Reweighting method applied in the EAPS

The Economically Active Population Survey (EAPS), like any other household survey, may have distortions in the estimates it produces, due to a series of causes related to field work and sample design. The causes and their main consequences on results are listed below:

a) Lack of response, which produces skews in the final estimates, because it affects households in a relatively different way than population groups of certain composition.

b) Possible tendency of the sample to over represent households from a determined structure in detriment of others. In this sense, the updating of the survey framework (and despite the updating process of the sectioning having been habitually introduced into the sample design) by the duration of the process, inevitably introduces a delay in the capture of the changes in population that may have an influence on the estimates. It is easier to detect a change in population levels (in aggregated figures) than the specific location of new population contingents.

On the other hand, there are often external statistical sources that may improve the quality of the survey estimates. A procedure to carry out the said procedure is reweighting.

Reweighting consists of correcting the original weights or elevation factors inferred from the survey design in such a way that some final factors are arrived at, such that, when applying them, the estimate of the variables for those that have information available from the reliable external source (reference data for reweighting) coincide with the data from the said source.

The reweighting of elevation factors obtained from the design of the sample, is a habitual practice in the European Union. In the case of the Work Force Survey (the European EAPS) it was foreseen as a technical improvement worthy of implementation once the necessary information was available (Council Regulation 577/1998), relative to the organisation of a sample survey in the active population in the community, in its article 3, section 5).

In order to carry out any reweighting procedure it is necessary to choose some auxiliary or explanatory variables X that exist both in the survey and in a statistical source foreign to this, whether this is a census or administrative register.

Variables X to choose from, apart from existing both in the survey and alternative statistical source. It is desirable that they present the strongest possible correlation with interest variables. Y.

On the other hand, the variables of interest Y are those whose estimates again have the greatest importance in the survey. In the case of the APS they would be those related directly with activity, as are the condition of employed, unemployed, etc. Without doubt these variables are strongly conditioned by the sex and age of the person.

Also bearing in mind the quarterly character of the survey, it will be necessary to have available a homogeneous series of variables. X in the alternative source with the same periodicity as the EAPS.

Generally, the favourite explanatory variables are the actual population classified by age groups and sex. In the case of the EAPS it has been opted to use the population projections relative to each Autonomous Community referring to each quarter.

Eleven age groups crossed with sex have been taken, which represents for each Autonomous Community a vector of twenty two actual population groups supplied by the projections in each quarter.

If in the survey we start with a sample size n , calling ω an original dimension weights vector $n \times 1$ and $\dot{\omega}$ the homologous vector of transformed weights, any reweighting procedure that is applied will give rise to a functional type relationship $\dot{\omega} = \dot{\omega}(\omega, X)$, in other words, the new weights will be a function of the originals ones and the auxiliary variables chosen.

The auxiliary information supplied by the survey will be contained in a matrix $X_{n \times p}$ where the auxiliary variables values appear in each row or the modality of these for each one of the sample individuals. In this case p would take the value 22.

The new weights have to fulfil the condition of balancing the sample. In other words $X'\dot{\omega} = x$, where: x the effective population vector provided by the projections.

With the weights $\dot{\omega}$, would change to obtain new estimates for any interest variable Y .

Procedure used

As a procedure to carry out the reweighting the INE has opted for CALMAR which is a framework in SAS developed by the INSEE in France.

With this method a distance function is previously defined $G(\omega, \dot{\omega})$ and it is required that $\sum_{k=1}^{n} \omega_k G(\omega_k, \dot{\omega}_k)$ is minimal for the whole of the sample with the tie $\sum_{k=1}^{n} \dot{\omega}_k = N$, in other words, the sum of transformed weights must recover a certain population total.

Calling h the quotient $\frac{\dot{\omega}_k}{\omega_k}$, the two distances most usually used are defined:

quadratic $G(h) = \left(\frac{h-1}{2} \right)^2$

and logarithmic $G(h) = h \log(h) - h + 1 \qquad h > 0$

Associated with these distance functions there exist the following functions of the transformation of new weights with respect to the original ones.

$\dot{\omega} = \omega(1+u)$ linear

$\dot{\omega} = \omega e^u$ exponential

With the linear function there exists the risk of obtaining negative weights while with the exponential function there may be more distortion of new weights with respect to the original ones.

The framework of CALMAR also offers the possibility of putting quotas for the transformation of the original weights, in other words, two values are searched for L and U such that $L < h_k < U \qquad k = 1,2....n$ where

$$h_k = \frac{\dot{\omega}_k}{\omega_k}$$

The INE has opted to use the linear truncated method taking as an inferior quota $L = 0.1$ and higher quota $U = 10.0$

The framework may apply from 1 to 7 dimensions; in this case it has been applied in 2 dimensions to a contingency table whose marginal row frequencies are the population projections of 16 years and over corresponding to each one of the provinces that constitutes a given Autonomous Community. The marginal column frequencies are the 16 and over population classified by sex and five year age groups for the whole of the Autonomous Community.

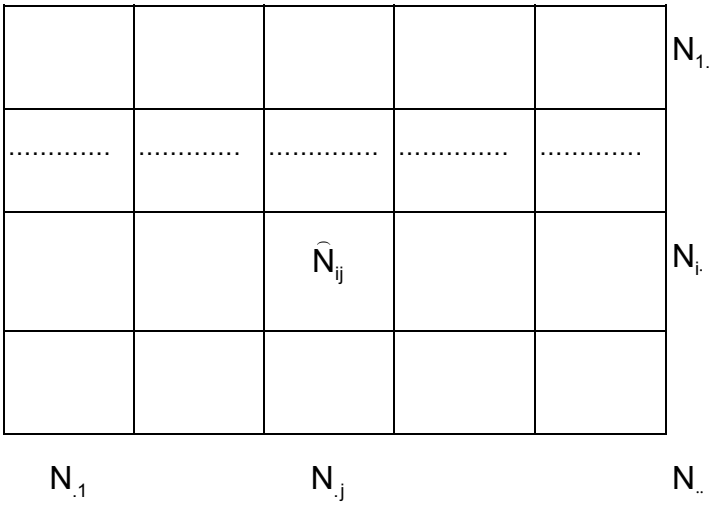
According to the diagram alongside (DIAGRAM 1), $\hat{N}_{ij}$ is the estimate that the survey provides with the original weights for the effective population of province I and age group j in the Autonomous Community considered.

Considering the linear transformation although with truncation, the relationship between the new and original weights relative to all registers k that belong to squares I, j, is

$$\dot{\omega}_k = \omega_k(1 + u_i + v_j)$$

The quantities u_i and v_j are the mysteries which the MACRO resolves. They may have any sign although normally when adding them up they give a certain a magnitude near to zero but in this way they satisfy the minimal variation condition in the transformation of the weights better.

DIAGRAM 1



$N_{i.}$ = Total population of 16 years old and over in the province I

$N_{.j}$ = Projection of the population for age and sex group j in the Autonomous Community considered.

The new frequencies of transformed squares $\hat{N}_{ij}$ will have to recover the marginal $N_{i.}$ and $N_{.i}$ supplied by the reference statistical source foreign to the survey. In our case, this deals with population projections elaborated by the INE.

The problem of single reweighting

The auxiliary variables used in the reweighting are quantitative (but specifically binary, they take the values 0 or 1) just like the practical total of the APS variables. In applying the CALMAR in 2 dimensions, each row of the matrix $X_{n\,x\,p}$ will have $p = 22$ columns; in one of them there will be a 1 and the remaining ones a 0 since each individual in the sample belongs to only one modality for age and sex groups. On the other hand, each individual will have k additional columns that indicate whether or not they belong to any of the k provinces that constitute the Autonomous Community to which CALMAR is applied.

The individuals of a same household who have different coordinates in the matrix $X_{n\,x\,p}$ will also see their original weights affected in a different way as the transformation factor $(1 + u_i + v_j)$, v_j varies according to age group and sex to which the individual belongs; this causes the loss of one of the original design qualities such as the single weighting for all individuals from the same household.

The parameter u_i for the transformation factor does not cause any problem, since it only depends on the province to which the household belongs.

To make the problem set out evident, a matrix is constructed Z_{sxp} where s is the number of households, which is easily obtained based on the X_{nxp}, simply adding the columns of coordinates for the individuals belonging to the same household.

Thus, for example, if we have a household with 3 persons and the following age and sex coordinates:

individuo 1	0	0	.....	1	.....	0
individuo 2	0	1	.....	0	.....	0
individuo 3	0	0	.....	1	.....	0

The age and sex coordinates for the household will be

0	1	.....	2	.....	0

The CALMAR will now be applied to matrix Z (matrix that takes the discreet values, 0, 1, 2 etc., depending on the number of household members of 16 and over) that moreover have the advantage of being of more reduced dimensions than X; the new transformed weight that is obtained for the household will be attributed to all its members in such a way that the initial quality of the design is recovered.

Upon applying the CALMAR macro problems can easily appear such as co-linearity between variables, a situation which can be resolved by deleting one of the columns of the explanatory matrix Z.

For Autonomous Communities with only one province the CALMAR has also been applied by deleting one of the coordinates from the auxiliary variables.

To get to know in detail the fundamental maths of CALMAR it is suggested that the following articles are consulted:

Deville, Särndal and Sautory, *Generalised raking procedures in survey sampling*; Journal of the American Statistical Association, September 93, Vol. 88 nº 423.

Deville and Särndal, *Calibration estimators in survey sampling*. Journal of the American Statistical Association; June 92, Vol. 87, nº 418.