Nonresponse Analysis of the Integrated Survey on Living Conditions (POLS)

Discussion paper 04004

Jelke Bethlehem & Barry Schouten

Statistics Netherlands

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.

Statistics Netherlands

NONRESPONSE ANALYSIS OF THE INTEGRATED SURVEY ON LIVING CONDITIONS (POLS)

Summary: This paper analyses the nonresponse of the Integrated Survey on Living Conditions, a large continuous survey of Statistics Netherlands. For this survey, more auxiliary variables were available than in regular situations. These variables could be obtained by linking registers and databases to survey data files. Moreover, also a number of fieldwork variables were included in the analysis. By analysing the enriched survey data file, more information could be obtained about a possible under- or over-representation of certain groups in the survey.

Keywords: nonresponse, fieldwork, auxiliary information

1. Introduction

Most social and demographic surveys in The Netherlands suffer from a substantial amount of nonresponse. Nonresponse is the phenomenon that units in the selected sample, and eligible for the sample survey, do not provide information, or that the provided information is unusable. The situation in which all requested information on a sample element is missing is called *unit nonresponse*. If information is missing on some items only, it is called *item nonresponse*. This paper focuses on the problem of unit nonresponse.

Due to nonresponse, the sample size is smaller than expected and, especially high nonresponse rates may lead to increased variances of estimators. This leads to less accurate, but still valid, estimates of population characteristics. A more serious problem caused by nonresponse is that estimates of population characteristics may be biased. This situation occurs if, due to nonresponse, some groups in the population are over- or under-represented, and these groups behave differently with respect to the characteristics to be investigated.

Bethlehem and Kersten (1985) describe some surveys where estimates for main survey variables were biased. They also show that magnitude of the bias increases with a higher nonresponse rate. Bethlehem and De Ree (1999) show that nonresponse rates of Statistics Netherlands surveys have risen over the years to a level of between 40 and 50%. Indeed, estimators must be assumed biased unless convincing evidence to the contrary is provided.

Because of the negative impact of nonresponse on the quality of population estimates, it is vital to reduce the amount of nonresponse in the fieldwork as much as possible. Nevertheless, in spite of all these efforts, usually a substantial amount of nonresponse remains. To avoid biased estimates, the realized response must be analysed. If it turns out that certain groups are over- or under-represented, some kind of correction procedure must be carried out.

This paper analyses the nonresponse of the Integrated Survey on Living Conditions. The survey will be denoted throughout this paper by its Dutch acronym POLS (*Permanent Onderzoek Leefsituatie*).

POLS is a large continuous survey of Statistics Netherlands. Every month a sample is selected. The survey consists of a number of thematic modules. The target population is not the same for every module. However, all target populations at least consist of people of age 12 and older. Persons are selected by means of a stratified two-stage sample. In the first stage, municipalities are selected within regional strata with probabilities proportional to the number of inhabitants. In the second stage, an equal probability sample is drawn in each selected municipality. Sampling frames are the population registers of the municipalities. All individuals of age 12 and older have the same probability of being selected in the sample. For younger people selection probabilities depend on age.

The survey has a modular structure. There is a base module with questions for all sampled persons. In addition, there are a number of modules about specific themes, such as employment situation, health, and justice. The sampled persons are distributed over these themes so that not every person has to answer questions about each theme. This paper restricts itself to analysis of nonresponse in the base module.

Nonresponse rates are high in the Netherlands. Therefore, nonresponse has been subject to much research. The nonresponse in the POLS survey has already been investigated thoroughly, see e.g. Geuzinge, Van Rooijen en Bakker (2000), Beukenhorst (2001), Ter Haar (2001), Schmeets, Michiels en Verber (2002) en Vollebregt (2002).

Auxiliary variables play an important role in establishing possible effects of nonresponse. Auxiliary variables can be defined as variables that have been measured in the survey, and for which also information on their distribution in the population is available. Ideally, the value of an auxiliary variable is available for all sampled and all non-sampled elements in the population. This requires a population database or register. In a less ideal, but still useful, situation the distribution of the auxiliary variables in the population is known (e.g. mean, or frequency distribution).

Auxiliary variables can be used to analyse survey response. If the distribution of such variables in the response differs significantly from the population distribution, this is an indication that groups are over- or under-represented. Moreover, if there is a substantially strong relationship between these auxiliary variables and the target variables of the survey, then population estimates are likely to be biased. If this is the case, these auxiliary variables can be used to reduce nonresponse bias, e.g. by carrying out an adjustment weighting procedure, see e.g. Bethlehem (2002).

The number of available auxiliary variables is limited in many surveys. Very often, only demographic variables like sex, age, marital status, and region can be used. Unfortunately, correlations between these auxiliary variables and target variables of the survey are usually not very strong. Consequently, they are not very helpful in reducing nonresponse bias. In case of POLS, more auxiliary variables were available. These variables could be obtained by linking registers and databanks to

the survey data file. Moreover, also a number of fieldwork variables were included in the analysis.

Section 2 of this paper describes which auxiliary variables where available for the analysis and treatment of nonresponse in POLS. Not all sampled persons have been contacted during the fieldwork period of POLS. Section 3 analyses the possible effects of this. The nonresponse itself is analysed in section 4. The results of this analysis give information about the over- or under-representation of specific groups. Section 5 summarizes the main findings in a number of conclusions. It also investigates the usefulness of these variables in the analysis of various types of nonresponse. This leads to a number of recommendations for the inclusion of additional auxiliary variables.

2. Data

2.1 Survey data

Starting point of the nonresponse analysis was the survey data file of POLS for the year 1998. The fieldwork covered a period of two months. In the first month, selected persons where approached with CAPI (*Computer Assisted Personal Interviewing*). For persons that could not be contacted or refused and who had a listed phone number, a second attempt was made in the second month using CATI (*Computer Assisted Telephone Interviewing*). For persons without a phone or without a listed phone number, a second attempt was made with CAPI. The results of the fieldwork are summarized in table 2.1.1.

Table 2.1.1. The fieldwork results of POLS in 1998

Result	Frequency	Percentage
Sample size	39 302	100,0 %
Response	24 008	61,1 %
Immediate response	9718	24,7 %
Converted refusers	14275	36,3 %
Other response	15	0,0 %
Nonresponse	15294	38,9 %
Unprocessed cases	2514	6,4 %
Not able (illness, handicap)	735	1,9 %
Perseverant refusers	8918	22,7 %
No contact	2093	5,3 %
Language problems	416	1,1 %
Moved	376	1,0 %
Other nonresponse	242	0,6 %

The sample size mentioned in table 2.1.1 is the net sample size. The initial gross sample size was larger. It consisted of 39,431 persons. In 129 cases persons did not belong to the target population of the survey. So, they were removed from the sample (over-coverage).

Ultimately, about 61% of the sampled persons responded. Note that almost 60% of these respondents (14,275 out of 24,008) refused one ore more times before they co-operated.

There were 2,514 cases in the sample that where not processed by the interviewers in the field. This phenomenon is denoted by *administrative nonresponse*. Reasons for this type of nonresponse are lack of capacity of the interviewer and interviewer not available (illness, holiday). Since there is a substantial amount of administrative nonresponse, it is important to find out whether this may be the cause of any bias. This is analysed in section 3.

2.2 Register information

In the early nineties of the last century, Statistics Netherlands started the development of an integrated system of social statistics. This system is called the Social Statistics Database (SSD). The SSD will ultimately contain a wide range of characteristics on each individual in the Netherlands. There will be data on demography, geography, income, labour, education, health, and social protection. These data are obtained by combining data from registers and administrative data sources. Moreover, also data from surveys are included. These data relate to attitude, behaviour, etc. For more information about the SSD, see Everaers & Van der Laan (2001).

Using internal personal identification numbers, SSD records could be linked to the survey data records, both for respondents and nonrespondents. Thus, demographic variables like sex, age, province of residence, and ethnicity became available for all sampled persons, and also socio-economic variables like employment and various types of social security benefits.

2.3 Postal code area information

The Netherlands is divided in approximately 420,000 postal code areas. A postal code area contains, on average, 17 addresses. These areas are homogeneous with respect to social and economic characteristics of its inhabitants.

Using information from the population register, Statistics Netherlands has computed some demographic characteristics for these postal code areas. Since postal codes are included in the survey data file for both respondents and nonrespondents, these characteristics can be linked to the survey data file. Among the variables used in this analysis, are degree of urbanisation, town size, percentage of people with a foreign background (non-natives). From another source also the average house value was included.

2.4 Fieldwork information

During the fieldwork period, interviewers kept record of all contact attempts. For each attempt its contact result was recorded (contact, or not). In case contact was established, the result of the co-operation request was recorded (response or

nonresponse, and in case of nonresponse the reason of nonresponse). Also other information was included, like the mode of the fieldwork attempt (CAPI or CATI), and whether there was contact with the person to be interviewed or another member of the household. All this fieldwork information was included in the analysis data file.

2.5 Other information

Two other variables also have been included in the survey data file. The first one is the interviewer district code. Thus, for every respondent and nonrespondent it is know which interviewer made the contact attempts. The second variable is an indicator whether a selected person has a listed telephone number or not.

3. Analysis of unprocessed cases

3.1 Unprocessed cases

There were 2,514 cases selected in the sample that where not processed by the interviewers in the field. Reasons for this type of nonresponse are lack of capacity (too high workload for the interviewer) and interviewer not available (illness, holidays). Since there is a substantial amount of this so-called *administrative nonresponse*, it is important to find out whether this may be the cause of any bias.

The unprocessed and processed cases where compared for the available auxiliary variables. For each variable separately a univariate logistic regression was carried out with status of the case (processed / unprocessed) as dependent variable. For the following explanatory variables significance differences were found between processed and unprocessed cases:

• Marital status of selected person;

• Percentage of people in the area of residence with a foreign background (non-natives);

• Province of residence;

• Size of town of residence;

• Degree of urbanization of town of residence;

• Whether or not the selected person has a listed telephone number;

• Interviewer district.

The logistic regressions showed that interviewer district had the highest explanatory power (8.9% explained variance). Second best was province of residence (6.4% explained variance). Indeed, a closer look at the data made clear that the highest percentages of unprocessed cases were found in interviewer districts in large cities.

Apparently, capacity problems in interviewer districts in large cities are the main cause of unprocessed cases.

To see whether explanatory power could be improved, other variables where added to the model (with interviewer district as explanatory variable). By adding average house value, the percentage of explained variance could be raised to 9.1%. Adding the percentage of non-natives instead had the same effect. Adding more variables did not show significant improvement. The results of the logistic regressions are summarized in table 3.1.1.

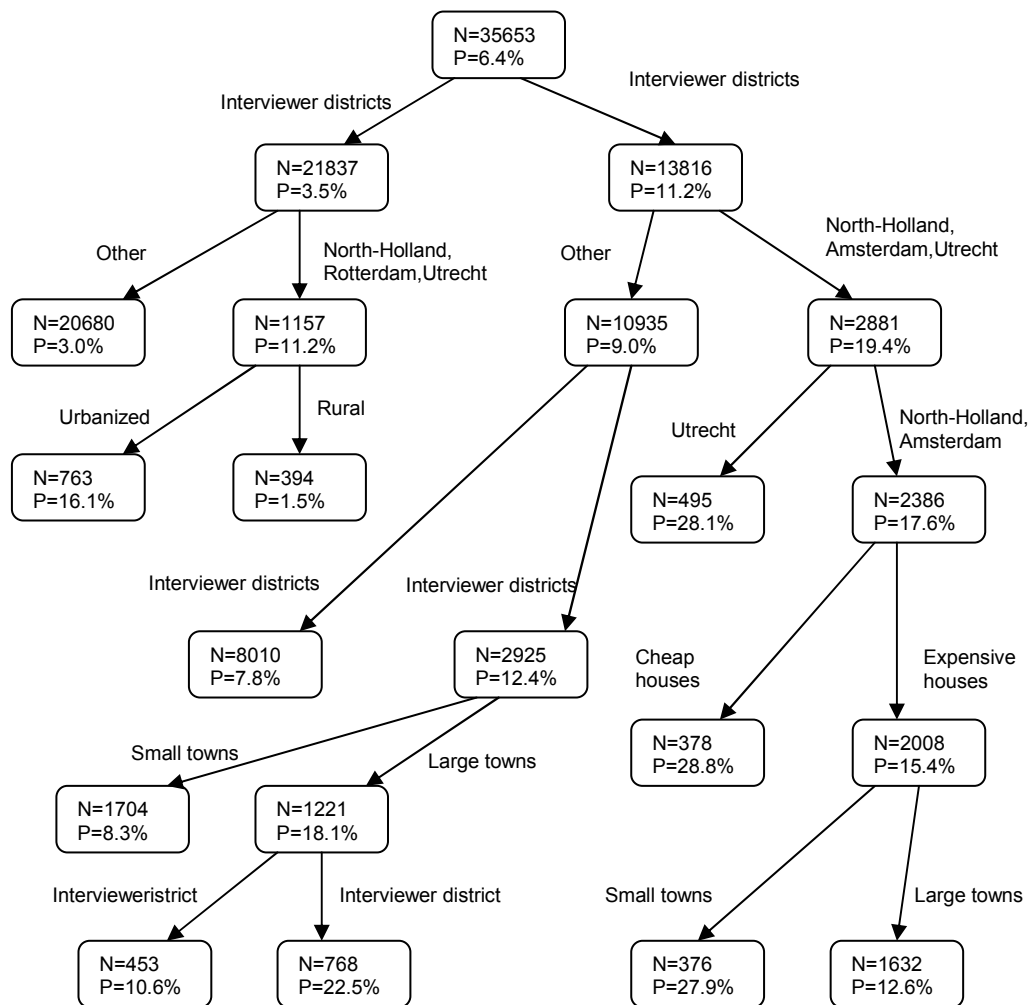*Table 3.1.1. Results of logistic regressions for processed/unprocessed*

Dependent variable	Explanatory variables	Percentage of explained variance
Processed/unprocessed	Province	6.4 %
	Interviewer district	8.9 %
	Interviewer district, Average house value	9.1 %
	Interviewer district, Percentage non-natives	9.1 %

To obtain more insight in the possible relationship between whether or not a case has been processed and auxiliary variables, also a classification tree has been build. Such a technique breaks up the data set into a number of homogeneous groups by means of a succession of binary splits. Thus, groups of cases are obtained that are increasingly more homogeneous with respect to the response variable. A more elaborate application of this technique can be found in Mesa et al. (2000). The result for the analysis of processed/unprocessed cases is shown in figure 3.1.1.

The conclusions from the analysis of the classification tree are similar to those of the logistic regression, but more detailed. For example, one can see that the highest percentage of unprocessed cases (28.8%) is found for cheap houses in the province of North-Holland and the town of Amsterdam. And the lowest percentage of unprocessed cases (1.5%) can be found in the rural areas of the province of North-Holland.

Also, the group of persons for which no contact attempts have been made in the first month of the fieldwork, turned out to be a selective group. This group (2,893 cases) resembles the group of unprocessed cases. This is not surprising as a large portion of this group (873 of the in total 2,893 cases) remained unprocessed at the end of the fieldwork period.

Figure 3.1.1. Classification tree for processed/unprocessed cases

3.2 Mode effects

The fieldwork of POLS 1998 covered a period of two months. In the first month, selected persons where approached with CAPI. For persons that could not be contacted or refused to co-operate, the following strategy was implemented:

- If an appointment had been made for the second month, they were re-approached with CAPI;

- If no appointment had been made, and they had a listed phone number, they were re-approached with CATI.

- If no appointment had been made, and they did not have a listed phone number, they were re-approached with CAPI.

The results of the fieldwork are summarized in table 3.2.1. The second column in the table refers to cases that were not processed in the second month, and that were also not processed in the first month. These numbers are not taken into account in the computation of response rates, and the analysis of the nonresponse. The third

column contain cases that were processed in the first month, and for which no further action was undertaken in the second month.

Table 3.2.1. Results of contact attempts for different interviewing modes

Result in Month 1	Not processed in month 2	Not attempted in month 2	Processed in month 2					
			No phone		Phone			
			CAPI		CAPI		CATI	
			R	NR	R	NR	R	NR
No attempt	873	0	240	423	713	386	68	183
Refusal, app.	23	21	302	221	1330	345	125	227
Refusal, no app.	3	2184	0	2	0	0	1277	4992
No contact	49	1036	132	371	276	261	374	676
Unprocessed	1566	0	14	17	37	9	3	10

The table shows that making appointments with first time refusers works. There are three groups:

- The response rate in the second month of the first-month refusers with a listed phone number, and with whom an appointment had been made, and who were re-approached with CAPI, is very high. 1,330 persons of the group of 1,675 responded, which is a response percentage of 79.4 %.

- Also the response rate in the second month of the first-month refusers without a listed phone number, and with whom an appointment had been made, and who were re-approached with CAPI, is high. 302 persons of the group of 523 responded, which is a response percentage of 57.7 %.

- The response rate was much lower for the first-month refusers with a listed phone number, and with whom an appointment had been made, and who were re-approached with CATI instead of CAPI. 125 persons of the group of 352 responded, which is a response percentage of 35.5 %.

Chances of success were much lower if no appointment could be made with a first-month refuser. All these people were re-approached with CATI. The response rate was low. Only 1,277 out of 6,269 responded, which is 20.4%.

When looking at the first-month no-contacts, a similar pattern can be observed, although response rates are lower:

- First-month no-contacts without a listed phone number were re-approached with CAPI: Their response rate was low. Only 132 out of 503 responded, which is a response percentage of 26.2%.

- First month no contacts with a listed phone number and who were re-approached with CAPI had a much higher response rate: 276 out of 537 responded, which is a response percentage of 51.4%.

- First month no contacts with a listed phone number and who were re-approached with CATI had a response rate comparable to that of corresponding refusers: 374 out of 1,050 responded, which is a response percentage of 35.6%.

All response rates have been summarized in table 3.2.2. A first conclusion that can be drawn is that making appointments helps. Once an appointment has been made for a next contact, the probability of getting a response is reasonably high. Interviewers were instructed to also attempt to make an appointment in the case of refusal. If they succeeded, the refusal was characterized as 'soft refusal'.

Table 3.2.2. Response rates in the second month by result of first month

First month result	Listed phone number	CAPI	CATI
Refusal, appointment	No	57.7 %	-
Refusal, appointment	Yes	79.4 %	35.5 %
Refusal, no appointment	Yes	-	20,4 %
No contact	No	26,2 %	-
No contact	Yes	51,4 %	35,6 %

A second conclusion is that there seems to be a difference between people with a listed and a non-listed telephone number. Both in the case of refusal and no-contact, the probability of getting a response at a next attempt is higher for people with a listed telephone number. This conclusion is similar to that of Beukenhorst (2001) that having a list / non-listed telephone number is an important auxiliary variable.

A third conclusion is that CAPI seems to work better than CATI. A further analysis of these differences showed that it is probably not caused by a mode effect but due to characteristics of the specific groups of people that were approached with CATI. In case of the first month refusers with an appointment and a listed phone number, it turned out in a high percentage of cases contact attempts were successful. However, contact attempts at the appointed time could fail. After such a failure, the case was handed over to the CATI unit. So, there was no appointment for the CATI attempt. In case of the first month no contacts with a listed phone number, CATI cases were mainly those cases where many CAPI attempts had already failed. So, these are really hard to contact cases.

4. Analysis of the various types of nonresponse

Now the nonresponse is analysed as it occurred in the field. This should give an idea of the type and size of the problems the interviewers are confronted with. The starting point is the sample obtained by leaving out the administrative nonresponse (unprocessed cases). This leaves a net sample of 39,302 − 2,514 = 36,788 cases. In 24008 cases there was a response. So, the response percentage was 65.3%
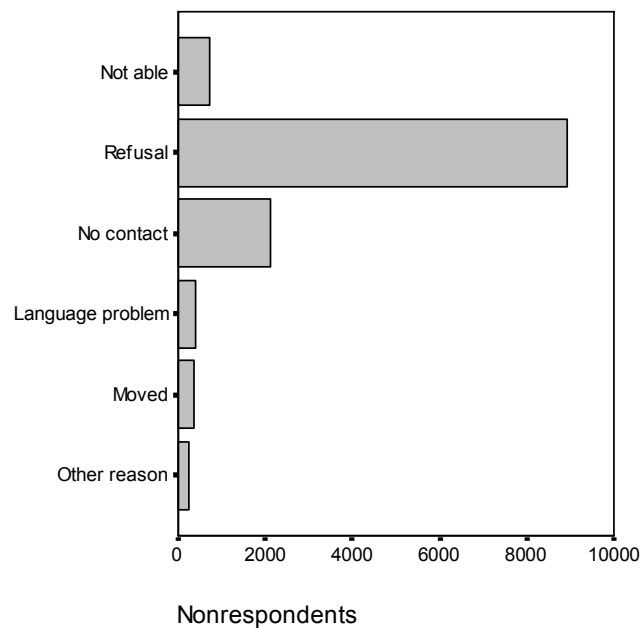
The POLS survey distinguishes the following main groups of nonrespondents:

- Persons not able to participate due to illness or handicap

- Refusers

- Persons that could not be contacted

- Persons with language problems

- Moved persons

Figure 4.1. shows how big the various nonresponse groups are. Clearly, the most important group is that of the refusers (69.8%). The non-contact group is much smaller (16.4%). The third group (5.8%) consist of the not-able.

Figure 4.1. Composition of the nonresponse

Nonrespondents

Language problems (3.3%) and moved persons (2.9%) constitute only a very small portion of the total nonresponse.

The various nonresponse groups are analysed in more detail in the following sub-sections. For the refusers, also a comparison will be made with persuaded nonrespondents.

4.1 Moved persons

After the sample had been selected, Statistics Netherlands asked the municipalities to provide recent address information of the selected persons. Municipal registrations are not completely up-to-date. They lag behind a few months with respect to moving. Therefore, it may happen that a selected person has moved to another address. In theory, if interviewers are able to determine the new address, e.g. by asking the new inhabitants, and this new address is within their region, they have to visit the new address. In practice, this almost never happens.

From table 2.1.1 it is clear that the moved persons form a relative small group of nonrespondents. They constitute about 1% of the total sample size.

The moved and non-moved persons where compared for the available auxiliary variables. For each variable separately a univariate logistic regression was carried out with status of the person (moved / not moved) as dependent variable. For the following explanatory variables significant differences were found:

- age;

- marital status;

- degree of urbanization.

The logistic regressions showed that age had the highest explanatory power (11.5% explained variance), followed by marital status (7.6%) and degree of urbanization (1.7%).

To see whether explanatory power could be improved, other variables were added to the model (with age as explanatory variable). By adding marital status, the percentage of explained variance went up to 14.3%. Adding degree of urbanization as main effect raised explained variance to 15.3%. The results of the logistic regressions are summarized in table 4.1.1.
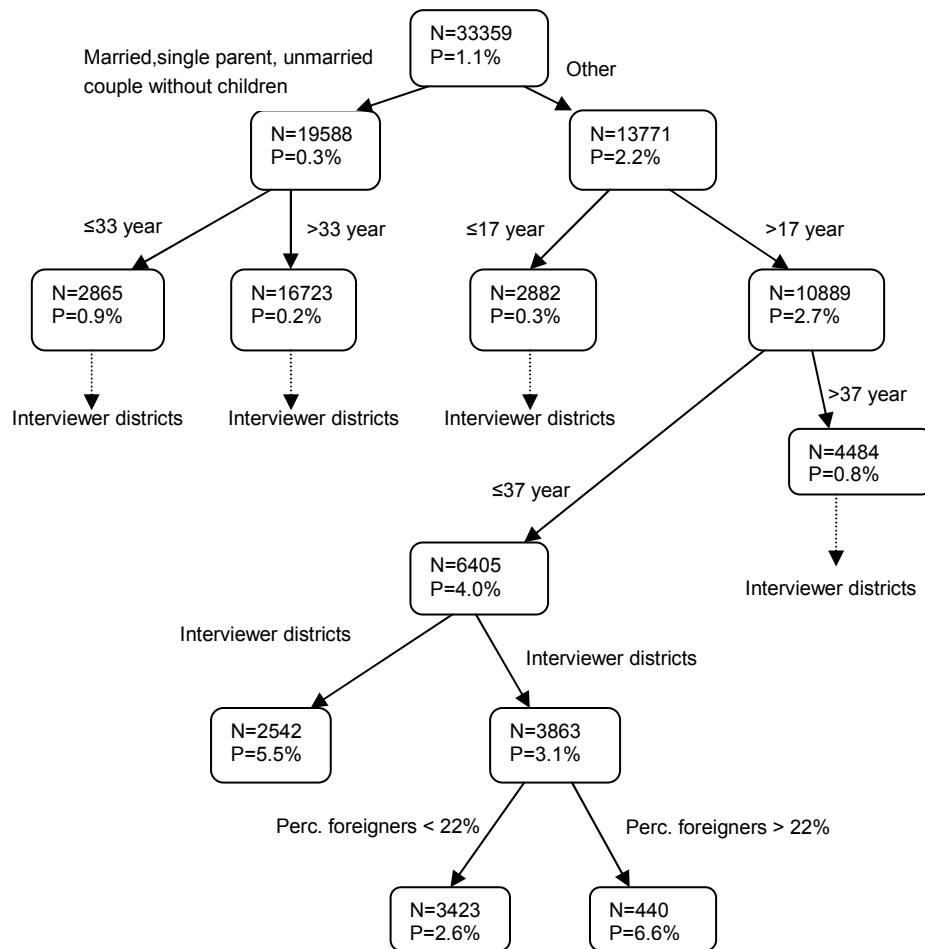
Table 4.1.1. Results of logistic regressions for moved/not moved

Dependent variable	Explanatory variables	Percentage of explained variance
Moved / not moved	Age	11.5 %
	Age, Marital status	14.3 %
	Age, Marital status, Degree of urbanization	15.3 %

The conclusion is that traditional auxiliary variables may be used to correct for a possible nonresponse bias. Another way to do it would be to use the most recent information about movings that is available in municipal registers. However, it is not very easy to acquire this information.

A classification tree analysis showed that another interesting variable is the composition of the household. The rate of moving is lowest (0.2%) among married couples, single parents and unmarried couples without children who are older than 33 years. The rate of moving of highest (6.6%) among single persons and unmarried couples of an age of between 17 and 37 years with children and who live in certain neighbourhoods with a high percentage of non-natives.

Figure 4.1.1. Classification tree for moved/non-moved persons

4.2 Hard-to-contact persons

The fieldwork of POLS 1998 was spread over two months. In the first month, contact attempts were made with 34,084 persons. Contact was established in approximately 90% of the cases.

In case of nonresponse in the first month, new contact attempts were made by telephone in the second month. This nonresponse group consisted of 7,895 persons. Contact was established in 84% of the cases.

In total, for the whole fieldwork period, there have been contact attempts with 36,412 persons. These attempts were successful in 94% of the cases. For 2,093 persons no contact at all could be established. This constitutes 5.3% of the sample.

The contacted and not contacted persons where compared for the available auxiliary variables. For each variable separately a univariate logistic regression was carried out with status of the person (contacted / not contacted) as dependent variable. For the following explanatory variables significance differences were found:

• Region of residence;

• degree of urbanization;

- listed phone number (yes/no);

- percentage of people with a foreign background in the neighbourhood;

- average house value;

- age;

- marital status.

The logistic regressions showed that listed phone number (yes/no) had the highest explanatory power (6.7% explained variance). To see whether explanatory power could be improved, other variables were added to the model (with listed phone as explanatory variable). By adding interviewer district, the percentage of explained variance went up to 11.0%. Adding age by marital status raised explained variance to 13.8%. After adding average house value and household size, a model was obtained that could account for 15.2% of the variance. The results of the logistic regressions are summarized in table 4.2.1.

Table 4.2.1. Results of logistic regressions for contact/no contact

Dependent variable	Explanatory variables	Percentage of explained variance
Contact / no contact	Listed phone	6.7 %
	Listed phone, Interviewer district	11.0 %
	Listed phone, Interviewer district, Age by marital status	13.8 %
	Listed phone, Interviewer district, Age by marital status, Average house value	14.2 %
	Listed phone, Interviewer district, Age by marital status, Average house value, Household size	15.2 %

An interesting question is whether no-contact cases resemble hard-to-contact cases. If so, the hard to contact persons can provide information about the no contacts. To answer this question, it is important to define what a hard to contact person is. One way to do it is to define it in terms of contact probabilities. The contact probability is the probability of success in a contact attempt at a randomly selected moment during the fieldwork period.
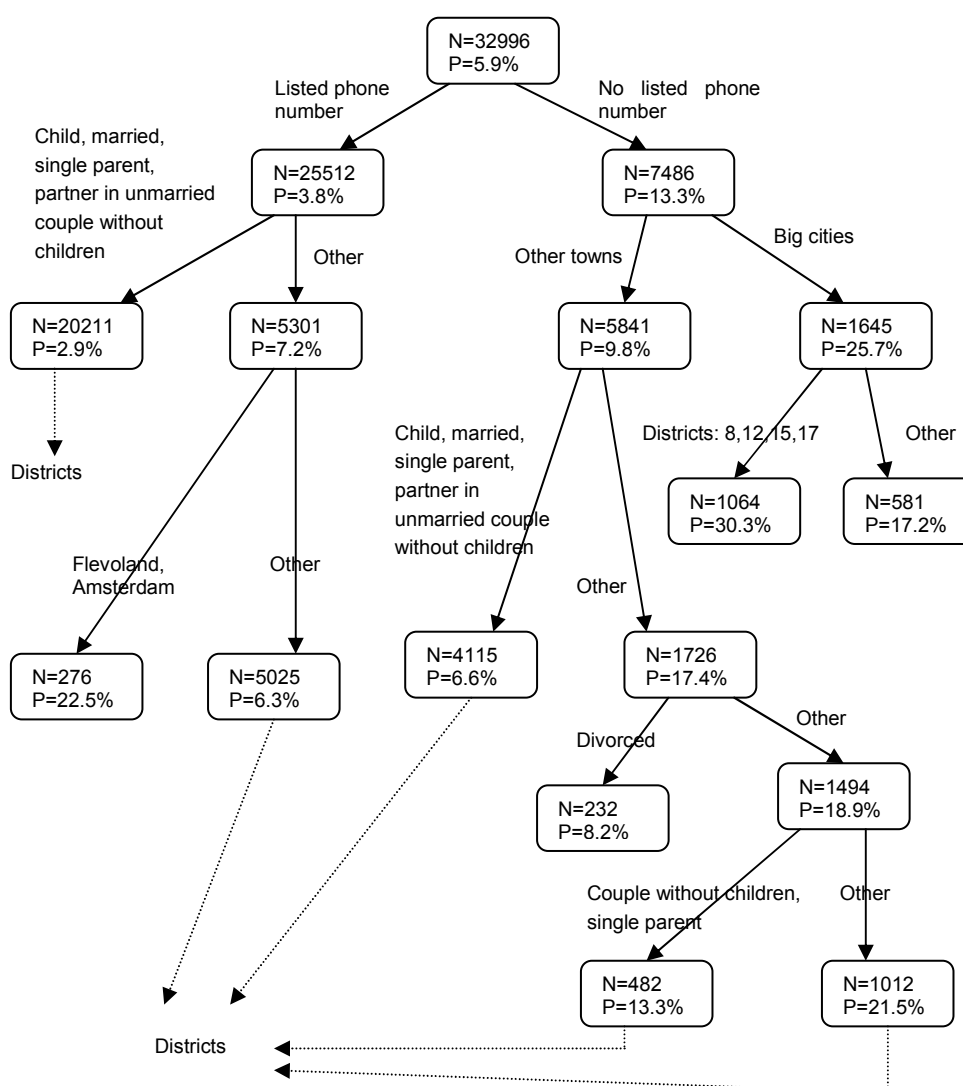
For each selected person, the contact probably could be estimated if a number of independent random contact attempts could be made. This can be seen as a series of Bernoulli trails with probability of success being equal to the contact probability.

Unfortunately, the fieldwork of POLS 98 was not organized in such a way that it allowed for easy estimation of contact probabilities. The contact attempts were

certainly not chosen at random moments. Also, the number of contact attempts changed form address to address.

Figure 4.2.1 shows the classification tree for hard to contact persons. The percentage of hard to contact persons is lowest (2.9%) for the group with a listed phone number, and consisting of children, married people, single parents, and partners in unmarried couple without children. Apparently, the percentage of hard to contact persons is highest (30.3%) for people without a listed phone number in certain interviewer districts in the four big cities.

Figure 4.2.1. Classification tree for hard to contact persons

4.3 Language problems

In about 1.1 % of the sample cases response was not possible due to language problems.

The Dutch speaking and non-Dutch speaking persons were compared for the available auxiliary variables. For each variable separately a univariate logistic

regression was carried out with status of the person (Dutch speaking / not Dutch speaking) as dependent variable. It did not come as a surprise that two auxiliary variables turned out to be correlated with the language problem variable. The first of these variables is ethnic origin (Native Dutch, Moroccan, Turkish, Surinamese, Antillean, Other western origin, Other non-western origin). The second variable was ethnic generation (Native Dutch, First generation non-native, Child of non-native parents, Child with one non-native parent).

The logistic regressions showed that ethnic group had the highest explanatory power (46.5% explained variance). The explanatory power of ethic generation was slightly less: 45.1%

To see whether explanatory power could be improved, other variables were added to the model (with age as explanatory variable). Only one significant improvement could be determined: by crossing ethnic group and ethnic origin, the percentage of explained variance could be raised to 51.2%. The results of the logistic regressions are summarized in table 4.3.1.

Table 4.3.1. Results of logistic regressions for Dutch/non-Dutch speaking

Dependent variable	Explanatory variables	Percentage of explained variance
Dutch spraking / not Dutch speaking	Age, Ethnic generation	45.1 %
	Age, Ethnic origin	46.5 %
	Age, Ethnic generation, Ethnic origin	51.2 %

There is some indication that employment status also has some explanatory power. However, the number of observations was too small for a definite conclusion.

Language problems are most serious for first generation Moroccans and Turks. The percentage of nonresponse due to language problems is 29% for the Turks, and 25% for Moroccans.

4.4 Refusals

After subtracting the non-contacts, the group of people remains for which contact was established. These people were invited to participate in the survey. The response rate in this group was 71.3%. A large part of the response (79.6% of the 71.3%) was already realized in the first month.

It should be noticed that the contacted group is not representative for the target population, because the no-contacts have been removed from it.

First, the refusals in the first month were analysed. The response rate among the contacts was 62.4%.

The participants and refusers were compared for the available auxiliary variables. For each variable separately a univariate logistic regression was carried out with

status of the person (participation / refusal) as dependent variable. Nearly all auxiliary variables had a significant contribution in these models. Due to the large sample size, even small contributions turned out to be significant.

There were only very small differences in participation rates for the various categories of the auxiliary variables, with a number of exceptions:

- Participation rates for people under 18 years were much higher than rates for older people;

- There are geographic differences. Co-operation rates in the big cities are much lower.

- Participation rates were much higher for persons with a listed telephone number. People with a none-listed telephone number tended to refuse more. Note that in the first month the listed telephone numbers did not play a role in the fieldwork strategy. So, this is not a fieldwork effect.

- Participation rates among Moroccans and Turks were higher. The on average lower response rate for these groups are apparently caused by contact problems (language problems), and not by refusal.

Separate analyses were carried for the group of age 18 year and older in the big cities, and for the group of age 18 and older in other parts of the country.

In the big cities, the variables with most explanatory power were the percentage of non-natives in the neighbourhood and the average house value. Participation decreases with an increasing percentage of non-natives and a decreasing housing value. So there seems to be a relationship between participation and social-economic status of the neighbourhood.

Outside the big cities, the interviewer district turned out to have most explanatory power. Apparently, the performance of interviewer plays a significant role here.

A second analysis was carried out on refusals in the complete fieldwork period. Persons below the age of 12 have not been included in this analysis. They have different inclusion probabilities. Consequently, percentages in the results reflect true response rates.

Figure 4.4.1 contains the classification tree for the refusers. The refusal rates in the homogeneous groups vary between 19.9% and 58.2%. Refusal rates under 20% are found for two groups:
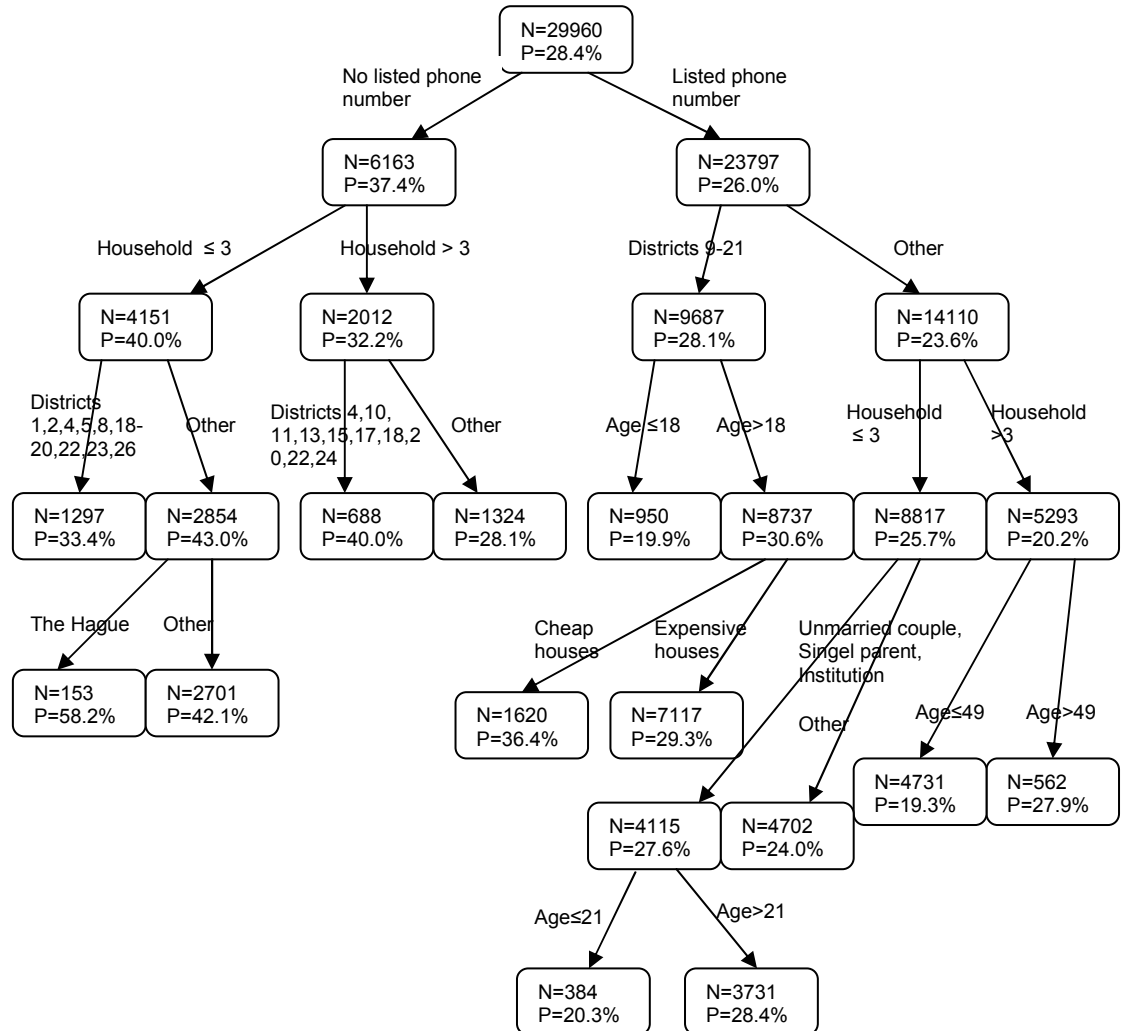
- A rate is 19.3% for people with an age of at least 50 in household consisting of at least four members, and with a listed telephone number;

- A rate of 19.9% is found for people in certain districts, with an age of at most 18 years, and who have a listed phone number.

Refusal rates of over 40% are found for three groups:

- A rate of 58.2% is found for people without a listed phone number in the town of The Hague who live in small households (less than four members);

- A rate of 42.1% is found for people without a listed phone number in other parts of country, and who live in small households (less than four members);

- A rate of 40.0% is found for people without a listed phone number in larger households (more than three members) in certain districts.

Figure 4.4.1. Classification tree for refusers

Particularly, there are strong effects for age and listed telephone number. Among the people of age 18 years and older, the variable listed telephone number (yes/no) has the highest explanatory power. The effect is even stronger than in the first month. This does not come as a surprise because of the role telephones play in the second month of the fieldwork.

It should be mentioned that response rates for first generation non-natives decrease in the second month of the fieldwork. This is due to the fact that telephone ownership among this group is much lower than for the rest of the population.

4.5 Direct respondents, persuaded refusers, and persistent refusers

An attempt has been made to use POLS 1998 to analyse response propensity. In this analysis, response propensity has been obtained as the complement of the nonresponse propensity. And this is defined as the effort required to persuade a contacted person to participate in the survey.

The group with the highest response propensity consists of the direct respondents. Their co-operation is obtained in the first contact. This group constitutes approximately 40% of the total response. This means that 60% of the response was realized in the second and subsequent attempts. This makes clear that interviewers should not stop after their first attempt.

The group of first time refusers can be divided in temporary refusers and permanent refusers. A temporary refusal occurs when a person does not want to co-operate at that particular point in time, but may want to co-operate at a different moment. The interviewers often succeed in such situations to make an appointment for an interview at a later moment in time.

In the case of POLS 1998, 14,275 persons refused to participate at the first contact moment, but were persuaded to co-operate at a later moment. For 9,164 this was realized after making an appointment. For 2,658 persons this was realized after two appointments, and for 865 persons after three or more appointments.

In using these data to define response propensity one has to realize that the effect of no contacts can not be separated from those of refusal. For example, the number of refusals in a series of contacts is not a good indicator of response propensity. For persons that are hard to contact, there are only a few contacts, and therefore the response propensity is incorrectly estimated as too low.

A consequence of this entanglement of no-contact and refusal effects is that it is hard to compare direct respondents with persuaded respondents. If we neglect effects of failed contact attempts, and compare direct respondents with respondents after one or two appointments, the following conclusion can be drawn:

- The proportion of direct respondents among the age group of 60 and older is relatively high. The percentage of appointments is low.

- Persuaded refusers more often have a job, less often have a social security benefit, even if corrected for age.

- Males more often make an appointment, even if corrected for age and having a job.

The first two conclusions are not surprising, but the third conclusion is new.

There is another effect that makes estimating response propensities difficult. This is the behaviour of the interviewers. They make decisions in the field about re-approaching refusers. When they feel a refusal is persistent, they may decide not to make a new attempt. Table 4.5.1 shows the results of the second contact by the results of the first contact.

Table 4.5.1. Results of the first and second contact.

	Result of second contact			
	No contact	Response	Appointment	Refusal
First contact	-	28.8%	48.0%	23.2%
Second attempt after appointment	4.5%	56.2%	27.6%	11.6%
Second attempt after refusal	39.2%	8.6%	17.2%	35.0%

The bottom line of the table shows that a substantial part of the first time refusers (39.2%) is not re-approached.

An analysis of the groups distinguished in table 4.5.1 leads to the following conclusions:

- There are relatively many elderly persons, women, widowed persons, and persons with a social benefit among both the first time respondents and the first time refusers. These persons do not tend to make appointments.

- Persons with a job more often make appointments.

- Persons living in the urbanized areas of the western part of the country more often refuse.

- First generation non-natives more often respond at the first contact than second generation non-natives.

4.6 Not able to participate due to illness or handicap

There is a small group of 735 (1.9%) people for which no response was obtained due to illness or handicap. In 416 cases there has been no direct contact between the interviewer and the selected person. Apparently, the contact was with another person in the household. From the fieldwork reports it is not clear whether illness or handicap was the direct cause of nonresponse. It is not unlikely that illness or handicap was used as a reason for nonresponse to avoid having to refuse.

Comparison of the not able persons with other persons shows a strong relationship with age, and a somewhat less strong relationship with marital status, having a job, receiving a social benefit due to disablement.

5. Conclusions

The analysis of the POLS 1998 data shows that additional auxiliary variables help to explain what is going on with respect to response and nonresponse. Not only demographic and socio-economic variables are useful in this respect, but also fieldwork variables that describe various contact attempts.

Traditionally, fieldwork reports are made to monitor fieldwork and interviewer performance. Use of this type of information in a nonresponse analysis requires this

information to be recorded in a more systematic way. Also, it is important that fieldwork information becomes a standard part of the survey data file.

It is a good idea to split the response mechanism in two sequential phases. The first phase is that of the contact attempt. The second phase is that of the co-operation attempt once contact has been established. Analysis of both phenomena may require different models and different auxiliary variables. However, in practical survey situations it is not easy to separate both mechanisms. Future survey design should attempt to take care of this in a better way. Of course, it also remains important to distinguish other groups of nonrespondents.

For the analysis of refusals it is important to be able to model response propensity. As a first step, it should become possible to separate in the survey data file the direct respondents from persuaded respondents and the persistent respondents.

The analysis of the nonresponse of POLS 1998 showed that that it helps to have the following auxiliary information available for both respondents and nonrespondents:

- Whether or not a person has a listed phone number.

- The composition of the household, and other characteristics of the household.

- Characteristics of the house

- Socio-economic status of the neighbourhood

References

Bethlehem, J.G. (2002), Weighting nonresponse adjustments based on auxiliary information. In: R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little (eds.), *Survey Nonresponse*, Wiley, New York.

Bethlehem, J.G. and De Ree, S.J.M. (1999), From population to sample. In: J.G.S.J van Maarseveen and M.B.G. Gircour (eds.), *A Century of Statistics*. Statistics Netherlands, Voorburg, The Netherlands, pp. 463-492.

Bethlehem, J.G. and Kersten, H.M.P. (1985), On the treatment of nonresponse in sample surveys. *Journal of Official Statistics* 1, pp. 287-300.

Beukenhorst, D.J. (2001), *Rendement van het mixed-mode design van POLS en vertekening*. Internal CBS report, Statistics Netherlands.

Everaers, P. & Van der Laan, P. (2001), *The Dutch virtual Census*. E-Proceedings of the 53th Session of the International Statistical Institute, Seoul, Korea.

Geuzinge, L., Rooijen, J. van, Bakker, B.F.M. (2000), The use of administrative registers to reduce nonresponse bias in household surveys, Netherlands Official Statistics 2000-2, pp. 32–39.

Groves, R.M., Couper, M.P. (1998), *Nonresponse in Household Interview Surveys*, John Wiley & Sons, New York.

Haar, D. ter (2001), *Non-response Analysis POLS*. Master Thesis,, Faculty of Economics, University of Groningen.

Mesa, D.M., Tsai, P. and Chambers, R.L. (2000), *Using Tree-Based Models for Missing Data Imputation: An Evaluation Using UK Census Data*. University of Southampton.

Schmeets, H., Michiels, J., Verber, K. (2002), *Effecten van nonrespons onder allochtonen in het POLS*. Internal CBS report, Statistics Netherlands, Heerlen.

Vollebregt, M. (2002), *Non-respons in POLS: een verkenning*, CBS report 699-02-TMO, Statistics Netherlands, Voorburg.