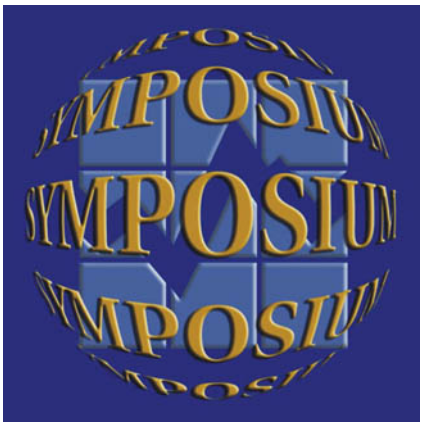
Catalogue no. 11-522-XIE

Statistics Canada International Symposium Series - Proceedings

Symposium 2004: Innovative Methods for Surveying Difficult-to-reach Populations

2004



Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2004
Innovative Methods for Surveying Difficult-to-reach Populations

A REVIEW OF STRATEGIES FOR SURVEYING RARE AND DIFFICULT TO REACH POPULATIONS IN ONS'S ESTABLISHMENT SURVEYS

Paul Smith and John Perry[1]

ABSTRACT

The Office for National Statistics' establishment surveys cover many subject areas, and several aim either to survey populations for which there is no good frame, or to measure characteristics that are rare in the population of businesses. ONS approaches these challenges in different ways. First, satellite registers are built for individual surveys (as in some surveys of the financial sector), which rely on the presence of an administrative source; these may not have complete coverage, but are generally good when the source is from a regulatory body. A second approach is to build up historic information into a panel. This presents challenges in the sampling, as supplementary coverage of the population outside the panel is needed. Both panel updating and "outlier" treatment can have an effect on the estimates with this approach. A third approach is to have a filter question from a more general survey, and to follow this up with a second phase sample. The two-phase sample approach has a better developed theory, but some practical issues remain, for example timing differences between the two phases, and how to deal with atypical observations. This can also be a way of extending or updating the panel. This paper reviews these methods and the challenges they present, and illustrates them with examples from the ONS's establishment surveys.

KEYWORDS: International Trade in Service; Pension Funds; Satellite Register; Two-Phase Sampling

1. INTRODUCTION

1.1 Surveys for rare characteristics

The needs of government for information on the activity of establishments (‘establishments’ and ‘businesses’ will be taken to be synonymous in this paper) are many, and they cover many sectors of economic activity. In some sectors the population is well-known as a result of legal requirements on businesses to register with an administrative scheme. The principal administrative sources in the UK are Value Added Tax, Pay As You Earn income tax and Companies House. The way in which the population which is the standard in ONS establishment surveys is constructed is outlined in section 1.2.

Some activities, however, are exempt from one or more of these administrative schemes, and in these cases the populations are less well-known, and therefore more difficult to survey. In these cases a combination of sources is used to generate a population from which to sample, and these give rise to a number of strategies for ensuring that coverage of the population is as good as possible.

A further challenge is presented by the situation where the population of establishments is well-defined, but the characteristic to be measured in the survey is rare in the population. In these cases the standard survey approach inevitably produces estimates with large variances which are difficult to use as economic indicators. In order to reduce the variability it is necessary to use some further information about the presence of the desired characteristic in the population.

[1]Paul Smith, Office for National Statistics, Cardiff Road, Newport, NP10 8XG, UK (paul.smith@ons.gov.uk); John Perry, Office for National Statistics, Cardiff Road, Newport, NP10 8XG, UK (john.perry@ons.gov.uk)

1.2 Identifying the population of businesses in the UK

The population of businesses in the UK is monitored for statistical purposes from the Inter-Departmental Business Register (IDBR) (Perry 1995). The register is based on three administrative sources:

- Value Added Tax (VAT), for which registration is compulsory for businesses with a turnover greater than a fixed threshold (£58,000 from April 2005). The registration provides both the classification, coded from a one-line description of the business activity, and an estimate, provided by the business, of its annual turnover. Certain activities (for example, pension funds) are exempt from compulsory registration, but may register voluntarily.
- Pay As You Earn. A scheme for collecting income tax/national insurance (including employers contributions). Any employer with employees whose earnings exceed a threshold (currently around £5,000) are required to register for PAYE.
- Companies House. There is a requirement for businesses that wish to incorporate to register at Companies House and appear on a public register and this information is additionally used in register construction. Such incorporated (usually identified as Limited or PLC) businesses file annual accounts with Companies House, although this information is less timely than the other two sources.

In many situations a business will be identifiable from all three sources, but in some sectors where there is an exemption from VAT, there may be only one source, PAYE, and this is particularly an issue in the financial sector. Since some financial businesses require relatively few employees and may have complex structures, it can be challenging to identify the units which need to be surveyed from such an administrative source based only on employment.

1.3 Outline

Section 2 deals with satellite registers and the challenges for maintaining them and keeping them in line with the main register. Section 3 outlines two-phase sampling approaches to surveying rare populations, and section 4 investigates the panel approach to obtaining a register. Section 5 compares and contrasts the different methods and highlights areas for further research.

2. SATELLITE REGISTERS

2.1 Survey integration through register integration

For the last decade the ONS has used the IDBR as the frame underlying its major business surveys, with the advantage that this unified frame allows survey co-ordination, through sample co-ordination and through ensuring that the range of surveys covers the range of activity identified on the register. This avoids the earlier challenges of separate frames for different sectors, which meant that businesses could be double counted or missed across a range of surveys.

The same challenges remain, however, in sectors which are not well-covered by the administrative data. To take a specific example, the measurement of activity and financial flows in the pensions sector is a topic of high current interest in the UK, but many pension funds are managed in such a way that they may or may not be separately identifiable as a business. In addition, the VAT system operates special rules for the treatment of pension funding, resulting in some exemption from registration for VAT purposes (HMCE 2002). The result of this is that it is not possible to identify all pension funds activity from the sources which form the IDBR. There are two main alternative sources of additional information, both specific to this sector of activity. The Occupational Pensions Regulatory Authority (OPRA) has statutory powers requiring any pension scheme to be registered with it. These schemes are not, however, grouped into funds, and there is some deficiency of coverage; for example local authorities are not required to register. The second source is from the National Association of Pension Funds (NAPF), a trade association with wide, but not complete, coverage of the industry in the UK, which has information on the pension funds amongst its membership, but not of which schemes make up these funds.

In an effort to get maximum coverage for minimum cost both within ONS and in compliance cost for those completing questionnaires, ONS is currently using the NAPF register as a source and basing its sampling at the fund level, although this strategy is currently under review. In order to make this work, a satellite register is constructed.

2.2 Satellite registers

A *satellite register* is a register which has an independent source, but which is linked to, or essentially "orbits around", the main register. The basis is simple, in that as many as possible of the cross-linkages between the satellite and the main register are produced, so that the updating of the main register and satellite are as synchronous as possible.

In the case of pension funds, the NAPF source is used to construct a register of pension funds and their addresses, which contains roughly 900 funds. Where possible, these are matched through address and trading style with businesses already existing on the IDBR. This process leaves some pension funds that are not on the IDBR, i.e. are not registered for VAT or PAYE purposes and these are then included within the population for the survey.

Because the IDBR is used as a sample control tool as well as a frame, additional units are created for the unmatched businesses. Both matched and unmatched units are flagged as units to be used specifically for the pension funds survey, and these units form the satellite register, embedded within the normal register database. Sample selection and maintenance can then proceed as normal with the reduced population defined by these "pensions businesses".

There are then two maintenance routes: the usual updating of administrative sources for the whole register, which may signal changes in the parent business structure; and changes in pension funds structure obtained periodically from NAPF. The former gives rise to proving activity (checking the business structure), and the latter provides updated units which are matched to existing register units and are either linked if a match is found, or created as a new/modified unit otherwise. The sources may not have the same properties, however. For example, NAPF has about 5% new creations over a two-year span, whereas the IDBR has around 7.5% creations in pensions and insurance businesses each year.

The NAPF source also provides information on the total assets of pension funds, which is used as an auxiliary variable for both stratification and estimation, and this is maintained directly from the NAPF updates.

2.3 Coverage and estimation issues for satellite registers

As described earlier, a satellite register can be used to supplement the units held on the main business register to provide additional variables for stratification or estimation. Such satellite registers, however, may themselves not be complete.

Following through the pensions example, the NAPF register has less than 100% coverage, and a previous comparison with OPRA suggested that the NAPF missed 1.6m pension scheme members from a known total of 12.2m (13%). This comparison is complicated by the lack of congruence between the units used in the two sources. In order to include some compensation for the undercoverage of NAPF, the estimation for the pension funds inquiry includes an additional inflation weight to account for the part of the population not covered. This factor is updated every two years.

The relatively low coverage of NAPF has led to some re-evaluation of whether using OPRA as a basis for sampling would be an improvement. One strategy would be to move to OPRA as the main basis for collection, but to use the NAPF data as a second frame as a means of identifying where the OPRA frame may be deficient, and to produce an improved estimate of total pensions activity.

3. TWO PHASE SAMPLING AND ESTIMATION

3.1 Filter questions

A filter question in this context is a question, typically on a survey with a large sample size and good coverage (the first phase survey), which identifies that a business has a certain characteristic, for example research and development (R&D) activity. The purpose of this question is to identify such activity for a wide range of businesses relatively cheaply, in particular cheaply for the business involved in terms of compliance cost. There is then a follow-up, the second phase survey, which asks more detailed questions of a subsample of those that have been identified (filtered) by the first process. This could be termed a 'covert filter question' since it is not clear to the business asking the first question that further information may be requested as a consequence, although ONS is careful to ensure and guarantee publicly that all information collected is used only for statistical purposes.

3.2 Two phase

In that this approach fits well with the two-phase structure, the methods for designing and analysing such surveys are well-known. There are two sampling weights, one expanding the second phase responses to the first phase sample total, and one expanding the first phase to the population. There is in fact a range of ways to use first-phase estimates and auxiliary totals, as described by Estevao & Särndal (2002). In the case where there is only a yes/no filter question, then their case (B2) is standard $\sum w_I \sum w_{II} y_{ki}$. However, where the filter question asks more detail, such as total activity in the filtered category, as has been the case with ITIS in the ONS's Annual Business Inquiry (ABI) since 2000, then w_{II} can be used to calibrate y_{kII} to $\sum w_I y_{kI}$ (case (A3)). In practice, ONS has not adopted this approach for ITIS, but instead uses the estimated first stage total $\sum w_I y_{kI}$ for data confrontation to identify anomalies in the data sources.

It is informative to consider the relative costs and benefits of the two-phase approach over the traditional solution of asking the detailed questions in the first survey. The overriding positive is that the process is relatively inexpensive, because detailed questions are only sent to businesses which are known to have some activity to measure. This reduces ONS costs, and also makes a small reduction in compliance costs, since businesses with no activity do not need to consider the detailed questions. Weighed against this are three principal costs. Firstly, there may be relatively few units in the phase I sample that have the required characteristic and non-response will mean that not all are identified. For example, only 1 in 10 of the smallest businesses have ITIS activity, rising to a third of the largest. Even where the characteristic is correctly identified, the activity may not take place in each year. For example, football teams may make an international transfer only once every few years. In fact 6 out of 10 responses from businesses identified as having ITIS from phase I were zero in phase II in 2001. If the sample is insufficient for measurement purposes, we may supplement this by sampling businesses not in phase I, but we will not know which have the activity, and this is therefore likely to both be expensive and produce an estimate with a large variance.

Secondly, there may be a timing mismatch between the phases, which may result in the phase I indicator being for activity one or even two years previously. This naturally contributes to the chances of not identifying all the activity in the survey year. It is possible to send phase II questionnaires after the phase I responses have been processed, for the same year as phase I, but this creates a very long delay between the end of the reference year and the results appearing.

Thirdly, there is the possibility of measurement error in the phase I responses. Many of the topics which are rare in the population are also complicated to define, and even with comprehensive notes and a good questionnaire design businesses may be confused and answer wrongly. Of course, some will not read the notes and then the chances of getting a wrong response are increased. Measurement error studies are expensive, and so no evidence has been gathered directly on this problem, but the estimates can be quite sensitive to this approach, not least because the population estimated at phase I could be poorly estimated.

3.3 A note on commodity-based sampling

Commodity-based surveys can be considered as a version of rare characteristic sampling in which every member of the population has a rare characteristic, but the rare characteristics are different for different units. In this sort of situation we could use a satellite register of businesses producing rare products, and in fact the ONS has some small surveys, run principally for other government departments, which adopt this approach, for example the Building Materials Inquiry.

More usually, however, commodity information is used in a two-phase sample targetting specific characteristics, for example the Producer Price Index, which has a two-phase sample based on the commodity information collected in the PRODCOM survey. The PPI requires a reasonable sample size in each of a long list of commodities, and although we have reasonable auxiliary information from classifications to help us predict this, only the product survey actually identifies the products being made. This is therefore just an example of the same two-phase approach, but with the filter questions being more demanding because they require quite detailed information on all commodities, which is then used for a relatively less onerous second phase (as well as being used in their own right to produce commodity production estimates, of course).

4. BUILDING HISTORIC INFORMATION INTO A PANEL

4.1 Markers for rare characteristics

Where businesses with rare characteristics are to form the basis for the population, there is typically no source at all to identify these businesses from the population of businesses on the IDBR, and in this case some processes are needed for capturing information on the rare characteristic when it is discovered, storing it and maintaining it. For example, ONS produces information on International Trade in Services (ITIS), which is undertaken by a small minority of services businesses. These types of surveys are generically termed “needle in a haystack” surveys within ONS.

An initial exercise has usually been set up to provide information on which the survey is based, often by using a filter question on a survey with wide coverage (see section 4 below). Businesses possessing the rare characteristic are identified by this route and a marker is added to their register entry. These marked businesses then form a take-all stratum within the subsequent survey design. However, it is unlikely that all the relevant businesses will be identified in this way, particularly since there are no censuses of businesses within ONS. Complete coverage is never attained, unless the characteristic is limited to the large businesses that are included in the take-all strata. Even then, non-response will be a factor that limits completeness. In order to make an estimate for other businesses with such activity (and incidentally to identify them), there must be some sampling within the rest of the population. This sampling is, however, typically rather sparse because of resource constraints, and this makes the estimates of activity within this part of the sample design particularly prone to large variability. Nevertheless, having a sample which covers the whole population of businesses has the advantage of correcting for the filter question measurement error problems detected in two-phase sampling (see section 3.2), since every member of the population has a non-zero probability of being included in the sample. Without further work it is unclear whether the mean squared error (mse) is smallest without measurement error but with large variability, or whether we might accept some measurement error for the sake of reducing the mse.

In other cases there may be a source which can be used as a basis for creating the markers, such as the press (Building trade journals are used for a number of surveys covering building materials, for example), Dun & Bradstreet listings, or the Yellow Pages.

4.2 Maintenance of markers

The markers relevant to activity are maintained by identifying businesses which have ceased to have this type of activity, and through addition of businesses newly identified as having such activity. Businesses which cease a special activity of this sort are relatively rare, and typically they are kept within the heavily sampled part of the population for a while (according to an "inertia rule") and only unmarked when it is clear that they will continue without this sort of activity. This is sensible if the cost of re-identifying them if they resume activity is high relative to the cost (in additional questionnaires and higher variance) of including them in the survey when they have no activity. Businesses newly identified with the activity come regularly from the sampling of the rest of the population, and these will move from the sparsely sampled to the heavily sampled part of the frame.

Updating the markers has the potential to create large differences in estimates of activity between years. So a business with a relatively high activity could, in successive surveys, be unsampled (and hence represented mainly by other businesses without any activity), sampled in the sparse part of the sample (in which case their activity will be weighted by a large weight because of the sparse sampling) and then sampled in the heavily sampled area (in which case a similar value will be weighted by a much smaller sampling weight and typically give a much lower total). Indeed this sort of pattern occurs in the ITIS data, and was particularly noticeable when the sample was expanded to cover an area of the business population hitherto assumed to have no activity in this rare characteristic. In the first year much new activity was identified, weighted by a large weight, which when transferred to the heavily sampled area in subsequent years indicated a drop in activity. The drop is due to sampling variability (and is a manifestation of the high variance for this sort of design), although in the case of a start-up year and the following year it would be reasonable to consider that the design changed and that the sampling variability would also change.

4.3 Consistent Estimation of Changes

In order to deal with the estimation of changes when there is so much period to period variability there are several options. The first is an outlier strategy, to reduce the contribution of the few businesses with large activity in the sparsely sampled part of the frame when estimating for non-sampled businesses. This is the classical outlier problem in survey sampling (see for example Hulliger 1995), although the challenge is more acute here because of the polarised nature of the responses. Karlberg (2000) has developed and tested some methods for estimating totals in populations such as these where a large proportion of the responses are zero, but these have not so far been applied to estimating changes in the ONS surveys described here, and it is less clear whether these provide sufficient stability for changes between years to be estimated efficiently.

A further strategy is to smooth over time. The main basis here would be to try to get an estimate of the number of "outliers" (positive responses) in the population by pooling data across a number of time periods. This could then be used to provide post-stratification population totals and hence to improve precision, at the cost of some loss of responsiveness if the proportion of businesses with the rare activity changes from year to year.

4.4 Exclusions

In some parts of the frame, the rare activity may be so sparse that it becomes impracticably expensive to include it in an efficient sampling scheme. In this case we could decide to exclude this part of the frame from sampling, and indeed this happens for the smallest businesses (with <10 employment) in ITIS, because they have such a low incidence. In these cases we either need sufficient information from "similar" businesses (typically those a little larger, and possibly with an auxiliary variable to assist us) to construct a model for estimation, or alternatively we may assume that there is no activity. Both of these approaches accept some (model) bias to exclude the sampling variance and keep the costs manageable.

5. DISCUSSION

Obtaining good coverage in "needle in a haystack" surveys is difficult without an available administrative source, but even where suitable sources are available there is a series of challenges to be overcome. Often there will be a choice of incomplete sources, in which case additional effort will be required to make them consistent. Sometimes there will be no linkage between the sources at all, and they may operate at different levels. In the extreme case, such as the difference between pensions funds and pension schemes described in section 2, it may be most sensible to contemplate network sampling in order to have a way of linking funds and schemes within the survey.

Two-phase surveys have a well-defined theory, but they are relatively slow. There are ways in which their performance can be improved, leading towards a panel type "post-stratified" design, such as those described in section 4, which have a reduced variance in the take-all part of the design, and no measurement error bias although this may lead to large overall variances because of the scarcity of observations.

All of these approaches have certain properties in common which make them stand out from other business surveys. There are typically few auxiliary variables available for use in business surveys, but the situation is often worse for rare characteristics where there is no independent source. Even where some auxiliaries can be obtained, through linkage of the satellite register to the main register, they often have poor predictive power for the main variables of interest, and cannot help to reduce the variability (in some of ONS's financial surveys the difference between the Horvitz-Thompson and ratio estimators is one of standardisation, since there is little difference in the variances).

The best estimate of level is not usually the focus of these surveys however – users wish to use them for economic monitoring and policymaking, and then the movement in the series is often of most interest. However, with all the challenges for estimation of levels, it is difficult to produce a method which is both approximately unbiassed and also has a small variance. In the end having some means to ensure that "most" of the units which have some of the rare activity are covered by the sample is the only way to produce reasonable estimates of movements to fulfil this need, and even then communicating the deficiencies in the estimates to the users if the data remains a challenge.

REFERENCES

Estevao, V. M. and Särndal C.-E. (2002), "The Ten Cases of Auxiliary Information for Calibration in Two-Phase Sampling", *Journal of Official Statistics*, 18, pp. 233-255.

HMCE (2002), "Notice 700/17 – Funded Pension Schemes", www.hmce.gov.uk

Hulliger, B. (1995), "Outlier-robust Horvitz Thompson estimators", *Survey Methodology* 21, pp. 79-87.

Karlberg, F. (2000), "Survey estimation for highly skewed populations in the presence of zeroes", *Journal of Official Statistics*, 16, pp. 229-241.

Perry, J. (1995), "The Inter-Departmental Business Register", *Economic Trends*, 505, November, pp. 27-30.