Accuracy of Regional Estimators

Sylvie Rousseau

1. Introduction

We aim to estimate the accuracy of regional statistics from a household survey including additional regional sampling. We also compare the results to those that would have been obtained without this regional over-selection. This study gives the first assessment -concerning accuracy- of a new sampling system especially conceived by the French statistical institute (INSEE) for national surveys with regional extensions. Our analysis is performed on the Health survey which was carried out in 2002-03 over 25 000 dwellings and included an over-selection of dwellings originating from 5 regions out of 22. Our approach consists in modelling the sampling design, then explaining how we compute the variance with POULPE application, developed by INSEE with SAS software. We mainly focus on sampling variance, including correction of non-response and calibration. Finally, we evaluate the contribution of the new sampling process to the accuracy of regional estimators.

2. The Health Survey

2.1 Overview of the Health survey

The Heath Survey is an individual survey, carried out with face to face interviews. The interviewer visited everyone three times, at one month apart. There was one individual questionnaire per visit, and one questionnaire common to the household. Data collection lasted for one year in order to cover all kinds of diseases.

The main characteristic of this survey is that several regions got an additional sample thanks to local financing. For these 5 regions, local financing allowed to double the sample size in comparison with the size provided by the national budget. In this way, the sample contained 25 000 dwellings, instead of 18 000 dwellings that would have been selected with national financing only.

1 Statistical Methodology Unit, Department of Social Statistics, Institut National de la Statistique et des Etudes Economiques, 18 bd A. Pinard, 75675 Paris cedex 14, France, sylvie.rousseau@industrie.gouv.fr.

2 Champagne-Ardennes, Ile-de-France, Nord-Pas-de-Calais, Picardie and Provence-Alpes-Côte-d’Azur.
2.2 The sampling bases

The sampling system used in France for household surveys calls on sampling bases of dwellings. To get in touch with people for the Heath survey, a sample of dwellings was selected at first, and then, everybody living in these dwellings -as a main home- was interviewed. Moreover, the sampling bases are geographically concentrated in given areas in order to control the traveling costs, since most of the surveys are face to face.

Practically, the sample of the Heath survey was drawn from four sampling bases. Two are traditionally used for national surveys and are the only ones used for regions without sampling extension:

− The main one is the Master Sample (EM99) which has been extracted from the general population census of March 1999, by a stratified sampling process to one or two degrees, depending on the stratum (Wilms, 2000). It defines the areas, called “EM-areas”, where all national surveys are done.
− The second sampling base is a list of dwellings whose construction ended after 1999 and which are located in EM-areas. This subset of new dwellings ensures a proper follow-up of the population of dwellings through time. These dwellings come from administrative sources (building authorizations).

However, since the Master Sample has been built to ensure good features for national surveys and reasonable data collection expenses, EM-areas do not cover the regional territory widely enough to provide reliable regional statistics. Besides, the size of the regional fractions of the national sample prevents from obtaining a sufficient accuracy at the regional level. Therefore, a bigger regional sample has to be selected. The method used to select this sample requires two additional subsets of dwellings:

− The Master Sample for national surveys with regional extensions (EMEX). It has been extracted from the census and defines EMEX-areas.
− A subset of new dwellings, located in the same EMEX-areas.

Thus, in regions with sampling extensions, the sample of dwellings taken in the census was drawn from the joint set “EM99 plus EMEX”, which has balancing properties with respect to the reference regional population.

EMEX was an innovation introduced in 2001 to provide a sufficient accuracy for regional topics without exhausting EM99 reserve. Thanks to this device, regional and national levels can be simultaneously handled with according to a harmonized methodology and comparisons between the regional and the national levels are possible. Furthermore, additional regional data can be used for national studies as well (Christine, Wilms, 2003).

2.3 The sampling design

Whatever sampling bases may be used, the sampling design of dwellings is stratified by regions and degrees of urbanization and includes several degrees, distinguishing groups of municipalities or urban units as primary sampling units (PSU) and municipalities or groups of districts as secondary sampling units (SSU).
The sampling design of individuals adds a degree to the sampling design of dwellings. This degree is simply a census of people mainly living in the selected dwellings. It also adds a phase due to total non-response. The total non-response has been modeled by a probability mechanism of Poisson (for households as for people).

**Figure 1: the individual sampling design**

The sampling design of dwellings can be described in the simplified following way which has been used to compute the accuracy:

- **Strata** are a crossing between regions and the presence of the dwelling in the census.
- New dwellings are systematically drawn in each region.
- For dwellings taken in the census of 1999, the design depends on the degree of urbanization and the presence of regional extensions:
  - For rural areas, small urban units and middle-sized urban units in regions without extension, the primary sampling units are selected proportionally to their size (in main houses). Then,
    - In rural areas, groups of municipalities are chosen proportionally to their size too. And then dwellings are selected by a systematic sampling.
    - For small urban units and middle-sized urban units in regions without extension, dwellings are directly and systematically selected in the primary sampling units.
For middle-sized urban units in regions with extension and big urban units, the first degree consists of a census of urban units. Then,
- In middle-sized units, the design is the same than the one performed in regions without extension.
- In big urban units, the second degree is a sampling of municipalities from which dwellings are systematically drawn.

At the last stage, everyone living in selected dwellings as a main home is interviewed.

We selected a sample of 25 021 dwellings in that way. Then, data collection revealed 21 655 main houses inhabited by about 45 500 people. At the end, we got nearly 40 000 respondents at the first visit and about 35 000 at the last one, like pictured in figure 2.

Figure 2 : the population interviewed and the total non-response behavior

<table>
<thead>
<tr>
<th>Population interviewed</th>
<th>Respondents Visit n°1</th>
<th>Respondents Visit n°2</th>
<th>Respondents Visit n°3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visit n°1</td>
<td>25 021 dwellings sampled</td>
<td>16 848 households = 40 865 people including 39 901 respondents</td>
<td>36 951 respondents</td>
</tr>
</tbody>
</table>

2.4 The sampling design under national financing only

In order to compare the accuracy of the survey with and without extensions, we tried to rebuild a fictitious sample we could have had without regional extensions. In this case, the sampling design would have been easier with one branch less due to middle-size urban units (figure 3). Indeed, these units would have been selected everywhere proportionally to their size, exactly like small urban units.

Practically, we selected a fictitious “national” sample from the effective one, both with respect to the national sampling design and the national budget. Thus, we drew a subsample of 18 000 dwellings according to the following method:
- First, we calculated the number of dwellings that would have been selected in every EM-areas according to the national sampling design.
- Then, we stratified the real sample into two groups to distinguish EM-areas and EMEX-areas, before drawing dwellings from the effective sample at equal probability with respect to the allocations calculated before.
At the end, we calculated the final weights, after correction of total non-response and calibration on auxiliary data. We used the same methods as for the effective survey except we give up the regional dimensions, there was no reason to favour any more (Caron, Rousseau, 2005).

Figure 3: the individual sampling design under national financing only

Finally, we got a sample of about 18 000 dwellings which suffers from the same non-response rates than the effective sample: it led us to about 28 800 individuals respondents at the first visit and about 25 000 at the last one, as shown in figure 4.

Figure 4: Structure of the “national” sample

- Population interviewed 1st visit
  17 974 dwellings
  15 480 households in the scope of the survey = 32 693 people

- Respondents 1st visit
  12 212 households = 29 425 people including 28 743 respondents

- Respondents 3rd visit
  25 377 respondents
3. Variance estimation

3.1 POULPE software

POULPE is a SAS macro-based application conceived by INSEE (Caron, Deville, Sautory, 1998). It computes the accuracy of simple estimators or complex functions with respect to many sampling designs. It also computes the accuracy of calibrated estimators and takes total non-response into account. It especially covers the current case of a several degrees and two phases sampling design, where the second phase is a Poisson sampling. It also provides the lower and upper bounds of the 95% confidence interval, assuming the estimator to be asymptotically normally distributed. Moreover, it computes the design effect which is the ratio of the variance obtained under the current sample design by the variance of a simple random sampling (without replacement) with the same size. Nevertheless, neither partial non-response, nor errors of measure are taken into account, which leads to over-estimate the accuracy.

3.2 Theoretical foundations

3.2.1 Variance estimation for a several degrees sampling

POULPE uses Raj's formula for a several degrees sampling: it calculates the variance due to each degree, one after the other. This recursion makes reckonings easier. Thus, the variance of $\hat{Y}$, Horvitz-Thompson estimator of a real total $Y$, is estimated without bias by $\hat{V}(\hat{Y}) = f(\hat{Y}_i | i \in s) + \sum_{i \in s} \omega_i \hat{V}_i$ where:

- $f(Y_i | i \in s) = \hat{V}_i \left( \sum_{i \in s} \omega_i Y_i \right)$ estimates the variance due to the 1st degree,
- $\hat{Y}_i$ estimates the sub-total $Y_i$ in the $i$-th primary unit,
- $\hat{V}_i = \hat{V}_{2|1}(\hat{Y}_i)$ estimates the variance due to the 2nd degree in the $i$-th primary unit,
- $\omega_i = \frac{1}{\pi_i}$ denotes the sampling weight of the $i$-th primary unit.

3.2.2 Variance estimation for unequal probability sampling

For unequal probability sampling, POULPE uses Deville’s formula, that is valid under a fixed-sized sampling design with large entropy. This is the formula used to calculate the variance due to the drawing of primary sampling units in rural areas for example:

$$\hat{V}(\hat{Y}) = \frac{n}{n-1} \sum_{k \in s} (1-\pi_k) \left( \frac{Y_k}{\pi_k} - \sum_{i \in s} a_i(s) \frac{Y_i}{\pi_i} \right)^2$$

where $a_i(s) = \frac{1-\pi_i}{\sum_{j \in s} (1-\pi_j)}$ and $\pi_i$ is the 1st order inclusion probability of $i$-th unit.
3.2.3 Variance estimation for a systematic sampling

For a systematic sampling, like the drawing of dwellings, POULPE uses the following Deville’s formula, which is valid assuming that data are in the same order than in the sampling base: \( \hat{V}(\hat{Y}) = N^2 \left(1 - \frac{n}{N}\right) \frac{t^2}{n} \) where \( t^2 = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} (y_i - y_{i-1})^2 \) and \( y_i \) denotes the value associated to \( i \)-th individual.

3.2.4 Variance estimation for a second phase of Poisson

For a sampling design including a second phase of Poisson, the variance is broken down into 2 terms, one by phase. Thanks to the mechanism of Poisson, the part due to the second phase only requires the probability of response, denoted by “\( p_i \)” for the \( i \)-th individual. Another advantage is that this formula can be used with only one respondent per household:

\[
\hat{V}(\hat{Y}) = \hat{V}_{\text{1st phase}}(\hat{Y}) + \sum_{i \in s} \left(1 - \frac{p_i}{\pi_i}\right) \left(\frac{y_i}{\pi_i}\right)^2 \quad \text{where} \quad \hat{Y} = \sum_{i \in s} \frac{y_i}{\pi_i} \text{ estimates } Y \text{ without bias.}
\]

3.2.5 Variance estimation for complex statistics

For non-linear statistics, like ratio for example, POULPE implements the linearization method. It consists of getting by derivation a linear statistic which has the same asymptotic variance than the complex parameter (Deville, 1999). For example, \( \hat{R} = \frac{\hat{Y}}{\hat{X}} \) estimates \( R = \frac{Y}{X} \) and its variance can be evaluated by:

\[
\hat{V}(\hat{R}) = \sum_{i \in s} \sum_{j \in s} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \left(\hat{z}_i - \hat{z}_j\right) \quad \text{where} \quad \hat{z}_i = \frac{1}{\hat{X}} \left(y_i - \hat{R}x_i\right).
\]

3.2.5 Variance estimation for calibrated estimators

For calibrated estimators, POULPE replaces the interest variable \( y \) with the estimated residuals \( \hat{u} \) from the regression, performed in the sample, of \( y \) on calibration variables. Thus it calculates: \( \hat{V}(\hat{Y}) = \hat{V} \left[ \sum_{i \in s} \frac{\hat{u}_i}{\pi_i} \right] \) (Deville, Särndal, Sautory, 1993).

3.3 Implementation

POULPE requires 3 files:
- The “tree diagram” describing the sampling design (figures 1 or 3),
- The geographical file giving the size of the stratums and of each sampling unit that intervenes in the sampling process,
- The individual data set, especially including the final weights and the probability of response associated to 2nd phase.

We implemented POULPE on two data sets:
- the real sample including 45 672 people in 1st phase,
- the national “fictitious” sample listing 32 693 individuals in 1st phase.
4. Results

4.1 Overview of a few results

The table 1 below lists by descending order the regions having the best accuracy for the proportion of people feeling healthy. Naturally, the best precision is obtained at the national level. Then come the regions with sampling extensions (written in blue) from the biggest to the smallest. There is an exception with Rhône-Alpes where the accuracy is quite good although there was no extension there: this region is one of the most populated in France and accordingly already received a large part of the sample size. When comparing Auvergne, a region without extension, to Champagne-Ardennes or Picardie that practised one, we can conclude that the presence of extension significantly improve the regional accuracy, as motivated at the creation of EMEX.

Table 1: Proportion of people feeling healthy

<table>
<thead>
<tr>
<th>Region</th>
<th>Estimator</th>
<th>95% Confidence interval (1st visit)</th>
<th>Coefficient of Variation (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>France</td>
<td>0.685</td>
<td>[0.68 - 0.69]</td>
<td>0.4</td>
</tr>
<tr>
<td>Ile-de-France</td>
<td>0.725</td>
<td>[0.71 - 0.74]</td>
<td>0.8</td>
</tr>
<tr>
<td>Provence-Alpes-Côte-d'Azur</td>
<td>0.65</td>
<td>[0.63 - 0.67]</td>
<td>1.4</td>
</tr>
<tr>
<td>Nord-Pas-de-Calais</td>
<td>0.66</td>
<td>[0.64 - 0.68]</td>
<td>1.5</td>
</tr>
<tr>
<td>Rhône-Alpes</td>
<td>0.7</td>
<td>[0.68 - 0.72]</td>
<td>1.6</td>
</tr>
<tr>
<td>Champagne-Ardennes</td>
<td>0.695</td>
<td>[0.67 - 0.72]</td>
<td>1.7</td>
</tr>
<tr>
<td>Picardie</td>
<td>0.69</td>
<td>[0.66 - 0.72]</td>
<td>2.3</td>
</tr>
<tr>
<td>Auvergne</td>
<td>0.63</td>
<td>[0.58 - 0.69]</td>
<td>4.5</td>
</tr>
</tbody>
</table>

4.2 The influence of the sample size

The main factor impacting the variance is obviously the sample size. The influence is O(1/n) where n denotes the effective sample size. This correlation trend appears while comparing the results between regions with and without extensions, or between the 1st and the 3rd visit, or also between the real sample and the “national” one (table 2).

4.3 The influence of calibration

Calibration is another factor that can improve the accuracy, especially for regional estimators, given the calibration dimensions are well correlated to the interest variables. Basically, increasing the number of calibration variables increases the accuracy of estimators of totals like pictured in table 3. However, the effect is limited on ratios.

Table 3: Coefficient of variation of the estimated number of people feeling healthy (%)

<table>
<thead>
<tr>
<th>Region</th>
<th>“Maximal” calibration</th>
<th>Without any regional dimensions</th>
<th>Without calibration</th>
</tr>
</thead>
<tbody>
<tr>
<td>France</td>
<td>0.4</td>
<td>0.4</td>
<td>0.8</td>
</tr>
<tr>
<td>Ile-de-France</td>
<td>0.9</td>
<td>1</td>
<td>1.4</td>
</tr>
</tbody>
</table>
Table 2: Number of G.P. consultations per year

<table>
<thead>
<tr>
<th>Coverage</th>
<th>Parameter</th>
<th>France</th>
<th>Champagne-Ardenne</th>
<th>Haute-Normandie</th>
</tr>
</thead>
<tbody>
<tr>
<td>After the 1st visit</td>
<td>Number of respondents</td>
<td>39 901</td>
<td>2 495</td>
<td>702</td>
</tr>
<tr>
<td></td>
<td>$\hat{Y}_i = \text{total number of G.P. consultations}$</td>
<td>232 700 000</td>
<td>6 030 000</td>
<td>5 820 000</td>
</tr>
<tr>
<td></td>
<td>$CV(\hat{Y}_i)$</td>
<td>0.8%</td>
<td>2.6%</td>
<td>17.4%</td>
</tr>
<tr>
<td>Real sample</td>
<td>Average number of G.P. consultations per person</td>
<td>4</td>
<td>4.4</td>
<td>4.1</td>
</tr>
<tr>
<td></td>
<td>$CV(\hat{\mu}_i)$</td>
<td>0.8%</td>
<td>2.6%</td>
<td>4.2%</td>
</tr>
<tr>
<td>After the 3rd visit</td>
<td>Number of respondents</td>
<td>35 073</td>
<td>2 281</td>
<td></td>
</tr>
<tr>
<td>Real sample</td>
<td>$CV(\hat{\mu}_3)$ where $\hat{\mu}_3$ denotes the average number of G.P. consultations per person</td>
<td>0.8%</td>
<td>2.9%</td>
<td></td>
</tr>
<tr>
<td>After the 1st visit</td>
<td>Number of respondents</td>
<td>28 743</td>
<td>779</td>
<td></td>
</tr>
<tr>
<td>“Fictitious” national sample</td>
<td>$CV(\hat{\mu}<em>{nat})$ where $\hat{\mu}</em>{nat}$ denotes the average number of G.P. consultations per person</td>
<td>1%</td>
<td>4.3%</td>
<td></td>
</tr>
</tbody>
</table>

4.4 The influence of the design effect

Another asset of EMEX is to reduce the design effect in regions with extensions (table 4). In fact, the system increases the number of primary sampling units so that the sampling process gets closer to a stratified sampling design. For instance, in regions with extension, the real sample contains about 50% more primary sampling units than the national one.

Table 4: Proportion of people feeling healthy

<table>
<thead>
<tr>
<th>Region</th>
<th>Estimator</th>
<th>1st Design effect</th>
<th>1st Design effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ile-de-France</td>
<td>0.725</td>
<td>1.03</td>
<td>1.21</td>
</tr>
<tr>
<td>Champagne-Ardenne</td>
<td>0.695</td>
<td>0.51</td>
<td>0.77</td>
</tr>
<tr>
<td>Picardie</td>
<td>0.69</td>
<td>1.14</td>
<td>1.55</td>
</tr>
<tr>
<td>Nord-pas-de-Calais</td>
<td>0.66</td>
<td>0.94</td>
<td>1.16</td>
</tr>
<tr>
<td>Provence-Alpes-Côte-d'Azur</td>
<td>0.65</td>
<td>0.91</td>
<td>1.23</td>
</tr>
<tr>
<td>Bourgogne</td>
<td>0.66</td>
<td>2.4</td>
<td></td>
</tr>
</tbody>
</table>
5. Conclusion

Our study shows that the sampling process based on EM99 and EMEX provides a proper accuracy for regional parameters stemmed from the Health survey. In this way, thanks to additional regional sampling, reliable regional analysis can be performed. Moreover, since sampling method is harmonized, national accuracy also gets better in the same time and comparisons between regions become possible. This accuracy gain is especially of interest when comparing different subpopulations or domains. Thus EMEX gives an answer to regional needs. These regional concerns are currently taken into account in the building of the next sampling device for household surveys, which is revised within the framework of the “rotating” population census settled in 2004 in France.

References


Accuracy of regional estimators

Sylvie Rousseau
INSEE, France
Motivations

› Compute the accuracy of different estimators for a survey including regional topics
› Compare the results to those that would have been obtained without additional regional sampling

Application to the French Health Survey carried out in 2002-03
Use of POULPE software
Summary

I. The Health Survey of 2002-03
II. Computation of accuracy
III. Results
I. The Health Survey of 2002-03
The Health Survey of 2003-03

› An individual survey
› Face-to-face interviews
   – 3 visits, once a month
   – 3 individual questionnaires + 1 for the household
› 5 periods for data collection

› 5 regions got an additional sample
   – Local financing doubled the sample size - at least
› A sample of 25 021 dwellings
The sampling design of dwellings

› A sample of dwellings where everyone is polled
› 4 sampling bases
  – The 2 traditional ones in regions without extension
    - The Master Sample, called EM99, extracted from the census of 1999
    - The sampling base for new dwellings
  – Completed by 2 subsets in regions with extension
    - The Master Sample for regional EXTensions of nationals surveys (EMEX)
    - A subset of new dwellings, geographically limited to « EMEX-areas »

› The sampling design of dwellings is
  – Stratified by regions × degrees of urbanisation
  – At several degrees
    - Primary Units = groups of communities or urban units (UU)
    - Secondary Units = communities
The sampling design of individuals

› An additional degree
  – Census of people living in the household

› A second phase due to total non-response
  – Total non-response is modelled by a probability mechanism of Poisson

› The sampling design of individuals
  – 2 phases
  – Several degrees
  – Stratified by regions \(\times\) degrees of urbanisation
The population interviewed and the total non-response behavior

**Population interviewed 1\(^{st}\) visit**
- 25 021 dwellings
  - 21 655 households in the scope of the survey
  - = 45 672 people

**Respondents 1\(^{st}\) visit**
- 16 848 households
  - = 40 865 people including 39 901 respondents

**Respondents 3\(^{rd}\) visit**
- 35 073 respondents
Schema of the individual sampling design

- Rural areas
- Small urban units
- Middle-sized urban units without extension

Stratum
Région ×1_dwelling taken in the census
- Middle-sized urban units in regions with extension
- Big urban units

New dwellings

PPS

Urban Unit (UU)

S.S.U.
Communities

SYS

Dwellings

EXH

Individuals

PPS

Rural areas

- Small urban units
- Middle-sized urban units without extension

PPS

P.S.U.
Group of communities or UU

SYS

Dwellings

EXH

Individuals

SYS

S.S.U.
Communities

SYS

Dwellings

EXH

Individuals

SYS

S.S.U.
Communities

SYS

Dwellings

EXH

Individuals

SYS

New dwellings

- Small urban units
- Middle-sized urban units in regions without extension

- Middle-sized urban units in regions with extension
- Big urban units

Individuals
Schema of the individual sampling design under national financing only

Stratum
Région \times 1_{\text{dwelling taken in the census}}

- Rural areas
- Small urban units
- Middle-sized urban units

Big urban units

PPS

EXH

P.S.U. Group of communities or UU

Rural areas

PPS

EXH

S.S.U. Communities

- Small urban units
- Middle-sized urban units

EXH

Dwellings

SYSTEXH

Individuals

New dwellings
The building of a « national » sample, only national financed

› A sample of about 18 000 dwellings without regional financing
› Firstly, reckoning of the number of dwellings that would have been selected in every traditional « EM-area » with respect to the sampling design
› Then, sampling of dwellings at equal probability among the real sample of 25 021 dwellings
› Finally, correction of total non-response and calibration on auxiliary data, as for the real survey, except for regional dimensions
Structure of this « national » sample

**Population interviewed**
1st visit
- 17,974 dwellings
- 15,480 households in the scope of the survey
  = 32,693 people

**Respondents**
1st visit
- 12,212 households
  = 29,425 people
  including 28,743 respondents

**Respondents**
3rd visit
- 25,377 respondents
II. Computation of accuracy
Poulpe software

› Conceived at Insee with SAS software
› Computes accuracy
  – For many sampling designs
    Application here to several degrees and 2 phases where the 2nd one a Poisson sampling
  – For simple estimators or complex functions
  – Takes calibration into account
  – Computes the design effect
  – Does not include influence of partial non-response nor errors of measure
Variance calculation for a 2-degrees sampling

\[ \hat{Y} = \text{Horvitz-Thompson estimator of real total } Y \]

\[ \hat{V}(\hat{Y}) = f(\hat{Y}_i | i \in s) + \sum_{i \in s} \omega_i \hat{V}_i \]

estimates its variance without bias [Raj]

\[ f(Y_i | i \in s) = \hat{V}_1 \left( \sum_{i \in s} \omega_i Y_i \right) \]

estimates the variance due to the 1st degree with \( \hat{Y}_i \) estimator of total \( Y_i \)

\[ \hat{V}_i = \hat{V}_{2|1}(\hat{Y}_i) \]

estimates the variance due to the 2nd degree in the \( i \)-th primary unit

\[ \omega_i = \frac{1}{\pi_i} \]

sampling weight of the \( i \)-th primary unit
Variance calculation with unequal probability sampling

\[ \hat{Y} = \text{Horvitz-Thompson estimator of real total } Y \]

\[ \hat{V}(\hat{Y}) = \frac{n}{n-1} \sum_{k \in s} (1 - \pi_k) \left( \frac{y_k}{\pi_k} - \sum_{i \in s} a_i(s) \cdot \frac{y_i}{\pi_i} \right)^2 \]

estimates its variance [Deville] under a fixed-sized sampling design with large entropy

- \( \pi_k = 1^{\text{st}} \text{ order inclusion probability of } k\text{-th individual} \)
- \( a_i(s) = \frac{1 - \pi_i}{\sum_{j \in s} (1 - \pi_j)} \)
Variance calculation for a systematic sampling

- The variance is estimated by

\[ \hat{V}(\hat{Y}) = N^2 \left( 1 - \frac{n}{N} \right) \frac{t^2}{n} \]

\[ t^2 = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} (y_i - y_{i-1})^2 \]

- \(y_i\) is the value associated to \(i\)-th individual

- Assuming that data are in the same order than in the sampling base [1]
Variance calculation for a 2\textsuperscript{nd} phase of Poisson \cite{1}

\[ \hat{Y} = \sum_{i \in s} \frac{y_i}{\pi_i p_i} \]

- $\pi_i$ = inclusion probability of $i$-th unit in 1\textsuperscript{st} phase
- $p_i$ = probability of response of $i$-th unit

\[ \hat{V}(\hat{Y}) = \hat{V}_{1^{st} \text{ phase}}(\hat{Y}) + \sum_{i \in s_2} \frac{1 - p_i}{p_i^2} \left( \frac{y_i}{\pi_i} \right)^2 \]
Variance calculation for complex statistics and calibrated estimators [1]

› Application of the linearization method

Ex.: $\hat{R} = \hat{Y} / \hat{X}$ estimates $R = Y / X$

$$
\hat{V}(\hat{R}) = \sum_{i \in s} \sum_{i \in s} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{\hat{z}_i}{\pi_i} \frac{\hat{z}_j}{\pi_j}
$$

$$
\hat{z}_i = \frac{1}{\hat{X}} \left( y_i - \hat{R} x_i \right)
$$

› For calibrated estimators

$$
\hat{V}(\hat{Y}) = V \left( \sum_{i \in s} \frac{\hat{u}_i}{\pi_i} \right)
$$

$\hat{u}_i$ is the estimated residual of the regression, in the sample, of $Y$ on calibration variables
III. Results
A few results

Proportion of people feeling healthy

<table>
<thead>
<tr>
<th>Region</th>
<th>Estimator</th>
<th>IC_{1.95%}</th>
<th>CV(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>France</td>
<td>0.685</td>
<td>[0.68 – 0.69]</td>
<td>0.4</td>
</tr>
<tr>
<td>Île de France</td>
<td>0.725</td>
<td>[0.71 – 0.74]</td>
<td>0.8</td>
</tr>
<tr>
<td>PACA</td>
<td>0.65</td>
<td>[0.63 – 0.67]</td>
<td>1.4</td>
</tr>
<tr>
<td>Nord</td>
<td>0.66</td>
<td>[0.64 – 0.68]</td>
<td>1.5</td>
</tr>
<tr>
<td>Rhône-Alpes</td>
<td>0.70</td>
<td>[0.68 – 0.72]</td>
<td>1.6</td>
</tr>
<tr>
<td>Champagne</td>
<td>0.695</td>
<td>[0.67 – 0.72]</td>
<td>1.7</td>
</tr>
<tr>
<td>Picardie</td>
<td>0.69</td>
<td>[0.66 – 0.72]</td>
<td>2.3</td>
</tr>
<tr>
<td>Auvergne</td>
<td>0.63</td>
<td>[0.58 – 0.69]</td>
<td>4.5</td>
</tr>
</tbody>
</table>

Real sample, 1st visit
The influence of the sample size

Comparing regions with or without extension, results of 1\textsuperscript{st} and 3\textsuperscript{rd} visit, « national » and real samples

– for « Number G.P’. consultations per year »

<table>
<thead>
<tr>
<th></th>
<th>France</th>
<th>Champagne</th>
<th>Hte-Normandie</th>
</tr>
</thead>
<tbody>
<tr>
<td>(r_1)</td>
<td>39 901</td>
<td>2 495</td>
<td>702</td>
</tr>
<tr>
<td>(\hat{Y}_1)</td>
<td>232 700 000</td>
<td>6 030 000</td>
<td>5 820 000</td>
</tr>
<tr>
<td>(CV(\hat{Y}_1))</td>
<td>0,8%</td>
<td>2,6 %</td>
<td>17,4%</td>
</tr>
<tr>
<td>(\hat{\mu}_{Y_1})</td>
<td>4</td>
<td>4,4</td>
<td>4,1</td>
</tr>
<tr>
<td>(CV(\hat{\mu}_{Y_1}))</td>
<td>0,8%</td>
<td>2,6%</td>
<td>4,2%</td>
</tr>
<tr>
<td>(r_3)</td>
<td>35 073</td>
<td>2 281</td>
<td></td>
</tr>
<tr>
<td>(CV(\hat{\mu}_{Y_3}))</td>
<td>0,8%</td>
<td>2,9%</td>
<td></td>
</tr>
<tr>
<td>(r_{nat})</td>
<td>28 743</td>
<td>779</td>
<td></td>
</tr>
<tr>
<td>(CV(\hat{\mu}_Y))</td>
<td>1%</td>
<td>4,3 %</td>
<td></td>
</tr>
</tbody>
</table>
The influence of calibration

› Increasing the number of calibration variables improves the accuracy for estimators of totals

› The effect is limited on ratios

– Ex.: Number of people feeling healthy

<table>
<thead>
<tr>
<th></th>
<th>CV(\hat{N}_1) in %</th>
<th>“Maximal” calibration</th>
<th>Without regional dimensions</th>
<th>Without calibration</th>
</tr>
</thead>
<tbody>
<tr>
<td>France</td>
<td>0,4</td>
<td>0,4</td>
<td>0,8</td>
<td></td>
</tr>
<tr>
<td>Ile-de-France</td>
<td>0,9</td>
<td>1</td>
<td>1,4</td>
<td></td>
</tr>
</tbody>
</table>

Visit 1, real sample
The design effect

Design effect is reduced in regions with extensions

- Ex: proportion of people feeling healthy

<table>
<thead>
<tr>
<th>Région</th>
<th>( \hat{p}_1 )</th>
<th>Deff(_{V1} )</th>
<th>Deff(_{1, nat.} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ile-de-France</td>
<td>0,725</td>
<td>1,03</td>
<td>1,21</td>
</tr>
<tr>
<td>Champagne</td>
<td>0,695</td>
<td>0,51</td>
<td>0,77</td>
</tr>
<tr>
<td>Picardie</td>
<td>0,69</td>
<td>1,14</td>
<td>1,55</td>
</tr>
<tr>
<td>Nord</td>
<td>0,66</td>
<td>0,94</td>
<td>1,16</td>
</tr>
<tr>
<td>PACA</td>
<td>0,65</td>
<td>0,91</td>
<td>1,23</td>
</tr>
<tr>
<td>Bourgogne</td>
<td>0,66</td>
<td>2,4</td>
<td></td>
</tr>
</tbody>
</table>

*Real and « national » samples, 1st visit*
Conclusion

› EMEX allows a better accuracy for regional parameters
  – especially in domains
› As an harmonized process, the additional regional sampling increases the national accuracy in same time.
› Next system of households surveys sampling, based on the census, takes regional concerns into account.
Bibliography

How to Counteract Problems of Quality in Household Surveys Caused by Intensive Migration Flows

Jorge Saralegui-Gil

1. Introduction

The great intensity of recent flows of immigrants into Spain and other European countries makes the appropriate administrative and, therefore, statistical recording of this phenomenon extraordinarily difficult. Shortly after the new subpopulations have moved into the national territory, distortions are bound to be produced in the quality of the sampling frames, in certain non-sampling errors and, more particularly, in the exogenous totals used to expand the sample weights. In light of these phenomena, it is essential to study and establish measures to absorb the undesired impacts of these new potential sources of error in the estimates.

2. Formulation of errors

Let $x_1$ be the survey target variable of subpopulation1 (i.e. proportion of extra-EU population working in the construction sector, etc.) in a domain, and $x_2$ the value for the complementary subpopulation. According to empirical evidence, it is well known that there are significant and systematic differences between their respective values for the main variables subject to observation, as well as in the behaviour in respect of non response and other measurement errors.

If $\hat{p}_1$ and $\hat{p}_2$ are the final estimators of subpopulations levels $p_1$ and $p_2$ (unknown), and $\hat{p}$ and $p$, respectively, the corresponding totals, then the total error of a sample or a proportion could be formulated as:

$$E = \left( (x_1 + \epsilon_{x_1}) - \frac{\hat{p}_1}{\hat{p}} + (x_2 + \epsilon_{x_2}) - \frac{\hat{p}_2}{\hat{p}} \right) \cdot \left( \frac{p_1}{p} + \frac{p_2}{p} \right)$$

(1)

The component of this difference that can be attributed to the effects due to the presence of the subpopulation 1 can be thus approximated by:

$$E \cong (x_1 - x_2) \left( \frac{\hat{p}_1}{\hat{p}} - \frac{p_1}{p} \right) + \epsilon_{x_1} \frac{\hat{p}_1}{\hat{p}} ;$$

(2)

---

1 Jorge Saralegui-Gil, National Institute of Statistics (INE), Rosario Pino,14-16, room 0502 Madrid 28020, Spain
Where \( \varepsilon \) synthesises non sampling errors in the means or proportions for that 'special' subpopulation, as well as the sampling error itself, due to inappropriate field work procedures to approximate respondents, the "informal" nature of the activity, and other observation and measurement errors, as well as the sampling error itself. Let us assume that, in the case of subpopulation 2, these errors \( (\varepsilon) \), which would have introduced the component \( \hat{P} \) in the above expression, are of no significance in comparison with those of subpopulation 1, which are controlled by the design and counteracted by the usual survey procedures, and consequently not to be dealt with here.

3. Error treatment

Formula (2) shows three error components, the control or cancellation of which require different, diverse actions.

a) \( \varepsilon \) can be reduced by the following procedures:

a.1) Increases in the sampling sizes of subpopulation 1 through additional samples or controlled selection. This type of sampling procedure, however, has several disadvantages that must be taken into account during project decision-making. The 'design effects' (increase in sampling error) due to the variability of weights introduced by the controlled selection of sub-samples may have an impact on the efficiency of other estimates related to the general population (Kish, 1979).

a.2) Improvements in interviewer training, collection instruments or mechanics and other field work techniques.

This procedure would possibly be the most beneficial for the correction of bias due to framework errors, the total or partial non response, hidden response errors, etc., but it requires an initial investment in resources for field work implementation not always affordable.

When it comes to the first term of (2), its analysis is to be made taking into account the exogenous population forecasts, necessary to expand sampling weights. Obviously, population forecasts are one of the statistical products which most 'suffer' in the presence of intensive and/or unexpected flows of foreign migration and, therefore, the presence of a potential model specification error must be taken into account when a now-cast population estimate is to be used by at the estimation step.

Let us assume, therefore, that we have the population now-casts and their ratios between the target subpopulation and the total, \( \left( \frac{\hat{P}}{P} \right) \) with its own error unknown, thus

\[
\left( \frac{\hat{P}}{P} - \frac{P}{P} \right) = \left[ \left( \frac{\hat{P}}{P} - \frac{\hat{P}}{P} \right) + \left( \frac{\hat{P}}{P} - \frac{P}{P} \right) \right]
\]

(3)
The first parenthesis of the right side in (3) is mainly caused by the higher propensity to non response by new residents in comparison with nationals. Its impacts on the total error depends on the structural differences in the ‘real’, unknown, means or proportions between subpopulation 1 and 2. Both effects have been empirically measured through estimates provided by the system of household surveys, and can be neutralized by using calibration or post-stratification techniques. 

Calibration of weights at household and individual level is the procedure chosen by the Spanish National Statistical Institute (INE) to adjust estimated populations to external structures by gender and age in both continuous and sporadic household surveys. As of the first quarter 2005, population distributions by nationality were included in the external vectors of auxiliary variables for calibration in some of the main surveys as INE LFS, using CALMAR freeware. These techniques are extremely useful for reducing total error in the estimates, although care must be taken in defining post-strata (in this case, groups of nationalities), to ensure they provide enough sample observations over any period, while remaining well balanced in relation to the target variables. However, it is important to note that the bias due to non response within post-strata are insensitive to calibration and, therefore, attention should always be paid to the aforementioned type a2 actions.

In terms of the second parenthesis of the second member of (3), its effect is controlled, although in practice it will never be nil, improving the model based estimates of population structures involved in the calculation of expansion factors. These improvements can be achieved through new external sources or additional forecast models adapted to the new conditions introduced by the intensification of the migratory flows. It is even possible to use the great potential of household surveys currently available within the official statistics system, in order to improve the quality of the models which provide demographic nowcast estimates relative to these new subpopulations, as developed in the following sections.

4. Estimates of net immigrant flows based on the change in household continuous surveys

Existing household surveys within the official statistics system provide a great potential to improve population now-cast models, as in the procedure, already tested in Spain, based on the use of a macro-survey such as the quarterly continuous Labour Force Survey.

There are two project dimensions which confer the LFS with a special profile among sampling surveys and which have proven very robust over the almost fifty years of existence of the survey:

a) Its two-stage design, the list of dwellings within the first stage units (PSU, geographic subdivisions within a municipality containing on average 500 households) being updated exhaustively and systematically during field work by specialised personnel every six quarters (one sixth every quarter), i.e. an update based on the real world, independently to the (administrative) Population Register.
b) Its rolling nature in the second stage, with six rolling sub-samples that allow for a complete renewal of the sample of dwellings (around 18 units within each PSU) every six quarters. First stage sample remains unchanged except for occasional updating due to changes in the selection probability. This allows for the accurate measurement of the ‘net’ dynamics of individuals entering or leaving the PSU. It is important to note that the first stage of the LFS, around 3,500 PSU’s, account for approximately 10% of the Country. Therefore, the coverage of the ‘net’ migrant population in the case of no subsampling within PSU can be very accurate. Actually, the analysis of the LFS first stage sample shows the extremely good coverage of the foreign population that would have been attained if all of the households within sampled PSU’s had been selected, thus showing that the under coverage effects of the foreign population in the LFS are due to second stage sample non response. Taking profit of these properties of LFS sample design, a robust estimation of the change in the structure of nationality of the population can be reached, based on the characteristics observed for the individuals at the wave when they are included for the first time in the sample.

5. The model

Let’s consider a population of reference SR (i.e. resident nationals, other choices being possible) for which demographic information of good quality is available on continuous basis. Let ST be the target subpopulation (i.e. residents non nationals). We refer here to an already tested solution based on the hypothesis that the differential total non response propensity SR versus ST (the latter much higher) is in practice constant between two periods of comparison (which must thus be as close as possible). More complex models, also under development, may ignore that hypothesis and introduce coefficients of differential non response propensity per nationality, using non response indicators from other current surveys based on samples selected from a list frame (Population Register), where the characteristic ‘nationality’ is available for the theoretical sample. In the simplified approach presented here, \( \Delta r_a \) is the estimator of change in the ST/SR ratio based on the subsample being interviewed for the first time, for a set of consecutive periods \( j \), measuring thus the change on the full LFS first stage sample:

\[
\Delta r_a = \left[ \sum_{j=7}^{j=12} \hat{x}_{aj} \sum_{j=1}^{j=6} \hat{x}_{aj} - \sum_{j=7}^{j=12} \hat{y}_j \sum_{j=1}^{j=6} \hat{y}_j \right]; (4)
\]

\( \hat{x}_{aj} \): estimator of subpop. ST, domain \( a \), based exclusively on the rotating group being interviewed for the first time; \( \hat{y}_j \): estimator of the SR (Nationals > 15 years), same PSU subsample. To obtain the ratio (change) estimates, the weights implicit in the estimators should not be calibrated to external totals disaggregated by nationality. The TS nowcast based on the change observed in LFS can be calculated as (several alternative formulations are also possible):

\[
\hat{X}_a (T) = \left( \frac{\hat{X}_a (T-6)* \Delta r_a * \hat{Y}(T) - \hat{X}_a (T-6)}{\hat{Y}(T-6)} \right) + (1/6) \hat{X}_a (T-1); (5)
\]
\( \hat{x} \) refers to adjusted totals of ‘net flows’ of the ST group under study and \( \hat{y} \) the original nowcasts of the SR group, both from quarter T-6, T-1 or T. To start the series, a base period b must be defined on which the target ratio is thought to be very reliable (e.g. in a census period or a special updating period for registers). Figure 1 shows how the series of official forecasts are very closely estimated by the LFS change estimator, while the direct LFS estimation underestimate more than 50% the foreign population, which is still arriving to Spain so intensively.

**Fig 1. Population forecasts versus estimator of change LFS. Foreign population in Spain.**

![Graph](image)

**References**


HOW TO COUNTERACT PROBLEMS OF QUALITY IN HOUSEHOLD SURVEYS CAUSED BY INTENSIVE MIGRATION FLOWS

Jorge Saralegui-Gil
Ine (Spain)
Q2006
CARDIFF, UK
Factors which determine the impact of Inmigration Flows in Measurement Errors.
Impacts depend on:

- Intensity of the flows. Concentration in time.
- Higher propensity to total non response and other observation errors of new entrants in comparison with resident population.
- Remarkable differences in structural behaviour in respect of the main survey variables.
- Poor quality of exogenous totals of inmigrants, provided by administrative sources.
- (Importance of impacts increases when those factors occur simultaneously)
Evolution of foreign residents total

- 1997: 542314
- 1998: 637085
- 1999: 748954
- 2000: 923879
- 2001: 1370657
- 2002: 1977946
- 2003: 2672596
- 2004: 3034326
- 2005: 3730610

Foreign residents Regional Distribution

SPAIN 2005
(8.5%)
Resident foreigners according to continent of nationality

- Asia: 4.8%
- EU: 22%
- Rest of Europe: 13%
- Africa: 19.6%
- South and Central America: 38.6%
- North America: 1.5%
- Unclassified: 0.4%
### NON RESPONSE PER NATIONALITY IN SURVEYS

#### SPAIN

**Surv. Use of Technologies in Households**

<table>
<thead>
<tr>
<th></th>
<th>Theoretical</th>
<th>Theoretical</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Effective sample</td>
<td>sample 1 (a)</td>
</tr>
<tr>
<td>Dwellings with at least 1 foreigner</td>
<td>339</td>
<td>775</td>
</tr>
<tr>
<td>Dw. with nationals only</td>
<td>10101</td>
<td>16343</td>
</tr>
<tr>
<td>Non response propensity (C)</td>
<td>1,41</td>
<td>1,22</td>
</tr>
</tbody>
</table>

**Time Use Survey**

<table>
<thead>
<tr>
<th></th>
<th>Theoretical</th>
<th>Theoretical</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Effective sample</td>
<td>sample 1 (a)</td>
</tr>
<tr>
<td>Dwellings with at least 1 foreigner</td>
<td>528</td>
<td>1132</td>
</tr>
<tr>
<td>Dw. with nationals only</td>
<td>14255</td>
<td>22312</td>
</tr>
<tr>
<td>Non response propensity (C)</td>
<td>1,37</td>
<td>1,13</td>
</tr>
</tbody>
</table>

(a) Including: effective sample, inhabited, not found, not at home, refusals, unable to respond
(b) Including: effective sample, not at home, refusals, unable to respond
(c) Calculated as (probability of non response foreigners) / (probability of non response nationals)
Error Formulation

\[ \hat{P}_1 \quad \hat{P}_2 \quad \text{Estimators of target & complementary subpopulations} \]

\[ \left( \frac{\hat{P}_1}{\hat{P}} \right) \quad \text{(Pop. Forecasts structures)} \]

\[ \mathcal{E}_{x_1} \quad \text{Measurement and sampling errors for target subpop.} \]

Then \[ E \cong (\bar{x}_1 - \bar{x}_2) \left( \frac{\hat{P}_1}{\hat{P}} - \frac{P_1}{P} \right) + \varepsilon_{\bar{x}_1} \frac{\hat{P}_1}{\hat{P}} \]

With \[ \left( \frac{\hat{P}_1}{\hat{P}} - \frac{P_1}{P} \right) = \left[ \left( \frac{\hat{P}_1}{\hat{P}} - \frac{\hat{P}_1}{\hat{P}} \right) + \left( \frac{\hat{P}_1}{\hat{P}} - \frac{P_1}{P} \right) \right] \]
Technical actions to counteract the impacts on Hds. Surveys quality in the presence of intensive in-migration flows (I)

Action on $\mathcal{E}_{\bar{x}_1}$

- Field Work actions: Special targeted investment on field work costs to improve the performance of immigrants in respect of non response and other observation errors.

- Supplementary samples of immigrants.
Technical actions to counteract the impacts on Hds. Surveys quality in the presence of intensive in-migration flows (II)

Action on

\[
\left( \frac{\hat{P}}{\hat{P}} - \frac{\hat{P}}{\hat{P}} \right)
\]

- Calibration/post-stratification of the sample per nationality/origin.

Action on

\[
\left( \frac{\hat{P}}{\hat{P}} - \frac{P}{P} \right)
\]

- Improving the quality of external totals on in-migration flows, provided by administrative sources.
Monitoring migration flows with a current household survey like the LFS given its SAMPLING SCHEME:

- Two Stages: first stage PSU’s, areas of 500 hds., on average. Selection pps.
- Second Stage dwellings s.r.s.
- 3400 PSU’s (aprox. 12% national territory)
- 68000 dwellings.
- Rotating Scheme: First stage fixed in time (except changes in selection probability).
- Second stage, within 1/6 PSU’s the subsample of dwellings rotates every quarter.
Rotating Sample of dwellings within fixed AREAS

What makes it suitable to measure Migration Flows?

• First Stage units are ‘AREAS’ aprox. fixed in time, whose limits are updated regularly.
• Lists of dwellings within first stage PSU’s are updated exhaustively every 6 quarters.
• Thus all new residents within first stage sampled areas do have probability of selection during the rotating period.
• In absence of non response and with no subsampling, first stage sample will suffice to estimate population structure (i.e. Per nationality, date of arrival, others) with high quality.
Rotating Sample of dwellings within fixed AREAS:
Steps to measure net migration flows

• Choose a subpopulation of reference Y, measured currently with high quality. ex.: Spaniards resident in Spain.
• For target subpopulation X, estimate Change in X/Y based only on the sample interviewed for the first time each wave of the rotating scheme.
Rotating Sample of dwellings within fixed AREAS.
Change in Subpopulations Ratio estimated in LFS
For a set of consecutive periods \( j \) ratio change can be estimated using the full LFS first stage sample in a domain ‘a’ as:

\[
\Delta \hat{r}_a = \left[ \sum_{j=7}^{j=12} \hat{X}_{aj} : \sum_{j=1}^{j=6} \hat{X}_{aj} \right] \left[ \sum_{j=7}^{j=12} \hat{Y}_j : \sum_{j=1}^{j=6} \hat{Y}_j \right]
\]
Rotating Sample of dwellings within fixed AREAS:
Estimation of level, based on ratio change

Thus, level of subpopulation X in quarter T can be estimated from the ‘moving’ change estimator of the ratio X/Y centred in T-6 as (double hat means we are estimating the ‘official’ forecast):

$$\hat{X}'_a(T) = \left( \frac{\hat{X}'_a(T-6)}{\hat{Y}(T-6)} \right) \Delta_a \cdot \hat{Y}(T) - \hat{X}'_a(T-6) \right) \frac{1}{6} + \hat{X}'_a(T-1)$$
<table>
<thead>
<tr>
<th>Year/Quart</th>
<th>Nationals</th>
<th>Foreigners</th>
<th>Change in ratio (X/Y)</th>
<th>LFS estimator based on Change first interview</th>
<th>Exogenous Oficial Forecast of &gt; 15 years old (base 2001)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000/1</td>
<td>5.214.794</td>
<td>58.606</td>
<td></td>
<td>750.377</td>
<td>32.943.158</td>
</tr>
<tr>
<td>2000/2</td>
<td>5.260.275</td>
<td>64.570</td>
<td></td>
<td>826.120</td>
<td>32.977.161</td>
</tr>
<tr>
<td>2000/3</td>
<td>5.192.891</td>
<td>62.040</td>
<td></td>
<td>901.865</td>
<td>33.011.142</td>
</tr>
<tr>
<td>2000/4</td>
<td>5.389.563</td>
<td>91.468</td>
<td></td>
<td>977.608</td>
<td>33.045.139</td>
</tr>
<tr>
<td>2001/1</td>
<td>5.383.945</td>
<td>81.432</td>
<td></td>
<td>1.057.946</td>
<td>33.074.548</td>
</tr>
<tr>
<td>2001/2</td>
<td>5.589.995</td>
<td>108.967</td>
<td></td>
<td>1.135.648</td>
<td>33.099.392</td>
</tr>
<tr>
<td>2001/3</td>
<td>5.546.540</td>
<td>95.748</td>
<td></td>
<td>1.219.056</td>
<td>33.124.229</td>
</tr>
<tr>
<td>2001/4</td>
<td>5.548.325</td>
<td>110.599</td>
<td></td>
<td>1.308.589</td>
<td>33.149.065</td>
</tr>
<tr>
<td>2002/1</td>
<td>5.536.149</td>
<td>118.097</td>
<td></td>
<td>1.404.257</td>
<td>33.163.503</td>
</tr>
<tr>
<td>2002/2</td>
<td>5.514.094</td>
<td>119.230</td>
<td></td>
<td>1.505.487</td>
<td>33.146.436</td>
</tr>
<tr>
<td>2002/3</td>
<td>5.366.136</td>
<td>130.593</td>
<td>1,523088</td>
<td>1.614.016</td>
<td>33.129.391</td>
</tr>
<tr>
<td>2002/4</td>
<td>5.541.740</td>
<td>176.738</td>
<td>1,558170</td>
<td>1.719.779</td>
<td>33.112.348</td>
</tr>
<tr>
<td>2003/1</td>
<td>5.559.972</td>
<td>180.970</td>
<td>1,623173</td>
<td>1.846.401</td>
<td>33.125.030</td>
</tr>
<tr>
<td>2003/2</td>
<td>5.582.325</td>
<td>169.244</td>
<td>1,604222</td>
<td>1.978.170</td>
<td>33.148.054</td>
</tr>
<tr>
<td>2003/3</td>
<td>5.334.193</td>
<td>174.058</td>
<td>1,572809</td>
<td>2.112.316</td>
<td>33.171.085</td>
</tr>
<tr>
<td>2003/4</td>
<td>5.620.413</td>
<td>189.020</td>
<td>1,615206</td>
<td>2.267.263</td>
<td>33.194.109</td>
</tr>
<tr>
<td>2004/1</td>
<td>5.604.041</td>
<td>190.294</td>
<td>1,574468</td>
<td>2.422.878</td>
<td>33.213.998</td>
</tr>
<tr>
<td>2004/2</td>
<td>5.545.763</td>
<td>188.232</td>
<td>1,445340</td>
<td>2.552.007</td>
<td>33.230.751</td>
</tr>
</tbody>
</table>
Population forecasts versus estimator of change LFS. Foreign population in Spain.

![Graph showing population forecasts versus estimator of change LFS. Foreign population in Spain.](graph.png)

- **Estimator of change foreigners first wave LFS**
- **Foreigners Official Forecasts**
- **Direct estimation full LFS**

The graph illustrates the comparison between the Population forecasts and the estimator of change for foreigners in Spain, highlighting the changes from 2000 to 2004.
HOW TO COUNTERACT PROBLEMS OF QUALITY IN HOUSEHOLD SURVEYS CAUSED BY INTENSIVE MIGRATION FLOWS

Jorge Saralegui-Gil
Ine (Spain)
Q2006
CARDIFF, UK
The R ‘sampling’ package

Alina Matei and Yves Tillé
University of Neuchâtel

Cardiff, Q2006
April 2006
The R language

- Shareware available on http://cran.r-project.org/
The R language

- Shareware available on http://cran.r-project.org/
- The Comprehensive R Archive Network
The R language

- Shareware available on http://cran.r-project.org/
- The Comprehensive R Archive Network
- Installation: 10 minutes
The R language

- Shareware available on http://cran.r-project.org/
- The Comprehensive R Archive Network
- Installation: 10 minutes
- All the manuels are available in pdf
The R language

- Shareware available on http://cran.r-project.org/
- The Comprehensive R Archive Network
- Installation: 10 minutes
- All the manuels are available in pdf
- Everyone can write an additional package (600 packages are available)
The R language

- Shareware available on http://cran.r-project.org/
- The Comprehensive R Archive Network
- Installation: 10 minutes
- All the manuels are available in pdf
- Everyone can write an additional package (600 packages are available)
- Packages are loaded directly from R.
The R language

- Shareware available on http://cran.r-project.org/
- The Comprehensive R Archive Network
- Installation: 10 minutes
- All the manuels are available in pdf
- Everyone can write an additional package (600 packages are available)
- Packages are loaded directly from R.
- The manual of the package is available online and in pdf.
The R language

- Shareware available on http://cran.r-project.org/
- The Comprehensive R Archive Network
- Installation: 10 minutes
- All the manuals are available in pdf
- Everyone can write an additional package (600 packages are available)
- Packages are loaded directly from R.
- The manual of the package is available online and in pdf.
- Package ‘sampling’ written by Matei and Tillé.
EFTA course for public statisticians (April 2005).
Continuous distributions

- EFTA course for public statisticians (April 2005).
- Objective: to apply directly the theory with the R language.
Continuous distributions

- EFTA course for public statisticians (April 2005).
- Objective: to apply directly the theory with the R language.
- Theory + Exercises with a laptop and R.
Continuous distributions

- EFTA course for public statisticians (April 2005).
- Objective: to apply directly the theory with the R language.
- Theory + Exercises with a laptop and R.
- Writing of a large set of procedures.
Continuous distributions

- EFTA course for public statisticians (April 2005).
- Objective: to apply directly the theory with the R language.
- Theory + Exercices with a laptop and R.
- Writing of a large set of procedures.
- Finally, decision of submitting the package to the CRAN.
Content of the package

- Stratification, two-stage, unequal probabilities, balanced sampling
Content of the package

- Stratification, two-stage, unequal probabilities, balanced sampling
- Estimation: calibration and regression estimator
Content of the package

- Stratification, two-stage, unequal probabilities, balanced sampling
- Estimation: calibration and regression estimator
- Tools: computation of inclusion probabilities, crossing strata
Content of the package

- Stratification, two-stage, unequal probabilities, balanced sampling
- Estimation: calibration and regression estimator
- Tools: computation of inclusion probabilities, crossing strata
- Data bases, Swiss municipalities, Belgian municipalities.
Tools

▶ writesample: return the list of all the samples of fixed sample size
Introduction and aim

Content of the package

Tools

- **writesample**: return the list of all the samples of fixed sample size
- **cleanstrata**: renumbering of the strata
Introduction and aim

Content of the package

Topics
Tools
Data bases
Simple random sampling
Unequal probability sampling
Balanced sampling

Tools

- writesample: return the list of all the samples of fixed sample size
- cleanstrata: renumbering of the strata
- disjonctive: return a matrix with 0 and 1 that is the disjonctive representation of the stratum.

Alina Matei and Yves Tillé University of Neuchâtel
Tools

- **writesample**: return the list of all the samples of fixed sample size
- **cleanstrata**: renumbering of the strata
- **disjonctive**: return a matrix with 0 and 1 that is the disjonctive representation of the stratum.
- **inclusionprobabilities**: compute unequal inclusion probabilities from an auxiliary variable variable.
Data bases

- MU284 A data frame with 284 municipalities on the following 11 variables: populations, political results.
Data bases

- MU284: A data frame with 284 municipalities on the following 11 variables: populations, political results.
- swissmunicipalities: 2896 Swiss municipalities. Surfaces and population.
Data bases

- **MU284** A data frame with 284 municipalities on the following 11 variables: populations, political results.

- **swissmunicipalities**: 2896 Swiss municipalities. Surfaces and population.

- **belgianmunicipalities**: 589 Belgian municipalities 11 variables, population and taxes.
Simple random sampling

- srswor: Simple random sampling with replacement.
Simple random sampling

- `srswor`: Simple random sampling with replacement.
- `srswor1`: Simple random sampling without replacement (sequential method).
Simple random sampling

- `srswo`: Simple random sampling with replacement.
- `srswo1`: Simple random sampling without replacement (sequential method).
- `srswr`: Simple random sampling with replacement.
Unequal probability sampling

- UPbrewer,
Unequal probability sampling

- UPbrewer,
- UPmaxentropy, (set of function)
Unequal probability sampling

- UPbrewer,
- UPmaxentropy, (set of function)
- UPmidzuno, UPmidzunopi2,
Unequal probability sampling

- UPbrewer,
- UPmaxentropy, (set of function)
- UPmidzuno, UPmidzunopi2,
- UPmultinomial,
Unequal probability sampling

- UPbrewer,
- UPmaxentropy, (set of function)
- UPmidzuno, UPmidzunopi2,
- UPmultinomial,
- UPpivotal, UPrandompivotal,
Unequal probability sampling

- UPbrewer,
- UPmaxentropy, (set of function)
- UPmidzuno, UPmidzunopi2,
- UPmultinomial,
- UPpivotal, UPrandompivotal,
- UPpoisson,
Unequal probability sampling

- UPbrewer,
- UPmaxentropy, (set of function)
- UPmidzuno, UPmidzunopi2,
- UPmultinomial,
- UPpivotal, UPrandompivotal,
- UPpoisson,
- UPSampford,
Unequal probability sampling

- UPbrewer,
- UPmaxentropy, (set of function)
- UPmidzuno, UPmidzunopi2,
- UPmultinomial,
- UPPivotal, UPrandompivotal,
- UPpoisson,
- UPSampford,
- UPsystematic, UPrandomsystematic, UPsystematicpi2,

Alina Matei and Yves Tillé University of Neuchâtel

The R ‘sampling’ package
Unequal probability sampling

- UPbrewer,
- UPmaxentropy, (set of function)
- UPmidzuno, UPmidzunopi2,
- UPmultinomial,
- UPpivotal, UPrandompivotal,
- UPpoisson,
- UPSampford,
- UPsystematic, UPrandomsystematic, UPsystematicpi2,
- UPtille, UPtillepi2,
Balanced sampling

- Design that satisfies the balancing equations

\[
\sum_{k \in S} x_k \pi_k = \sum_{k \in U} x_k,
\]

where \( x_k \) is a vector of auxiliary variables.
Introduction and aim
Content of the package

Balanced sampling

- Design that satisfies the balancing equations
  \[ \sum_{k \in S} \frac{x_k}{\pi_k} = \sum_{k \in U} x_k, \]
  where \( x_k \) is a vector of auxiliary variables.
- Cube algorithm: flight phase and landing phase.

Alina Matei and Yves Tillé
University of Neuchâtel

The R ‘sampling’ package
Balanced sampling

- Design that satisfies the balancing equations

\[
\sum_{k \in S} \frac{x_k}{\pi_k} = \sum_{k \in U} x_k,
\]

where \( x_k \) is a vector of auxiliary variables.
- Cube algorithm: flight phase and landing phase.
- `samplecube`, `fastflightcube`, `landingcube`
Balanced sampling

- Design that satisfies the balancing equations

\[ \sum_{k \in S} \frac{x_k}{\pi_k} = \sum_{k \in U} x_k, \]

where \( x_k \) is a vector of auxiliary variables.

- Cube algorithm: flight phase and landing phase.
- `samplecube`, `fastflightcube`, `landingcube`
- Complex survey balancedstratification balancedcluster balancedtwostage
The Methodological Approaches to Definition of the Quality of External Information at Formation of Sample

Natalia Vashchaeva

Abstract (paper unavailable)

The basic sources of the information about the socio-economic characteristics of the population (households) of Ukraine are the state sample surveys, such as survey of households living conditions, survey of the population economic activity, survey of agricultural activity. At any stage of these surveys realization there is a necessity to use the external information of the appropriate quality.

Formation of the household sample is executed on the basis of the various information, which directly or indirectly is concerned with households. In particular, it is number of households or number of the residents in the inhabited locality, number of the voters on election centers, the average size of family etc. The correctness of this information, its quality for samples with complex design appreciably influences reliability of the estimation of general probability of household selection and, as consequence, on reliability of definition of household base weights as well as on sample mistakes and displacements at indicators estimation.

In the report the basic sources of reception of the external information are considered. They are systematized for a type, sort and structure. It allows to define opportunities of the external information use and to prove utility of its use at different stages of sample formation, namely at designing and formation of qualitative sample design, construction of sampling frame, realization of selection procedures, control of sample formation quality.

The external information, which is used at sample formation, differs on its quality. Quality of the external information can be defined by methodology of its receiving. But it is very difficult, and frequently it is impossible to estimate methodology of the information receiving. In the report it is shown, that the influence of the used external information (its quality) directly on results of sample survey in the quantitative form can be estimated, mainly, since a stage of the sample formation.

Thus, estimation of quality of the external information, which is used at sample formation, on the basis of survey data is an actual problem. Besides, for modern population sample surveys development of methods of the quality monitoring of the external information is important. The decision of these problems will allow to increase quality of surveys results as a whole, including due to correction of mistakes and displacements which didn’t manage to be corrected at sample formation.

Key words: sample survey, quality of external information, sample design.
Is quality at stake in Official Statistics?

Jelke Bethlehem

Statistics Netherlands
Methods and Informatics Department
Official Statistics

Why Official Statistics?

- The availability of reliable statistical information is indispensable for a democratic society.
- To fulfil their role, Official Statistical Institutes must have the trust of the public, of respondents and of users.
- Statistical figures should be unbiased and reliable. They should never be influenced by political or other interest groups
- They should be produced in accordance with professional quality standards.
European Statistics Code of Practice

- Purpose
  - Improving trust and confidence in the independence, integrity and accountability of Official Statistics, and in the credibility and quality of their statistics.
  - Providing a benchmark of statistical principles, values and best practices that should help them in producing and disseminating high quality statistics.
European Statistics Code of Practice

Principles

- Professional independence of statistical institutes
- Legal mandate for data collection
- Adequate resources
- Commitment to quality
- Sound methodology
- Relevance
- Accuracy and reliability
- Timeliness and punctuality
- Coherence and comparability
- Accessibility and clarity
ISI Declaration on Professional Ethics

**Principles**

- Statisticians should consider the likely consequences of collecting and disseminating various types of data and should guard against predictable misinterpretations or misuse.
- Statisticians should attempt to uphold their professional integrity without fear or favour.
- They should not engage or collude in selecting methods designed to produce misleading results, or in misrepresenting statistical findings by commission or omission.
Some threats

- **Budget cuts**
  - Can we maintain quality at lesser costs?
  - Should we rely more on models that cannot be checked?

- **Reducing administrative burden**
  - Should we and can we rely on register data?
  - Collected for different purpose by different agency with different definitions?
  - Do we control statistical quality?
  - Can we maintain professional independence?
Some threats, continued …

- **Commercial web surveys / panels**
  - More than 20 large online panels in the Netherlands
  - More than 1,700,000 panel members (> 10%).
  - Many in several panels. 47% completed another questionnaire less than one week ago.
  - People are paid for participation. Professional respondents? Fear of over-exploitation.
  - No focus on quality. Customers do not ask questions.
  - Survey climate spoiled?
  - Should statistical institutes should replace their surveys by large online panels?
So, the questions are …

- Will there be a point at which we cannot guarantee any more the quality of the statistics we produce?

- Will our professional integrity be at stake at this point?

- What could possible actions be to avoid this?
A model for statistical inference based on mixed-mode interviewing

Fannie Cobben and Barry Schouten
Statistics Netherlands
Introduction

Mixed Mode data collection to compensate the weakness of individual modes
Consequence: data from different processes
How to combine?
In this presentation…

Different mixed mode set-ups
Model for nonresponse
Combination of mixed mode and NR-model
Issues
Concurrent vs sequential MM

De Leeuw (2005):
Assume that the allocation probabilities $\eta$ to a specific mode are known:

$$\begin{align*}
\eta(x_i) & \quad 1-\eta(x_i) \\
\rho_1(x_i) & \quad 1-p_1(x_i) \\
R & \quad NR \\
p_2(x_i) & \quad 1-p_2(x_i) \\
R & \quad NR
\end{align*}$$
## Concurrent model

\[ \text{Mode}_i = \begin{cases} 
1, & \text{with probability } \eta(x_i) \\
0, & \text{with probability } 1 - \eta(x_i) 
\end{cases} \]

For \( m = 1, 2 \):

\[ I_m^* = \beta^m X_i^m + \delta_i^m \]

\[ I_m = I\{I_m^* \geq 0\} \]

\[ Y_i^* = \gamma X_i + \mu_i \]

\[ Y_i = \begin{cases} 
Y_i^* & \text{if } I_1 = 1 \text{ or } I_2 = 1 \\
- & \text{else}
\end{cases} \]

Use different variables in \( X, X^1 \) and \( X^2 \)

Equations are correlated

Ignore measurement error

---

Q 2006, Cardiff, April 24-26
Sequential Mixed Mode

Mode $m$: $\rho^1(x_i)$, $1 - \rho^1(x_i)$

Mode $n$: $\rho^2(x_i)$, $1 - \rho^2(x_i)$
Sequential model

Mode equations:
\[ I_i^* = \alpha X_i^I + \epsilon_i \quad I_i = I\{I_i^* \geq 0\} \]
\[ J_i^* = \beta X_i^J + \delta_i \quad J_i = I\{J_i^* \geq 0\} \]

Regression equation:
\[ Y_i^* = \gamma X_i + \mu_i \]
\[ Y_i = \begin{cases} 
Y_i^* \text{ if } I_i = 1 \text{ or } J_i = 1 \\
- \text{ else} 
\end{cases} \]

Again, use different variables in $X, X^I$ and $X^J$.

Correlation structure of error terms.
Nonresponse: one individual mode


‘Peel off’ the participation process:

Contact
Language problem
Not able due to illness
Participation
The participation process

- Full sample
  - Contact
  - Non-contact
    - No language problem
      - Able
        - Refusal
      - Not able
        - Participation
    - Language problem
The model

Participating equations:
\[ I_{ij}^* = \beta_j X_{ij} + \varepsilon_{ij} \]
\[ I_{ij} = I\{I_{ij}^* > 0\} \]
\[ \theta_j = (\alpha_j, \beta_j) \]

Regression equation:
\[ Y_i^* = \gamma X_i + \mu_i \]
\[ Y_i = \begin{cases} Y_i^* & \text{iff } I_{i4} = 1 \\ - & \text{else} \end{cases} \]

Different variables in \( X, X_1, \ldots, X_4 \)
Thus far...

We have:

Two different mixed mode set-ups: *concurrent* and *sequential*

A model for nonresponse / participation.

These are our ‘building blocks’.

For a statistical model that can be used in a mixed mode setting we have to combine them.
Combination

Take a mixed mode set-up and ‘plug in’ the participation process.

*Concurrent MM*: participation process in leaves of tree.

*Sequential MM*: participation process before allocation to a different mode. All sources of nonresponse proceed to the next mode.
Example 1: Basic Question Approach in the Dutch LFS

Full sample: CAPI

- $\rho_1(x_i)$
- $1-\rho_1(x_i)$

- R
- NR

$\eta(x_i)$

- CATI
- PAPI/CAWI

- $\rho_2(x_i)$
- $1-\rho_2(x_i)$
- $\rho_3(x_i)$
- $1-\rho_3(x_i)$

$\eta(x_i)$

- R
- NR
- R
- NR
Example 2: Pilot Safety Monitor

See paper by Björn Janssen

Full sample: CAWI

\[ \rho^1(x_i) \quad 1-\rho^1(x_i) \]

\[ I_1 \]

\[ R \quad NR \]

\[ \eta(x_i) \quad 1-\eta(x_i) \]

\[ \text{CATI} \quad \text{CAPI} \]

\[ \rho^2(x_i) \quad 1-\rho^2(x_i) \quad \rho^3(x_i) \quad 1-\rho^3(x_i) \]

\[ I_2 \quad I_3 \]

\[ R \quad NR \quad R \quad NR \]
Issues

Structure of covariance matrix
How to solve the model by maximum likelihood?
Inclusion probabilities?
Different modes; different questionnaires
Categorical target variables
Abstract: Household surveys can be conducted using various data collection modes. Each individual data collection mode has its shortcomings. Face-to-face interviewing is expensive. Not every household has a telephone/Internet connection and can be approached by CATI resp. a Web survey. Mail surveys have a low response rate. Mixing data collection modes provides an opportunity to compensate for the weakness of each individual mode. This can reduce costs and at the same time increase the response. It is even possible to reduce the selectivity of the response beforehand. For this purpose, sampled persons or households can be allocated to a specific mode based on known background characteristics.

An optimal mixed mode strategy may still be in the future, but suppose that we have a survey administration system with decision rules that can support any strategy of mixed mode data collection. Such a system gives us a number of datasets, collected through different modes, for the same survey. How can we combine this data, so that we can use it for statistical inference?

We extend the sample selection model (see Heckman (1979)) so that it can be used to aggregate data from general mixed mode strategies and at the same time adjust for non-response bias.

Keywords: Non-response bias, mixed mode data collection, sample selection model

1. Introduction

1.1 Mixed mode data collection

In this paper, we assume that the data collection covers the entire population. We restrict ourselves to data collection during the response phase, i.e. we disregard the contact phase (e.g. notification letters, screener calls). We also do not regard the mixed mode variant where the choice of data collection mode is left to the respondent. De Leeuw (2005) describes two different mixed mode data collection designs. The first design is a concurrent system. The sample is divided in groups that are approached by
different mode, but at the same time. See Figure 1. The other design is a sequential
design. All sample elements are approached by one mode. The non-respondents are
then followed up by a different mode than used in the first approach. This process can
be repeated, see Figure 2.

Figure 1: Concurrent mixed mode design.

Modes differ in various aspects, e.g. timeliness or costs. Biemer and Lyberg (2003)
discuss optimal designs for unimode data collection. See Pierzchala (2006) for an
overview of the differences in cognition and response. Because of these disparities,
there are mode effects. A mode effect occurs if the answers of a respondent differ when
asked the same question in a different mode. It is difficult to evaluate mode effects. In
this paper, we make some assumptions regarding these effects. First, we assume that
there are no questionnaire effects. As De Leeuw (2005) notes, the questionnaires need
to be equivalent in a cognitive way (and can thus vary by mode without causing a mode
effect). In fact, we do not include measurement errors in our models in general.
1.2 Response process

Before a person actually participates in a survey, there are some hurdles to be taken. First, contact has to be made before a person can decide to comply with the survey request or not. When contact has been made, the person must be able to participate. A person can be unable to participate due to language problems, or may be able to speak the language but is unable to cooperate due to a longtime illness. When contact has been made, and a person is also able to participate, the last hurdle is the willingness to comply with the request.

The response process can be decomposed in a number of stages, see Figure 3. These decompositions differ between modes. One important distinction is the assistance of an interviewer. Especially the last step, i.e. refusal or cooperation, is influenced by the interaction between the interviewer and the sample person. See Groves et al. (1992). But there are other distinctions as well. For instance, in a Web as well as a paper survey it is only observed whether a person participates in the survey or not. The actual reason for the non-response is unclear. It could be a non-contact, a language problem, a longtime illness as well as a refusal.

Figure 3: The response process

Each of these steps relates to different characteristics of the respondent and the data collection. E.g. the non-respondents due to language problems will have a different profile than the refusers. By distinguishing between these types of non-response and the incorporation of instrumental variables we can better explain the process and eventually better adjust for non-response bias. In the literature, this approach is
suggested as well, see e.g. Lepkowski and Couper (2002) or Nicoletti and Peracchi (2005).

1.3 Outline

The aim of our research is to develop a model that combines data collected by mixed modes (both concurrent and sequential), thereby accounting for the different response processes in each of the individual modes. We translate and extend the model by Heckman (1979) to mixed mode and response models. We follow the bivariate probit model like in Van de Ven and Van Praag (1981).

In section 2, we present models to combine data in both the concurrent and the sequential mixed mode design. Section 3 discusses a model for the response process as described in subsection 1.2. A combination of the designs and the model for the response process is outlined in section 4, as well as an agenda for future research.

2. Mixed mode models

2.1 Concurrent

Recall that the concurrent mixed mode design assigns sample persons to a specific mode. All sample persons are thus approached at the same time but in different modes. One can think about optimal allocation strategies that reduce non-response bias and increase response. We assume that the allocation probabilities of the sample persons to a specific mode are known beforehand. The allocation probability for sample person \( i \) is denoted by \( \eta(x_i) \). We do not include the entire response process yet but only distinguish between response (R) and non-response (NR). In every mode \( m \) there is an underlying participation decision that determines the response probability \( \rho^m(x_i) \). This participation decision in mode \( m \) is denoted by \( l_m \). The model for the concurrent mixed mode with two modes looks like Figure 4.
Figure 4: Concurrent mixed mode model.

The answer to a survey question is obtained when the sample person is assigned to mode 1 and participates in mode 1 or when he/she is allocated to mode 2 and participates in this mode. The model can be described in a two-step manner. Conditional on the mode allocation and the response process for the person in the assigned mode, an answer to the survey question is obtained. This can be described in a similar way as the sample selection model proposed by Heckman (1979). Let us first introduce some notation. Let the target population of a sample survey consist of \( N \) individuals 1, 2, ..., \( N \). Let \( Y \) denote a target variable of the survey. Associated with each individual \( k \) is a value \( Y_k \) of this target variable. Assume that the aim of the sample survey is to estimate the population mean of the target variable

\[
\bar{Y} = \frac{1}{N} \sum_{k=1}^{N} Y_k.
\]  

(2.1.1)

Furthermore, let \( X \) be a vector of auxiliary variables or covariates, with values \( X_k \), for \( k = 1, 2, \ldots, N \). The sample selection model consists of two stages. In the context of survey participation the first stage models the participation of a person in the survey. Consequently, in the second stage the outcome to the survey is estimated while making use of the information from the first stage by correcting for the persons that did not participate. With the concurrent mixed mode design this process occurs for groups of persons in different modes; determined by the allocation probabilities \( \eta(x) \). There is a latent variable \( I^* \) that determines the participation. However, this variable is not observed. We only observe the outcome of the process (a response or a non-response in this case). In equation (2.1.2) the model for the first stage is described.
\[
Mode_i = \begin{cases} 
1, \text{with probability } \eta(x_i) \\
2, \text{with probability } 1 - \eta(x_i)
\end{cases}
\]

For \( m = 1, 2 \):
\[
I^*_{i,m} = \beta^*_i X^m_i + \delta^m_i
\]
\[
I_{i,m} = I \{ I^*_{i,m} \geq 0 \}
\]

The parameters \( \beta^*_i, \delta^m_i \) are resp. the vector of coefficients and the random error term for mode \( m \). We assume that each person has an answer to the survey and thus a value for the target variable. We just do not always observe it. This can also be modelled as a latent variable equation, where the target variable \( Y \) is the latent variable that can be explained by auxiliary variables \( X \) and a certain random error \( \mu \). See equation (2.1.3). This variable is observed conditional on the outcome of the participation process in the second part of equation (2.1.2).

\[
Y^*_i = \mu X_i + \mu_i
\]
\[
Y_i = \begin{cases} 
Y^*_i \text{ if mode 1 and } I_1 = 1 \text{ or mode 2 and } I_2 = 1 \\
- \text{ else}
\end{cases}
\]

Equation (2.1.2) and (2.1.3) are linked by their error terms \( \delta^1, \delta^2, \mu \). This is expressed in the correlation structure. Usually a multivariate normal distribution is assumed. See equation (2.1.4).

\[
\begin{pmatrix} \delta^1 \\ \delta^2 \\ \mu \end{pmatrix} \sim N \left( \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & \rho_{13} \\ 0 & 1 & \rho_{23} \\ \rho_{31} & \rho_{32} & \sigma^2 \end{pmatrix} \right)
\]

This model can e.g. be estimated by maximum likelihood, where the likelihood equals

\[
L = \sum_{i=1}^{n} \left[ \eta(x_i) I_{i,1} P(Y_i = y_i; I_{i,1} = 1 | X_i) + (1 - I_{i,1}) P(Y_i = y_i; I_{i,1} = 0 | X_i) \right] \\
+ (1 - \eta(x_i)) I_{i,2} P(Y_i = y_i; I_{i,2} = 1 | X_i) + (1 - I_{i,2}) P(Y_i = y_i; I_{i,2} = 0 | X_i) \right]
\]

2.2 Sequential

In a sequential mixed mode design, the entire sample is first approached by one specific mode. The non-respondents to that mode are then followed-up by another mode. This process can be repeated. Like in the concurrent design, there is a latent variable that determines the participation. The difference is that the process is now
repeated in time, for different modes and only for the non-respondents in earlier modes. See Figure 5.

**Figure 5: Sequential mixed mode model.**

In the same line of reasoning as the concurrent model, we can describe this model as follows in equations (2.2.1) – (2.2.4).

\[
I_{i,1}^* = \alpha Y_i + \varepsilon_i \quad I_{i,1} = I\{I_{i,1}^* \geq 0\}
\]

\[
I_{i,2}^* = \beta X_i + \delta_i \quad I_{i,2} = I\{I_{i,2}^* \geq 0\} \quad (2.2.1)
\]

\[
Y_i^* = \gamma X_i + \mu_i
\]

\[
Y_i = \begin{cases} 
Y_i^* & \text{if } I_{i,1} = 1 \text{ or } I_{i,2} = 1 \\
- & \text{else}
\end{cases} \quad (2.2.2)
\]

\[
\begin{pmatrix} 
\varepsilon_i \\
\delta_i \\
\mu_i 
\end{pmatrix} 
\sim N_3 \left( \begin{pmatrix} 
0 \\
0 \\
0 
\end{pmatrix}, \begin{pmatrix} 
1 & \rho_{12} & \rho_{13} \\
\rho_{21} & 1 & \rho_{23} \\
\rho_{31} & \rho_{32} & \sigma 
\end{pmatrix} \right) \quad (2.2.3)
\]

\[
L = \sum_{i=1}^{n} \left[ I_{i,1} P(Y_i = y_i; I_{i,1} = 1 | X_i) + (1 - I_{i,1}) I_{i,2} P(Y_i = y_i; I_{i,1} = 0; I_{i,2} = 1 | X_i) + (1 - I_{i,1})(1 - I_{i,2}) P(Y_i = y_i; I_{i,1} = 1; I_{i,2} = 0 | X_i) \right] \quad (2.2.4)
\]

The main difference between the concurrent and the sequential mixed mode model is in the correlation structure of the error terms and, consequently, in the likelihood. It
becomes clear in the distribution of the error terms from equations (2.2.1) and (2.2.2) that now the process is sequential. For a person to be a respondent in mode $m$, $m > 1$, this person has had to be a non-respondent in all the modes $< m$. This condition is translated by the correlation between the participation processes whereas in the concurrent mixed mode model these equations have a zero correlation, see equation (2.1.4).

3. Response model

In subsection 1.2 we describe the response process. We ‘peel off’ the entire process to distinguish between different types of non-response. Figure 3 gives a graphical display. As we already motivated, it is important to make a distinction between these groups because they can be very different. If, and only if, there are instrumental variables available that partially explain the various stages, we believe it to be beneficiary to model them separately. It can be used to better estimate the survey outcomes, i.e. to better adjust for non-response bias.

The model that is described in this section is able to combine all sources of non-response and to estimate a survey outcome incorporating instrumental information. Without this information one does not need to model stages separately from a non-response adjustment perspective. However, if one is interested in the separate stages one could leave the distinction. In that case the equations can be estimated separately as the correlations are introduced by the incorporation of instrumental variables alone. This is an important feature of the response model, because by combining the different non-response types into one model, the mass of the observations is preserved which means that the model has more explanatory power. See Figure 6 for a graphical representation. Of course, other decompositions can be modelled analogously.

Nicoletti and Peracchi (2005) make a similar distinction. They distinguish non-contact and refusal as causes of non-response. They do, however, not model the answers to the survey questions and focus solely on the characteristics of these two types of non-response.
Figure 6: The response model.

The model is again similar to the models in section 2.1 and 2.2, in that respect that we use a multiple stage structure as in the concurrent mixed model and a dependent structure as in the sequential mixed mode model. There are \( j = 4 \) selection equations that determine whether a person passes through to a next stage in the response process, which ends with refusal or participation. Hence there are 5 possible outcomes to the response process: Non-contact, language problem, not able due to long-time illness, refusal or participation. Equation (3.1) displays the selection equations for the \( j = 4 \) stages in the response process. The final probability of response (\( \pi \)) can be described by the probabilities determined by the selection equations (\( \pi^1, \pi^{11}, \pi^{111}, \pi^{1111} \)). See equation (3.2).

\[
I_{ij}^* = \beta_j X_{ij} + \varepsilon_{ij}
\]

\[
I_{ij} = I\{I_{ij}^* > 0\}
\]
\[ \pi_1 = P(I_1 = 1) \]
\[ \pi_{11} = P(I_2 = 1 | I_1 = 1) \]
\[ \pi_{111} = P(I_3 = 1 | I_2 = 1; I_1 = 1) \]
\[ \pi_{1111} = P(I_4 = 1 | I_3 = 1; I_2 = 1; I_1 = 1) \]
\[ \pi = P(\text{respond}) = \pi_1 \pi_{11} \pi_{111} \pi_{1111} \] (3.2)

The outcome for the target variable is only observed when a person participates in the survey. This is modelled in (3.3).

\[ Y_i^* = \pi X_i + \mu_i \]
\[ Y_i = \begin{cases} Y_i^* & \text{iff } I_{i4} = 1 \\ - & \text{else} \end{cases} \] (3.3)

The distribution of the error terms is again assumed to be multivariate normal, see equation (3.4).

\[
\begin{pmatrix}
\varepsilon_{1i} \\
\varepsilon_{2i} \\
\varepsilon_{3i} \\
\varepsilon_{4i} \\
\mu_i
\end{pmatrix}
\sim N_5 \left( \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \
\begin{pmatrix}
1 & \rho_{12} & \rho_{13} & \rho_{14} & \rho_{15} \\
\rho_{21} & 1 & 0 & \rho_{24} & \rho_{25} \\
\rho_{31} & 0 & 1 & \rho_{34} & \rho_{35} \\
\rho_{41} & \rho_{42} & \rho_{43} & 1 & \rho_{45} \\
\rho_{51} & \rho_{52} & \rho_{53} & \rho_{54} & \sigma
\end{pmatrix} \right) \] (3.4)

Only those auxiliary variables are used in equations (3.1) that relate to the corresponding causes of non-response. This enables us to use the information of interest exactly where it adds explanatory power. For instance, information about the interviewer can be included when explaining participation or refusal but has nothing to do with a person being longtime ill or not.

Again we can estimate this model by maximum likelihood, see (3.5). When these models are estimated, the estimated parameters are used to estimate the mean of the target variable (2.1.1).

\[ L = \sum_{i=1}^n \prod_{j=1}^4 I_j P(Y_i = i; I_{i1} = 1; I_{i2} = 1; I_{i3} = 1; I_{i4} = 1 | X_i) + \]
\[ + I_{i1} I_{i2} I_{i3} (1 - I_{i4}) P(Y_i = i; I_{i1} = 1; I_{i2} = 1; I_{i3} = 1; I_{i4} = 0 | X_i) \]
\[ + I_{i1} I_{i2} I_{i3} P(Y_i = i; I_{i1} = 1; I_{i2} = 1; I_{i3} = 1 | X_i) + \ldots \] (3.5)
4. Future research: combination of mixed mode- and response models

In section 2 we present two models for a mixed mode strategy, a concurrent- and a sequential model. Additionally, in section 3 a response model to adjust for non-response bias is discussed. This response model makes a clear distinction between different causes of non-response and uses instrumental variables to explain the stages in the response process.

The mixed mode models and the response model are the basic ingredients for a general framework to combine data from different mixed mode designs, thereby accounting for the different response processes of these modes and making optimal use of all available information. Response models can simply be inserted into the mixed mode models wherever necessary or needed. In the concurrent design, this means response models are inserted at the leaves of the tree. In the sequential design it is somewhat more complicated. Extending the model with different types of non-response would imply that all sources of non-response proceed to a next mode. In general this is not true. One may for instance follow up non-contacts only. Insertion there depends on the follow up strategy chosen.

There are a number of important issues that need further research. First, the estimation procedure of the models needs to be worked out. Second, we need to find a strategy to select variables. Third, we need to extend the models to designs with unequal inclusion probabilities.

Estimation by maximum likelihood, as suggested in sections 2 and 3, is an option but the models can become very large, hence leading to high dimensional parameter spaces. Furthermore, the likelihoods cannot be written in closed form so that we need to resort to numerical methods making the estimation complex and burdensome. A possible alternative is Bayesian estimation of the models, see Albert and Chib (1993) and Groenewald and Mokgatlhe (2004).

The number of auxiliary variables is in general too large to enter them all in the models simultaneously. We, therefore, need a procedure to enter variables to each of the equations. Since there can be many such equations in a complicated mixed mode design, this is not straightforward.

In many cases samples are stratified based on the expected variances within strata. We cannot simply add the inverse inclusion probabilities as weights to the equations. Future research will be directed at efficient estimation schemes, strategies for the selection of variables, and at general sampling designs.
References


