

Use of the CONTROL SAMPLE to measure the impact of the changes in the EAPS in the first quarter of 2005.

1. Introduction.

The CONTROL SAMPLE **aims** to measure the impact produced in the EAPS given the modification of the questionnaire and the collection method.

The sampling theory establishes that the measurement of the change between two estimates should be performed using the same sample and the same measuring instruments with a view to comparing homogeneous amounts and controlling, to a certain extent, both sampling and non-sampling errors.

When it becomes necessary to measure the impact in the estimates brought about by the modification of the questionnaire and the collection methods, it is established that the best measure for this impact would be to interview the same families that are already part of the sample using the same interviewers. This approach was rejected in practice, since on the one hand it is impossible to use the same interviewers since they are continuing their regular tasks in the survey, and on the other, the bother it would entail for the families would increase the number of incidents and would introduce a response bias.

Considering the aforementioned elements, a **control sample** has been designed independently from the general sample, with a sample size of about 15.000 dwellings, in order to **measure the effects of these changes in the estimation of the main characteristics** of the survey, Active people, Employment and Unemployment in the Autonomous Communities.

With a view to the design of this sample being as similar as possible to the survey design, dwellings have been selected in the same primary units as those in the general sample. These dwellings have been interviewed successively, with the questionnaire and procedures used up until the fourth quarter of 2004 in the survey, hiring and training temporary interviewers to undertake these tasks.

Estimates from the control sample present a greater sample error than that of the general survey, given the lower sample size. Therefore, the survey results should be compared cautiously.

2. Sample characteristics.

A sample of **six** dwellings has been selected in 2920 census sections. The sections are those of the current EAPS sample not including the sections of the rotation shift that will renovate their sample of dwellings during the first quarter of 2005.

The first interviews were carried out in the third quarter of 2004 in three of the six selected dwellings. In the fourth quarter, the first interview was performed in the other three and the second interview was carried out in the dwellings that had already been interviewed in the previous quarter.

The second and third interviews, respectively, were carried out in the first quarter of 2005.

3. Direct estimation method.

Direct estimates obtained from the control sample in the first quarter of 2005 are obtained using the **same** estimators as in the survey, that is to say, **ratio estimators** calibrated for the auxiliary variables, age group and sex by autonomous communities and calibration of foreigners. The control sample uses the same external sources as the general survey.

Furthermore, in order to ensure the results are as precise as possible, **as well as the aforementioned variables**, the sample also uses those corresponding to three activity situations: employed, unemployed and inactive, in the fourth quarter of 2004, by autonomous communities. As regards employed persons, the sample has considered all four economic sectors (agriculture, industry, construction and services).

By using this method to calculate raising factors, the control sample provides exactly the same estimate as the EAPS for the fourth quarter of 2004 as regards the main variables that describe relationship with activity.

In order to undertake the described calibration process, the sample employs the same CALMAR software used for the EAPS. The calibration process acts as a generalised regression estimator (GREG), which in its multidimensional form is expressed by:

$$\hat{Y}_{\text{GREG}} = \hat{Y}_{\text{HT}} + \beta'(X - \hat{X}_{\text{HT}})$$

When using qualitative variables, the GREG estimator equals a post-stratification estimator that is incomplete as regards the calibration variables.

This procedure **has only been used to obtain estimates on a national level**, on considering that they present an acceptable level of precision.

4. Use of combined estimators.

As mentioned previously, direct estimates obtained using the control sample are affected by higher sample errors than those obtained from the EAPS, given the smaller sample size.

Sample errors increase as regards estimates in more broken down domains, Autonomous Communities. Therefore an alternative method has been used, aiming to overcome the lack of precision derived from a smaller sample size.

A **Combined Estimator** has been used to achieve more precise estimates with this level of breakdown, and therefore more soundness when measuring the change.

This type of estimator attempts to improve the efficiency of the direct estimator when the sample size is not big enough to guarantee the precision required in the estimates. To do so, a synthetic estimator is also used, since it uses sample information from a space-time domain different to that of the objective population, albeit similar to the same.

Consequently, there are two different estimators: the direct estimator, which is unbiased but presents a vast variability in the estimates, and the synthetic estimator, which presents less variability but also introduces a bias that is difficult to specify albeit considered minor a priori.

The **combined or compound** estimator is obtained as a lineal convex combination of a direct estimator and a synthetic estimator. It aims to decrease the variance of the estimates even if this contributes a bias to the latter, always aiming to ensure that the mean square error achieved is substantially lower than that obtained via direct estimate, that is to say, to improve the accuracy of the estimates.

The direct estimator used has been obtained from the control sample in line with the process detailed in the previous section. The synthetic estimator is a **predictor** based on an ARIMA time series model, for which the whole series of data from the EAPS since the year 1976 have been used.

Trials carried out using the synthetic estimator in the first quarter of 2004 confirmed the process' soundness, i.e., the proximity of its predictions and the estimates of the survey for said quarter as regards objective variables and the level of breakdown for the Control Sample.

When selecting coefficients, the proximity (or similarity) between the predictor and the control sample has been considered, in accordance with the sign of the evolution of the characteristic, implementing the following criterion:

The general expression for the combined estimator \hat{X} is as follows:

$$\hat{X} = \alpha \cdot \hat{T}_x + (1 - \alpha) \cdot \hat{P}_x$$

where

\hat{T}_x represents the estimator of characteristic X obtained using the control sample and \hat{P}_x represents the estimator obtained using the predictor, both in the first quarter of 2005.

α is the coefficient or weighting used in the combined estimator.

The following cases have been considered:

$$\text{A. Sign} (\hat{T}_x - \hat{X}_{4/04}^{\text{EPA}}) = \text{Sign} (\hat{P}_x - \hat{X}_{4/04}^{\text{EPA}})$$

Yes:

$$\hat{P}_x - \hat{\sigma}(\hat{P}_x) \leq \hat{T}_x \leq \hat{P}_x + \hat{\sigma}(\hat{P}_x) \quad \alpha = 0.7$$

$$\hat{P}_x - 2\hat{\sigma}(\hat{P}_x) \leq \hat{T}_x \leq \hat{P}_x + 2\hat{\sigma}(\hat{P}_x) \quad \alpha = 0.4$$

$$\hat{P}_x - 3\hat{\sigma}(\hat{P}_x) \leq \hat{T}_x \leq \hat{P}_x + 3\hat{\sigma}(\hat{P}_x) \quad \alpha = 0.2$$

$$\text{B. Sign} (\hat{T}_x - \hat{X}_{4/04}^{\text{EPA}}) \neq \text{Sign} (\hat{P}_x - \hat{X}_{4/04}^{\text{EPA}})$$

Yes:

$$\hat{P}_x - \hat{\sigma}(\hat{P}_x) \leq \hat{T}_x \leq \hat{P}_x + \hat{\sigma}(\hat{P}_x) \quad \alpha = 0.6$$

$$\hat{P}_x - 2\hat{\sigma}(\hat{P}_x) \leq \hat{T}_x \leq \hat{P}_x + 2\hat{\sigma}(\hat{P}_x) \quad \alpha = 0.3$$

$$\hat{P}_x - 3\hat{\sigma}(\hat{P}_x) \leq \hat{T}_x \leq \hat{P}_x + 3\hat{\sigma}(\hat{P}_x) \quad \alpha = 0.1$$

Estimates calculated in this manner are calibrated to ensure national totals coincide with the direct estimates obtained from the control sample at a national level.

In order to allocate the coefficient, the sample considers that the predictor, synthetic estimator, introduces a certain smoothing on the distribution of the sampling of the direct estimator which is, therefore, sounder than the former. Therefore, the further the estimate of the control sample, direct estimate, moves away from the predictor, the lower the weighting.

This methodology is based on Small Area Estimation Techniques which the INE has worked on via its participation in the international project EURAREA (Enhancing Small Area Estimation Techniques) in the framework of the 5th R&D Programme, developed by the European Union.

5. Final considerations.

The measurement of the impact based on the comparison of the estimates obtained in both samples, control and general, have the following limitations:

- Since both samples have different sizes, **sample errors** have very different effects on level estimates.
- Both estimates are affected by **non-sampling errors** that are hard to measure. In this sense it is necessary to note the following:
 - a) The changes implemented in the questionnaire aim to reduce these errors, as they make it easier for the persons interviewed to understand the questions.

- b) The effect of employing a larger proportion of temporary interviewers in the control sample compared to those used for the normal EAPS is unknown.
- c) When the comparison between successive quarters is performed using the same operation, the biases due to these errors are compensated and therefore do not affect the comparison; this does not happen when different operations are involved.

Nevertheless, the use of new calibration variables and of combined estimators provides sound estimates for the measure of the impact, due to the changes, in the main characteristics of the sample.

SGMYTE.

Madrid, June 3rd 2005