STUDY AND ADJUSTMENT OF NON-RESPONSE IN HOUSEHOLD SURVEYS

EUSKAL ESTATISTIKA ERAKUNDEA
INSTITUTO VASCO DE ESTADISTICA

Donostia-San Sebastián, 1
01010 VITORIA-GASTEIZ
Tel.: 945 01 75 00
Fax.: 945 01 75 01
E-mail: eustat@eustat.es
www.eustat.es

Presentation

This document presents the work carried out by Eustat during 2006 and 2007 in relation to the study and treatment of non-response.

The work carried out forms part of the operations of the Basque Statistics Plan 2005-2008, related to R&D&I in statistical methods, which sets out to investigate and apply new statistical-mathematical methods in statistical operations. The spirit underlying these studies comes within the framework of management excellence, which is defined by a process of continual improvement.

The inclusion of the study into non-response in the operations of the Plan is a reflection of the concern that EUSTAT shares with other international statistics offices as regards this matter. The lack of response is a phenomenon which is happening, to different extents, in many statistical operations, always reducing the accuracy of the estimators of the results and which, furthermore, introduces bias.

EUSTAT has always studied and monitored the lack of response in household surveys. However, the need to establish common measurement standards or criteria has been put forward, based on the European methodology published for this purpose by the Institute for Social & Economic Research, of the University of Essex (United Kingdom).

This research also sets out to study as to how to improve the estimations in household surveys, in situations where there is a lack of response, via auxiliary information, information which is external to the survey.

I hope that this publication proves useful to all those interested in this area of statistics.

Vitoria-Gasteiz, September 2008

Josu Iradi Arrieta

General Director.

STUDY AND ADJUSTMENT OF NON-RESPONSE IN HOUSEHOLD SURVEYS

SUMMARY

The document before you[1] is divided into the following chapters:

The first chapter makes an introduction and sets out the objectives that have governed the production of this technical Notebook.

The second chapter covers standardisation of the calculation of response rate and its results in the Survey on the Population in Relation to Activity (henceforth PRA). This standardisation has been dealt with in greater detail in the Work Notebook dedicated to this subject, published by Eustat.

The objective of the third chapter is to set out the analysis of auxiliary information, which is the information available, both for respondent and non-respondents, and which is related to the lack of response. This analysis is centred on selecting the most appropriate auxiliary variables to reduce the effect of the lack of response in the survey. For this we have applied the methodology proposed by Statistics Sweden.

The fourth chapter tackles calibration based on specific auxiliary variables, using the generalised regression estimator. This estimator is explained extensively in this chapter, as is the CLAN[2] programme.

The fifth chapter presents the conclusions which have been reached and the final point refers to the bibliography employed.

[1] EUSTAT would like to thank Susana Sanz Abrego for her excellent research work carried out in the framework of the Grant for Resarch into Statistical-Mathematical Methodology promoted by EUSTAT, specifically in the field of non-response.

[2] CLAN is a macro developed by Statistics Sweden which operates in the SAS environment

Index

Index of tables

Index of graphs

Chapter 1

Introduction

Introduction and objectives

In recent years, statistical offices have shown some concern for the lack of response produced in replying to official statistics.

The importance of the lack of response stems chiefly from two motives. The first, the introduction of bias or a measurement error, due to the fact that the non-respondents may differ from the respondents in important characteristics. Secondly, that non-response reduces the accuracy of the estimations, since there are fewer cases available for analysis.

Faced with the possibility that a lack of response will occur, two non-exclusive courses of action can be taken. Firstly, the offices can act beforehand and try and reduce it in the field of work. To do so, everything relating to the data collection is checked, from questionnaires to incentives, the means of contacting the population, collection by telephone, via the Internet, etc.

Secondly, action can be taken afterwards, when the phenomenon has occurred and then the statistics offices consider the nature of estimation in non-response situations. In general, there are two procedures to deal with the lack of information once collected. The first procedure is imputation, which consists of substituting the blank data for acceptable known data (Puerta 2002). This procedure is usually considered in cases of partial lack of response, which is to say when the individual has given some information.

The second procedure is the calibration or adjustment with auxiliary information and is usually used in cases of total non-response. It is carried out at the point of estimation, when the weights of the responses of individuals to surveys are calculated. The statistics offices then set out to study which is the auxiliary information available for respondents and non-respondents, which allows the improvement of the estimation which would be obtained, applying the appropriate estimation methods when a lack of response does not occur or when it is negligible.

Eustat mainly considers the study of non-response in two senses. Firstly, to measure the lack of response in response to standardised criteria, which allows rates between surveys from different statistics offices and also of a temporary manner to be compared.

As regards this aspect, this Technical Notebook shows the definition of the Survey on the Population in Relation to Activity (PRA), from the 4th quarter of 2004 to the 3rd one of 2006.

In reality this is a summary of the Work Notebook Standardisation and Systemisation of the Calculation of Response Rates (Eustat 2007) which explains in greater detail the methodology applied (Lynn et al. 2001), as well as the applied set of response rates (rate of co-operation, contact, rejection and eligibility).

Secondly, it sets out to study the method of calibration or reweighting for the lack of response, which allows for improvement in the estimations of the surveys, referring throughout to those aimed at households. This study has in turn the following two closely related aspects:

- that related to the estimators, in the sense that some may have more appropriate properties for the object of the study

- that related to auxiliary information, in the sense that there are criteria to select information which is more efficient to combine with the previous estimators

This Technical Notebook is centred on the second objective and dedicates a chapter to the analysis of the auxiliary information available. It is based on the methodology proposed by Statistics Sweden, with a series of conditions that must satisfy the information that exists for respondents and non-respondents. The trials based on this methodology are applied once more to the PRA.

In a subsequent chapter, we move on to the calibration of this survey using this auxiliary information. For the calibration specific software is used, the CLAN macro, developed in the SAS environment by Statistics Sweden. Prior to this, we also introduce the estimator that is to be used in situations of lack of response, the generalised linear (GREG). And finally, a section is dedicated to the demonstration of the results with the new calibration, which are not very different from those already published with the method of general calibration, with is evaluated in the conclusions.

Chapter 2

Standardisation and systemisation of response rates

EUSTAT has always studied and monitored the lack of response in household surveys. However, the need to establish common measurement standards or criteria has been put forward and it was decided to follow the European methodology published for this purpose by the Institute for Social & Economic Research, of the University of Essex (United Kingdom) (Lynn et al. 2001).

Taking said methodology as a reference, the incidences obtained in EUSTAT for household surveys have been encoded. In particular, the study of non-response has been carried out for the Survey on the Population in Relation to Activity.

2.1 ISER notation used and definition of response rate

Once the incidences are encoded with the ISER classification (Lynn et al. 2001), as shown in detail in the Work Notebook (Eustat, 2007), the following standard definitions are applied which will be used for the calculation of the rates. The number in brackets refers to the ISER classification.

I = Complete interview (1)

P = Partial interview (2)

NC = Non-contact (3)

R = Refusal (4)

O = Other type of non-response (5)

UC = Unknown eligibility, no contact (641,651,661 and part of 67).

UN = Unknown eligibility, no contact (61,62,63,642,652,662, 68 and rest of 67).

NE = Not eligible (7)

Ec = Estimated proportion of the contacted cases of unknown eligibility that are eligible

En = Estimated proportion of the non-contacted cases of unknown eligibility that are eligible

The definition of response rate is as follows:

Response rate indicates the proportion of interviews carried out with all the eligible cases. This is to say it shows the number of interviews carried out, whether complete or partial, over the total number of interviews carried out, plus the refusals, non-contacts, those classified as other type of non-response and, in a specific proportion, those possibly eligible, whether contacted or not.

The formula is as follows:

$$RR_O = \frac{I + P}{(I + P) + (R + NC + O) + e_C UC + e_N UN)}$$

In the aforementioned Work Notebook the rest of the rates (co-operation, contact, rejections or refusals and eligibility) are set out, with the results obtained.

2.2 Results obtained

The previous definitions have been applied to the sample of the Survey on the Population in Relation to Activity (PRA). The purpose of the survey is, in general, to find out the relation to activity of the population aged 16 and over.

As regards its sample design, The PRA is a longitudinal survey that is conducted every three months. The sample consists of 5088 dwellings in each of the different surveying quarters. Each dwelling remains in the panel for 8 rotation waves, which is equivalent to 2 years, after which the dwelling is removed from the panel. The sample is therefore renewed by an 1/8 each quarter.

Between the 4th quarter of 2004 and the 3rd one of 2006, response rates by quarter obtained were higher than 80%. In the 4th quarter of 2004, when the panel was completely renewed, the response rate was 83.9%. This rate fell slightly during the following 3 quarters and from 2006 onwards began to increase in all the quarters, reaching 86.9% in the third quarter of 2006.

Without making a systematic comparison of these results, it seems that the response rates could be considered as acceptable.

T.1 Response rates by quarter

Quarter/Year	Response Rate
4/2004	83,9
1/2005	82,1
2/2005	82,6
3/2005	82,8
4/2005	83,8
1/2006	85,1
2/2006	86,0
3/2006	86,9
MEDIA	84,2

Source: Eustat.PRA

Chapter 3

Analysis of the auxiliary information for reweighting

As was stated in the introduction, there are two main ways to treat the lack of response in surveys. One way is that of imputation, by which, using statistical methods, values are assigned to the missing information. The other way is that of reweighting. This consists of calculating new weightings, bearing the auxiliary information in mind, for the whole of the sample that has replied to the survey. The success of this reweighting consists of possessing efficient auxiliary information which somehow takes into account the lack of response mechanism.

The aim of this chapter is to determine the auxiliary information that is going to intervene in the reweighting. This means using it in the calculation of the weights of the individuals who have replied, in order to reduce the bias of non-response in the survey. To do so, we have applied the methodology put forward by Statistics Sweden, in their publication, "Estimation in presence of Nonresponse and frame imperfections" (Lundström and Särndal (2002).

In the aforementioned manual the conditions that should be fulfilled in order to determine the auxiliary variables to be considered are the following:

1. They must explain the variation of the probabilities of response

2. They must explain the variation of the main study variables

3. They must identify the most important domains

3.1 Auxiliary information and response rates

Firstly, a series of auxiliary variables were analysed to see if they fulfil the first principle. The auxiliary variables available for the study were: sex, age, province, size of household, number of active people in the household, size of municipality and age of oldest person in the household. Later the response rates per category for these variables were calculated, and also the combination of the auxiliary variables size of household and age of the oldest person in the household.

Below we can see the response rates of the categories of the auxiliary variables which initially show differences:

3.1.1 Response rate of the combination of the variables "size of household" and "age of the oldest person in the household"

In order to see more clearly the differences between the categories of the variables combining size of household and age of the oldest person in the household. Carrying out the corresponding study, it was decided to divide into four categories.

The selected categories were as follows:

1. Households with one person in the household where the oldest person in the household is under 45

2. Households with one person in the household where the oldest person in the household is over 45

3. Households with two or more persons in the household where the oldest person in the household is under 45.

4. Households with two or more persons in the household where the oldest person in the household is over 45.

Below is the corresponding table.

T.2.Response rate by size of household and age of the oldest person in the household

		20044	20051	20052	20053	20054	20061	20062	20063
All	0 - 44 year	85,3	84,2	83,5	84,0	85,2	87,3	87,6	87,6
	More than 45	83,6	81,5	82,6	82,9	84,0	85,2	86,5	88,0
	All	84,0	82,1	82,8	83,2	84,3	85,7	86,8	87,9
1 Person	0 - 44 year	65,9	64,5	65,9	67,7	68,3	71,1	71,0	73,4
	More than 45	79,7	78,9	79,8	80,8	81,9	83,3	83,7	84,9
	All	76,4	75,5	76,5	77,9	79,0	80,9	81,2	82,4
2 or more	0 - 44 year	86,7	85,8	85,0	85,3	86,5	88,4	88,7	88,7
	More than 45	83,9	81,7	82,8	83,1	84,2	85,4	86,8	88,2
	All	84,6	82,6	83,3	83,6	84,7	86,0	87,2	88,3

Source: Eustat.PRA

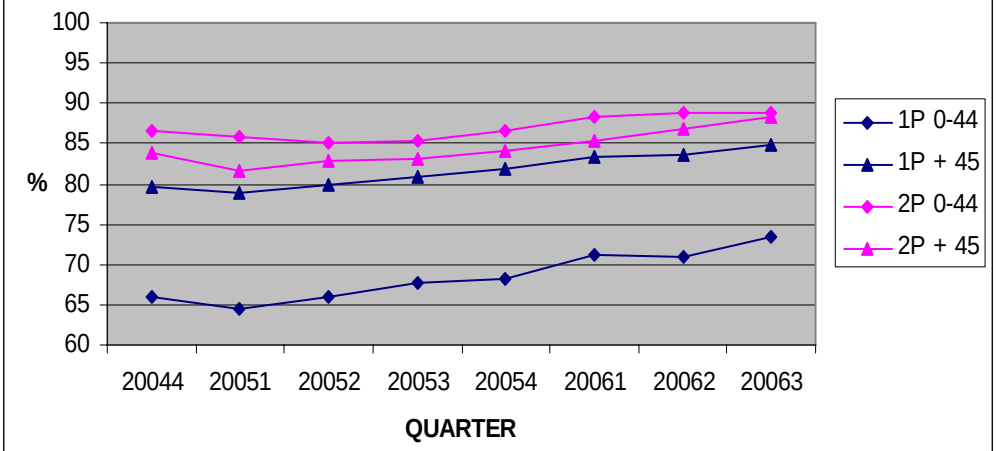
Analysing response rates in the new categories we see the following:

1. There is a great difference between the response rate of households with one person aged 0-44 and the rest. This difference is approximately 15 points compared to the other categories, which means it could be said that the households with only one person aged under 45 have a much lower response rate than the rest. This rate ranges between 66% and 73%, corresponding to the first and last surveying quarters respectively.

2. On the other hand would be the remaining three categories with a much higher response rate. Among these, the lowest corresponds to the households with one person aged over 45, followed by the households with two or more persons and the age of the oldest person from 0 to 44, and finally, households with two or more persons older than 45.

Below is the corresponding graph of the response rate by size of household and age of the oldest person.

G.1. RESPONSE RATES BY SIZE OF HOUSEHOLD AND AGE OF THE OLDEST PERSON IN THE HOUSEHOLD



3.1.2 Response rate of the variable "size of municipality"

The following table shows the evolution of response rates by size of municipality for each of the quarters.

Observing the 3 categories, the table shows that there were differences between the 3 types. The highest response rates corresponded to the municipalities with less than 5000 inhabitants, and in contrast, the lowest rates corresponded to the municipalities with the greatest number of inhabitants.

On the other hand, we can see that the rates of the smallest municipalities remained practically unchanged, while the response rate of the largest municipalities rose by 6 points.

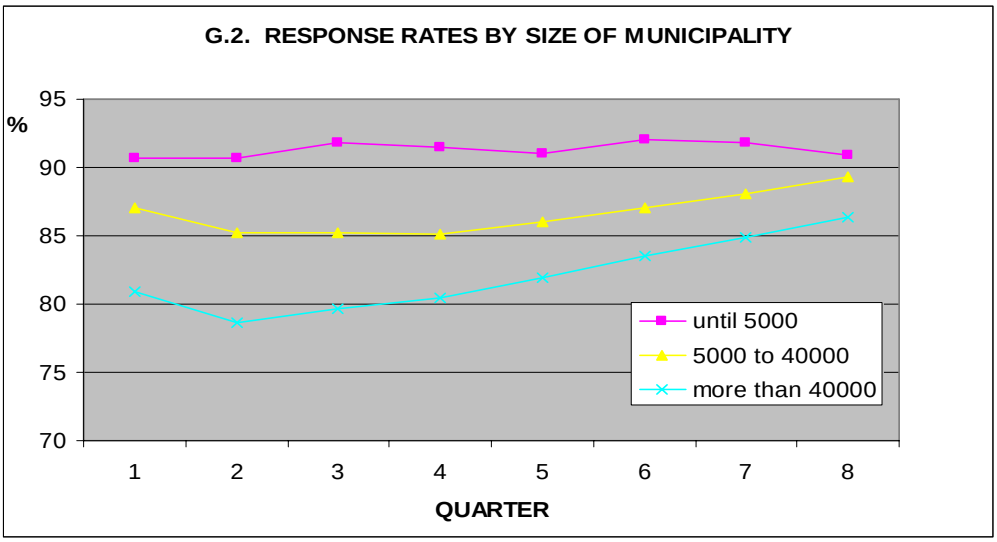
This variable will be a candidate for selection, due to the existence of differences between the different categories, although the differences were smaller in the last two quarters.

T.3. Response rate by size of municipality

Quarter	All	until 5000	5000 to 40000	More than 40000
20044	84,0	90,7	87,1	80,9
20051	82,1	90,7	85,2	78,7
20052	82,8	91,9	85,3	79,7
20053	83,2	91,5	85,1	80,5
20054	84,3	91,1	86,0	81,9
20061	85,7	92,1	87,1	83,6
20062	86,8	91,8	88,1	84,9
20063	87,9	91,0	89,3	86,3

Source: Eustat.PRA

The following graph shows the evolution of response rates by size of municipality in each quarter.



3.2 Auxiliary information and study variables

Once the variables that fulfil the first principle are chosen, the next phase within the selection of the auxiliary information is to see its effect on the study variables. Following the aforementioned manual, we need to check that the variables selected with the previous criteria also fulfil the condition that they have some relation to the study variables.

The study variable selected in this case is the variable “relation to activity” encoded in three categories: Employed, Unemployed and Inactive.

Below we can see the results obtained for the variables selected with the first criteria.

3.2.1 Auxiliary variable “size of municipality” by activity

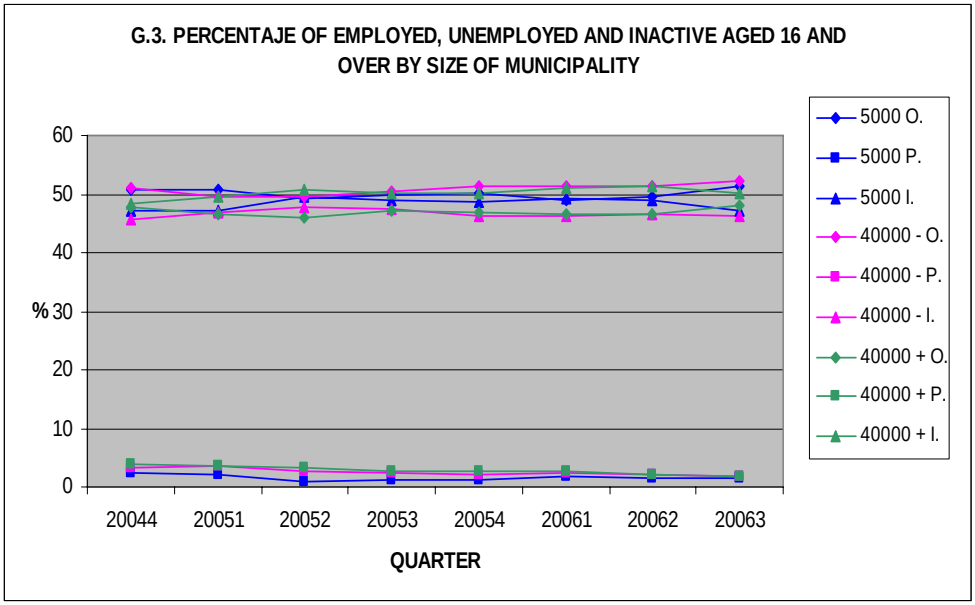
Analysing the variable “size of municipality” we see that the percentage distribution of employed, unemployed and inactive in municipalities of less than 5,000 inhabitants is similar to that of municipalities of 5,000-40,000 inhabitants. In municipalities of more than 40,000 inhabitants these differences are a little greater, but they are not significant.

T.4. Percentage of employed, unemployed and inactive aged over 16 by size of municipality

		20044	20051	20052	20053	20054	20061	20062	20063
All	All	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	Employed	48,9	47,8	47,4	48,3	48,6	48,2	48,4	49,6
	Unemployed	3,6	3,6	2,9	2,5	2,4	2,5	2,0	1,8
	Inactive	47,5	48,6	49,7	49,2	49,0	49,3	49,6	48,6
until 5000	All	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	Employed	50,8	50,7	49,4	49,8	50,0	49,1	49,5	51,4
	Unemployed	2,5	2,2	1,0	1,1	1,3	1,7	1,6	1,5
	Inactive	47,1	47,0	49,6	49,1	48,7	49,2	49,0	47,1
5000-40000	All	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	Employed	51,0	49,4	49,6	50,4	51,4	51,3	51,3	52,1
	Unemployed	3,4	3,7	2,6	2,2	2,2	2,5	2,0	1,8
	Inactive	45,6	46,9	47,7	47,4	46,4	46,3	46,7	46,1
More than 4000	All	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
	Employed	47,8	46,6	46,0	47,1	47,0	46,5	46,7	48,0
	Unemployed	3,8	3,7	3,3	2,7	2,7	2,6	2,1	1,9
	Inactive	48,4	49,7	50,7	50,1	50,3	50,9	51,2	50,1

Source: Eustat.PRA

Below is the corresponding graph of the percentages of employed, unemployed and inactive by size of municipality.



3.2.2 Combination of "size of household" and "age of the oldest person in the household" by activity

Now we will analyse the results of the combination of the auxiliary variables size of household and age of the oldest person in the household by activity.

Below we see the results by the activity variable. The following tables and graphs are disaggregated by each of the three categories of the activity variable: employed, inactive and unemployed.

3.2.2.1 Percentage of employed aged 16 and over by size of household and oldest person in the household

In the following graph and table the following can be seen:

1. As regards the households with one person, the percentage of employed of the people aged 0-44 is much greater than that of people aged over 45. The proportion of the former ranges between 83% and 89%, whereas the percentage of the latter ranges between 15% and 17%.

2. In households with 2 people or more, this difference is also significant, but not so much so as the previous case. On the one hand, the percentage of employed of those aged under 44 varied between 77% and 81%, and that of those aged over 45 is 44%.

3. Examining both results, we can see that in the 0-44 year-old category there are differences of approximately 6 or 8 points between the households with a single person in the household and the households with 2 or more people in the

household. However in the category of over 45 years old these differences are highly significant, with the percentage of employed in dwellings with two or more people being greater than those with one single person.

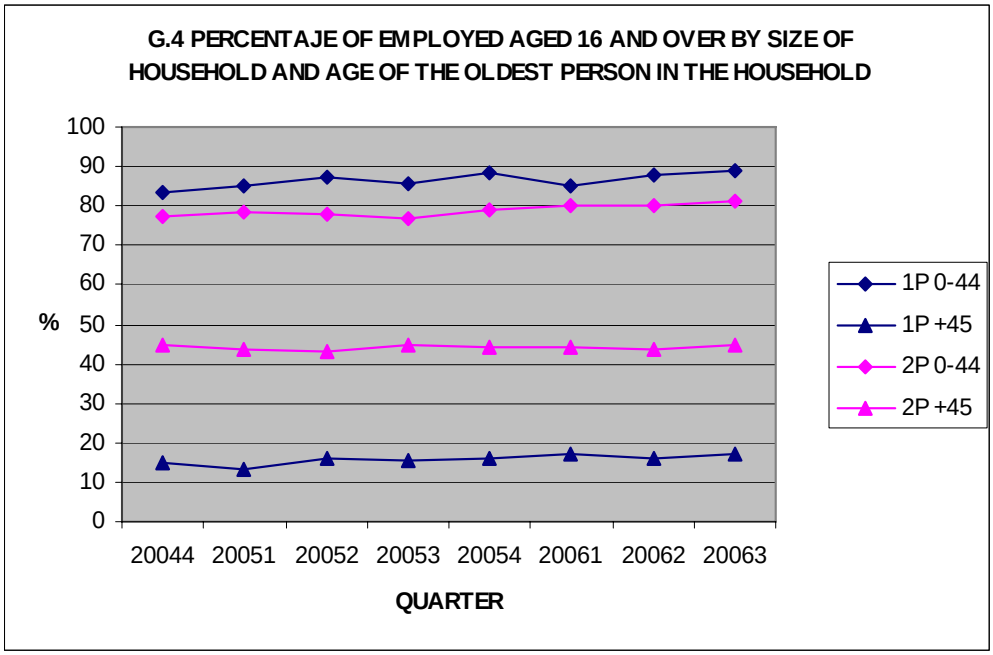
4. By way of a summary, it is worth pointing out that within the over-45s there are highly significant differences between the percentage of employed in dwellings with two or more people and the percentage of the dwellings with a single person in the household. In the category of under-45s this difference is not as large as in the previous case, but it is significant.

T.5. Percentage of employed aged 16 and over by size of household and oldest person in the household

	20044	20051	20052	20053	20054	20061	20062	20063
0 to 44	77,9	79,1	78,7	77,8	79,9	80,4	80,9	81,9
More than 45	42,5	41,5	41,1	42,3	42,0	41,8	41,5	42,5
All	48,9	47,8	47,4	48,3	48,6	48,2	48,4	49,6
0 to 44	83,6	85,1	87,4	85,7	88,4	84,8	88,0	89,2
More than 45	15,2	13,3	15,8	15,4	16,3	17,2	16,3	17,0
All	28,4	26,5	29,8	28,4	29,8	28,9	28,9	31,1
0 to 44	77,4	78,5	77,8	77,1	79,0	80,0	80,2	81,1
More than 45	44,5	43,7	43,1	44,5	44,3	44,0	43,8	44,9
All	50,5	49,5	48,8	50,1	50,3	49,9	50,1	51,3

Source: Eustat.PRA

The following graph shows the differences between the 4 categories:



3.2.2.2 Percentage of unemployed aged 16 and over by size of household and oldest person in the household

The table and graph corresponding to the percentage of unemployed shows the following:

1. On the one hand, we can see that the percentage of unemployed within the households where only one person lives is different for those aged under 44 compared to those aged over 44. The former have a rate that ranges between 3% and 7%, while the proportion of the latter ranges between 0.25% and 1.3%

2. On the other hand, examining the households of two or more people, we can conclude that the differences between these two categories are barely noticeable. Both categories vary between 2% and 4%.

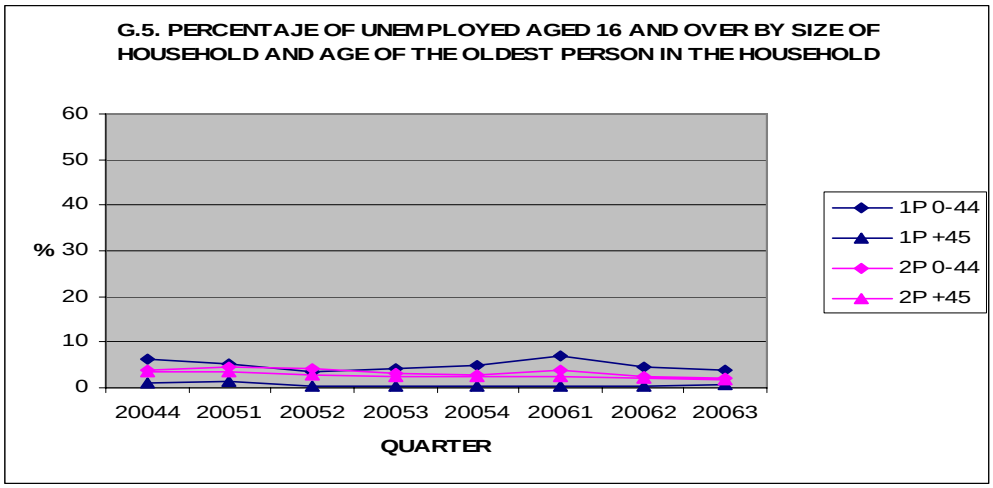
3. Therefore, comparing the two results, it is obvious that in the households with one person, the unemployment rate is much more marked, with those aged under 44 showing an unemployment rate that is somewhat higher than normal and those older than 45 a very low rate.

T.6. Percentage of unemployed aged 16 and over by size of household and oldest person in the household

		20044	20051	20052	20053	20054	20061	20062	20063
All	0 to 44	4,1	4,6	4,0	3,2	2,9	4,0	2,7	2,2
	More than 45	3,4	3,4	2,7	2,3	2,3	2,2	1,9	1,7
	All	3,6	3,6	2,9	2,5	2,4	2,5	2,0	1,8
1 Person	0 to 44	6,2	5,2	3,5	4,1	5,0	7,0	4,5	3,7
	More than 45	1,2	1,3	0,3	0,3	0,4	0,4	0,4	0,6
	All	2,2	2,0	1,0	1,0	1,2	1,5	1,2	1,2
2 or more	0 to 44	4,0	4,5	4,1	3,1	2,7	3,7	2,5	2,1
	More than 45	3,6	3,6	2,9	2,5	2,5	2,4	2,0	1,9
	All	3,7	3,7	3,1	2,6	2,5	2,6	2,1	1,9

Source: Eustat.PRA

The following graph shows the percentage of unemployed by size of household and age of the oldest person in each of the quarters.



3.2.2.3 Percentage of inactive aged 16 and over by size of household and oldest person in the household

Finally we see the results of the proportion of inactive people:

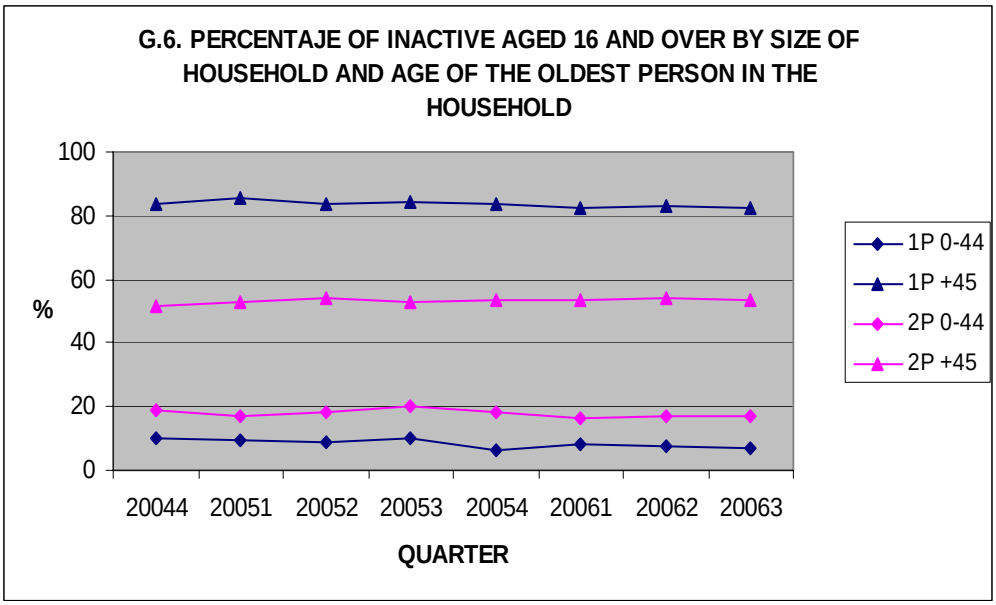
1. Firstly, examining the households where only one person lives, we see that there is a great difference between those aged over 44 and those aged under 45. The proportion of inactive people in the former varies between 82% and 85%, whereas that of the latter ranges between 7% and 10%.

2. On the other hand, analysing the results of the households with 2 or more people, there are also significant differences between those over and under 45, but these differences are not so extreme. The percentage of those aged over 45 inactive, ranges between 51% and 54%, while the proportion of inactive people aged under 45 ranges approximately between 16% and 19%.

T.7. Percentage of inactive aged 16 and over by size of household and oldest person in the household

		20044	20051	20052	20053	20054	20061	20062	20063
All	0 to 44	18,0	16,4	17,3	19,1	17,3	15,6	16,4	15,9
	More than 45	54,1	55,1	56,2	55,4	55,6	56,0	56,6	55,7
	All	47,5	48,6	49,7	49,2	49,0	49,3	49,6	48,6
1 Person	0 to 44	10,3	9,7	9,1	10,2	6,6	8,2	7,5	7,1
	More than 45	83,6	85,4	83,8	84,3	83,4	82,4	83,3	82,4
	All	69,4	71,4	69,3	70,7	69,0	69,6	69,9	67,7
2 or more	0 to 44	18,6	17,0	18,1	19,9	18,3	16,4	17,3	16,8
	More than 45	51,9	52,8	54,0	53,0	53,2	53,6	54,1	53,3
	All	45,8	46,8	48,1	47,4	47,2	47,5	47,8	46,8

Source: Eustat.PRA



To conclude the graphs and tables commented upon above, it is clear that **the percentages for unemployed, employed and inactive are different in households with one person in the household than in households with two people or more**.

In the three cases, households with one person in the household showed percentages at the extreme ends of the scale and in many of them the differences as regards the same age bracket in households with two or more people is considerable.

It should be pointed out that the differences that exist in the tables of employed and inactive, as regards the category of those aged over 45, among those households with one person or two or more people, are highly significant.

One example of this is that the proportion of employed in the category of those aged over 45 is much greater in those households with 2 or more people than in those with one. Conversely, the proportion of inactive aged over 45 is much higher in households with just one person.

Therefore, it has been decided to **select the combination of variables size of household by age of the oldest person in the household**, and rule out the variable size of municipality since the differences in the different categories are not significant.

3.3 Domains

After selecting the combination of the variables size of household by age of the oldest person in the household due to it also fulfilling the second principle, we move on to the following phase within the selection of auxiliary information.

The third step in the Swedish manual states that the auxiliary variables selected must identify the most important domains. To do so, we have added the population by

province, sex and age as auxiliary information, since these variables are used thus far to calculate the calibration weights and for their importance in the publication of results.

Chapter 4

Application of the calibration for non-response

This chapter will set out the results obtained after carrying out the calibration with the auxiliary information analysed. This calibration will be carried out with the CLAN macro, by Statistics Sweden, which uses a generalised linear estimator (GREG).

The auxiliary variables are determined that can intervene in the calculation of the weights of the individuals that have replied, which in our case are:

- on the one hand, the combination of the variables size of household and age of the oldest person in the household, and
- on the other, the combination of province, sex and age,

The CLAN macro was used to calculate the corresponding accurate estimations and the standard deviations of interest in the PRA. CLAN is a programme written in SAS macro language and was designed by the Swedish Statistics Office (Andersson and Nordberg 2007).

The CLAN macro can calculate the accurate estimations based on the Horvitz-Thompson (H.T) estimator or on the calibration or generalised regression estimator (GREG). The latter is an estimation method that uses auxiliary information in the estimation stage and has been used in the different trials that we have conducted. The idea of using auxiliary information is based on the correlation of the auxiliary variables with the study variable.

The auxiliary information is used for:

- o reducing sampling errors
- o reducing bias and variance due to the lack of response

The previous definitions have been applied to the sample of the Survey on the Population in Relation to Activity (PRA).

4.1 The generalised regression estimator (GREG)

The estimation of regression means that for element *k* in the sample, the pair (y_k, $\mathbf{x}_k$) is observed, where y_k is the value observed of *y* (the variable of interest), while $\mathbf{x}_k$ is the vector of auxiliary information. The methodology also requires that the population total of vector ***x*** is known.

For a more detailed description of regression estimators, see Särndal C, Swensson B and Wretman (1992). Below there is a brief outline of the GREG estimator in a situation of lack of response.

A random sample *s* of size n_s, is extracted from a population U consisting of N individuals, in accordance with the sampling design *p(.)*, where all the individuals have a probability >0 of being included in the sample. Due to a non-response, the data of the variable *y* can be collected only for a sub-set *r* of size m_r. The sample design *p(.)* in the PRA implies, for example, that the population is divided into H strata, where stratum *h* contains N_h dwellings. In each stratum *h*, a random sample of size n_h, is extracted giving all the dwellings the same possibility of being included in the sample.

The **regression estimator, in general** for a total $t_y = \sum_U y_k$ would be:

$$\hat{t}_y = \sum_r w_k y_k$$

y_k = the value of the variable *y* for the element *k.*

If we apply it to situations of non-response, the notation is as follows:

$w_k = g_k \times d_k$ = the weight depends on the sampling design, the auxiliary vector $\boldsymbol{x}_k$ and the model used for the adjustment of non-response.

$$d_k = 1/\left(\pi_k \hat{\theta}_k\right)$$

π_k = the probability of inclusion of individual *k*. In the PRA $\pi_k = \frac{n_h}{N_h}$ for all the individuals that belong to stratum *h.*

$\hat{\theta}_k$ = The probability of response estimated for individual *k*, $\hat{\theta}_k = \frac{m_h}{n_h}$.

In reality, what we have applied is $\hat{\theta}_k = \frac{m_{hg}}{n_{hg}}$, assuming that all the individuals of the stratum respond independently and with the same probability.

The strata were made taking into account the Response Homogenous Groups (RHG), obtained from the study of the auxiliary information, which is to say the groups formed by the combination of the variables size of household and age of the oldest person in the household.

This estimation is based on the assumption that the individuals respond independently to each other and with the same probability in stratum *h*. (in our case, within stratum h and each RHG)

$$g_k = 1 + \left(\mathbf{t}_x - \hat{\mathbf{t}}_x\right)' \left(\sum_r \frac{\mathbf{x}_k \mathbf{x}_k' q_k}{\pi_k \hat{\theta}_k} \right)^{-1} \mathbf{x}_k q_k$$

g_k, is a correction factor which reflects the contribution of the auxiliary information to reduce the bias due to non-response and sampling error.

$\mathbf{x}_k = (x_{1k}, \ldots, x_{jk}, \ldots x_{Jk})'$, is a vector of extension J, where J is the number of auxiliary variables.

q_k, is a known constant.

$\mathbf{t}_x = (t_{x1}, \ldots, t_{xj} .., t_{xJ})$, is a vector of extension J, which contains the known totals of the register.

$\hat{\mathbf{t}}_x = (\hat{t}_{x1}, \ldots, \hat{t}_{xj} .., \hat{t}_{xJ})$, contains the estimations of the totals. $\hat{t}_x = \sum_r d_k x_k$

The **variance, in situations of lack of response,** for $\hat{t}_y$, is estimated as:

$$\hat{V}(\hat{t}_{yGREG}) = \sum\sum \frac{\pi_{kl}\hat{\theta}_{kl} - \pi_k \hat{\theta}_k \pi_l \hat{\theta}_l}{\pi_{kl}\hat{\theta}_{kl}} w_k e_k w_l e_l$$

π_{kl}, is the probability of inclusion of second order.

$\hat{\theta}_{kl} = \frac{m_h}{n_h} \frac{m_h - 1}{n_h - 1}$, is the estimated probability that *k and l* belong to *r* (respondents)

$$e_k = y_k - \mathbf{B}'\mathbf{x}_k \qquad \mathbf{B} = \left(\sum_r \frac{\mathbf{x}_k \mathbf{x}_k' q_k}{\pi_k \hat{\theta}_k} \right)^{-1} \sum_r \frac{\mathbf{x}_k y_k q_k}{\pi_k \hat{\theta}_k}$$

4.2 Results obtained with the CLAN macro

To calculate the GREG estimator we used Response Homogenous Groups (RHG). The calculation of the RHGs was based on the auxiliary information selected in the study of the non-response, which is to say the combination of variables size of household and age of the oldest person in the household. Thus 4 Response Homogenous Groups were obtained:

1. Size of household= 1 person and age of the oldest person<45 years old.
2. Size of household= 1 person and age of the oldest person>=45 years old.

3. Size of household= More than one person and age of the oldest person<45 years old.

4. Size of household= More than one person and age of the oldest person>=45 years old.

Using RHGs in our case, each stratum (territory or province) was divided into the 4 Response Homogenous Groups. With this model of non-response it is assumed that the dwellings for each group respond independently and with the same probability of response. The population projections for province, sex and age were taken as additional auxiliary information for the regression estimator.

Results were obtained for all the series of quarters. Included here are those corresponding to the first quarter of 2005 (20051) and for the third quarter of 2006 (20063) with this method, the published results (without treatment of non-response), and the differences between them by province and sex.

The data shown is that relative to the total number of employed, total number of unemployed and the unemployment rate.

T.8. Employed and unemployed population and unemployment rate by province and sex. Results obtained, results published and differences between them.

			Clan			Argitaratua			BJA-CLAn aldeak		
Year/Quarter	Province	Sex	Employed	Unemployed	Unemployment rate	Employed	Unemployed	Unemployment rate	Employed	Unemployed	Unemployment rate
20051	All	Man	547,0	34,4	5,9	547,0	33,8	5,8	0,0	-0,6	-0,1
20051	All	Woman	392,7	36,3	8,5	389,1	36,9	8,7	-3,6	0,6	0,2
20051	All	All	939,7	70,7	7,0	936,1	70,7	7,0	-3,6	0,0	0,0
20051	Araba	Man	82,5	1,8	2,1	82,7	2,0	2,4	0,2	0,2	0,3
20051	Araba	Woman	55,9	2,5	4,3	56,6	2,8	4,7	0,7	0,3	0,4
20051	Araba	All	138,4	4,3	3,0	139,3	4,8	3,3	0,9	0,5	0,3
20051	Gipuzkoa	Man	179,8	10,6	5,6	177,9	10,4	5,5	-1,9	-0,2	-0,1
20051	Gipuzkoa	Woman	130,1	10,3	7,4	131,6	10,0	7,1	1,5	-0,3	-0,3
20051	Gipuzkoa	All	309,9	21,0	6,3	309,5	20,4	6,2	-0,4	-0,6	-0,1
20051	Bizkaia	Man	284,7	22,0	7,2	286,3	21,4	7,0	1,6	-0,6	-0,2
20051	Bizkaia	Woman	206,7	23,4	10,2	200,8	24,1	10,7	-5,9	0,7	0,5
20051	Bizkaia	All	491,4	45,4	8,5	487,2	45,5	8,5	-4,2	0,1	0,0
20063	Guztira	Man	556,7	19,0	3,3	556,1	18,6	3,2	-0,6	-0,4	-0,1
20063	Guztira	Woman	408,1	16,5	3,9	403,6	17,6	4,2	-4,5	1,1	0,3
20063	Guztira	All	964,8	35,5	3,6	959,7	36,2	3,6	-5,1	0,7	0,0
20063	Araba	Man	83,3	1,8	2,1	83,9	2,1	2,4	0,6	0,3	0,3
20063	Araba	Woman	60,7	2,5	3,9	59,3	2,5	4,0	-1,4	0,0	0,1
20063	Araba	All	144,0	4,3	2,9	143,3	4,6	3,1	-0,7	0,3	0,2
20063	Gipuzkoa	Man	187,2	4,9	2,6	183,9	5,1	2,7	-3,3	0,2	0,1
20063	Gipuzkoa	Woman	134,7	3,8	2,7	135,9	3,6	2,6	1,2	-0,2	-0,1
20063	Gipuzkoa	All	322,0	8,8	2,6	319,8	8,7	2,6	-2,2	-0,1	0,0
20063	Bizkaia	Man	286,2	12,2	4,1	288,3	11,4	3,8	2,1	-0,8	-0,3
20063	Bizkaia	Woman	212,7	10,3	4,6	208,4	11,4	5,2	-4,3	1,1	0,6
20063	Bizkaia	All	498,9	22,5	4,3	496,6	22,8	4,4	-2,3	0,3	0,1

Source: Eustat.PRA. (Note: The totals of employed and unemployed are given in thousands, and the unemployment rate in %)

The differences between the two methods are generally small ones, meaning that we cannot say that the adjustment of non-response using this treatment with the selected auxiliary variables influences the results obtained.

Chapter 5

Conclusions

This Technical Notebook has covered some aspects related to the lack of response, especially in household surveys. Firstly, is has dealt with the issue of how to measure the lack of response and secondly, the issue of adjustment or calibration when there is a lack of response in a household survey.

In relation to the first point, how to measure the lack of response, the standard proposed by the Institute for Social & Economic Research (Lynn et al. 2001) was applied to the Survey on the Population in Relation to Activity (PRA), which is a panel aimed at households.

One of the main conclusions at this point is the advantage of using a common or standard method of classifying incidences or field results for household surveys and also for the calculation of rates. Its use in different surveys allows the comparison between them, provided that the designs are similar and the differences in response rates are not totally attributable to their different characteristics.

As a consequence of the above, Eustat is currently working on the means of extending this to other surveys aimed at households. This work has two aspects: one, that related to the computing procedures for the calculation of these rates using the incidences; and the other, that related to the documentation and dissemination of the method. With this aim, a manual of incidences has been made, adapting the categorisation proposal by ISER. Given the increasingly generalised use of mixed methods of information gathering, we have made the most of this opportunity to extend the manual of incidences to telephone interviews.

Turning to the second point, calibration or adjustment for the lack of response in surveys, two aspects have been covered. Firstly, we have set out the method of Statistics Sweden for the selection of auxiliary information, available for respondents and non-respondents of a survey, so as to allow improved estimations in a situation of lack of response. This study was made for the variables available in the PRA and we selected those that best fulfilled the conditions, which, very briefly, are: being related to the lack of response and also with the study variables.

Secondly, we applied the generalised regression estimator to this auxiliary information, and the results were estimated again. This was done via the Statistics Sweden CLAN macro, which produces estimations, with its sampling errors. This means another step forward regarding the measurement of the level of lack of response.

The results that were obtained for the PRA, with these new estimations, are not in reality very different from those already published. This similarity of results could be interpreted chiefly in two ways. Although there is no direct correspondence between lack of response and bias, it could be understood that with a response rate higher than 80% in the panel of the PRA in the period under study (4th quarter of 2004 under the 3rd quarter of 2006), perhaps the bias that is caused is not very great. Another interpretation could be that the auxiliary information used in the study, the most appropriate of that available, according to the method, is not sufficient to introduce differences to the results.

The main conclusion of this phase of the study is that the method that has been applied here offers many possibilities. Especially in the immediate future, when administrative information is becoming more accessible and is being used more often in the administrative process. In fact, steps are already being taken to try to apply it to other auxiliary variables in this same survey.

Capítulo 6

Bibliography

CLAES ANDERSSON AND LENNART NORDBERG

A USER´S GUIDE TO CLAN 97.

1998, STATISTICS SWEDEN

CLAES ANDERSSON AND LENNART NORDBERG

SUPPLEMENT TO "A USER´S GUIDE TO CLAN 97".

2006, STATISTICS SWEDEN

EUSTAT

STANDARDISATION AND SYSTEMISATION OF THE CALCULATION OF RESPONSE RATES

WORK NOTEBOOK, 2007 http://www.eustat.es/document/datos/ct_16_c.pdf

SIXTEN LUNDSTRÖM AND CARL-ERIK SÄRNDAL

ESTIMATION IN THE PRESENCE OF NONRESPONSE AND FRAME IMPERFECTIONS

2001, STATISTICS SWEDEN.

SIXTEN LUNDSTRÖM AND CARL-ERIK SÄRNDAL

ESTIMATION IN SURVEYS WITH NONRESPONSE

2005, WILEY,

PETER LYNN, ROELAND BEERTEN, JOHANNA LAIHO AND JEAN MARTIN.

RECOMMENDED STANDARD FINAL OUTCOME CATEGORIES AND STANDARD DEFINITIONS OF RESPONSE RATE FOR SOCIAL SURVEYS.

ISER WORKING PAPERS NUMBER 2001-23.

AITOR PUERTA GOICOECHEA

CLASSIFICATION TREE-BASED IMPUTATION

TECHNICAL NOTEBOOK, EUSTAT, 2002
http://www.eustat.es/document/datos/ct_04_c.pdf

CARL-ERIK SÄRNDAL, BENG SWENSSON, JAN WRETMAN

MODEL ASSISTED SURVEY SAMPLING

1992, SPRINGER-VERLAG NEW YORK, INC.

.