Adjustment for Bias in the Integrated Survey on Household Living Conditions (POLS) 1998

Discussion paper 04001

Barry Schouten

Statistics Netherlands

Voorburg/Heerlen, March 2004

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
2003–2004	= 2003 to 2004 inclusive
2003/2004	= average of 2003 up to and including 2004
2003/'04	= crop year, financial year, school year etc. beginning in 2003 and ending in 2004

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: 6008304001

Statistics Netherlands

ADJUSTMENT FOR BIAS IN THE INTEGRATED SURVEY ON HOUSEHOLD LIVING CONDITIONS (POLS) 1998

Summary: Estimates for population statistics can be seriously biased in case response rates are low and the response to a survey is selective. Methods like poststratification or propensity score weighting are often employed in order to adjust for bias due to nonresponse.

One problem that many adjustment methods have in common is the choice which of the available auxiliary variables are used. For poststratification, it must be decided what strata are defined. For propensity score weighting, groups must be formed that have comparable response probabilities in order to keep a low variance.

In this report we use the generalised regression estimator to adjust for nonresponse and propose a selection strategy of weighting variables that accounts for the relation with response behaviour and the relation with survey questions simultaneously. The selection strategy is applied to the Integrated Survey on Household Living Conditions (POLS) 1998.

We make a weaker assumption than the usual Missing-at-Random (MAR) assumption. We assume that stratum means may be biased by nonresponse, but that the relative distances between the stratum means are preserved.

Keywords: Bias, bias adjustment, household survey, nonresponse, weighting model, poststratification, POLS

1. Introduction

Nonresponse can affect the quality of estimates in case nonrespondents are different from respondents for the topics of a survey. This threat may be serious if the size of the nonresponse is large relative to the sample size. In the Netherlands response rates are often low. Adjustment methods for potentially selective response, therefore, play an important role in improving quality.

Adjustment methods are based on auxiliary information from population databases, censuses and registers. This information may be available on the population level or on the individual level. In the first, the distribution of for instance age, gender and marital status is known for the target population of the survey. More ideally, however, the auxiliary information is available on person or household level and can be linked directly to the sample. In this report we assume the latter situation, i.e. auxiliary variables can be linked to both respondents and nonrespondents.

Under nonresponse we may distinguish three groups of variables, namely the survey questions of interest, the auxiliary variables that are linked from external sources and the response indicator. The response indicator is a 0-1-variable indicating whether a sampled person responded or not and stands by itself. In the survey literature a lot of research has been devoted to the relation between the response indicator and auxiliary variables. Furthermore, it is known by long that weighting methods using auxiliary variables that are correlated with the important survey questions may considerably reduce the variance of estimators. As a consequence also the relation between survey questions and auxiliary variables has been analysed. The only relation, however, that is not and can not be investigated directly is the relation between the response indicator and the target variables of the survey.

In the literature various adjustment methods are given that incorporate auxiliary variables. For a recent overview and a comparison of methods, see Kalton and Flores-Cervantes (2003). An estimator that is often used is the generalised regression estimator modified to nonresponse, see Bethlehem (1988). When using only crossings of categorical auxiliary variables this estimator reduces to poststratification. The population is divided into a number of subpopulations, the so-called strata, and the missing answers of the nonrespondents are predicted by the 'average' answers of the respondents in the same stratum. Another method that is often used is propensity score weighting. This technique was introduced by Rosenbaum and Rubin (1983) in the setting of studies for causal effects. In the nonresponse setting the propensity score is the response probability, i.e. the probability that a person or household responds given that he or she is selected in the sample. The response probability is usually fitted by means of a logit or probit model. The answers by the respondents are weighted by the inverse of their estimated response probabilities.

One problem that many adjustment methods have in common is the selection of informative auxiliary variables. In the case of poststratification it must be decided how strata are defined. In the case of propensity score weighting groups must be formed that have comparable response probabilities in order to keep a low variance. Since there are three groups of variables involved, the choice of strata or response groups is often performed in two steps. In the first step auxiliary variables are selected that explain the response indicator. In the second step a further selection of variables is performed in which variables are chosen that also relate to the important target variables of the survey. Little (1986) proposes to form so-called adjustment cells by modelling the response probability, forming response groups and clustering response groups based on the differences between the 'average' answers to the survey questions. See also Rosenbaum and Rubin (1984), Ekholm and Laaksonen (1991) and Czajka et al. (1992). Eltinge and Yansaneh (1997) compare several criteria for the formation of adjustment cells. Geuzinge, Van Rooijen and Bakker (2000) propose to use the product of the correlation between the response indicator and the auxiliary variables and the correlation between a target variable and the auxiliary variables as a measure for the relevance of auxiliary variables in a weighting model.

Crucial in the adjustment for nonresponse are the assumptions that are made about the nonresponse or missing data mechanism. The nonresponse mechanism is called

Missing Completely at Random (MCAR) whenever the probability of response is independent of the survey questions. In case the probability of response is independent of the survey questions when conditioned on a set of auxiliary variables, the mechanism is called Missing at Random (MAR). For most surveys the MCAR assumption does not hold for the auxiliary variables. In practice it is usually assumed that the nonresponse can be made ‘sufficiently’ MAR by incorporation of the available auxiliary variables in a weighting model.

In this report we use the generalised regression estimator to adjust for nonresponse and we select strata by minimising the maximal absolute bias under a weaker assumption than MAR. We assume that stratum means all have the same bias, i.e. the relative distances between the stratum means are preserved under nonresponse. Schouten (2003a) gives an interval for the maximal absolute bias and proposes to use those strata that minimise the width of the interval for the bias. Furthermore, he shows that the poststratification estimator shifts the response mean approximately by a value equal to the center of the interval.

As a criterion we use the absolute bias of an estimator and we employ poststratification to recover sample means. In the analysis no population totals or means are used. Thus the response is calibrated to the sample for the auxiliary variables that are selected in the weighting model. Additional weights may be used to further calibrate the sample to the population. However, we did not perform this extra step. We concentrate on the bias because we believe that nonresponse affects especially the location of means and not their variation. We believe that variance reduction is most effective in the calibration from sample to population.

Since the absolute bias falls within an interval of known form but unknown size, the criterion of absolute bias reduces to minimisation of the absolute bias that is maximally possible. In other words an estimator is to be favoured to another estimator in case it corresponds to a smaller interval.

We apply the selection strategy to the 1998 Integrated Survey on Household Living Conditions abbreviated to POLS (Permanent Onderzoek Leefsituatie) in Dutch. We only regard the persons of 12 years and older.

In section 2 we first give some theoretical background and introduce the selection strategy in more detail. In section 3 we give results for POLS 1998. Finally, in section 4 we discuss the outcomes.

2. The selection of weighting variables

In this section we briefly summarise the selection strategy described in Schouten (2003a). In section 2.1 we first introduce some notation. Next, in section 2.2 we derive the interval for the maximal absolute bias under this assumption. In section 2.3 we turn to the criteria for selecting auxiliary variables in the poststratification estimator.

2.1 Notation

Let n be the size of the sample and r be the size of the response. Furthermore, let y_i be the value of a survey question for individual i and $x_i = (x_{1,i}, x_{2,i}, \ldots, x_{p,i})'$ be a vector of p auxiliary variables for individual i that are linked from some external source.

We assume that the auxiliary variables are linearly independent, i.e. there exists no vector c so that $c' x_i = 0, \forall i$. We also assume that x_i does not contain excess information. In other words, there exists no vector c so that $c' x_i = 1, \forall i$. Implicitly, the second assumption means that none of the auxiliary variables is a constant.

By r_i we denote the response indicator of individual i, i.e. taking the value 1 if a person did respond and 0 otherwise, with expectation $E(r_i) = \rho_i$. Here, ρ_i is the individual response probability. We assume that the individual response probabilities $\rho_i \in [0,1]$ are unknown.

In the following, we assume that at least two persons or households responded, so that response means and variances can be calculated. We will distinguish response-based statistics from sample-based statistics using an asterisk as extra index.

Let for some variable z, $\bar{z}$ be the sample mean and $\bar{z}^*$ be the response mean, i.e.

$$\bar{z} = \frac{1}{n} \sum_{i=1}^{n} z_i \tag{1}$$

$$\bar{z}^* = \frac{1}{r} \sum_{i=1}^{n} r_i z_i \ . \tag{2}$$

Let S_Z^2 and S_Z^{*2} be, respectively, the sample variance and the response variance of variable z

$$S_Z^2 = \frac{1}{n-1} \sum_{i=1}^{n} (z_i - \bar{z})^2 \tag{3}$$

$$S_Z^{*2} = \frac{1}{r-1} \sum_{i=1}^{n} r_i (z_i - \bar{z}^*)^2 \ . \tag{4}$$

Furthermore, C_{Z_1,Z_2} and $C^*_{Z_1,Z_2}$ denote the sample and response covariance

$$C_{Z_1,Z_2} = \frac{1}{n-1} \sum_{i=1}^{n} (z_{1,i} - \bar{z}_1)(z_{2,i} - \bar{z}_2) \tag{5}$$

$$C^*_{Z_1,Z_2} = \frac{1}{r-1} \sum_{i=1}^{n} r_i (z_{1,i} - \bar{z}_1^*)(z_{2,i} - \bar{z}_2^*) \ . \tag{6}$$

Analogously, Γ_{Z_1,Z_2} and $\Gamma^*_{Z_1,Z_2}$ will be used for, respectively, the sample and response correlation between z_1 and z_2.

Finally, the vector of covariances between the response probabilities and the auxiliary variables x is $C_{\rho,X}$. Since the response probabilities ρ_i are unknown, estimators for $C_{\rho,X}$ can only be based on the response indicators r_i. Let

$$C_{\rho,X} = \frac{1}{n-1}\sum_{i=1}^{n}(r_i - \bar{r})(x_i - \bar{x}). \tag{7}$$

The *generalised regression estimator* $\bar{y}^*_{gr}$ for the sample mean of y using auxiliary variables x, see for instance Bethlehem (1988), is

$$\bar{y}^*_{gr} = \bar{y}^* + (\bar{x} - \bar{x}^*)' \beta^*_{X,Y}, \tag{8}$$

with $\beta^*_{X,Y}$ the regression vector

$$\beta^*_{X,Y} = (S^{*2}_X)^{-1} C^*_{X,Y}. \tag{9}$$

Note that in (8) we calibrate to the sample mean and not to the population mean. The two assumptions at the beginning of this section prevent $(S^{*2}_X)^{-1}$ from being singular.

2.2 Maximal absolute bias

We will use absolute bias as a criterion for the selection of a weighting model. In this section we give intervals for the bias of the response mean and the generalised regression estimator.

In Schouten (2003a) the following interval for the population correlation between the target variable y and the response probabilities ρ is given

$$\Gamma_{\rho,Z}\Gamma_{Z,Y} - \sqrt{1-\Gamma^2_{\rho,Z}}\sqrt{1-\Gamma^2_{Z,Y}} \le \Gamma_{\rho,Y} \le \Gamma_{\rho,Z}\Gamma_{Z,Y} + \sqrt{1-\Gamma^2_{\rho,Z}}\sqrt{1-\Gamma^2_{Z,Y}}, \tag{10}$$

for any auxiliary variable z.

The bias of the response mean is approximately

$$B(\bar{y}^*) = E(\bar{y}^*) - \bar{y} \cong \frac{C_{\rho,Y}}{\bar{\rho}}, \tag{11}$$

where $\bar{\rho}$ is the sample mean of the response probabilities. Hence, with (10)

$$\begin{aligned} &\frac{C_{\rho,Z}C_{Z,Y}}{S^2_Z\bar{\rho}} - \frac{S_\rho S_Y}{\bar{\rho}}\sqrt{1-\Gamma^2_{\rho,Z}}\sqrt{1-\Gamma^2_{Z,Y}} \le B(\bar{y}^*) \\ &\qquad \le \frac{C_{\rho,Z}C_{Z,Y}}{S^2_Z\bar{\rho}} + \frac{S_\rho S_Y}{\bar{\rho}}\sqrt{1-\Gamma^2_{\rho,Z}}\sqrt{1-\Gamma^2_{Z,Y}}. \end{aligned} \tag{12}$$

Schouten (2003a) shows that under the assumption that (9) is a conditionally unbiased estimator for the regression vector in the sample, i.e.

$$E(\beta^*_{X,Y}) = \beta_{X,Y} = (S_X^2)^{-1} C_{X,Y}\,, \tag{13}$$

the bias of the generalised regression estimator can be bounded approximately by

$$\begin{aligned}\frac{C_{\rho,Z}C_{Z,Y}}{S_Z^2\overline{\rho}} - \frac{C_{\rho,X^{'}\beta_{X,Y}}C_{X^{'}\beta_{X,Y},Y}}{S^2_{X^{'}\beta_{X,Y}}\overline{\rho}} - \frac{S_\rho S_Y}{\overline{\rho}}\sqrt{1-\Gamma^2_{\rho,Z}}\sqrt{1-\Gamma^2_{Z,Y}} \le B(\overline{y}^*_{gr}) \\ \le \frac{C_{\rho,Z}C_{Z,Y}}{S_Z^2\overline{\rho}} - \frac{C_{\rho,X^{'}\beta_{X,Y}}C_{X^{'}\beta_{X,Y},Y}}{S^2_{X^{'}\beta_{X,Y}}\overline{\rho}} + \frac{S_\rho S_Y}{\overline{\rho}}\sqrt{1-\Gamma^2_{\rho,Z}}\sqrt{1-\Gamma^2_{Z,Y}}\,,\end{aligned} \tag{14}$$

for any variable z. Comparing (14) with (12) it can be seen that the generalised regression estimator produces a shift of the interval for the bias.

Now, we return to the criterion of absolute bias. The maximal absolute bias of the regression estimator is half the width of (14) taking $z = x^{'}\beta_{X,Y}$. Consequently, we must search for the vector of auxiliary variables that minimises

$$2\frac{S_\rho S_Y}{\overline{\rho}}\sqrt{1-\Gamma^2_{\rho,X^{'}\beta_{X,Y}}}\sqrt{1-\Gamma^2_{X^{'}\beta_{X,Y},Y}}\,. \tag{15}$$

Since $2\frac{S_\rho S_Y}{\overline{\rho}}$ is independent of the choice of auxiliary variables, it suffices to minimise

$$w(x) = \sqrt{1-\Gamma^2_{\rho,X^{'}\beta_{X,Y}}}\sqrt{1-\Gamma^2_{X^{'}\beta_{X,Y},Y}}\,. \tag{16}$$

Finally, let $w^*(x)$ be the estimator for $w(x)$ based on the response

$$w^*(x) = \sqrt{1-\Gamma^2_{\rho,X^{'}\beta^*_{X,Y}}}\sqrt{1-(\Gamma^*_{X^{'}\beta^*_{X,Y},Y})^2}\,. \tag{17}$$

2.3 The selection strategy

In the previous section we gave an approximate interval for the bias of the generalised regression estimator under the assumption that the regression parameter is an unbiased estimator of the true regression parameter. Here, we elaborate on the implications of this assumption and we introduce a strategy for minimising (17).

In case an auxiliary variable is categorical and x_i consists of the 0-1-dummy variables for the categories of this variable, then regression estimation reduces to poststratification, and the regression parameters equal the response means of the strata. The intercept or constant term in the regression estimator is zero. The MAR assumption implies that the response means of the strata are unbiased estimates of the true stratum means.

In section 2.1 we assumed that x_i does not contain excess information. This means that for every categorical variable at least one dummy must be omitted. Suppose we include one categorical variable in the weighting model and exclude the dummy variable corresponding to the last category. Now the intercept or constant term in the regression estimator does not vanish, and the regression parameters $\beta_{X,Y}$ equal the stratum response means minus the average of the stratum response means.

Hence, assumption (13) implies that the stratum means are not affected by nonresponse relative to the average stratum means. In other words, relative distances between the stratum response means are unbiased. However, it does not imply that the stratum response means themselves are unbiased. We only assume that all stratum response means are biased in the same direction. This assumption is weaker than MAR. We will, therefore, refer to this assumption as NMAR-R, since it is a Not-Missing-at-Random assumption and it concerns the relative distances between stratum means.

In practice it is often assumed that the MAR assumption holds whenever the weighting model contains a sufficient amount of auxiliary variables. In Schouten (2003b) it is shown, however, for two examples of auxiliary variables that stratum response means may still have excess bias. The two examples also show that the remaining bias of the stratum response means is mostly in one direction but not always of approximately the same size. Hence, neither the MAR assumption nor NMAR-R holds for these examples. The NMAR-R assumption is, however, weaker and closer to the observed effect of nonresponse.

Bias adjustment is carried out because response means are biased by selective response. However, why then should stratum response means not be biased. We believe that 'first-order derivatives' like distances between stratum response means are less affected by nonresponse than the means themselves. A more empirical motivation for the NMAR-R assumption is that often the construction of weighting models leads to a series of estimates that move in one direction but with decreasing stepsizes. The first variable produces the largest shift, the second variable a somewhat smaller shift but in the same direction, and so on. We, therefore, assume the NMAR-R assumption and seek for the vector of auxiliary variables that minimises (17).

We will first introduce some notation. Let the available auxiliary variables be labeled 1 to p and let $\mathrm{M} = \{1,2,\dots,p\}$ be the total set of auxiliary variables. For any two subsets $\mathrm{M}_1 \subseteq \mathrm{M}$ and $\mathrm{M}_2 \subseteq \mathrm{M}$ we let $\Delta w^*(\mathrm{M}_1,\mathrm{M}_2)$ denote the difference of the widths (17) between the two models, i.e.

$$\Delta w^*(\mathrm{M}_1,\mathrm{M}_2) = w^*(\{x_l\}_{l\in \mathrm{M}_1}) - w^*(\{x_l\}_{l\in \mathrm{M}_2}). \tag{18}$$

Furthermore, $\hat{s}_{\Delta w^*(\mathrm{M}_1,\mathrm{M}_2)}$ will represent an estimator for the standard deviation of (18) and $\xi_{1-\alpha}$ is the $100(1-\alpha)\%$-quantile of the standard normal distribution. We omit the two models in the index of the estimator unless it is not clear from the context which models are taken. Finally, the "empty" model is denoted by ϕ. (Note that $w^*(\phi)=1$.)

We will use the following forward-backward selection strategy that is similar to stepwise regression with forward inclusion and backward elimination:

1. Take the auxiliary variable k for which $\frac{\Delta w^*(\phi,\{k\})}{\hat{s}_{\Delta w^*}}$ is largest for all the auxiliary variables, but only if $\frac{\Delta w^*(\phi,\{k\})}{\hat{s}_{\Delta w^*}} > \xi_{1-\alpha}$. Set $i=1$ and let $\mathrm{M}_1 = \{k\}$. If a variable is added to the empty model go to step 2, otherwise go to step 4.
2. Add the auxiliary variable l for which $\frac{\Delta w^*(\mathrm{M}_i, M_i \cup \{l\})}{\hat{s}_{\Delta w^*}}$ is largest for all the remaining auxiliary variables, but only if $\frac{\Delta w^*(\mathrm{M}_i, M_i \cup \{l\})}{\hat{s}_{\Delta w^*}} > \xi_{1-\alpha}$, and let $\tilde{\mathrm{M}}_{i+1} = \mathrm{M}_i \cup \{l\}$. Otherwise, let $\tilde{\mathrm{M}}_{i+1} = \mathrm{M}_i$.
3. Remove auxiliary variable $m \in \mathrm{M}_i$ for which $\frac{\Delta w^*(\tilde{\mathrm{M}}_{i+1} \setminus \{m\}, \tilde{\mathrm{M}}_{i+1})}{\hat{s}_{\Delta w^*}}$ is largest, but only if $\frac{\Delta w^*(\tilde{\mathrm{M}}_{i+1} \setminus \{m\}, \tilde{\mathrm{M}}_{i+1})}{\hat{s}_{\Delta w^*}} > \xi_{1-\alpha}$, and let $\mathrm{M}_{i+1} = \tilde{\mathrm{M}}_{i+1} \setminus \{m\}$. Otherwise, let $\mathrm{M}_{i+1} = \tilde{\mathrm{M}}_{i+1}$.
4. If no auxiliary variable has been added or removed stop, otherwise repeat from step 2 with $i := i+1$.

The proposed selection strategy starts with a simple model with only one weighting variable, namely the variable that minimises (17). In the following steps iteratively variables are added and removed. Variables are only added or removed in case the difference in width (18) is larger than $\xi_{1-\alpha}$ estimated standard deviations. The significance level α may be chosen differently for the addition and removal step.

In the present form the selection strategy is not directly suited for categorical target variables y with more than two categories. The interval width in (15) is a vector in case y is a vector. If y is a nominal or ordinal variable with more than two categories, then correlations do not make much sense and y may better be represented by a vector of dummy variables for each category. The vector in (15) then contains the widths for all categories. If there are two categories, then it can be shown that the two intervals must have exactly the same size. However, in case there are more categories, the intervals have different sizes in general. Hence, we can only favour one auxiliary variable to another if we introduce some ordering, for instance by using the maximum or average width over the categories. For convenience we will use as a criterion the sizes of the shifts that are generated by the weighting model. In case of variables y with more than three categories, an auxiliary variable is to be preferred if it produces larger shifts relative to the response means. In the future we will replace this criterion by introducing an ordering in the proposed selection strategy.

We must make three additional remarks. First, in Schouten (2003a) it is proposed to use the vector b that minimises

$$w^*(x) = \sqrt{1-\Gamma_{\rho,X'b}^2}\sqrt{1-(\Gamma_{X'b,Y}^*)^2}, \tag{19}$$

instead of the regression parameter $\beta_{X,Y}^*$. The resulting interval (13) is in general smaller than the interval using the regression parameter. However, we do not have a closed form for the optimal b in (19) so that we are confined to numerical optimisation techniques. Since numerical optimisation can be quite time consuming and it can only be guaranteed that the optimisation leads to local minima, we decided not to search for the optimal b. Alternatively, we may also use the parameter that follows from regression of the response indicator on the auxiliary variables, i.e.

$$\beta_{\rho,X} = (S_X^2)^{-1} C_{\rho,X},$$

but we are interested in the prediction of the survey questions and not in the prediction of the response behaviour. We use the regression parameter (9) because poststratification also serves variance reduction.

The second remark concerns the calibration to sample means. Statistics about a population are gathered by means of a sample. Nonresponse reduces the sample. The sampling design, however, fixes the (first-order) inclusion probabilities of all individuals in the population. Hence, sample means may be biased with respect to population means whenever inclusion probabilities are not the same for all individuals, but the source of bias is known and can be corrected for. We believe that calibration to population means is especially interesting for variance reduction of estimators. When it comes to bias it is more interesting to calibrate to sample means. An individual needs to be sampled before a nonresponse can occur. Calibration to population means can be incorporated by using additional weights. Here we focus, however, on the effects of selective nonresponse.

Finally, we remark that from (12) and (14) it may be noticed that (in expectation) the poststratification estimator produces a shift of the bias interval. This shift involves the product of the correlations $\Gamma_{X'\beta_{X,Y}^*,Y}^*$ and $\Gamma_{\rho,X'\beta_{X,Y}^*}$. It is not true in general that the weighting model that produces the smallest width of interval (14) also produces the largest shift, since maximising $\Gamma_{X'\beta_{X,Y}^*,Y}^*\Gamma_{\rho,X'\beta_{X,Y}^*}$ is not the same as minimising (17). Take for example $\Gamma_{X'\beta_{X,Y}^*,Y}^* = 1$ and $\Gamma_{\rho,X'\beta_{X,Y}^*} = 0$. Geuzinge, van Rooijen and Bakker (2000) suggest to use the product of the correlations as a criterion for selecting auxiliary variables which comes down to maximising the shift.

3. Results

In this section we apply the proposed selection strategy to the 1998 Integrated Survey on Household Living Conditions (POLS). Section 3.1 gives some background to this survey. In section 3.2 we describe the set of available auxiliary variables and the survey questions that we selected for bias adjustment. The weighting models are given in section 3.3.

3.1 The Integrated Survey on Household Living Conditions (POLS) 1998

The Integrated Survey on Household Living Conditions is a large continuous survey with questions about issues like health, social participation, justice and recreational activities. In the following we will abbreviate the survey by its Dutch acronym POLS (Permanent Onderzoek LeefSituatie).

The survey is modular and consists of a base questionnaire and a number of questionnaires that deal with one seperate 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 report we regard all persons of 12 years and older and omit only the nonresponse due to frame errors. The sample then consists of 36136 persons.

The 1998 POLS has a fieldwork period of two months. The first month is CAPI, and the second month is a mixture of CAPI and CATI. After the first month the size of the response was 17039 persons. In the second month the response increased with 4532 persons to a total response of 21571 persons, i.e. a response rate of 59.7%. It is important to remark that we included also the addresses that were not processed ("onbewerkt retour") due to insufficient interviewer capacity.

For a more detailed description of POLS we refer to Vousten and De Heer (1998). The nonresponse in POLS 1998 has also been investigated by Beukenhorst (2001), Ter Haar (2001), Vollebregt (2002), Schmeets, Michiels and Verber (2002) and Bethlehem and Schouten (2003).

3.2 Target variables and auxiliary variables

From the POLS 1998 survey we selected six important target variables:

- Employment in three classes: employed for 12 hours or more per week, employed but less than 12 hours per week, or unemployed.
- Owner of a house: yes or no.

- At least one activity per month in a club (sports, music, etc.): yes or no.
- Owner of a pc or laptop: yes or no.
- Highest educational level successfully completed: primary (basis), junior general secondary (Mavo), pre-vocational (Vbo), senior general secondary/pre-university (Havo/Vwo), secondary vocational (Mbo), higher professional (Hbo) or university (Wo).
- Religion: none, Roman-Catholic, Protestant, Islamic and other.

We will apply the selection strategy of section 2.3 to this six target variables and to two auxiliary variables that we will treat as target variables, namely receiving a form of social allowance (yes or no) and ethnic origin (foreign or native).

In the bias adjustment we will make use of the following auxiliary variables: gender (male or female), age, marital status (not married, married, divorced or widowed), ethnic group (native, Moroccan, Turkish, Surinam, Netherlands Antilles/Aruba, other non-western non-native or other western non-native), ethnic generation (native, non-native 1st generation, non-native 2nd generation one parent or 2nd generation two parents), having a job (yes or no), receiving a form of social allowance (yes or no), province of residence, region in the Netherlands (north, east, west or south), children in the household (yes or no), household type (single, couple, couple with children, single parent or other), household size (1, 2, 3, 4 or >4), degree of urbanisation (5 levels), size of town (8 levels), interviewer district (27 districts), having a listed telephone number (yes or no), average value of houses in 6-digit postcode area and percentage of non-natives in 6-digit postcode area.

In section 3.3 we will use the variables *Woz (house value)*, *%foreign (percentage foreign)*, *age-15 (age)*, *age-ms (combination of age and marital status)* and *provplus (province and largest cities)*. For a description of the classifications used, see the appendix.

For a more detailed description of the available auxiliary information we refer to Bethlehem and Schouten (2003).

3.3 Selection of weighting models

We will apply the selection strategy to the variables of section 3.2. Table 1 gives the response means for the eight selected variables after the first month of fieldwork and after the full two months. The table shows that for most variables the response means after one month and after two months are approximately the same. The only exception is employment. The percentage of respondents that is employed for 12 hours or more increases with almost two percent.

In the complete sample 12.1% of the persons receives some form of social allowance and 15.0% is foreign. Hence the response means are quite different from the sample means and the differences have increased slightly in the second month of the survey. See also Beukenhorst (2001).

Next we will illustrate the selection strategy for the selected variables with two categories. Tables 2 a) to e) describe the selection process for these variables. The final weighting models are depicted by bold letters. In the tables the variables household type, houeshold size, ethnic group and marital status are abbreviated to *hhtype*, *hhsize*, *ethgr* and *mstat*. We use the jackknife method with group size 100 to estimate the standard deviations in the selection strategy, see Miller (1974), and take $\alpha = 0.01$ as the significance level for both additions and removals.

Table 1: The response means after one month and after two months for six target variables in POLS 1998 and for two auxiliary variables. For variables with two categories (yes/no) only the response mean for the answer 'yes' is given.

Variable	*After 1 month*	*After 2 months*
Employment	(48.6%, 7.4%, 44.0%)	(50.4%, 7.2%, 42.3%)
Owner of a house	63.0%	63.3%
Active in club	46.4%	46.6%
Owner of a pc	59.6%	59.8%
Educational level	(25.0%, 10.6%, 16.3%, 6.4%, 25.2%, 12.0%, 4.6%)	(24.7%, 10.6%, 16.3%, 6.4%, 25.3%, 12.1%, 4.7%)
Religion	(37.2%, 33.4%, 22.0%, 2.0%, 5.4%)	(37.5%, 33.1%, 22.1%, 1.9%, 5.5%)
Social allowance	10.5%	10.4%
Ethnic background	12.9%	12.5%

In first instance in the selection of weighting variables we also crossed auxiliary variables, but we found in all investigated cases that the estimates based on weighting models including interaction effects differ at most 0.1% of the estimates based on weighting models that incorporate only the main effects. This important finding led us to the decision not to model interaction effects at all. Consequently, in tables 2 a) - e) only models with main effects are given. It must be noted, however, that crossing auxiliary variables does narrow the interval for the bias in general.

We will explain table 2 a), the results of the selection process for ownership of a house. The variable that produces the smallest interval is the average house value (Woz). Introduction of this variable gives a considerable reduction of 0.125 of $w^*(x)$. This variable is thus added to the (empty) weighting model. Next all remaining variables are tested in combination with the average house value. The 'best' variable is household type, which gives a further reduction of 0.026 of $w^*(x)$. By itself household type gives $w^*(x) \cong 0.939$, which is much bigger than the $w^*(x) \cong 0.875$ of average house value. In the following step we add a third variable to average house value and household type. It turns out that the percentage of foreigners is the choice leading to the smallest interval. $w^*(x)$ decreases from 0.849 to 0.829, and the variable is therefore added to the model. Next, the weighting models with one variable removed are compared to the three-variable model. However, both models lead to an increase of $w^*(x)$ that is not significant at $\alpha = 0.01$ and the variables are not removed. In the fourth and fifth iteration, respectively, the variables provplus and age-ms are added. In the fifth iteration the variable age-ms is replaced by age-15, indicating that marital status does not significantly affects the width of the interval. Finally, in the sixth iteration no auxiliary variables can be found that significantly decreases $w^*(x)$.

Table 2: Results of the selection strategy for the survey questions a) ownership of a house b) active in a club, c) ownership of a pc and the auxiliary variables d) social allowance and e) ethnic background. Also given are the correlations between the survey question and the weighting model, the correlation between the response indicator and the weighting model and the selection criterion.

a)

Model	$\bar{y}^*_{gr}$	$\Gamma^*_{X'\beta^*_{X,Y},Y}$	$\Gamma_{\rho,X'\beta^*_{X,Y}}$	$w^*(x)$
Woz	0.612	0.47	0.11	0.875
Woz + hhtype	0.605	0.52	0.13	0.849
Woz + hhtype + %foreign	0.598	0.54	0.15	0.829
Hhtype + %foreign	0.602	0.44	0.16	0.886
Woz + hhtype + %foreign + provplus	0.594	0.56	0.16	0.820
Hhtype + %foreign + provplus	0.599	0.45	0.17	0.881
Woz + %foreign + provplus	0.599	0.53	0.15	0.841
Woz+hhtype+%foreign+provplus+age-ms	0.593	0.57	0.16	0.810
Hhtype + %foreign + provplus + age-ms	0.598	0.47	0.17	0.869
Woz + %foreign + provplus + age-ms	0.594	0.56	0.16	0.816
Woz + hhtype + provplus + age-ms	0.596	0.55	0.16	0.825
Woz+hhtype+%foreign+provplus+age-15	0.594	0.57	0.16	0.813
Woz+hhtype+%foreign+provplus+mstat	0.594	0.56	0.16	0.819

b)

Model	$\bar{y}^{*}_{gr}$	$\Gamma^{*}_{X'\beta^{*}_{X,Y},Y}$	$\Gamma_{\rho,X'\beta^{*}_{X,Y}}$	$w^{*}(x)$
Age-ms	0.462	0.16	0.08	0.983
Age-ms + %foreign	0.455	0.20	0.14	0.969
Age-15 + %foreign	0.455	0.20	0.14	0.970
Mstat + %foreign	0.459	0.15	0.11	0.983
Age-15 + %foreign + telephone	0.451	0.21	0.18	0.962
%foreign + telephone	0.455	0.14	0.18	0.975

c)

Model	$\bar{y}^{*}_{gr}$	$\Gamma^{*}_{X'\beta^{*}_{X,Y},Y}$	$\Gamma_{\rho,X'\beta^{*}_{X,Y}}$	$w^{*}(x)$
Age-ms	0.585	0.50	0.06	0.866
Age-ms + woz	0.579	0.52	0.09	0.850
Mstat + woz	0.586	0.34	0.08	0.938
Age-15 + woz	0.582	0.51	0.08	0.858
Age-ms + woz + ethgr	0.574	0.53	0.11	0.842
Woz + ethgr	0.585	0.22	0.14	0.966
Mstat + woz + ethgr	0.583	0.34	0.10	0.934
Age-15 + woz + ethgr	0.577	0.52	0.10	0.850
Age-ms + woz + ethgr + hhsize	0.573	0.54	0.11	0.837
Woz + ethgr + hhsize	0.572	0.42	0.14	0.898
Mstat + woz + ethgr + hhsize	0.573	0.45	0.13	0.883
Age-15 + woz + ethgr + hhsize	0.573	0.53	0.11	0.840
Age-ms + ethgr + hhsize	0.575	0.53	0.10	0.846
Age-15 + woz + ethgr + hhsize + allowance	0.572	0.54	0.12	0.835
Woz + ethgr + hhsize + allowance	0.571	0.42	0.15	0.896
Age-15 + ethgr + hhsize + allowance	0.574	0.53	0.11	0.844
Age-15 + woz + hhsize + allowance	0.577	0.53	0.10	0.844

d)

Model	$\bar{y}^{*}_{gr}$	$\Gamma^{*}_{X'\beta^{*}_{X,Y},Y}$	$\Gamma_{\rho,X'\beta^{*}_{X,Y}}$	$w^{*}(x)$
Age-ms	0.109	0.34	-0.05	0.940
Age-ms + woz	0.112	0.36	-0.08	0.930
Mstat + woz	0.109	0.22	-0.08	0.973
Age-15 + woz	0.109	0.33	-0.06	0.943
Age-ms + woz + telephone	0.114	0.37	-0.11	0.925
Woz + telephone	0.111	0.16	-0.16	0.975
Mstat + woz + telephone	0.111	0.23	-0.11	0.968
Age-15 + woz + telephone	0.113	0.34	-0.10	0.937

e)

Model	$\bar{y}^*_{gr}$	$\Gamma^*_{X'\beta^*_{X,Y},Y}$	$\Gamma_{\rho,X'\beta^*_{X,Y}}$	$w^*(x)$
%foreign	0.140	0.36	-0.14	0.925
%foreign + telephone	0.144	0.38	-0.16	0.915

Twice did we remove a variable, namely for ownership of a house and active in a club at least once per month. See tables a a) and 2 b). In both cases age-ms has been replaced by age-15, so that marital status is removed from the weighting model.

Table 3 contains the final weighting models for all variables. Also depicted are the weighting models for the variables with three or more categories. As we remarked before, these models are constructed in an ad hoc fashion by looking at the sizes of the shifts relative to the response means.

Table 3: The weighting models that follow from our selection strategy.

Variable	*Model*
Employment	Age-15 + ethnic group + telephone + woz + gender
Owner of a house	Woz + household type + %foreign + provplus + age-15
Active in a club	Age-15 + %foreign + telephone
Owner of a pc	Age-15 + woz + ethnic group + household size + allowance
Educational level	Age-15 + %foreign + provplus + woz
Religion	Ethnic group + provplus + telephone + job
Social allowance	Age-ms + woz + telephone
Ethnic background	%foreign + telephone

The weighting model that is currently used for POLS has the following form

(Gender X age-3 X married/not married) + (gender X age-15) + (region X age-3) + urbanity + provplus + household size, (20)

where the variable age-3 is age in three classes (12-34 years, 35-54 years, 55 years and older). The variable region is a recoding of the province of residence, see the appendix. The particular form of the weighting model originates from the calibration to certain population totals. Effectively, the current model contains six variables: age, gender, marital status, degree of urbanity, provplus and household size.

Table 4 compares the estimates using the models of table 3 for the first month and the full two months of fieldwork to the estimates using the current weighting model for the full two months. Two conclusions can be drawn from table 4. First, the weighting models do not seem to be able to completely remove the differences between the response means after the first and second month. The differences between the first and second month estimates in table 4 are smaller than were found in table 1, but are still present and have the same sign except for ownership of a pc/laptop. Secondly, the differences between the estimates of the selected weighting models and the estimates of the current weighting model can be more than one percent, which relatively large.

Clearly, we can only draw strong conclusions from table 4 when we have some idea of the size of the variance of the estimators. It may be that most differences are not significant at the usual levels. It must be remarked, however, that it is not at all straightforward to approximate variances under the NMAR-R assumption. We did approximate variances for some of the models under the stronger MAR assumption using bootstrap-methods. These simulations revealed that most standard deviations are smaller than 0.001.

Table 4: The estimates for sample means using the weighting models that follow from the selection strategy after the first month and after the second month, and the estimates with the current weighting model after the second month.

Variable	*Estimate after month 1*	*Estimate after month 2*	*Estimate with current model*
Employment	49.6%, 7.0%, 43.4%	50.6%, 6.8%, 42.6%	51.4%, 6.8%, 41.8%
House owner	59.1%	59.4%	60.4%
Active club	45.1%	45.1%	45.6%
Pc-owner	57.3%	57.2%	58.3%
Educational level	25.3%, 10.1%, 16.0%, 6.5%, 25.0%, 12.2%, 4.9%	25.6%, 10.2%, 16.0%, 6.4%, 24.8%, 12.1%, 4.9%	25.1%, 10.2%, 15.8%, 6.5%, 24.8%, 12.5%, 5.2%
Religion	38.6%, 31.4%, 21.2%, 2.9%, 5.8%	38.5%, 31.4%, 21.2%, 2.9%, 5.9%	39.0%, 31.7%, 21.5%, 2.1%, 5.7%
Social allowance	11.6%	11.4%	11.0%
Ethnic background	14.6%	14.4%	13.3%

Since in practice it is more convenient to have a single weighting model, we combine the weighting models of table 3 and get the following model

Age-15 + woz + telephone + %foreign + ethnic group + provplus + household type. (21)

In the proposed weighting model all variables are included except having a job, receiving some form of social allowance and gender. These variables gave only small changes and were selected only once. We do include household type, however, since household type and household size were both selected once but are strongly related to each other. Furthermore, the variable age-15 is taken and not age-ms because marital status only once had a significant impact.

Alternative weighting models are

Age-15 + woz + telephone + %foreign + provplus

Age-15 + woz + telephone + ethnic group + provplus,

in which household type is omitted and either the variable %foreign or ethnic group is added. The latter two variables are collinear.

Table 5 gives the estimates for the six selected target variables using the proposed weighting model (21). Some of the estimates slightly changed with respect to the estimates in table 4.

Table 5: The estimates for the sample means using the proposed weighting model.

Variable	*Estimate with proposed model after month 2*
Employment	50.6%, 6.8%, 42.6%
House owner	59.0%
Active club	44.8%
Pc-owner	57.3%
Educational level	25.9%, 10.2%, 15.9%, 6.4%, 24.6%, 12.2%, 4.9%
Religion	38.5%, 31.4%, 21.2%, 3.0%, 5.9%

The weighting model that we propose in (21) shares the variables age and provplus with the current weighting model (20). New are the average house value, the percentage of non-natives, ethnic group, household type and having a listed telephone number. The estimates using (20) and (21) differ at most 1.4. The largest differences are for proportions of house owners (1.4), pc owners (1.0%), Islamics (0.9%), persons that are active in a club at least once per month (0.8%), persons that are employed for 12 hours or more (0.8%) and the persons that are unemployed (0.8%).

When we compare table 4 to table 1 we see that the largest shift is 4.3%, namely for the proportion of house owners. In general the shifts are rather small.

4. Discussion

First we summarise the findings for POLS 1998:

- The differences between the response means after the first and second month are small in most cases. This is an encouraging result, because it implies that the additional response in the second month is not very different from the response in the first month. However, the response in the second month consists of more persons that are employed for 12 hours or more, while there are few non-native persons and persons with a form of social allowance.
- The biases that we encountered for the first and second month response means are reduced by the use of weighting models but did not disappear. This is somewhat disappointing, because it means that the available auxiliary variables do not completely explain the difference between the first and second month respondents.
- For all investigated cases we found that weighting models with interaction effects give estimates very similar to weighting models with only main effects. This means we can substantially reduce the number of parameters in the model without affecting the outcomes and hence incorporate more auxiliary information.
- For the two auxiliary variables that we regarded, i.e. having a form of social allowance and ethnic background, the selected weighting models produce estimates that are much closer to the sample means than the response means.
- The shifts that are produced by the weighting models relative to the response means are usually not very large. The largest shift that we found is 4.3% for ownership of a house.
- The differences between the proposed weighting model and the current weighting model for POLS may run up to more than 1%. This seems especially due to the 'new' auxiliary variables: average house value at 6-digit postcode level, percentage of non-natives at 6-digit postcode level and having a listed telephone number. These three variables seem very interesting for use in weighting models.
- Analysis of the nonresponse of POLS 1998 revealed that the auxiliary variables interviewer district and degree of urbanisation relate quite strongly to response behaviour. However, these variables did not turn out to be very interesting for weighting nonresponse. It seems that these variables hardly relate to the important survey questions.

The last finding seems to indicate the usefulness of the proposed strategy for selecting auxiliary variables in weighting models. This strategy focuses simultaneously on the relation between auxiliary variables and survey questions, and auxiliary variables and response behaviour. Auxiliary variables are only interesting in case both relations exist.

Another benefit of the strategy is that the construction of strata can be done in one step. In this report we used poststratification as a method to adjust for bias. However, the selection strategy may equally well be used to form cells in propensity score weighting.

There are a number of issues that need to be resolved. First, we need to test other estimation methods for the standard deviations in the selection strategy. Also, it is necessary to investigate to what extent correlations between target and auxiliary variables are affected by nonresponse. Furthermore, it is necessary to adapt the strategy to categorical target variables with more than two categories. The most obvious choices are the maximum of the widths or the average widths. More research is, however, necessary.

In practice it is not very convenient to have a weighting model for each survey question. For this reason the construction of weighting models is often solely based on the prediction of the response behaviour. It is then believed that there will always be at least one survey question that relates to the auxiliary variables that best explain the nonresponse mechanism. As we indicated before this approach may lead to weighting models containing excess variables. Alternatively, one may select the most important survey questions or do a principal component analysis and take the first, say five, components. See for instance Geuzinge, van Rooijen and Bakker (2000). Still it remains unclear how to construct a weighting model that summarises several weighting models.

So far we neglected the variance of estimators because we believe it to be less important than bias. However, variance plays a role in the quality of the estimator. If the number of strata or adjustment cells is large, then the variance may be considerable. We believe, however, that for sample sizes like in POLS it is especially interesting to take variance into account in the calibration of the sample to the population. In the calibration of the response to the sample we think that bias is dominant to variance. Future research must give more insight into the role of variance. It should also be remarked that the approximation of the variance is not at all straightforward under the NMAR-R assumption.

Another important aspect is the assumption underlying the response mechanism. Here we assumed that stratum response means are biased but in the same direction. As a consequence relative distances between stratum means are preserved under nonresponse. There seems to be some evidence that this assumption is closer to the truth than the MAR assumption. More evidence is needed, however.

Finally, in the future we would also like to investigate whether estimates can be improved by using the vector that minimises the width of the interval for the bias. In this report we used the parameter vector that follows from regression of the survey question on the auxiliary variables. In general the optimal vector will produce smaller intervals and thus a smaller maximal absolute bias under NMAR-R.

References

Bethlehem, J.G. (1988), Reduction of Nonresponse Bias through Regression Estimation, *Journal of Official Statistics*, 4, 251–260.

Bethlehem, J.G. and Schouten, J.G. (2003), Nonresponse Analysis of the Integrated Survey on Living Conditions (POLS), Research paper, Statistics Netherlands, Voorburg, *To appear*.

Beukenhorst, D.J. (2001), Rendement van het Mixed-mode Design van Pols en Vertekening, Research paper, Statistics Netherlands, Heerlen.

Czajka, J.L., Hirabayashi, S.M., Little, R.J.A., and Rubin, D.B. (1992), Projecting from Advance Data Using Propensity Modelling: An Application to Income and Tax Statistics, *Journal of Business & Economic Statistics*, 10, 117–131.

Ekholm, A. and Laaksonen, S. (1991), Weighting via Response Modelling in the Finnish Household Budget Survey, *Journal of Official Statistics*, 7, 325–337.

Eltinge, J.L. and Yansaneh, I.S. (1997), Diagnostics for Formation of Nonresponse Adjustment Cells, with an Application to Income Nonresponse in the U.S. Consumer Expenditure Survey, *Survey Methodology*, 23, 33–40.

Geuzinge, L., Rooijen, J. van, and Bakker, B.F.M. (2000), The Use of Administrative Registers to Reduce Non-response Bias in Household Surveys, *Netherlands Official Statistics* 2000-2, 32–39.

Haar, D. ter (2001), *Non-response Analysis POLS*, Masters thesis, Faculteit Economische Wetenschappen, Rijksuniversiteit Groningen.

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

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

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

Rosenbaum, P.R. and Rubin, D.B. (1983), The Central Role of the Propensity Score in Observational Studies for Causal Effects, *Biometrika*, 70, 41–55.

Rosenbaum, P.R. and Rubin, D.B. (1984), Reducing Bias in Observational Studies using Subclassification on the Propensity Score, *Journal of the American Statistical Association*, 79 (387), 516–524.

Schmeets, H., Michiels, J., and Verber, K. (2002), Effecten van Non-respons onder Allochtonen in het POLS 1998, BPA-nr.: H3125-02-SOO, Research paper, Statistics Netherlands, Heerlen.

Schouten, J.G. (2003a), Reduction of Nonresponse Bias using Auxiliary Variables, BPA nr.: 1814-02-TMO, Research paper, Statistics Netherlands, Voorburg.

Schouten, J.G. (2003b), Maximal Absolute Bias under Two Not-Missing-at-Random Assumptions, Research paper, Statistics Netherlands, Voorburg, *To appear*.

Vollebregt, M. (2002), Non-respons in POLS: een Verkenning, BPA nr.: 699-02-TMO, Research paper, Statistics Netherlands, Voorburg.

Vousten, R. and Heer, W. de (1998), Reducing Non-response: the POLS Fieldwork Design, *Netherlands Official Statistics* 1998-2, 16–19.

Appendix: The categories of a number of auxiliary variables

WOZ - Average value of houses (in €) at the 6-digit postcode level in 12 classes:

1: 0 – 50 thousand	5: 125 – 150 thousand	9: 250 – 300 thousand
2: 50 – 75 thousand	6: 150 – 175 thousand	10: 300 – 350 thousand
3: 75 – 100 thousand	7: 175 – 200 thousand	11: 350 – 400 thousand
4: 100 – 125 thousand	8: 200 – 250 thousand	12: 400 thousand and more

%FOREIGN - Percentage of foreigners at the 6-digit postcode level in 8 classes:

1: 0 – 5%	3: 10 – 15%	5: 20 – 30%	7: 40 – 50%
2: 5 – 10%	4: 15 – 20%	6: 30 – 40%	8: 50% and more

PROVPLUS - Province of residence and the four largest cities:

1: Friesland	7: Utrecht (excl. Utrecht city)	13: Amsterdam
2: Groningen	8: Noord-Holland (excl. A'dam)	14: Rotterdam
3: Drente	9: Zuid-Holland (excl. R'dam en Den Haag)	15: Den Haag
4: Overijssel	10: Zeeland	16: Utrecht city
5: Gelderland	11: Noord-Brabant	
6: Flevoland	12: Limburg	

AGE-15 - Age in 15 classes:

1: 12-14	4: 20-24	7: 35-39	10: 50-54	13: 65-69
2: 15-17	5: 25-29	8: 40-44	11: 55-59	14: 70-74
3: 18-19	6: 30-34	9: 45-49	12: 60-64	15: >74

AGE-MS - Combination of age and marital status in 36 classes:

1: 12-17	13: 40-44 not married	25: 60-64 not married
2: 18-19	14: 40-44 married	26: 60-64 married
3: 20-24 not married	15: 40-44 other	27: 60-64 other
4: 20-24 other	16: 45-49 not married	28: 65-69 widowed
5: 25-29 not married	17: 45-49 married	29: 65-69 married
6: 25-29 other	18: 45-49 other	30: 65-69 other

7: 30-34 not married	19: 50-54 not married	31: 70-74 widowed
8: 30-34 married	20: 50-54 married	32: 70-74 married
9: 30-34 other	21: 50-54 other	33: 70-74 other
10: 35-39 not married	22: 55-59 not married	34: >74 widowed
11: 35-39 married	23: 55-59 married	35: >74 married
12: 35-39 other	24: 55-59 other	36: >74 other

REGION – Region of residence in the Netherlands:

1: North (Groningen, Friesland, Drenthe)

2: East (Overijssel, Gelderland

3: West (Noord-Holland, Zuid-Holland

4: South (Noord-Brabant, Limburg)