The Selection of Strata in Nonresponse Adjustment

Barry Schouten (Statistics Netherlands)

1. Introduction

Nonresponse in household surveys can be a threat to the quality of statistics. Research shows that often the response to these surveys is selective with respect to demographic characteristics like age and household composition. For this reason estimators are usually adjusted to account for nonresponse. The Bascula tool is added to Blaise in order to support nonresponse adjustment.

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. First, candidate covariates are selected that relate either to the nonresponse mechanism or to the key survey topics. Second, adjustment is performed using these candidate covariates using for instance Bascula.

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.

The method is computationally intensive but it facilitates nonresponse adjustment in a single step. The method may be included in Blaise to combine selection of strata and weighting in one tool.

In section 2 we give some background to the selection problem. In section 3 we discuss results. Section 4 concludes.

2. The selection of strata using classification trees

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 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.

Here we do not make assumptions about the missing data mechanism. 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.

Schouten (2005a and b) 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. 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.

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.

However, we still need a strategy to select strata based on the criterion above. 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. This method enables adjustment for nonresponse in a single step as the resulting classification tree represents the stratification to be used in the weighting of the response.

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 bias interval, 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.

In our case we choose the following splitting and stopping rules, where α, K, R_1 and R_2 are prespecified parameters:

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.

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.

For details about the methodology and algorithms we refer to Schouten (2005a and b).

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). 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.

3. Results

We apply the classification tree method 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.

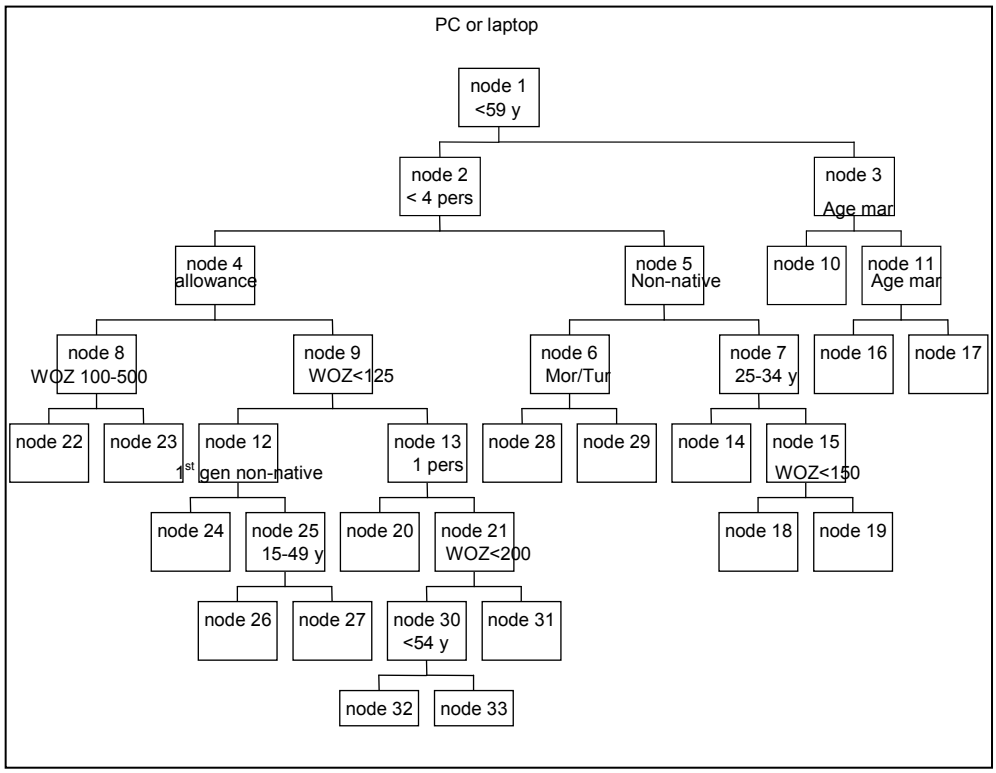
The survey is a two-stage sample, in which the clusters in the first stage are formed by municipalities. From the clusters simple random samples without replacement

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 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 one survey question from the POLS questionnaire, namely whether a person owns a personal computer or laptop. 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. The auxiliary variables that we used are gender, age, ethnic origin, ethnic generation, marital status, children living in the household, household type, household size, 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, job, old-age pension, disability allowance, unemployment benefit, social security, interviewer district, interviewer seniority and interviewer gender. 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.
Figures 1 and 2 show the classification trees for the two selected variables. The trees both accidentally contain 33 nodes and 17 leaves. The leaves are used as strata in the poststratification estimator.

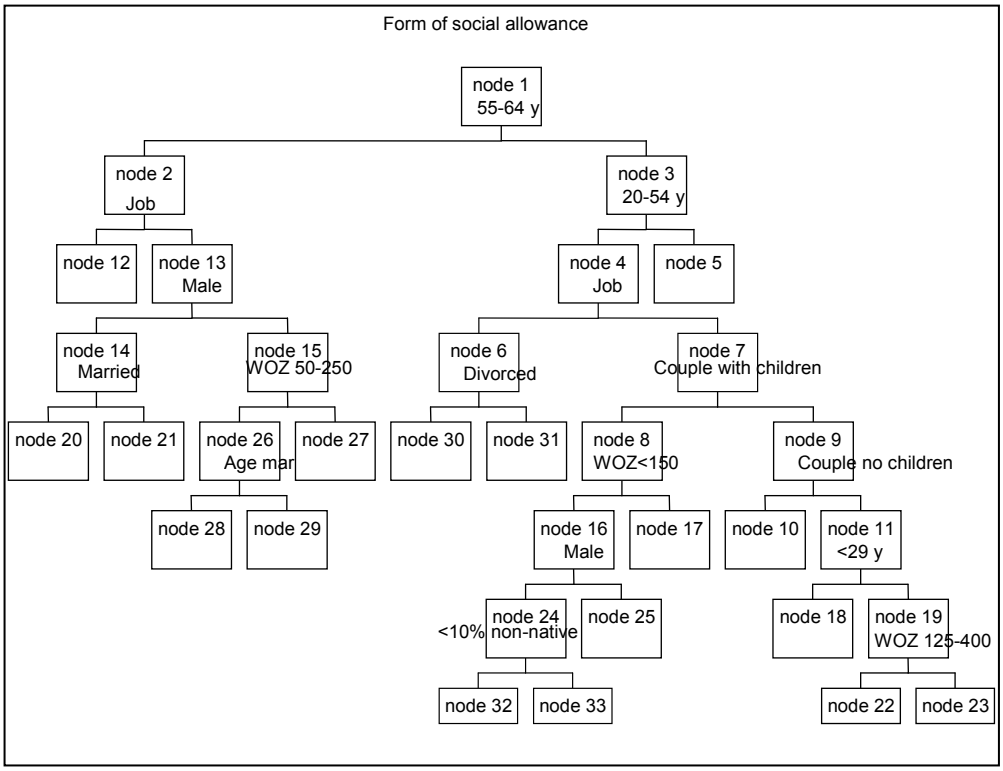
Figure 1: The classification tree for ownership of a personal computer or laptop.



We take figure 2 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 2 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 2 the variable age is twice used as a classifier. Consequently, the strata may be rather exotic when compared to usual weighting models.

Figure 2: The classification tree for social allowance.



In table 1 we compare the classification tree estimates to the estimates using the current POLS weighting model. In table 1 for the classification estimates also the 95%-confidence interval is constructed using approximated jackknife standard deviations.

Table 1: The estimates according to the classification tree method and the current POLS weighting model. 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*
Owner of PC	57.6% (± 0.6%)	58.3%
Social allowance	11.5% (± 0.4%)	11.0%

From table 1 we can see that the classification tree estimates do deviate from the estimates using the alternative model. Most differences fall, however, within the 95%-confidence intervals. This is true in general for almost all survey questions that we investigated. Furthermore, we must remark that the current POLS weighting model does not contain all the auxiliary variables that were used in the classification tree method. In case these variables are added, then differences are very small. Importantly, the number of strata in the classification tree stratification is much smaller than that in the current weighting model.

4. Conclusions

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.

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 tree structure turns 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. We propose to implement the algorithm in existing software like Bascula.

The stability of the resulting trees is a more difficult problem. However, to our opinion it is not a serious problem as long as estimates are stable. A promising technique in this respect is the Random-Forest method developed by Breiman (2001). Vanhommerig (2005) investigated the use of this technique for our purposes.

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.

5. References

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.

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.

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.

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. (2005a), A Selection Strategy for Weighting Variables under a Not-Missing-at-Random Assumption, to appear in the Journal of Official Statistics.

Schouten, J.G., Nooij, G. de (2005b), Nonresponse Adjustment using Classification Trees, Research paper 05001, Methods and Informatics Department, Statistics Netherlands.

Vanhommerig, R. van (2005), Nonresponse Adjustment using Random Forests, Masters thesis, Mathematics Department, Universiteit van Utrecht.