WEIGHTING NATIONAL SURVEY DATA

ELFE Survey:

Weighting national survey data

Thierry Siméon – First version. December 2019

Summary

Weighting method used in survey waves up to the age of 2 years 3

Weighting of data from maternity units 3

Weighting of the following survey waves: at 2 months, 1 year, 2 years 7

Why maintaining this method for the subsequent survey waves is problematic....... 8

The chosen weighting method: Simultaneous calibration....... 9

Comparison of the new method with the previously obtained results 10

Appendix 1: SAS software procedure....... 16

WEIGHTING NATIONAL SURVEY DATA

The aim of this note is to describe the weighting methods used for the data from the ELFE survey: both the method initially implemented to calculate weights up to the survey wave when the participating children were 2 years old, and a new method used beginning with the survey wave at the age of 3 ½ years.

This note opens by presenting all of the elements needed to understand this change in methodology. First, it outlines the weighting used in the previous survey waves (in maternity units and at the ages of 2 months, 1 year, and 2 years) to infer estimates for the population as whole from the data provided by the respondents.

For more detail on the results of the method implemented at earlier stages, see the following notes:

- Weighting inclusion
- Weighting 2 months
- Weighting 1 year*
- Weighting 2 years*

*= Not yet translated

The present note then sets out the main principles of the weighting method chosen for the subsequent survey waves, when the children were 3 ½ years old and thereafter.

For more details on the weighting method implemented at age 3 ½, please see:

- National Survey: Weighting at the age of 3 ½ years – Siméon

These notes are available on the pages for the corresponding survey wave on the PANDORA platform.

Note: Since the new method does not require specific knowledge on nonrespondents, as weights are estimated using direct calibration on the respondents' data alone, new weights for a representative sub-sample of the population can be produced on demand.

If you are analysing results from the ELFE surveys, do not hesitate to contact Thierry Siméon (thierry.simeon@ined.fr) who can guide you and generate the weights for you, allowing you to draw inferences from the analyses on your specific sub-population to the target population.

Etude Longitudinale
Française depuis
l'Enfance

WEIGHTING NATIONAL SURVEY DATA

Weighting method used in survey waves up to the age of 2 years

Weighting of data from maternity units

The weighting method for the data from maternity units is described in detail in the document "Weighting of the ELFE survey at inclusion". Here we will simply recall the main principles:

The infants included in the cohort were selected as follows: their date of birth was one of a selection of days in 2011, and they were born in one of a sample of maternity units in metropolitan France.

For the ELFE survey, the selection consisted of:

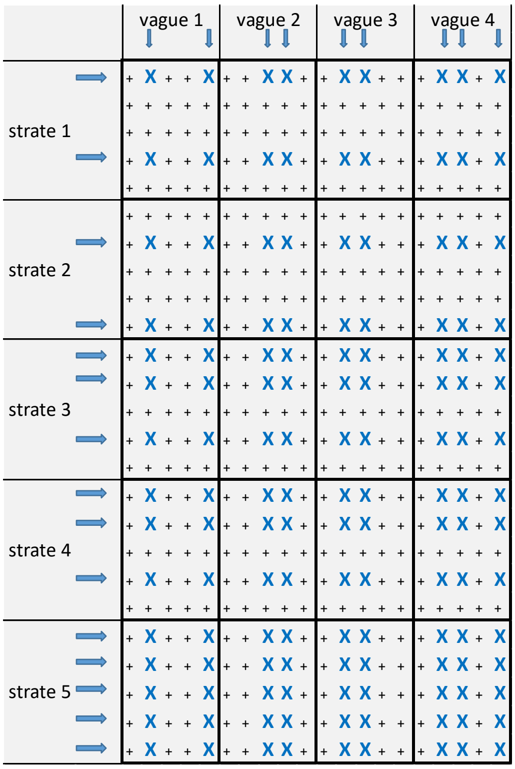
- A sampling frame consisting of the list of maternity units (public and private) in metropolitan France in 2008: 544 maternity units were catalogued. The stratification variable was the size of the maternity unit (by annual number of births). Five equal-sized strata were constructed. The sample included 349 maternity units.

Strates *g*	Nb d'accouchements par maternité en 2008	Taille dans la population N_g	Taille de l'échantillon n_g
1	[145-699[	108	28
2	[700-1009[	108	47
3	[1010-1418[	109	66
4	[1422-2187[	108	97
5	[2197-5215[	111	111
TOTAL		544	349

- From a set of days (all days in the year 2011). To represent each season, the 25 survey days (4, 6, 7 and 8 days) were divided into waves. For logistical reasons, these days were not drawn randomly, but fixed: from 1 to 4 April, from 27 June to 4 July, from 27 September to 4 October, and finally from 28 November to 5 December.

Vague *h*	Taille dans la population M_h	Taille de l'échantillon m_h
1	90	4
2	91	6
3	92	7
4	92	8
TOTAL	365	25

The selection of days was the same for each maternity unit (or vice versa: the sample of maternity units was the same for each selected day). The final sample consists of the intersection of the selected locations and the selected days. The selection used for the ELFE survey can thus be schematized as follows:



Given the independence of the selection of rows and columns, each "maternity x day" unit in the sample of infants is simply assigned an initial weight based on the effect of the sampling plan, $\frac{N_g M_h}{n_g m_h}$

		Vague *h*			
		1	2	3	4
Strate *g*	1	86,79	58,50	50,69	44,36
	2	51,70	34,85	30,20	26,43
	3	37,16	25,05	21,71	18,99
	4	25,05	16,89	14,63	12,80
	5	22,50	15,17	13,14	11,50

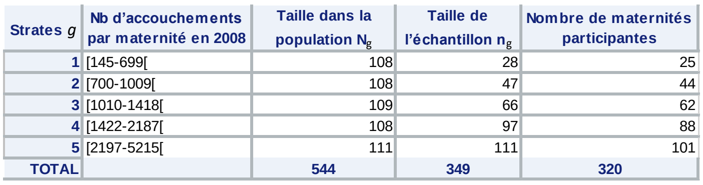
Initial weights drawn from the sampling plan

Two major causes of nonparticipation, at the level of the maternity unit and the day, were then analysed.

Some of the selected maternity units did not participate in the survey for various reasons. Among the 349 maternity units in the selected sample, 25 did not participate in any wave. Four further randomly selected maternity units were ultimately not invited. This makes a total of 29 nonparticipating maternity units to be taken into account in this phase.

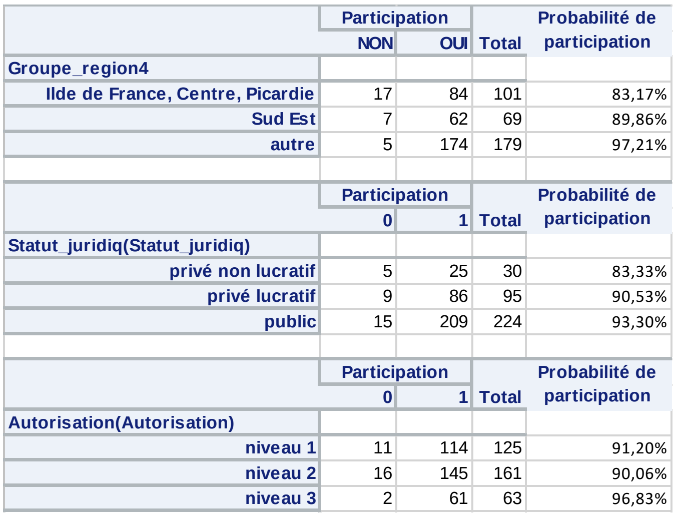
Other maternity units only participated in a subset of days. Of the expected 320 x 25 = 8,000 "maternity unit x day" combinations, only 7,741 took place.

The maternity units which refused to participate in the survey were distributed as follows:



Strates *g*	Nb d'accouchements par maternité en 2008	Taille dans la population N_g	Taille de l'échantillon n_g	Nombre de maternités participantes
1	[145-699[	108	28	25
2	[700-1009[	108	47	44
3	[1010-1418[	109	66	62
4	[1422-2187[	108	97	88
5	[2197-5215[	111	111	101
TOTAL		544	349	320

In addition to the stratum of the maternity unit, there are three further variables to characterize participating and nonparticipating maternity hospitals: region, level of medical specialization, and legal status.



	Participation			Probabilité de participation
	NON	OUI	Total	
Groupe_region4				
Ilde de France, Centre, Picardie	17	84	101	83,17%
Sud Est	7	62	69	89,86%
autre	5	174	179	97,21%
	Participation			Probabilité de participation
	0	1	Total	
Statut_juridiq(Statut_juridiq)				
privé non lucratif	5	25	30	83,33%
privé lucratif	9	86	95	90,53%
public	15	209	224	93,30%
	Participation			Probabilité de participation
	0	1	Total	
Autorisation(Autorisation)				
niveau 1	11	114	125	91,20%
niveau 2	16	145	161	90,06%
niveau 3	2	61	63	96,83%

The proportion of nonparticipating units was high in Ile-de-France (where 15 of the 29 nonparticipating maternity units were located) and among private non-profit maternity units (5 out of 30 maternity units with this status did not participate).

The hypothesis of independence between the response mechanism and the variables for region and legal status was rejected for maternity units overall; but the decision was made to retain all of the available information to characterize nonparticipation at the maternity unit level. (It may be wondered whether, for example, level of medical specialization is important for characterizing children and their future development.) To do this, the score method was used, with participation rates weighted by the initial weights. Ten participant groups with equal numbers of units were constructed using the variables stratum, region, level of medical specialization, and legal status. The 10 selected groups were modelled, with the probability of participation for a maternity unit in a given group ranging from 71% to 100%. The extreme groups (the two groups with the lowest score and the two with the highest score) were grouped together. Analyses were thus run on eight homogeneous participation groups.

In addition, even among the maternity units that agreed to participate, some could not be surveyed during all four waves for logistical reasons. To remedy this issue, we simply chose to adjust the initial weights, by wave and stratum, based on the ratio of the number of maternity units that were surveyed to the number of participating maternity units.

At this point it is crucial to recall that the units surveyed in ELFE were not "maternity unit x day" combinations. The sample unit was the individual infant. The plan was simply to select all infants born in any one of the selected maternity units on the survey days (cluster sampling of infants).

Mothers who wished to participate in the survey responded to a face-to-face questionnaire. A number of variables on nonparticipating mothers were collected in a “refusal file”. These variables are common to participating and nonparticipating mothers. This enables reweighting for nonparticipation that takes mothers' characteristics into account.

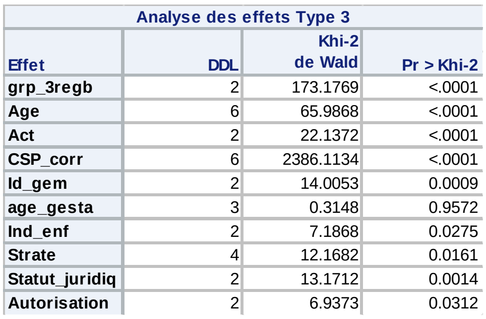
Infant nonparticipation was due to two effects.

The first effect is gaps in coverage: some eligible mothers were not approached. On the survey days, it was sometimes impossible for the interviewers to speak to all mothers, as when there were several births at the same time, or the mother left the maternity unit too early. This phenomenon, where individuals from the target population are absent from the sampling frame, is known as undercoverage. The number of eligible births per maternity unit was collected in the delivery room and is known (although it is approximate and is doubtless an overestimate). In order to correct for this problem, a coefficient was calculated by region (number of eligible infants/number of infants surveyed). Each infant was thus assigned this coefficient, slightly above 1, in order to correct for undercoverage.

The second effect is evidently larger: nonparticipation due to the mother's refusal to take part in the survey. After correcting for partial nonresponse, a logistic model was produced to create homogeneous participant groups using the unweighted score method. The included variables are:

- Maternity unit stratum, legal status, and level of medical specialization.
- Mother's age: [18; 22], [23; 24], [25; 29], [30; 34], [35; 39], 40 or above;
- Gestational age (in weeks): [33; 37], [38; 40], more than 40 weeks;
- Mother's *département* of residence, grouped by region, and then by group of regions: Ile-de-France, Centre, Picardie, Northeast, Northwest, Southeast, Southwest;
- SPC (occupations and socio-professional categories), with an eight-group classification: Farmers / Self-employed (non-farming) / manager or higher-level occupation / Intermediate occupation / Clerical or sales worker / Manual worker / No occupation / Cannot classify occupation;
- Activity at the time of pregnancy: yes/no;
- Indicator for twin birth: gave birth to twins or to a single baby;
- Primiparity (having a baby for the first time): yes/no;

The hypothesis of independence between the participation mechanism and a series of variables can be rejected.



Analyse des effets Type 3			
Effet	DDL	Khi-2 de Wald	Pr > Khi-2
grp_3regb	2	173.1769	<.0001
Age	6	65.9868	<.0001
Act	2	22.1372	<.0001
CSP_corr	6	2386.1134	<.0001
Id_gem	2	14.0053	0.0009
age_gesta	3	0.3148	0.9572
Ind_enf	2	7.1868	0.0275
Strate	4	12.1682	0.0161
Statut_juridiq	2	13.1712	0.0014
Autorisation	2	6.9373	0.0312

Fifty homogeneous participant groups were created using all of the included variables. The extreme groups (the 10 with the highest scores and the 5 with the lowest scores) were merged. In the end, 35 groups were produced, with modelled probabilities of participation between 13% and 75%.

Finally, a calibration process was performed. The National Perinatal Survey (ENP) is a regular survey (1995, 1998, 2003, 2010, 2016) in France. It aims to provide information on the health status of and perinatal care for infants and mothers, their characteristics, and associated risk factors. It is performed regularly in order to track how these are changing over time. The 2010 ENP took place on 15-21 March 2010 in all maternity units in metropolitan France as well as in overseas departments. We worked on the subsample that fulfilled the ELFE criteria: 14,492 infants (filtered for gestational age, mother's age, twin indicator, and birth in metropolitan France).

The following calibration variables were chosen: mother's age, region, marital status, immigrant status, level of education, and primiparity. Mother's age and region were grouped into 5 and 6 categories respectively, level of education into 3 categories, and the other variables are binary.

To limit the dispersion of the weights, they were truncated at 200 (% of weights). Finally, the weights were adjusted to maintain the total for the population as a whole.

In summary, the maternity unit weighting was produced as follows:

- Effect of the sampling plan

- This weighting was corrected to take into account nonparticipation at the "maternity unit x day" level (nonparticipation of a maternity unit on either all or a subset of survey days).
  Variables used: the size of the maternity unit (stratum), wave of the day of birth, legal status, level of medical specialization, and regional group**.**

- This weight was corrected again to take into account nonparticipation at the infant level.
  Variables used: maternity unit's stratum, legal status, and level of medical specialization; mother's age; Gestational age; mother's *département* of residence by regional group, SPC, Activity at the time of pregnancy, twin indicator, primiparity.

- Finally, these infant weights were calibrated on the known data on the population as a whole, and then truncated at 200. They were then adjusted to maintain the total of infants born in 2011, the target of the ELFE survey.
  Variables used: Age, Region, Marital status, Immigrant status, Level of education, and Primiparity

Weighting of the following survey waves: at 2 months, 1 year, 2 years

The weighting method used in the subsequent waves to date is described in detail in the documents "Weighting of the ELFE survey at time 1 (2 months of age)", "National survey age 1: Weighting at time 2 (1 year of age)" and "National survey age 2: Weighting at time 3 (age of 2 years)".

The principle of the weighting is the same in each of these cases. The maternity unit weights described above are adjusted to take into account nonparticipation in the survey wave (starting again from the maternity unit weighting, and not from the last known weighting, in order to avoid too great a dispersion of weights and ensure longitudinal consistency).

Thus, participation in each new survey wave is modelled using a logistic procedure. Note that the variables used to model this nonparticipation are those provided by the survey in maternity units. The method used is the creation of homogeneous participant groups on the basis of probabilities estimated by logistic regression. A number of groups (15 groups) are created on the basis of the sorted scores resulting from this regression. Within each of these groups, a response probability is estimated simply by the proportion of initially participating infants for which responses were received. The variables used are the following:

Wave at 2 months of age:

Mother's age, regional group, mother's SPC, marital status, couple's immigrant status, birth preparation sessions, took holidays during pregnancy

Wave at 1 year of age:

Variables used at 2 months + mother's level of education, father's SPC, mother’s tobacco consumption before pregnancy, mother's BMI.

Wave at 2 years of age:

Variables used at 2 months + mother's level of education, father's SPC, tobacco consumption before pregnancy, mother's BMI.

In addition, where the weight relates to the non-reference parent (non-primary carer), two variables were added to the nonresponse model: the father's presence for the birth and the father's activity at the time of the birth.

The weight for each infant is then adjusted by a coefficient that is equal to the inverse of this response probability.

These data were then calibrated to the same totals as described above for the maternity unit level. Finally, extreme weights (above 250) were truncated, and the overall distribution of weights slightly adjusted to obtain the desired total number of infants.

Why maintaining this method for the subsequent survey waves is problematic

The basic process consists, as we have seen, in first reweighting for nonresponse (NR), followed by calibration. An individual thus goes from an initial weight d to another weight d × F(X) / R , with R the (estimated) response probability and F (X) a function of the calibration variables X.

The process is relatively simple to understand, but requires knowing all of the information about nonrespondents that is required to model NR – which may no longer match the data collected on nonparticipants in the maternity units (activity, couple, sibship, etc.). Moreover, if there are too few respondents in certain classes of calibration variables (more precisely, for certain combinations of categories of the calibration variables), the weights will be disproportionate (risk of non-convergence).

Thus, in a longitudinal survey, maintaining the current method and applying it to future waves raises two issues.

- Producing new weights by calculating, for each survey wave, homogeneous participant groups between the initial wave in maternity units and the subsequent waves, and then “simply” multiplying the initial weights by the inverse of these response probabilities, risks generating very large weights whose truncation will strongly distort the chosen calibration in the long term. There is thus a risk both of weighting a small group of individuals too heavily or, by truncating these weights, of making the adjustment – and thus the factoring-in of nonresponse – potentially ineffective.

- Modelling response probabilities for ELFE infants in future waves (increasingly distant from birth) requires knowing the variables that allow nonresponse to be modelled. But by definition we do not know these variables, which are often sociodemographic, for nonrespondents. The calculations must thus be based on variables dating from the last survey wave with a response, which can quickly turn out to be problematic (modelling nonresponse using values of variables such as living with a partner, activity status, etc. that are several years old is quite risky).

We may then wonder whether calibration on a set of well-chosen variables, based directly on the sample of respondents, might not serve to deal with nonresponse. In other words, the initial process could be replaced with simultaneous calibration. In this case, the calibration is applied to the maternity unit weights from the selection (adjusted by the overall response rate for reasons of convergence).

Note: these two methods are strictly equivalent if the response probability is fully accounted for by the calibration variables alone.

This process is relatively easy to implement, but care must be taken: the explanatory variables that model nonresponse are the same ones that are used for the calibration. We thus assume that infant NR can be correctly explained by variables whose true total is known.

Moreover, the simultaneous method provides definite advantages for surveys where the data “ages”. If the variables used to explain “overall” NR (and not necessarily NR at a given time) are well chosen, with known total values, we do not need to know each of their values for nonrespondents. Instead we can directly calibrate with respondents alone to obtain the required totals. This assumes that the variables explaining NR do not depend on the time of the later survey wave (at 2 years, 3 years, 5 years, etc.).

An equivalent response probability is thus estimated a posteriori for each infant, by the ratio [weight before calibration/weight after calibration]. Unlike the maternity unit weighting, this response probability is thus “individual” (each infant has their own specific response probability), rather than being common to infants in the same homogeneous participant group.

The chosen weighting method: Simultaneous calibration

As explained in the previous paragraph, the quality of the simultaneous calibration method depends strongly on the variables used. These must explain as much of NR as possible.

We have seen that the variables explaining NR at the ages of 2 months, 1 year, and 2 years are more or less the same. There is also a complete study on attrition from the ELFE survey which analyses the relationship of this phenomenon to many variables. After numerous descriptive analyses, step-by-step regression was used to seek a better model of attrition over time. After the imputation of values where data are missing at low rates (below 7%), the process yielded a 10-variable model integrating, in order of importance: mother's SPC, father's SPC, acceptance of the transmission of data on the child, birth preparation sessions, alcohol consumption, father's age, holidays during pregnancy, mother's level of education, father’s activity, mother’s activity.

It is interesting to note that all of these studies still propose to model nonresponse or attrition using the same variables concerning socio-demographic and health characteristics, as well as the couple's involvement during the pregnancy.

In order not to include too many calibration variables, we focused on variables that predicted nonparticipation in maternity units (nonparticipation rate of around 50%, vs. 10-20% nonresponse in each subsequent wave). We also had to give preference to variables with data available in the National Perinatal Survey or in vital records for a population comparable to the coverage of the ELFE survey.

Because SPC is sometimes poorly codified (reported by respondents who sometimes misunderstand the categories, father's SPC provided by the mother, etc.), for activity status we chose to use the most reliable variables, combining mother's level of education with both parents' age group and employment status (active or not).

We thus propose the use of simultaneous calibration on 13 variables.

The variables include:

6 "contextual" variables (variables already used for calibration since the initial survey wave in maternity units):

- Regional group of residence (out of 5: Ile-de-France, Centre et Picardie/Northeast/Northwest/Southeast/Southwest)
- Primiparous mother (yes/no)
- Marital status (parents married/unmarried at the time of birth)
- Mother's age (18-24 / 25-29 / 30-34 / 35 and +)
- Mother's level of education (no schooling, primary, lower secondary vocational (CAP), secondary vocational (BEP)/ three last grades of high school/higher education
- Mother's immigrant status (yes/no)

To which we add the following variables, which are known to be associated to attrition:

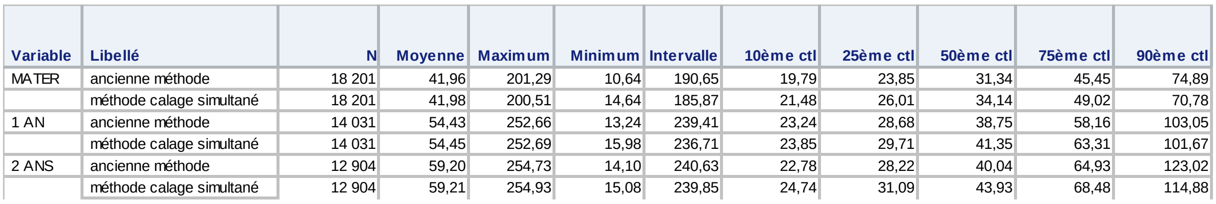
- Birth preparation sessions (yes/no)
- Father's activity status at the time of birth (in employment/other):
- Father's age (18-24 / 25-29 / 30-34 / 35 and +)
- Mother living with partner at birth (yes/no)
- Mother's alcohol consumption during pregnancy (yes/no)
- Twin birth (yes/no)
- Mother's activity status at the time of childbirth (in employment/other)

Note that for these new variables, missing data (5% for father's age, below 2% for the rest) were imputed before the calibration. As for the previous methods, all of the totals for the ELFE population come from vital records or the 2010 National Perinatal Survey.

Comparison of the new method with previously obtained results

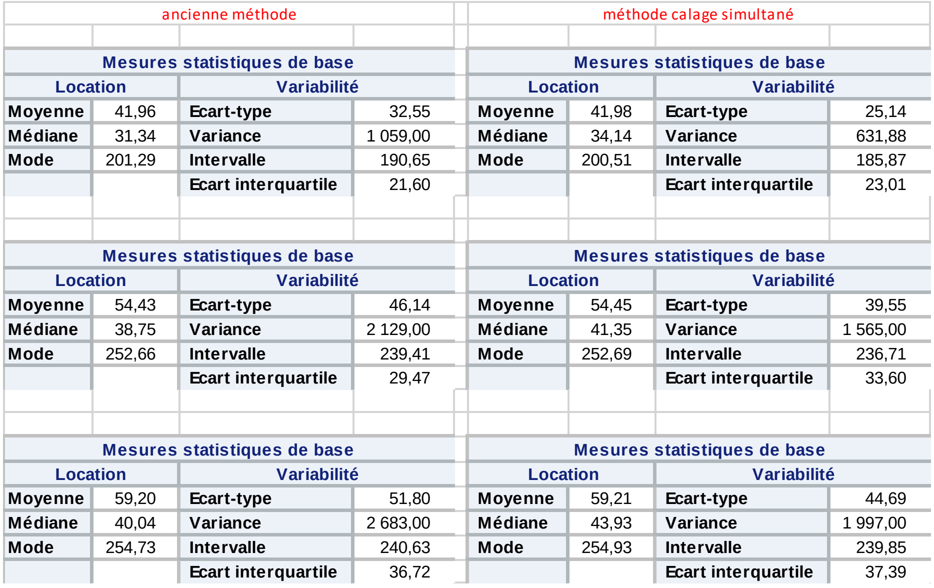
In order to check the legitimacy of this method and measure its impact on previously published results, it was applied to participating infants from the survey waves in maternity units, at 1 year, and at 2 years. This allows us to compare the weights obtained with the two different methods for a given individual at different times, and the overall results obtained with data from the respondents in each survey wave.

The basic statistics are equivalent.



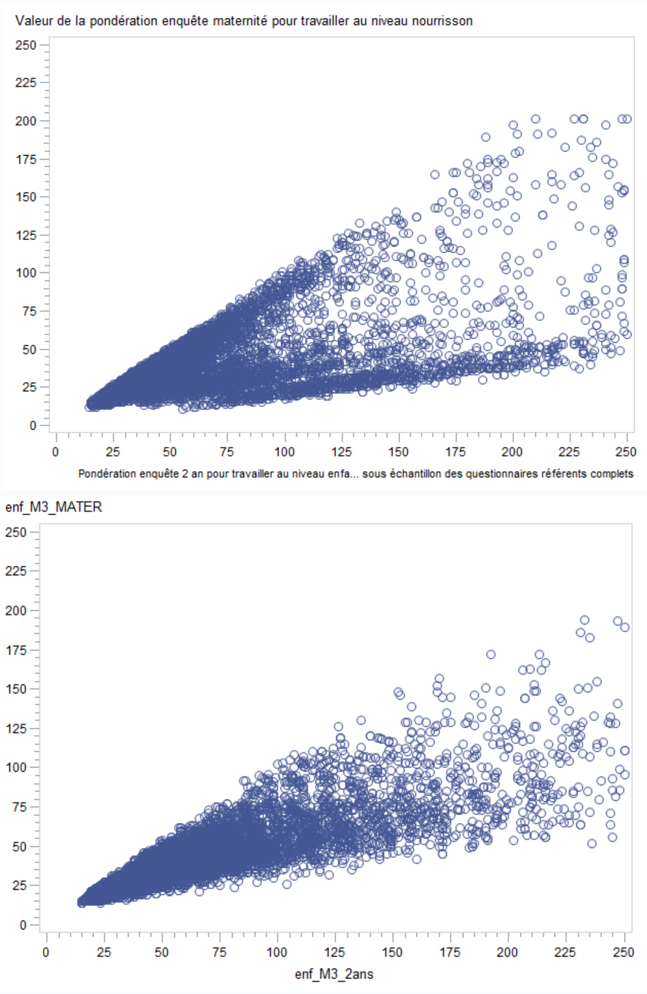
Variable	Libellé	N	Moyenne	Maximum	Minimum	Intervalle	10ème ctl	25ème ctl	50ème ctl	75ème ctl	90ème ctl
MATER	ancienne méthode	18 201	41,96	201,29	10,64	190,65	19,79	23,85	31,34	45,45	74,89
	méthode calage simultané	18 201	41,98	200,51	14,64	185,87	21,48	26,01	34,14	49,02	70,78
1 AN	ancienne méthode	14 031	54,43	252,66	13,24	239,41	23,24	28,68	38,75	58,16	103,05
	méthode calage simultané	14 031	54,45	252,69	15,98	236,71	23,85	29,71	41,35	63,31	101,67
2 ANS	ancienne méthode	12 904	59,20	254,73	14,10	240,63	22,78	28,22	40,04	64,93	123,02
	méthode calage simultané	12 904	59,21	254,93	15,08	239,85	24,74	31,09	43,93	68,48	114,88

But the new method immediately offers an important advantage: the weights are more compact. While their extent under the two approaches is similar, as is their inter-quartile interval, with the previous method the extreme weights are more widely dispersed. This leads the variance of the weights to be much (20-25%) higher.



ancienne méthode				méthode calage simultané			
Mesures statistiques de base				Mesures statistiques de base			
Location		Variabilité		Location		Variabilité	
Moyenne	41,96	Ecart-type	32,55	Moyenne	41,98	Ecart-type	25,14
Médiane	31,34	Variance	1 059,00	Médiane	34,14	Variance	631,88
Mode	201,29	Intervalle	190,65	Mode	200,51	Intervalle	185,87
		Ecart interquartile	21,60			Ecart interquartile	23,01
Mesures statistiques de base				Mesures statistiques de base			
Location		Variabilité		Location		Variabilité	
Moyenne	54,43	Ecart-type	46,14	Moyenne	54,45	Ecart-type	39,55
Médiane	38,75	Variance	2 129,00	Médiane	41,35	Variance	1 565,00
Mode	252,66	Intervalle	239,41	Mode	252,69	Intervalle	236,71
		Ecart interquartile	29,47			Ecart interquartile	33,60
Mesures statistiques de base				Mesures statistiques de base			
Location		Variabilité		Location		Variabilité	
Moyenne	59,20	Ecart-type	51,80	Moyenne	59,21	Ecart-type	44,69
Médiane	40,04	Variance	2 683,00	Médiane	43,93	Variance	1 997,00
Mode	254,73	Intervalle	240,63	Mode	254,93	Intervalle	239,85
		Ecart interquartile	36,72			Ecart interquartile	37,39

A second result: comparing weights for the same individual from the maternity units, at 1 year, and at 2 years shows greater drift in this relationship using the previous weighting method. For example, in the graphs below it can be seen that infants with a weighting of 100 (on the x-axis) at the age of 2 years could have a weighting from the maternity unit of between 15 and 100. Under the previous method, then, the weighting of some infants is thus multiplied by 6 over these successive waves, while that of others is unchanged. With the new method, this range is between 30 and 100. The weights thus remain more similar over time. This fact is all the more important as the drift over the course of future waves is likely to be even greater. This could lead, in time, to a situation where the results of the analyses of data from maternity units would be very different if they were analysed with a subsample consisting of the respondents from a later wave.



Comparison of weights for individual infants from maternity units and at age 2 years. Previous method (top) / new method (bottom)

Finally, it should be noted that for the data from the maternity units, there is very little difference between the distributions of variables calculated using the two methods. For example, the tables below present the distributions of weights for maternity units (MATER), at the ages of 1 year and 2 years, with the method used until the age of 2 years (called Elfe_) versus the simultaneous calibration method (called Nle_).

Note: the observed differences between certain calibration variables and the final distributions result from the truncation following the calibration.

Distributions of variables concerning the child's mother and father:

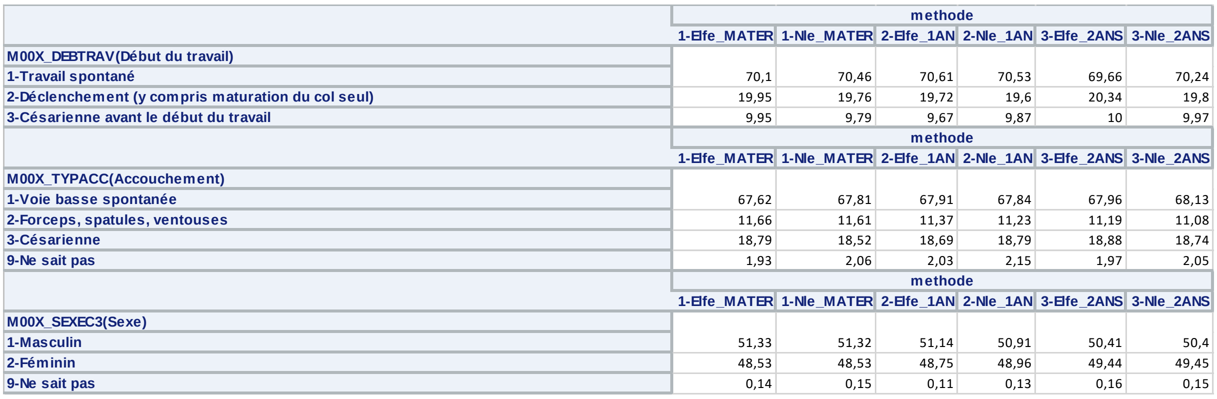
	methode					
	1-Elfe_MATER	1-Nle_MATER	2-Elfe_1AN	2-Nle_1AN	3-Elfe_2ANS	3-Nle_2ANS
M00M2_LIEUNAISM(Lieu de naissance mère)						
1-En France	81,55	81,39	81,79	81,79	82,19	81,84
2-Dans un autre pays	18,45	18,61	18,21	18,21	17,81	18,16
	methode					
	1-Elfe_MATER	1-Nle_MATER	2-Elfe_1AN	2-Nle_1AN	3-Elfe_2ANS	3-Nle_2ANS
M00M2_NATIOM(Nationalité mère)						
1-Française de naissance (y compris par réintégration)	82,95	83,07	83,32	83,65	83,5	83,48
2-Française par acquisition (naturalisation, mariage, déclaration, ou option à la majorité)	4,1	4,81	4,6	4,94	4,88	5,11
3-Etrangère	12,9	12,08	12,06	11,39	11,58	11,38
4-Apatride	0,04	0,04	0,02	0,02	0,04	0,03
	methode					
	1-Elfe_MATER	1-Nle_MATER	2-Elfe_1AN	2-Nle_1AN	3-Elfe_2ANS	3-Nle_2ANS
M00M2_ETATMAT(Etat matrimonial mère)						
1-Mariée ou remariée (y compris séparée légalement)	45,08	44,99	44,78	45,13	44,94	45,26
2-Pacsée	12,7	12,91	13,79	13,75	14,01	14,13
3-Divorcée	1,28	1,29	1,28	1,32	1,22	1,27
4-Célibataire	40,76	40,67	40,08	39,74	39,74	39,26
5-Veuve	0,18	0,14	0,06	0,06	0,09	0,07
	methode					
	1-Elfe_MATER	1-Nle_MATER	2-Elfe_1AN	2-Nle_1AN	3-Elfe_2ANS	3-Nle_2ANS
M00M2_COUPLE(La mère vit en couple)						
0-non	7,63	7,31	7,15	7,07	6,5	6,82
1-oui	92,37	92,69	92,85	92,93	93,5	93,18
	methode					
	1-Elfe_MATER	1-Nle_MATER	2-Elfe_1AN	2-Nle_1AN	3-Elfe_2ANS	3-Nle_2ANS
M00M2_NIVET(Niveau d'études mère)						
1-Ecole primaire	1,33	1,04	0,85	0,74	0,72	0,78
2-Collège (classes de la 6e à la 3e)	7,58	6,71	6,3	6,01	5,6	5,91
3-Classes préparant à un CAP ou à un BEP	17,98	19,53	18,84	20,23	17,86	19,77
4-Classes de seconde, première ou terminale générales	8,26	7,92	8,01	7,86	8,3	7,9
5-Classes de seconde, première ou terminale techniques	3,4	3,5	3,58	3,62	3,69	3,53
6-Classes de seconde, première ou terminale professionnelles	8,36	8,5	8,27	8,44	8,81	8,51
7-Etudes supérieures (facultés, IUT, etc,)	52,53	52,38	53,89	52,81	54,69	53,26
8-Vous n'avez jamais été scolarisée	0,56	0,43	0,26	0,28	0,32	0,34
	methode					
	1-Elfe_MATER	1-Nle_MATER	2-Elfe_1AN	2-Nle_1AN	3-Elfe_2ANS	3-Nle_2ANS
M00M2_PROFESS(Catégorie profession mère)						
1-Agriculteur, exploitant	0,31	0,33	0,39	0,39	0,38	0,4
2-Artisan, commerçant ou chef d'entreprise	2,81	3,21	3,07	3,24	3,26	3,29
3-Cadre ou profession intellectuelle supérieure	11,54	13,57	12,15	13,88	12,1	13,97
4-Profession intermédiaire (instituteur, infirmier, technicien, contremaître…)	16,41	17,05	17,7	17,64	17,75	17,87
5-Employé	37,4	41,7	41,58	42,28	43,09	42,67
6-Ouvrier	1,91	2,15	2,31	2,23	2,34	2,19
7-Sans profession	8,96	6,51	6,92	6,16	6,33	5,96
9-Ne peut classer la profession	20,66	15,49	15,88	14,19	14,74	13,64

	methode					
	1-Elfe_MATER	1-Nle_MATER	2-Elfe_1AN	2-Nle_1AN	3-Elfe_2ANS	3-Nle_2ANS
M00M2_LIEUNAISP(Lieu de naissance père)						
1-En France	81.48	82.35	81.96	83.84	82.54	83.87
2-Dans un autre pays	18.52	17.65	18.04	16.16	17.46	16.13
	methode					
	1-Elfe_MATER	1-Nle_MATER	2-Elfe_1AN	2-Nle_1AN	3-Elfe_2ANS	3-Nle_2ANS
M00M2_NATIOP(Nationalité père)						
1-Française de naissance (y compris par réintégration)	82.56	83.26	82.99	84.75	83.46	84.61
2-Française par acquisition (naturalisation, mariage, déclaration, ou option à la majorité)	5.30	5.07	5.26	4.77	5.06	4.93
3-Etrangère	11.84	11.40	11.54	10.28	11.21	10.19
4-Ne sait pas	0.30	0.27	0.21	0.21	0.27	0.27
	methode					
	1-Elfe_MATER	1-Nle_MATER	2-Elfe_1AN	2-Nle_1AN	3-Elfe_2ANS	3-Nle_2ANS
M00M2_EMPLOIC(Situation professionnelle père)						
1-A un emploi	88.35	87.48	89.48	87.93	90.30	88.27
2-Est homme au foyer	0.40	0.45	0.35	0.40	0.32	0.38
3-Est élève, étudiant ou en formation	1.41	1.52	1.55	1.77	1.61	1.85
4-Est au chômage	6.60	7.13	5.78	6.64	5.33	6.44
5-Est en congé parental	0.11	0.13	0.10	0.11	0.11	0.12
6-Est retraité	0.16	0.18	0.13	0.15	0.11	0.17
7-Est dans une autre situation	2.98	3.11	2.62	2.99	2.22	2.77
	methode					
	1-Elfe_MATER	1-Nle_MATER	2-Elfe_1AN	2-Nle_1AN	3-Elfe_2ANS	3-Nle_2ANS
M00M2_PEREACC(Le père a assisté à l'accouchement)						
0-non	21.77	21.12	20.12	19.79	20.33	19.27
1-oui	78.23	78.88	79.88	80.21	79.67	80.73

Distributions of variables concerning the pregnancy:

	methode					
	1-Elfe_MATER	1-Nle_MATER	2-Elfe_1AN	2-Nle_1AN	3-Elfe_2ANS	3-Nle_2ANS
M00M2_TABAG(Tabagisme pendant la grossesse)						
0-Non	78,38	78,05	77,79	78,74	78,84	79,22
1-Oui	21,62	21,95	22,21	21,26	21,16	20,78
	methode					
	1-Elfe_MATER	1-Nle_MATER	2-Elfe_1AN	2-Nle_1AN	3-Elfe_2ANS	3-Nle_2ANS
M00M2_TABA3G(Tabagisme pendant le 3e trimestre)						
0-Non	15,01	15,3	16,67	16,46	15,92	15,85
1-Oui	81,31	81,03	80,3	80,18	81,18	81,22
9-Non renseigné	3,68	3,67	3,03	3,36	2,89	2,93
	methode					
	1-Elfe_MATER	1-Nle_MATER	2-Elfe_1AN	2-Nle_1AN	3-Elfe_2ANS	3-Nle_2ANS
M00M2_FQALCOOL(Consommation d'alcool)						
0-Jamais	77,88	76,57	76,36	76,45	76,32	76,4
1-1 fois par mois ou moins souvent, ou lors d'occasions particulières comme les fêtes	15,16	16,15	16,44	16,49	16,6	16,64
2-2 à 4 fois par mois	1,57	1,6	1,55	1,48	1,59	1,55
3-2 à 3 fois par semaine	0,22	0,29	0,24	0,28	0,18	0,21
4-4 fois par semaine ou plus, mais pas tous les jours	0,04	0,05	0,02	0,02	0,05	0,03
5-Tous les jours	0,09	0,05	0	0	0,03	0,02
6-Seulement avant de se savoir enceinte	4,98	5,23	5,35	5,22	5,17	5,11
7-Ne souhaite pas répondre	0,06	0,07	0,04	0,06	0,05	0,04
	methode					
	1-Elfe_MATER	1-Nle_MATER	2-Elfe_1AN	2-Nle_1AN	3-Elfe_2ANS	3-Nle_2ANS
M00X_HTAG(Hypertension artérielle pendant la grossesse)						
0-Non	96,14	96,34	96,4	96,35	96,1	96,23
1-Oui avec protéinurie (?0,3g/l ou par 24h)	1,69	1,55	1,51	1,54	1,74	1,72
2-Oui sans protéinurie	2,16	2,11	2,09	2,11	2,16	2,05
	methode					
	1-Elfe_MATER	1-Nle_MATER	2-Elfe_1AN	2-Nle_1AN	3-Elfe_2ANS	3-Nle_2ANS
M00X_DIABGEST(Diabète gestationnel)						
0-Non	92,33	92,52	92,46	92,51	92,25	92,53
1-Oui	7,67	7,48	7,54	7,49	7,75	7,47

Distributions of variables concerning the birth:



	methode					
	1-Elfe_MATER	1-Nle_MATER	2-Elfe_1AN	2-Nle_1AN	3-Elfe_2ANS	3-Nle_2ANS
M00X_DEBTRAV(Début du travail)						
1-Travail spontané	70,1	70,46	70,61	70,53	69,66	70,24
2-Déclenchement (y compris maturation du col seul)	19,95	19,76	19,72	19,6	20,34	19,8
3-Césarienne avant le début du travail	9,95	9,79	9,67	9,87	10	9,97

	methode					
	1-Elfe_MATER	1-Nle_MATER	2-Elfe_1AN	2-Nle_1AN	3-Elfe_2ANS	3-Nle_2ANS
M00X_TYPACC(Accouchement)						
1-Voie basse spontanée	67,62	67,81	67,91	67,84	67,96	68,13
2-Forceps, spatules, ventouses	11,66	11,61	11,37	11,23	11,19	11,08
3-Césarienne	18,79	18,52	18,69	18,79	18,88	18,74
9-Ne sait pas	1,93	2,06	2,03	2,15	1,97	2,05

	methode					
	1-Elfe_MATER	1-Nle_MATER	2-Elfe_1AN	2-Nle_1AN	3-Elfe_2ANS	3-Nle_2ANS
M00X_SEXEC3(Sexe)						
1-Masculin	51,33	51,32	51,14	50,91	50,41	50,4
2-Féminin	48,53	48,53	48,75	48,96	49,44	49,45
9-Ne sait pas	0,14	0,15	0,11	0,13	0,16	0,15

Appendix 1: SAS software procedure

This appendix presents the code used to generate a weighting using SAS 9.4 (SAS Institute Inc, 2013).

This procedure requires the use of the SAS CALMAR macro, whose aim is to adjust a survey sample by reweighting individuals using auxiliary information available on some calibration variables: https://www.insee.fr/fr/information/2021902. *(in french)*

After creating the distribution of totals on which the calibration is to be based, this procedure requires:

table_selection = name of the table containing the data to be weighted. Caution: this table should only include these infants. The user should thus first generate this table using a DATA step.
This table must contain at least the following fields:

- *identifiant* = field identifying the records (by default ID...);
- *strate* = field identifying the stratum of the maternity unit (by default M00M1_MATSTRATEC1) ;
- *vague* = field identifying the survey wave (by default M00M1_VAGUE);
- the calibration variables CS_1 to CS_13.

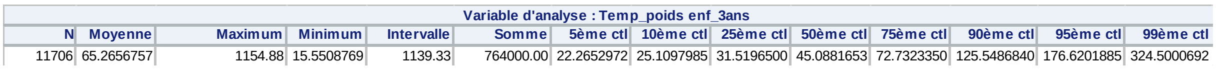
Tsortie = name of the output table
Psortie = name of the Tsortie field with the calculated weight
poidsMAX = max weight for truncation
Tmarge = name of the table with the calibration margins (known population totals). By default, margesM3
Label = Label of Psortie
tronc = 1 if weights are to be truncated at poidsMAX. 0 otherwise

For example, the SAS table "pond3ans" contains data for the 11,706 children who participated in the survey at the age of 3 ½ years. This weighting was generated using the following command:

```
%CALAGE
(table_selection=pond3ans,
strate=M00M1_MATSTRATEC1,
vague=M00M1_VAGUE,
identifiant=IDXX_XX,
Tsortie=pond_3ans,
Psortie=enf_3ans,
poidsMAX=250,
Tmarge=margesM3,
Label=enf_3ans,
tronc=1);
```

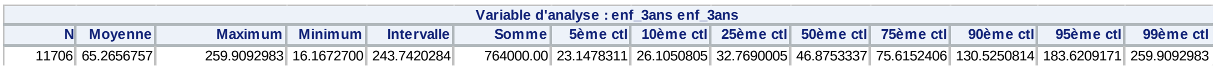
This procedure generates the table "pond_3ans", which includes only the two variables IDXX_XX and enf_3ans.

This procedure also provides some statistics on the generated weights (before and after truncation).



Variable d'analyse : Temp_poids enf_3ans

N	Moyenne	Maximum	Minimum	Intervalle	Somme	5ème ctl	10ème ctl	25ème ctl	50ème ctl	75ème ctl	90ème ctl	95ème ctl	99ème ctl
11706	65.2656757	1154.88	15.5508769	1139.33	764000.00	22.2652972	25.1097985	31.5196500	45.0881653	72.7323350	125.5486840	176.6201885	324.5000692



Variable d'analyse : enf_3ans enf_3ans

N	Moyenne	Maximum	Minimum	Intervalle	Somme	5ème ctl	10ème ctl	25ème ctl	50ème ctl	75ème ctl	90ème ctl	95ème ctl	99ème ctl
11706	65.2656757	259.9092983	16.1672700	243.7420284	764000.00	23.1478311	26.1050805	32.7690005	46.8753337	75.6152406	130.5250814	183.6209171	259.9092983

SAS code:

```
data margesM3;
input var $ n mar1-mar6;
cards;
CS_1 5 29.96 19.15 15.42 19.93 15.54 .
CS_2 2 43.1 56.9 . . . .
CS_3 2 45 55 . . . .
CS_4 4 13.96 31.22 33.25 21.57 . .
CS_5 3 27.8 19.9 52.3 . . .
CS_6 2 81.25 18.75 . . . .
CS_7 2 50.2 49.8 . . . .
CS_8 2 87.36 12.64 . . . .
CS_9 2 7.3 92.7 . . . .
CS_10 4 6.66 22.65 32.94 37.75 . .
CS_11 2 76.6 23.4 . . . .
CS_12 2 97.25 2.75 . . . .
CS_13 2 69.7 30.3 . . . .
;


%MACRO CALAGE (table_selection, identifiant, strate, vague, Tsortie, Psortie, poidsMAX,
Tmarge,Label, tronc );
title  ' ';


%let listeCALAGE= CS_1 CS_2 CS_3 CS_4 CS_5 CS_6 CS_7 CS_8 CS_9 CS_10 CS_11 CS_12 CS_13;
/* table avec uniquement les variables de calage nécessaires */
data calageSIMU (keep=&identifiant
&strate  &vague
&listeCALAGE
);
set &table_selection;
run;


data calageSIMU; set calageSIMU;
if &strate = 1 then _TOTALMAT_=108;
else if &strate = 2 then _TOTALMAT_=108;
else if &strate = 3 then _TOTALMAT_=109;
else if &strate = 4 then _TOTALMAT_=108;
else if &strate = 5 then _TOTALMAT_=111;

if &strate = 1 then _MATsel_=25;
else if &strate = 2 then _MATsel_=44;
else if &strate = 3 then _MATsel_=62;
else if &strate = 4 then _MATsel_=88;
else if &strate = 5 then _MATsel_=101;


if &vague = 1 then _TOTALJOUR_=90;
else if &vague = 2 then _TOTALJOUR_=91;
else if &vague = 3 then _TOTALJOUR_=92;
else if &vague = 4 then _TOTALJOUR_=92;

if &vague = 1 then _JOURsel_=4;
else if &vague = 2 then _JOURsel_=6;
else if &vague = 3 then _JOURsel_=7;
else if &vague = 4 then _JOURsel_=8;

pondAVANT_calage = (_TOTALMAT_ * _TOTALJOUR_) / (_MATsel_ * _JOURsel_) ;
run;
```

```
proc sort data=calageSIMU; by &identifiant; run;

*  écrit en dur NB=36028 = nb de nourrissons si aucune NR;
%let NB = 36028;

/* stocker dans NBP le total nourrissons à traiter = présents dans la table*/
proc sql;
create table totalP
as select count(*) as NBP from calageSIMU ;
quit;

data _null_;set totalP;
CALL SYMPUT('NBP', NBP);
run;

/* on redresse la pond avant calage par le taux de réponse global. NB/NBP.
c'est juste pour améliorer la convergence de calmar*/
data calageSIMU;
set calageSIMU;
pondAVANT_calageR=pondAVANT_calage*&NB/&NBP;
run;
proc delete data= totalP;


%let Ttemp=Tmp_&Tsortie;

%CALMAR (data= calageSIMU , poids=pondAVANT_calageR , ident=&identifiant ,
                datamar=&Tmarge , m=2 , /*editpoi=oui, */  /* 3 : logit */
               /* 2 : raking ratio */
               /* LO=0.6 , UP=1.3 ,*/
                datapoi=&Ttemp , poidsfin=Temp_poids , labelpoi=&Label,
                PCT=oui , EFFPOP=&TOTPOP);


/* si besoin, pour tronquer poids calés */
data Temp_sortie;
set &Ttemp ;
if (&tronc=1 and Temp_poids>&poidsMAX) then Temp_poids_T=&poidsMAX;
else Temp_poids_T=Temp_poids;
run;

proc sql ;
create table tronque
as select sum(Temp_poids_T) as tt from Temp_sortie;
quit;

data _null_;set tronque;
CALL SYMPUT('total_Tronque', tt);
run;
proc delete data= tronque; run;

DATA Temp_sortie ;
set Temp_sortie ;
Temp_poids_T = Temp_poids_T *&TOTPOP/ &total_Tronque;
run;

* génération de la table finale;
data &Tsortie (keep=&identifiant &Psortie);
set Temp_sortie;
&Psortie= Temp_poids_T;
label &Psortie=&Label;
run;

proc sort data=&Tsortie; by &identifiant;run;
```

```
title 'statistiques poids AVANT troncature';
*données avant/apres troncature;
proc means data=Temp_sortie  n mean max min range sum p5 p10 p25 p50 p75 p90 p95 p99;
var Temp_poids; run;

title  'statistiques poids APRES troncature';
*données avant/apres troncature;
proc means data=&Tsortie  n mean max min range sum p5 p10 p25 p50 p75 p90 p95 p99;
var &Psortie; run;

proc delete data= &Ttemp;
proc delete data= Temp_sortie;
proc delete data= calageSIMU;

title ' ';
%mend;
```