eurostat

Doc. Class/05/UK/03 1 July 2005

Methodological challenges in implementing the new industrial classification

EN

NACE/CPA Implementation Task Force meeting 4-6 July 2005
Agenda item 4

STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES

02 Statistical Governance, quality and evaluation
Tel. (352) 4301-32023, Fax (352) 4301-33899

Methodological challenges in implementing the new industrial classification

1. Introduction

This paper is an excerpt from a larger paper written on implementing the new industrial classification. This paper focuses only on the methodological challenges of implementaiton, although the section on the Business Register is retained since it is of some use to the implementation task force due to the automatic coding issue.

The paper deals with implementation in the context of:

- The Business Register
- Sampling
- Weighting (in the context of estimation)
- Backdata

2.1 Business Register

The amount of Business Register work that can be carried out on implementing the new classification is quite limited until the UK version of NACE, SIC(2007), is ready. A major part of the Business Register's work will be in equipping our electronic coding tool to effectively code businesses on the new classification.

Until now ONS has used an automatic coding tool called the Precision Data Coder (PDC) which reads business descriptions, usually obtained from the Annual Register Inquiry, and automatically codes the business into a five-digit SIC(2003) industry. ONS has decided to change the coding tool, from the PDC to one built by Statistics Canada called Automatic Coding by Text Recognition (ACTR). This change is independent of the new classification in the sense that it will be introduced initially to code to SIC(2003), with an updated version to be introduced at the same time as the new classification.

ACTR compares incoming business descriptions against a reference database when seeking to obtain a match and get an SIC code. The new reference database will be updated using codes from the new classification. The ACTR work will concentrate on those classifications where one SIC(2003) translates into several SIC(2007) codes ('one to many') whilst those that map to a single SIC(2007) ('one to one' or 'many to one') should be re-coded automatically from SIC(2003) to SIC(2007). If ACTR cannot supply a code (including cases for which we do not have a business description) another method will be required. Currently the most likely solution is to reassign such cases using probabilistic correlation tables.

The new SIC structure, including subclasses, will be available in the latter half of 2005, with the explanatory notes that accompany them shortly afterwards. These notes are essential for coders to interpret and code business descriptions accurately, the first actual electronic publications will be available by July 2006. Against this, Business Register colleagues' preliminary requirements suggests they will need a version of ACTR built on the new basis by around summer 2006. Development of ACTR and the electronic publications will be taken forward in parallel.

As covered above, a big benefit of developing the ACTR knowledge base for SIC(2003) is that future work for SIC(2007) will move more quickly.

2.2 Sampling

All business surveys are currently selected from the IDBR according to SIC(2003). It will be necessary to redesign these surveys to be able to be selected according to SIC(2007).

Most of ONS's business surveys operate with stratified simple random samples. Stratification is usually by a fairly fine level of SIC(2003) detail and between four and six sizebands based on employment values held on the IDBR. Allocation of the total sample size to strata is usually done by the Neyman Optimal Allocation method (Neyman 1934) where the sample size, n_h, in stratum h is:

$$n_h = n \frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h}$$

where L is the number of strata in the population, N_h is the number of elements in stratum h in the population and S_h^2 is the variance of elements in stratum h in the population according to the estimation model chosen for the survey.

Given that each business in the population will be reclassified to SIC(2007) and will therefore have a new code, we can determine the population size in each of the new strata. Since S_h^2 relates to the value in the population at large, we usually estimate this by s_h^2 , the variance of elements in stratum h in the sample.

However, under newly-defined strata, we may not have these for some strata, so alternative approaches will need to be examined. One option is to produce these estimates from the relevant businesses making up each new stratum, according to the weight each business had in the original survey. In practice however, we find that values of s_h^2 are often too variable between strata to use them directly, so it is necessary to use an average of previous sample variances, or to model stratum-level estimates of variance against Business Register counts such as stratum size and the totals of employment and turnover. Such a modelled approach would likely work well in the situation where we have reconstituted strata since new 'variances' can be produced according to the characteristics of any of stratum, however designed. Other alternatives such as x-optimal allocation (Sarndal et al 1992) could also be considered on a similar basis.

A further complication to the redesign of samples under a new classification is that redesigning samples is resource intensive, and it may be impractical to reallocate all samples adequately in the time allowed between new SIC(2007) becoming available and the need to select samples. In this case, alternative proxies may need to be sought to transition between the SICs. One option that may be possible is for the existing sample to be tabulated against the new strata, and using the number in each stratum as the new sample size. Of course this won't lead to an optimal solution, but

the allocation procedure is such that reasonably large deviations can be made from optimality with only a small impact on the quality of estimates produced.

2.3 Weighting

2.3.1 Theory

This section sets out some options relating to weighting during the change to SIC(2007).

To prepare the way, we present a summary of calibration estimation, as implemented in ONS.

Let {*1,…,k,...,N*} be the set of labels that uniquely identify the *N* distinct elements of a target finite population **U**. Without loss of generality, let **U** = {*1,…,k,...,N*}. A survey is carried out to measure the values of *J* survey variables. Denote by $\mathbf{y}_k = (y_{k1}, \ldots, y_{kJ})'$ the *J×1* vector of values of the survey variables for the *k*th population element.

We assume that the primary purpose of the survey is to estimate the population vector of totals $\mathbf{T}_y = \sum_{k \in \mathbf{U}} \mathbf{y}_k = \mathbf{Y}_\mathbf{U}' \mathbf{1}_N$ where $\mathbf{Y_U}$ denotes the *N×J* population matrix of *y* values given by $\mathbf{Y}_\mathbf{U} = [\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_N]'$, and $\mathbf{1}_N$ denotes the *N×1* vector of ones.

We assume that *n* distinct elements in **U** are included in a sample *s*, $s = \{k_1, \ldots, k_n\} \subset \mathbf{U}$, which is selected for observation in the survey. Hence the purpose of the survey is to estimate $\mathbf{T}_y$ on the basis of the available survey data {$\mathbf{y}_k$; $k \in s$}. The "standard" estimator for totals when these are the only data available from the sample is the Horvitz-Thompson (H-T) estimator defined as

$$\hat{\mathbf{T}}_y = \sum_{k \in s} d_k \mathbf{y}_k$$

where $d_k = 1/\pi_k$ is the design weight for unit *k*, and π_k is the sample inclusion probability for unit *k*. In most survey applications, however, some auxiliary variables $\mathbf{x}_k = (x_{k1}, \ldots, x_{kp})'$ may be available, which may help improve the estimation of the target parameter $\mathbf{T}_y$.

One way to do this is by calibration. The key idea behind calibration estimation is as follows. Although we know the population totals for the *x* variables, suppose we would try to estimate them from the sample, using the H-T estimator. This would lead to the estimation of $\mathbf{T}_x$ by $\hat{\mathbf{T}}_x = \sum_{k \in s} d_k \mathbf{x}_k$. However, these estimates $\hat{\mathbf{T}}_x$ often would not match the corresponding population totals $\mathbf{T}_x$ exactly, leading to the so-called "calibration error" $\hat{\mathbf{T}}_x - \mathbf{T}_x$. We modify the estimator to avoid this "error", and use a "calibrated" estimator where the design weights d_k are modified, leading to new weights w_k to be used in the calibrated estimator

$$\hat{\mathbf{T}}_{xC} = \sum_{k \in s} w_k \mathbf{x}_k$$

where $\{w_k, k \in s\}$ are case weights such that there is no calibration error, i.e. satisfying

$$\hat{\mathbf{T}}_{xC} - \mathbf{T}_x = \sum_{k \in s} w_k \mathbf{x}_k - \mathbf{T}_x = \mathbf{0}$$

These conditions are called the "calibration constraints". The idea is that if the "calibrated" weights $\{w_k, k \in s\}$ succeed in reducing or avoiding error when "estimating" the x totals, they may also reduce the error when estimating the y totals, using the calibration estimator:

$$\hat{\mathbf{T}}_{yC} = \sum_{k \in s} w_k \mathbf{y}_k$$

A large number of sets of weights $\{w_k, k \in s\}$ may satisfy the calibration constraints given the sample data $\mathbf{X}_s$, the design weights $\{d_k, k \in s\}$ and the population totals $\mathbf{T}_x$. One way of selecting those that lead to "reasonable" sets of weights is to think of calibration weights w_k as modifications to the design weights d_k that change them the least. This is justified because using the design weights d_k provides the corresponding H-T estimator with desirable properties such as design-unbiasedness and consistency (in the sense that as the sample size increases, the estimator converges in probability towards the right target $\mathbf{T}_y$).

Deville and Särndal (1992) defined a family of calibration estimators for $\mathbf{T}_y$ where the weights w_k are chosen such that specified distance functions measuring how far the w_k are from the d_k are minimised. Their idea is to minimise

$$E_P\left(\sum_{k \in s} G_k(w_k, d_k)\right)$$

or equivalently minimise, for every sample s,

$$\sum_{k \in s} G_k(w_k, d_k)$$

subject to $\hat{\mathbf{T}}_{xC} - \mathbf{T}_x = \sum_{k \in s} w_k \mathbf{x}_k - \mathbf{T}_x = \mathbf{0}$, where $G_k(w_k, d_k)$ is a measure of the distance between w_k and d_k satisfying some regularity conditions to be specified later, and E_P denotes the expectation with respect to the probability distribution induced by the sampling design used to select the sample s.

One popular choice for the distance function is to take

$$G_k(w_k, d_k) = \frac{(w_k - d_k)^2}{q_k d_k} \quad k \in s$$

for some known constants $q_k > 0$, $k \in s$, to be specified. In this case, the solution is given by

$$w_k = d_k \times g_k$$

where

$$g_k = 1 + q_k \left(\mathbf{T}_x - \hat{\mathbf{T}}_x\right)' \left(\sum_{i \in s} q_i d_i \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \mathbf{x}_k$$

With the weights w_k, the resulting calibration estimator for the total of a survey variable y_j can be written as

$$\hat{T}_{y_j C} = \sum_{k \in s} w_k y_{kj} = \hat{T}_{y_j} + \left(\mathbf{T}_x - \hat{\mathbf{T}}_x\right)' \hat{\mathbf{B}}_j$$

where $\hat{T}_{y_j} = \sum_{k \in s} d_k y_{kj}$ is the H-T estimator for $T_{y_j} = \sum_{k \in \mathbf{U}} y_{kj}$ and $\hat{\mathbf{B}}_j$ is defined as

$$\hat{\mathbf{B}}_j = \left(\sum_{k \in s} q_k d_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \left(\sum_{k \in s} q_k d_k \mathbf{x}_k y_{kj} \right)$$

2.3.2 A simple example

Suppose a sample of size n=50 is drawn from a population of N=1000 units. The design weight ($d_k = 1/\pi_k$) is therefore 20 and the H-T estimator of the population total of the study variable y is:

$$\hat{\mathbf{T}}_y = \sum_{k \in s} 20 \mathbf{y}_k$$

Now suppose that we have an auxiliary variable, for which the population total $\mathbf{T}_x = \sum_{k \in \mathbf{U}} \mathbf{x}_k = 41{,}000$ and the sample total $\sum_{k \in s} \mathbf{x}_k = 2{,}000$. Now the H-T estimator of the auxiliary total based on the sample is $\hat{\mathbf{T}}_x = \sum_{k \in s} 20 \times 2{,}000 = 40{,}000$, which has a calibration error of $\hat{\mathbf{T}}_x - \mathbf{T}_x = 1{,}000$. The modifying g-weight is calculated in this case as the simple $g_k = \dfrac{\hat{\mathbf{T}}_x}{\mathbf{T}_x} = 1.025$, whose use leads to a calibration error-free estimate.

2.3.3 Application to classification change

We have identified three options for applying calibration weighting in the context of the classification change. First, we first outline some basic assumptions as follows.

- There will be a year during which the frame will be classified to both systems at the unit level - assume this is year 1 (changeover year). Note that as sample selection will be based on a design incorporating only one of these systems (probably the former classification) then the design weights (a-weights) will be fixed by this design.
- There will be a requirement for aggregates to be produced on both old and new classifications for all years prior to the change year. This is described in the following section on back series.
- That during year 1 that selection is based on the old classification system and that for following years on the new system.
- There will be a requirement for aggregates to be produced on both old and new classifications during the changeover year.
- There will be a requirement for aggregates to be produced only on the new classifications after the change year.

We now outline the three options. Note that in each case, the calibration approach results in a single weight (the product of a and g) for each business, so aggregates,

for whatever domain, are simply the products of the weight and the survey variable, summed over all relevant businesses in the domain.

Option 1

Year 1

- Calculate calibration factors (g-weights) using the old classification.
- Produce results using conventional estimation for the old classification and by domain estimation for the new classification.

Year > 1

- Calculate calibration factors (g-weights) using the new classification.
- Produce results using conventional estimation for the new classification..

Pros

- Completely consistent with the old series (years earlier than 1- i.e. no discontinuity in the time series going backwards)
- Gives the new classification on the Business Register time (a year) to settle down
- Totals for equivalent classifications (those that haven't changed between SIC(2003) and SIC(2007)) will be the same.
- Weighting is consistent with design (selection).

Cons

- There may be a discontinuity in the year following the change; this depends on the size of the difference between the classification systems.

Option 2

Year 1+

- Calculate calibration factors (g-weights) using the new classification.
- Produce results using conventional estimation for the new classification and by domain estimation for the old classification.
- Variances for the old classification domains would need to be calculated differently (domain estimates) to those under the new system.

Pros

- Completely consistent with the new series (no discontinuity in the time series going forwards)
- Any discontinuity taken as one hit in the changeover year.
- Weighting for subsequent years is the same as for year 1.
- Again totals for equivalent classifications (those that haven't changed between SIC(2003) and SIC(2007)) will be the same.

Cons

- The new classification on the Business Register may not have settled down so there may be issues with outliers or other unusual results during year 1.
- Weighting is not consistent with design (selection) in year 1.

Option 3

Year 1

- Calculate calibration factors (g-weights) using both classification systems. In this case the population totals are reproduced by summing the weighted employment (turnover) for both classifications.

- Note that the variances for both classifications would need to be calculated using Statistics Canada's Generalized Estimation System (GES) (Estevao et al 1995).

Year > 1

- Calculate calibration factors (g-weights) using the new classification only.
- Produce results using conventional estimation.

Pros

- Discontinuity should be minimised in both years since the calibration totals are reproduced under both classification systems in year 1. This is conditional on there being some correlation between the output variables and the chosen auxiliary.
- Gives the new classification on the Business Register time (a year) to settle down
- Totals for equivalent classifications (those that haven't changed between SIC2003 and SIC2007) will be the same.
- Weighting is consistent with design (selection).

Cons

- If the classifications are radically different there may be a problem with extreme weights in year 1. (For example if there happens to be a very small sample in one of the new classifications in year 1 since selection was carried out using the old classification).

Summary and Discussion of Alternatives

All three options are can be sensibly applied during a classification change and have been listed in increasing order of risk and benefit.

For option 1 the main disadvantage is that discontinuity will arise in the year following the classification change, whereas it may be considered more sensible to have the discontinuity coincide with the strict date of the changeover. The main advantage to option 1 is that the maximum time is allowed for the new classification to settle down before it is used for weighting.

Option 2 moves the discontinuity a year earlier so that there should be consistency between years 1 and 2; the discontinuity therefore takes place during the same period that the classification is changed. There is some risk here due to using the new classification on the Business Register a year earlier than in option 1.

The main risk with option 3 is that some unexpected weights are produced in year 1. This is especially true for variables that are not correlated (or negatively correlated) with the auxiliary variable (employment or turnover).

2.4 Back series

A consistent back series is important for many users. There are clear difficulties with producing such a back series since it is impossible to be completely sure of the classifcation of any business at any point in history. Therefore, we need to ask what is practicable within the constraints of data and systems/resources availability. Two main options are considered.

First, if the SIC(2003) codes are known for individual businesses, new codes can be assigned at the individual level according to the following criteria:

- for as far back as the business has the same classification, we can assume that the current new SIC(2007) would be appropriate and use this value;
- for other situations, individual SIC(2007)s can be imputed according to a look-up table.

Then, once each record on the historical dataset has an SIC(2007) code, domain estimates can be produced as required.

A second, simpler, alternative is to use correlation matrices that record the relationship between the new industries and the old. Then, estimates can be produced on the new basis by taking old estimates and multiplying them by the appropriate conversion factors.

Both methods rely on assumptions that the current relationship between the old and new classifications is appropriate to apply to old data. Clearly, the farther back in time series are converted, the more quality issues will be associated with that. However, both provide a reasonable way of producing back series to meet users' needs.

The first method was recently applied in the Annual Survey of Hours and Earnings when the Standard Occupational Classification (SOC) changed from its 1990 revision to a 2000 revision. For records in the same job between years t and t-1 the SOC00 code from year t was carried back to year t-1. In addition for any records with the same SOC90 code in years t and t-1 the SOC00 code from year t was carried back to year t-1. The remaining codes were imputed according to the frequencies derived from the 2002 dataset, which was coded according to both classifications.

Good quality back series will depend on good quality measures of the correlation between the two systems. The use of the IDBR to code businesses on both levels means that Business Register data may be used for this purpose. Such a correlation matrix can also be formed based on the number of businesses, their employment or their turnover. The Annual Business Inquiry is a further source of possible correlation information, although the sample here is much smaller than that from the IDBR.

References

Estevao, V., Hidiroglou, M. A. and Särndal, C. E. (1995) Methodological principles for a generalized estimation system at Statistics Canada. *J. Off. Stat.* **11** 181-204.

Neyman, J. (1934) On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *J. R. Statist. Soc.* **97** 558-606.

Särndal, C. E, Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*. New York: Springer.