THE HEALTH INTERVIEW SURVEY
1997

Protocol for the Selection of the Households and the Respondents

Center for Operational Public Health Research
Department of Epidemiology
Scientific Institute of Public Health

National Institute of the Statistic

Department of Biostatistics
University of Limburg

S.P.H. / EPISERIE N° 12

THE HEALTH INTERVIEW SURVEY
1997

Protocol for the Selection of the Households and the Respondents

P. Quataert (S.P.H.)
H. Van Oyen (S.P.H.)
J. Tafforeau (S.P.H.)
E. Schiettecatte (N.I.S.)
L. Lebrun (N.I.S.)
L. Bellamammer (N.I.S.)
G. Molenberghs (L.U.C.)

S.P.H. / EPISERIE N° 12

Reference : Health Interview Survey, 1997. Protocol for the selection of the households and the respondents. P. Quataert, H. Van Oyen, J. Tafforeau, E. Schiettecatte, L. Lebrun, L. Bellamammer, G. Molenberghs. S.P.H. / EPISERIE N° 12, S.P.H., Brussels, 1997.

D/1997/2505/09

This report was commissioned by Ministries of Health of the Federal Government, the Flemish Community, French Community, the Walloon Region, the Brussels Region and the German Community.

Table of Contents

1. Introduction

Each observation can be seen as the result of a two step process. First a subject (a sampling-unit, a respondent, ...) is *selected*, then various characteristics of interest are *measured* on this subject:

Observation = Selection + Measurement.

This document develops the design of the first step, i.e. the selection. The theoretical considerations of the selection process have been elaborated in a previous report(1). The design for the second step, the development of the questionnaire of the 1997 Belgian Health Interview Survey (HIS), is discussed elsewhere (2).

The selection will be tackled from a double perspective:

- the development of the sampling scheme; i.e. the mechanism to get a random sample of households and respondents.
- the rules to contact the respondents; i.e. the various aspects of the fieldwork such that the respondents sampled are retrieved and effectively interviewed.

The final sampling scheme of the households and respondents is a combination of many sampling techniques : stratification, multistage sampling and clustering[1]:

- First there is a *regional stratification*. Belgium is divided into 3 regions, the Flemish Region, the Walloon Region and the Brussels Region, for which the number of interviews has been predetermined. The reason for this stratification is to be able to produce results by regions.
- The second *stratification* is at the level of *the provinces*. This second level of stratification is done to improve the quality of the sample over a simple random sample. In particular a geographical spread is achieved. The sample size within the provincial stratification is proportional to the population size of the province. There is the special case of the province of Liège as the sample size of the German Community (which is geographically located in the province of Liège) has been predetermined.

- Then within the strata units are accessed in two (for the households (HH)) or three (for the individuals) stages:
  - First, within each stratum municipalities are selected with a selection chance proportional to their size. These municipalities are called the *Primary Sampling Units* (PSU). Each time a PSU is selected a group of 50 individuals have to be interviewed successfully during the year 1997.

---

1 A detailed discussion is given in the chapters I to IV. The concepts are defined and explained in the appendices.

- Then within each municipality an equiprobable sample of households, the *Secondary Sampling Units* (SSU), is drawn such that 50 individuals per group can be interviewed in total.
- Finally within each household at most four individuals, the *Tertiary Sampling Units* (TSU), are chosen.

To reflect all these aspects the presentation and argumentation of the protocol consists of five chapters :

I. Stratification ;
II. Selection of the municipalities in each stratum ;
III. Selection of a cluster of households within each municipality ;
IV. Selection of a cluster of respondents within a household ;
V. Contacting the respondents.

For each chapter the discussion is divided into three parts:

- First the relevant requirements (related to the general objectives) and the most important boundary conditions (in practice) are discussed [2].
- Then the sampling scheme is described along with a short discussion why some choices are made.
- Finally annotations are made. This can be a further elaboration of some arguments, results of the sampling, some additional requirements or points of attention, alternatives not chosen, cross-references to the other chapters,...

To preserve the logical flow of the text, some technical aspects are moved into the appendices.

Under the heading 'Selection in a Nutshell' (appendix 1.1) a short overview of the main issues of the sampling is presented. Appendix 1.2 serves as a 'file rouge' or guide throughout this document. It consists of a systematic compilation of the possibilities and limits of the fundamental options and tools used for sampling for the HIS. Those remarks originated from an ad hoc working group.
In the second part of the appendices a detailed description of the sampling schemes used for the 1997 HIS is given.

[2] See also in the appendix where the requirements and boundary conditions are discussed from another perspective.

I. Stratification

Requirements and Boundary Conditions

Following elements were specified in advance:

- a total number of successful interview, equally spread over the year 1997, is predetermined to be 10000 ;
- for the three regions of Belgium (Flemish region, Walloon region and Brussels region), the number of individuals who need to be successfully interviewed is fixed : 3500, 3500 and 3000 respectively {10126} ;
- additionally, within the Walloon region an oversampling should be done for the German Community of Belgium (in the district Eupen-Malmédy). The total number of successful interviews is 300.

Further it was decided to fix the number of interviews by province proportional to their size[3] to:

- allow the provinces for an oversampling in a transparent way (currently, this option is not followed) ;
- prevent problems related to the over- or underrepresentation of some provinces by chance in the random sample. This may be especially the case for smaller provinces as the probability with which this may occur is a function of the population size of the province.

Also one should take into account that in order to keep the fieldwork manageable the number of interviews to be done in each municipality should be at least 50. Hence it was decided to work with multiples of 50 within each stratum.

The Sampling Scheme (a subdivision into strata)

To comply with the requirement above, Belgium was subdivided into 12 strata:

- 5 provinces in the Flemish region (5)
- the Brussels region (1)
- 5 provinces in Walloon region (the province of Liège without the German Community in the district Eupen-Malmédy) and the German Community (5 + 1).

---

3 This does not provide a guarantee that the results can be extrapolated on the level of province. Therefore an larger sample is necessary in most cases.

As already stated :

1. for the Brussels region and the German Community the number is fixed to be respectively 3000 and 300.
2. For the other strata (the provinces) the number of interviews is distributed proportional to the size of the provinces within each region but such that the numbers are multiples of 50 and add up to 3500 in both the Flemish and Walloon Region (3200 for the Walloon Region without the German Community and 300 for the German Community).

To achieve this some rounding was necessary and as a consequence the probability for selection within each region was not perfectly equal.

The results are represented in table 1. For each stratum the following items are specified: the population, the fraction relative to the total number of inhabitants of the region, the theoretical number of individuals to be interviewed , the effective number of individuals to be interviewed (a multiple of 50), the corresponding number of groups of 50 individuals to be interviewed and the probability of being selected (the sampling rate/1000).

Notes

From the table it can be derived that the chance of being selected differs appreciably from region to region. In total 10 000 interviews will be done. Hence - as the total population of Belgium is about 10 000 000 - the overall chance of being selected is almost 1/1000. In the Flemish region the relative chance is about a half (0.6), in the Walloon region it is close to 1. In the Brussels region this factor is 3.2; i.e. that the chance of an individual in the Brussels region to be selected is approximately 3.2 times the overall probability of being selected. Within each region the selection probabilities by province differ too because of the rounding process described above. This variation is very small, with the exception of the German Community where the chance of being selected is augmented by a factor of 4.3 because of the predetermined oversampling.

It should be stressed that this difference in selection probability does not affect the re-presentativity of the samples. Instead by taking in each region nearly the same number of respondents, the precision and hence the quality of resulting inferences at regional level is made about equal. However for the combination of the results to the national level some precision is lost but a valid estimate and inference can be obtained by reweighting each region by stratum with weights inversely proportional to the selection probability.

Table 1. The distribution of the sample size by provinces.

Province	**(A) Population***	**(B) Fraction (%)**	**(C) Theoretical number of individuals to be interviewed**	**(D) Effective number of individuals to be interviewed (multiple of 50)**	**(E) Number of Groups of 50 individuals**	**(F = (D/A) * 10^3) The probability for an individual to be selected**
Antwerpen	1631243	27.7	971	950	19	0.58
Vlaams Brabant	999186	17.0	595	600	12	0.60
Limburg	775302	13.2	461	450	9	0.58
Oost-Vlaanderen	1351777	23.0	805	800	16	0.59
West-Vlaanderen	1122849	19.1	668	700	14	0.62
Flemish Region (5 strata)	**5880357**	**100.0**	**3500**	**3500**	**70**	**0.60**
Brabant Wallon	339062	10.5	334	350	7	1.03
Hainaut	1284761	39.6	1267	1250	25	0.97
Liège without the German Community (GC)	944291	29.1	931	900	18	0.95
Luxembourg	241339	7.4	238	250	5	1.04
Namur	435677	13.4	430	450	9	1.03
Walloon Region without the GC (5 strata)	**3245130**	**100.0**	**3200**	**3200**	**64**	**0.99**
German Community (1 stratum)	69438	100	300	300	6	4.32
Walloon Region (6 strata)	**3314568**	**100**	**3500**	**3500**	**70**	**1.06**
Brussels region (1 stratum)	**948122**	**100.0**	**3000**	**3000**	**60**	**3.16**
BELGIUM (12 strata)	**10143047**	**100.0**	**10000**	**10000**	**200**	**0.99**

*: NIS (population 01.01.1996)

(C) = (3500 * (B))/100 within the Flemish region; (C) = (3200* (B))/100 within the Walloon Region without the GC

II. Selection of the municipalities (PSU)

Requirements and Boundary Conditions

In a first step the municipalities are selected within each stratum. This selection is made for the whole year (1997) at once. In this way time is less confounded with place and it also facilitates the planning of the fieldwork (e.g. the recruitment of the interviewers).

To guarantee the efficiency of the sample, some additional rules should be built into the random selection:

- The chance of selection of a municipality should be proportional to its population size.
- The larger cities and metropoles should be included at least once with certainty in the sample. The number of times a city or a metropolis has to be included is determined by its actual population size.
- A similar remark holds for the smaller towns and villages. Also from this group elements should be present. By grouping smaller communities through ordering the whole set of communities according to size the representation of smaller communities out of the pool of smaller communities is ensured. The assumption made is that smaller communities of about the same size are exchangeable with respect to the items of interest.

The Sampling Scheme

Above requirements are achieved by a weighted systematic sampling where the municipalities are ordered (from large to small) and expanded proportional to their size (area probability sampling). As a consequence the chance for a municipality to be selected is proportional to the number of inhabitants. By ordering and systematic sampling also one implicitly stratifies the municipalities in blocks of a certain size and from each block just one municipality is chosen[4]. The sample will contain municipalities ranging from small to large. Further the systematic sampling guarantees that the larger cities are selected with certainty. In fact some large cities will be selected more than once because their size is a multiple of the step size by which the systematic sample is taken.

[4] As example, the selections of the municipalities of the province of Namur is worked out in detail on page 47.

The list below specifies the towns which are selected at least once :

Flemish region: 5: Antwerpen; 2: Gent; 1: Leuven, Brugge

Walloon region: 3: Charleroi; Liège; 2: Namur;
1: Mons, Mouscron, La Louvière, Seraing, Tournai, Verviers, Eupen

Brussels region: all municipalities are selected at least once (see page 63 for details).

Notes

- By the sampling scheme described above a random sample of municipalities was selected. The result is summarised in a map and table 2. This provides an overview of the geographical dispersion of the selected municipalities The detailed list of the municipalities and the number of times they are selected is given on page 57.

- Of the 589 municipalities 144 are selected. The number of municipalities (PSU) selected is smaller than the number of groups (200) of 50 individuals because several municipalities are selected more than once.

Selected municipalities (PSE) en number of groupes per PSE, Health Interview Survey, Belgium, 1997.

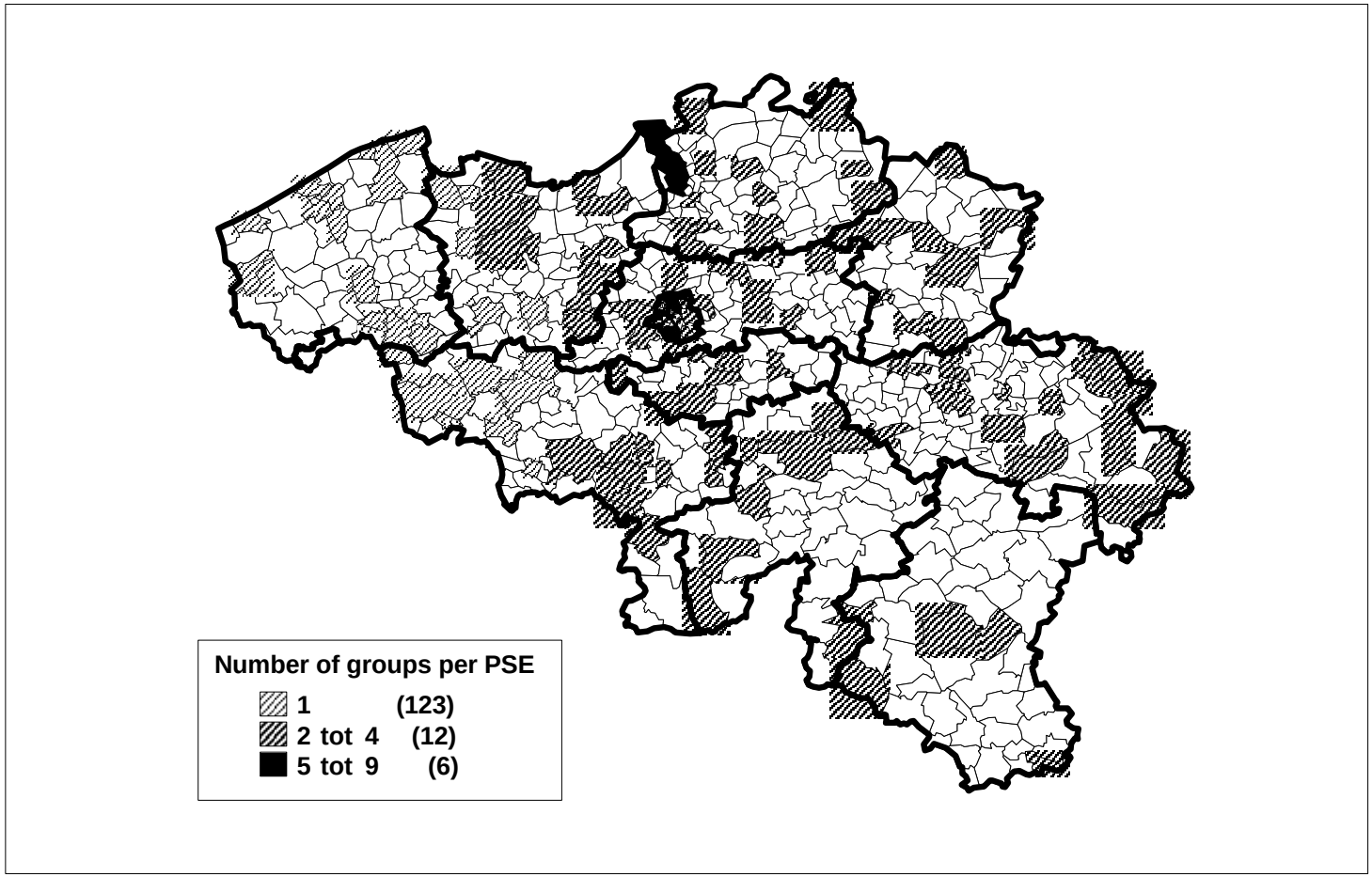


Table 2. The number of selected municipalities (PSU) by provinces.

Province	**Effective number of interviews**	**Number of groups of 50 individuals**	**Number of municipalities (PSU)**	**Municipalities which are selected more than once**
Antwerpen	950	19	14	Antwerpen (6 times)
Vlaams Brabant	600	12	12	
Limburg	450	9	9	
Oost-Vlaanderen	800	16	14	Gent (3 times)
West-Vlaanderen	700	14	14	
Flemish Region	**3500**	**70**	**63**	
Brabant Wallon	350	7	7	
Hainaut	1250	25	21	Charlerloi (4 times) Mons (2 times)
Liège without the German Community	900	18	15	Liège (4 times)
Luxembourg	250	5	5	
Namur	450	9	8	Namur (2 times)
Walloon Region without the German Community	**3200**	**64**	**56**	
German Community	300	6	6	
Walloon Region	**3500**	**70**	**62**	
Brussels Region	**3000**	**60**	**19**	see page 63
BELGIUM	**10000**	**200**	**144**	

III. Selection of the Households (SSU)

Requirements and Boundary Conditions

At this stage the municipalities in which the households have to be sampled are known and the corresponding number of groups (or the number of respondents to be interviewed (a multiple of 50)) is given. The sampling frame is a copy of the National Register. For the first quarter of 1997 the copy of the National Register is of the first of January 1996. For the 3 other quarters of 1997 the copy of the National Register is of the first of January 1997. The units of this sampling frame are the reference person[5] of the households: in this step households are selected.

Because about 6 % to 20 % (in Brussels) of the people move every year, the information of the copy of the National Register will be rapidly out of date. As the cost for a copy is rather high (approximately 400 000 Bfr), only two different copies will be used:

- sample 1 (1st quarter of '97): copy of the first of January '96
- samples 2 - 4 (2nd, 3rd and 4th quarter of '97): copy of the first of January '97

To cope with the decaying quality it was decided to control systematically the vital status of the reference person and the address with the latest available data (within approximately one month before contacting the household) in the National Register. Household not included in the sampling frame are those which are newly created or moved recently into the municipality.

To plan the sampling of the households two further points should be taken into account:

1. Not all households sampled will result in an interview. There may be a considerable drop-out (about 50 %). The reason can vary from not eligible (e.g. moved out the PSU) over impossible to locate to a refusal. To compensate for this, an a priori 'larger than necessary' sample is required.

5 The reference person is the administrative reference of the household within the National Register. Although it may be in general an adult member of the household, it is not necessary so. It does not refer to the concept of 'head' of the household.

2. The units of the sampling frame (the National Register) are households with a variable number of members. This is an issue because not the number of households but the number of individuals to be interviews is fixed. Hence an estimation should be made of the number of households needed. Additionally one should take into account that a (small) fraction of the household members will refuse. This shifts the distribution of the (effective) household members to the left and increases the variability further.

1. To tackle, at least partially, systematic trends in drop-out, it was decided not to replace the households in a simple random fashion, but to seek for matches based on :
   a) the statistical sector[6] within the municipality ;
   b) the size of the household and
   c) the age of the reference person[7].

   To facilitate the organisation of the fieldwork three matches per household will be generated immediately. A group of matched quadruples is defined in this document as a cluster of households.

2. To cope with the variable household membership, twice the expected number of clusters of households (the household with its 3 possible replacements) will be sampled. Hence the total sample size will be 8 times as large. The expected number of households can be calculated, because for each municipality the distribution of the individuals by household is known and hence the average number of household members. This household membership factor varies between 1.8 (Brussels) and 2.5 to 3 (smaller municipalities). Then the expected number of households needed per group, equals 50 divided by the mean household size. E.g. if the mean household size is 2.5 then the number of households needed is 50 / 2.5 or 20 households.

---

6 The Belgian municipalities are divided into statistical sectors. The aim of this division is to create more homogenous quarters with respect to demographic, economic and social characteristics as well as related to housing and transport.

7 These 3 variables are the only relevant information available in the sampling frame. Other parameters cannot be trusted sufficiently.

Sampling Scheme: a clustered systematic sample of households of the National Register.

To sample the households the reference persons in the national Register are ordered hierarchically by :

- statistical sector ;
- the size of the household and
- the age of the reference person.

In this way households close to each other according to the first variable will be located close to each other on the list. To achieve the same desirable feature for the other variables, the order of the variable j + 1 is alternated when the level of the variable is changed from the current to the next level (j = 1,2). For example, in statistical sector 1, the household sizes are given in a increasing order, while they decrease in statistical sector 2. For household size 1 in statistical sector 1, the reference persons are listed according to increasing age, while the ages decrease for the households of size 2 in statistical sector 1. A schematic presentation is given in table 3.

Table 3. The ordered sampling frame for the selection of the households

PSU	**Statistical sector**	**Household size**	**Age reference person**
PSU : X	1	1	youngest ... ... oldest
		2	oldest ... ... youngest
		3	youngest ... ... oldest
		4	...
		4+	
	2	4+	
		4	
		3	
		2	
		1	
	3	1	
		...	

From this list a clustered systematic sample is taken. Instead of taking one household in each step of the sampling, a series of four consecutive households is taken. In this way one obtains the four matched households required. To account for the additional uncertainty that the number of personal interviews can be estimated from the number of households, but not determined with certainty (variable household membership, within household refusal), one has to ensure that a larger number of matched quadruples is selected to enable the inclusion of additional households. By setting the step size to half its original value the number of clusters is doubled.

With this sample a four columns table (see page 66) is formed where each row represents a cluster of the (four) households. There are as many rows as there are clusters selected. The first column is the first household to contact. If this household is not eligible or if it does not result into an interview, the next household (in the second column) is contacted and so on. When all the households of the cluster are used and further replacement is necessary, the first eligible household of the next available row is selected. To prevent any order effect the households within each row are randomised. Also the rows themselves are randomised. Then there are no row-effects and it is possible to work from top to bottom until a sufficient number of interviews is realised.

Above sampling scheme solves many issues at once:

- By taking a systematic sample from an ordered list, it is ensured that the characteristics of the sample will be close to that of the municipality with respect to the variables statistical sectors, household size and age of the reference person.
- By taking in each step a cluster of four households, one forms in a natural way homogeneous groups of households which can be used to replace each other in case of non-response. If none of the four households in a row result in an interview, then a new row can be started. The latter is only necessary if the other rows already started are not sufficient.
- By taking twice as many clusters as needed the variability of the household size is anticipated. One starts with a first random group of the selected clusters. If with this group the required number of interviews is not achieved, then the next row (cluster) is started.
- By making a list in advance, the organisation of the fieldwork is facilitated because no algorithm is necessary to decide about the next replacement and all information about contacting is present.

An algorithm is developed to link the number of households to be contacted with the household size as known from the National Register and in a second phase with the real household size. This is necessary because only 12.5 individuals may be interviewed per quarter.

- The algorithm continues to select the next eligible cluster until the sum of the individuals in the households to be interviewed is the closer to 12.5, with possible range 11 to 14. In case of equal distance to 12.5, there is an random process to remain beneath 12.5 or to exceed it.
- The program identifies the replacement household.
    - In case of a household non-response, the replacement is the next eligible household within the cluster. Once a cluster is initiated, the algorithm continues to select the next eligible household within the cluster independently of the number of successful interviews attained. So once a cluster is started, all efforts are taken to have a successful contact with a household within the cluster.
    - In case there are no more eligible households in the cluster or there are less interviews within the household than expected the next eligible cluster(s) will provide the replacement household(s). In case of equal distance to 12.5 the household is selected to reach the upper value.

Notes

1. *Spread of the interviews*

The interviews should be distributed as evenly as possible over the year. In this way possible confounding between time and place is reduced. To keep the organisation of the fieldwork feasible, it was decided to control the number of interviews at least by quarter. Hence every three months on average 12.5 interviews per group should be realised. It is clear that this number will vary inevitably. As two samples (a first for the first quarter, a second for the 2^{nd} to 4^{th} quarter) are taken, there is a (small) chance the same respondents are selected. To control for this following steps are used :

- the second sample for the 2^{nd} to 4^{th} quarter is taken, disregarding the fact that there may be an overlap ;
- once the sample has been taken, it is checked against the sample of the first quarter. Duplicates are removed. This procedure has the advantage that one does not need to remove units from the whole 1997 National Register database.

Statistically this amounts to a somewhat elaborate version of sampling without replacement. Of course, the previous statement overlooks the fact that the 1996 National Register database and the 1997 National Register database differ. In particular, an new household in 1997 has got a slightly increased selection probability. The influence of this phenomenon is minor, compared to most other issues (non-response, households in the population but not in the sampling frame) that it is probably not necessary to account for it.

2. *Oversampling*

The oversampling is not a problem. Within this sample a random sample of the households to be contacted is selected. However one should take extreme care not to replace a household without using strict criteria. Otherwise difficult to contact households will be substituted by easy to contact households.

3. *Matching of households for replacement*

An alternative to the method described above to form clusters of 4 households is to make a random sample and to cluster afterwards, e.g. by introducing a distance formula:

$$d(o,r)^2 = \frac{(ss_0 - ss_r)^2}{spread(ss)} + \frac{(age_o - age_r)^2}{spread(age)} + \frac{(size_o - size_r)^2}{spread(size)}$$

where ss_0 - ss_r : distance between statistical sectors. This would be a qualitative variable. The other two distances would be quantitative variables

An alternative for the ordering in the systematic sample was to use the addresses. In this way the geographical spread is maximised. However matching on household size and age was considered more relevant.
These last methods were considered to be complex and not offering substantial advantage over the method used to select clusters with comparable households with respected to the 3 selected variables.

4. *Verification of the vital status and the address of the reference person*

Seen from the perspective of the reference person (which is the sampling unit and the key to the respondents) two situations are possible:

<u>1. Reference person died</u>		
a: household size = 1	: household died	==> stop
b: household size = 2 +	: household remains	
	* remained at same address	==> household is still eligible
	* moved within PSU	==> household is still eligible
	* moved out of PSU	==> household is no longer eligible
	* HH can not be located in National Register	==> household is no longer eligible
<u>2. Reference person moved</u> (for any reason: total HH moved, only reference person moved (e.g. divorce)	* moved within PSU	==> household is still eligible
	* moved out of PSU	==> household is no longer eligible
	* HH can not be located in National Register	==> household is no longer eligible

IV. Selection of the Household Members (TSU)

Requirements and Boundary Conditions

At most four members of the household[8] will be interviewed because :

- interviewing more persons is inefficient because of the familial correlation: members of the same family tend to resemble each other more closely than members from different households. By augmenting the number of interviews nearly no new information is obtained for the global sample ;
- the burden on the household would be to large.

Hence if a family contains more than four members a selection rule is necessary. This selection should in principal be random. Always selecting the reference person of the household might lead to bias since the reference person is not a random member of the household. He or she might have special characteristics. In the literature this person is sometimes denoted as the gatekeeper. Even including the partner might not totally compensate for this. However there are some practical limitations :

- it may be difficult to explain that the reference person will not be interviewed, while other members are and
- there is a general household questionnaire. This information should come from the reference person (or the partner).

Therefore following selection rules are used within a household to select the individuals to be interviewed.

1. In a household of no more than 4 members all individuals are interviewed.
2. In a household with 5 or more members only 4 members will be interviewed :
    - in a household with a reference person and a partner both the reference person and the partner are interviewed and only two additional individuals will be selected using the birthday rule ;
    - in a household with a reference person without a partner, the reference person and 3 additional members, selected using the birthday rule, are interviewed.

[8] Selecting individuals through the houshold and interviewing more than one person in a houshold has a substantial impact on the organisation and the cost of the fieldwork. However as members of a houshold are more likely to be alike than individuals out of different household there is a loss of precision and the variance of the estimate will be larger, especially for those variables with a strong familial correlation.

Using the birthday rule the individuals having their birthday first (month, day) from the date of the first contact onwards, are selected.

Within a household (SSU) that agreed to take part in the survey, non-response at personal level (TSU) is still possible. Possible reasons are:

- refusal
- unable to participate (children, mentally disabled persons, ...)
- one of the members is not at home for a (extended) period (e.g. students, persons in a hospital, outside the country, ...).

It is specified for which cases a proxy is allowed:

1. Obligatory:
   person younger than 15 years ;
   person older than 60 with negative score on the introductory question ;
   person too sick or with mental disabilities.
2. Person cannot be reached for an extended period (at least more than 1 month).
3. Person refused and does not refuse that a proxy answers for him/her.

Sampling Scheme

If selection is necessary, the reference person and his/her partner is selected automatically. A randomisation will be done for the two or three remaining persons only. The selection itself is based on the birthday-rule. The two or three persons having their birthday first from the date of the first contact onwards, are included in the sample.

When the reference person (and the partner) is always selected but the other members have a probability of less than 1 to be selected, then the selection probability between members of the same household with a size of at least 5 is variable. This difference needs to be taken into account using the appropriate weights in order to avoid the potential bias as mentioned above :

- For a household of size k = 1,2,3,4 there is no additional weight required as everybody will be taken and the probability of selection for all members is 1.
- For a household of size $k \geq 5$ with reference person and partner the selection probabilities once the household is selected are :
  1. for the reference person : p = 1 ;
  2. for the partner : p = 1 ;
  3. for the k - 2 remaining persons : p = 2/(k - 2).

  The inverse of these quantities should be multiplied to the weight already calculated for this household.

- For a household of size $k \geq 5$ with reference person but without partner the selection probabilities once the household is selected are :

    1. for the reference person : p = 1 ;
    2. for the k - 3 remaining persons : 3/(k - 1).

    Again, the inverse of these quantities should be multiplied to the weight already calculated for this household[9].

Concerning the non-response the rule is that NEVER replacements are allowed because bias is very probable here (e.g. household members having less time will be more reluctant to answer and will be replaced by members having more time).

Notes

1. Enumeration of the Household Members

The selection within a household should be done by the interviewer as one cannot fully rely on the information of the National Register[10].

It is important to know the precise number of members of the household, because this determines the inclusion probability. A danger is that some members of the household are not enumerated (e.g. a child living partly in an institution, ...).

2. Conclusion / exclusion criteria

Collective households.

People living in a small collective household with at most 8[11] persons (e.g. small religious community) consider themselves as a household, hence they should be treated as such.

Individuals living in other collective household (e.g. prisons, cloister, health institutions) are not eligible for this survey.

---

[9] This solution is intuitively appealing since it amounts to extrapolating the 4 interviews back to the whole household. The reference person (and the partner) represent themself, while the 2 (3) other individuals represent the rest of the household. In the case of the household with reference person and partner, the sum of the weights equals k [= 1 + 1 + ((k-2)/2) + ((k-2)/2)] . There the net effect of the procedure is that the corresponding potential source of bias is completely eliminated. However there is a small loss in efficiency.

[10] If the information of the National Register is considerably different from the household information it may affect the quota of the 50 of individuals to be interviewed within the group. If the quota would not be reached, a new household from the next available cluster of household needs to be interviewed.

[11] To consider 8 persons in a collective household as the cut-off is an arbitrary cut-off.

Immigrants.

Immigrants with the children speaking one of the three national languages : when the communication is sufficient (e.g. children of 15 years or older) then they can be interpreters. The interviewer should also contact the field manager is case communication is too difficult because of language problems.

Elderly people . Following rules were agreed upon :

Elderly people, with their address still in the original household, are considered members of that household. This implies that they will be interviewed, together with the other members of that household (at most 4).

Elderly people, with their address in the institution, will be interviewed as if they were a single person household.

Elderly people, with their address still outside the institution, but living in an institution :

1. when a reference person : if the new address is still within the PSU, the person will be contacted at the new address; if not the household is no longer eligible
2. when not a reference person : the elderly person is no longer a member of the household.

V. Contacting the Respondents

Once the households are selected, the reference person should be contacted. To maximise collaboration and to minimise efforts a contact schedule (called the history page) has been worked out. Also explicit rules are established to replace a certain household/individuals without any effective interview. A detailed description is given in the Interviewer Guide {11310}.

All attempts to contact a household should be recorded in a structured way: the household history page. This so-called monitoring information can be used to supervise the interviewers and to evaluate the contact process. There are a minimum number of attempt of contact:

by telephone: the number of attempt of contact is 5 ;
at the door: the number of attempt of contact is 3.

In the process of contacting the households three stages can be distinguished:

- the announcement of the interview ;
- the first personal contact ;
- the personal interview (and the self-administrated questionnaire).

(1) THE ANNOUNCEMENT OF THE INTERVIEW

- To facilitate co-operation it is necessary to send a letter or leaflet in advance. Without such an introduction many people will refuse the interview because they are suspicious. This will especially be true in the larger cities. In the letter and leaflet the objectives of the interview will be explained and an indication will be given that the interviewer will contact the household within one or two weeks. Also a telephone number for further information is given. This can increase the trust in the soundness of the interview. The letter will be addressed to the reference person.
- A very important remark is that sending the letter *starts the process of making contacts*. Once the letter is sent, self-selective forces can play and threaten the representativity of the sample. From this perspective it is not an issue that a too large sample is drawn from the sampling frame. *As long as no letter is sent, the units selected are not 'active' and cannot interfere*.

(2) THE FIRST PERSONAL CONTACT: GETTING CONSENT & MAKING APPOINTMENTS

- About one week after the announcement of the interview, the interviewer will seek contact with the reference person of the household to solicit for co-operation and to make an appointment for the interview (in most cases it will not be possible to start the interview immediately). Refusal results in drop-out.
  - If the household has a telephone, an appointment is (or can be) made by telephone.
  - Alternatively, if this is not possible (because there is no telephone, or no telephone contact is established), it is necessary to make a first "at the door" contact to make an appointment. If nobody is home, it is necessary to come back at other hours. For this a schedule can be made.
- It should be controlled whether the address corresponds to the household mentioned. If not, the household moved or a wrong address was obtained. In this case the household should be recontacted by a letter at the new address if it is in the same PSU. If the new address is outside the PSU, the household is no longer eligible.
- During the first contact, also the number of household members should be asked for. If more than 4 persons are present, a random selection should be made by asking the persons their birthday (the reference person and its partner are always interviewed). Then an appointment should be made such that all persons are home. If this is not possible, more than one appointment may be necessary.

(3) THE INTERVIEW & THE SELF-ADMINISTRATED QUESTIONNAIRE

- While one person is interviewed, the others can already fill out the self-administrated questionnaire.
- It is possible that some members of the household refuse. In this case, proxy interviewing is allowed but no replacements.
- If some members are not capable to participate (e.g. in hospital or on a long journey) and they are selected, then proxies are allowed.

Appendices

A1. Selection Issues

Appendix 1.1

Selection in a Nutshell

Designing a major survey is a complex matter. This is because a lot of conflicting objectives and practical constraints are to be reconciliated. Also a lot of factors (known and unknown) play a role and interact with each other. Hence for the design a lot of decisions should be made without having full knowlegde of the consequences of the various alternatives. Similarly during the conduct not everything can be controlled and choices should be made to balance quality and feasibility.

For this purpose, a summary will be presented of all important design issues. This will serve as a 'fil rouge' for a systematic discussion of the possibilities and limits of the choices made for selection protocol of the Health Interview Survey (HIS).

The figure on page 27 outlines how decisions in the design and events during the conduct of the survey threaten the representativity, such that the sample realized is not a perfect mirror of the original population aimed at. These problems however can never be avoided totally, but being concious of them will improve the quality of the design and the conduct of the selection because clear tradeoffs can be made.

<u>The choice of the Study Population</u>

The target population (i.e. the population determined by the general objectives of the study) need not be the same as the study population (i.e. the population that can be defined accurately and reached in study), thus creating a discrepancy from the very start. This results form the fact that defining the population and delineating it unequivocally can be very difficult.

Further, budget and time constraints force to make a tradeoff between:

- the original objectives of the study; especially with respect to the population subgroups or the domains of public health for which results need to be extrapolated from the study;
- the feasibility (a function of resources, such as cost and available methodology).

As a result the target population and the study population are likely to differ. Some groups are excluded from the study and in principle it is not allowed to extrapolate the results of this study to these specific groups. For this reason it is important to document carefully which groups are eliminated and why.

The Sampling Frame

A further feature of surveys is that it is generally not possible to take *directly* a random sample from the population. A sampling frame needs to be constructed. The simplest implementation is a list which enumerates the entities of the target population as completely as possible. In our case such a list is not present or inaccessible and the population should be approached in a more sophisticated fashion. As a consequence, the sampling frame seldom coincides with the target population. **It should be documented as explicitly as possible which individuals are not covered (undercoverage) :** e.g. an estimation of the size of the homeless population in the selected PSUs**.** Also, it is possible to have overcoverage, i.e. people not belonging to the target population are selected as well. Of course, this is often a less severe issue, since one can always exclude a subset of units from the analysis. This implies that clear inclusion/exclusion rules should be defined at the design stage, and that a strategy should be worked out to detect these cases as early as possible.

The Sampling Scheme

A sampling mechanism need to be defined by which the units are selected from the sampling frame. This is necessary, but not sufficient, to guarantee that the sample selected is representative/random, and is not guided by the (implicit) preferences of the researcher. In addition, one has to ensure that a mechanism is chosen for which it has been proven (mathematically) that the sampling mechanism leads to unbiased[12] estimators with minimal variance. In principle such a mechanism involves a random step. But a random step in itself is no guarantee for the validity of the procedure. The selection probability of each unit should be known.

In addition, the sampling design needs to be reflected at analysis time. The estimators, and especially the precision estimates associated with them, should reflect the sampling mechanism. For example, a simple random sample is taken from a different universe of possible samples than a systematic sample.

---

[12] Unbiasedness implies that the sample should be as close to reality as possible. This implies that one wants to minimise both bias and variability. A very precise but slightly biased estimator may be preferred over an unbiased but very variable estimator.

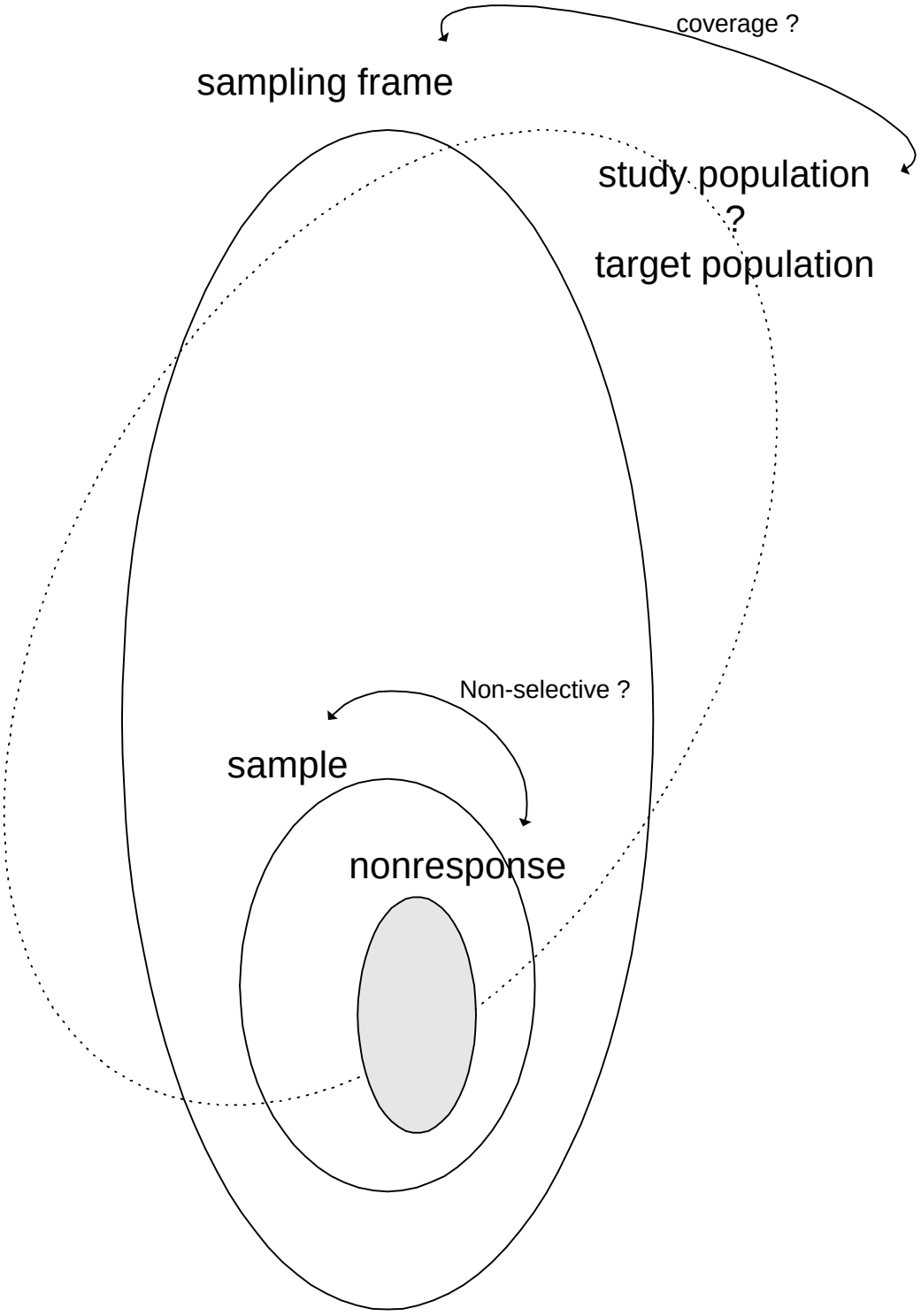
Unit Non-Response

Once the individuals are selected, they are contacted and their cooperation is solicited. A subset of them will not be reached at all, will not be able or will not want to collaborate (for various reasons), a problem known as *unit non-response*. Indeed, for some people it is difficult to locate them, the available coordinates are out of date.

As it is very unlikely that this process of *self-selection* is entirely uninformative (in the sense that several variables of interest might influence non-response), non-response should be limited at a minimum. However the best designed study, with extreme care for the conduct, will still face some of these issues. Therefore, after non-response has occured, the quality of the sample has decreased in two aspects:

- efficiency due to a decreased sample size (can be overcome by additional sampling)
- bias due to non-random non-response.

SELECTION


coverage ?
sampling frame
study population
?
target population
Non-selective ?
sample
nonresponse

Appendix 1.2

Analysis of the Requirements and Boundary Conditions

For the development of a protocol for the selection of the respondents a working group was started in collaboration with the SPH, the NIS and the LUC. The members were H. Van Oyen (SPH), J. Tafforeau (SPH), P. Quataert (SPH),L. Bellamammer (NIS), L. Lebrun (NIS), E. Schiettekatte (NIS), G. Molenberghs (LUC).

The main purpose was to make the possibilities and constraints of the tools for selection of the respondents as explicit as possible such that during the implementation of the protocol a good balance could be made between the requirements (derived from the objectives) and the (practical) boundary conditions.

In this appendix a structured compilation of the discussion and conclusion of the working group are given :

-A- TARGET POPULATION

-B- THE SAMPLING FRAME

1 Data structure

2 Timeliness

3 Differential coverage problems

-C- SAMPLING SCHEME

A. THE TARGET POPULATION

– *Inclusion criteria:*

- All people having a residence in Belgium; the country is stratified in 3 regions:

  Flemish region (3500 personal interviews)

  Walloon region (3500 personal interviews)

  Brussels region (3000 personal interviews)
- No restrictions for nationality.

  For Belgium: on average 10 % of the population consists of the immigrants, but there is a large difference between regions (Flemish region : 4.5 %, Walloon region : 11.4 %, Brussels region : 28.5 %)
- No age limits (proxies allowed)
- Elderly living in an institution should be included
- Small (religious) collective communities (maximum 8 persons) living in a house consider themselves as a household. Those people should be included.

– *Exclusion criteria:*

- People living in institution: religious communities, prison, ... with the exception of the institutions for elderly.

– *Notes:*

- *For elderly the rule is as follows:*
  * Elderly people, with their address still in the original household, are considered members of that household. This implies that they will be interviewed, together with the other members of that household (at most 4).
  * Elderly people, with their address in the institution, will be interviewed as if they were a single person household.

- For people living in Belgium not able to speak Dutch, French or German cannot be interviewed. No special questionnaires are planned for this situation. In this case this will result in a drop-out. However immigrants whose children do speak one of those languages, it should be tried to accomplish an interview anyhow if one of these children is older than 15 and is used as an interpreter.

B. THE SAMPLING FRAME

During the preparation of the Health Interview Survey it was decided to use the National Register as the sampling frame. Basically the National Register may provide a list of reference persons of the households. They are the key by which the information about a household can be accessed. Hence the sampling of the household comes down to a sampling of *the reference persons*. This aspect is crucial and should be taken into account for many decisions.

Another very important point is that the respondents are situated at two levels. Information about the household itself as well as of its members (at most 4) is of interest. The distinction between both levels of respondents has implications for the statistical analysis since the sampling probabilities will be different.

For the information about the households one should know how many persons declared themselves as the reference person. E.g. a household with two reference persons has a 2-times increased probability of being selected. However this information is very hard to get (if not impossible) both from the National Register or by questionning the household members. It is proposed to ignore this issue.

For the characteristics on an individual level, an additional correction is necessary when there are more than 4 members in a household (say k). Then the individual selection probability for the 2 or 3 members to be selected by the birthday rule is 2/(k - 2) or 3/(k - 1) and for the analysis the inverse of this number should be used as a weight.

1. Data structure

– *The contents of the National Register is ruled by the law of 8/8/1983 (Art. 3).*

Definition of the Reference Person:

When no special declaration is made, it is the oldest person or the husband. However it's depends upon the use and the appreciation of the municipalities.

If the reference person dies, the partner automatically becomes the reference person for a household of 2 persons and this situation is adapted in the Register. For larger household the municipality may contact the household and ask the question to the household.

– *Information about the reference person of the household*

- National number
- Last name, first name
- Place of birth and birth date (hence age is known)
- Sex
- Nationality
- Main residence (the address: municipality, street, number)
- Profession : a 5-class coding
- Marital status
- Composition of the household
    * size of the HH
    * same information as reference person
- Place and date of death (in this case the copy of the National Register obtained does not contain this household anymore)

– *The information attached can be used to*

- locates the households
- guides the sampling process (e.g. a systematic sample could be taken of a list ordered by address, or by size of the household and age of the reference person)

– *The information used :*

- National number
- Main residence : address

<u>For stratification</u>

- Main residence : statistical sector of each address
- Size of the household (1, 2,3,4, 4+)
- The age of the reference person (in years)

<u>Other information used for statistical purposes</u>

- Sex
- Nationality

– *The precision of the information:*

  - No precise information is available about the quality of the National Register. When change occurs with the household or within the household it may takes up to ± 1 months[13] before the information in the National Register is updated. In reality this delay may be greater and depends also on the level of informatisation of civil register witin the municipality. Hence is preferable that the interviewers themselves find out the composition of the household when visiting.
  - It might be interesting to compare to what extend the information known in advance corresponds to the real situation. However self-reporting itself might lead to errors. Of course, variables such as sex and main residence will not be prone to high error rates. More fundamentally, one will be able to check the situation only for those who are attained. In other words, the individuals with wrong information, which lead to the fact that they cannot be contacted any more, provide no means to check the information. If not applied with extreme caution, this procedure will underestimate the problem of incorrect information.

2. Timeliness

– Households are born and die. In principle (this should be distinguished from new(born) members in a household or members that left the household recently. As there may be some delay, the sampling frame does not cover those households. In the case of a household which died, a control process with the latest available National Register (one month before each quarter), or the interviewer will identify those households. This only affects the efficiency of the sampling process as a household is selected which is not eligigble. In case of a new household; they are not identifiable and they are not included in the sampling frame. In order to have a maximal coverage by sampling frame and not missing new households, the date of the creation of the sampling frame should be close as possible with the date of the sampling. As the price of each new copy is high (approximately 400.000 Bfr.) it was decided to work with only two copies that the NIS already obtains for its own purposes and not to buy a copy(ies) specifically for the Health Interview Survey.

[13] Here is a legal time to declare to the municipality is a newborn and death and a change in address. Municipalities which are not online connected to the National Register transfer the information on a monthly basis.

– *One may not underestimate the turn over:*

in Brussels: about 20 % migration in one year
in general: about 7 %

– *Time delay*: as the first copy of the National Register is from 1/1/96 for the sample of the first quarter of 1997 and from 1/1/97 for the sample for the 2nd to 4th quarter a verification of the sample is necessary to identify households which are no longer eligible or households which needs an update of the reference person or the address. This is done as late as possible prior to first contact of the households each quarter, approximately one month in advance. Two criteria are checked for: the vital status of the reference person and the current address of the main residence of the reference person. The results of this control for the sample of the first quarter of 1997 is given in table 4.

– *The cost of the verification is low* (2.5 Bfr. by reference person number or 21.850 Bfr for the 8740 reference persons selected for the first quarter). The control occurs automated. Only in the case the reference persons of a two or more persons household is dead (less than 1 percent of the selected households), a manual on-line search is necessary. The time needed for the control process is approximately 1 week.

– *Seen from the perspective of the sampling frame two situations are possible:*

1. The reference person died. If he was living alone then the household does not exist anymore, if not, then, in principle, the household will be contacted if its members still live within the same PSU.
2. The reference person moved. If the new address is located in the same PSU then the household is contacted at the new address. In the other situation there is a drop-out. It should be clear however that moving of the reference person can cover very different situations. It can be a result of a move of the household with all its members. It can be a disintegration of the household (e.g. a divorce). In such case the reference person and the remaining members of the household constitutes the eligible household.

3. Differential coverage problems

- If the household not covered by the sampling frame were a random sample from the households having a residence in Belgium and included in the sampling frame, there would be no problem. However this non-coverage is probably linked with variables under study.
- E.g. it is clear that the lower social classes will be less covered. The lowest ones are known to be very hard to reach because they might not have no residence (homeless), or they might comply less with the administrative procedures.

Table 4 Status of the household after the control with the national register on the 31th of November 1997.
*8740 households: sample 1ste quarter 1997 + 12 household non-resident**

Code	Reference person	Action	N (%)
1	alive and at same address		7573 (86.65)
2	died, 1 person household : household died	Household drops out	64 (0.73)
3	died, 2 or more person household household still at same address	Change name of reference person	53 (0.61)
4.	died, 2 or more person household household moved within municipality	Change name of reference person and address	2 (0.02)
5	died, 2 or more person household household moved out municipality	Change name of reference person and address / household drops out	3 (0.03)
6	died, 2 or more person household household can not been traced	Household drops out	1 (0.01)
7	alive and moved within municipality	Change address	478 (5.47)
8	alive and moved out municipality	Change address / household drops out	555 (6.35)
9	alive and moved but can not been traced	Household drops out	11 (0.13)
	reference person : non resident*	not eligible	12

* There were 12 households excluded initially because they are registered as non-resident. To prevent in the future that those type of households are selected, all households with a statistical sector equals to Axxx (where x equals a blanc) or with the code of the street ending by 9999 will be excluded before the sampling procedure starts.

C. SAMPLING SCHEME:

It was decided to devise the country into strata with a fixed number of interviews to be realised and to sample within each stratum with a three-stage sampling scheme:

1. First municipalities are selected within each stratum. These municipalities can be considered as clusters of households and will be denoted as the primary sampling units (PSU).
2. Next within each municipality the households are selected. For the information linked to the household itself this is the last step.
3. The household itself can be considered as a cluster for the individual respondents. If the size of the family is smaller or equal than 4, the total cluster is taken, if not, subsampling is necessary.

The sampling scheme should take into account the following constraints:

1. Sample Size by Region

   Total sample size: 10 000; for each region the number of interviews is fixed:

   - Flemish region: 3500
   - Walloon region: 3500
   - Brussels region: 3000

   This is to ensure that for each region inference is possible with about the same precision. Roughly speaking the confidence intervals for Brussels will be 1.1 as large as for the other two regions.

2. Sample Size by Province

   Although in principle a weighted random sample within each region would result in a representative sample on the regional level it is possible that just by chance the smaller provinces are not represented well. Therefore within each region the number of interviews will be fixed by province in proportion to the number of inhabitants. An additional advantage is that this option allows the provinces to pay in a transparent way for extra interviews.

3. The selection of the municipalities

   The selection chance of each municipality should be proportional to its size. This warrants the selection of important municipalities (metropoles and larger cities) but also of some smaller municipalities as the sampling scheme is a systematic procedure after classification of the municipalities by size.

4. The German community in Belgium

   At least one German municipality should be present in the sample. Further a political decision was taken to do an over-sampling and to fix the number of interviews to 300, even though proportional to the number of inhabitants (66 527) only about 70 interviews can be expected.

5. Representativity over time

   To have representativity over time, every quarter 2 500 individuals should be interviewed. Within each quarter, the number of interviews realised should be as equally spread as possible. This should be achieved within each municipality selected.

6. The number of interviews within a municipality

   To keep fieldwork feasible the number of interviews by municipality should be at least 50. Then every quarter on average 12.5 interviews should be realised.

7. The respondents

   The respondents are situated at two levels. The members of the households and the household itself for questions relating to the household as a whole.

8. The number of households within a community

   The sampling frame used from the National Register is organised around the reference persons. Hence households are selected and not individuals.

9. The respondents within a household

   Within a household at most 4 members may be interviewed as interviewing more persons is not expected to offer much more information (because of the familial correlation) and would put a too heavy burden on the household. In any case the reference person and, if applicable, its partner should be interviewed.

A2. Sampling Schemes

Appendix 2.1

Sampling Schemes

-A- Simple Random Sampling without replacement
-B- Systematic sampling
-C- Clustering and Multi-stage Samplin
-D- Stratification

A. Simple Random Sampling without replacement

In a simple random sample without replacement each unit is selected with equal probability. Without replacement means that each unit is selected at most once.

Although very simple and straightforward in theory (and for the statistical analysis) this way of sampling is seldom used in practice. Some of the most important reasons are (along with some solutions):

- Very often the sampling units are unknown in advance and to enumerate them totally would be prohibitively expensive. A possible solution is first to sample from a population of larger entities and then to select within each cluster some target units. The gain is that in this way only the individuals within each selected cluster should be enumerated.
- Another problem with simple random sampling is that this sampling scheme can be statistically very inefficient. For instance (a lot of) precision can be gained by subdividing a heterogeneous population into more homogeneous blocks and by taking a separate sample from each of them (stratification).
- Also from a practical point of view simple random sampling can be inefficient. If the location of the selected individuals is too scattered then the fieldwork costs can be high. Here clustering can offer a solution (the price to pay is that more interviews needs to be made to have the same precision).
- Another reason to abandon simple random sampling is to guarantee some subgroups are present in a sufficient large number. For instance to have about equal precision in the three regions of Belgium, the sample size in each region is of the same order of magnitude although this is not in proportion to the number of inhabitants.

A common feature for all these modifications is that the estimation techniques become less standard. This holds for the point estimates as well as for the interval estimates. For the point estimates often a reweighting is necessary to take into account the different selection probabilities. For the standard errors often no closed mathematical expression exists and approximations are necessary.

Below a summary will be given of the sampling schemes used for the HIS (Health Interview Survey). Each time a short definition will be followed by some advantages and disadvantages of the techniques and the implications for the statistical analysis are discussed shortly.

B. Systematic sampling

A procedure to generate a sample of a fixed size n out of a population of size N is to draw every N/nth unit of a list, where the starting point is choosen at random between 1 and N/n. This procedure is valid if the list shows no unwanted systematic trend or periodicity.

A special application is to order the units as a function of some combination of characteristics. This results in an implicit stratification. Such a design are a guarantee that the characteristics of the sample will be close to the characteristics choosen to order the list. E.g. by ordering the municipalities by their size, one can assure that the larger towns are certainly selected and that the division over smaller and larger towns (and hence the contrast urban-rural) is approximatively equal to the real distribution.

For the statistical analysis one can use the same formulas as for the simple random sample, provided no periodicity (increasing variability) nor systematic trend (decreasing variability) is present. For most variable of the HIS questionnaire this will be the case.

C. Clustering and Multi-stage Sampling

For cluster sampling the population is grouped into subpopulations or clusters. A sample of clusters is selected and every element in the cluster is selected. The sampling becomes two-stage if within each cluster a second sampling is done and only a few individuals are chosen[14]. Then the groups selected first are called primary sampling units (PSU) and the units selected in the next step are called the secondary sampling units (SSU). This scheme can be extended to three or even more stages (multi-stage sampling).

A drawback of a clustered sampling design is that in most cases it decreases the precision, because the individuals in the cluster tend to be more alike to each other. However this procedure is highly recommended when:

- not all the units of the population are given in advance and/or enumerating them would be prohibitively expensive (or in short no sampling frame exists). As a solution clusters are sampled first because often they are better known and/or easier to enumerate (at least less units should be enumerated). Because it is very often not very informative or possible to take all elements of the cluster a subsample is taken (multistage design).
- when the population elements are scattered over a large region. Then the fieldwork is facilitated because the length of trips to be made by the interviewers is reduced. Also supervision can be more efficient.

The main issue is that inferential techniques become less standard. An approximate procedure is to multiply the precision estimates by the so-called design factor, i.e. the factor to correct the standard errors calculated under the hypothesis of simple random sampling without replacement. In this respect simple random sampling still remains - although not often used - the reference.

[14] If in the limit only one individual is choosen, the clustering disappears (although to realize the multistage sampling clusters should be defined).

D. Stratification

The population (or the sampling frame[15]) is partioned and a sample is drawn from each of the partition constituents.

There are two main reasons of stratification:

- to increase the precision. This will be the case when the parts are more homogeneous than the total. Techniques exist to find out the optimal allocation of sample size over the different parts.
- to ensure that the precision of the estimates in each stratum is sufficient to draw inferences. This can be at the expense of precision.
- to respect important administrative or political boundaries.

If the stratification is not proportional to the size of the various strata, then one has to reweight the sample from each stratum in order to produce optimal (for estimators) and valid (for precision) inferences for the global population.

[15] It is possible to sample each part by a different sampling frame.

Appendix 2.2

The Selection of the Municipalities

-A- Introduction

-B- The distribution over the regions

-C- The distribution over the provincies

-D- The systematic sampling of the municipalities

-E- The selection

A. Introduction

As discussed before, the selection of the respondents is a three step process:

1. First the municipalities are selected (PSU)
2. Then 50 households are selected in each municipality chosen (SSU)
3. Finally within a selected household at most 4 persons are selected for the interview (TSU)

In this chapter the selection of the municipalities is described.

An important decision in this respect was to take the municipalities (589 in total, with an average of 17 221 inhabitants) and not the old municipalities which are more homegeneous (about 2500 in total with on average 4000 inhabitants). Although this last option seems attractive, it was rejected, as this information is not (directly) available in the sampling frame.

The basic idea was to subdivide Belgium into strata and to take a systematic sample of the municipalities within each stratum. For the three different regions the number of interviews was fixed: 3500 for the Flemish region, 3500 for the Walloon region and 3000 for the Brussels region.

However there are other additional requirements:

- The German community should be oversampled (300 interviews). To guarantee this a fourth stratum can be introduced by splitting the German Community off.
- For the provinces no clear quota were negociated and hence it is possible to draw a random sample proportional to the size.

These two points lead to make a subdivision of 12 strata: the 10 provinces, Brussels and the German Community. Within each region the number of interviews is distributed proportionnaly to their size.

B. The distribution over the regions

To make fieldwork feasable it is chosen to have in each primary sampling unit at least 50 interviews (a group). Hence in total 200 groups should be selected in the first step. As some municipalities might be selected more than once the number of groups may be greater than the number of PSUs (municipalities) selected. However the sampling probabilities are not equal. In the first place the number of interviews in each regio is fixed: 3500 for the Flemish region, 3500 for the Walloon region and 3000 for Brussels region. Also it is specified that an oversampling should be taken of 300 in the german part of Belgium (the German Community : the arrondissement Eupen-Malmedy, denoted by 'Eupen').

To realize these boundary conditions Belgium is subdivided into 4 parts (within brackets the number of groups to be selected is given:

- Flemish region: 3500 (70)
- Walloon region : 3500 (70 = 64 + 6 (Eupen))
- Brussels region : 3000 (60)

If we relate these figures to the population size of each region (see table below, figures of INS, january 1st 1996) it is seen that the probability of being selected differs considerably from region to region. Currently the Belgian population size is about 10 million people. Hence on average the chance of being selected is 1/1000. However in the Flemish region this is only about 0.6/1000, for the Walloon provinces (minus Eupen), it is close to the global average, for Brussels this is more than 3 times as large and for the German Community it is 4 times as large.

These calculations illustrate clearly that when calculating global estimates for the country appropriate weights should be taken into account.

On the other hand for the precision this disproportionate sampling makes (nearly) no difference (except for a correction factor, which is very small here) as it is not the sampling proportion which is relevant, but the absolute size of the sample. In fact by making the sampling sizes in each region close to each other, it is guaranteed that in each of the three main regions the same inferences and conclusions are possible. If Flemish region is taken as the reference, then the precision in the Walloon part is a factor 0.98 and for Brussels it is 0.94.

2	RegC	RegN	RegL	Pop.Size	Intv	# Groups of 50	P(S)	Gap	Prec
1	VLA	Flanders	1	5880357	3500	70	0.595	84005	1.00
2	WAL	Wallony	2	3245130	3200	64	0.986	50705	0.98
4	GER	German	4	69438	300	6	4.320	11573	0.29
3	BRU	Brussels	3	948122	3000	60	3.164	15802	0.94
				10143047					

* Wal region : denotes the Walloon region without German Community
P(S) : sampling probability per 1000 ((Intv/Pop.size) *1000)
Gap : stepsize in systematic sampling of municipalities
(Pop.size/# Groups of 50)

C. The distribution over the provincies

In principle the subdivision proposed above is sufficient to select the municipalities within each region. However

- to guarantee that the number of interviews realized in each province is (nearly) proportional to the population and
- to allow that in some provinces an oversampling is possible,

it was decided also to fix also in advance the number of PSU to be selected in each province.

This comes down to make 12 strata: the 10 provinces, Brussels and Eupen (the German Community of Belgium). The table below summarizes the main results.

It is impossible to assign the number of PSU such that the probability of selection is totally equal. In Flemish region the probability varies between 0.580 /1000 (Limburg) and 0.623/1000 (West-Vlaanderen). This was the best that could achieved and in fact all are close to the overall average (0.595). The implication is that a (small) reweighting will be necessary to reconstruct the total for that region. A similar remark holds for the Walloon provinces (without Eupen). Here the figures vary between 0.953/1000 (Liege) and 1.036/1000 (Luxembourg).

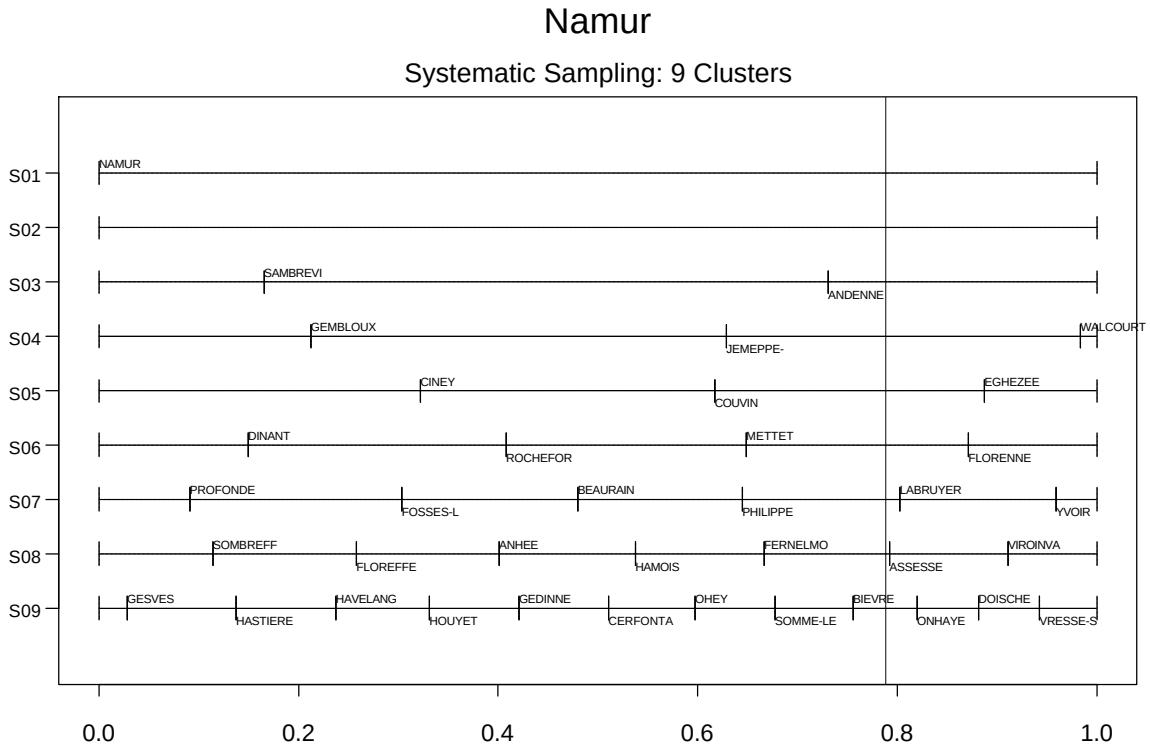
ProvN	Abb	ProvL	Reg	Pop	Frac	#of groups	Gap	P(S)
Antwerpen	ANT	1	1	1631243	27.74	19	85855	0.582
Vlaams Brabant	BrV	2	1	999186	16.99	12	83266	0.600
Limburg	LIM	3	1	775302	13.18	9	86145	0.580
Oost-Vlaanderen	OVL	4	1	1351777	22.99	16	84486	0.592
West-Vlaanderen	WVL	5	1	1122849	19.09	14	80204	0.623
TOTAL			1	5880357	100.00	70	84005	0.595
Brabant Wallon	BrW	6	2	339062	10.45	7	48437	1.032
Hainaut	HAI	7	2	1284761	39.59	25	51390	0.973
Liege	LIE	8	2	944291	29.10	18	52461	0.953
Luxembourg	LUX	9	2	241339	7.44	5	48268	1.036
Namur	NAM	10	2	435677	13.43	9	48409	1.033
TOTAL			2	3245130	100.00	64	50705	0.986*
Brussel	BRU	11	3	948122	100.00	60	15802	3.164
Eupen	EUP	12	4	69438	100.00	6	11573	4.320

* : Walloon region minus the German Community

D. The systematic sampling of the municipalities

Once the quota by stratum (by province) are defined, the sampling is straightforward. Within each province (region) the cities are ordered by size from large to small. To give the cities a weight proportional to its size, they are repeated in the list as many times as their number of inhabitants. Hence the probability of selection is proportional to the size.

To make this procedure more transparant a graphical representation was developed. Each city is represented by a segment proportional to the size. The sum of the segments represent the total population of the stratum. Then this line is cut into a number of pieces equal to the number of groups of 50 individuals one wants to draw from the stratum and these segments are put one below the other. Then taking a systematic sample comes down to drawing an arbitrary vertical line. The process is illustrated for the province of Namur.



[1] In the province of Namur nine groups of 50 inidividuals should be drawn, which is translated into 9 horizontal lines. On this horizontal lines the municipalities are represented by segments which can extend over many horizontal lines; e.g. Namur is so large in comparison with the other cities that it is distributed over more than two horizontal lines[16]. The implication is that Namur as a PSU will be selected at least twice and there is also a chance it is selected 3 times. The same happens in other provinces and below a summary is given of the cities which are selected at least once (see pictures in next section).

[16] The proportion of inhabitants living in Namur is 105059/435677 = 24.1 % which is larger than 2/9 = 22.2 % indeed.

	NIS	Name	Region	Prov	# of time to be selected with certainty
1	11002	ANTWERPEN	V	ANT	5
2	24062	LEUVEN	V	BrV	1
3	44021	GENT	V	OVL	2
4	31005	BRUGGE	V	WVL	1
5	53053	MONS	W	HAI	1
6	54007	MOUSCRON	W	HAI	1
7	55022	LA LOUVIERE	W	HAI	1
8	57081	TOURNAI	W	HAI	1
9	52011	CHARLEROI	W	HAI	3
10	62096	SERAING	W	LIE	1
11	63079	VERVIERS	W	LIE	1
12	62063	LIEGE	W	LIE	3
33	63023	EUPEN	W	LIE	1
13	92094	NAMUR	W	NAM	2
14	21002	AUDERGHEM OUDERGEM	B	BRU	1
15	21003	BERCHEM-SAI SINT-AGATHA	B	BRU	1
16	21006	EVERE	B	BRU	1
17	21008	GANSHOREN	B	BRU	1
18	21011	KOEKELBERG	B	BRU	1
19	21014	SAINT-JOSSE SINT-JOOST-	B	BRU	1
20	21017	WATERMAEL-B	B	BRU	1
21	21005	ETTERBEEK	B	BRU	2
22	21007	FOREST VORST	B	BRU	2
23	21010	JETTE	B	BRU	2
24	21013	SAINT-GILLE SINT-GILLIS	B	BRU	2
25	21018	WOLUWE-SAIN SINT-LAMBRE	B	BRU	2
26	21019	WOLUWE-SAIN SINT-PIETER	B	BRU	2
27	21009	IXELLES ELSENE	B	BRU	4
28	21012	MOLENBEEK-S SINT-JANS-M	B	BRU	4
29	21016	UCCLE UKKEL	B	BRU	4
30	21001	ANDERLECHT	B	BRU	5
31	21015	SCHAERBEEK SCHAARBEEK	B	BRU	6
32	21004	BRUXELLES BRUSSEL	B	BRU	8

[2] On the other hand one notes that all small villages of the province of Namur are located on the bottom line. Hence the procedure assures that one of these villages will be selected as they form one block or ‘stratum’. A possible issue with the smaller villages however can be that there are not enough people. However for all of them the number of inhabitants is larger than 50. The table below gives the five smallest villages in the Flemish region and the Walloon region, the only village with could give problems is Herstappe in the province of Limburg. However the selection probability is very small.

NIS code	Name	Pop. size	District	Prov	Reg
73028	HERSTAPPE	90	73	LIM	V
33016	MESEN	992	33	WVL	V
34043	SPIERE-HELKIJN	1887	34	WVL	V
23009	BEVER	1893	23	BrV	V
45062	HOREBEKE	1924	45	OVL	V
84029	HERBEUMONT	1427	84	LUX	W
84016	DAVERDISSE	1431	84	LUX	W
81013	MARTELANGE	1513	81	LUX	W
82009	FAUVILLERS	1759	82	LUX	W
85047	ROUVROY	1875	85	LUX	W

[3] Another point which is illustrated is that the systematic sampling cannot result in any combination of the municipalities. In fact by fixing the order and working in a systematic way restricts the number of possibilities and the cities are grouped in ‘strata’ from which just one will be selected. E.g. either Namur is selected a third time or Sambreville is selected. And if Namur is selected a third time also Andenne is selected ... . Using a systematic sampling procedure the univers of possible samples is reduces and the the more extreme samples are ruled out.

A full list of all these ‘strata’ is given below.

```
Namur: 9
********


Always Selected: 2
      Mun Min
574 NAMUR   2

(Remaining) Strata: 7
S01: NAMUR SAMBREVILLE ANDENNE
S02: ANDENNE GEMBLOUX-SUR-ORNEAU JEMEPPE-SUR-SAMBRE WALCOURT
S03: WALCOURT CINEY COUVIN EGHEZEE
S04: EGHEZEE DINANT ROCHEFORT METTET FLORENNES
S05: FLORENNES PROFONDEVILLE FOSSES-LA-VILLE BEAURAING
     PHILIPPEVILLE LA BRUYERE YVOIR
S06: YVOIR FLOREFFE SOMBREFFE ANHEE HAMOIS FERNELMONT ASSESSE
     VIROINVAL
S07: VIROINVAL GESVES HASTIERE HAVELANGE GEDINNE HOUYET
     CERFONTAINE OHEY SOMME-LEUZE BIEVRE ONHAYE DOISCHE VRESSE-SUR-SEMOIS
```

[4] Finally the vertical line on the plot is drawn at random by generating a random variate from a uniform distribution. These municipalities whose segment is crossed are selected. For the current line, the selected municipalities are (in the list above these municipalities are underlined):

```
NAMUR 2 ANDENNE JEMEPPE-SUR-SAMBRE COUVIN METTET PHILIPPEVILLE FERNELMONT
BIEVRE
```

Above procedure can be repeated for each province. The corresponding lists can be found at the end of the appendix.

E. The selection

By the sampling scheme described above a random sample was selected. The result is given below. Within brackets the total number of groups for each stratum is given. If a municipality (PSU) is selected more than once it is underlined and the number of groups to be selected in that town is specified. After this list the corresponding pictures are included.

Flemish region [70 groups = 3500 individuals]

Antwerpen [19]: ANTWERPEN 6 MECHELEN HEIST-OP-DEN-BERG SCHOTEN MORTSEL
EDEGEM ZOERSEL BALEN KALMTHOUT DUFFEL RUMST RAVELS GROBBENDONK DESSEL

Vlaams Brabant [12]: LEUVEN DILBEEK GRIMBERGEN SINT-PIETERS-LEEUW
ZAVENTEM SCHERPENHEUVEL-ZICHEM ST-GENESIUS-RODE TERNAT HAACHT
BOORTMEERBEEK BEGIJNENDIJK BOUTERSEM

Limburg [9]: GENK BERINGEN TONGEREN HOUTHALEN-HELCHTEREN MAASEIK
TESSENDERLO HAMONT-ACHEL BORGLOON NIEUWERKERKEN

Oost-Vlaanderen [16]: GENT 3 AALST SINT-NIKLAAS DENDERMONDE NINOVE
OUDENAARDE ZOTTEGEM MALDEGEM EEKLO HAALTERT STEKENE ASSENEDE NEVELE
WAARSCHOOT

West-Vlaanderen [14]: BRUGGE KORTRIJK OOSTENDE ROESELARE KNOKKE-HEIST
WEVELGEM ZWEVEGEM KOKSIJDE BLANKENBERGE ICHTEGEM DEERLIJK OUDENBURG
GISTEL ALVERINGEM

Walloon region [70 groups = 3500 individuals]

Brabant Wallon [7]: WAVRE OTTIGNIES-LLN NIVELLES GENAPPE REBECQ
LA HULPE INCOURT

Hainaut [25]: CHARLEROI 4 MONS 2 LA LOUVIERE TOURNAI MOUSCRON CHATELET
BINCHE ATH FLEURUS MANAGE BOUSSU MORLANWELZ FONTAINE-L'EVEQUE LESSINES
BELOEIL FARCIENNES FRASNES-LEZ-ANVAING ERQUELINNES ANTOING BEAUMONT
MERBES-LE-CHATEAU

Liège minus German Community [18]: LIEGE 4 SERAING VERVIERS ANS FLEMALLE
GRACE-HOLLOGNE HUY FLERON WAREMME SPRIMONT SPA AWANS WAIMES REMICOURT
STOUMONT
Liège (german speaking part) [6]: EUPEN KELMIS RAEREN ST-VITH BULLINGEN
LONTZEN

Luxembourg [5]: AUBANGE LIBRAMONT-CHEVIGNY BOUILLON VAUX-SUR-SURE
DAVERDISSE

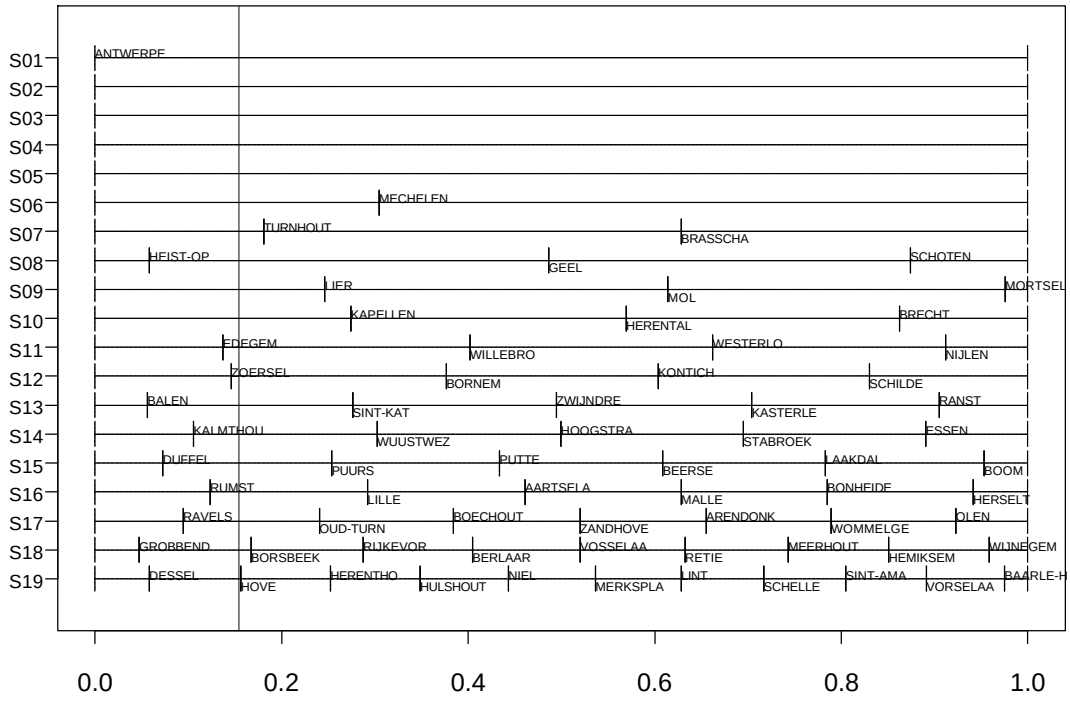
Namur [9]: NAMUR 2 ANDENNE JEMEPPE-SUR-SAMBRE COUVIN METTET
PHILIPPEVILLE FERNELMONT BIEVRE

Brussel/Bruxelles [60 groups = 3000 individuals]

BRUSSEL 9 SCHAARBEEK 6 ANDERLECHT 6 UKKEL 5 ELSENE 4
ST-JANS-MOLENBEEK 5 ST-LAMBRECHTS-WOLUWE 3 VORST 2 ST-GILLIS 3 JETTE 3
ETTERBEEK 2 ST-PIETERS-WOLUWE 2 EVERE 2 OUDERGEM 2 WATERMAAL-BOSVOORDE 2
ST-JOOST-TEN-NODE GANSHOREN ST-AGATHA-BERCHEM KOEKELBERG

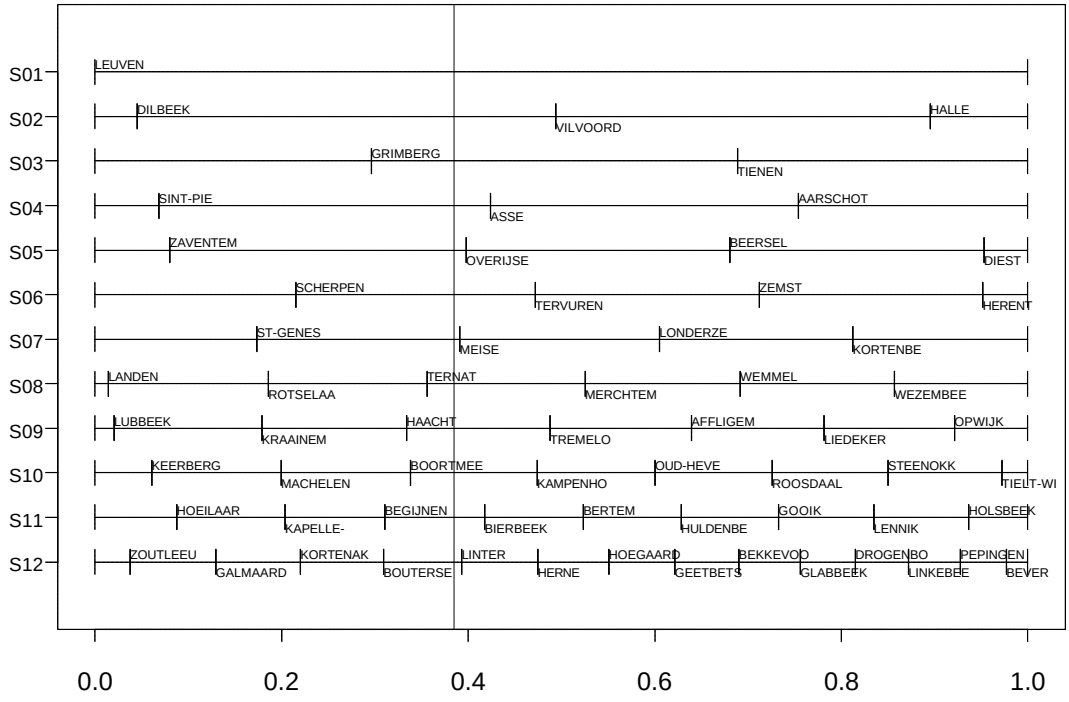
Antwerpen

Systematic Sampling: 19 Clusters



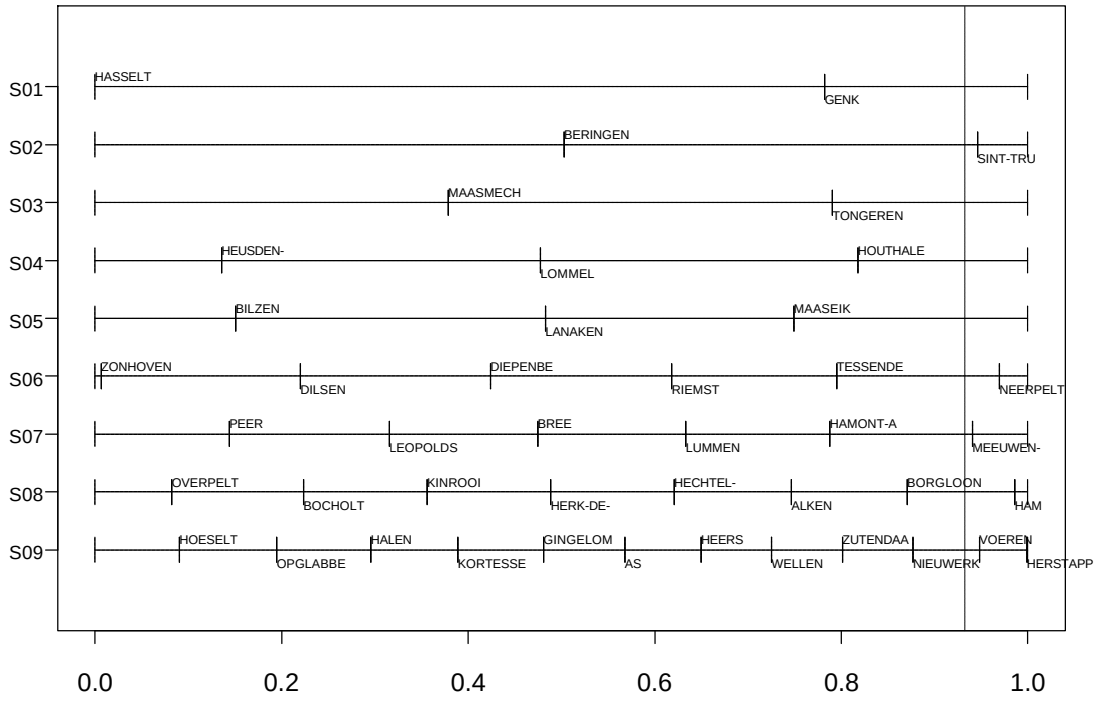
Vlaams Brabant

Systematic Sampling: 12 Clusters



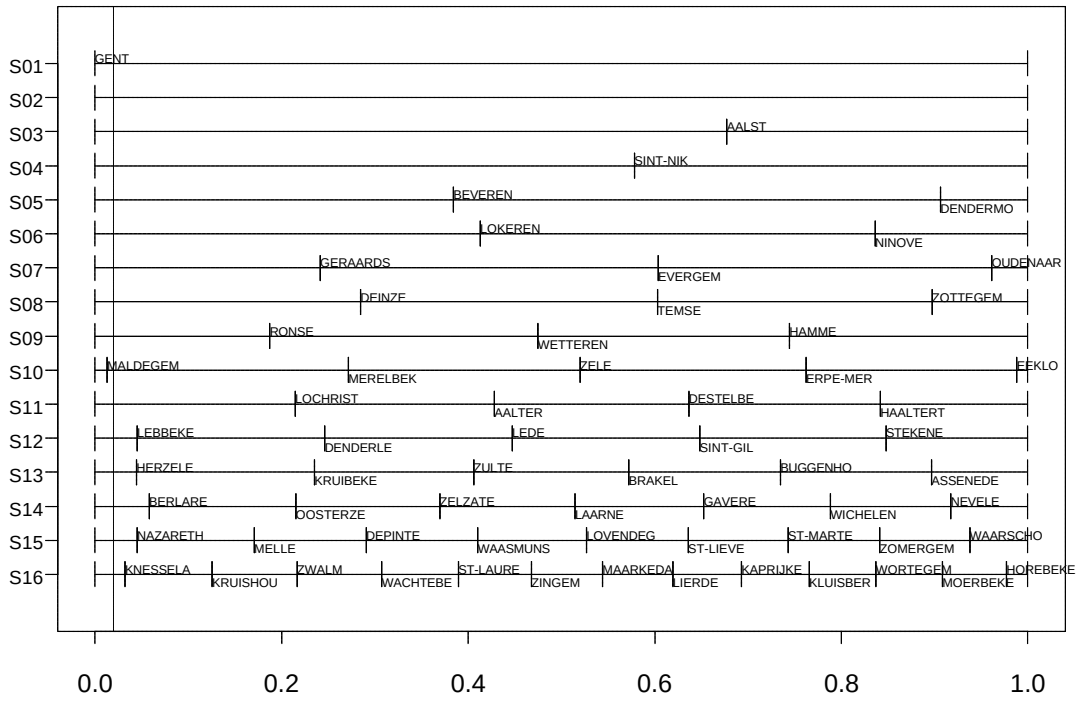
Limburg

Systematic Sampling: 9 Clusters



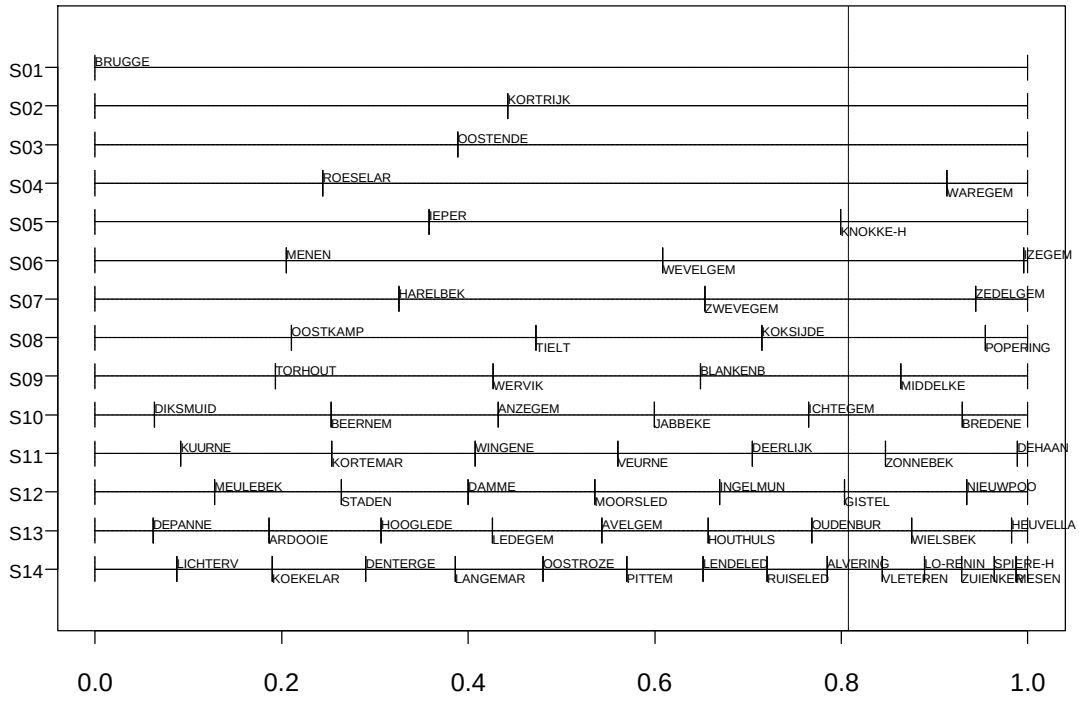
Oost-Vlaanderen

Systematic Sampling: 16 Clusters



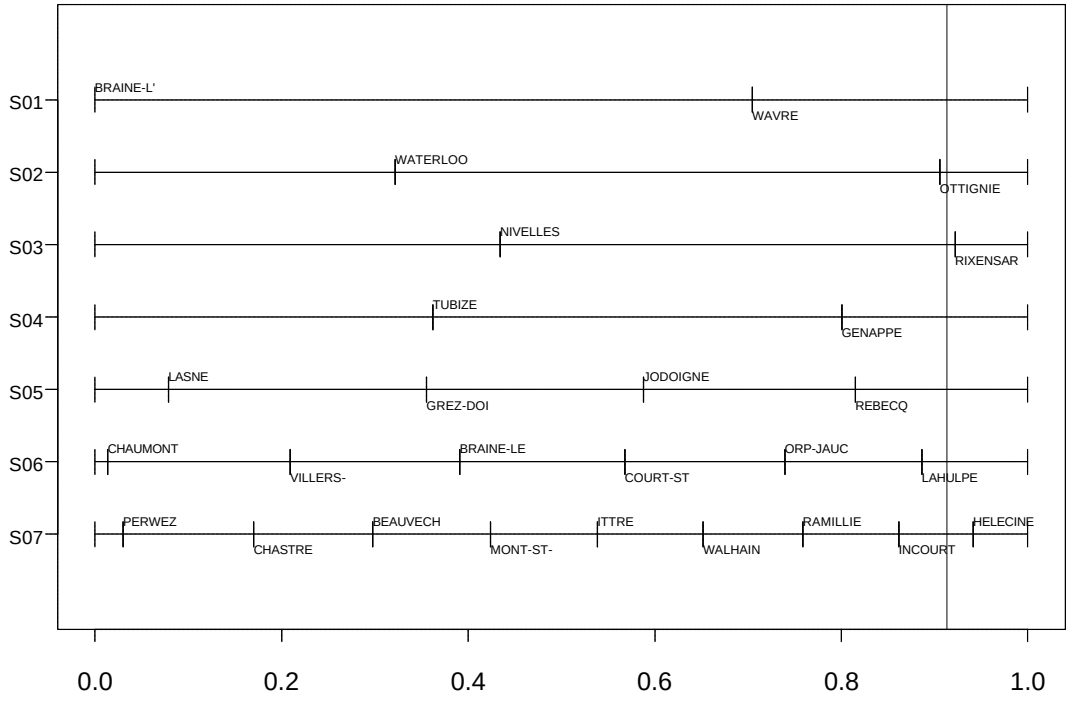
West-Vlaanderen

Systematic Sampling: 14 Clusters



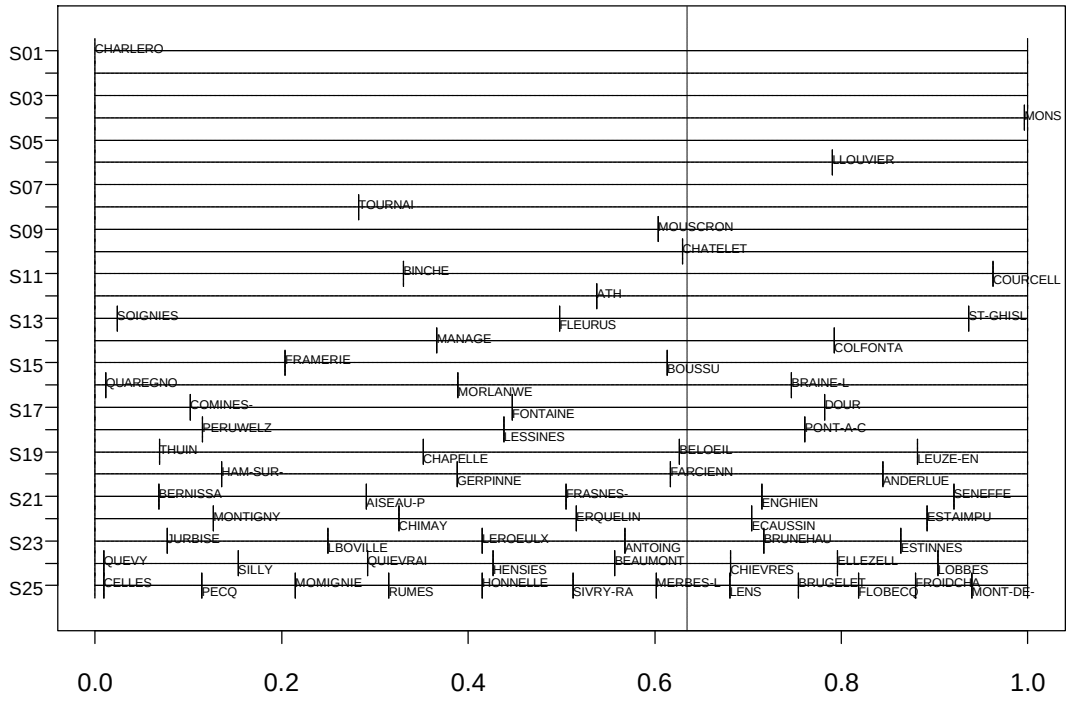
Brabant Wallon

Systematic Sampling: 7 Clusters



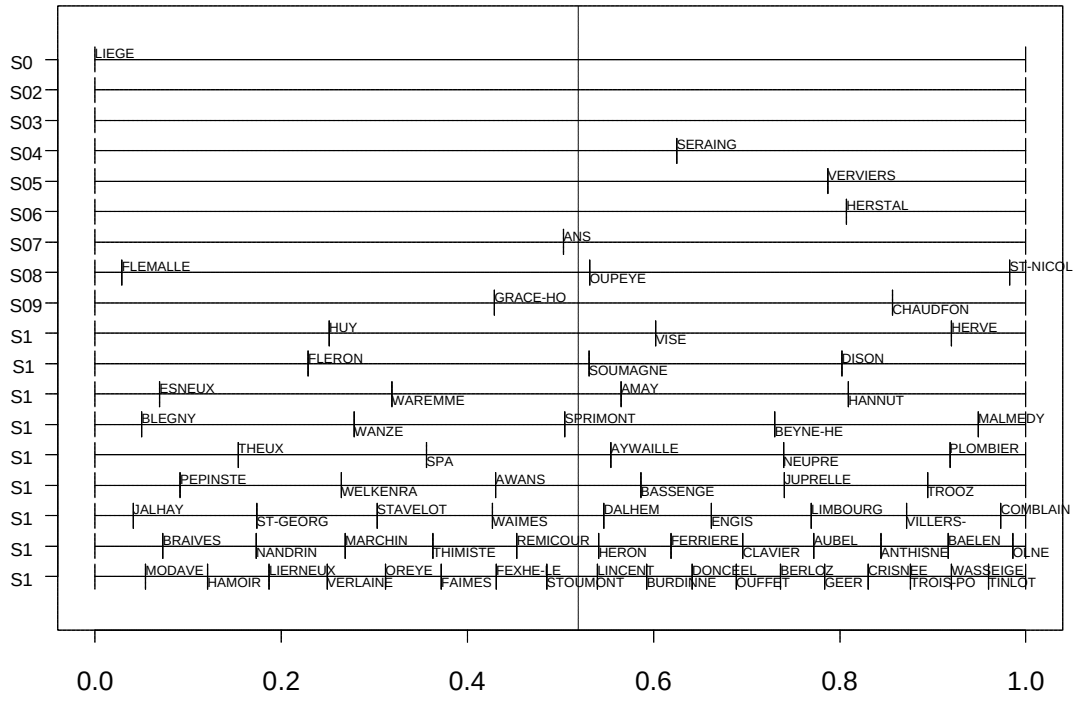
Hainaut

Systematic Sampling: 25 Clusters



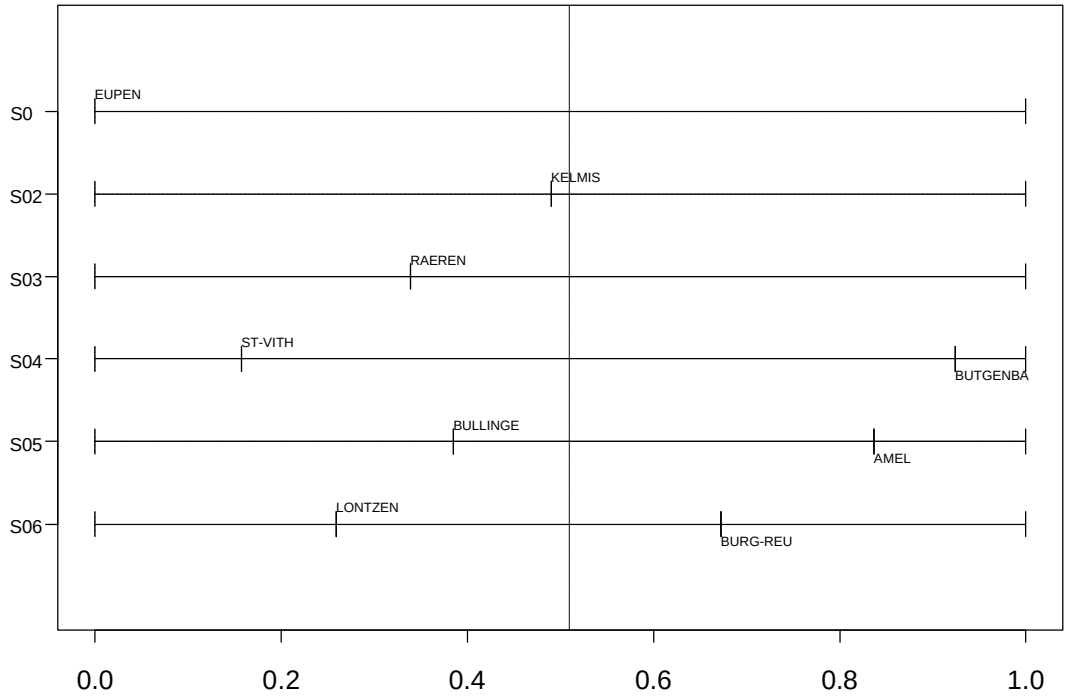
Liège

Systematic Sampling: 18 Clusters



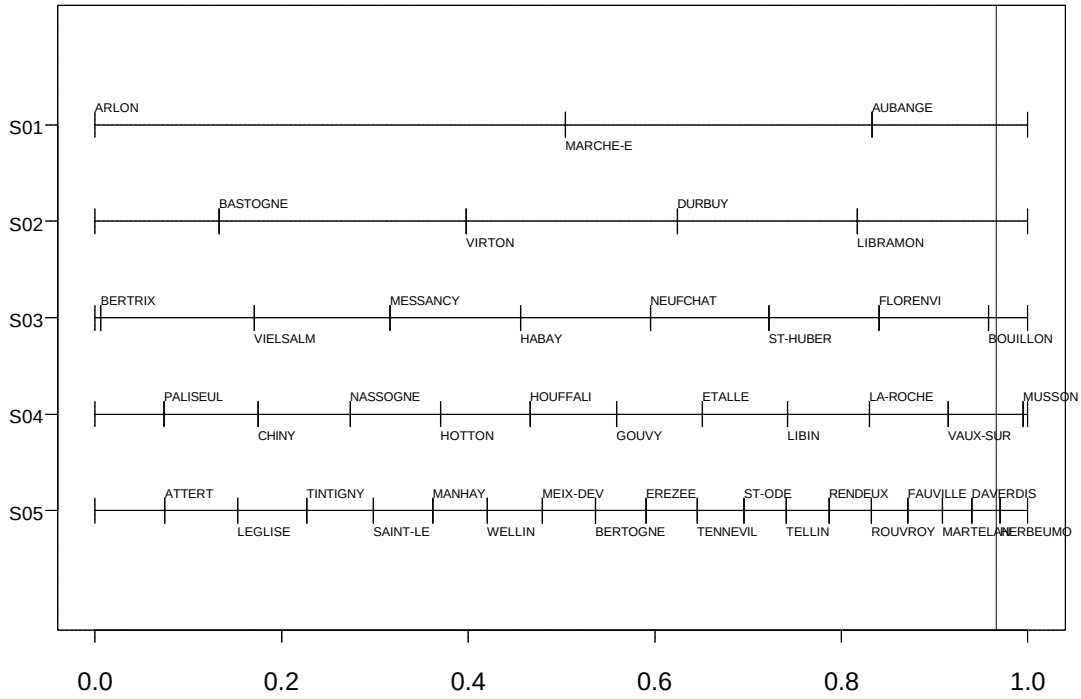
German Community

Systematic Sampling: 6 Clusters



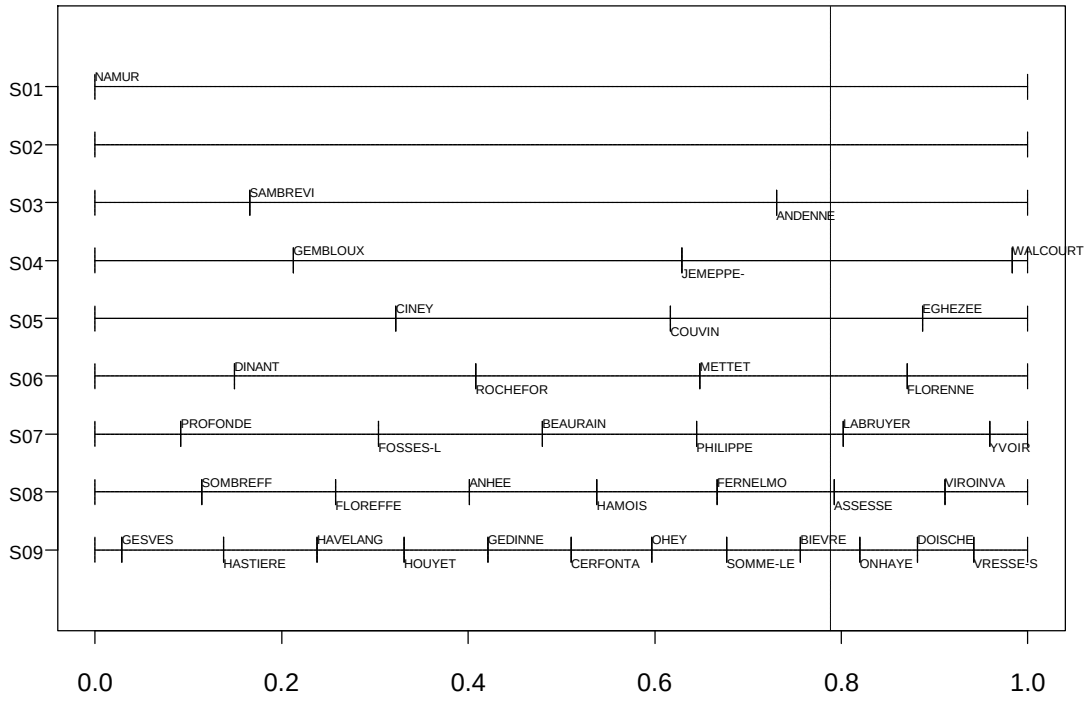
Luxembourg

Systematic Sampling: 5 Clusters



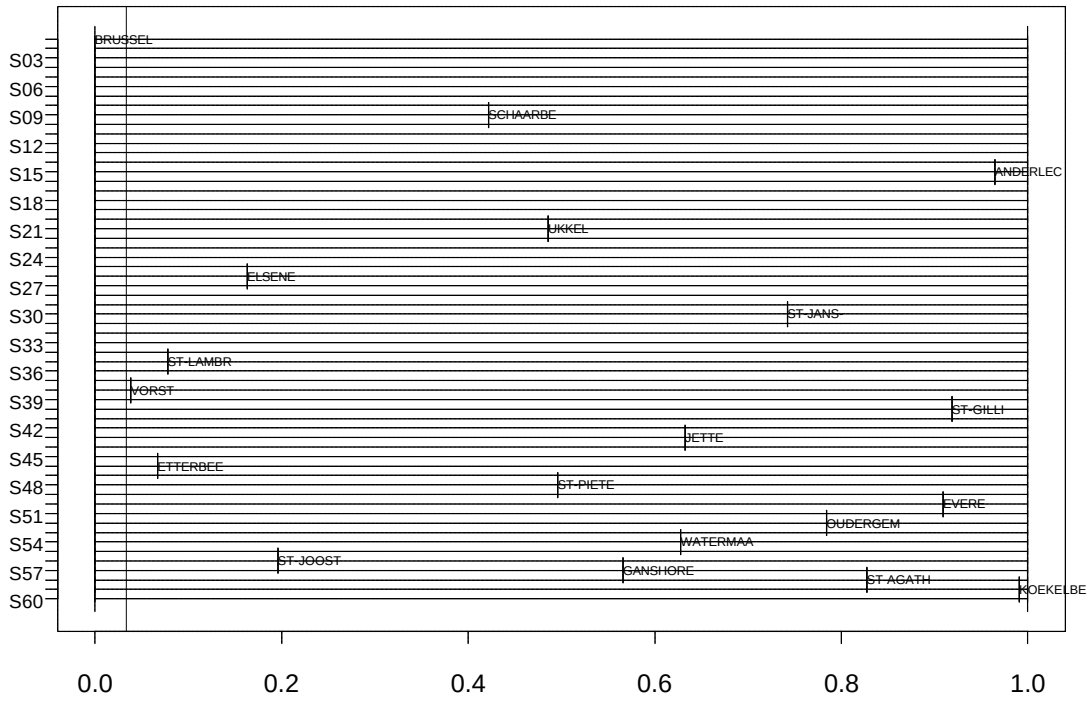
Namur

Systematic Sampling: 9 Clusters



Brussel

Systematic Sampling: 60 Clusters



Systematic sampling : Strata + sample selected

The list below give a description of the above given graphical representation of the systematic sampling. First it is given how many time some municipalities are selected with certainty. Next within the remaining blocs (the horizontal lines on the graphs), the selected municipality is underlined.

Antwerpen: 19

Always Selected: 5
ANTWERPEN 5
(Remaining) Strata: 14
S01: ANTWERPEN MECHELEN
S02: MECHELEN TURNHOUT BRASSCHAAT
S03: BRASSCHAAT HEIST-OP-DEN-BERG GEEL SCHOTEN
S04: SCHOTEN LIER MOL MORTSEL
S05: MORTSEL KAPELLEN HERENTALS BRECHT
S06: BRECHT EDEGEM WILLEBROEK WESTERLO NIJLEN
S07: NIJLEN ZOERSEL BORNEM KONTICH SCHILDE
S08: SCHILDE BALEN SINT-KATELIJNE-WAVER ZWIJNDRECHT RANST KASTERLEE
S09: KASTERLEE KALMTHOUT WUUSTWEZEL STABROEK HOOGSTRATEN ESSEN
S10: ESSEN DUFFEL PUURS BEERSE PUTTE LAAKDAL BOOM
S11: BOOM RUMST LILLE AARTSELAAR BONHEIDEN MALLE HERSELT
S12: HERSELT RAVELS OUD-TURNHOUT BOECHOUT ZANDHOVEN ARENDONK WOMMELGEM OLEN
S13: OLEN GROBBENDONK BORSBEEK RIJKEVORSEL BERLAAR VOSSELAAR RETIE HEMIKSEM MEERHOUT
WIJNEGEM
S14: WIJNEGEM DESSEL HOVE HERENTHOUT HULSHOUT NIEL MERKSPLAS LINT SCHELLE SINT-AMANDS
VORSELAAR BAARLE-HERTOG

Vlaams Brabant: 12

Always Selected: 1
LEUVEN 1
(Remaining) Strata: 11
S01: LEUVEN DILBEEK VILVOORDE HALLE
S02: HALLE GRIMBERGEN TIENEN
S03: TIENEN SINT-PIETERS-LEEUW ASSE AARSCHOT
S04: AARSCHOT ZAVENTEM OVERIJSE BEERSEL DIEST
S05: DIEST SCHERPENHEUVEL-ZICHEM ZEMST TERVUREN HERENT
S06: HERENT ST-GENESIUS-RODE MEISE LONDERZEEL KORTENBERG
S07: KORTENBERG LANDEN ROTSELAAR TERNAT MERCHTEM WEMMEL WEZEMBEEK-OPPEM
S08: WEZEMBEEK-OPPEM LUBBEEK KRAAINEM HAACHT TREMELO AFFLIGEM LIEDEKERKE OPWIJK
S09: OPWIJK MACHELEN KEERBERGEN BOORTMEERBEEK KAMPENHOUT OUD-HEVERLEE ROOSDAAL
STEENOKKERZEEL HOEILAART
S10: HOEILAART TIELT-WINGE KAPELLE-OP-DEN-BOS BEGIJNENDIJK BERTEM BIERBEEK HULDENBERG
GOOIK LENNIK HOLSBEEK
S11: HOLSBEEK ZOUTLEEUW GALMAARDEN KORTENAKEN BOUTERSEM LINTER HERNE HOEGAARDEN
GEETBETS BEKKEVOORT GLABBEEK-ZUURBEMDE DROGENBOS LINKEBEEK PEPINGEN BEVER

Limburg: 9

Always Selected: 0

(Remaining) Strata: 9
S01: HASSELT GENK
S02: GENK BERINGEN SINT-TRUIDEN
S03: SINT-TRUIDEN MAASMECHELEN TONGEREN
S04: TONGEREN HEUSDEN-ZOLDER LOMMEL HOUTHALEN-HELCHTEREN
S05: HOUTHALEN-HELCHTEREN BILZEN LANAKEN MAASEIK
S06: MAASEIK ZONHOVEN DILSEN DIEPENBEEK RIEMST TESSENDERLO NEERPELT
S07: NEERPELT PEER LEOPOLDSBURG BREE LUMMEN HAMONT-ACHEL MEEUWEN-GRUITRODE
S08: MEEUWEN-GRUITRODE OVERPELT KINROOI BOCHOLT HERK-DE-STAD HECHTEL-EKSEL ALKEN
BORGLOON HOESELT
S09: HOESELT HAM OPGLABBEEK HALEN KORTESSEM GINGELOM AS WELLEN HEERS ZUTENDAAL
NIEUWERKERKEN VOEREN HERSTAPPE

Oost-Vlaanderen: 16

Always Selected: 2
GENT 2

(Remaining) Strata: 14
S01: GENT AALST
S02: AALST SINT-NIKLAAS
S03: SINT-NIKLAAS BEVEREN DENDERMONDE
S04: DENDERMONDE LOKEREN NINOVE
S05: NINOVE GERAARDSBERGEN EVERGEM OUDENAARDE
S06: OUDENAARDE DEINZE TEMSE ZOTTEGEM
S07: ZOTTEGEM RONSE WETTEREN HAMME
S08: HAMME MALDEGEM MERELBEKE ZELE ERPE-MERE EEKLO
S09: EEKLO LOCHRISTI AALTER DESTELBERGEN HAALTERT
S10: HAALTERT LEDE DENDERLEEUW LEBBEKE SINT-GILLIS-WAAS STEKENE
S11: STEKENE HERZELE KRUIBEKE ZULTE BRAKEL BUGGENHOUT ASSENEDE
S12: ASSENEDE BERLARE OOSTERZELE ZELZATE LAARNE GAVERE WICHELEN NEVELE
S13: NEVELE NAZARETH MELLE DE PINTE WAASMUNSTER LOVENDEGEM
ST-LIEVENS-HOUTEM ST-MARTENS-LATEM ZOMERGEM WAARSCHOOT
S14: WAARSCHOOT KNESSELARE KRUISHOUTEM ZWALM WACHTEBEKE
ST-LAUREINS ZINGEM MAARKEDAL LIERDE KAPRIJKE WORTEGEM-PETEGEM
KLUISBERGEN MOERBEKE HOREBEKE

West-Vlaanderen: 14

Always Selected: 1
BRUGGE 1

(Remaining) Strata: 13
S01: BRUGGE KORTRIJK
S02: KORTRIJK OOSTENDE
S03: OOSTENDE ROESELARE WAREGEM
S04: WAREGEM IEPER KNOKKE-HEIST
S05: KNOKKE-HEIST MENEN WEVELGEM IZEGEM
S06: IZEGEM HARELBEKE ZWEVEGEM ZEDELGEM
S07: ZEDELGEM OOSTKAMP TIELT KOKSIJDE POPERINGE
S08: POPERINGE TORHOUT WERVIK BLANKENBERGE MIDDELKERKE
S09: MIDDELKERKE DIKSMUIDE BEERNEM ANZEGEM JABBEKE ICHTEGEM BREDENE
S10: BREDENE KUURNE KORTEMARK WINGENE VEURNE DEERLIJK ZONNEBEKE DE HAAN
S11: DE HAAN MEULEBEKE DAMME STADEN INGELMUNSTER MOORSLEDE GISTEL NIEUWPOORT
S12: NIEUWPOORT DE PANNE ARDOOIE HOOGLEDE LEDEGEM AVELGEM HOUTHULST WIELSBEKE
OUDENBURG HEUVELLAND
S13: HEUVELLAND LICHTERVELDE KOEKELARE DENTERGEM LANGEMARK-POELKAPELL OOSTROZEBEKE
PITTEM LENDELEDE RUISELEDE ALVERINGEM VLETEREN LO-RENINGE ZUIENKERKE
SPIERE-HELKIJN MESEN

Brabant Wallon: 7

Always Selected: 0

(Remaining) Strata: 7
S01: BRAINE-L'ALLEUD WAVRE
S02: WAVRE WATERLOO OTTIGNIES-LLN
S03: OTTIGNIES-LLN NIVELLES RIXENSART
S04: RIXENSART TUBIZE GENAPPE
S05: GENAPPE LASNE GREZ-DOICEAU JODOIGNE REBECQ
S06: REBECQ CHAUMONT-GISTOUX VILLERS-LA-VILLE BRAINE-LE-CHATEAU COURT-ST-ETIENNE
ORP-JAUCHE LA HULPE
S07: LA HULPE PERWEZ CHASTRE BEAUVECHAIN MONT-ST-GUIBERT ITTRE WALHAIN RAMILLIES
INCOURT HELECINE

Hainaut: 25

```
Always Selected: 8
       CHARLEROI   3
            MONS   1
     LA LOUVIERE   1
         TOURNAI   1
        MOUSCRON   1

(Remaining) Strata: 17
S01: CHARLEROI MONS
S02: MONS LA LOUVIERE
S03: LA LOUVIERE TOURNAI MOUSCRON CHATELET
S04: CHATELET BINCHE COURCELLES
S05: COURCELLES ATH
S06: ATH SOIGNIES FLEURUS ST-GHISLAIN
S07: ST-GHISLAIN MANAGE COLFONTAINE
S08: COLFONTAINE FRAMERIES BOUSSU
S09: BOUSSU QUAREGNON MORLANWELZ BRAINE-LE-COMTE
S10: BRAINE-LE-COMTE COMINES-WARNETON FONTAINE-L'EVEQUE DOUR
S11: DOUR PERUWELZ LESSINES PONT-A-CELLES
S12: PONT-A-CELLES THUIN CHAPELLE-LEZ-HERLAIM BELOEIL LEUZE-EN-HAINAUT
S13: LEUZE-EN-HAINAUT HAM-SUR-HEURE-NALINN FARCIENNES GERPINNES ANDERLUES
S14: ANDERLUES BERNISSART AISEAU-PRESLES FRASNES-LEZ-ANVAING SENEFFE ENGHIEN
S15: ENGHIEN MONTIGNY-LE-TILLEUL CHIMAY ERQUELINNES ECAUSSINES ESTAIMPUIS
S16: ESTAIMPUIS JURBISE LES BONS VILLERS LE ROEULX ANTOING BRUNEHAUT ESTINNES
S17: ESTINNES QUEVY SILLY QUIEVRAIN HENSIES BEAUMONT CHIEVRES ELLEZELLES LOBBES
S18: LOBBES CELLES RUMES MOMIGNIES PECQ HONNELLES SIVRY-RANCE MERBES-LE-CHATEAU
       LENS BRUGELETTE FLOBECQ MONT-DE-L'ENCLUS FROIDCHAPELLE
```

Liège: 18

```
Always Selected: 5
        LIEGE   3
      SERAING   1
     VERVIERS   1

(Remaining) Strata: 13
S01: LIEGE SERAING VERVIERS HERSTAL
S02: HERSTAL ANS
S03: ANS FLEMALLE OUPEYE ST-NICOLAS
S04: ST-NICOLAS GRACE-HOLLOGNE CHAUDFONTAINE
S05: CHAUDFONTAINE HUY VISE HERVE
S06: HERVE FLERON SOUMAGNE DISON
S07: DISON ESNEUX WAREMME AMAY HANNUT
S08: HANNUT BLEGNY WANZE SPRIMONT BEYNE-HEUSAY MALMEDY
S09: MALMEDY THEUX SPA AYWAILLE NEUPRE PEPINSTER
S10: PEPINSTER PLOMBIERES WELKENRAEDT AWANS JUPRELLE BASSENGE TROOZ
S11: TROOZ JALHAY ST-GEORGES-SUR-MEUSE STAVELOT WAIMES DALHEM
       ENGIS LIMBOURG VILLERS-LE-BOUILLET BRAIVES
S12: BRAIVES COMBLAIN-AU-PONT NANDRIN MARCHIN THIMISTER-CLERMONT
       REMICOURT HERON FERRIERES CLAVIER ANTHISNES AUBEL BAELEN OLNE
S13: OLNE MODAVE HAMOIR VERLAINE LIERNEUX OREYE FAIMES STOUMONT
       FEXHE-LE-HAUT-CLOCHE LINCENT BURDINNE BERLOZ OUFFET DONCEEL GEER
       CRISNEE TROIS-PONTS TINLOT WASSEIGES
```

German Community: 6

```
Always Selected: 1
     EUPEN   1

(Remaining) Strata: 5
S01: EUPEN KELMIS
S02: KELMIS RAEREN
S03: RAEREN ST-VITH BUTGENBACH
S04: BUTGENBACH BULLINGEN AMEL
S05: AMEL LONTZEN BURG-REULAND
```

Luxembourg: 5

Always Selected: 0

(Remaining) Strata: 5
S01: ARLON MARCHE-EN-FAMENNE AUBANGE
S02: AUBANGE BASTOGNE VIRTON DURBUY LIBRAMONT-CHEVIGNY
S03: LIBRAMONT-CHEVIGNY BERTRIX VIELSALM MESSANCY HABAY
NEUFCHATEAU ST-HUBERT FLORENVILLE BOUILLON
S04: BOUILLON PALISEUL CHINY NASSOGNE HOTTON HOUFFALIZE ETALLE
GOUVY LIBIN LA-ROCHE-EN-ARDENNE VAUX-SUR-SURE MUSSON
S05: MUSSON ATTERT LEGLISE TINTIGNY SAINT-LEGER WELLIN MANHAY
MEIX-DEVANT-VIRTON EREZEE BERTOGNE TENNEVILLE RENDEUX TELLIN ST-ODE
ROUVROY FAUVILLERS MARTELANGE HERBEUMONT DAVERDISSE

Namur: 9

Always Selected: 2
NAMUR 2

(Remaining) Strata: 7
S01: NAMUR SAMBREVILLE ANDENNE
S02: ANDENNE GEMBLOUX-SUR-ORNEAU JEMEPPE-SUR-SAMBRE WALCOURT
S03: WALCOURT CINEY COUVIN EGHEZEE
S04: EGHEZEE DINANT ROCHEFORT METTET FLORENNES
S05: FLORENNES PROFONDEVILLE FOSSES-LA-VILLE BEAURAING PHILIPPEVILLE LA BRUYERE
YVOIR
S06: YVOIR FLOREFFE SOMBREFFE ANHEE HAMOIS FERNELMONT ASSESSE VIROINVAL
S07: VIROINVAL GESVES HASTIERE HAVELANGE GEDINNE HOUYET
CERFONTAINE OHEY SOMME-LEUZE BIEVRE ONHAYE DOISCHE VRESSE-SUR-SEMOIS

Brussel: 60

Always Selected: 50

BRUSSEL	8
SCHAARBEEK	6
ANDERLECHT	5
UKKEL	4
ELSENE	4
ST-JANS-MOLENBEEK	4
ST-LAMBRECHTS-WOLUWE	2
VORST	2
ST-GILLIS	2
JETTE	2
ETTERBEEK	2
ST-PIETERS-WOLUWE	2
EVERE	1
OUDERGEM	1
WATERMAAL-BOSVOORDE	1
ST-JOOST-TEN-NODE	1
GANSHOREN	1
ST-AGATHA-BERCHEM	1
KOEKELBERG	1

(Remaining) Strata: 10

S01: BRUSSEL SCHAARBEEK ANDERLECHT
S02: ANDERLECHT UKKEL
S03: UKKEL ELSENE ST-JANS-MOLENBEEK
S04: ST-JANS-MOLENBEEK ST-LAMBRECHTS-WOLUWE
S05: ST-LAMBRECHTS-WOLUWE VORST ST-GILLIS
S06: ST-GILLIS JETTE
S07: JETTE ETTERBEEK ST-PIETERS-WOLUWE EVERE
S08: EVERE OUDERGEM
S09: OUDERGEM WATERMAAL-BOSVOORDE
S10: WATERMAAL-BOSVOORDE ST-JOOST-TEN-NODE GANSHOREN ST-AGATHA-BERCHEM KOEKELBERG

Appendix 2.3
Oversampling to cope with household drop-out and variable household membership factor.

A. Requirements

Oversampling is necessary, because:

- *Household drop-out*: households which cannot be contacted or refusals.
- *Variable household membership factor*: the number of interviews to be made is fixed. However households are sampled, not individuals. In advance, it can only be predicted roughly how many respondents will be achieved. Also it is possible that respondents within a household refuse.

B. Boundary Conditions

- For practical reasons it is necessary to have in advance a list with all respondents sufficiently large.
- Oversampling is not an issue. However one should ensure that a difficult to contact household is not replaced by an easy to contact household.

C. Sampling Scheme to Generate the Oversampling

- To cope with above problems it is choosen to generate a sample which contains 8 times as much households as necessary:
  1. For each household 3 possible candidates for replacement are selected.
  2. For each municipality the distribution of the household members is known. Based on this one can estimate the number of households necessary to contact for realizing a sufficient number of interviews. It is proposed to take twice as much households as expected.
- This sample is realized by a clustered systematic sampling within the PSU. First the households are hierachically ordered by 1) statistical sector, 2) their size (1, 2, 3, 4,...,) and the age (in years) of the reference person. To eliminate boundary effects, the ordering of the lower order variable (household size; age) is alternatively increasing and decreasing. The step of the systematic sampling is chosen twice as small as necessary to realize the expected number of households. By each step four consecutive households are selected.

D. Rules to use the sample

- Above sample is organized in a table with four columns. Each row (the horizontal direction of the table) consists of a cluster of four consecutive households to cope with drop-out. The extra rows (the vertical direction) are buffer against the uncertainty about the number of members in the households. To prevent from any order effects a randomization is done within rows and also the rows themselves are randomized.
- A row is stopped when at least one interview is realized. All remaining households are considered as never been sampled and hence they should not be included in response rates.
- The randomization of the rows is necessary because in advance it is not known how many rows are needed to achieve the number of interviews needed to start with a first series of households. If this series does not result in the required number of interviews, then the next row(s) should be started. If there is any order present in the table, then the replacement will be biased.
- In addition, to spread the interviews also in time during a quarter, one should start only with a few households. As a first estimate the number of households given by the National Register can be used.
- Because the sampling frame from which the households are dynamic and the copy used is a not recent one the vital status and the address of reference person have to be controled. This can be done in an automatic way. Only in the case of the death of a reference person of a more than one person household a manual online verification to trace the remaining of the household is necessary. Following situations are possible

    1. Reference person <u>died</u>.

        If one person household : stop (household disappears)

        If 2 or more person household: try to trace the rest of the household

            If untracable: stop

            If address still within the municipality: contact new reference person

            If outside: stop

    2. Reference person <u>moved</u> (for any reason: total household moves; part of the household moves (e.g. divorce), ...).

            If address still within municipality: contact household

            If not: stop

E. Example

1. For a PSU 50 interviews (= one group) should be done.

mean household size is 1.9

the number of households to be interviewed is (12.5/1.9) = 6.6 = 7.

the number of clusters of households (n = 4) is 7 * 2= 14

the number of households to be sampled = 14 * 4 = 56 households

Only some of these 56 households will be contacted following a strict protocol as it is estimated that 7 household need to be interviewed successfully in order to obtain 12.5 individuals interviewed.

2. Step size for the systematic sampling.

Say N = the number of households within a PSU

n = number of clusters of households (to be selected)

the step size y = (N-3)/n. In the selection program (for the first quarter) the step size was approximated by N/n as N is in general much larger than n.

start first step at a random number x between 1 and y

the first cluster is : x, x + 1, x + 2, x + 3

the second cluster is : (x + y), (x+ y) + 1, (x + y) + 2), (x + y) + 3

...

the nth row: (x + (n-1)y); (x + (n-1)y) + 1; (x + (n-1)y) + 2; (x + (n-1)y) + 3;

Step	**Replacement households**			
	1	2	3	4
1	nr 011	nr 012	nr 013	nr 014
2	nr 021	nr 022	nr 023	nr 024
3	nr ij*			
4				
5				
...				
...				
13				
14	nr 141	nr 142	nr 143	nr 144

* : nr ij: I is the ith step, j is the order of the household in the ordered national register.

3. Reorganizing the sample.

The n-rows are randomized. Then within each row (a cluster of 4 households), the columns are randomized. The result of this step is given in table 2.

Table 2

Row	**Step**	**HH size**	**Replacement households**			
			1	2	3	4
1	3	1	nr 034	nr 033	nr 031	nr 032
2	10	4	nr 102	nr 103	nr 104	nr 101
3	6	2				
4	4	1				
5	8	3				
6	14	4+				
...	...					
13	3					
14	12		nr 123	nr 121	nr 124	nr 122

4. The check of the sample information with the last available information of the National Register.

Information on all selected households will be cross-checked with the more up-to-date information in the National Register. This is done within the month prior to the first mailing to the selected households every quarter.

- A file with the selected reference persons is sent to the National Register. The actual address and the vital status of the reference person is matched to that list and the matched file will return to the NIS.
- The matched updated file is matched to the information of the sample and the address will be up-dated

For the reference persons of a more than one person household who is death an on-line search in the National Register is done to identify the remaining individuals of the household and the actual address (table 3).

Table 3. Status of the household after the control with the National Register

Code	Description	Action
1	Reference person alive at same address	
2	Reference person died, 1 person household : household died	household drops out
3	Reference person died, 2 or more person household household still at same address	change name of reference person
4.	Reference person died, 2 or more person household household moved within municipality	change name of reference person and address
5	Reference person died, 2 or more person household household moved out municipality	change name of reference person and address / household drops out
6	Reference person died, 2 or more person household household can not been traced	household drops out
7	Reference person alive, moved within municipality	change address
8	Reference person alive, moved out municipality	change address / household drops out
9	Reference person alive, moved but can not been traced	household drops out

F. Management of the sample.

An algorithm is developed to link the number of households to be contacted with the household size as known from the National Register and in a second phase with the real household size. This is necessary because only 12.5 individuals may be interviewed per quarter. The algorithm continues to select the first eligible household within the cluster until the sum of the individuals in the households to be interviewed is the closed to 12.5, with possible range 11 to 14. In case of equal distance to 12.5, there is an random process to remain under 12.5 or to go over it.
The program identifies the replacement household: in case of a household non-response, the replacement is the next eligible household within the cluster; in case there are no more eligible households in the cluster or there are less interview within the household as was expected the next eligible cluster(s) will provide the replacement household(s). In case of equal distance to 12.5 the household is selected to reach the upper value.

For the second trimester and in the 4th trimester it was decided to adopt the algorithm in such a way that the number to contact individuals was no longer 12.5 an average per group and quarter but higher (table 1). This was done to speed up the field work by giving the interviewer a larger batch of households to contact.

Table 4

See-sow point in algorithm to initiate a cluster of households

	Flemish region	Walloon region	Brussels region
1 Quarter	12.5	12.5	12.5
2 Quarter	13.0	13.0	13.0
3 Quarter	12.5	12.5	12.5
4 Quarter	13.0	17.0	19.0

For the 4th quarter and within the Walloon en Brussels region this adaptation may introduce some selection towards the most easy to reach household as the data collection will be terminated when the regional quata are reached.

Appendix 2.4
Results of the sampling 1997.

1. Quarter 1

A total of 8752 households have been selected. However 3 clusters of 4 households were identified later as non-resident, and are not eligible. To avoid this problem, all references with a statistical sector equal to A followed by 3 blancs or with a street code which ends on 9999 should be excluded from the sampling frame.

The remaining 8740 households respresent 19599 individuals. They had a mean household size of 2.24. The households were selected within 144 different PSUs or municipalities. 22 PSUs have been selected more than one time.

92% of the sample of reference persons remains eligible to have their household invited. The reason to be no longer eligible is mainly because the household moved out the PSU: 6.4% of the reference persons (table page 35). Only in 11 cases could the household not been matched with the updated file. Online verification was necessary in only 59 cases, or in less than 1% of the selected household.

	Region	Number of households	Number of individuals	Mean House-hold size
Quarter 1				
1/1/96 sampling frame	Flemish region	2832	6797	2.40
	Brussels region	3064	6280	2.05
	Walloon region	2844	6522	2.29
	Belgium	8740	19599	2.24
Quarter 2				
1/1/97 sampling frame	Flem	2832	6902	2.44
	Bruss	3080	6207	2.02
	Wall	2864	6847	2.39
	Belgium	8776	19956	2.27
Quarter 3				
1/1/97 sampling frame	Flem	2832	6855	2.42
	Bruss	3080	6393	2.08
	Wall	2864	7160	2.50
	Belgium	8776	20408	2.32
Quarter 4				
1/1/97 sampling frame	Flem	2826	7076	2.50
	Bruss	3065	6213	2.03
	Wall	2848	6934	2.44
	Belgium	8739	20223	2.31

2. Quarter 2, 3, 4

No controle was done for these quarters.

Bibliography

1. Quataert, P. Van Oyen, H. Gegevensinzameling i.v.m. middelengebruik d.m.v. CATI. IHE/Episerie n°6 Ed. Brussel. C.O.O.V., Instituut voor Hygiëne en Epidemiologie, 1995;

2.Hermans H, Lambert M, Reginster G, Tafforeau J, and Van Oyen H. Naar een gezondheidsenquête door middel van interview in België. Brussel. IHE, 1995; 1-193.

3. Demarest, S., Tellier, V., Van der Heyden, J., Schiettecatte, E., Tafforeau, J., Van Oyen, H. Health Interview Survey, 1997. Interviewers Guide. Brussels. Department of Epidemiology, SIP, 1996; 1-65.