Task Force on the implementation of NACE Rev. 2

Handbook on methodological aspects related to sampling designs and weights estimations

Version 1.0

July 2006

Preface

Implementing the revised classification of economic activities, NACE Rev 2, will be a major task for all Member States. A substantial amount of work will need to be carried out in the fields of business registers, business surveys and national accounts. Some household surveys also will be affected. National Statistical Institute will face several major challenges, which include:

- coordination of the timing of the move across Member States
- dependencies of statistics
- handling of the national accounts move to the revised NACE
- reclassification of all units on the business register according to the revised NACE
- the difficulties of maintaining two classifications
- sampling under the new NACE
- estimation (weighting) under the new NACE
- simultaneous estimation and results assessment under both new and old classification
- construction of industry weights for short term statistics
- construction of back series in terms of the revised classification.
-

This document is the third of a set of four handbooks promoted by the Task Force "Implementation of NACE Rev. 2", devoted to:

1. The setting up of the implementation plan
2. Implementation of NACE Rev.2 in Business Registers
3. The methodological aspects related to sampling designs and weights estimations
4. Back-casting methodologies for the reconstruction of time series broken by a change of classification.

This set of handbooks aims at providing colleagues of National Statistical Institutes with suggestions, common practices, "checking lists", methodologies and similar tools which can be used during the complex project of the implementation of NACE Rev. 2.

This paper was drafted by Mark Williams (ONS- the United Kingdom), advised by other participants to the Task Force.

This handbook, as well as the others of the series, will be updated each time there is reason for that. The electronic version on the "Operation 2007" website http://forum.europa.eu.int/irc/dsis/nacecpacon/info/data/en/index.htm will always be the actual one.

Contents

1. Introduction

This paper is concerned with the challenges faced in sampling and estimation. It is worth noting at the outset that the change of classification requires extra resources for sampling and estimation due to:

- the re-design of sampling and estimation methods because of the change of domains (new economic activities in scope etc.)
- special modification of sampling and estimation methods to produce estimates simultaneously both by new and old classification
- (preferably) an increase in the overall sample size:
- The need of producing estimates reliable enough to produce results simultaneously both by new and old classification
- to compensate for the decrease in accuracy due to higher misclassification of economic activity in the first years of the new classification
- to allow for the increase of detail required under the new classification, according to the various European Regulations (eg Structural and Short Term Statistics Regulations). (This sample size increase may be permanent)
- evaluation of coherence between estimates by old and by new classification

The amount of work regarding sampling and estimations also depends strongly on the level of detail for back-casting.

On the other hand the change of NACE provides an excellent opportunity to substantially improve sample and estimation design.

2. Sampling under the new NACE

All business surveys are currently operated under NACE Rev 1.1, with the sample normally selected with the industrial classification as one of the stratification variables. It will be necessary to redesign these surveys so that NACE Rev 2 is the industrial classification by which samples are selected.

There are different types of sampling schemes available. These include cut off designs, the use of panels, simple random sampling, stratified simple random sampling, systematic sampling etc. Probably the most common designs in use in National Statistics Institutes around the world are cut-off designs, stratified designs and combinations of the two.

2.1 Impact of the move on cut-off designs

Cut-off sampling involves sampling all units above a certain threshold, and no units below. The thresholds are set according to a certain variable, such as employment or turnover, and may differ by industry.

Thresholds are chosen to find the balance between the number of units in scope and the coverage of the economic activity (for example the threshold may be set to guarantee that enterprises that contribute greater than or equal to eighty per cent of turnover in an industry are covered).

Methodologists have a strong preference for random sampling, so a change in industrial classification provides an opportunity to change the design from cut-off to one where random sampling takes place. There are two main categories of cut-off design, and we look at each separately below.

case i - a single cut-off threshold which is applied to all industries

If the cut-off thresholds do not differ by industry then there is little work to be done. The (single) threshold can remain the same even under the new coding structure. All industries have the same threshold so there is no change except in re-coding computer systems to accept the new codes. Production of back data is an issue of course but that is not a sampling problem and is dealt with in the backcasting handbook.

We know that the new industrial classification has extra four digit industries, and it is possible that some of these industries were not covered before but will be now. Care will need to be exercised on the overall sampling size allowed and this may lead to a change in the threshold applied to all industries. Register information on counts of businesses in each of the new industries will be required in order to determine whether there is any need to amend the cut-off threshold. Also register information on relevant auxiliary variables (turnover, employment) may be used. The information on counts should be available during 2008 at the latest.

case ii - different thresholds by industry

This is the more complicated case since it is necessary to determine what the thresholds should be for the industries in the new industrial classification. Once again it will be important to examine Register information to carry out some simple analyses of the numbers of businesses in each industry above certain thresholds.

This information, combined with correspondence tables mapping NACE Rev 1.1 to NACE Rev 2, can be used to determine estimated thresholds for the new industries. These initial thresholds may need to be reviewed once sampling on the new NACE begins properly. Many industries will map one-to-one from the old NACE to the new and these will be easy to handle, but care will need to be taken on the other types of correspondence. An iterative approach may be necessary, since the counts of businesses and sums of relevant auxiliary variable in each industry will differ as the quality of that information on the business register improves.

2.2 Impact on panels

The panel design is one where the same businesses are in the sample each period. There is no estimation to find population totals but rather the change in the variable being measured, from one period to the next, is what is sought.

Often panels are used in short-term statistics to measure change. Results derived from the panel are applied to a benchmarked total from a more reliable annual survey and are revised every year or so when the latest annual data become available. Much of the approach for cut-off designs can be applied to panels since the challenges faced are similar.

One thing is certain, a change in industrial classification provides an excellent opportunity to refresh and update the panel. This will be essential to ensure good coverage is obtained across the range of industries in the new classification.

Modelling and benchmarking techniques may need to be used to estimate missing totals for new industries or new size bands.

2.3 Simple random sampling

It is possible for surveys to be random without following a stratified design. Indeed, the industrial classification may not feature in the sample design, but is used when the information received from respondents is post-stratified by industry.

In the UK a good example of this is the annual structural earnings survey. The survey is a 1 in 100 random sample of all jobs registered in the Pay as You Earn Scheme administered by the UK's tax department. Since the sample size is so large, at around a quarter of a million employees, the quality of the results when the information is post-stratified by industry is very good. Surveys like this where the industrial classification does not form part of the design are easy to deal with since there is no work to do when sampling according to the new industrial classification since all the work is in the post-stratification of the businesses which have responded. There is of course an issue in terms of producing back series, if this is required. That issue is not dealt with in this paper.

Simple random sampling where there is selection according to the industrial classification but there is no stratification by size of business is a special case of a stratified design. Discussion of stratified designs is given below.

2.4 Impact of the move on stratified designs

Many Member States use stratified random sampling in their operation of business surveys. In the UK, stratification is usually by a fairly fine level of NACE Rev 1.1 detail and between four and six size bands based on employment values held on the business register. 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 NACE Rev 2 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. If some suitable auxiliary information (eg turnover, employment) is available on the register other alternatives such as x-optimal allocation (Sarndal et al 1992) could also be considered on a similar basis.

The optimal sample size in stratum *h* is:

$$n_h = n \cdot \frac{N_h \cdot S_{xh}}{\sum_{h=1}^{L} N_h \cdot S_{xh}}$$

2.5 Resource issue in re-designing samples

Re-designing samples is resource intensive, and it may be impractical to reallocate all samples adequately in the time allowed between the new NACE becoming available and the need to select samples. In this case, alternative proxies may need to be sought to transition between the old and new classifications

2.6 A practical method for reallocating the sample

One option that may be possible is for the existing sample to be tabulated against the new strata, and the number in each new stratum to start off as the new sample size in that stratum. 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. Furthermore, it will be possible to identify those parts of the new sample which appear weak. This may well be the case for new industries. The sample may be far too small to deliver meaningful results in some of the new industries. Under these circumstances, it would be wise to carry out over-sampling (leading to 'top-up samples') for a period, maybe of one year. Over-sampling is a process whereby the weak parts of the sample are supplemented by extra units, thus ensuring the quality of the estimates in these industries is maintained.

This assessment of how the current NACE Rev 1.1 sample maps across to NACE Rev 2 can be carried out as soon as all of the units on the Register have been assigned a new NACE code. In the UK this will be from the start of 2008, thus allowing work on the sample allocation to begin from this time. This means that by the time sampling on the new classification begins in earnest we will have already carried out analyses which show us where sample sizes should be increased, if only for a short time (maybe one year), in order to maintain quality. In the case of short term statistics, this would be through 2009. If the top-up samples are used then good estimates of the sample variances can be obtained by the end of this first year and this will allow an optimal allocation based on Neyman to be achieved in time for the start of 2010. Of course, agreement that over-sampling can be undertaken must be obtained from the leaders in the NSIs and money set aside for this to happen. Member States may not be able to achieve this agreement, and if this is so then the early periods on the new NACE may prove difficult, with poor results in some industries due to low sample sizes. It may be possible to re-allocate the (new) sample in this instance, by taking from the industries which are reasonably well covered and adding to those which are not. The guidance on this use of top-up samples is summarised below:

Top-up samples:

- Take an iterative approach, in terms of the sample allocation, and analyse with new information when it becomes available.
- Conduct analyses of numbers (mainly) of businesses in old and new codes on the register and in survey samples.
- See if any look particularly small and try to re-allocate some of the sample to these, or increase the sample size - timing is important here.
- If publishing on both old and new codes simultaneously, it may be necessary to boost the total sample size for this period, to ensure an adequate sample for both. Once publication under the old codes ceases, the sample could be cut back to its original size.
- A complete re-allocation should be considered once enough information to estimate variances under the new codes is available.
- Care needs to be taken when weighting the sample to correctly represent the sample design.

2.7 How the timetable might look at the transition to the new NACE

Sometimes it is helpful to look at a practical example. Imagine how the move to the new NACE might be organised for a short-term survey. The timetable below aims to illustrate how this would look.

It is assumed that there would be a minimum period of double coding and updating both by new and old NACE in calendar years 2008 and 2009. If the length of the period of double coding and double updating of register is longer then 2 years, further modifications may be carried out.

Jan 2008	All businesses on the business register are coded according to two classifications, NACE Rev 1.1 and NACE Rev 2.
Through 2008	Continue sampling according to NACE Rev 1.1 but use Register information to tabulate the existing sample against the new strata in NACE Rev 2. Use this scheme as the first attempt at the sample on the new basis. For strata that are weak, in terms of their sample size, estimate the numbers required for acceptable results on the new NACE and seek approval to carry out this top-up of the sample.
Jan 2009	Draw the sample on NACE Rev 2. As shown above, this is initially simply the old simple tabulated against the new industries and new strata but with, hopefully, the strata which are weakest under the new classification boosted by a top-up of the sample.
Through 2009	As information is returned on the new basis from the businesses so it becomes possible to calculate the variance of the elements in the sample in each of the strata. These sample variances are used as approximations of the population variances and enable the Neyman allocation to be carried out.
Jan 2010	Using the information obtained through 2009 it is now possible to re-allocate the sample in a far more efficient way. The top-up sample can cease and the survey can be thought of as being properly conducted according to NACE Rev 2.

3. Weighting (as applied to estimation of variables in sampling theory) under the new NACE

*3.1 **Theory** (for a description of weighting theory, see Annex A)*

In Annex A a description of weighting theory is set out. This is weighting theory in the context of estimation of variables using sample surveys, not in terms of weighting together indices to form higher aggregates, such as occurs in the domain of short term statistics. A summary of calibration estimation as implemented in the Office for National Statistics in the UK is presented in Annex A. The key idea behind calibration estimation is that of finding weights which will modify the usual estimator such that the calibrated estimator for the auxiliary variable is one for which there is no calibration error. For more discussion of the theory please see the annex.

3.2 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 time series according to the new classification in the year following the change; this depends on the size of the difference between the classification systems. The breaks may be significant especially in strata poorly represented in the sample

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.
- Breaks in time series according to the old classification.

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).

3.3 Summary and discussion of alternatives

All three options 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).

It is probably easiest for Member States to follow option 1. It would be good to do the work necessary so that option 3 is an option, but this depends on having the resources necessary to carry out the work.

Annex A

Weighting theory (in the context of estimation of variables using sample surveys)

This section sets out some options relating to weighting during the change to NACE Rev 2.

To prepare the way, we present a summary of calibration estimation, as implemented in the Office for National Statistics, UK.

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 $\mathbf{y}_k = (y_{k1}, \mathsf{K}, y_{kJ})'$ 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 Y_U denotes the *N×J* population matrix of *y* values given by

$\mathbf{Y}_\mathbf{U} = [\mathbf{y}_1, \mathbf{y}_2, \mathsf{L}, \mathbf{y}_N]'$, 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*,

$s = \{k_1, \mathsf{K}, k_n\} \subset \mathbf{U}$,

which is selected for observation in the survey. Hence the purpose of the survey is to estimate T_y on the basis of the available survey data {**y***k* ; k_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 $d_k = 1/\partial_k$ is the design weight for unit *k*, and ∂_k is the sample inclusion probability for unit *k*. In most survey applications, however, some auxiliary variables

$\mathbf{x}_k = (x_{k1}, \mathsf{K}, x_{kp})'$,

that are strongly positively correlated with variable *y*, may be available. Using this information may help improve the accuracy of estimation of the target parameter 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 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 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 $\{ w_k , k_s \}$ 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 $\{ w_k , k_s \}$ 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 $\{ w_k , k_s \}$ may satisfy the calibration constraints given the sample data **X***s*, the design weights $\{ d_k , k_s \}$ and the population totals 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 T_y).

Deville and Särndal (1992) defined a family of calibration estimators for 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 *EP* 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 *qk* > 0, *ks*, 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 $y_j j$ 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

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

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.