Working Papers

02/2011

On the error of backcast estimates using conversion matrices under a change of classification

Ignacio Arbués

Natalia López

First draft: February 2011

This draft: February 2011

On the error of backcast estimates using conversion matrices under a change of classification

Abstract

The classifications used by statistical agencies are sometimes updated. Hence, for the sake of comparability, it is necessary to estimate data from past periods according to the new classification. A frequently used method to calculate the estimates is through the use of Conversion Matrices. We present a theoretical analysis of this method and show with a practical example that it is possible to obtain useful estimates of the error.

Keywords

Change of Classification, Backcasting

Authors and Affiliations

Ignacio Arbués

Natalia López

Dirección General de Metodología, Calidad y Tecnologías de la Información y las Comunicaciones, Instituto Nacional de Estadística

On the error of backcast estimates using conversion matrices under a change of classification

Ignacio Arbués* and Natalia López
Instituto Nacional de Estadística, Spain

February 18, 2011

Abstract

Official statistics agencies produce disaggregated data according to different classifications (of economic activities, products, occupations, ...). When these classifications become obsolete, their replacement in the statistical production is a difficult task from many points of view. One of the difficult issues is the necessity to provide retrospective data according to the new classification, for otherwise the users would not have comparable data nor long time series. The calculation of these retrospective data (backcasting) is performed most often either by a micro approach, that is, by reclassifying the micro-data of previous periods according to the new classification or by the Conversion Matrices Method (CMM), that consists of using data classified according to both classifications (since usually there is an overlapping period) to calculate the coefficients of some conversion matrices that are used to estimate the unknown aggregates of the past as linear combinations of the known ones. This method lacks not only theoretical support but also diagnostic tools to assess the quality of the estimates. In this paper, we propose a method to estimate the error of the CMM and present the results of a practical application of the method to the change from revision 1.1 to revision 2 of the Statistical Classification of Economic Activities in the European Community (NACE).

*Corresponding author. Email: iarbues@ine.es; Instituto Nacional de Estadística, Castellana 183, 28071, Madrid, Spain

1 Introduction

Together with the main aggregates that describe the whole economy of the geographical area of interest, statistical agencies usually produce more detailed data. These data are obtained as a breakdown of the large aggregates according to different classifications. For instance, we can obtain the production activity of industrial branches using the National Classification of Economic Activities CNAE-09, that is, the Spanish version of NACE Rev. 2 –Statistical classification of economic activities in the European Community–, employment by occupation according to CNO (National occupation classification), the household expenditures by expenditure group COICOP/HBS (Classification Of Individual Consumption by Purpose, used in the Household Budget Surveys). Since the economy is always changing, after some time, these classifications become obsolete and are no longer considered as useful tools to provide an accurate description of the subject under study. Then, a new classification is designed –increasingly often by a supra-national institution– and statistical agencies are requested to adapt their statistics to it.

The adaptation of the statistical production to a new classification is largely an organizational issue, not so much a methodological one. However, the users often need to compare the disseminated data with data from the past. If the data of different periods are in different classifications it is not possible, for example, to apply econometric methods or, at best, it becomes extremely difficult. Thus arises a requirement to the producer of statistics to provide at least an estimation of aggregates according to the new classification corresponding to time periods prior to the change of classification. This is what is known among the official statisticians as 'backcasting'.

NACE is the statistical classification of economic activities in the European Community and is part of an integrated system of classifications developed under the vigilance of United Nations Statistical Division. It has a hierarchical structure that codes the universe of economic activities. This structure is as follows:

- Sections: Alphabetical code.
- Divisions: two-digit numerical code.
- Groups: three-digit numerical code.

- Classes: four- digit numerical code.

The first version of NACE appeared in 1970, but it did not allowed comparison with other international classifications and that was the reason why in 1990 NACE rev. 1 was produced, with ISIC rev. 3 as a starting point. In 2002 was established NACE rev. 1.1, that introduced some updates and in that year started the procedure of full revision the NACE. The new version, NACE rev. 2, is applied since January 1st, 2009. The changes of NACE have as a consequence a break of continuity of time series data. That is why it is necessary to 'backcast'.

Roughly, it can be said that there are two different approaches to perform the task. One of them is generally known as micro-approach and it is essentially a repetition of the calculations that were originally performed when the data according to the old classification were produced, but now with the micro-data classified according to the new classification. Besides sampling problems (when the classification is used for stratification) discussed for example in [1], what is produced with this method can be considered not as an estimate, but an exact recalculation. Unfortunately, to recode the micro-data is usually very costly and in many cases not feasible.

The most popular alternative approach is known as Conversion Matrices Method (CMM) and consists of estimating the aggregates according to the new classification as a linear combination of the aggregates according to the old one. The aggregates can be arranged in a vector and thus, the new vector can be obtained as the product of a matrix of coefficients times the old vector. This method is discussed, for example in [2], [3] and [4].

There are some decisions to take when producing backcast data, but most of them belong to the next two kinds:

- Choice of the method to use for the estimation.
- Trade-offs between quality and amount of information provided. For example, how far in the past or what level of detail can be achieved without seriously compromising the quality.

The main problem here is that it is very difficult to take such decisions without a measure of the quality of the estimates and that is precisely what the CMM lacks. Consequently, in the past, the estimates have been analyzed only in empirical and informal ways, such as visual inspection of the time series

mainly focused on their stability. The fact that a time series is stable in time does not guarantee at all that it is an accurate estimate of the theoretical one (that is, the one that would had been calculated if the new classification were been used at the time).

In fact, there is a third approach, in which no use of the micro-data is made at all, using instead econometric models. This happens to be the most developed approach in the literature (see [5] or [6]), but it is mostly of interest for analysts outside the statistical institutions, that do not have access to the micro-data.

In this paper, we tackle the problem of estimating the error of the estimates produced by the CMM, thus providing a tool to take the above-mentioned decisions with greater awareness of the consequences on quality. In section 2, we describe the CMM as is usually applied, in section 3, we present our estimates of the error, in section 4 the results of an application to real data are discussed.

2 Calculation of conversion matrices

Let us introduce some notation. We will denote by P_t the population of period t, with $t = 1, \ldots, T$, and then, we identify with an index i ranging from 1 to N the units in $P = \cup_{t=1}^{T} P_t$. The value of the variable of interest measured at time t and unit i is denoted by $x_{i,t}$.

Every unit i belongs to a class[1] according to the old classification, which we denote by $\alpha(i)$ and according to the new one, $\beta(i)$. Let us assume that when the old classification was being used, the objects of interest were the population totals

$$X_t^{\ell}(\alpha) := \sum_{\alpha(i)=\ell} x_{i,t} \tag{1}$$

and after the change of classification the new objects of interest are

$$X_t^{j}(\beta) := \sum_{\beta(i)=j} x_{i,t}. \tag{2}$$

We denote by S_t^{ℓ} the subsample of class ℓ at time t. In order to avoid inessential complications, we assume that the sample is obtained by simple random sampling and the totals according to the old classification were estimated

[1] We use now the word 'class' in a generic sense, not specifically the four-digit activities of NACE.

by the Horvitz-Thompson estimator

$$\hat{X}_t^\ell(\alpha) := \frac{M^\ell}{m^\ell} \sum_{i \in S_t^\ell} x_{i,t}, \tag{3}$$

where M^ℓ is the size of the strata (class) ℓ and m^ℓ is the size of the subsample.

We may also consider

$$X_t^{\ell,j}(\alpha,\beta) = \sum_{\alpha(i)=\ell,\beta(i)=j} x_{i,t}, \tag{4}$$

that is, the total obtained using the units that belong to class ℓ according to the old classification and to j according to the new one.

Let us now consider the following expression

$$\pi_t^{j/\ell} = \frac{X_t^{j,\ell}(\alpha,\beta)}{X_t^\ell(\alpha)}. \tag{5}$$

Depending on the nature of the variable $x_{i,t}$, $\pi_t^{j/\ell}$ can be interpreted as a conditional probability. In the example of section 4, $x_{i,t}$ is the turnover of industrial enterprises. Then, we may regard $\pi_t^{j/\ell}$ as the conditional probability of a currency unit to be spent to purchase a product of class j according to β, given that it is spent to purchase a product of class ℓ of α.

The numerator in (5) cannot be calculated if we do not know the classification of the units in both α and β. We will assume that this is known for a time D (double classification period). In fact, rather than $\pi_D^{j/\ell}$, we will have,

$$\hat{\pi}_D^{j/\ell} = \frac{\hat{X}_D^{j,\ell}(\alpha,\beta)}{\hat{X}_D^\ell(\alpha)}, \tag{6}$$

where

$$\hat{X}_D^{j,\ell}(\alpha,\beta) = \frac{M^\ell}{m_\ell} \sum_{i \in S_D^\ell, \beta(i)=j} x_{i,D}, \tag{7}$$

and consequently,

$$\hat{\pi}_D^{j/\ell} = \frac{\sum_{i \in S_D^\ell, \beta(i)=j} x_{i,D}}{\sum_{i \in S_D^\ell} x_{i,D}}. \tag{8}$$

That means that the 'conditional probability' is estimated in the sample S_D^ℓ. Using $\hat{\pi}_D^{j/\ell}$, the estimate of the totals $X_t^j(\beta)$ by the conversion matrix method would be,

$$\hat{\hat{X}}_t^j(\beta|\alpha) = \sum_\ell \hat{\pi}_D^{j/\ell} \hat{X}_t^\ell(\alpha). \tag{9}$$

We can decompose the error of these estimates as

$$X_t^j(\beta) - \hat{X}_t^j(\beta|\alpha) + \hat{X}_t^j(\beta|\alpha) - \hat{\hat{X}}_t^j(\beta|\alpha), \tag{10}$$

where

$$\hat{X}_t^j(\beta|\alpha) = \sum_\ell \pi_D^{j/\ell} \hat{X}_t^\ell(\alpha). \tag{11}$$

In this paper, we will focus on the first difference in (10), assuming that the estimates of $\pi_D^{j/\ell}$ are good enough.

We can write (11) in matrix form as

$$\hat{X}_t(\beta|\alpha) = \Pi_D \hat{X}_t^\ell(\alpha),$$

where the vector $\hat{X}_t(\beta|\alpha)$ and $\hat{X}_t(\alpha)$ contain the estimates corresponding to all activities, and Π_D is the conversion matrix with as many rows as classes is β and as many columns as classes in α.

2.1 Seasonal variant

When the variable under study has seasonal behaviour, (11) may provide a poor estimate. In order to see this, let us consider that the method is applied to monthly data from y complete years, and D is December of year y. If for certain j and ℓ, the intersection of classes α^ℓ and β^j contains units with a seasonal pattern different from the remainder of α^ℓ, then $\pi_D^{j/\ell}$ will not represent a good estimate in the 'conditional probability' sense.

As an example, consider $\ell = 15$ and $j = 11$, where α is NACE rev. 1.1 and β is NACE rev. 2. Class α^ℓ comprises both food and beverages and β^j includes only beverages. The production of beverages shows a markedly different seasonal pattern from food. Consequently, the sales of beverages, $X_t^{\ell,j}(\alpha,\beta)$ varies strongly among the months of the year as a share of the total $X_t^\ell(\alpha)$. In particular, the $\hat{\pi}_D^{11/15}$ computed with December of year y will be too small to estimate $X_t^{11}(\beta)$ in the summer months, when the sales of beverages are greatest.

On the other hand, for short-term indicators (i.e., with more than one data per year), the units are classified according to α and β for at least a whole year. Let us assume for simplicity that the double classification period ranges from $t = T - s + 1$ to T, where $s = 12$ for monthly data and $s = 4$ for quarterly data.

Then, we will use the estimate

$$\hat{X}_t^j(\beta|\alpha) = \sum_\ell \pi_{D(t)}^{j/\ell} \hat{X}_t^\ell(\alpha), \tag{12}$$

where $D(t)$ is $T - s + \text{rem}(t, s)$ and $\text{rem}(t, s)$ is the remainder of the integer division of t by s. In order to avoid cumbersome notation, we will make the dependency of D on t implicit, thus writing D when no ambiguity arises.

3 Estimation of the error

In this section, we will estimate the error of the conversion matrix estimate (CME) of $X_t(\beta)$. We can write the error of β^j as

$$E_t^j = X_t^j(\beta) - \hat{X}_t^j(\beta|\alpha). \tag{13}$$

The estimator for the total of α^ℓ can be also written as

$$\hat{X}_t^\ell = \frac{M^\ell}{m_\ell} \sum_{\alpha(i)=\ell}^{N} x_{it} I_i,$$

where I_i is a sample membership indicator that equals one if the unit i is in the sample and zero otherwise.

Taking into account that $\hat{X}_t^j(\beta|\alpha) = \Pi_T \hat{X}_t^\ell(\alpha)$, we obtain

$$E_t^j = \sum_{\beta(i)=j} x_{it} - \sum_\ell \pi_D^{j/\ell} \frac{M^\ell}{m_\ell} \sum_{\alpha(i)=l} x_{it} I_i = \sum_{i=1}^{N} b_i^j x_i - \sum_{i=1}^{N} \sum_\ell \pi_D^{j/\ell} \frac{M^\ell}{m_\ell} a_i^\ell x_i I_i,$$

where b_i^j and a_i^ℓ are indicator variables such that $b_i^j = 1$ and $a_i^\ell = 1$ if and only if i belongs to α^ℓ and β^j. Let us introduce the following notation,

$$e_i^j = b_i^j - \sum_\ell \pi_D^{j/\ell} \frac{M^\ell}{m^\ell} a_i^\ell I_i.$$

Then the error can be expressed more succinctly as

$$E_t^j = \sum_{i=1}^{N} x_{i,t} e_i^j. \tag{14}$$

Our aim is to estimate the mean squared error, that is, $\mathbf{E}[(E_t^j)^2]$.

Now, we propose a model for the relationship between the values of the variable under study at the double classification period D and their values at t.

In particular, we express $x_{i,t}$ as the product of the value at time D, $x_{i,D}$ times the sum of one plus a random term $\zeta_{i,t}$, that is,

$$x_{i,t} = x_{i,D}(1 + \zeta_{i,t}) \quad \text{or} \quad x_{i,t} = x_{i,D} + \eta_{i,t}, \tag{15}$$

where $\eta_{i,t} = x_{i,D}\zeta_{i,t}$.

Let us denote by $\eta_t^{\alpha,\ell} = \sum_{\alpha(i)=\ell} \eta_{i,t}$, $\eta_t^{\beta,j} = \sum_{\beta(i)=j} \eta_{i,t}$, where i runs along the population. By ζ, we mean the matrix $\{\zeta_{i,t}\}_{i,t}$. We will make the following assumptions on their moments.

Assumption 1. *The moments of $\zeta_{i,t}$ depend only on the class of i according to α, that is, we can write*

$$\mathbf{E}[\zeta_{i,t}] = \mu_t^{\alpha(i)} \tag{16}$$

$$\mathbf{Var}[\zeta_{i,t}] = \sigma_t^{2.\alpha(i)}. \tag{17}$$

This assumption is necessary to prevent the results to depend on unknown moments that we would not be able to estimate without having the units classified according to β.

Assumption 2. *If $i \neq i'$, then $\zeta_{i,t}$ and $\zeta_{i',t}$ are independent.*

This assumption is restrictive, but we consider that it is an acceptable simplification for a first approach. It remains for future work to allow for a certain degree of dependence across i.

Assumption 3. *The selection of the sample is independent from ζ.*

We will first decompose the error in two terms.

Proposition 1. *If assumptions 1, 2 and 3 hold, then the expected squared error can be decomposed as*

$$\mathbf{E}\left[\left(E_t^j\right)^2\right] = \mathbf{E}[\mathbf{E}[E_t^j|\zeta]^2] + \mathbf{E}[\mathbf{V}[E_t^j|\zeta]], \tag{18}$$

where

$$\mathbf{E}[E_t^j|\zeta] = \eta_t^{\beta,j} - \sum_\ell \pi_D^{j/\ell} \eta_t^{\alpha,\ell} \tag{19}$$

$$\mathbf{V}[E_t^j|\zeta] = \sum_\ell (\pi_D^{j/\ell})^2 V^\ell \tag{20}$$

and V^ℓ is the variance in the sampling distribution of the estimator $\hat{X}_t^\ell(\alpha)$.

Proof. We can see that (18) holds as follows.

$$\begin{aligned}
\mathbf{E}\Big[\big(E_t^j\big)^2\Big] &= \mathbf{E}\Big[\mathbf{E}[\big(E_t^j\big)^2|\zeta]\Big] = \\
&= \mathbf{E}\Big[\mathbf{E}[\big(E_t^j\big)^2|\zeta]\Big] - \mathbf{E}\Big[\mathbf{E}[\big(E_t^j\big)|\zeta]^2\Big] + \mathbf{E}\Big[\mathbf{E}[\big(E_t^j\big)|\zeta]^2\Big] = \\
&= \mathbf{E}\left[\mathbf{E}[\big(E_t^j\big)^2|\zeta] - \mathbf{E}[\big(E_t^j\big)|\zeta]^2\right] + \mathbf{E}\Big[\mathbf{E}[\big(E_t^j\big)|\zeta]^2\Big] = \\
&= \mathbf{E}\Big[\mathbf{V}[E_t^j|\zeta]\Big] + \mathbf{E}\Big[\mathbf{E}[\big(E_t^j\big)|\zeta]^2\Big]
\end{aligned}$$

where the first identity holds by the law of total expectation and the last one by the definition of $\mathbf{Var}[\cdot|\zeta]$.

Let us now analyze $\mathbf{E}[E_t^j|\zeta]$. From (12), (13) and the unbiasedness of $\hat{X}_t^\ell(\alpha)$ we have

$$\mathbf{E}[E_t^j|\zeta] = X_t^j(\beta) - \sum_\ell \pi_D^{j/\ell} X_t^\ell(\alpha). \tag{21}$$

On the other hand,

$$X_t^\ell(\alpha) = X_D^\ell(\alpha) + \eta_{\alpha,t}^\ell \qquad X_t^j(\beta) = X_D^j(\beta) + \eta_{\beta,t}^j.$$

Thus, we can replace $X_t^j(\beta)$ and $X_t^\ell(\alpha)$ in (13) and we get

$$\mathbf{E}[E_t^j|\zeta] = \Big\{X_D^j(\beta) - \sum_\ell \pi_D^{j/\ell} X_D^\ell(\alpha)\Big\} + \Big\{\eta_t^{\beta,j} - \sum_\ell \pi_D^{j/\ell}\eta_t^{\alpha,\ell}\Big\}.$$

We can see that the first term is equal to zero,

$$\begin{aligned}
X_D^j(\beta) - \sum_\ell \pi_D^{j/\ell} X_D^\ell(\alpha) &= X_D^j(\beta) - \sum_\ell \frac{X_D^{j,\ell}(\alpha,\beta)}{X_D^\ell(\alpha)} X_D^\ell(\alpha) = \\
&= X_D^j(\beta) - \sum_\ell X_D^{j,\ell}(\alpha,\beta) = 0,
\end{aligned}$$

so we arrive to (19). It only remains to check (20), but this amounts to realize that conditional to ζ, E_t^j is a linear combination of the errors of the estimators $\hat{X}_t^\ell(\alpha)$. □

For the practical application of proposition 1, of the two terms of the decomposition (18), the second can be replaced by its sample counterpart, but in order to obtain the first one, we will have to do some further work, which is summarized in the following proposition.

Proposition 2. *In the assumptions of proposition 1,*

$$\mathbf{E}[\mathbf{E}[E_t^j|\zeta]^2] = \sum_\ell \sigma_t^{2,\ell}\big\{(1-2\pi_D^{j/\ell})Q_D^{\ell,j} + (\pi_D^{j/\ell})^2 Q_D^\ell\big\} \tag{22}$$

where

$$\begin{aligned} Q_D^{\ell,j} &= \sum_{\alpha(i)=\ell,\beta(i)=j} x_{i,D}^2, \\ Q_D^\ell &= \sum_{\alpha(i)=\ell} x_{i,D}^2. \end{aligned}$$

Proof. We have to calculate the expectation of the squared conditional expectation in (19).

$$\begin{aligned} & \mathbf{E}[(\eta_t^{\beta,j} - \sum_\ell \pi_D^{j/\ell}\eta_t^{\alpha,\ell})^2] = \\ = \;& \mathbf{E}\Big[\big(\sum_{i=1}^N b_i^j \eta_{i,t} - \sum_{i=1}^N \sum_\ell \pi_D^{j/\ell} a_i^\ell \eta_{i,t}\big)^2\Big] = \\ = \;& \mathbf{E}\left(\Big[\sum_{i=1}^N \eta_{i,t}(b_i^j - \sum_\ell \pi_D^{j/\ell} a_i^\ell)\Big]^2\right) = \\ = \;& \mathbf{E}\left(\sum_{i,i'} \eta_{i,t}\eta_{i',t}(b_i^j - \sum_\ell \pi_D^{j/\ell} a_i^\ell)(b_{i'}^j - \sum_\ell \pi_D^{j/\ell} a_{i'}^\ell)\right) = \\ = \;& \underbrace{\sum_{i,i'}(b_i^j - \sum_\ell \pi_D^{j/\ell} a_i^\ell)(b_{i'}^j - \sum_\ell \pi_D^{j/\ell} a_{i'}^\ell)\mu_t^{\alpha(i)} x_{i,D}\mu_t^{\alpha(i)} x_{i',,D}}_{(A)} + \\ & \underbrace{+\sum_i (b_i^j - \sum_\ell \pi_D^{j/\ell} a_i^\ell)^2 \sigma_t^{2,\alpha(i)} x_{i,D}^2}_{(B)} \end{aligned}$$

We calculate the two terms separately.

(A)

$$
\begin{aligned}
& \sum_{i,i'}(b_i^j - \sum_\ell \pi_D^{j/\ell} a_i^\ell)(b_{i'}^j - \sum_\ell \pi_D^{j/\ell} a_{i'}^\ell)\mu_t^{\alpha(i)} x_{i,T} \mu_t^{\alpha(i)} x_{i',T} \\
= & \Big[\sum_i (b_i^j - \sum_\ell \pi_D^{j/\ell} a_i^\ell)\mu_t^{\alpha(i)} x_{i,D}\Big]^2 \\
= & \Big[\sum_\ell \sum_{\alpha(i)=\ell} \big(b_i^j - \pi_D^{j/\ell}\big)\mu_t^{\alpha(i)} x_{i,D}\Big]^2 \\
= & \left[\sum_\ell \mu_t^\ell \Big\{ \sum_{\alpha(i)=\ell} b_i^j x_{i,D} - \pi_D^{j/\ell} \sum_{\alpha(i)=\ell} x_{i,D}\Big\}\right]^2 \\
= & \left[\sum_\ell \mu_t^\ell \Big\{ X_D^{\ell,j}(\alpha,\beta) - \pi_D^{j/\ell} X_D^\ell(\alpha)\Big\}\right]^2 = 0
\end{aligned}
$$

(B)

$$
\begin{aligned}
& \sum_i (b_i^j - \sum_\ell \pi_D^{j/\ell} a_i^\ell)^2 \sigma_t^{2,\alpha(i)} x_{i,D}^2 \\
= & \sum_\ell \sum_{\alpha(i)=\ell} \big(b_i^j - \pi_D^{j/\ell}\big)^2 x_{i,D}^2 \sigma_t^{2,\ell} \\
= & \sum_\ell \sigma_t^{2,\ell} \sum_{\alpha(i)=\ell} \big(b_i^j - 2 b_i^j \pi_D^{j/\ell} + (\pi_D^{j/\ell})^2\big) x_{i,D}^2 \\
= & \sum_\ell \sigma_t^{2,\ell} \sum_{\alpha(i)=\ell} \big(b_i^j x_{i,D} - \pi_D^{j/\ell} x_{i,D}\big)^2 \\
= & \sum_\ell \sigma_t^{2,\ell} \big\{(1 - 2\pi_D^{j/\ell}) Q_D^{\ell,j} + (\pi_D^{j/\ell})^2 Q_D^\ell\big\},
\end{aligned}
$$

so we conclude. □

Among the quantities involved in (22), $\pi_D^{j/\ell}$, $Q_D^{\ell,j}$ and Q_D^ℓ can be estimated with the micro-data of the double-coding periods. The variances $\sigma_t^{2,\ell}$ pose more difficulties, since they require at least a subsample of units common to both periods t and D. In section 4, we show that this sample does not need to be very large.

If we had the full sample at both t and D, we might estimates μ_t^ℓ and $\sigma_t^{2,\ell}$

as follows,

$$\hat{\mu}_t^\ell = \frac{1}{m^\ell} \sum_{i \in S^\ell} \zeta_{i,t} \tag{23}$$

$$\hat{\sigma}_t^{2,\ell} = \frac{1}{m^\ell} \sum_{i \in S^\ell} (\zeta_{i,t} - \hat{\mu}_t^\ell)^2. \tag{24}$$

If we only have a subsample S_1^ℓ of m_1^ℓ units common to t and D, we would replace m^ℓ and S^ℓ by m_1^ℓ and S_1^ℓ, in (23)–(24).

4 Application

In this section, we will show the results of an application of the previous analysis to real data. Now, the old and new classifications will be NACE rev. 1.1 and NACE rev. 2, both of them at the division (two digit) level.

The part of (19) due to the sampling errors of $\hat{X}_t^\ell(\alpha)$ can be easily calculated and it is not the main concern for us. Hence, we will assume that our sample is the whole population, so the sampling errors are null and we focus on the first term in (19).

On the other hand, if we intend to assess the validity of the estimates of proposition 2, we need to compare the estimates to the true errors, that is, we need the actual values of $X_t^j(\beta)$ for some t outside the double classification period. We can do this because we have available data classified according to both classifications for several years.

In 4.1, we describe the survey to which we apply the previous results and the process of reclassification of the individual units, in 4.2, we analyze the estimation of the variances $\sigma^{2,\ell}$ and in 4.3 we discuss the results.

4.1 The data

We have used for our study the data from the Industrial Turnover and New Orders survey. Let us describe the main features of the survey.

- The data are requested in a monthly basis and the series start in January 2002.

- The sample contains around 13.500 units (the statistical unit is the establishment). It is not a probabilistic sample, but a cutoff one. The criterion

for a unit to be included is the number of employees. Consequently, the sample is quite stable along the time.

In order to reclassify the units, we have used the following sources,

- One - to - one relationships between NACE rev. 1.1 and NACE rev. 2.
- PRODCOM data that provided value of production for all products of each establishment.
- Structural Statistics that provided double-classification at the enterprise level (only 2005).
- Manual work

The sample of the Turnover/New Orders survey is not a probabilistic one. Instead, all units with the variable 'number of employees' over a certain threshold are included. Consequently, but for the small number of unclassified units, the recalculated totals are the same as those we would have produced during the lifetime of the survey if we had used the NACE rev. 2. For a few divisions, the number of units was too low to obtain reliable results and therefore they have been excluded of the study.

4.2 Estimation of the variances

The estimation of the variances $\sigma_t^{2,\ell}$ posed some difficulties. The distributions of the ζ_t^i have long tails on the right side (skewness often above ten and reaching in some cases 15) and the naive estimator proposed in (23)–(24) is not robust enough. On the other hand, the logarithm-transformed distributions look more familiar, as we can see in figure 1. Thus, we assume a log-normal model. In order to make the estimates even more robust, we use the 0.01-trimmed mean, that is, we exclude the 1% extreme values on the right and the left tails. Thus,

$$\xi_{i,t} = \log \zeta_{i,t} \tag{25}$$

$$\hat{\mu}_{\xi,t}^{\ell} = \frac{1}{m^{\ell}} \sum_{i \in S^{\ell,*}} \xi_{i,t} \tag{26}$$

$$\hat{\sigma}_{\xi,t}^{2,\ell} = \frac{1}{m^{\ell}} \sum_{i \in S^{\ell,*}} (\xi_{i,t} - \hat{\mu}_{\xi,t}^{\ell})^2 \tag{27}$$

$$\hat{\mu}_t^{\ell} = \exp\left(\hat{\mu}_{\xi,t}^{\ell} + \frac{1}{2}\hat{\sigma}_{\xi,t}^{2,\ell}\right) \tag{28}$$

$$\hat{\sigma}_t^{2,\ell} = \left\{ \exp\left(\hat{\sigma}_{\xi,t}^{2,\ell}\right) - 1 \right\} \exp\left(2\hat{\mu}_{\xi,t}^{\ell} + \hat{\sigma}_{\xi,t}^{2,\ell}\right), \tag{29}$$

where by $S^{\ell,*}$ we mean the sample excluding the extreme values.

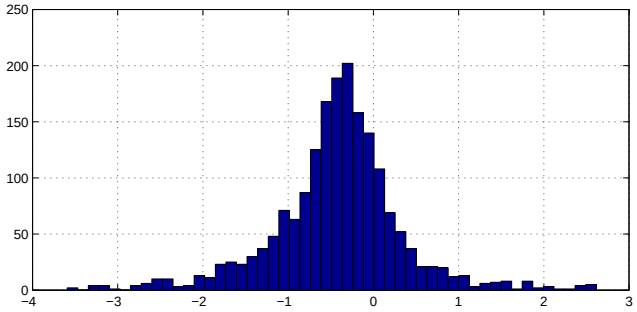

Figure 1: Histogram of $\log \zeta$ for $t = 1$ and $\ell = 5$. The 1% extreme values on both sides are removed.

The assumption that we have at our disposal the micro-data to estimate μ_t^ℓ and $\sigma_t^{2,\ell}$ according to the method we have described is in contradiction with an underlying assumption that makes necessary the CMM, namely, that we do not know how to relate the units in t and D (since if we were able to, then most of the unit reclassification work would be done, besides the problem of units that change their activity along time). Consequently, we have made an estimation under the more restrictive assumption of equal moments for all activities. In other words, we estimate common μ_t and σ_t for all ℓ. This allows for a more realistic application of the method, since we can draw a subsample and estimate σ_t with it.

4.3 Results

We have obtained the real and estimate error for the divisions, using different $\hat{\sigma}^2, \ell_t$ estimates for each division (table 1). In tables 2, 3 and 4, we present the results when a common $\hat{\sigma}_t^2$ is estimated for all ℓ, using the whole sample, a subsample of about 5% and one of about 1%, respectively.

Even in the worst case (table 4), the estimates provide a reasonable guide about the level of magnitude of the error. If we regress the 28 real errors on their estimates, we find that the estimates explain most of the variation of the error ($R^2 = 0.82$), but the trend is 0.34, so we are overestimating the error by a factor of around three. This means that at this point, the estimates are useful

to: (a) know the level of magnitude of the errors and (b) detect the classes where the greatest quality problems are. The analysis of the sources of the detected overestimation of the error remains for future research, together with the generalization to estimators other than the Horvitz-Thompson.

References

[1] Van den Brakel, J. (2009). Sampling and estimation techniques for the implementation of the NACE Rev. 2 in Business Surveys. Statistics Netherlands Discussion paper (09013).

[2] Yuskavage, R. (2007). Converting historical industry time series data from SIC to NAICS. BEA Papers, Bureau of Economic Analysis.

[3] James, G. (2008). Backcasting, for use in Short Term Statistics, UK Office for National Statistics.

[4] Buiten, G., J. Kampen and S. Vergouw (2009). Producing historical time series for STS-statistics in NACE Rev. 2: Theory with an application in industrial turnover in the Netherlands (1995 - 2008). Statistics Netherlands Discussion paper (09001).

[5] Caporin, M. and D. Sartore (2006). Methodological aspects of time series back-calculation. Proceedings of the Workshop on frontiers in benchmarking techniques and their application to official statistics, Luxembourg, 6–7 April, Eurostat Working Paper series.

[6] Angelini, E., J. Henry and M. Marcellino (2003). Interpolation and backdating with a large information set. European Central Bank Working Paper series, no. 252.

div.	real	estimate	div.	real	estimate
5	0.0222	0.0549	21	0.0898	0.1416
6	0.4223	3.5794	22	0.0166	0.0166
8	0.1159	0.1695	23	0.0101	0.0250
10	0.0352	0.0251	24	0.0028	0.0093
11	0.1528	0.1071	25	0.0270	0.0276
12	0.0000	0.0000	26	0.0211	0.1051
13	0.1172	0.0595	27	0.1567	0.1300
14	0.1045	0.0420	28	0.1332	0.1368
15	0.0278	0.0409	29	0.0143	0.0207
16	0.0039	0.0115	30	0.0783	0.3252
17	0.0109	0.0068	31	0.0813	0.0791
18	0.0967	0.1159	32	0.0698	0.1449
19	0.0065	0.0083	33	0.1437	0.4889
20	0.0414	0.0611	35	0.0819	0.0112

Table 1: Relative error estimated and real (moments depending on ℓ).

div.	real	estimate	div.	real	estimate
5	0.02223	0.04334	21	0.08979	0.17317
6	0.42233	1.05881	22	0.01658	0.02009
8	0.11588	0.15840	23	0.01007	0.01965
10	0.03519	0.02851	24	0.00285	0.00954
11	0.15281	0.12107	25	0.02698	0.02651
12	0.00000	0.00000	26	0.02108	0.06543
13	0.11717	0.05375	27	0.15671	0.12403
14	0.10445	0.03690	28	0.13318	0.12386
15	0.02779	0.02650	29	0.01426	0.01942
16	0.00386	0.01082	30	0.07833	0.17680
17	0.01088	0.00754	31	0.08131	0.07141
18	0.09674	0.12690	32	0.06984	0.12924
19	0.00652	0.01671	33	0.14367	0.27365
20	0.04138	0.07629	35	0.08187	0.00900

Table 2: Relative error estimated and real (same moment estimates for every ℓ).

div.	real	estimate	div.	real	estimate
5	0.02223	0.04419	21	0.08979	0.17635
6	0.42233	1.08211	22	0.01658	0.02044
8	0.11588	0.16108	23	0.01007	0.02000
10	0.03519	0.02899	24	0.00285	0.00969
11	0.15281	0.12315	25	0.02698	0.02695
12	0.00000	0.00000	26	0.02108	0.06666
13	0.11717	0.05466	27	0.15671	0.12609
14	0.10445	0.03753	28	0.13318	0.12593
15	0.02779	0.02703	29	0.01426	0.01976
16	0.00386	0.01100	30	0.07833	0.18005
17	0.01088	0.00767	31	0.08131	0.07268
18	0.09674	0.12924	32	0.06984	0.13151
19	0.00652	0.01694	33	0.14367	0.27834
20	0.04138	0.07763	35	0.08187	0.00918

Table 3: Relative error estimated and real (same moment estimates for every ℓ) in a subsample of 5% of the whole sample.

div.	real	estimate	div.	real	estimate
5	0.02223	0.04997	21	0.08979	0.19948
6	0.42233	1.20207	22	0.01658	0.02295
8	0.11588	0.18255	23	0.01007	0.02252
10	0.03519	0.03289	24	0.00285	0.01106
11	0.15281	0.13883	25	0.02698	0.03041
12	0.00000	0.00000	26	0.02108	0.07526
13	0.11717	0.06198	27	0.15671	0.14275
14	0.10445	0.04268	28	0.13318	0.14273
15	0.02779	0.03001	29	0.01426	0.02227
16	0.00386	0.01240	30	0.07833	0.20300
17	0.01088	0.00862	31	0.08131	0.08182
18	0.09674	0.14550	32	0.06984	0.14815
19	0.00652	0.01946	33	0.14367	0.31525
20	0.04138	0.08778	35	0.08187	0.01020

Table 4: Relative error estimated and real (same moment estimates for every ℓ) in a subsample of 1% of the whole sample.