Re-design of short term surveys at the UK Office for National Statistics

Myrto Miltiadou
Gareth James

Summary

The UK is in the process of implementing a new revised standard industrial classification (SIC). The short term surveys which collect turnover and employment data will be selecting under the new SIC for the first time in January 2010. This paper describes the work undertaken by the UK Office for National Statistics to re-design these surveys.
The surveys were re-designed as part of a single project that aimed to improve coherence of methods and promote the use of best practice.
The paper describes how the new classification groups and employment thresholds were defined and agreed with the customers. The paper also explains the optimal sample allocation techniques and estimation methods used. A multivariate macro was applied for the optimisation of the sample across all three surveys that ensured an appropriate balance between Manufacturing and Services sectors without placing additional burden on responders.
The paper also explains the different methods used for backcasting survey estimates onto the new classification.

1. Introduction

Economic activity in UK official statistics is classified according to the Standard Industrial Classification or SIC. The SIC is derived from the EU-wide industrial classification, NACE which is also based on the International Standard Industrial Classification of the United Nations. The SIC is a hierarchical classification system with each code containing five digits. The first two digits denote the Section, the third digit the Group, the fourth digit the Class and the fifth digit the sub_class. SIC and NACE are identical at the four digit level whereas the fifth digit level is UK specific.
Every 10 to 15 years the UK classification is reviewed and updated to reflect changes in the economy. New codes and descriptions are introduced to reflect the development of new industries and existing codes are combined or removed as industries become smaller or less significant. The last major revision of the SIC took place in the mid-1990s with some smaller revisions following in the early 2000s which resulted in SIC(2003). A revised version of the SIC(2003), the SIC(2007), is now been implemented in the UK. This is part of a co-ordinated move with other EU member states which are also revising their classification.
The implementation of the new SIC(2007) has been phased over a number of years. This is mainly to meet European legislation requirements but also for reasons of practicality.
The short term surveys used to collect turnover and employment data will be selected under the new SIC for the first time in January 2010. The paper describes the re-design of the Monthly Production Inquiry (MPI), the Monthly Inquiry into Distribution and Services Sector (MIDSS) and the Retail Sales Inquiry (RSI). The three surveys collect similar information but cover different parts of the economy.
The main customers of these surveys are National Accounts who use the outputs to produce the Short Term Indicators: Index of Production, the Index of Services and the Retail Sales Index. They all feed into the 'Blue Book', which is the main annual National Accounts publication.
The change in SIC affects all survey processes and to implement the change all the processes have to be reviewed. There was also the need for improving efficiency of the sample, increasing the coherence of methods between the surveys and promoting use of best practice. The change in SIC provided a good opportunity for a complete re-design of the surveys. Together with introducing new SIC groups, new employment thresholds were defined to form the new sample strata.
Methodology Directorate has worked closely with the business areas, the teams that process these surveys, to ensure that customer requirements are going to be met. This included lengthy consultations to define quality requirements, in terms of coefficients of variation, and to agree the level of detail needed for the output groups. The customers had to prioritise their needs in order to achieve acceptable quality with feasible sample sizes.

Estimates are needed on both old and new SICs. The office will be using both macro methods (conversion matrices) and micro methods (domain estimation) to produce estimates on both SIC basis.

2. UK Sampling Frame

The Inter-Departmental Business Register (IDBR) is maintained by the ONS and it is a list of UK businesses. It is the sampling frame used for most of ONS business surveys. It is based on three administrative sources, Value Added Tax and Pay As You Earn data from Her Majesties Revenue and Customs (HMRC), and Incorporated Business Registered data from Companies House. The Business Register Survey and other ONS surveys are used to maintain the IDBR.
Information held on the IDBR includes the SIC, the employment and turnover of the business. The SIC and the employment are usually used to form the strata in business surveys where the turnover is used by most surveys as the auxiliary variable for estimation purposes.
Since January 2008 each business on the IDBR has carried both SIC(2003) and SIC(2007) codes.
A new coding tool had to be installed that converted textual descriptions supplied by the businesses into SIC codes. For the businesses where descriptions were not available the SIC(2007) code was imputed using probabilities derived from the businesses that had descriptions available. Since January 2008 changes have been made to the SIC(2007) codes assigned. This is mainly due to imputations being replaced by real codes or corrections being made after contacting the business. Both sets of codes will be maintained on the IDBR until the transition to the new SIC is completed.

Table 1 below shows the structures of the old and new SIC and the number of SIC groups at each level. The new SIC includes more detail in the Services Sector, changes in hierarchy, for example Publishing Sector has moved from Manufacturing to Services. New industries were created, like information and communication.

Table 1 Structures of the old and new SIC

	SIC(2003)	SIC(2007)
Sectors (letter)	17	21
Division (2-digit)	62	88
Group (3-digit)	225	272
Class (4-digit)	514	615
Sub-class (5-digit)	699	728

3. Survey Design

The change in SIC presented an ideal opportunity to improve the design of our surveys. The three short term surveys had to be selected under the new SIC for the first time, in January 2010. This enabled us to consider integrating and reviewing all methodological aspects of the surveys together and form the 'Monthly Business Survey'. All questionnaires for the three surveys had to be reviewed and be rolled out as the Monthly Business Survey questionnaires. The aim for this was to prevent the confusion for responders if they move from one sector to another under the SIC(2007) changes. The SIC implementation plan also included improving the sample design, increasing the consistency across the different surveys and standardise the methodology where possible.
The processing and operational systems used by the surveys are similar. Ratio estimation is used, calibrating to known IDBR totals of the auxiliary variable. The differences between the surveys are mainly due to the fact that they developed separately over the years. One main difference was the way the surveys collected employment information. MPI and RSI collected employment information from every business in the sample, whereas MIDSS only collected employment information from a subset of the sample.
Since the surveys are so similar it was decided that for optimisation purposes the sample would be re-designed across all three surveys and the whole process would be implemented as one project.

3.1 Definition of SIC groups

The surveys are stratified by SIC and employment size. The level of SIC detail that is needed for stratification can vary from survey to survey and in many cases, like in the three surveys mentioned, it is not individual SIC codes that form the strata but a combination of codes. Since the three surveys are changing from SIC(2003) to SIC(2007) new groups of SIC codes had to be defined. In order to achieve this it was essential to establish the customer requirements. This was an important part of the process as the number of the groups and the detail of the groups depended on the quality requirements of the customers. The areas that processed these surveys had lengthy and detailed discussions with their main customer, National Accounts, in order to establish the outputs required. At the same time a number of iterations between Methodology and the business output areas were required in order to agree the final groupings. The new strata on SIC(2007) were not fixed at a specific level but varied on a group by group basis depending on the customer requirements. National accounts had specified that they wanted the output levels to be comparable to the outputs levels produced from the structural ONS surveys and specifically to the Annual Business Inquiry. This inquiry, among other things, produced annual turnover estimates for businesses across the whole economy. The level of detail of each group also depended on the number of businesses available in each group. In cases where the population of a group was not sufficient for sampling purposes, groups had to be merged together.
The number of SIC groups used for SIC(2007) was 185 compared to 355 used for SIC(2003).

3.2 Definition of employment thresholds

The arrangements of the employment band thresholds that are used for stratification were decided on a group by group basis. Different options were considered for each group and the best one was applied. Four employment bands were assigned for each group. The cumulative root frequency method was used to decide where the thresholds should be applied so that each group is assigned the best combination possible. For reasons of practicality the number of combinations had to be kept at a reasonable level so that the survey areas can manage them. The final result included seven different band combinations for MBS. In the past one set of size bands was applied for MPI and RSI and three sets of size bands for MIDSS.
Table 2 below shows the employment band combinations assigned for MBS

Table 2 Employment band combinations

Employment band combinations for MBS
0-9, 10-149, 150-249, 250+
0-9, 10-99, 100-249, 250+
0-9, 10-49, 50-249, 250+
0-9, 10-49, 50-149, 150+
0-9, 10-19, 20-99, 100+
0-4, 5-19, 20-149, 150+
0-4, 5-19, 20-99, 100+

The top band for each group is completely enumerated, that is all businesses within this band are selected in the sample. According to the SIC(2003) allocation around 21358 businesses were completely enumerated. Under the SIC(2007) allocation around 16897 are now completely enumerated. The other bands were sampled with rotation rules depending on the size of the business.

3.3 Employment Information

The three surveys collect employment information that feeds into the Work Force Jobs estimates. The MPI and the RSI collect employment information every quarter from all the businesses in the sample where the MIDSS only collects employment information from a subset of businesses. When reviewing the quality requirements for the Work Force Jobs estimates it was clear that the quality requirements

were met just by sampling a subset. This was a very positive step as it meant that fewer businesses will receive the employment questions and therefore the burden on responders will be reduced. This brought more consistency into the way employment information is collected across different parts of the economy.

3.4 Estimation methods

The three surveys use ratio estimation but it had to be decided whether to estimate per each stratum separately (separate ratio estimation) or estimate for all the non-completely enumerated strata together (combined ratio estimation). In the past the choice of the method used was an arbitrary split between MPI and MIDSS. Part of the re-design work was to optimise the methodology and apply the method that best fits each SIC group.
Two standard errors of the estimated totals were calculated for each group, one based on separate and another based on combined estimation. Each group adopted the method that provided the lowest standard error. In cases where there was not a sufficient number of observations in the strata or the difference between separate or combined estimation was not clear, scattered plots were used. If the slopes for each strata were significantly different and sufficiently large then separate estimation was used.
For Manufacturing and Services Sectors the results show that overall under SIC(2007) 59 groups use separate ratio estimation where 97 use combined. In the past under SIC(2003) 106 groups that relate to Services used separate and 222 groups that relate to Manufacturing used combined.

4. Sample allocation

In order to optimise the stratified random sample, a multivariate sample allocation macro was used. The macro produces near optimal integer sample allocations such that desired accuracy requirements for sub-group populations are satisfied.
In the past the Neyman allocation method was used for top level optimisation. After the allocation the quality of the lower level estimates was checked and adjustments were made accordingly. This time with the use of the multivariate macro it was possible to specify the requirements at lower levels. The level used for this allocation was the output SIC group level agreed by the customers. A Neyman allocation method was then applied below this level.
The limitation of this multivariate allocation method is that the overall sample size cannot be specified. Therefore the relative standard errors or coefficients of variations (CVs) for each group have to be adjusted after every run until a feasible sample size is achieved. The areas that process the surveys had to specify the CVs that they wanted to achieve for each SIC group. These were then used to form the constraints for the multivariate allocation. The process was lengthy and complicated as it took several allocations to achieve the appropriate accuracy at the appropriate sample size. The fact that the sample was allocated across the three surveys helped with the optimisation and the balancing between manufacturing and services sectors. Given the limit on the sample size not all the initial CVs could be met, especially in areas with high volatility, so compromises were made in order to come up with an allocation that satisfies all three sets of customers.
As well as determining appropriate CVs the macro also required variances for each strata. To obtain a robust estimate of the variances several months of survey data were used and an average variance was calculated. Then variance was weighted to take in to account the fact that the data are going to be sampled under a different design (SIC(2007)). The estimates from the SIC(2003) had to be calibrated to the SIC(2007) stratums and then recalculated.
Variances had to be estimated for stratums that were not sampled under SIC(2003). A lot of SIC codes that were not in scope under SIC(2003) came in scope under SIC(2007). For cases where data were missing for the entire SIC, variances were imputed based on the data from their respective sectors. For cases where data were missing from certain strata, variance figures were produced using the percentage increase in variances from the populated cells.
The allocation achieved the quality requirements by maintaining around the same sample sizes as before. The sample size achieved for the MBS is around 36500 for the turnover estimates and 25200 for employment estimates. The previous sample sizes were 37000 for turnover and 34500 for employment. The employment sample has reduced since sub-sampling is introduced with MBS.

A lot of information had to be estimated and assumptions had to be made for an allocation to be achieved on SIC(2007) basis. The allocation will need to be reviewed in a years time when data on SIC(2007) becomes available.
Table 3 shows the sample allocation for turnover by Industry group and table 4 shows the sample allocation for employment by Industry group.

Table 3 Sample allocation for turnover

Industry	Sample under SIC(2003)	Sample under SIC(2007)
Production	6664	5996
Services	25463	25512
Retail	5008	5007
Total	37144	36515

Table 4 Sample allocation for employment

Industry	Sample under SIC(2003)	Sample under SIC(2007)
Production	7170	5845
Services	22222	17264
Retail	5008	2126
Total	34400	25235

5. Estimation on both SICs

For a number of periods, estimates of total turnover need to be made available on both SIC(2003) and SIC(2007). Since January 2009, estimates (current period and historical) have been provided to Eurostat, although the surveys have remained stratified by SIC(2003). From January 2010 the surveys will be stratified by SIC(2007), but estimates on SIC(2003) are still required as inputs to National Accounts processes until publication of the UK's 'Blue Book' in mid-2011.
The UK will be using two methods of estimation to achieve estimation on both bases: a micro-method approach and a macro-method approach.

5.1 Micro-method approach

Micro methods involve re-working micro data. In the context of the short term surveys, this means using data collected on SIC(2003) in periods up to December 2009 to produce estimates on SIC(2007); and for later periods data collected on SIC(2007) to produce estimates on SIC(2003). In order to achieve this, the data sets all need to be dual-coded, i.e. each unit listed in the sample file needs to have both SIC(2003) and SIC(2007) codes available, and this has been achieved by matching with extracts from the IDBR, which has been dual-coded since January 2008.

Calibration estimation, in the form of a ratio estimator, is used in the short term surveys. Since there are parts of the economy that are in-scope of the surveys under SIC(2007) but out-of-scope under SIC(2003), and vice versa, the decision was taken to define two sets of calibration groups - one for use when estimating SIC(2003) totals and one for estimating SIC(2007) totals. Therefore, different calibration weights will be applied to units' survey returns depending on the estimate required. Estimated standard errors can also be calculated for the domain estimates.

Such domain estimates will be made for SIC(2003) estimates up to publication of the 2011 ‘Blue Book’, and will also be made for SIC(2007) estimates back to January 2009. Estimates on SIC(2007) before this will be made using macro methods instead, a decision taken as the SIC(2007) codes on the IDBR were observed to be still 'settling' during 2008, and are not available at all prior to this.

5.2 Macro method approach

Macro methods are an alternative to the micro methods; the primary difference being that survey micro data are not used in the estimation process. ONS has chosen to use conversion matrices to produce estimates where use of micro methods is not possible or practical. Conversion matrices

simply apportion already-existing aggregate estimates on one SIC to the categories of the other SIC according to a set (matrix) of proportions, before re-aggregating them to form the required estimates on the other SIC. The proportions are derived from data held on the IDBR for the entire survey universe of businesses, and take into account the size of each business in terms of its turnover or employment.

The same set of proportions can be applied to all periods of historical estimates, allowing back series on SIC(2007) to be constructed, even though dual-coded data do not exist for these periods. Although the quality of the macro-method estimates in recent periods appears quite reasonable (from comparison with domain estimates), it is acknowledged that the quality of estimates further back in time will be lower due to possible inappropriateness of current-period factors to apportion historical estimates.

5.3 Comparison of methods

Conversion matrices are simple to apply and can be used to create estimates in periods for which dual-coded data are not available. Empirical evidence suggests that the quality of estimates is reasonable too, at least for more recent periods, and for variables such as employment and turnover as collected by the short-term surveys. However, there may be biases present in the estimates, and the application of just one set of proportions means that the effect of re-classification of businesses can be missed. This approach means that the dual SIC(2003)-SIC(2007) estimates are not entirely consistent, but it was felt better to achieve the best possible quality separately for each set of estimates.

The SIC(2007) series will be formed by linking together the three series of estimates. The first part, prior to 2009 will be created using conversion matrices and will be linked to estimates for periods throughout 2009 created using domain estimation (replacing conversion estimates currently been calculated). Then, from January 2010 the estimates will be derived directly from the surveys stratified by SIC(2007):

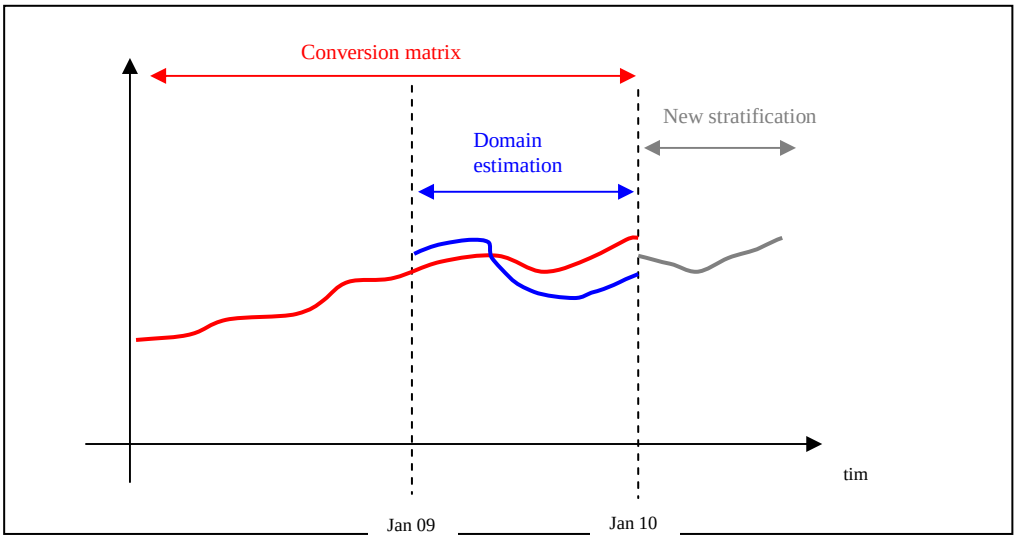

Figure 1: the different sections of a SIC(2007) series for the short term statistics. Note that the domain estimates during 2009 will replace the conversion matrix estimates also available for this period, and that linking of the series, at January 2009 and January 2010 may be required.

Further discussion of the methods can be found in James (2008a), Nolan et al (2008) and James (2008b).

6. Conclusion

The change in SIC(2007) provided an opportunity for improving the design of the short term surveys, MPI, MIDSS and RSI. Joining the surveys together enabled us to introduce consistency and where possible standardised the methodology used.
The new stratification is not fixed at a specific level but varies industry by industry based on the level of detail required by the customers. Decisions on the methodology were made on a case by case basis. This proved to be a complex process as balance needed to be achieved between what is best methodologically and what is practical to implement.
All the analysis was based on data produced by conversion methods therefore it is essential to carry out a review in about a years time when actual data become available.

7. References

James G. (2008) Backcasting for use in short-term statistics: interim report from the UK Office for National Statistics, *http://circa.europa.eu/*

Nolan L., Šova M.G., Brown G., James G., Lewis P. (2008) Backcasting for use in short-term statistics: final report from the UK Office for National Statistics, *http://circa.europa.eu/*

ONS(2007) UK Standard Industrial Classification of Economic Activities 2007 (SIC 2007): Structure and explanatory notes.

Preston, J. (2004). Optimal Sample Allocation In Multivariate Surveys: An Integer Solution. *Unpublished document,* Australian Bureau of Statistics.