WP.
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Paris, France, 28-30 April 2014)

Topic (i): Selective editing / macro editing.

USING R-INDICATORS TO MONITOR HOUSEHOLD SURVEYS AND PRIORITIZE DATA COLLECTION : AN APPLICATION TO THE 2010 HOUSEHOLD WEALTH SURVEY IN FRANCE

Prepared by Thomas MERLY-ALPA (thomas.merly-alpa@ens-lyon.org), INSEE

I. ABSTRACT

The French National Institute for Statistics and Economic Studies (INSEE) realizes lots of face to face surveys drawn from a master sample. Each primary unit is managed by one or many pollsters. We may face logistical problems such as a loss of response rate, uniformely in all primary units or in some particular areas. In this particular case, we generally only have the capacity to collect a few more questionnaires in these primary units. So we need prioritization techniques to select the households that need to be investigated in order to reduce as far as possible the non-response bias and loss of precision induced by this fall in collection rates in such units.

To fulfill this goal, we use R-indicators, for indicators of representativeness, which were introduced by Shouten & al. in 2009 [1] to analyse the representativeness of the sample during the survey. The concept of representativeness used here isn't related to the subject studied or auxiliary variables but only to dispersion within the sample of response propensities, based on how the survey is monitored. This means that the method exposed here can be widely adapted to any survey.

The R-indicators offer us only a measure of a global representativeness, which can't be used to monitor directly the survey. Following [2] we consider partial R-indicators, both conditionally and unconditionally, which represent within and between dispersion of the response propensities. The different subgroups are based on the variables used in the logit model to estimate the response model. These R-indicators offer us a way to differentiate population in terms of representativeness in order to prioritize the collect of data : as shown in [6], the groups which have negative unconditional partial R-indicators and huge (in absolute value) partial R-indicators are the one to be focused on if we want to greatly increase global representativeness, even if it is at the cost of a loss in response rate.

The purpose of this presentation will be to apply these methods to the data which were collected for the 2010 Household Wealth survey ("Enquête Patrimoine 2010"). First, we calculate the R-indicators and partial R-indicators of samples collected at some point, which show the natural evolution of a survey without any kind of prioritization. Then, we simulate the vacancy of pollsters attributed to some geographical areas and compare some prioritization strategies and their effects on the quality of estimation of the distribution of wealth in France.

II. DEFINITIONS

A. R-INDICATORS

R-indicators [1] are designed to be a measure of the representativeness of a sample. They are based on the estimated response propensities and aren't directly linked to the interest variables, which means they are useful in most scenarios. The R-indicator is a measure of the quality of the data collection in regard to auxiliary variables, and is closer from 1 when the sample is balanced, which means we are close to a CMAR (Completly Missing At Random) situation [3].

We consider a population of N households and define the response propensity $\theta_i = \mathbb{P}[r_i = 1 \mid s_i = 1]$ for each household i. The average response propensity is denoted $\bar{\theta} = \frac{1}{N}\sum_{i=1}^{N}\theta_i$. We now consider the following definition.

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

$$R(\theta) = 1 - 2S(\theta)$$

in which :

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

We have the following inegality about the dispersion of response propensities :

$$S(\theta) \leq \sqrt{\bar{\theta}(1-\bar{\theta})} \leq \frac{1}{2}$$

which states that the R-indicators is between 0 and 1, 1 being the perfect case of all equal response propensities. This scenario doesn't imply that all the households reacts equally to the survey, but means that the monitoring of data collection was calibrated to allow more times and attempts to the least compliant (or reachable) units.

In most of the cases, the response propensities θ_i are unknown. We use logit models based on auxiliary variables to estimate them. Moreover, we only have information on sampled units - we denote $s_i = 1$ when a unit i is sampled, 0 otherwise, and Π_i the design weight.
Here is an estimator of the R-indicator :

$$\hat{R}(\theta) = 1 - 2\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\frac{s_i}{\Pi_i}(\hat{\theta}_i - \hat{\bar{\theta}})^2}$$

with the following notation :

$$\hat{\bar{\theta}} = \frac{1}{N}\sum_{i=1}^{N}\hat{\theta}_i\frac{s_i}{\Pi_i}$$

The R-indicator is linked to boundaries on non-response bias : if we denote Y the interest variable and $\hat{\bar{Y}}_{HT}$ its Horvitz-Thompson average estimator, we have then using Cauchy-Schwarz :

$$\left|B(\hat{\bar{Y}}_{HT})\right| \leq \frac{1 - \hat{R}(\theta)S(Y)}{2\bar{\theta}}$$

That implies that for $\gamma > 0$, as long as $\hat{R}(\theta) \geq 1 - 4\hat{\hat{\theta}}\gamma$, the non-response bias will satisfy $B \leq \gamma$.

B. PARTIAL R-INDICATORS

Partial R-indicators [2] measure the influence of one particular variable on the representativeness. There is two kind of partial R-indicators : unconditional partial R-indicator 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, the one used in the non-response logit model. They are both based on response propensities and strongly linked to global R-indicator. They can be seen as inner and outer variance from the decomposition of the dispersion of response propensities in regard to modalities of a given variable.

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 unconditional partial R-indicator is estimated by

$$\hat{R}_U(Z) = \sqrt{\sum_{h=1}^{H} \frac{\hat{N}_h}{N}(\hat{\hat{\theta}}_h - \hat{\hat{\theta}})^2}$$

in which :

$$\hat{N}_h = \sum_{i=1}^{N} \frac{s_i}{\Pi_i}\mathbf{1}_{z_i=h} \quad \text{and} \quad \hat{\hat{\theta}}_{\mathrm{h}} = \frac{1}{\hat{\mathrm{N}}_{\mathrm{h}}}\sum_{\mathrm{i}=1}^{\mathrm{N}} \hat{\theta}_{\mathrm{i}}\frac{\mathrm{s}_{\mathrm{i}}}{\Pi_{\mathrm{i}}}\mathbf{1}_{\mathrm{z_i=h}}$$

We can also defined the unconditional partial R-indicator associated to the modality h of the variable Z, whose estimator is :

$$\hat{R}_U(Z,h) = \sqrt{\frac{\hat{N}_h}{N}}(\hat{\hat{\theta}}_h - \hat{\hat{\theta}})$$

This estimator can be positive or negative, depending on the stratum h to be over-represented or under-represented.

Definition 3. *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 denoted X^-, assumig this set is stratified in J different strata. The conditional partial R-indicator is based on the inner variance of this stratification :*

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

The conditional partial R-indicator is estimated by :

$$\hat{R}_C(Z) = \sqrt{\frac{1}{N-1}\sum_{j=1}^{J}\sum_{i \in U_j}\frac{s_i}{\Pi_i}(\hat{\theta}_i - \hat{\hat{\theta}}_j)^2}$$

There is also an estimator of the conditional partial R-indicator related to the modality h of variable Z :

$$\hat{R}_C(Z,h) = \sqrt{\frac{1}{N-1}\sum_{j=1}^{J}\sum_{i\in U_j}\frac{s_i}{\Pi_i}\mathbf{1}_{z_i=h}(\hat{\theta}_i - \hat{\bar{\theta}})^2}$$

This estimator is always positive.

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 ; once the variables are selected, we just have to rank the modalities related unconditional partial R-indicator to choose strata to be priorized.

Precision and bias of the estimation of R-indicators and partial R-indicators will not be considered here. Researchs such as [5] indicate that in large samples the bias of estimation can be ignored. Variance of R-indicators is an open topic ([7],[5]) which is not the main purpose of this article.

III. THE 2010 HOUSEHOLD WEALTH SURVEY IN FRANCE

A. DESCRIPTION OF THE SURVEY

The objective of the french 2010 Household Wealth Survey is to describe the household situation with regard to financial, real-estate and professional assets as well as outstanding debts. It is used to observe the distribution of assets as well as the different asset holding patterns across households. It also provides particularly comprehensive information on factors accounting for wealth accrual over the life cycle:

- family and professional biography;
- gifts and inheritances;
- income and financial situation;
- motives for the holding (or non-holding) of the different assets;
- comprehension of the processes involved in the creation and transfer of economics assets.

The 2010 survey also provides insights into non monetary dimensions of wealth, including the cultural, social and family dimensions of capital. Similar surveys are conducted in other European countries for the purpose of international comparisons.

The 2010 household Assets survey was conducted between October 19, 2009 and February 26, 2010. The collection was computer-assisted. The survey has been conducted approximately every six years since 1986. Data for the 2010 Assets survey was collected from a little less than 18,000 households (denoted the **standard sample**). A specific approach protocol was tested, in order to determine whether it could improve the survey response rate. An additional sample (denoted **non-standard sample**) of 3,000 households was drawn in order to have more respondants in the last decile of wealth distribution. Income data was also partially collected through matching with tax service data.

B. REPRESENTATIVENESS

Estimations of the R-indicators were done at different stages of the survey, using SAS. They were computed separately for the two samples because the variables used to deal with the non-response were different. The results show that the R-indicators seems to decrease over time in the non-standard sample (the one most relevant to estimate wealth distribution), and decreasing until the third month in the standard sample. This effect might be explained by the release of a reserve sample at this moment.

R-indicator	1 month	2 months	3 months	end of survey
Standard sample	0.824	0.777	0.756	0.819
Non-std sample	0.851	0.757	0.733	0.721

In order to calculate the unconditional partial R-indicator and the conditional partial R-indicators, we used a stepwise version of the logistic regression to estimate the response propensities. This was decided because the classical way of dealing with non-response in this survey used lots of variables ; that leads to irrelevant definition of the conditional partial R-indicator, because the stratification was too thin. The variables selected by the stepwise logistic regression are classical household attributes (household type, dwelling type. . .), information on their property income and the strata in which the household has been classified for the sampling. These stratum are different for the two samples : for the standard one, it is stratum based on data from previous surveys : farmers, senior citizens. . . ; for the non-standard one, the 4 stratum were built upon a classification algorithm in order to match the sample, because this was the first time this sample was drawn.

These partial R-indicators offer us a way to differentiate population in terms of representativeness in order to prioritize the data collection. We follow [6] and decided to select firstly variables with important partial R-indicators. Once this done, we select the groups with negative unconditional partial R-indicators, which means they are under represented, from the modalities of the selected variables. They are the one to be focused on if we want to greatly increase global representativeness.

IV. IMPACT OF A LOSS IN RESPONSE RATE

The main purpose of this paper is to evaluate the impact of a loss of response rate in a survey. This reflexion was launched after unusually large non response rates were observed in the 2013 edition of the annual survey on victimisation (10% more than the year before) in the context of a major change in the pollsters work system. We wanted to know if priorization with the help of R-indicators allows us to compensate for the loss of response rate in terms of precision of the estimations.

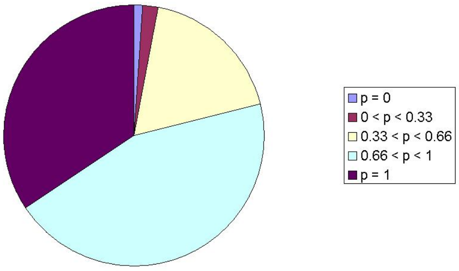

FIGURE 1. Distribution of the rates of loss p_k in primary units.

A. SIMULATING THE LOSS OF HOUSEHOLDS

In order to test this hypothesis, we decided to simulate this loss of response rate on the data collection of the 2010 household Assets survey after 3 months of collect. We denote for each primary unit k the rate of loss p_k which is calculated by :

$$p_k = \frac{2RR_{2013}}{RR_{2012} + RR_{2011}}$$

i.e the ratio between the bad response rate in 2013 and the average response rate in 2011 and 2012. The distribution of the p_k over the primary units is represented in Figure 1.

We carried out $N = 100$ simulations of the loss of response rate on the data. First, we needed to compute the unconditional partial R-indicator and the conditional partial R-indicators in order to select the groups which are to be prioritized. Using stepwise logistic regression, we obtained the response propensities associated to the new data collection, and we were able to compute the partial R-indicators. The average results obtained are shown in these tabulars. We denoted "Unc." and "Con." the unconditional partial R-indicator and the conditional partial R-indicators associated to the variable Z or its modalities.

Standard sample.

Variable	**Unc.**	**Con.**
Strata	7.62	0.20
Farmers	1.46	
Senior Citizens	-3.84	
Independants	-2.34	
Property Income	-1.42	
Middle Management	-3.13	
Others	4.89	
Type of Dwelling	3.33	0.07
Flat	-3.08	
House	1.27	
HLM[1]	1.13	0.05
No	-0.44	
Yes	1.04	
Type of Household	4.01	0.08
Full	-0.35	
Couple	-1.61	
Family	2.32	
Single parent	0.73	
Single	-2.73	

Non-standard sample.

Variable	**Unc.**	**Con.**
Strata	6.76	0.26
Rich City-dwellers	-1.75	
Personal Property	-0.17	
Real Property	0.10	
Others	6.53	
Type of Dwelling	9.93	0.14
Flat	-4.23	
House	8.99	
Type of Household	6.89	0.26
Full	2.54	
Couple	6.26	
Family	0.10	
Single parent	0.35	
Single	-1.33	
High Property Income	5.25	0.12
No	5.14	
Yes	-1.08	
Very High Property Income	5.02	0.12
No	4.76	
Yes	-1.60	

These results allowed us to select some of the groups which are going to be priorized for the simulation.

Standard sample	**Non-standard sample**
Senior Citizens	Rich City-dwellers
Middle Management	Flat
Independants	Single
Single	
Flat	

B. IMPACT ON THE R-INDICATOR

We ran the $N = 100$ simulations of loss of response rate ; then we add some households to the data collection. We have a few parameters to fix : the number n of households to be added per primary unit and the percentage x of primary units in which we prioritize households. We decided to study a few scenarios for x :

(1) We use prioritization for every primary unit.
(2) We concentrate our efforts on the few primary unit whose loss rate p_k verifies $p_k < 0.66$. This concerns 20% of the primary units.
(3) We select only the most impacted primary units in which $p_k < 0.33$. This concerns less than 3% of the primary units.

For all of these scenarios, we decided to compare two strategies. The first one is to choose at random n households in the non-standard sample (according to the survey design, this sample should be more crucial regarding the precision of the estimations). The second one is to choose in the non-standard sample up to n households which belong to prioritized groups, and if needed, add some households from prioritized groups of the standard sample to have n households. Then we compute the R-indicator in both cases. The results are displayed in these tabulars, with the best R-indicator bolded to ease lecture.

Scenario (1)

	Std Sample		**Non-std Sample**	
n	Prioritized	Random	Prioritized	Random
1	**0.7651**	0.7638	**0.6929**	0.6914
2	**0.7687**	0.7672	**0.6954**	0.6911
3	**0.7722**	0.7641	**0.6969**	0.6954
4	**0.7715**	0.7672	**0.6988**	0.6881
5	0.7666	**0.7672**	**0.7015**	0.6840
10	0.7495	**0.7696**	**0.7034**	0.6770

Scenario (2)

	Std Sample		**Non-std Sample**	
n	Prioritized	Random	Prioritized	Random
1	0.7624	**0.7637**	**0.6999**	0.6921
2	0.7665	**0.7666**	**0.7050**	0.6989
3	**0.7664**	0.7658	**0.7071**	0.6938
4	**0.7699**	0.7696	**0.7099**	0.6945
5	**0.7756**	0.7706	**0.7118**	0.6972
10	**0.7906**	0.7809	**0.7255**	0.7027

Scenario (3)

	Std Sample		**Non-std Sample**	
n	Prioritized	Random	Prioritized	Random
1	0.7615	**0.7618**	**0.6962**	0.6954
2	0.7621	**0.7626**	**0.7037**	0.6957
3	**0.7635**	0.7629	**0.7072**	0.6963
4	0.7641	**0.7658**	**0.7076**	0.6960
5	0.7649	**0.7667**	**0.7112**	0.6973
10	**0.7725**	0.7706	**0.7200**	0.7040

If we consider firstly the non-standard sample, which is mainly concerned by the procedure, we directly see that the prioritization method offers a better representativeness than the random selection. This also points out that increasing the number of primary units added has a positive impact on representativeness. The scenario (2) seems to obtain best results than the others.

Concerning the standard sample, the analysis is far more complex as the selection of households occurs only in this sample if the non-standard sample hasn't enough dwellings in the prioritized groups. We can see that prioritization isn't always the best method, but obtains always close results to the random selection. Moreover, increasing n can lead to saturation effects (the prioritized groups become over represented) which have negative impact on the representativeness. Regarding this sample, the scenario (2) is also the best one.

C. IMPACT ON THE PRECISION

We are now trying to evaluate the impact of prioritization routines on precision of the data. To this concern, we decide to focus on some key parameters such as the mean over french population of the gross estate (denoted here after "gross"), the net assets (denoted "net"), the financial assets (denoted "fin"), the real-estate (denoted "real") and the professional assets (denoted "prof"), i.e the value of compagnies and tools owned by the household. We tried to evaluate an extra term of variance related to the loss of response rate by considering the dispersion of the computed values $\hat{y}$ for each of the key indicators in the 100 simulations. These $\hat{y}$ were calculated using the same non-response correction as in the initial survey :

(1) We used an adapted score method [4], i.e a response propensity stratification, to evaluate the response propensities.
(2) We use marginal calibrations with data from others studies and administrative register to re-calibrate the weight of the households.

We ran three different simulations to evaluate the impact of priorization.

C.1. *PRIORITIZING AFTER DATA COLLECTION.* First, we want to compare a situation where the data collection is stopped at some point (for the purpose of our simulations, this will be after the third month of the survey) and the same situation with some more households being surveyed, chosen at random or according to R-indicators. This prioritization occurs only in the few primary unit whose loss rate p_k verify $p_k < 0.66$, following scenario (2) of paragraph 4.A. The next table shows the numerical standard deviation of the $\hat{y}$ in all this different cases.

	gross	net	fin	real	prof
Data Collection Stopped	3160	3082	618	928	2779
Random, $n = 5$	2708	2644	571	855	2425
Prioritized, $n = 5$	2795	2695	650	876	2359
Random, $n = 10$	2323	2136	568	922	1947
Prioritized, $n = 10$	2175	2056	615	861	1732

As we obvioulsy increase the response rate by surveying some more households, the diminution of this variance is clearly explained. But this table shows that choosing households according to R-indicators have a better effect than a random selection when $n = 10$, which leads to think that prioritization might be an efficient strategy as far as enough households are concerned.

C.2. *PRIORITIZING DURING DATA COLLECTION.* For the second case, we let the data collection finish during the last month in two different ways : we let the real survey occurs, or we choose to select the households in the most impacted primary units (following scenario (2) of paragraph 4.A), and let the data collection follow its path in the other primary units. This scenario of partial prioritization forces us to reduce the data collection in other primary units in order to be able to survey all the chosen households : in the simulations, we decided to shrink the number of surveyed households by 25%. This rate was chosen because the primary units selected represent one in five primary unit, so we might need 1 in 4 pollsters in the other areas to deal with the prioritization. The next table shows the standard deviation of the $\hat{y}$ in all these different cases.

	gross	net	fin	real	prof
Data Collection Finished	2175	2179	536	699	1904
Random, $n = 5$	2436	2270	539	766	1983
Prioritized, $n = 5$	2521	2337	552	741	2059
Random, $n = 10$	2327	2132	589	731	1870
Prioritized, $n = 10$	1967	1771	592	676	1504

In this study, the response rate in prioritization scenarios is lower than the final one :

n	Data Collection Finished	Random	Prioritized
5	50.48%	49.21%	49.07%
10	50.48%	50.14%	50.03%

This explains why in most cases the precision is lower than in the initial study ; but this highlights that the procedure of prioritization with $n = 10$ reduced significantly the additionnal term of variance, even with a lower response rate.

C.3. *PRIORITIZING IN EVERY PRIMARY UNIT.* The last study concerns the scenario (1) of paragraph 4.A, in which we use a prioritization routine in every primary unit. When we compare the results obtained with this method to the final data, we have the following standard deviations :

	gross	net	fin	real	prof
Data Collection Finished	2175	2179	536	699	1904
Random, $n = 5$	2712	2668	559	879	2460
Prioritized, $n = 5$	2768	2682	599	821	2418
Random, $n = 10$	2553	2431	516	892	2284
Prioritized, $n = 10$	2367	2223	520	763	1891

This procedure implies a loss of precision (as we have a higher additional term of variance) even when we prioritize $n = 10$ households. This is mostly linked to the global loss of response rate induced by this method (with lower rates than in C.2), but this also may be explained by a phenomenon of over representation ; as we focus on same groups in all primary units, these groups might become predominant.

V. CONCLUSION

We decided to use R-indicators and partial R-indicators to choose which groups of households to prioritize. Partial R-indicators offer us a better way to determine these groups than just looking at the response rate, given that they are strongly linked to the global R-indicator. The analysis done on the french 2010 Household Wealth Survey lead to select some groups who were under represented after the third month of data collection.

These informations were used in simulations in order to evaluate the impact on surveying more households, and how to select them. As we consider the effects of prioritization on representativity, we conclude that in this case, focusing our efforts on the most affected primary units would be more efficient. The analysis done on precision of some variables lead to the same conclusion, as a global prioritization implies a loss in precision.

One of the main results is that, given a minimum number of household selected, it seems always better to select them with the help of R-indicators than randomly. This means that the prioritization method based on partial R-indicators seems to be a good way to increase accuracy when facing low response rates.

Another idea was to study a prioritization during the data collection. We pursue the data collection in the least affected primary units, but at a lower rate in order to do some prioritization routine in the other primary units. This strategy appears to be useful to reduce the additionnal variance due to a loss in response rate, which means that even if we can't logistically survey more households at the end of the data collection, we could consider to do this priorization routine in the last month (or week, according to the length of the survey considered) with respect to the data collected so far.

All these results must be taken carefully as the simulation were made with some assumptions (such as the possibility to transfer pollsters from one primary unit to another) and that the results only concern an additional term of variance : the total variance might change otherwise as the design variance can vary within these scenarios. Some more research needs to be done on this aspect of the subject. We might also consider studying potential bias induced by these methods.

ACKNOWLEDGEMENTS

I'd like to thank the Survey Department of INSEE for their precious help to analyse data and compute these methods, and the Wealth Team of the Demographic and Social Statistics Directorate for all the information they gave me about the 2010 Household Wealth Survey.

References

[1] Schouten B., Cobben F., and Bethlehem J. Indicators for the representativeness of survey response. *Survey Methodol.*, 35(1):101–113, Jun. 2009.

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

[3] Rubin D. Inference and missing data. *Biometrika*, 63(3):581–592, Dec. 1976.

[4] David Haziza and Jean-François Beaumont. On the construction of imputation classes in surveys. *International Statistical Review*, 75(1):25–43, 2007.

[5] Shlomo N. and Schouten B. Theoretical properties of partial indicators for representative response. *Technical Report*, 176:169–189, 2013.

[6] Shlomo N., Schouten B., and de Heij V. Designing adaptative survey design with R-indicators. *New Techniques and Technologies for Statistics Conference*, 2013.

[7] Natalie Shlomo, Chris Skinner, and Barry Schouten. Estimation of an indicator of the representativeness of survey response. *Journal of Statistical Planning and Inference*, 142(1):201 – 211, 2012.