Nonresponse adjustment using classification trees

Discussion paper 05001

Barry Schouten and Guido de Nooij

Statistics Netherlands

Voorburg/Heerlen, February 2005

Explanation of symbols

.	= data not available
*	= provisional figure
x	= publication prohibited (confidential figure)
–	= nil or less than half of unit concerned
–	= (between two figures) inclusive
0 (0,0)	= less than half of unit concerned
blank	= not applicable
2004–2005	= 2004 to 2005 inclusive
2004/2005	= average of 2004 up to and including 2005
2004/'05	= crop year, financial year, school year etc. beginning in 2004 and ending in 2005

Due to rounding, some totals may not correspond with the sum of the separate figures.

Publisher
Statistics Netherlands
Prinses Beatrixlaan 428
2273 XZ Voorburg
The Netherlands

Printed by
Statistics Netherlands - Facility Services

Cover design
WAT ontwerpers, Utrecht

Information
E-mail: infoservice@cbs.nl

Where to order
E-mail: verkoop@cbs.nl

Internet
http://www.cbs.nl

ISSN: 1572-0314
Key figure: X-10
Production code: 6008305001

Statistics Netherlands

NONRESPONSE ADJUSTMENT USING CLASSIFICATION TREES

Summary: Nonresponse adjustment methods make use of covariates that are available for both respondents and non-respondents. A problem is the selection of covariates that relate both to the key survey questions and to the response behaviour. Therefore, often the process of selection is performed in two steps.

We present a classification tree method that allows for the construction of weighting strata that simultaneously account for the relation between response behaviour, survey questions and covariates.

We apply the classification trees to survey data of Statistics Netherlands.

Keywords: Data Mining; Bias; Linear weighting; Poststratification; Strata.

1. Introduction

Nonresponse to surveys affects population estimators in case on average respondents and non-respondents give different answers to the survey questions. Auxiliary information is usually linked to the survey so that potential bias can be detected and corrected for. Commonly used techniques are linear weighting, multiplicative weighting and propensity score weighting. For an overview of adjustment methods we refer to Bethlehem (2002) and Kalton and Flores-Cervantes (2003).

Crucial in the successful employment of adjustment methods is the validity of the assumptions underlying the methods. Most techniques assume that conditionally on a set of available auxiliary variables respondents cannot be distinguished from non-respondents when it comes to the survey topics. Hence, in case the values of these variables are fixed response is at random, a feature called Missing-at-Random in the literature. Although, it seems reasonable that fixing a number of characteristics makes respondents resemble non-respondents, there is little empirical evidence in practice to support the assumption. In fact, when more auxiliary information becomes available as was the case at Statistics Netherlands, it follows that current weighting models can be improved (see Schouten 2003). The additional variables either give a better explanation of response behaviour or are better predictors of the survey questions.

In this paper we do not make assumptions about the missing data mechanism and, hence, make a Not-Missing-at-Random assumption. Schouten (2004) shows that in general an interval can be set up for the bias of the response mean and the bias of the poststratification estimator. The width of the bias interval depends on the correlation between the 0-1 response indicator and auxiliary variables and the correlation between a survey question and auxiliary variables.

We have two motivations for employing the weaker Not-Missing-at-Random assumption. First, as we already mentioned we do not believe the Missing-at-Random assumption to always hold since at Statistics Netherlands newly available auxiliary information indicated that current weighting models may still lead to biased estimates. Also, one can never preclude that in the future more relevant auxiliary information can be deployed in the adjustment of nonresponse. Secondly, even when the final set of weighting variables comes close to being Missing-at-Random, we still need a rule to decide which variables to use and which to omit in the adjustment. It seems straightforward to choose the width of the bias interval as a criterion to compare weighting models, because this interval accounts simultaneously for the relation between response behaviour and auxiliary information and the relation between survey questions and auxiliary information.

The objective of this paper is the efficient search for a weighting model that minimises the width of the bias interval as given in Schouten (2004). We propose a classification tree method with the interval width as a splitting rule and the significance of the decrease in interval width as stopping rule.

Classification trees and its continuous counterpart regression trees are today assigned as data mining technique but go back to the sixties. It all started with the so-called Automatic Interaction Detector (AID) proposed by Morgan and Sonquist (1963). They suggest to partition a population in homogeneous subpopulations by means of repeated binary splits. In the succeeding decades several variants of their technique were developed, e.g. the well-known CHAID by Kass (1980). For an overview of classification and regression trees see for example Breiman et al. (1984) and Murthy (1998). Classification and regression trees have also been designed for the imputation of missing data, see Mesa, Tsai and Chambers (2000). We employ these techniques to efficiently adjust for unit nonresponse in surveys.

The classification tree that we have developed does not resemble any other existing classification tree method when it comes to the splitting and stopping rules. This is because we want to minimise the width of the bias interval. The leaves of our classification tree form the weighting strata in the poststratification estimator. A stratum is split into two new strata whenever it leads to the largest decrease in width of the bias interval for the poststratification estimator. However, in case this decrease is not significant on a prescribed level, then the split is not permitted and the classification stops.

In section 2 we give some background to the poststratification estimator and its properties. We motivate and describe the proposed classification tree algorithm in section 3. Next in section 4 we apply the algorithm to the Dutch Integrated Survey on Household Living Conditions (in Dutch Permanent Onderzoek Leefsituatie). We round off the paper with a discussion of the results in section 5.

2. Poststratification

2.1 The poststratification estimator

The classification that we propose amounts to the formation of strata in the poststratification estimator. We form these strata by repeated splits of the target population of the survey.

Poststratification is a method often used to assign weights to respondents. For a reference see Cochran (1977). The target population of a survey is partitioned into so-called strata. The strata are disjoint sets formed by the categories of auxiliary variables and it is assumed that these strata are more homogeneous with respect to the survey questions. Originally, poststratification was designed for the case where there is no nonresponse in order to reduce the variance of population estimators. However, in surveys where part of the sample does not respond, poststratification can also be used to reduce bias. In that case the objective is twofold. The categorical auxiliary variables that form the strata should relate both to the survey questions and to response behaviour. In essence, the poststratification estimator predicts the answers of non-respondents by the average answer of respondents in the same stratum.

First, we introduce some notation. In the following we distinguish random variables from their realisations by using upper-case and lower-case letters, i.e. the realisation of a random variable Z is denoted by z. We let μ_Z en σ_Z be the expectation and standard deviation of variable Z, and $c(Z_1, Z_2)$ and $\gamma(Z_1, Z_2)$ correspond to, respectively, the covariance and correlation between variables Z_1 and Z_2.

Let $(R_i, \Delta_i, Y_i)_{1 \le i \le n}$ be independent and identically distributed. Here, R_i stands for the 0-1 indicator for response, i.e. $R_i = 1$ in case unit i responds in the survey and $R_i = 0$ otherwise. The vector $\Delta_i = (\Delta_{1,i}, \Delta_{2,i}, \ldots, \Delta_{H,i})^{/}$ corresponds to the H disjunct categories of a nominal or ordinal auxiliary variable, i.e. $\Delta_{h,i} = 1$ if unit i belongs to category h and $\Delta_{h,i} = 0$ otherwise, and $\sum_{h=1}^{H} \Delta_{h,i} = 1$. The vector Δ_i gives a stratification and we will refer to the categories as strata. Finally, Y_i represents an answer to a survey question.

We let q_h denote the marginal probability for stratum h, $q_h = P[\Delta_{h,i} = 1]$, let p_h be the conditional probability of response in stratum h, $p_h = P[R_i = 1 \mid \Delta_{h,i} = 1]$, and μ_Y^h be the conditional expectation of Y in stratum h, $\mu_Y^h = E(Y_i \mid \Delta_{h,i} = 1)$. The overall response probability is denoted by p.

We assume that the values of the auxiliary variables are observed for all units. However, the answers to the survey question are only observed for the respondents. Furthermore, we assume that the marginal probabilities q_h are known.

The objective of this paper is the estimation of the unknown expectation of the survey question Y, i.e. of $\mu_Y = EY_i$.

For this purpose we use the poststratification estimator $\hat{\bar{y}}_{post}$, defined by

$$\hat{\bar{y}}_{post} = \sum_{h=1}^{H} q_h \bar{Y}_h^* , \tag{1}$$

with $\bar{Y}_h^*$ the response mean to survey question Y in stratum h, i.e.

$$\bar{Y}_h^* = \frac{\sum_{i=1}^{n} R_i Y_i \Delta_{h,i}}{\sum_{i=1}^{n} R_i \Delta_{h,i}} . \tag{2}$$

2.2 Bias of the poststratification estimator

The bias of the poststratification estimator (1) can be derived under the condition that each stratum contains at least one respondent

$$\sum_{i=1}^{n} R_i \Delta_{h,i} > 0, \quad \forall h. \tag{3}$$

First, we look at the bias of the stratum response mean $\bar{Y}_h^*$ under (3) relative to μ_Y^h. This bias equals

$$\begin{aligned} B(\bar{Y}_h^*) &= E\left(\bar{Y}_h^* \mid \textstyle\sum_{i=1}^{n} R_i \Delta_{h,i} > 0\right) - \mu_Y^h \\ &= \sum_{i=1}^{n} \left(P[R_i \Delta_{h,i} = 1 \mid \textstyle\sum_{i=1}^{n} R_i \Delta_{h,i} > 0] \; E\left(Y_i (1 + \textstyle\sum_{\substack{j=1 \\ j \neq i}}^{n} R_j \Delta_{h,j})^{-1} \mid R_i \Delta_{h,i} = 1 \right) \right) - \mu_Y^h, \end{aligned}$$

since each term vanishes for which $R_i \Delta_{h,i} = 0$ and (3) is satisfied in stratum h in case $R_i \Delta_{h,i} = 1$.

The sum $\sum_{i=1}^{n} R_i \Delta_{h,i}$ is binomially distributed with parameters n and $q_h p_h$. Hence, by application of Bayes' rule

$$\begin{aligned} P[R_i \Delta_{h,i} = 1 \mid \textstyle\sum_{i=1}^{n} R_i \Delta_{h,i} > 0] &= \frac{P[R_i \Delta_{h,i} = 1, \textstyle\sum_{i=1}^{n} R_i \Delta_{h,i} > 0]}{P[\textstyle\sum_{i=1}^{n} R_i \Delta_{h,i} > 0]} \\ &= \frac{P[R_i \Delta_{h,i} = 1]}{1 - P[\textstyle\sum_{i=1}^{n} R_i \Delta_{h,i} = 0]} \\ &= \frac{q_h p_h}{1 - (1 - q_h p_h)^n} . \end{aligned} \tag{4}$$

Furthermore, since Y_i and $\left(1+\sum_{\substack{j=1\\ j\neq i}}^{n} R_j\Delta_{h,j}\right)^{-1}$ are independent we have that

$$\begin{aligned} E\left(Y_i(1+\sum_{\substack{j=1\\ j\neq i}}^{n} R_j\Delta_{h,j})^{-1} \mid R_i\Delta_{h,i}=1\right) &= E\left(Y_i \mid R_i\Delta_{h,i}=1\right)E\left(1+\sum_{\substack{j=1\\ j\neq i}}^{n} R_j\Delta_{h,j}\right)^{-1} \\ &= \frac{E\left(Y_iR_i \mid \Delta_{h,i}=1\right)}{p_h} E\left(1+\sum_{\substack{j=1\\ j\neq i}}^{n} R_j\Delta_{h,j}\right)^{-1} \\ &= \frac{E\left(Y_iR_i \mid \Delta_{h,i}=1\right)}{p_h} \frac{(1-(1-q_hp_h)^n)}{nq_hp_h}. \end{aligned} \tag{5}$$

The last step in (5) can be obtained using the fact that $\sum_{\substack{j=1\\ j\neq i}}^{n} R_j\Delta_{h,j}$ is again binomially distributed, however, with parameters $n-1$ and q_hp_h. If Z is binomially distributed with parameters m and ϕ, then it can be shown that $E\frac{1}{Z+1}=\frac{1-(1-\phi)^{m+1}}{(m+1)\phi}$.

Now, if we combine (4) and (5) we get

$$B(\bar{Y}_h^*)=\sum_{i=1}^{n}\frac{E\left(Y_iR_i \mid \Delta_{h,i}=1\right)}{np_h}-\mu_Y^h=\frac{E(Y_iR_i \mid \Delta_{h,j}=1)}{p_h}-\mu_Y^h. \tag{6}$$

We define $c_h(Z_1,Z_2)$ and $\gamma_h(Z_1,Z_2)$ as the conditional covariance and correlation, respectively, between Z_1 and Z_2 in stratum h. Furthermore, we let σ_Z^h be the conditional standard deviation of Z in stratum h.

We can now rewrite (6) to

$$B(\bar{Y}_h^*)=\frac{c_h(Y_i,R_i)+\mu_Y^hp_h}{p_h}-\mu_Y^h=\frac{c_h(Y_i,R_i)}{p_h}=\gamma_h(Y_i,R_i)\sqrt{\frac{1-p_h}{p_h}}\sigma_Y^h, \tag{7}$$

using the fact that the conditional standard deviation of R_i in stratum h equals $\sqrt{p_h(1-p_h)}$.

Consequently, the bias of the poststratification estimator under condition (3) follows from (1) and (7)

$$\begin{aligned} B(\hat{\bar{y}}_{post}) &= \sum_{h=1}^{H} q_hB(\bar{Y}_h^*) \\ &= \sum_{h=1}^{H} q_h\gamma_h(Y_i,R_i)\sqrt{\frac{1-p_h}{p_h}}\sigma_Y^h. \end{aligned} \tag{8}$$

2.3 A local bias interval for the poststratification estimator

First, we derive an interval for the bias in each stratum of the poststratification estimator. We will call this interval the local bias interval. In the next section we derive an alternative interval that we will refer to as the global bias interval.

We can construct sharp lower and upper limits for the bias of the poststratification estimator using the following lemma.

Lemma 2.3: If variables Z_1, Z_2 and Z_3 have a finite variance, then the correlation $\gamma(Z_1, Z_2)$ between Z_1 and Z_2 is bounded by

$$\begin{aligned}\gamma(Z_1,Z_3)\gamma(Z_2,Z_3) - \sqrt{1-\gamma(Z_1,Z_3)^2}\sqrt{1-\gamma(Z_2,Z_3)^2} \le \gamma(Z_1,Z_2) \\ \le \gamma(Z_1,Z_3)\gamma(Z_2,Z_3) + \sqrt{1-\gamma(Z_1,Z_3)^2}\sqrt{1-\gamma(Z_2,Z_3)^2},\end{aligned} \quad (9)$$

and the bounds in (9) are sharp.

We refer to Schouten (2002) for a proof of lemma 2.3. By sharp bounds we mean that we can find Z_1, Z_2 and Z_3 that correspond to any value in interval (9) while fixing $\gamma(Z_1, Z_3)$ and $\gamma(Z_2, Z_3)$. In the following we omit the index i to the random variables.

Let X be some random variable with finite σ_X. If we combine (8) and (9) we get

$$\begin{aligned}&\sum_{h=1}^{H} q_h \sqrt{\frac{1-p_h}{p_h}}\sigma_Y^h\left(\gamma_h(X,R)\gamma_h(X,Y) - \sqrt{1-\gamma_h(X,R)^2}\sqrt{1-\gamma_h(X,Y)^2}\right) \\ &\le B(\hat{\bar{y}}_{post}) \\ &\le \sum_{h=1}^{H} q_h \sqrt{\frac{1-p_h}{p_h}}\sigma_Y^h\left(\gamma_h(X,R)\gamma_h(X,Y) + \sqrt{1-\gamma_h(X,R)^2}\sqrt{1-\gamma_h(X,Y)^2}\right).\end{aligned} \quad (10)$$

Hence, any auxiliary variable defines an interval for the bias of the poststratification estimator. Since the bounds in (10) are sharp, we may take the intersection over all available auxiliary variables other than the variable that is used to construct the stratification. We may, however, also take different auxiliary variables for each stratum. Let $X_1, X_2, \ldots, X_H$ be H, not necessarily different, auxiliary variables. We can again use (8) and (9) to get the following lower and upper limit for the bias

$$\begin{aligned}&\sum_{h=1}^{H} q_h \sqrt{\frac{1-p_h}{p_h}}\sigma_Y^h\left(\gamma_h(X_h,R)\gamma_h(X_h,Y) - \sqrt{1-\gamma_h(X_h,R)^2}\sqrt{1-\gamma_h(X_h,Y)^2}\right) \\ &\le B(\hat{\bar{y}}_{post}) \\ &\le \sum_{h=1}^{H} q_h \sqrt{\frac{1-p_h}{p_h}}\sigma_Y^h\left(\gamma_h(X_h,R)\gamma_h(X_h,Y) + \sqrt{1-\gamma_h(X_h,R)^2}\sqrt{1-\gamma_h(X_h,Y)^2}\right).\end{aligned} \quad (11)$$

Since we can bound $\gamma_h(X_h,R)\gamma_h(X_h,Y) - \sqrt{1-\gamma_h(X_h,R)^2}\sqrt{1-\gamma_h(X_h,Y)^2} \ge -1$ and $\gamma_h(X_h,R)\gamma_h(X_h,Y) + \sqrt{1-\gamma_h(X_h,R)^2}\sqrt{1-\gamma_h(X_h,Y)^2} \le 1$, it always holds that

$$-\sum_{h=1}^{H} q_h \sqrt{\frac{1-p_h}{p_h}}\sigma_Y^h \le B(\hat{\bar{y}}_{post}) \le \sum_{h=1}^{H} q_h \sqrt{\frac{1-p_h}{p_h}}\sigma_Y^h, \quad (12)$$

and the maximal absolute bias equals $\sum_{h=1}^{H} q_h \sqrt{\frac{1-p_h}{p_h}} \sigma_Y^h$.

We call interval (11) the local bias interval as it allows for a local search for new strata. In the next section we set up an alternative bias interval.

2.4 A global bias interval for the poststratification estimator

In this section we argue that the bias that is maximally possible in absolute sense is the same for the poststratification estimator $\hat{\bar{y}}_{post}$ and the response mean defined by

$$\bar{Y}^* = \frac{\sum_{i=1}^{n} R_i Y_i}{\sum_{i=1}^{n} R_i}.$$

First, we derive the bias of $\bar{Y}^*$ by assuming only one stratum. Similarly to (7) we have

$$B(\bar{Y}^*) = \frac{c(Y_i, R_i) + \mu_Y p}{p} - \mu_Y = \frac{c(Y_i, R_i)}{p} = \gamma(Y_i, R_i)\sqrt{\frac{1-p}{p}}\sigma_Y, \qquad (13)$$

where σ_Y is the standard deviation of Y. Similarly to (10) we then get the following bias interval of the response mean when using auxiliary variable X

$$\begin{aligned} &\sqrt{\frac{1-p}{p}}\sigma_Y\left(\gamma(X,R)\gamma(X,Y) - \sqrt{1-\gamma(X,R)^2}\sqrt{1-\gamma(X,Y)^2}\right) \\ &\le B(\bar{Y}^*) \le \sqrt{\frac{1-p}{p}}\sigma_Y\left(\gamma(X,R)\gamma(X,Y) + \sqrt{1-\gamma(X,R)^2}\sqrt{1-\gamma(X,Y)^2}\right). \end{aligned} \qquad (14)$$

Next, we write the poststratification estimator $\hat{\bar{y}}_{post}$ in the form of a regression estimator. The estimator can be rewritten to

$$\begin{aligned} \hat{\bar{y}}_{post} &= \sum_{h=1}^{H} q_h \bar{Y}_h^* = \sum_{h=1}^{H-1} q_h \bar{Y}_h^* + (1 - q_1 - \ldots - q_{H-1})\bar{Y}_H^* \\ &= \bar{Y}_H^* + \sum_{h=1}^{H} q_h (\bar{Y}_h^* - \bar{Y}_H^*) \\ &= \bar{Y}_H^* + \sum_{h=1}^{H} q_h (\bar{Y}_h^* - \bar{Y}_H^*) - \sum_{h=1}^{H} \bar{\Delta}_h^* \bar{Y}_h^* + \bar{Y}^* \\ &= \bar{Y}^* + \sum_{h=1}^{H} (q_h - \bar{\Delta}_h^*)(\bar{Y}_h^* - \bar{Y}_H^*), \end{aligned} \qquad (15)$$

where $\bar{\Delta}_h^*$ is the response mean of the auxiliary stratum indicator Δ_h. Hence, the poststratification estimator amounts to the regression estimator with the first $H-1$ strata indicator variables Δ_h as predictors, assuming the stratum probabilities q_h to be known and incorporating an intercept in the model. If we let α be the intercept

and β be the vector of slope parameters, then the estimators for these parameters in (15) are $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_{H-1})' = (\bar{Y}_1^* - \bar{Y}_H^*, \ldots, \bar{Y}_{H-1}^* - \bar{Y}_H^*)'$ and $\hat{\alpha} = \bar{Y}^* - \sum_{h=1}^{H-1} \bar{\Delta}_h^* \hat{\beta}_h$.

Let us now suppose that the true differences between the stratum means defined by $\lambda_{h,k} := \mu_Y^h - \mu_Y^k$ are known, but that the stratum means μ_Y^h themselves are unknown. If we would know one of the stratum means, say μ_Y^H, then clearly we also know the others since we can simply add $\lambda_{h,H}$. Hence, we only need to estimate one stratum mean. We take μ_Y^H as the unknown stratum mean.

An obvious estimator for μ_Y^H is

$$\hat{\mu}_Y^H = \sum_{h=1}^{H} \bar{\Delta}_h^* (\bar{Y}_h^* - \lambda_{h,H}) . \tag{16}$$

If all strata would have the same mean, i.e. $\lambda_{h,H} = 0$, then $\hat{\mu}_Y^H$ is simply the response mean. In case there would be no nonresponse, then $\hat{\mu}_Y^H$ is the maximum likelihood estimator. Provided the $\lambda_{h,H}$ are known and using (16) we get the following estimator for μ_Y

$$\hat{\mu}_Y = \sum_{h=1}^{H} q_h (\hat{\mu}_Y^H + \lambda_{h,H}) . \tag{17}$$

We can rewrite (17) to

$$\begin{aligned} \hat{\mu}_Y &= \sum_{h=1}^{H} q_h \Big(\sum_{k=1}^{H} \bar{\Delta}_k^* (\bar{Y}_k^* - \lambda_{k,H}) + \lambda_{h,H} \Big) \\ &= \sum_{k=1}^{H} \bar{\Delta}_k^* (\bar{Y}_k^* - \lambda_{k,H}) + \sum_{h=1}^{H} q_h \lambda_{h,H} \\ &= \bar{Y}^* + \sum_{h=1}^{H} (q_h - \bar{\Delta}_h^*) \lambda_{h,H} . \end{aligned} \tag{18}$$

We can, thus, conclude that $\hat{\mu}_Y$ is equal to the regression estimator when the observed differences between the strata, i.e. $\bar{Y}_h^* - \bar{Y}_H^*$, are replaced by the true stratum differences $\lambda_{h,H}$.

We construct a bias interval for $\hat{\mu}_Y$ and show that it has the same width as that of the response mean given by (14). Conditionally on $\sum_{i=1}^{n} R_i > 0$ the bias of estimator $\hat{\mu}_Y$ equals

$$\begin{aligned} B(\hat{\mu}_Y) &= B(\bar{Y}^*) - \sum_{h=1}^{H} E(\bar{\Delta}_h^* - q_h) \lambda_{h,H} \\ &= B(\bar{Y}^*) - \sum_{h=1}^{H} B(\bar{\Delta}_h^*) \lambda_{h,H} \end{aligned}$$

$$= B(\overline{Y}^*) - \sum_{h=1}^{H} \frac{c(\Delta_h, R)}{p} \lambda_{h,H}$$

$$= B(\overline{Y}^*) - \frac{c(\sum_{h=1}^{H} \lambda_{h,H}\Delta_h, R)}{p}$$

$$= B(\overline{Y}^*) - \sqrt{\frac{1-p}{p}}\, \gamma(\sum_{h=1}^{H} \lambda_{h,H}\Delta_h\ , R)\, \sigma_{\sum_{h=1}^{H} \lambda_{h,H}\Delta_h}. \qquad (19)$$

The expression for the bias of $\overline{\Delta}_h^*$ in the third line of (19) is derived analogously to the bias of the response mean.

Now, let $X = \sum_{h=1}^{H} \lambda_{h,H}\Delta_h$ in (14). From (19) it follows that the bias of the estimator $\hat{\mu}_Y$ falls in an interval with the same width as the bias interval of the response mean $\overline{Y}^*$, since in lemma 2.3 all correlations are fixed including $\gamma(X, R)$. The interval width of (19) equals

$$2\sqrt{\frac{1-p}{p}}\sigma_Y \sqrt{1-\gamma(\sum_{h=1}^{H} \lambda_{h,H}\Delta_h, R)^2}\ \sqrt{1-\gamma(\sum_{h=1}^{H} \lambda_{h,H}\Delta_h, Y)^2}. \qquad (20)$$

As a consequence the bias interval has the same width (20) for both estimators. We, therefore, strongly believe that for the poststratification estimator $\hat{\bar{y}}_{post}$ also a bias interval with the same width as (20) can be constructed, since contrary to the response mean this estimator does make use of auxiliary information but it does not use the true differences $\lambda_{h,H}$ between the strata. In other words, if the parameters $\lambda_{h,H}$ are replaced by their corresponding response based estimators $\overline{Y}_h^* - \overline{Y}_H^*$, then the width of the bias interval remains as it is. We do not have a proof, however.

It is not difficult to show that (20) is equal to

$$2\sqrt{\frac{1-p}{p}}\sigma_Y \sqrt{1-\gamma(\sum_{h=1}^{H} \mu_Y^h\Delta_h, R)^2}\ \sqrt{1-\gamma(\sum_{h=1}^{H} \mu_Y^h\Delta_h, Y)^2}. \qquad (21)$$

We refer to (19) as the global bias interval, since we can use it for a global search for new strata. This interval serves as the basis for the classification algorithm in this paper. In the next section we elaborate on the algorithm.

3. The classification tree algorithm

3.1 Motivation

In practice the formation of strata is not straightforward. When the categories of auxiliary variables are crossed it may occur that empty or almost empty strata are formed. In that case the answer of non-respondents in the same stratum cannot be predicted or the prediction is based on too small a number of respondents causing

variance to grow. Furthermore, it may not be efficient to cross every category of one auxiliary variable with every category of another auxiliary variable. Finally, a criterion must be chosen that makes it possible to compare different sets of strata. When should one choice of strata be favoured to another choice of strata?

The problem of forming strata is addressed extensively in the literature. Little (1986) for instance suggests forming so-called adjustment cells by first dividing respondents that have similar response behaviour into separate cells and then pooling those cells that give similar answers to the survey questions.

In this paper we make no assumption about the missing data mechanism. In sections 2.3 and 2.4 we derived general intervals for the bias of the poststratification estimator. The intervals give us the maximal absolute bias, i.e. the bias under the worst case scenario. We can use the maximal absolute bias as a criterion for the selection of strata. Strata are only subdivided into new strata in case the new stratification leads to a significant decrease in the maximal absolute bias.

In this paper we restrict ourselves to the construction of strata using the global bias interval. The approach is similar to Schouten (2004) where a forward inclusion – backward elimination strategy is applied.

The width of the bias interval (19) is proportional to

$$w = \sqrt{1-\gamma(\sum_{h=1}^{H}\mu_Y^h\Delta_h, R)^2}\sqrt{1-\gamma(\sum_{h=1}^{H}\mu_Y^h\Delta_h, Y)^2}. \qquad (22)$$

Both the standard deviation of the survey question σ_Y and the response probability p do not depend on the stratification chosen. Hence, we focus on the stratification that minimises (22).

Clearly, the expectations μ_Y^h are not known nor are the correlations in (22). We, therefore, have to rely on estimators that may themselves be biased. We propose to minimise

$$w^* = \sqrt{1-\Gamma(\sum_{h=1}^{H}\overline{Y}_h^*\Delta_h, R)^2}\sqrt{1-\Gamma^*(\sum_{h=1}^{H}\overline{Y}_h^*\Delta_h, Y)^2}, \qquad (23)$$

with $\Gamma(\sum_{h=1}^{H}\overline{Y}_h^*\Delta_h, R)$ the sample correlation between the predictor $\sum_{h=1}^{H}\overline{Y}_h^*\Delta_{h,i}$ and the response indicators R_i

$$\Gamma(\sum_{h=1}^{H}\overline{Y}_h^*\Delta_h, R) = \frac{\sum_{i=1}^{n}(\sum_{h=1}^{H}\overline{Y}_h^*\Delta_{h,i} - \sum_{h=1}^{H}\overline{Y}_h^*\overline{\Delta}_h)(R_i-\overline{R})}{\sqrt{\sum_{i=1}^{n}(\sum_{h=1}^{H}\overline{Y}_h^*\Delta_{h,i} - \sum_{h=1}^{H}\overline{Y}_h^*\overline{\Delta}_h)^2}\sqrt{\sum_{i=1}^{n}(R_i-\overline{R})^2}}$$

and $\Gamma^*(\sum_{h=1}^{H}\overline{Y}_h^*\Delta_h, Y)$ the response correlation between the predictor $\sum_{h=1}^{H}\overline{Y}_h^*\Delta_{h,i}$ and survey question Y_i

$$\Gamma^*(\sum_{h=1}^{H}\overline{Y}_h^*\Delta_h, Y) = \frac{\sum_{i=1}^{n}R_i(\sum_{h=1}^{H}\overline{Y}_h^*\Delta_{h,i} - \sum_{h=1}^{H}\overline{Y}_h^*\overline{\Delta}_h^*)(Y_i-\overline{Y}^*)}{\sqrt{\sum_{i=1}^{n}R_i(\sum_{h=1}^{H}\overline{Y}_h^*\Delta_{h,i} - \sum_{h=1}^{H}\overline{Y}_h^*\overline{\Delta}_h^*)^2}\sqrt{\sum_{i=1}^{n}R_i(Y_i-\overline{Y}^*)^2}}.$$

If we have two stratifications, i.e. two sets of strata, then we let Δw^* denote the difference in width given by (23).

A classification tree is an ideal method to enforce categories to be crossed only when this really leads to a more optimal set of strata. More optimal in the present setting means that the new stratification gives a smaller value of (23), i.e. the absolute bias that is maximally possible is smaller for the new stratification.

Classification trees are constructed top-down. We let the root of the tree consist of all respondents. Hence, the starting point is only one stratum and the poststratification estimator reduces to the response mean. The first step, then, is a bisection of all respondents into two disjoint groups, so-called nodes. In each subsequent step one of the nodes is selected and split again into two disjoint groups. This process is repeated until no more node is allowed to be split. The end nodes, called leaves, will be the strata in the weighting of the response. The splits are made using classifiers, in our case the categories of auxiliary variables that are available for both respondents and non-respondents.

Next in order to run the classification we need two rules, a splitting rule and a stopping rule. In section 3.2 we propose splitting and stopping rules and set up the classification algorithm.

3.2 The splitting rule, stopping rule and classification algorithm

In the previous section we argued that classification should be directed at the width of the bias interval. We want to split that stratum using that classifier or category of an auxiliary variable that gives the largest decrease in (23).

> Splitting rule:
>
> Choose that bisection of a node that leads to the largest decrease in the width of the bias interval of the corresponding poststratification estimator.

Note that in making a bisection we optimise over all current nodes and over all classifiers. The splitting rule is, therefore, a global rule, while for some other classification tree methods the splitting rule is local. The latter means that the bisection of a node is executed independent of the other nodes. For each node it is decided whether it will be split and optimisation is over the classifiers only.

It is less obvious how we should choose a stopping rule. We propose four rules of which the first rule plays the dominant role.

Stopping rules:

1. A split is not allowed if the *p*-value corresponding to the standardised decrease in interval width is larger than α.
2. The maximum number of leaves is K.
3. A node cannot be split if the number of respondents in that node is smaller than R_1.
4. A candidate node cannot be formed if the number of respondents in that node is smaller than R_2.

The last three rules are straightforward. We may want to limit the number of strata and the number of respondents in the strata. The first stopping rule, however, needs further explanation.

If we forget for a moment the other three stopping rules, then it is clear that we can split every stratum until only strata with one respondent are left. However, we only want to create a new set of strata in case Δw^* is systematic and not caused by taking a sample out of the population. We want to distinguish those strata that lead to a decrease of the width of the bias interval at a predefined significance level. We, therefore, must take the variance of Δw^* into account.

If we let $s_{\Delta w^*}$ be an estimator for the standard deviation of Δw^*, then we want the standardised decrease in the bias interval width

$$\Delta w_s^* = \frac{\Delta w^*}{s_{\Delta w^*}} \tag{24}$$

to be significantly smaller than zero. Clearly, we do not know the true probability distribution of the standardised bias decrease in (24). We, therefore, assume that the distribution of Δw_s^* can be approximated by a standard normal distribution and we let the *p*-value in the first stopping rule be the left quantile of the standard normal distribution.

We estimate the standard deviation $s_{\Delta w^*}$ by the jackknife estimator, see Miller (1974). All sample units are omitted once and Δw^* is computed based on the sample minus the omitted sample unit. This leads to n estimates. The jackknife estimator is the standard deviation of those estimates corrected for dependence between the estimates. One may, however, also resort to other variance approximation methods like bootstrap.

The classification algorithm becomes:

1. Select the classifier that produces the largest decrease Δw^* when splitting the total population. Let $k=1$ and go to step 2.
2. Make the split unless the p-value corresponding to Δw_s^* is larger than α. If the p-value is smaller then let $k := k+1$ and go to step 3. Otherwise stop.
3. Go to step 4 if $k < K$. Otherwise stop.
4. For each of the leaves with more than R_1 respondents select the classifier that produces the largest decrease Δw^*. If all k leaves are smaller than R_1 respondents, then stop. Otherwise go to step 5.
5. Remove all candidate splits that lead to at least one node with less than R_2 respondents. If no candidate splits are left, then stop. Otherwise go to step 6.
6. Select the best split and go to step 2.

Ideally, we should also like to use (24) in the splitting rule instead of the unstandardised Δw^*. Unfortunately, this leads to much longer computation times as the jackknife estimate for the standard deviation needs to be estimated for all classifiers and all nodes. Therefore, we have chosen only to approximate standard deviations in case a candidate split is selected.

4. Results

We apply the algorithm of section 3.2 to the Dutch Integrated Survey on Household Living Conditions (Permanent Onderzoek Leefsituatie in Dutch) for the years 1998 and 2002. We will abbreviate the survey by its Dutch acronym POLS. POLS is a large continuous survey with questions about issues like health, social participation and recreational activities.

The survey is modular and consists of a base questionnaire and a number of questionnaires that deal with one separate topic. The base questionnaire is to be filled in by all persons. However, each person only fills in one topical questionnaire. The base questionnaire contains general questions and a number of basic questions that are used for allocation of the topical questionnaires. These basic questions are also used in weighting models for the topical questions. Here, we will focus on questions from the base questionnaire.

The survey is a two-stage sample, in which the clusters in the first stage are formed by municipalities. From the clusters simple random samples without replacement are drawn consisting of persons. The first-order inclusion probabilities differ only for age. All persons of 12 years and older have the same probability to end up in the

sample. In this paper we regard all persons of 12 years and older and omit only the nonresponse due to frame errors. The 1998 and 2002 samples then consist of, respectively, 36136 persons and 39170 persons.

The 1998 and 2002 POLS had a fieldwork period of two months. In 1998 the first month was CAPI, and the second month was a mixture of CAPI and CATI. In 2002 both months were CAPI. After two months the size of the response was 21571 persons in 1998 and 22259 persons in 2000, i.e. a response rate of respectively 60% and 57%.

We selected two survey questions from the POLS questionnaire, namely whether a person owns a personal computer or laptop and whether a person is active in sports. We also selected one auxiliary variable, whether a person receives some form of social allowance (disability, unemployment, social security), and treated this variable as if it was a survey question.

To the survey we linked demographic and regional variables, information about jobs and social allowances and fieldwork information. We refer to the appendix for an overview of the variables used in the classification. All non-categorical variables like age and the average value of house in the postal code area were made categorical. The dummy-variables corresponding to the categories of the auxiliary variables were used as classifiers in the algorithm. However, one dummy-variable was omitted for every variable for redundancy reasons.

We divided all auxiliary variables into two groups, nominal and ordinal, and treated the two groups differently in the classification. The categories of ordinal variables are arranged according to some natural order, e.g. degree of urbanisation or age in classes. The categories of nominal variables lack any order.

In case of ordinal variables we did allow categories to be clustered as long as the clustering followed the ordering. For example, degree of urbanisation has five classes: 1) very strong, 2) strong, 3) moderate, 4) little and 5) not. Examples of classifiers that are allowed are for instance 1 to 4, 2 to 3 and 3 to 5. However, 1 to 2 plus 4 is not allowed, since category 3 is missing.

The decision to allow clustering of ordinal variables was motivated by the fact that without clustering some ordinal variables were only rarely used as classifiers. The number of categories of some of the available ordinal variables is quite large and the average number of respondents per category small. As a consequence these categories are less interesting in the formation of substrata. As we did not want to form clusters ourselves, we decided to let the algorithm search over all possible clusters.

Figures 1 to 3 show the classification trees for the three selected variables. The trees contain, respectively, 33, 27 and 33 nodes, and 17, 14 and 17 leaves. The leaves are used as strata in the poststratification estimator. Table 1 shows the values of w^*, Δw^*, $s_{\Delta w^*}$ and Δw_s^* for the first 17 iterations of the classification algorithm corresponding to figure 1.

We take figure 3 as an example. Whenever a node is split, the classifier is attached to the node. The labelling of the nodes indicate the order in which nodes where created in the classification tree algorithm. The root in figure 3 is split based on the question whether a person is between 55 and 64 years of age. All persons having an age in this interval go to node 2. All other persons go to node 3. Next node 3 is split into nodes 4 and 5 based on a further classification on age. The persons younger than 54 go to node 4, the persons older than 64 go to node 5. Node 4 is then split based on the question whether a person has a job. Node 5 is a terminal node and is not split. This node corresponds to the stratum 65 years and older. In the sixth iteration finally node 2 is split based again on having a job. Here node 12 is a stratum, and consists of all persons younger than 55 years that have a job.

When we move down along the branches of the tree, the nodes are based on an increasing number of classifications. Some of these may be nested, e.g. in figure 3 the variable age is twice used as a classifier. Consequently, the strata may be rather exotic when compared to usual weighting models.

Figure 1: The classification tree for ownership of a personal computer or laptop.

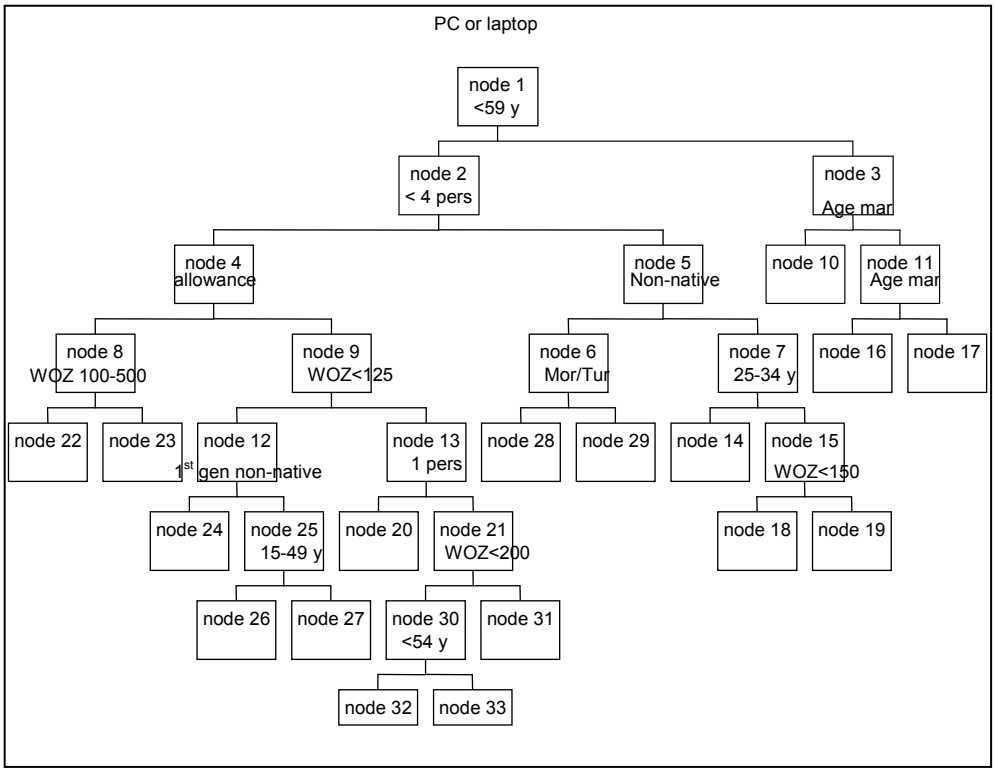


Figure 2: The classification tree for activity in sports.

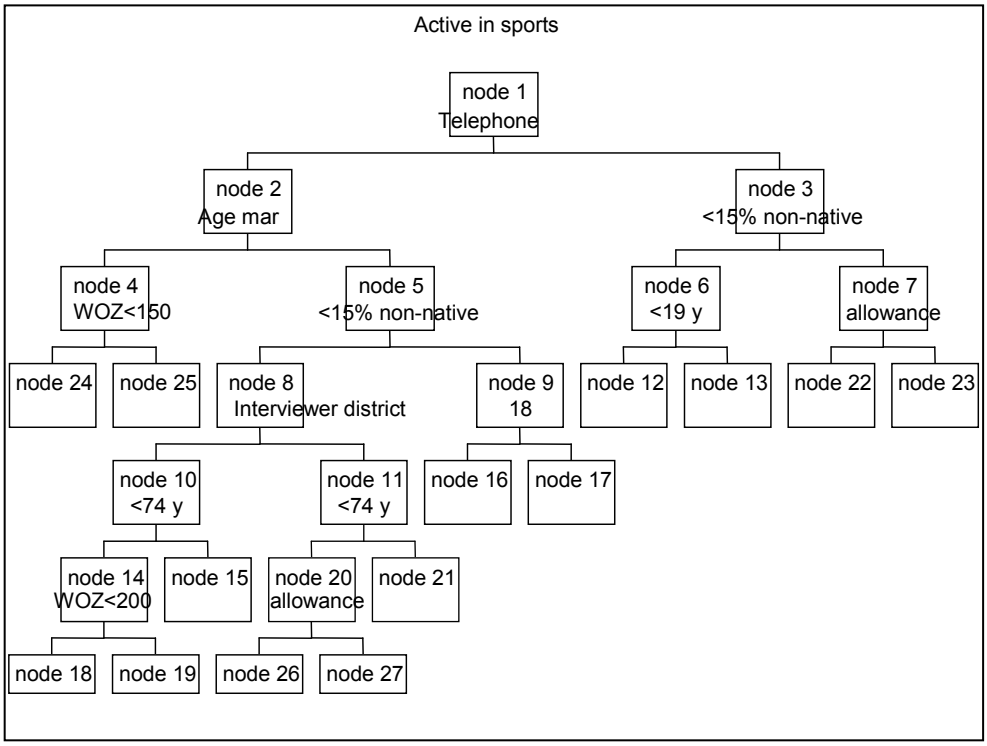
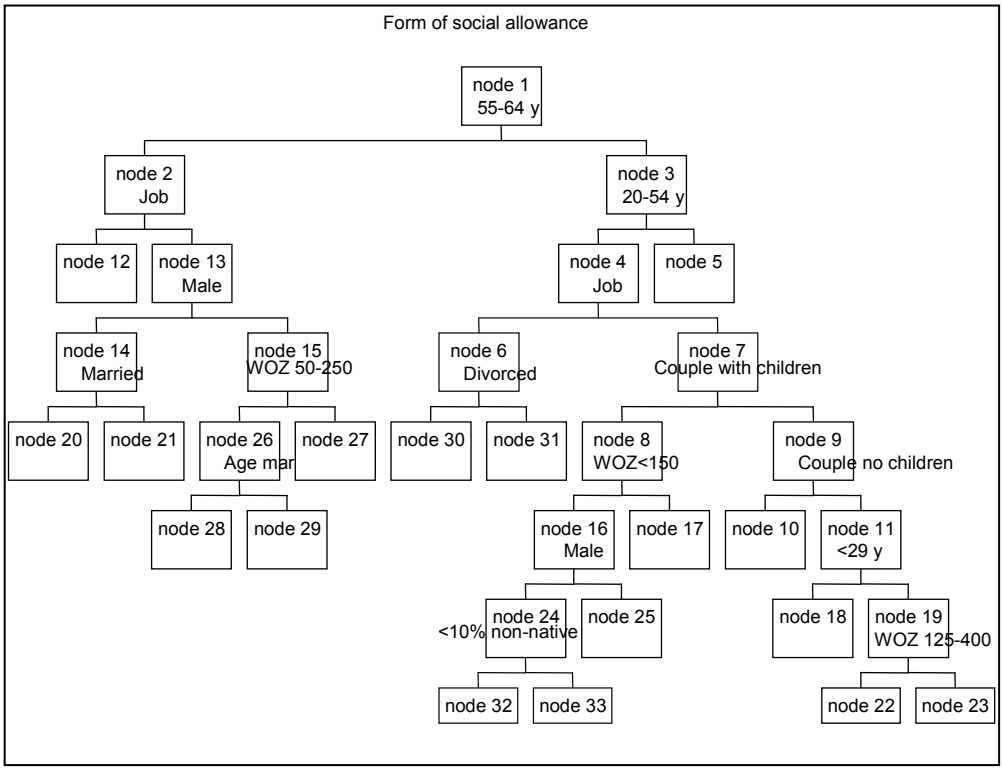


Figure 3: The classification tree for social allowance.



In table 1 we can see that the main decrease in the width of the bias interval for ownership of a personal computer or laptop is obtained in the first iteration. The other interations lead to much smaller decreases. The 17th interation is the first iteration that gives a value of Δw_s^* smaller than the 99%-quantile of the standard normal distribution which is approximately equal to 2.32.

In table 2 we compare the classification tree estimates to the estimates using the current POLS weighting model and the weighting model proposed in Schouten (2004). In Schouten (2004) an alternative forward inclusion – backward elimination algorithm is proposed, however, based on the same objective function, the width of the bias interval. In table 2 for the classification estimates also the 95%-confidence interval is constructed using the approximated jackknife standard deviations.

From table 2 we can see that the classification tree estimates do not deviate much from the estimates using the alternative model. Most differences fall within the 95%-confidence intervals. This is true in general for almost all survey questions that we investigated. The differences with the estimates of the current weighting model are in most cases somewhat larger.

Table 1: The values of w^*, Δw^*, $s_{\Delta w^*}$ and Δw_s^* for the first 17 iterations of the classification algorithm applied to the POLS survey variable ownership of a personal computer or laptop.

Iteration	w^*	Δw^*	$s_{\Delta w^*}$	Δw_s^*
1	0.879	0.121	0.0045	26.77
2	0.862	0.017	0.0013	13.43
3	0.854	0.008	0.0010	8.46
4	0.847	0.006	0.0010	6.73
5	0.841	0.006	0.0012	5.17
6	0.838	0.003	0.0008	3.91
7	0.835	0.003	0.0005	5.03
8	0.833	0.002	0.0006	4.23
9	0.831	0.002	0.0006	3.53
10	0.829	0.002	0.0004	4.30
11	0.828	0.001	0.0005	2.77
12	0.826	0.001	0.0003	4.84
13	0.825	0.001	0.0004	3.59
14	0.823	0.001	0.0004	3.07
15	0.822	0.001	0.0005	2.53
16	0.821	0.001	0.0003	2.79
17	0.820	0.001	0.0004	2.13

Table 2: The estimates according to the classification tree method, the current POLS weighting model and the weighting model proposed in Schouten (2004). For the classification tree estimates also the 95%-confidence interval is given that is computed using the jackknife approximations.

	Classification tree stratification	*Current POLS model*	*Forward – backward model*
Owner of PC	57.6% (± 0.6%)	58.3%	57.2%
Active in sports	44.9% (± 0.6%)	45.6%	45.1%
Social allowance	11.5% (± 0.4%)	11.0%	11.4%

We summarise some other findings related to optimality, stability and computing times of the classification algorithm:

- For some survey questions we were able to build trees by hand that correspond to a smaller value of the criterion function (23) than the trees built by the proposed algorithm. Differences are, however, very small and estimates only slightly change. Also, it was quite cumbersome and labour-intensive to find those more optimal trees. The changes in estimates are within the 95%-confidence interval.
- We divided the POLS sample into two and into 10 disjoint groups and applied the classification tree method to each of the subsamples. This way we could investigate the stability properties of the method. We found that the tree structures can be quite different even for the two halfsamples that have a size of approximately 18000 sample persons. However, the estimates computed with the different sets of strata were in most cases quite similar. We also applied the tree that followed from one halfsample to the other halfsample. Again differences were acceptable.
- The computation time of the algorithm can be quite long, up to 1 hour or more for the complete sample of 36000 persons and a set of approximately 165 classifiers. We must also remark that the categories of age (15 classes), average house value of the postal code area (12 classes), degree of urbanisation (5 classes) and proportion of non-natives in postal code area (9 classes) were allowed to form clusters. For instance, 24 clusters can be formed for the proportion of non-natives in postal code area apart from the 9 classifiers for each of the categories. Ideally, we would like the splitting rule to depend on the standardised difference in the width of the bias interval rather than the unstandardised width. However, computation times will in that case not be feasible in practice.

5. Discussion

We argue that the usual missing-at-random assumption is not always valid in surveys. However, even if the assumption is true for the final weighting model, we need a criterion to form strata or adjustment cells in non-response adjustment. In this paper the missing-at-random assumption is, therefore, not made. We show that general intervals can be set up for the bias of the response mean and the poststratification estimator. We propose to minimise the width of these intervals.

We have set up two bias intervals for the poststratification estimator, which we refer to as the local and the global bias interval. We do not have a proof that the global bias interval is valid. However, we argued that the width of this interval must be the same as the width of the bias interval for the response mean. Even if we know the true differences between the stratum means, then we showed that the width of the bias interval is not reduced. In this paper we minimised the width of the global bias interval. In the future we will investigate algorithms for the minimisation of the local bias interval.

Classification trees are candidate tools to form strata economically and in an automated way. Strata are divided into substrata only in case there is a significant decrease in interval width, leaving those strata alone that do not lead to any further decrease. Since the prediction of survey questions and the relation to response behaviour is combined in one splitting criterion, the strata can be formed by an automated algorithm. Hence, the classification tree method provides a tool to perform weighting in one step.

Approximations for the variance of the poststratification estimates come as a useful by-product of the jackknife-method. A proposed split of a tree node is executed only in case the decrease in interval width is significant. The jackknife-method is employed to approximate the variance of this decrease. However, at the same time the variance of the poststratification estimates can be computed while only marginally increasing the computation times.

There are also some drawbacks to the proposed classification tree method. First, the trees turn out not to be very stable. Even for two quite large samples the resulting trees may have quite different forms. However, due to multicollinearity in the variables the estimates are rather stable. In case the tree of one sample is applied to another sample, the estimates do not change much. Second, the computation times of the classification tree algorithm are considerable, since the number of nodes and splits to be investigated can become quite large. For practical purposes the current software is too slow and need to be made more sophisticated. Third, the classification gives a set of strata for each survey question and it does not seem straightforward how to combine those sets into one set of strata that suit all survey questions. Fourth, we found that the algorithm is not optimal. Examples can be constructed where trees exist that give a smaller bias interval. In most cases these trees can be formed by choosing splits in the first iterations that are close to not being significant.

Summarising, we distinguish the following advantages and disadvantages:

Advantages:

- The selection of auxiliary variables can be done in one step.
- The construction of weighting models can easily be automated.
- The formation of strata is economical.
- The variance of the poststratification estimator can be approximated synchronically.

Disadvantages:

- Computation times are considerable.
- The stability of the tree structures is poor.
- The combination of sets of strata corresponding to different survey questions is not straightforward.
- The proposed algorithm is suboptimal.

We begin with the last disadvantage. We belief the suboptimality of the algorithm to be marginal. Only after careful and labour-intensive analysis we were able to construct trees that correspond to a slightly more optimal set of strata.

The computation times may be shortened by more efficient programming and using more specialised software. The routines for the classification trees were written and coded in S-plus by the authors.

The stability of the resulting trees is a more difficult problem. This problem may be solved by methods that create ensembles of classification trees, see e.g. Bauer and Kohavi (1999) and Dietterich (2000). A promising technique in this respect is the Random-Forest method developed by Breiman (2001). Future research is necessary to investigate whether such methods can improve the stability of the classification.

Finally, the combination of weighting models for different survey questions may be circumvented by the use of a multidimensional splitting rule. Instead of splitting the population separately for each survey question, we may choose the node and classifier that correspond to the largest decrease in the interval width over all survey questions. Also, this question needs further research.

References

Bauer, E., Kohavi, R. (1999), An empirical comparison of voting classification algorithms: bagging, boosting, and variants, *Machine Learning* 36, 105–139.

Bethlehem, J.G. (2002), Weighting nonresponse adjustments based on auxiliary information, In *Survey Nonresponse (ed. R.M. Groves, D.A. Dillman, J.L. Eltinge, R.J.A. Little)*, 275–288, Wiley Series in Probability and Statistics, John Wiley & Sons, New York, NY, USA.

Breiman, L. (2001), Random forests, Technical paper, Statistics Department, University of California, Berkeley, CA, USA. Available via http://www.stat.berkeley.edu/users/breiman .

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. (1984), *Classification and Regression Trees*, Chapmann & Hall,Boca Raton, FL, USA.

Cochran, W.G. (1977), *Sampling Techniques (3rd edition)*, Wiley Series in Probability and Mathematical Statistics–Applied, John Wiley & Sons, New York, NY, USA.

Dietterich, T.G. (2000), An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting and randomization, *Machine Learning* 40, 139–157.

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

Kass, G.V. (1980), An explanatory technique for investigating large quantities of categorical data, *Journal of the Royal Statistical Society C, Applied Statistics* 29, 119–127.

Little, R.J.A. (1986), Survey Nonresponse Adjustments for Estimates of Means, *International Statistical Review*, 54, 139–157.

Mesa, D.M., Tsai, P., Chambers, R.L. (2000), Using tree-based models for missing data imputation : An evaluation using UK census data, Technical paper AUTIMP-project, University of Southampton, UK.

Miller, R.G. (1974), The Jackknife – a Review, *Biometrika*, 61, 1–15.

Morgan, J.A., Sonquist, J.N. (1963), Problems in the analysis of survey data, and a proposal, *Journal of the American Statistical Association* 58, 415–434.

Murthy, S.K. (1998), Automatic construction of decision trees from data: A multidisciplinary survey, *Data Mining and Knowledge Discovery* 2, 345–389.

Nooij, G. de (2004), A classification tree method for the construction of weighting strata, Thesis, Hogeschool van Amsterdam, Amsterdam, The Netherlands.

Schouten, J.G. (2002), Grenzen voor de correlatie tussen een doelvariabele en de responsindicator in enquêtes, Research paper 0225, Sector Methoden en Ontwikkeling, CBS, Voorburg.

Schouten, J.G. (2003), Adjustment for bias in the Integrated Survey on Living Conditions (POLS) 1998, Paper presented at the 14th International Workshop on Household Survey Nonresponse, August 2003, Leuven, Belgium.

Schouten, J.G. (2004), A selection strategy for weighting variables under a Not-Missing-at-Random assumption, paper submitted to journal.

Appendix: Auxiliary variables in the classification trees

Here, we give an overview of the auxiliary variables that are used in the classification tree algorithm applied to POLS surveys of 1998 and 2002.

Demographic (personal level):

Gender, age, ethnic origin (native, Moroccan, Turkish, Surinam, other non-western non-native, other western non-native), *ethnic generation* (native, 1st generation, 2nd generation and one parent not born in The Netherlands, 2nd generation and both parents not born in The Netherlands), *marital status, nationality, country of birth*.

Demographic (household level):

Children living in the household, household type (single, single parent, couple, couple with children, other type), *household size*.

Regional level:

Degree of urbanization, province in the Netherlands (separate categories for four largest cities), *size of municipality, average value of houses at postal code area, proportion non-native in postal code area.*

Income and allowance:

Job, old-age pension, disability allowance, unemployment benefit, social security.

Fieldwork information:

Interviewer district, interviewer seniority, interviewer gender.