household Wealth in Europe: The Household Finance and Consumption Survey" in the JSM 2018 and including us in the programme. We also thank Arthur B. Kennickell and Tobias Schmidt for helpful comments and suggestions on an earlier draft of this paper.

References

Andy Peytchev and Kristen Olson (2007). Using interviewer observations to improve non-response adjustments: NES 2004. Sociology Department, Faculty Publications 145, University of Nebraska - Lincoln.

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

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

Brick, J. M. (2013). Unit nonresponse and weighting adjustments: A critical review. *Journal of Official Statistics*, 29(3).

Buskirk, T. D. and Kolenikov, S. (2015). Finding respondents in the forest: A comparison of logistic regression and random forest models for response propensity weighting and stratification. *Survey Insights: Methods from the Field*.

Chen, Q., Gelman, A., Tracy, M., Norris, F. H., and Galea, S. (2015). Incorporating the sampling design in weighting adjustments for panel attrition. *Statistics in medicine*, 34(28):3637–3647.

da Silva, D. N. and Opsomer, J. D. (2009). Nonparametric propensity weighting for survey nonresponse through local polynomial regression. *Survey Methodology*, 35(2):165–176.

Earp, M., Mitchell, M., McCarthy, J., and Kreuter, F. (2014). Modeling nonresponse in establishment surveys: Using an ensemble tree model to create nonresponse propensity scores and detect potential bias in an agricultural survey. *Journal of Official Statistics*, 30(4).

Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems. *J. Mach. Learn. Res*, 15(1):3133–3181.

Friedman, J., Hastie, T., and Tibshirani, R. (2009). *The elements of statistical learning*. Springer series in statistics, New York, NY, USA.

Gouskova, E., Heeringa, E., McGonagle, K., Schoeni, R., and Stafford, F. (2008). Panel Study of Income Dynamics revised longitudinal weights 19932005. PSID technical report, University of Michigan, Ann Arbor, USA.

Hapfelmeier, A., Hothorn, T., Ulm, K., and Strobl, C. (2014). A new variable importance measure for random forests with missing data. *Statistics and Computing*, 24(1):21–34.

Janitza, S., Strobl, C., and Boulesteix, A.-L. (2013). An auc-based permutation variable importance measure for random forests. *BMC bioinformatics*, 14(1):119.

Kalton, G. and Flores-Cervantes, I. (2003). Weighting methods. *Journal of Official Statistics*, 19(3):81–97.

Kennickell, A. B. and Woodburn, R. L. (1999). Consistent weight design for the 1989, 1992 and 1995 SCFs and the dsitribution of wealth. *Review of Income and Wealth*.

Knerr, P., Forkert, A., Chudziak, N., Gilberg, R., and Kleudgen, M. (2015). Methodenbericht. private haushalte und ihre finanzen (PHF. 2.Erhebungswelle 2014). Technical report, infas, Bonn.

Kreuter, F., Olson, K., Wagner, J., Yan, T., Ezzati-Rice, T. M., Casas-Cordero, C., Lemay, M., Peytchev, A., Groves, R. M., and Raghunathan, T. E. (2010). Using proxy measures and other correlates of survey outcomes to adjust for non-response: Examples from multiple surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 173(2):389–407.

Lee, B. K., Lessler, J., and Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in medicine*, 29(3):337–346.

Little, R. J. and Rubin, D. B. (2014). *Statistical analysis with missing data*. John Wiley & Sons.

Little, R. J. and Vartivarian, S. (2003). On weighting the rates in non-response weights. *Statistics in medicine*, 22(9):1589–1599.

Little, R. J. and Vartivarian, S. (2005). Does weighting for nonresponse increase the variance of survey means? *Survey Methodology*, 31(2):161–168.

Little, R. J. A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review / Revue Internationale de Statistique*, 54(2):139–157.

Lohr, S., Hsu, V., and Montaquila, J. (2015). Using classification and regression trees to model survey nonresponse. *Survey Research Methods Section*.

Lynn, P. (1996). Weighting for non-response. In et al., R. B., editor, *Survey and Statistical Computing*, pages 205–214. Chesham.

Phipps, P. and Toth, D. (2012). Analyzing establishment nonresponse using an interpretable regression tree model with linked administrative data. *The Annals of Applied Statistics*, 6(2):772–794.

Rizzo, L., Kalton, G., and Brick, J. M. (1996). A comparsion of some weighting adjustment methods for panel nonresponse. *Survey Methodology*, 22(1):43–53.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*.

Schouten, B. and de Nooij, G. (2005). *Nonresponse adjustment using classification trees*. CBS, Statistics Netherlands.

Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(1):25.

Torsten Hothorn, Kurt Hornik, and Achim Zeileis (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, 15(3):651–674.

Tzamourani, P. (2015). Constructing cross-sectional weights for the german panel on household finances. In *JSM Proceedings, Social Statistics Section*, pages 3462–3469, Alexandria. VA: American Statistical Association.

von Kalckreuth, U., Eisele, M., Blanc, J. L., Schmidt, T., and Zhu, J. (2012). The PHF: a comprehensive panel survey on household finances and wealth in Germany. Discussion Paper 13, Deutsche Bundesbank, Frankfurt am Main.