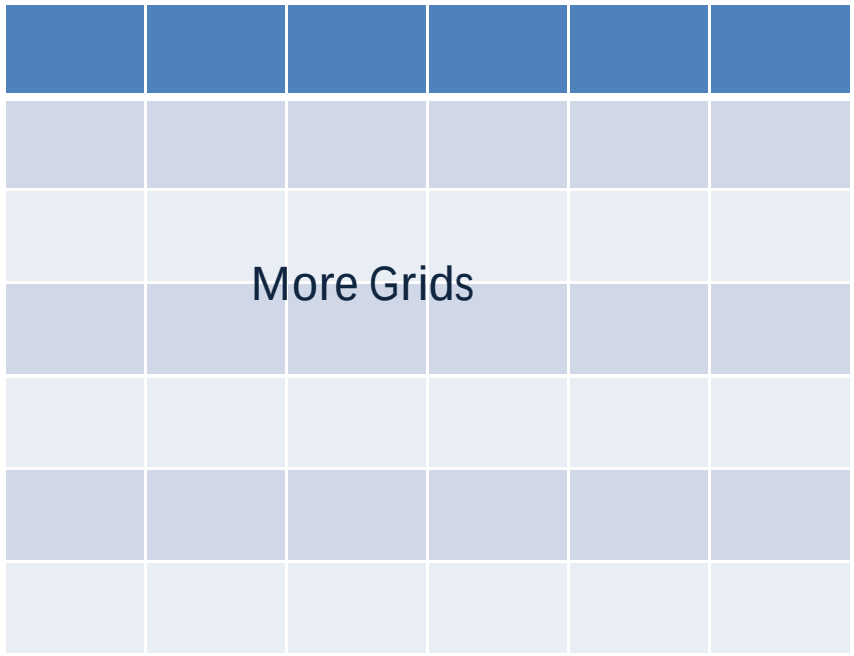

A Grid in Lithuania

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

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



My title includes thus the concept

MIXED-MODE SURVEY

I am not going to talk much about this currently popular issue but more about sampling and also on responding to my reference mixed-mode survey that is an ongoing project initiated by the sociologist Matti Kortteinen and the geographer Mari Vaattovaara, both from the Helsinki University. There are also in this project such researchers as Teemu Kemppainen and Henrik Lönnqvist, and some sub-contractors as Statistics Finland, the Central Population Register assisted by Logica, the Finnish Taxation register and the Employment register.

Note that there are two major effects in mixed-mode surveys needed to be carefully considered (ISR 2012, 306-322, *)

Selection effects

And

Measurement effects

I will here consider only selection effects to some extent.

*Vannieuwenhuyze, Loosveldt and Molenbergs

As far as the MIXED-MODE SURVEYS are concerned, these can be a mixture of two data collection modes at minimum like
- Mail + F2F
- Web + Phone
- Web + Mail.

Our survey uses the latest strategy that is maybe the cheapest possible strategy but not necessarily best. Our overall response rate was just above 35% that is the same as our recent historical attitudes postal mail survey. Naturally, the web makes everything cheaper and easier to handel although the web response rate is not high. Our success with web was not excellent but it was NOT maybe well motivated. Next page

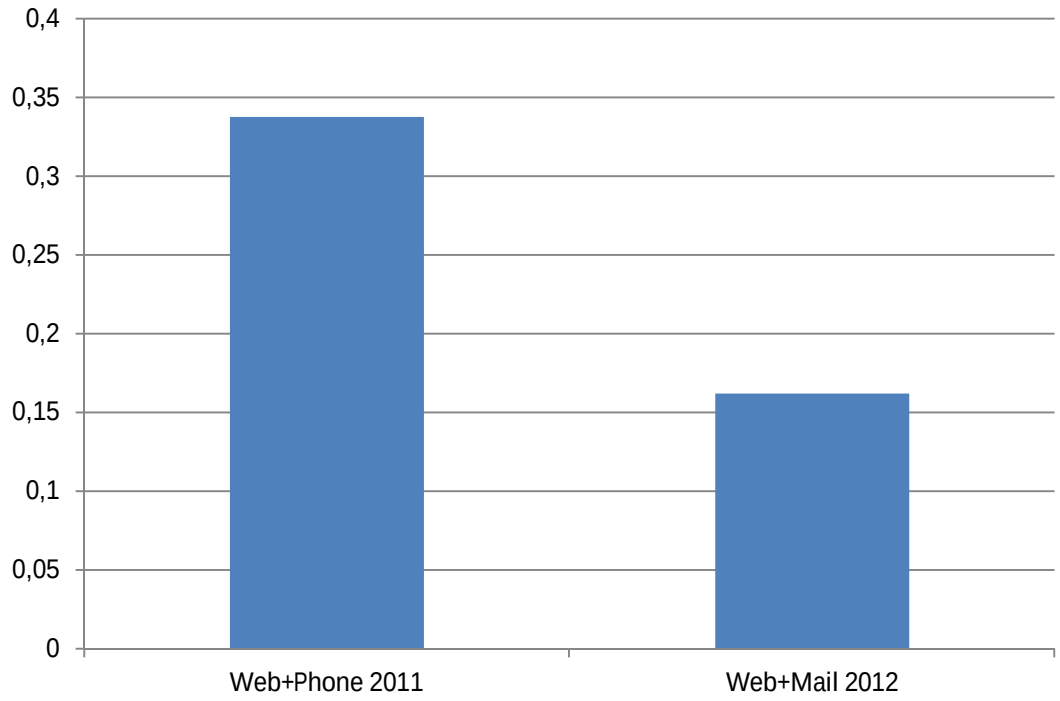
Proportion of the web responders in two recent mixed-mode surveys

Question: Is it easier to motivate web when the alternative is phone?

More details on next page

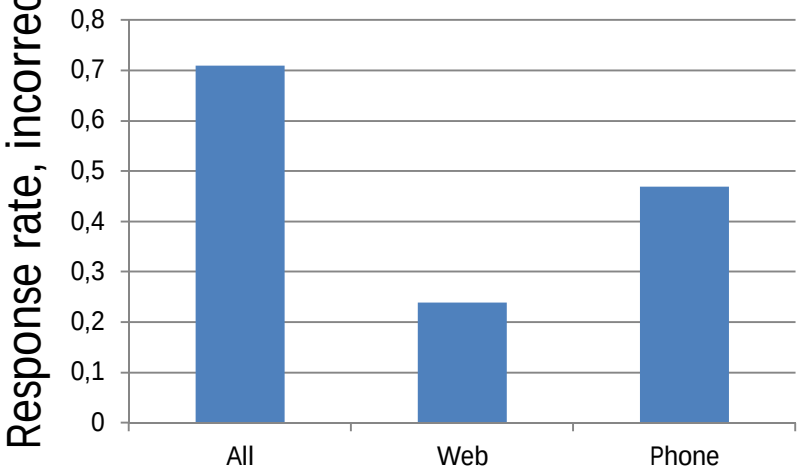
LEFT: the survey for Finnish consumers
Statistics Finland 2011

RIGHT: This survey

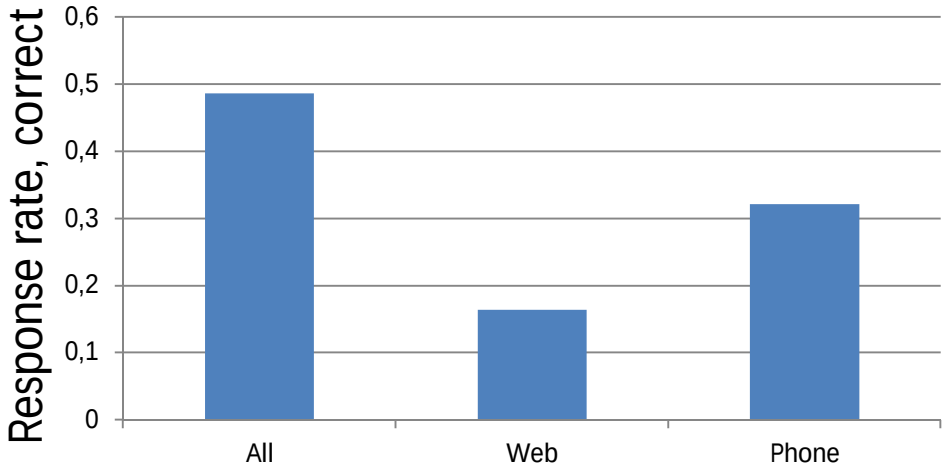


Statistics Finland Mixed-Mode Pilot for Consumer Barometer 2011

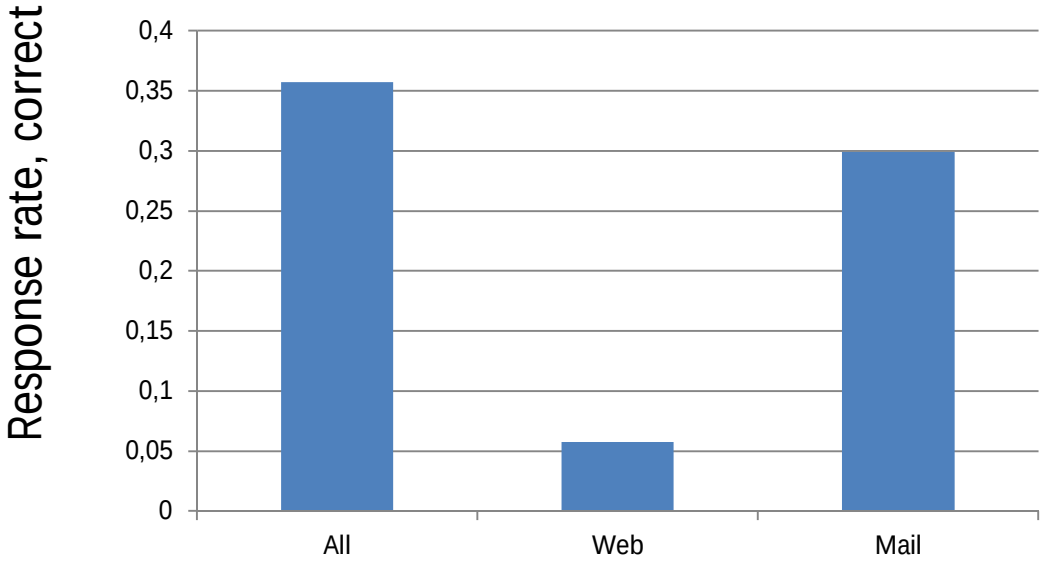
With Phone Numbers



With All



This Mixed-Mode Study 2012



This presentation continues with two issues

A. Sampling
Stratified sampling
-Explicit vs Implicit: both are possible but here explicit
Stratum type
- Administrative regions/areas
- More or less statistical, partially administrative (census areas)
- Geographical like GIS based

In this presentation both administrative and GIS based are applied that are then combined.

B. Response in this survey

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 applied 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.

Target population and sampling frame

The statistical units of the target population are 25-74 years old residents of 16 Finnish southern municipalities whose mother tongue is either Finnish or Swedish. The information is based on the January 2012 population register. Our sampling frame was 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 decided not to participate in the whole study completely.

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 information 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 (RED) vs. 'poor' people (BLUE) in the municipalities of the survey. The remaining grids are between those two ones or empty of people

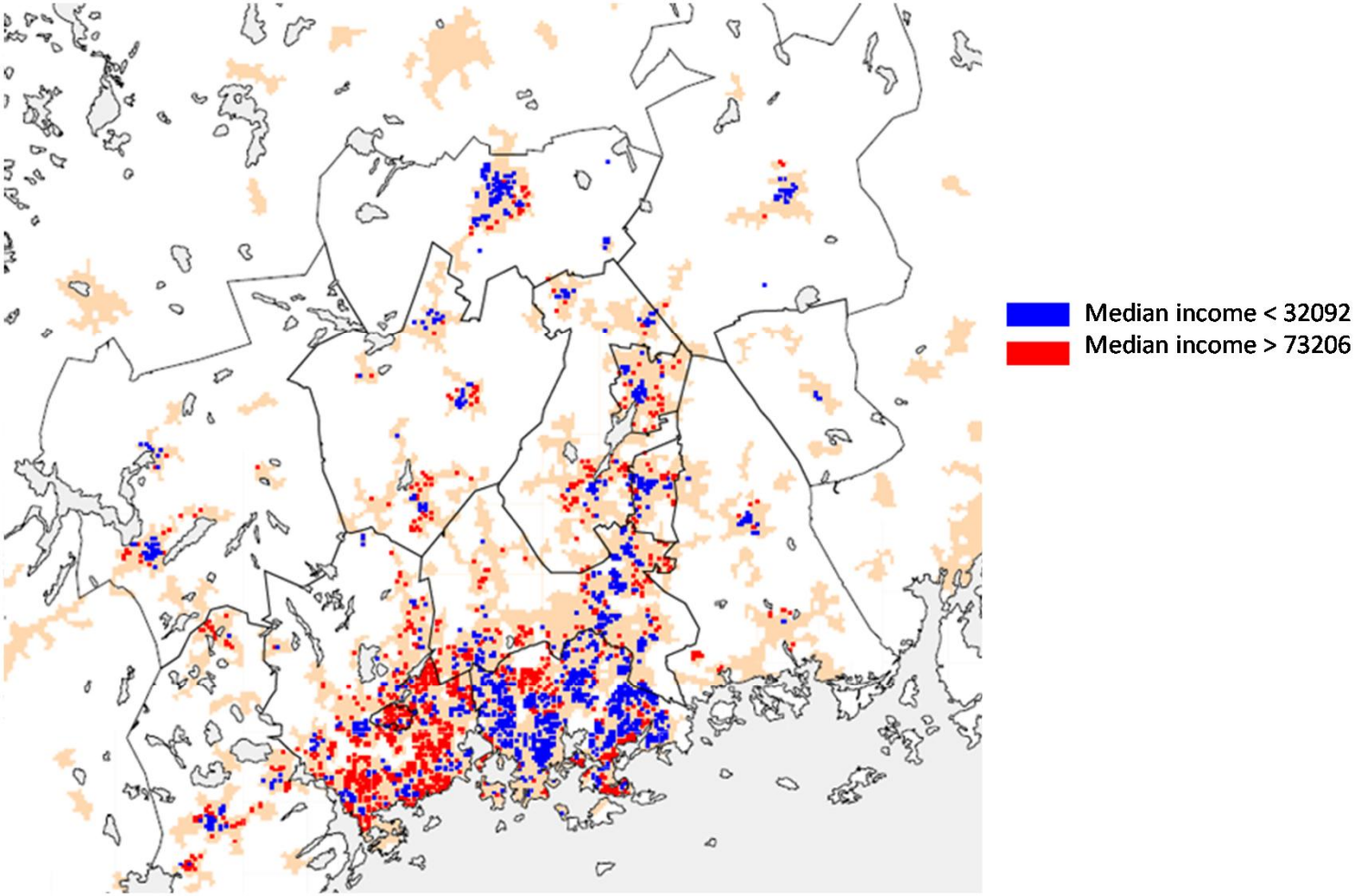


Table 1. Allocation of gross sample

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**

These two types of strata are overlapping, i.e. dependent on each other

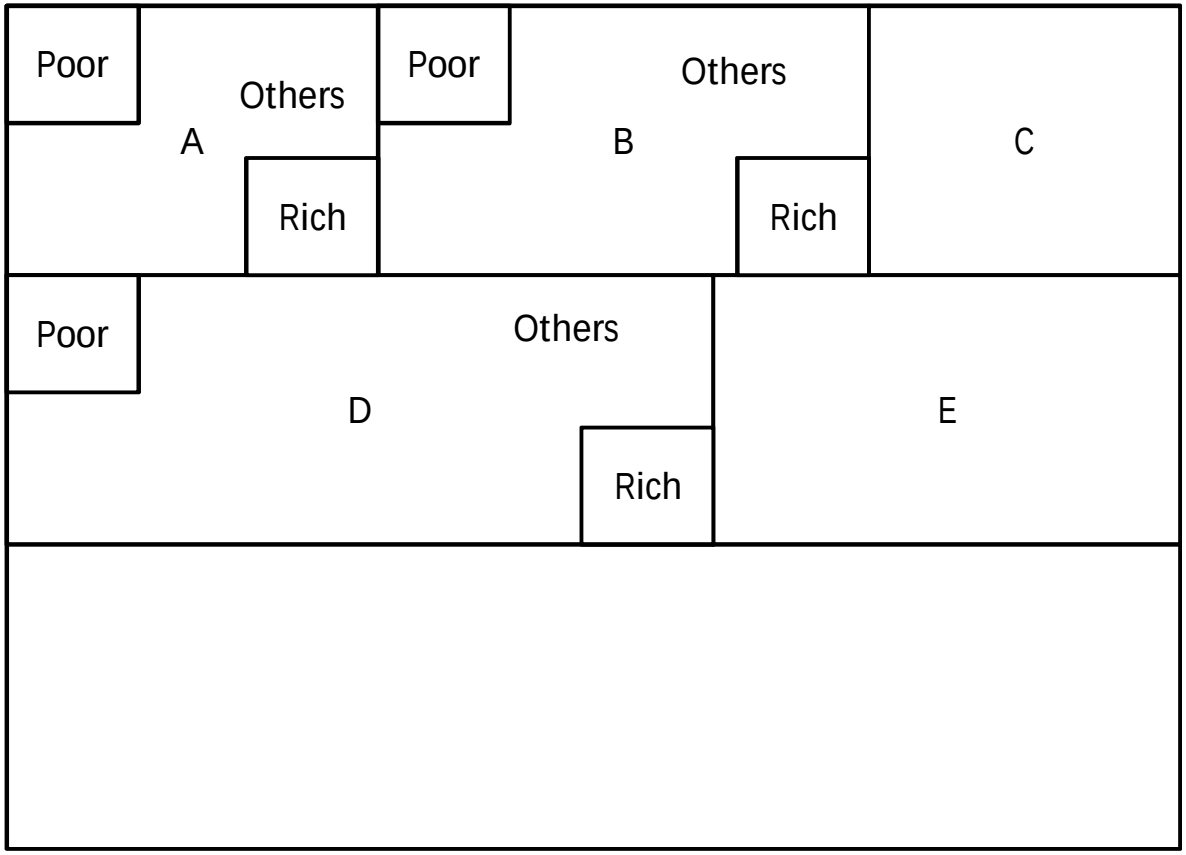
The inclusion probabilities are straightforwardly computable for Lahti and Lohja since any conditionality problem does not exist. They are as usually:

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

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

The remaining municipalities are more difficult. We look at an illustration on next page.

Municipality Strata A, B, C, ... and Grid-based strata within each of them



The graph is not ideal, since OTHERS includes all types of grids, thus Rich, Poor, Medium and Confidential

Note that it was not automatic to create these overlapping strata since this required to match together those two data sets by the equal postal zip codes, first.

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.

There are different approaches to solve that problem.

One strategy is presented in my written paper. It is workable but not a best possible one.

The second strategy is presented below:

We have to constitute the three different formulae, one for each of these groups. Our problem is that we have no information about all three populations at stratum level but only at the whole population level. Fortunately, we have been able to compute the gross sample sizes at stratum level. Hence our strategy is as follows:

- We assume that the each frame (poor, rich and others) is proportional to the gross sample size, and thus compute the frame population with this assumption for each stratum. Basically this is a valid assumption since sampling in each case is random within explicit strata.

For this reason our stratum populations are like *estimates*, and respectively we use the symbols with 'hat'. The numbers without hats are known

Frame populations of rich strata *h (for 16 municipalities)*:

$$\hat{N}_{rich,h} = N_{rich,h} \frac{n_{ruch,h}}{n_{rich}}$$

And similarly to poor strata

$$\hat{N}_{poor,h} = N_{poor,h} \frac{n_{poor,h}}{n_{poor}}$$

For the ‘others’ strata we have no the same information but we can compute these population figures as follows:

$$\hat{N}_{others,h} = N_h - \hat{N}_{rich,h} - \hat{N}_{poor,h}$$

Here N_h *are known and thus they are population figures for municipality strata (their sum in Table 2 = 699725).*

Now we can straightforwardly to compute the inclusion probabilities to each out of three population groups:

$$\pi_k = \frac{n_{rich,h}}{\hat{N}_{rich,h}} \qquad \pi_k = \frac{n_{poor,h}}{\hat{N}_{poor,h}} \qquad \pi_k = \frac{n_{others,h}}{\hat{N}_{others,h}}$$

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

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

After the fieldwork, when we thus know the numbers of the unit-respondents, symbolised by *r,* we will get the basic weights assuming that the response mechanism is ignorable within explicit strata. In this case, we replace the symbols '*n*' with the respective symbols '*r*'.

BUT there are more strategies still. The third one is as follows:

-First to constitute the ordinary inclusion probabilities to each explicit stratum, thus both for the grid part and for the municipality part.

- And then the sampling weights, respectively. The sum of the both sampling weights is too high, due to overlapping.

-These weights need to be benchmarked so that their overall sum is exactly = the municipality target population in each respective stratum. How to do this?

This benchmarking is done as follows:

- Compute the sum of all subgroup (Rich, Poor, Others) sampling weights for each explicit stratum.

-Compute the shares of each subgroup in each stratum like

$$q_{poor} = \frac{sum(poor)}{(sum(poor) + sum(rich) + sum(others)}$$

-And similarly to rich and others

-Next multiply the initial sampling weights with this share.

Next I will give some numbers from our data, they are just fresh and will be maybe revised. It is good to remember that these weights are not exact and constitute a small additional uncertainty to the results. These are more essential in small strata (in gross sample figures, see Table 2).
The second comment on these figures is also interesting:

-The sampling for municipalities was in fact conditional so that those who were included in the grid part sample were excluded from the municipality part. Unfortunately, we cannot take into account this question due to lack of data.
-Also, the whole family/dwelling unit was excluded at the same time, not only one person. This constitutes an interesting additional question as the next page table shows.

Dwelling size	Respon-dents	Nonres-pondents
1	0,222188	0,249971
2	**0,437097**	**0,338643**
3	0,149407	0,1613
4	0,13298	0,165095
5	0,044708	0,059862
6	0,009773	0,014549
7	0,002183	0,004715
8	0,000832	0,00184
9	0,00052	0,00138
10	0,000104	0,00046
11	0,000104	0,000345
12	0,000104	0,000518
14	0	0,000288
15	0	0,000173
17	0	5,75E-05
18	0	5,75E-05
19	0	0,000115
20	0	5,75E-05
21	0	0,000115
22	0	0,000115
30	0	5,75E-05
33	0	5,75E-05
34	0	5,75E-05
39	0	5,75E-05
48	0	0,000115

Maximum dwelling unit (DU) sizes are large; so they are not household sizes, especially for non-respondents.

Table 3. Some statistics of the gross sample design weights, strategy 2 and 3

Statistics	Strategy 2	Strategy 3
Observations	27000	27000
Mean	25.9	25.9
Total = sum	699725	699725
Minimum	13.1	13.1
Maximum	57.1	58.4
CV (%)	31.7	60.1

The following results are for strategy 2 that is however the best we have invented. If we would get the exact population figures for overlapping grids, the weights could be exact as well but these are afterwards hard to get.

Table 4. Some statistics of the gross/net sample design weights, strategy 2

Statistics	Gross	Net	Grid part Gross	Grid part Net	Municipality part Gross	Municipality part Net
Observations	27000	9628	12000	4387	15000	5231
Mean	25.9	72.8	24.4	66.7	27.1	77.8
Total = sum	699725	699725	292615	292615	407110	407110
Minimum	13.1	24.9	13.1	24.9	13.1	39.0
Maximum	57.1	167.8	37.2	151.4	57.1	167.8
CV (%)	31.7	37.0	20.5	29.1	36.4	39.9

Table 5. Some statistics gross/net sample design weights by grids Strategy 2

Statistics	Rich Gross	Rich Net	Poor Gross	Poor Net	Intermediate Gross	Inter-mediate Net
Observations	6994	2715	9576	3288	10438	3615
Mean	24.4	64.5	24.3	68.5	28.4	72.9
Total = sum	170856	170856	232493	232493	296595	296595
Minimum	13.1	24.9	13.1	37.1	13.1	39.0
Maximum	37.2	151.4	37.2	134.1	57.1	167.8
CV (%)	18.3	22.0	22.3	31.5	28.4	82.9

Response modeling

Our respondent data are thus partially available. I already gave the first sampling weights for the respondents. Such initial or base weights are easy to compute. This weighting is only the start for constructing good sampling weights. These require also to analyse non-response and to adjust for it. I have already started this by using the response propensity modelling first and then to calibrating the sums of the resulted weights into the sums of the gross sample weights. This is done at each stratum level so that overlapping strata are covered too (e.g. Laaksonen 2007, Survey methodology 2007, Laaksonen&Chambers 2006, Journal of Official Statistics).

Response modeling

The response propensity modeling is more advantageous if good auxiliary variables are available. Our pattern will not be 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. We are also getting useful information from the taxation register at individual level, and hopefully also from the employment register. (such as being unemployed).

Response modeling

So far, I have tested the system with currently available auxiliary variables. I used a probit link function and estimated thus response propensities with this model (logit link is more used but it is not necessarily best).

Next page shows the estimates by municipality strata.

The subsequent page, respectively, gives some other estimates, applied until now. As soon as we get tax and employment register data, these will be used.

One week ago I got the data on the respondents.
Some figures already above. Now more

By municipality (or strata): Surprise = Helsinki area rates are highest that is not usual in other surveys. Maybe they are more interested in the topic.

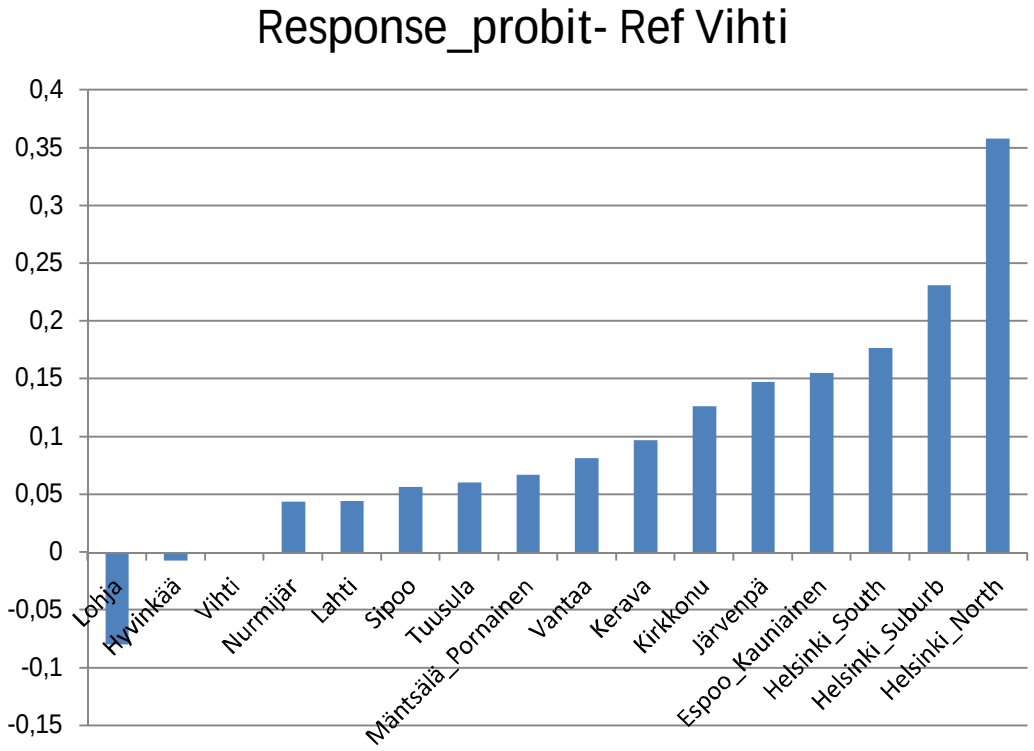


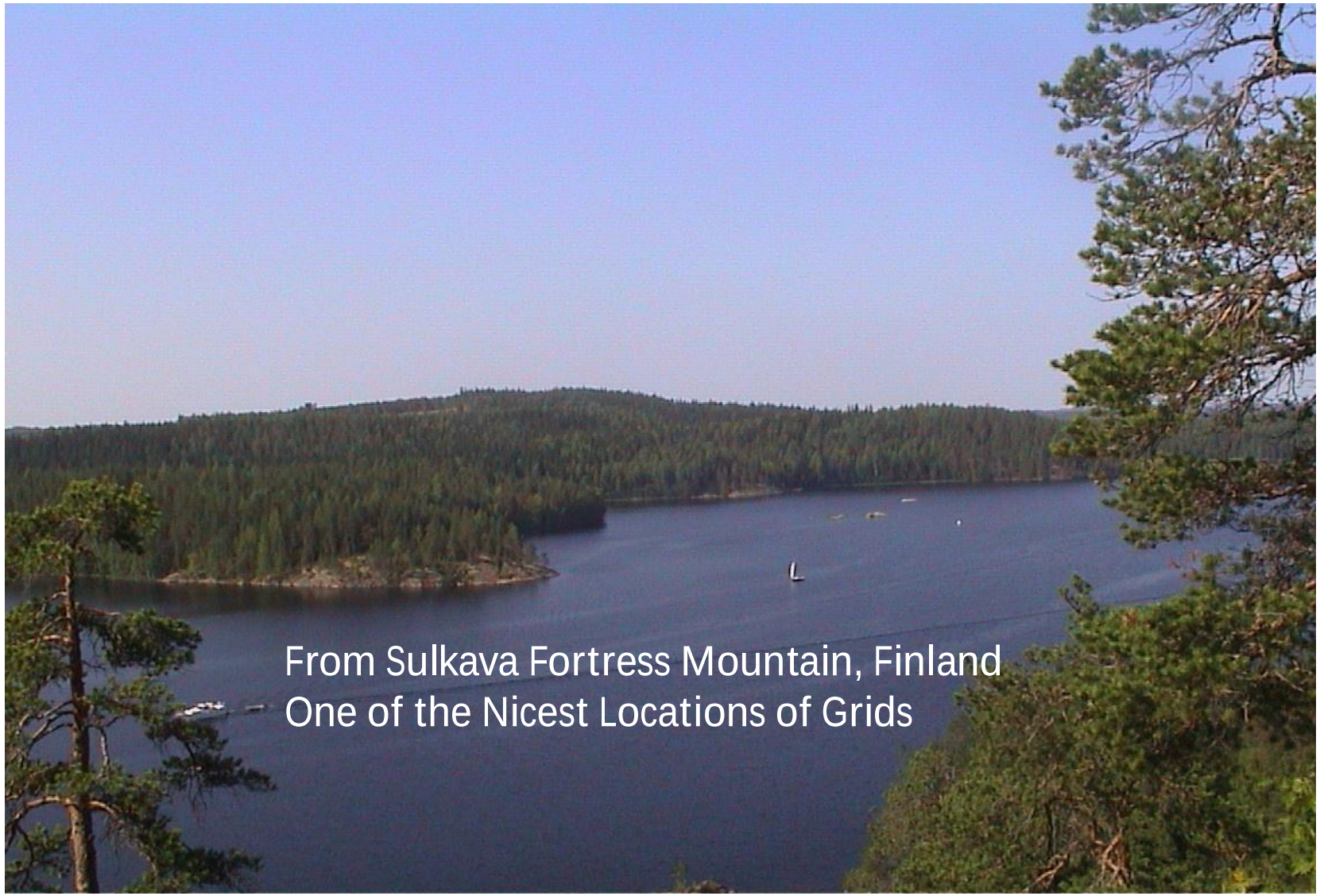
Table 6:

Results from my probit model

Auxiliary variable	Estimate	P-value
Rich grid	-0.061	<0.001
Poor grid	-0.15	<0.001
Intermediate grid	.	.
Males	-0.276	<0.001
Females	.	.
25-34 age old	-0.623	<0.001
35-44 age old	-0.569	<0.001
45-54 age old	-0.440	<0.001
55-64 age old	-0.183	<0.001
65+	.	.
Finnish speaking	0.008	0.228
Swedish speaking	.	.
One person DU	0.144	<0.001
2 persons DU	0.325	<0.001
3 persons	0.248	<0.001
4 persons	0.241	<0.001
5 persons	0.184	<0.001
6+ persons	.	.

Table 7. Finally, some comparisons between the two weights for the respondents. A quick comment: maybe workable but the maximums look quite big. Maybe good to collapse some overlapping strata. Variation with web vs mail respondents is not big.

Statistics	Basic weights	Adjusted weights with response propensity	Adjusted weights for mail respondents	Adjusted weights for web respondents
Observations	9628	9628	8054	1565
Mean	72.8	72.8	70.9	82.4
Total = sum	699725	699725	570900	128824
Minimum	24.9	18.1	19.0	23.9
Maximum	167.8	386.8	339.0	389.0
CV (%)	37.0	48.9	49.2	49.2


From Sulkava Fortress Mountain, Finland
One of the Nicest Locations of Grids

THANK YOU