Working paper 30

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Budapest, Hungary, 14-16 September 2015)

Topic (iv): Evaluation and feedback

Using the CURIOS Algorithm to Manage the Prioritization of CAPI Surveys

Prepared by Antoine Rebecq[1], Thomas Merly-Alpa[2], INSEE

I. Abstract

This paper presents the CURIOS (Curios Uses Representativity Indicators to Optimize Samples) algorithm (see Rebecq and Merly-Alpa [2015] for a more precise presentation, in french) used for the prioritization (as explained in Merly-Alpa [2014]) of CAPI surveys led at the French National Institute for Statistics and Economic Studies (INSEE). We can't use StatCan methods on CATI prioritization (see Laflamme [2009]) as the organization of CAPI surveys in France forbid on-the-fly changes. Our method is based on a two-wave sampling design, with some learning during the first wave and an adjustment of the sampling design in the second wave. The second-wave sample is computed by minimizing the expected value of a loss function, which is a linear combination of several factors related to the quality of the sample. Monte Carlo methods are used to compute the expected values.

The main objective of the algorithm is to minimize the dispersion of the weights taking account non-response phenomenon, which leads to more robust estimations, in particular when some calibration routine is applied to the data. Some others goals introduced in the minimization problem are linked to the quality of the final sample, using for instance the partial R-Indicators (introduced by Schouten et al. [2011]) to analyse the behaviour of respondants in the sample.

This algorithm has already been applied to a few household surveys in France, such as the 2014 Household Wealth survey ("Enquête Patrimoine 2014"). The algorithm can be used for any CAPI survey and the objectives can be designed to approach any quality goal set for the sample.

II. The Prioritization of CAPI Surveys

A. Prioritization of Surveys

The term "Prioritization" usually refers to the process of adjusting the call order during the collection of a CATI survey (see for instance, the set of techniques used at StatCan in Laflamme [2009]). These methods are based on indicators linked to some sort of balance of the set of respondants within the sample. At StatCan, the indicator used gives insight on the expected precision of the estimator of a variable of interest (Beaumont et al. [2014]). We could also think about the method using R-indicators

[1] antoine.rebecq@insee.fr
[2] thomas.merly-alpa@insee.fr

which was exposed in Merly-Alpa [2014] during the 2014 Work Session on Data Editing.

However, we can't actually use any of these methods for our surveys, because most surveys led by INSEE in France are CAPI, i.e require the physical presence of an interviewer. This implies a much more complex organization. In fact, we have logistic issues to deal with: we need the prioritization to be done with some respect to the geographic distribution of the household surveyed, and we need to allow time after any change in the list of household affected to an interviewer, because the interviewers need time to investigate locations and set up dates for meetings. As shown in Figure 1, the data collection process starts very slowly in the first few weeks, and accelerates after: this is due to the time needed for the interviewer to investigate households. Stopping the survey process at a wrong time might lead to a loss of data and therefore we need to be careful with any change in the list of households given to an interviewer.

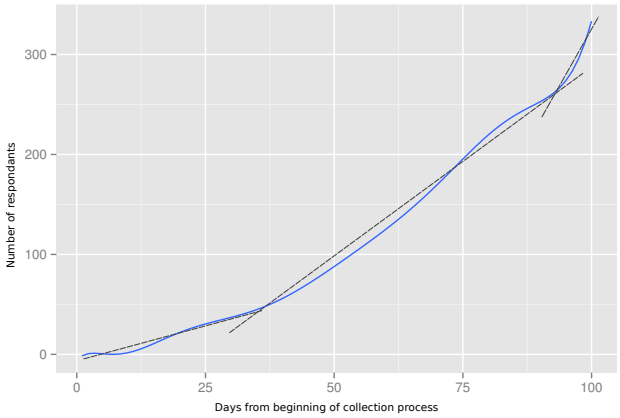

FIGURE 1. Collect rate as a function of time in a french survey.

These issues led us to abandon the idea of an on-the-fly prioritization, and we decided to investigate others means to achieve this goal such as adaptive sampling.

B. Two-Wave Sampling Design

As explained before, we can't manage on-the-fly prioritization. We then decided to organize our sampling process in order to include some kind of adaptive sampling. This process is similar to the prioritization of calls in CATI surveys, because it means that the interviewers are going to focus on some groups of households which will be over-representated in the samples, but it won't be as reactive as the original prioritization method. In order to adapt the sample to the data describing the collection process, we need to split the sample in multiple waves. Each wave lasts a fair amount of time, so that the process described in figure 1 is not troubled. The idea behind this subdivision is to learn about the process of response and collect during the first wave in order to draw a second-wave sample using all this information.

Our algorithm CURIOS (Curios Uses Representativity Indicators to Optimize Samples) aims at taking into account all the data gathered during the previous waves of the surveys (for instance the first one when we're dealing with a two-wave sampling design, which is the most frequent case) in order to draw a new sample for the next wave. The method used is kind of similar to the prioritization techniques previously exposed, because our algorithm starts by checking every group of people characterized by the variables given to the model in order to find the underrepresentated ones. Then the allocation of

every group is computed in order to fit some constraints and to more or less match the initial sampling method. The exact computation is described in the section below.

III. The CURIOS Algorithm

A. Objectives of the Algorithm

The CURIOS algorithm tries to minimize a loss function on the total sample by modifying the structure of the sample drawn for the second wave. The minimization program writes:

$$\arg\min_{S_2} \; l(S) = \mathbb{E}\left[\Sigma(w_{NRA}) + \lambda_1 \cdot \Gamma(S) + \lambda_2 \cdot D(S)\right]$$

where S_2 is the sample drawn in second wave, S is the total sample, w_{NRA} a vector of non-reponse adjusted weights and the λ are positive parameters. The sample S_2 is selected using **unequal probability sampling**, so that all combinations of units and associated weights can be explored during the optimization phase (see B). Functions Σ, Γ and D refer to the objectives of the algorithm which are described below:

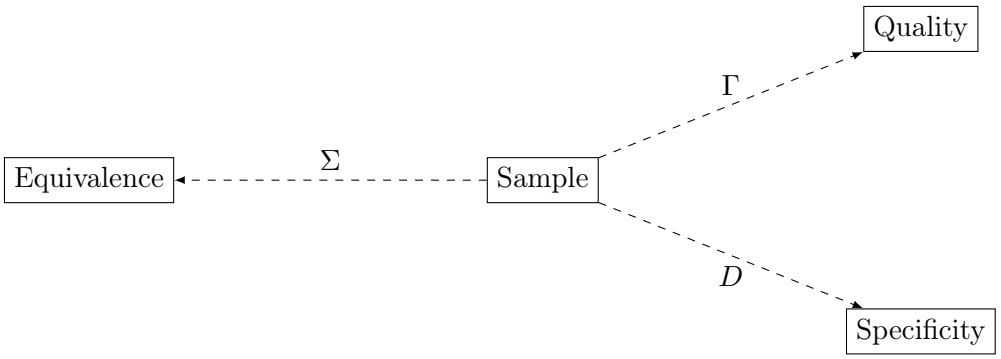

FIGURE 2. Objectives of the CURIOS algorithm.

These three objectives aren't equivalent, and pull the optimum sample in opposite directions. The next paragraphs will specify the role and nature of each objective.

A.1. *Equivalence.* The philosophy behind the concept of **Equivalence**, which is the most important objective of the algorithm, is that every household has to be as important as any other in order to compute estimations on a general population (see Rebecq and Merly-Alpa for a much more detailed discussion on this objective, in french). That's why the main goal of this objective is to reduce the dispersion of the impact of the characteristics of an household on the global result. In the context of sampling theory, this means that we want to minimize the dispersion of the sampling weights of the selected units. As we're dealing with a major issue of non-response, we focus on the non-response adjusted weights, which are computed using reweighting methods such as HRG (Homogenous Response Groups).

The objective isn't only used to match some kind of ideal scenario because of some beliefs we might have. The minimization of the dispersion of the sampling weights is a classic goal of sampling theory, because it leads to more robust estimation, in particular when the survey concern a wide spectrum of topics. Moreover, the calibration routine applied to the data in order to balance the sample regarding

some margins is much more efficient when there isn't any outlier with a larger weight than any other. And even if our study is focused on the estimation of only one parameter of interest, we should keep in mind that econometrical estimations might be done on our sample data. The precision of these estimation is directly linked to the dispersion of the sampling weights of the unit (see Solon et al. [2015] for a wider discussion on econometrics and sampling weights).

A.2. *Quality.* The **Quality** objective is based on some kind of balance between the respondants to the survey. For instance, we might want them to match the characteristics of the larger population in order to be balanced in some X variables which might explain the non response or the variable of interest. We took a different approach here, following Schouten et al. [2011] and using R-Indicators and (mainly) partial R-indicators to monitor the sample's balance. We can see Merly-Alpa [2014] for a more precise introduction on the R-indicators in the context of the prioritization of surveys, but let's recap the definitions of these indicators, where θ_i is the response propensity, i.e $\theta_i = \mathbb{P}(i \in R)$:

Definition 1. *The **R-indicator** is a measure of lack of association between reponding and auxiliary variables :*

$$R(\theta) = 1 - 2 \cdot \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\theta_i - \bar{\theta})^2}$$

Partial R-indicators measure the influence of one particular variable on the representativeness. There are two kinds of partial R-indicators : unconditional partial R-indicators measure the contribution of single variables to a lack of representativeness, and conditional partial R-indicators measure the contribution of single variables to a lack of representativeness *given the other variables* X used in the non-response model.

Definition 2. *The **unconditional partial R-indicator** associated to the variable Z with H strata measures the contribution of Z to a lack of representativeness, and is based on outer variance :*

$$R_U(Z) = \sqrt{\sum_{h=1}^{H}\frac{N_h}{N-1}(\bar{\theta}_h - \bar{\theta})^2}$$

*where N_h denotes the size of stratum h, and $\bar{\theta}_h$ the average response propensity within strata h. The **conditional partial R-indicator** associated to Z measures the contribution of single variables to a lack of representativeness given the set of other variables stratified in J different strata. The conditional partial R-indicator is based on the inner variance of this stratification :*

$$R_C(Z) = \sqrt{\frac{1}{N-1}\sum_{j=1}^{J}\sum_{i\in U_j}(\theta_i - \bar{\theta}_j)^2}$$

We can also defined the unconditional partial R-indicator associated to the modality h of the variable Z, which can be positive or negative, depending on the stratum h to be over-represented or under-represented.

The values of the partial R-indicators associated to a variable Z allow us to know how much this variable is useful in the analysis of the representativeness of the sample ; the modalities related to unconditional partial R-indicators order strata by priorization usefulness. Our Quality objective is then to reduce the higher values of modalities related unconditional partial R-indicators and raise the negative ones, in order to have no group left to prioritize. The goal is to achieve an R-Indicator of 1, which means that anyone has the same response propensity θ_i.

A.3. *Specificity.* Surveys are often designed to achieve specific goals (such as precise estimations on specific variables of interest). That's why we included the **Specificity** objective to our algorithm. CURIOS will try to find the sample S_2 which is the closest to the original sample, with respect to the others objectives. Any distance between allocations obtained from CURIOS and from the initial sampling design can be used for this purpose. In the example developed in section IV - the 2014 Household Wealth survey - this objective wasn't directly included, but it was taken into account because we respected the structure of the sampling by keeping it stratified (inclusion probabilities within the strata were equal).

B. Search for the optimum

B.1. *The over-representation vector ϵ.* In the general case, the sampling design for the second-wave sample S_2 is (fixed size) unequal probability sampling. We define the selection probabilities for the unequal sampling design from the probabilities of the "regular" sampling design, multiplied a vector ϵ called the **over-representation vector**:

$$\pi^{wave\ 2} = \epsilon \cdot \pi \tag{1}$$

where π is the vector of selection probabilities for the second-wave sample if we applied the exact same sampling design than in the first wave. This allows us to understand easily which units or profiles were over- or under-represented (prioritized or not) in the final optimal sample derived by CURIOS. For some values of $\epsilon \in \mathbb{R}^{+*}$, some components of $\pi^{wave\ 2}$ might not be in $[0, 1]$. If this happens, vector $\pi^{wave\ 2}$ is reshaped proportionally until all its components are valid probabilities.

In practice, when we draw the second-wave sample, we don't start over the whole sampling process from the sampling frame (mostly because it would be very ineffective as French surveys very often use two-degree sampling). Instead, we draw a much larger first-wave sample than needed, so that we can keep a "stock" of units, from which the second-wave sample is drawn (two-phase sampling).

B.2. *Simulating the data collection process.* Given an ϵ vector, the data collection process for the second wave is simulated using Monte-Carlo methods, using a model of the response phenomenon. Any modeling or learning technique can be used to create the model, although logit regression is often used in practice, mainly because feature selection is very important when designing a non-reponse model. The model chosen can also be the same that the one used for the R-indicators, assuming the model for wave 1 is still valid for wave 2. Non-reponse adjustment techniques are used (for example Homogeneous Response Groups using the same non-reponse model that was used for the simulations) to compute the NRA weights. Final sampling weights are computed using a formula by Ardilly (Ardilly [2006]):

$$w_{final} = \frac{n_1}{n_1 + n_2} w_1 + \frac{n_2}{n_1 + n_2} w^{wave\ 2}$$

where n_i is the sample size of wave i.

B.3. *Searching and analyzing minima.* Simulations described in section B.2 are done numerous times to provide a stable estimation for the loss function given a certain ϵ. In order to find optimal values of ϵ, we search for minima using linear optimization algorithms, triggered from various points (*grid search*) so we don't let the algorithm get caught on a high local minimum. Empirical evidence show that the loss function is smooth enough to yield good performance from classic linear optimization algorithms. However, estimation of l is done by Monte-Carlo, which means optimization has to be done on $\hat{l}$, which is not as smooth. We thus use the Nelder-Mead algorithm (Nelder and Mead [1965]), which performs well on noisy functions.

One a minimum is found, we can compute $\hat{l}_{min}$. Keeping in mind that $\hat{l}_{min}$ is obtained with some variance, we use a sort of "confidence interval" to decide whether choosing this sampling design for S_2 seems worth it:

$$\hat{CI}(\hat{l}_{min}) = \left[\hat{l}_{min} - 2\hat{\Sigma}_{MC}(\hat{l}_{min}); \hat{l}_{min} + 2\hat{\Sigma}_{MC}(\hat{l}_{min})\right] \tag{2}$$

If $l_1 = l(\epsilon = 1)$ is included in $\hat{CI}(\hat{l}_{min})$, then we'd probably be better off sticking to the original sampling design.

In practice we expect to find several local minima. Some of these minima will not be "significant" regarding the confidence interval defined in the above paragraph. The remaining minima define possible scenarios for the second-wave sampling design (see section IV). We recommend choosing between them not only by comparing the values of predicted $\hat{l}$, but by analyzing thourougly the proposed samples. Keeping in mind that the CURIOS algorithm relies on model-based estimations, should we end up with a scenario leading to a remarkably low $\hat{l}$ but a very high distorsion of the sample structure compared to wave 1, it might be much more careful to choose a scenario leading to a slightly higher $\hat{l}$, but with lower distorsion of the sample structure.

C. A Brief Theoretical Study of CURIOS

In this section we'll try to give a short analysis of some mathematical results linked to the CURIOS algorithm (see Merly-Alpa and Rebecq [2015]). As it is very hard to analyse the whole minimization problem, we focus on a smaller part of the issue by considering a two-strata population whose size is $N = N_1 + N_2$ with (known) uniform response probabilities ρ_i within each strata. We are interested in a variable X which is more heterogeneous in one strata, and we want to draw a sample of size n using the CURIOS algorithm to determine the value of the allocations n_1 and n_2 in both strata. In this particular context, the minimization program can be rewritten as:

$$n_f^1 = \underset{n_1}{\operatorname{argmin}} \quad \mathrm{Disp}(\mathrm{p}_{\mathrm{CNR}}^{\mathrm{k}}) + \lambda\, \mathrm{Dist}((\mathrm{n}_1, \mathrm{n} - \mathrm{n}_1), (\mathrm{n}_{\mathrm{init}}^1, \mathrm{n}_{\mathrm{init}}^2)) \tag{3}$$

This equation leaves out the Quality term, because we assumed that the response probabilities were uniform within the strata. Also, we don't have to compute expected values here, as we suppose response rates are fixed and known. The Specificity term used here is the distance from the initial allocation $(n_{\mathrm{init}}^1, n_{\mathrm{init}}^2)$, which is the Neyman allocation taking non response into account:

$$n_{\mathrm{init}}^i = n \frac{\frac{N_i S_i}{\sqrt{\rho_i}}}{\sum_{i=1}^{2} \frac{N_i S_i}{\sqrt{\rho_i}}} \tag{4}$$

The main parameter which has to be set in (3) is the λ, which can be interpretated as the importance attributed to the Specificity objective relatively to the Equivalence one. We can easily see that if we choose $\lambda = 0$, the sample will be only drawn from one strata, while if we choose $\lambda \to +\infty$, we'll end up with the initial allocation. Therefore, we have to find a good value in this window. We decided to try to use a λ such as the optimality of the Neyman allocation in term of variance of an Horvitz-Thompson estimator of the variable of interest X will not be altered much. We derive the following theorem:

Theorem 1. *Let $V(\lambda)$ be the variance of the Horvitz-Thompson estimator of the variable of interest X which is obtained by the allocation n_1, n_2 minimizing equation with the chosen λ. Under structural hypothesis on the population, $V(\lambda)$ is decreasing in λ and his second derivative $V''(\lambda)$ has a maximum within $]0, +\infty[$ which is called λ_{opt} the torsion point of V.*

Figure 3 shows that there is a wide range of acceptable values for λ. The goal of the theorem is to prove the existence of the optimal one λ_{opt} which is the one the further left on the flat part of the curve, in order to minimize the dispersion of the sampling weights.

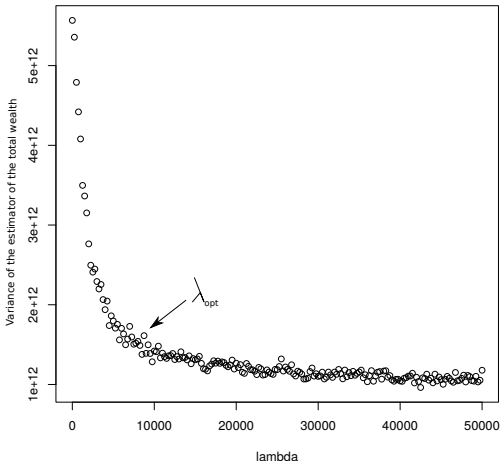

FIGURE 3. Empirical form of $V(\lambda)$.

IV. Application to the 2014 Household Wealth Survey in France

A. Context

Just like many surveys aiming at estimating a quantitative variable whose distribution is heavy-tailed, the 2014 Household wealth survey uses a stratified sampling design. The sample is also divided in two sub-samples (which are drawn from two different sampling frames):

(1) the “standard" sub-sample
(2) the “non-standard" sub-sample, where the wealthiest individuals are selected with higher priority

Each of these two sub-samples are stratified. As we explain further in section B, we focus here on the “standard" sub-sample. Its strata are:

- Senior citizens
- Self-employed
- Individuals whose revenue is mostly property income
- Executives
- Others

Allocation between these strata are determined based on expected response rate (in the same logic as the Neyman allocation taking non-response into account (4)) and dispersion of wealth measured on previous Household wealth surveys.

Due to internal organization constraints, the conditions of the data collection process were poor in the region that includes Paris and its suburbs. As this region is very populous and holds a high diversity of situations regarding wealth and estate, these conditions were excpected to lead to high-variance estimations. It was thus decided to divide the data collection process in Île-de-France into two waves,

and use the CURIOS algorithm, which was already proven effective on simulations (Merly-Alpa [2014]). The second-wave sample contains **820 units**.

B. Results

The sample S_2 is computed using the CURIOS algorithm. As explained in section III.A the "specificity" objective was not directly included as an optimization parameter.

B.1. *Equivalence objective.* In the context of the Household Wealth Survey, the "equivalence" objective should in fact relate to units which belong to the same stratum. Indeed, it makes sense (especially for econometric studies, as explained in Rebecq and Merly-Alpa and in section III. A) that units with similar profiles should have close weights. However, the survey is led jointly with the European Central Bank, who imposes that variances of various estimators can be computed using bootstrapped weights. This can only be done by keeping the original sampling design within each stratum. Only the allocations between strata can be modified. Thus, weights can't be modified within strata, and the "equivalence" objective chosen is the one described in section III. A. Trying to minimize the dispersion of the weights for the sub-sample containing the welthiest individuals wouldn't make much sense, as they contribute very much to the quality of the estimator of the total wealth. Consequently, our prioritization only focuses on the "standard" sub-sample.

B.2. *Quality objective.* At the end of the data collection process of the first-wave sample, we model the fact of being or not respondant to the survey using a logit regression :

$$\begin{aligned} \text{respondant} \sim \text{city category} + \text{stratum} + \text{house owner} \\ + \text{type of dwelling} + \text{sex of respondant} + \text{dwelling surface (categorical)} \end{aligned}$$

and we compute the R-indicators yielded by the model (table 1). All variables are self-explanatory, except for "city category", which consists in a 4-category clusterization of cities in the region Île-de-France, introduced so that the model could contain a geographic parameter.
As we can see, the R-indicator is very high (0.96), which means we can't really detect any imbalance in the sample as a whole, mostly because the sample is too small and the response rate is low. However, we still keep the "quality" objective in our optimization because:

(1) We can detect significant imbalance on a few modalities studied separately (see table 1)
(2) Trying to keep the R-indicator close to 1 ensures that the final sample won't be imbalanced

C. Final allocation

Using linear optimization and grid search, CURIOS found two "significant minima" (as defined in section III. B), which gives us two scenarios to choose from. Final allocations for the two scenarios are listed in table 2.
The first scenario is expected to reduce the dispersion of NRA weights by **3.5%** while the second scenario is expected to reduce it by **7.5%**. However, the final allocations given by the second scenario are much more different from the initial allocation than the allocations in the first scenario. If our model were inaccurate (which is somewhat likely because the sample size is quite low), choosing the second scenario could yield poor estimators. We thus make a conservative choice, by retaining **scenario 1**.

We can see that the two scenarios reduce allocations for 4 of 5 strata and increase allocation for the remaining stratum ("Others"). This is expected since we know from the preliminary analysis of the data collection process (table 1) that profiles that need to be prioritized are people who do not own their

R-indicator
0.96

Variable	R-indicator (× 100)
City category	0.66
Stratum	0.83
House owner	**0.98**
Housing type	**1.16**
Sex of respondant	0.05
House surface (categorical)	**1.30**

House owner	R-indicator (× 100)	Pct. responding
No	-1.31	30%
Yes	1.85	38%

Type of dwelling	R-indicator (× 100)	Pct. responding
Flat	-1.00	34%
House	2.48	41%

Dwelling surface	R-indicator (× 100)	Pct. responding
Less than $40m^2$	-2.1	31%
Between 40 and $100m^2$	0.7	34%
Greater than $100m^2$	2.0	37%

TABLE 1. R-indicators for the "standard" sub-sample of the 2014 Household Wealth survey

Strate	Initial Allocation	Scenario 1	Scenario 2
Senior citizens	81	65	44
Self employed	34	25	16
Property Income	20	14	10
Executives	183	151	108
Others	158	221	298

TABLE 2. Allocations for the second wave sample, per stratum

home, live in flats or whose homes are rather small. These profiles are mostly found in the "others" stratum.

References

P. Ardilly. *Les techniques de sondage*. Editions Technip, 2006.

J.-F. Beaumont, C. Bocci, and D. Haziza. Une procédure adaptative pour la priorisation des appels téléphoniques lors de la collecte des données. *8eme colloque francophone de Sondages*, 2014.

F. Laflamme. Experiences in assessing, monitoring and controlling survey productivity and costs at statistics canada. 2009.

T. Merly-Alpa. Using R-indicators to monitor household surveys and prioritize data collection: an application to the 2010 household wealth survey in France. *Work Session on Statistical Data Editing*, 2014.

T. Merly-Alpa and A. Rebecq. L'algorithme CURIOS pour l'optimisation du plan de sondage en fonction de la non-réponse. *Journées de la Statistique*, 2015.

J. A. Nelder and R. Mead. A simplex method for function minimization. *The computer journal*, 7(4): 308–313, 1965.

A. Rebecq and T. Merly-Alpa. Pourquoi minimiser la dispersion des poids en sondage ? *preprint*.

A. Rebecq and T. Merly-Alpa. Algorithme CURIOS et méthode de "priorisation" pour les enquêtes en face-à-face. Application à l'enquête Patrimoine 2014. *JMS*, 2015.

B. Schouten, N. Shlomo, and C. Skinner. Indicators for monitoring and improving representativeness of response. *Journal of Official Statistics*, 27(2):231–253, 2011.

G. Solon, S.J Haider, and J.M Wooldridge. What are we weighting for? *Journal of Human Resources*, 50(2):301–316, 2015.