Grid sampling with an application to a mixed-mode human survey

Seppo Laaksonen[1]

[1]University of Helsinki, e-mail: Seppo.Laaksonen@Helsinki.Fi

Abstract

Two types of strategies are used for sampling designing. One strategy is a standard stratified random sampling with regional strata, but the other uses special strata. These strata are based on 250mx250m grids so that all the grids are sorted by the income medians of the residents and two explicit strata are constituted. One of these grid strata consists of the grids with low income whereas the other of the grids with high income. These two types of samples are overlapping partially. However, there is need to use both samples in one framework. This leads to a non-trivial strategy for sampling and estimation.
Keywords: Strata, stratum overlapping, conditional inclusion probability

1 Introduction

The European Social Survey (ESS) is one of the most qualified surveys in Europe. Its sampling design varies from one country to the next, but we can still recognise the following basic features from these:

- Simple random sampling (srs) so that the study units (15+ aged residents) are explicitly available from a register.
- Random sampling with explicit strata, using often registers as well.
- Two stage cluster sampling so that the first stage units are small-area primary sampling units (psu's), whereas the two-stage units are directly as study units.
- Three stage cluster sampling so that the first-stage units are small-area primary sampling units (psu's), but the second-stage is needed to draw households or addresses before drawing the study units.

Srs naturally does not use stratification but the other three strategies often use, but not always. The main line in the ESS is that if stratification is used, it is explicit stratification and sample allocation is proportional, exactly or approximately. There are however many countries that use non-proportional allocation and even so that the anticipated response rates has an effect on the gross sample size. This has been made cautiously, for example so that the gross sampling fraction for large cities is higher than for rural areas. This thus, since the response rates seem to be low in large cities. In some cases, non-proportional allocation is made in order to obtain enough accurate results for certain explicit strata; the reason for this is national, it is not required by the ESS coordinating committee. See some information about the ESS sampling, Lynn et al 2007. The ESS website includes also useful information such as the sampling design principles (http://www.europeansocialsurvey.org/index.php?searchword=sampling&ordering=&searchphrase=all&Itemid=217&option=com_search).

We can consider the ESS as a standard survey in the sense that even though explicit strata are used, these are rather traditional such as administrative regions of a country. In this study, we go forward although we also use a very standard explicit stratification. On the other hand, our sample allocation is not proportional at all, but such that gives opportunity to get enough accurate estimates for specific strata. It should be noted that the use of anticipated response rates cannot be here used well, since our survey is rather unique and any a priori information does not exist. So, we hope that our 'intuition' for sample allocation was enough good from this point of view[1].

Administrative regions are important but people within these regions may be very different and the results obtained from these do not tell much about their attitudes or feelings, among others. Hence we try to go on to smaller areas and without administrative constraints. One strategy is to use Geographic Information System (GIS) so that small areas are the grids of 250 metres times 250 metres. People living within such small squares are expected to be as neighbours of each other, and hence their attitudes, feelings and opinions are maybe clustered to some extent. The whole data when being available in late 2012 give opportunity to analyse in very details urbanization vs ruralisation issues in the south Finland, in 16 municipalities totally, including Helsinki and its neighbour municipalities. This subject-matter analysis is forthcoming. This paper describes the sampling design strategies of the study.

The paper is organised so that we explain in Section 2 the target population and the sampling frame. Section 3 concentrates on the sampling design itself. It is good already to notice that we use the two different designs in fact. This leads to certain technical challenges that are solved in an interesting way in the next section. We present our solutions also empirically, using the gross sample data. Section 4 presents an interesting solution to calculate the inclusion probability. The final section discusses further steps that are possible to specify after the data from the respondents are available. The fieldwork has been conducted using such a mixed-mode design that gives for a potential respondent to participate either by web or by postal mail. This was considered to be best because it is inexpensive. In the forthcoming paper we also analyse the effects of this mixed-mode strategy that is rather new but becoming more common (see e.g. the ESS website: http://www.europeansocialsurvey.org/index.php?option=com_content&
view=article&id=67&Itemid=552).

2 Target population and sampling frame

The statistical units of the target population are 25-74 years old residents of 16 Finnish southern municipalities those mother tongue is either Finnish or Swedish. The information is based on the January 2012 population register. Our sampling frame has also constructed from this register.

From the regional point of view we have however two target populations, one being just those 16 municipalities. But the second is more complex and it is based on 250m x 250m grids of 14 out of these 16 municipalities. The reason for this is that two municipalities decide not to participate in this second study.

The first target population is divided into 19 explicit strata that are equal to the municipalities except that Helsinki consists of the three strata (most urbanised southern area, most urbanised northern area, suburb area). These are also administrative areas.

For the second regional target population, the income of the grids was used. The income concept is the taxable income from the 2010 taxation register. The median income of all the grids was computed and then the grids were sorted by this order, from the lowest median to the highest median. Consequently, two groups or strata were formed, the lowest quintile (called also 'poor') vs the highest quintile (called also 'rich'). This in-

[1] This study is initiated by Matti Kortteinen and Mari Vaattovaara from the University of Helsinki. My role has been and will be to help in methodological issues including sampling design that is the focus of this paper now.

formation was received from Statistics Finland who maintains the grid data base with population and taxation statistics data. Before determining the final strata, some robustness was made so that some initial grids were omitted. The basic reason was to protect people of too small grids. This was based on the confidentiality declaration of Statistics Finland.

When the set of grids was made robust, the two strata were ready to use. The first quintile thus constitutes one stratum and the fifth quintile the second, respectively. The map of Figure 1 shows how these two strata are spread around our municipalities. It is easy to see that 'rich' grids are concentrated on certain areas, and 'poor' grids on the other, respectively. However, any of them do not cover any whole municipality. There are empty areas from both types of grids, that is, their median income is somewhere in the middle (no poor, no rich) or the grids are 'closed' for confidentiality reasons.

Figure 1. Grids for 'rich' people vs. 'poor' people in the municipalities of the survey. The remaining grids are between those two ones

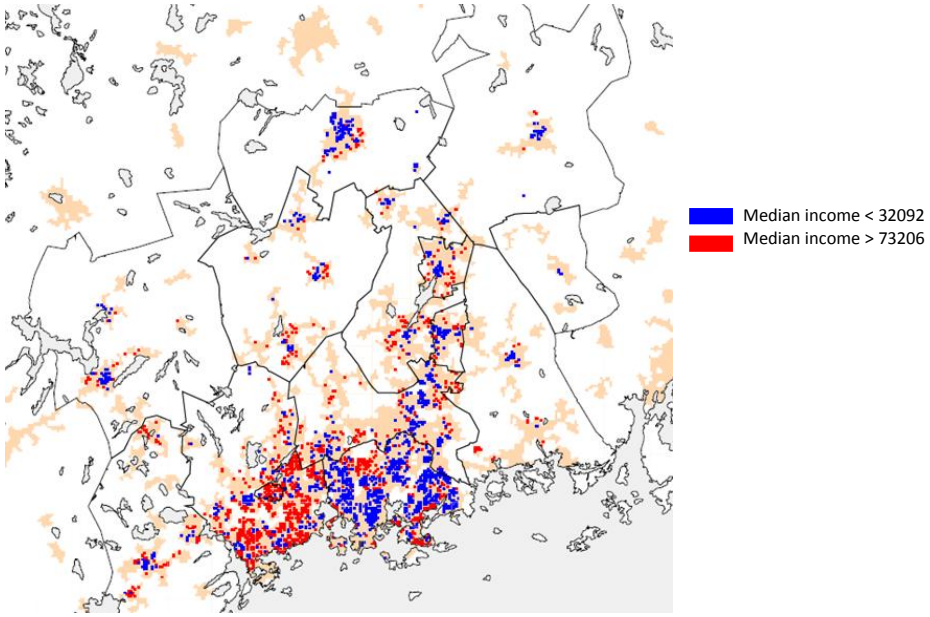


Table 1 shows what has been the 'intuition' of our research group. The grid-based stratum sizes were desired to be enough big in order to get enough accurate estimates. In the next section we come back to this issue and observe that the actual gross sample sizes are even higher due to sampling selection process used. The municipality based stratum sizes have been allocated much with a minimum principle that in our case means the 600 gross sample size, at minimum. Obviously this ensures that we will have enough respondents to estimate results reasonably well. If the response rate in such small municipalities would be for example 50%, we will get 300 respondents from this site of the data. This number is expected to increase from the grid site to some extent.

Table 1. Allocation of gross sample

Stratum	**Gross sample size**
Grids of 5th quintile income (High income grids, 'Rich')	6 000
Grids of 1th quintile income (Low income grids, 'Poor')	6 000
All income based strata	**12 000**
Espoo and Kauniainen	2 000
Helsinki, most urbanised southern area	1 000
Helsinki, most urbanised northern area	1 000
Helsinki, suburb	2 500
Hyvinkää	600
Järvenpää	600
Kauniainen	600
Kerava	600
Kirkkonummi	600
Lahti	1 000
Lohja	600
Mäntsälä	600
Nurmijärvi	600
Pornainen	600
Sipoo	600
Tuusula	600
Vantaa	1 500
Vihti	600
All municipality based strata	**15 000**
The whole gross sample	**27 000**

3 Sampling design

The sampling design for the both two parts of the survey is stratified random sampling. However, the design is not any standard such design, since these both samples are dependent. That is, the grid-part residents can be drawn to the sample also from the municipality part. Lahti and Lohja are the exceptions, their sampling design is exactly stratified random sampling.

The sample selection was performed by a sub-contractor who has access to the population register and to the grid information. The sub-contractor received the instructions to draw a sample but this could not be done so that all sampling principles were possible to take into account. The sample selection process was as follows.

First, the grid sample part was selected with the desired amount of respondents. This was done by addresses, so that one valid person from one address only was accepted. At the same time, this address selected was marked for the second round of the sampling selection that was concerned municipality samples. Thus, this second round was conditional to the first round, and hence it was not possible to draw the same person twice in the sample.

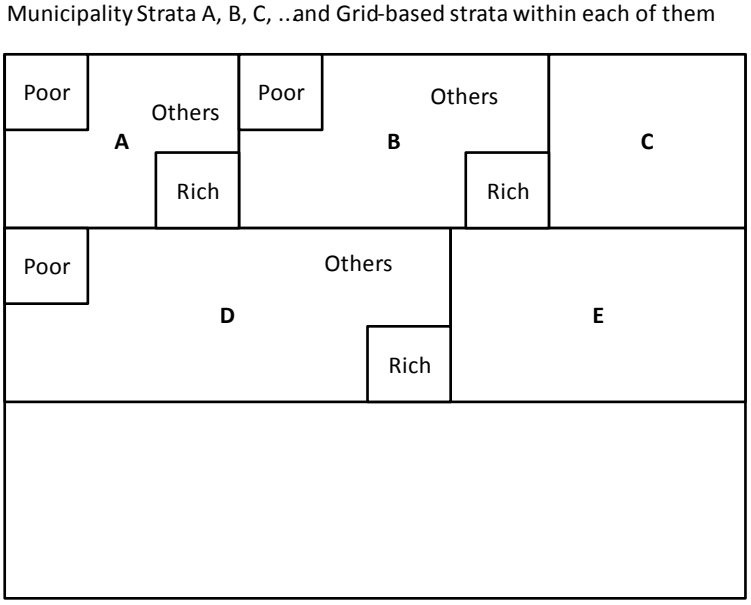
The inclusion probabilities are straightforwardly computable for Lahti and Lohja since any conditionality problem does not exist. They are as usually:

$$\pi_k = \frac{n_h}{N_h}$$

Here *h* is stratum (Lahti or Lohja), *n* is desired gross sample size and *N* = number of 15-74 years old residents, respectively.

The inclusion probabilities for the other municipalities and strata are more difficult to compute, since we have to know what is the probability for a selected person to be included in the sample? This probability was

not available and hence we estimated it using the received gross sample that was possible after both samples were available. In order to illustrate the problem better, I use the following scheme:



First, we matched those two gross samples together so that each grid-sample person is identified to its municipality stratum. This was made by using postal zip codes of both data files. This identification worked well, although it was not definitely sure in advance. Now we were able to calculate the overlapping gross sample sizes and get good opportunities for estimating the inclusion probabilities. It is good to remind that sample selection is random within both grid strata, and it can be assumed that the distribution of the gross sample into municipalities corresponds approximately to the target population distribution. This assumption is in any way used in this study even not being perfectly true. It is clear that this uncertainty should be taken into account in variance estimation that is not included in this paper.

Table 2. Distribution of gross sample to strata. The group 'Others' in the above scheme is equal to municipality gross sample size.

	Poor grids	Rich grids	Municipality	Total	25-74 year Population
Helsinki, most urbanised southern area	110	46	1000	1156	27465
Helsinki, most urbanised northern area	1142	8	1000	2150	40206
Helsinki, suburb	2501	1324	2500	6325	147098
Espoo-Kauniainen	546	3127	2000	5673	131840
Hyvinkää	248	64	600	912	24944
Järvenpää	115	38	600	753	21717
Kerava	124	48	600	772	18874
Kirkkonummi	89	173	600	862	20065
Lahti	0	0	1000	1000	57059
Lohja	0	0	600	600	22613
Mäntsälä-Pornainen	49	22	600	671	13850
Nurmijärvi	85	120	600	805	21924
Sipoo	48	134	600	782	10269
Tuusula	118	201	600	919	20948
Vantaa	746	574	1500	2820	104930
Vihti	81	121	600	802	15923
All	6000	6000	15000	27000	699725

The inclusion probabilities are required to calculate separately to the three groups:

- for poor grids areas
- for rich grids areas
- for others who however can live either in poor grids, in rich grids or in intermediate poor/rich areas.

The sampling design for municipalities is independent of richness or poorness of their living grids, and hence the inclusion probabilities need to be calculated following this fact. For the analysis, it is of course possible to identify the respondents correctly by their grid. This information is also included in the data file.

In fact, we cannot calculate the inclusion probabilities straightforwardly. We had to 'estimate' them as explained below. I present them as the following formula:

$$\pi_k = \frac{n_{hc}}{\hat{N}_{hc}}$$

Here *c* is income level stratum (poor, rich, others). We have gross sample sizes for each strata as presented in Table 2, but we cannot know precisely population sizes for these overlapping strata. Hence we estimate them assuming that the gross sample size represents correctly to the corresponding population size. The formula for these statistics, e.g. for the stratum *h1* is as follows:

$$\hat{N}_{h1} = N_h \frac{n_{h1}}{(n_{h1} + n_{h2} + n_{h3})}.$$

4 Sampling design weights

When we have the inclusion probabilities, we can easily calculate the gross sample design weights:

$$w_k = \frac{1}{\pi_k}$$

Table 3 illustrates these design weights from our data. We see that the weights vary quite much that is due to the desired targets for the sample sizes. The variation for the grid part is smaller than for the municipality part.

Table 3. Some statistics of the gross sample design weights

Statistics	The whole sample	Grid part	Municipality part
Observations	27000	12000	15000
Mean	25.9	24.4	27.1
Total	699725	292615	407110
Minimum	13.1	13.1	13.1
Maximum	57.1	37.2	57.1
CV (%)	31.7	20.5	36.4

Note: The overlapping is useful thus for the our big point, to compare people's attitudes, living conditions etc within different types of very small areas, such as 250m x 250 grids. I already mentioned that the gross sample size (and net sample size consequently) will be increased from the initial targets due to the overlapping. When identifying people of the municipality sample into poor vs rich grids, our gross sample size was increased essentially, from 6000 to 9572 in poor grids, but only from 6000 to 6992 in rich grids. We can thus observe that a random selection provides relatively much more people from poor grids than from rich grids.

5 Concluding remarks and future

This is a new and obviously innovative approach to survey sampling, especially for stratification. The GIS data are used for many purposes but not so much for sampling and estimating. At least, I have not seen the approach like this in literature.

Our respondent data are soon becoming to be available. This gives opportunity to create the sampling weights for the respondents. Such initial or base weights are easy to compute, that is, just to change gross sample sizes *n* to the corresponding net sample sizes, let say *r*. This weighting is only the start for constructing good sampling weights. These require also to analyse non-response and to adjust for it. My plan is to use the response propensity modelling first and then to calibrate the sums of the resulted weights into the sums of the gross sample weights. This will be done at each stratum level so that overlapping strata are covered too (e.g. Laaksonen 2007).

The response propensity modeling is more advantageous if good auxiliary variables are available. Our pattern is not perfect, thanks for the problem that we are outside Statistics Finland who has more such variables easily available. We have not obtained for example education that is too hard to get for outsiders, but we have many population register variables fortunately, such as age, gender, mother tongue, dwelling unit structure, previous living area, house type and house size. Our group is also going to ask basic information from the taxation register and the employment register.

References

Laaksonen, S. (2007). Weighting for Two-Phase Surveyed Data. *Survey Methodology*, December Vol. 33, No. 2, pp. 121-130, Statistics Canada.

Lynn, P. & Gabler, S. & Häder, S. & Laaksonen, S. (2007). Methods for Achieving Equivalence of Samples in Cross-National Surveys. *Journal of Official Statistics*, 27, 1, 107-124.