

# Contents

1	Introduction .....	2
2	Evaluation of variance estimation software .....	4
2.1	Requirements on software for business statistics .....	4
2.1.1	Introduction .....	4
2.1.2	Parameters .....	4
2.1.3	Point estimators .....	5
2.1.4	Variance estimation methods .....	9
2.1.4.1	The Taylor linearisation method .....	9
2.1.4.2	The Jackknife method .....	10
2.1.4.3	The Bootstrap method .....	10
2.1.4.4	The Balanced Repeated Replication (BRR) method .....	10
2.1.5	Summary of requirements .....	10
2.2	Critical comparison of software packages .....	11
2.2.1	Sample designs .....	12
2.2.2	Nonresponse models and outlier treatment .....	13
2.2.3	Parameters .....	14
2.2.4	Estimators .....	15
2.2.5	Variance estimators .....	16
2.2.6	Interfaces, documentation and help .....	18
2.2.6.1	Initial reactions of new users to the software .....	21
2.2.7	Correctness and speed .....	21
2.2.8	Ease of integration with processing systems .....	21
2.2.9	Costs .....	22
2.3	Recommendations for variance estimation software for use in EU member states .....	22
3	Simulation study of alternative variance estimation methods .....	24
3.1	The simulated population .....	24
3.1.1	A model for data generation .....	24
3.1.2	Domains and estimators .....	25
3.1.3	Data features .....	25
3.2	Processing .....	26
3.3	Results .....	26
3.3.1	Comparison of estimators .....	26
3.3.2	Comparison of variance estimators .....	27
3.3.2.1	Naïve variance estimators .....	29
3.3.3	Comparison of software package outputs .....	30
3.4	General conclusions .....	31
4	Variances in STATA/SUDAAN compared with analytical variances .....	33
4.1	Expansion estimator .....	33
4.2	Ratio estimator .....	33
4.3	What does SUDAAN do? .....	34
5	References .....	36

# 1 Introduction

*Paul Smith, Office for National Statistics*

One of the key indicators of quality in sample surveys is the sampling variance arising from the random sampling mechanism through the randomisation distribution. This indicates the variability introduced by choosing a sample instead of enumerating the whole population, assuming that the information collected in the survey is otherwise exactly correct. For a discussion of the theory underlying these calculations, see chapters M2<sup>1</sup> and M3 of the methodology report (volume I). For any given survey, an estimator of this sampling variance can be evaluated and used to indicate the accuracy of the estimates. The forms of these estimators are often complex, especially when the design contains strata or clusters, and when the estimation model uses auxiliary information to improve the accuracy.

In order to make these calculations feasible, appropriate software is required, and although it is possible to construct a program within most survey processing systems to do this for a specific survey, there has been a recent trend towards the production of generalised software which will calculate the appropriate variances in a wide range of commonly met survey situations. These must then be incorporated into the survey process. Sampling variances are often not time-critical information, and any difficulties with data transfer to or setup of this software are offset by the generalised nature of the programs.

In this paper we evaluate five generalised packages which are publicly available: CLAN, GES, SUDAAN, STATA and WesVar PC. There are four main variance estimation methods, Taylor, jackknife, bootstrap and balanced repeated replication (these are explained in section 2.1.4), and between them these packages cover all the available methods except the bootstrap (Table 1.1). These are the packages which were available at the time of putting together the tender for this study, with the exception of PC-CARP which was available but has not been studied. Other packages are being developed; those known to the Model Quality Report team are BASCULA and POULPE but neither of these seems to be fully functional in its current version.

Method	Direct + Taylor series methods	Jackknife	Bootstrap	Balanced repeated replication
Software	CLAN		None	
	GES	GES		
	STATA			
	SUDAAN	SUDAAN		SUDAAN
		WesVarPC		WesVarPC

Table 1.1: Variance estimation methods available in the evaluated software packages.

<sup>1</sup> Reference is made throughout this document to the Methodology report by prefixing section references with an “M”.

The requirements for a variance estimation package are discussed in section 2.1, and there is a comparative description of the packages in section 2.2. Section 2.3 draws conclusions about the suitability of the packages for general use in business surveys in EU member states, and makes recommendations for which should be adopted. A separate simulation study has been undertaken to look at the properties of the available variance estimators, and this is presented in chapter 3 of this report. A more detailed description of the differences in underlying methods between STATA/SUDAAN and the other packages for the Taylor linearisation approach to ratio estimation is given in chapter 4.

## 2 Evaluation of variance estimation software

*Paul Smith, Office for National Statistics  
Sixten Lundström, Statistics Sweden  
Ceri Underwood, Office for National Statistics*

### 2.1 Requirements on software for business statistics

#### 2.1.1 Introduction

The units in business surveys can be of various types, such as enterprises and kind-of-activity units. Mostly a Business Register (BR) is used as the frame for the survey. There is a set of units on the BR, such as enterprises, legal units, local units, and possibly kind-of-activity units. There is a set of variables for each type of unit, some common to other types of unit, some unique. Ordinarily, the BR contains information on which industry each unit belongs to and a measure of the “size” of the unit. The size variable is often the number of employees, or perhaps a measure of turnover (depending on unit level). These variables and their reference dates affect the use of auxiliary information in the sampling design and in the estimation process.

In business surveys two typical kinds of probability sampling design can be identified, namely (i) one-step element and (ii) one-step cluster. Typical examples are (i) surveys with the enterprise as both the sampling unit and observation unit, and (ii) surveys with the enterprise as the sampling unit and all its kind-of-activity units or all its local units as the observation units.

The population is often stratified by industry and size, and from each stratum a simple random sample is drawn. The stratification variable ‘industry’ is used with regard to the domains of estimation that are mostly defined by industry. Size is usually an effective variable for reducing the sampling variability (see chapter M2).

Business surveys are ordinarily carried out continuously, either annually, quarterly or monthly. The samples may be co-ordinated over time, using a panel system or possibly a technique based on permanent random numbers (Ohlsson 1995). Units in business statistics typically change fairly rapidly; they can “die”, they can merge with another unit and they can split into several units. The industrial classification may change, and the size of the unit can vary.

#### 2.1.2 Parameters

Let us look at the various types of finite population parameters that are typical for a business survey. Consider the finite population of  $N$  units  $U = \{u_1, \dots, u_k, \dots, u_N\}$ . Sometimes we are interested in the population total

$$t_y = \sum_U y_k \quad (2.1)$$

where  $y_k$  is the value of the study variable,  $y$ , for the  $k$ th element. Moreover, totals for domains – typically industries – are also common. Let us denote the domain set by  $U_d$ ,

$d = 1, \dots, D$ , and set  $y_{(d)k} = \begin{cases} y_k & \text{if unit } k \in U_d \\ 0 & \text{otherwise} \end{cases}$ . Then the total for domain  $d$  is

$$t_{y_d} = \sum_U y_{(d)k} = \sum_{U_d} y_k \quad (2.2)$$

Ratios of different types are common in business statistics. To define these types let  $z$  be another study variable and let the population total for  $z$  be denoted  $t_z$  and the domain total  $t_{z_d}$ . One type of ratio is

$$R_d = t_{y_d} / t_{z_d} \quad (2.3)$$

A typical example here is production per head with industry as domain. Another type of ratio is

$$R = t_{y_d} / t_y \quad (2.4)$$

showing for example the production of an industry, relative to the whole population.

Another parameter of interest is

$$I_d = t_{y_d} / t'_{z_d} \quad (2.5)$$

where ‘prime’ (‘) indicates ‘relative to another population’. A typical application of (2.5) is the relative change in production (say) by industry from one period to another, that is, the totals in the numerator and the denominator have different reference times, but otherwise relate to the same variable and domain. The sample units (involved in the numerator and denominator) are partly the same, partly different, and units that contribute to the total on both occasions may have changed domain (industry) in between.

Indices of production (say) are examples of complex sets of parameters, typically built up from components like (2.5), and usually also deflated by price indices. Yet (2.5) is already a challenge for the available software. The complexity also depends on the way samples are co-ordinated over time.

### 2.1.3 Point estimators

To estimate the parameters defined in section 2.1.2, a sample  $s$  of size  $n$  is drawn from  $U$  (or actually from the frame). Stratification is commonly used in business surveys, that is, a simple random sample  $s_h$  of size  $n_h$  is drawn from the stratum  $U_h$ ,  $h = 1, \dots, H$ , where

$U = \bigcup_{h=1}^H U_h$ . Let the stratum sizes be  $N_h$ ,  $h = 1, \dots, H$ , and the design weights are  $d_k = N_h / n_h$  for  $k \in s_h$ .

However, nonresponse occurs in the survey process, and the response set  $r$  of size  $m$  is obtained, where  $r \subseteq s$ . There are two main ways of treating this problem, namely weighting and imputation. In weighting, the nonresponse compensation adjustment weight  $v_k$  is

constructed primarily with the aim of reducing the nonresponse bias, but is also used to reduce the additional component of sampling error caused by nonresponse (see chapter M8). When using the weighting approach the estimator consists of the sum of the weighted values for elements in  $r$ , where the weight consists of the product of  $d_k$  and  $v_k$ , where  $v_k$  is the tool for making the inference from  $r$  to  $s$  and  $d_k$  from  $s$  to  $U$ . When imputation is used, values for all  $n$  elements are used in the estimation, but now  $n-m$  of these values are estimates (approximations) of the real values.

None of these methods is expected to completely eliminate the bias. When a substantial nonresponse bias is still present the variance estimate and the confidence interval will be an irrelevant and incomplete measure of the quality of the point estimate. As indicated above, nonresponse will also cause an additional component of sampling error. This is obvious in weighting, since the number of observations is reduced from  $n$  to  $m$ .

In the following, we describe estimators used in business surveys. Here we describe the estimator using a nonresponse compensation adjustment weight, which has a more complex form than the estimator based on imputation.

The nonresponse compensation adjustment weight is an approximation of the inverse of the response probability. That is, one seeks a relevant model of the response probabilities. Commonly, this model consists of a grouping of the sample  $s$ . Särndal, Swensson & Wretman (1992) denote them Response Homogeneity Groups (RHGs). In the following we will choose among three different types of RHGs, namely

- |   |   |       |
|---|---|-------|
| (i) strata and RHGs coincide<br>(ii) RHGs are subgroups of strata<br>(iii) RHGs cut across the strata | } | (2.6) |
|---|---|-------|

The simplest estimator is the Horvitz-Thompson estimator, combined with nonresponse model (i). That means that we find it plausible that each sampled element in the stratum responds with the same probability. In this case the nonresponse compensation weight is

$v_k = \frac{n_h}{m_h}$  and since  $d_k = N_h/n_h$  the resulting weight is  $N_h/m_h$  and the estimator has the form

$$\hat{t}_y = \sum_{h=1}^H N_h \bar{y}_{r_h} \quad (2.7)$$

where  $\bar{y}_{r_h} = \frac{1}{m_h} \sum_{r_h} y_k$ .

A somewhat more complex estimator is obtained when using nonresponse model (ii), namely

$$\hat{t}_y = \sum_{h=1}^H \frac{N_h}{n_h} \sum_{q=1}^{L_h} n_{hq} \bar{y}_{r_{hq}} \quad (2.8)$$

where  $n_{hq}$  is the size of the part of  $s$  that falls into RHG  $hq$ ;  $m_{hq}$  is the size of  $r_{hq}$ , the response set in RHG  $hq$ , and  $\bar{y}_{r_{hp}} = \frac{1}{m_{hp}} \sum_{r_{hp}} y_k$ .

When using nonresponse model (iii) an even more complex estimator is obtained. Let us here express it by the general version

$$\hat{t}_y = \sum_r d_k v_k y_k \quad (2.9)$$

Frames used in the Member States regularly contain more information than industry and number of employees, for example, the turnover from a previous time of reference. Moreover, geographical information for the local units is commonly available. Thus, there may be register information, which is correlated with the study variables and/or the response probabilities, but not used in the estimator of the form (2.9). A simple version of such information is a partition of the population. To demonstrate estimators based on such a partition we let  $U_1, \dots, U_p, \dots, U_p$  be groups that form a mutually exclusive and exhaustive partition of the population. Assume that we know the sizes of these groups,  $N_1, \dots, N_p, \dots, N_p$ . Then they can be used as poststrata. Such an estimator, using the nonresponse model (i) mentioned above, has the form

$$\hat{t}_{yr} = \sum_{p=1}^P \frac{N_p}{\hat{N}_p} \sum_{h=1}^H \frac{N_h}{m_h} \sum_{r_{hp}} y_k \quad (2.10)$$

with  $\hat{N}_p = \sum_{h=1}^H \hat{N}_{hp}$ , where  $\hat{N}_{hp} = \frac{N_h}{m_h} m_{hp}$ ;  $m_{hp}$  is the size of  $r_{hp}$ , the response set that belongs to the union of  $U_h$  and  $U_p$ .

Estimator (2.10) is a special case of the following general estimator

$$\hat{t}_{yr} = \sum_r d_k v_k g_k y_k \quad (2.11)$$

where

$$g_k = 1 + \left( \sum_U \mathbf{x}_k - \sum_r d_k v_k \mathbf{x}_k \right)^T \left( \sum_r d_k v_k \mathbf{x}_k \mathbf{x}_k^T / \sigma_k^2 \right)^{-1} \mathbf{x}_k / \sigma_k^2 \quad (2.12)$$

By choosing the positive factors  $\sigma_k^2$  the approach can be made very flexible. This will become apparent in subsequent sections. The vector  $\mathbf{x}_k$  is called the *auxiliary vector* in what follows. Estimator (2.11) is based on a general approach to regression for two-phase sampling following Särndal & Swensson (1987). It is here used in the nonresponse situation, but since we do not know the response probabilities the second-phase inclusion probabilities have to be estimated in some way (see also M2.3.1.5). The inverse of this estimate is denoted by  $v_k$ . In what follows the estimator (2.11) is called the *GREG estimator*.

In the case of poststratification the auxiliary vector is defined by  $\mathbf{x}_k = (\gamma_{1k}, \dots, \gamma_{pk}, \dots, \gamma_{Pk})^T$

where, for  $p = 1, \dots, P$ ,  $\gamma_{pk} = \begin{cases} 1 & \text{if unit } k \in U_p \\ 0 & \text{otherwise} \end{cases}$  and  $\sigma_k^2 = 1$  for all  $k$ . This poststratification

approach gives us one simple method of dealing with outlying observations in a survey, since they can be moved into an appropriate poststratum for estimation.

Most of the classical estimators can be derived as special cases from the GREG estimator. For example, if  $\mathbf{x}_k = x_k$  for all  $k$  and  $\sigma_k^2 \propto x_k$ , where  $x_k$  is a continuous variable, and when nonresponse model (i) is used, then the following estimator is obtained:

$$\hat{t}_{yr} = \frac{\sum_{h=1}^H N_h \bar{y}_{r_h}}{\sum_{h=1}^H N_h \bar{x}_{r_h}} \sum_U x_k \quad (2.13)$$

Estimator (2.13) is sometimes called the *combined ratio estimator*.

Sometimes the group totals  $\sum_{U_p} \mathbf{x}_k$  are known and, in this general case, the  $p$ -groups are called *model groups*. Let us present a simple example. As before assume that  $x_k$  is a continuous variable, but here we know the quantities  $\sum_{U_p} x_k$ ;  $p = 1, \dots, P$ . Let  $\mathbf{x}_k = (\gamma_{1k} x_k, \dots, \gamma_{pk} x_k, \dots, \gamma_{Pk} x_k)^T$ ,  $\sigma_k^2 \propto x_k$  for each  $p$ -group, and the RHGs coincide with strata (nonresponse model (i)) then the GREG estimator takes the form

$$\hat{t}_{yr} = \sum_{p=1}^P \frac{\sum_{h=1}^H \hat{N}_{hp} \bar{y}_{r_{hp}}}{\sum_{h=1}^H \hat{N}_{hp} \bar{x}_{r_{hp}}} \sum_{U_p} x_k \quad (2.14)$$

If strata and model groups coincide then estimator (2.14) can be written

$$\hat{t}_{yr} = \sum_{h=1}^H \frac{\bar{y}_{r_h}}{\bar{x}_{r_h}} \sum_{U_h} x_k \quad (2.15)$$

Estimator (2.15) is sometimes called the *separate ratio estimator*.

When  $\mathbf{x}_k = (1, x_k)$  for all  $k$  and  $\sigma_k^2 = \text{constant}$ , then the classical regression estimator is obtained.

Many business surveys are subject to occasional unusual observations, or outliers, which can have a large effect on the estimates. In these cases, robust versions of point estimators are often used, with the simplest being the poststratification estimator with the outliers in their own (completely enumerated) poststratum. This follows from the method above (2.13). Other methods involve adjusting the weights or the responding values, and winsorisation is becoming widely used within the UK for treating outliers. This leads to a different estimator, which does not necessarily fit completely into the GREG framework.

The parameters (2.1)-(2.4) are totals or functions of totals from the same period of reference. Estimators for these parameters can be obtained by replacing these totals by their estimators. Parameter (2.5) is much more complex since it contains totals from two periods of reference. In most surveys two consecutive samples are drawn in such a way that they overlap each other. That makes it possible to construct combined estimators that are more effective than just replacing the totals by their estimators. However, variance estimation becomes complicated. We do not go deeper into this problem but just refer to Nordberg (1998), who has found a solution to the special sampling procedure used at Statistics Sweden.

So far we have only discussed one-step element sampling designs, but it is easy to see how the one-step cluster alternative affects the formulas. Auxiliary information can be known at the cluster level or at the unit level. In the latter case we can choose to use the auxiliary information either at cluster level or at unit level. When the auxiliary information is known only at the cluster level the model groups are, of course, defined for that level.

#### **2.1.4 Variance estimation methods**

There are four principal ways of calculating variances (Wolter 1985), each unbiased or asymptotically unbiased in most widely-used design-estimation strategies if full response is assumed, but each (in general) producing a different value for the unbiased estimate:

- direct calculation and Taylor linearisation;
- jackknife;
- bootstrap;
- balanced repeated replication method.

Before we discuss these methods just a few words about variance estimation when imputation is used, following the discussion in section 2.1.3. The literature describes many imputation methods such as nearest neighbour donor, current ratio, current mean, auxiliary trend, etc. However, the theoretical development of variance estimators when data contain imputations is still in its initial phase. Two examples of articles on this problem are Särndal (1992) and Deville & Särndal (1994). In surveys where the ‘complete data set’ is treated as if it were the full-response set, however, this will commonly underestimate the variance (see, for example, Rubin 1986).

##### *2.1.4.1 The Taylor linearisation method*

Direct calculation involves application of (normally) the Sen-Yates-Grundy estimator (Sen 1953, Yates & Grundy 1953) to form the variances of simple survey estimates. More complex survey estimates are first linearised by taking the first-order terms in an appropriate Taylor-series expansion, and then the SYG estimates are inserted into the linearised formula.

This is basically a set of appropriate linear expressions for the variances of estimators, which has to be coded into the software. Every different design-estimand<sup>2</sup> combination requires a different formula which must be (essentially) hard-coded; separate formulae are not required for different estimation models if the GREG estimator (see equation (2.11)) is present, as all the commonly used models are either GREG or special cases of it.

#### 2.1.4.2 *The Jackknife method*

The jackknife involves dropping an observation and recalculating the estimates from the remaining observations, repeating successively until all observations have been dropped, and then finding the variance of the resulting series of estimates (with a suitable multiplier to give approximate unbiasedness). The drop-one jackknife is usually used, as it can be shown to give the variance estimate with the smallest sampling variability, although it is possible to drop pairs of observations (or even more) too; this strategy is usually adopted to speed up processing since drop-one is the most processor-intensive method. We consider only drop-one methods here. More information on the jackknife estimator is in M2.4.2.2-M2.4.2.3.

It should be noted that the jackknife is only strictly applicable in with-replacement designs. It can be used in without-replacement designs where the sampling fractions are “sufficiently small” (Wolter 1985, p168), but in many business survey designs, the sampling fractions are relatively large. The dangers of this approach are illustrated in the simulation in chapter 3 below.

#### 2.1.4.3 *The Bootstrap method*

The bootstrap involves resampling a number of times with replacement from the sampled observations, and calculating an estimate for each of the bootstrap samples. The variance of these “bootstrap” estimates is then calculated, again with a suitable multiplier to ensure unbiasedness. The method is described in more detail in M2.4.2.4.

#### 2.1.4.4 *The Balanced Repeated Replication (BRR) method*

This is derived from the balanced half samples (BHS) method which has a very specific application in cluster designs where each cluster has exactly two final stage units. By successively deleting one of these units and changing the weight of the other to compensate, a range of estimates can be produced whose variance can be calculated and suitably adjusted to give an appropriate variance estimator (Wolter 1985). Various adaptations of this can be applied in designs where the clusters have variable numbers of units, based on dividing these into two groups. Recent research (Rao & Shao 1996) shows that only by using repeated divisions (“repeatedly grouped balanced half samples” (RGBHS)) can an asymptotically correct estimator be obtained. This method, then, can only be used for the usual stratified designs in business surveys if we are prepared to treat a stratum as if it were a cluster, and to run the package a number of times with different divisions of the elements into two groups; where there is an odd number of elements in the stratum the results are biased, and ways of reducing this bias (but not eliminating it) are described in Slootbeek (1998). There are ways in which this can be done, but the results are typically unsatisfactory and the manipulation of both data and software becomes very involved.

### 2.1.5 **Summary of requirements**

There is a number of requirements for point and variance estimation in business surveys which any software should satisfy. We have pointed out several such requirements in the

---

<sup>2</sup> thing to be estimated

previous sections. However, in order to simplify the evaluation we will here present a structured summary of these requirements. The demands on the software will certainly vary between Member States (MS). Consequently, packages which only meet some of the requirements mentioned ahead may be sufficient for a particular MS, provided that they meet the requirements of this MS.

The packages will be evaluated with respect to their ability to cope with the following situations.

*Sampling designs:* One-step stratified sampling of units or clusters. In each stratum a simple random sample is drawn. In some strata the finite population correction (fpc) has a large effect; in take-all strata it reduces the sampling variance to zero. Panels or random number techniques are used in the sampling procedure.

*Nonresponse models and outlier adjustment:* Weighting within RHGs (i)-(iii), as described in 2.1.3 and equation (2.6) or imputation as described in sections 2.1.3 and 2.1.4, and outlier treatment using poststratification or winsorisation as described in 2.1.3.

*Parameters:* Parameters for measuring levels as in (2.1)-(2.4) and parameters for measuring change as in (2.5). More complex parameters such as indices are also of great interest.

*Estimators:* Estimators for totals as defined in (2.7) to (2.15). Ratios and other functions of these estimators are also of interest. Point estimates and the corresponding variance estimates for parameters such as (2.5), for example measures of change between two consecutive periods (a demanding task for the packages) are of interest.

*Variance estimators:* availability of different variance estimation methods (Taylor, Jackknife, BRR, Bootstrap).

The packages will also be evaluated with respect to:

- interface, documentation and help functions;
- whether computations are correctly done;
- execution time;
- simplicity to integrate into production systems;
- cost for purchase or licenses.

## **2.2 Critical comparison of software packages**

The software packages evaluated here fall into two distinct groups based on the way they are designed and the type of situations in which they can be used. It makes sense to structure the discussion around these two groups, as the methods employed within the packages are very similar within groups, and quite different between them.

*Group I:* CLAN and GES are designed for stratified designs with estimation models up to the complexity of the generalised regression (GREG) estimator. They are characterised by having two parts to their processing, one in which the appropriate weights are calculated for the survey observations, and then a second phase where the estimates and their associated variances are produced. The variances specifically take account of these weights, and are based on the variances of the residuals from the GREG model (or a specific (simpler) case).

*Group II:* STATA, SUDAAN and WesVar are designed principally for cluster designs with versions of the Horvitz-Thompson (HT) estimator (in most cases optionally involving poststratification); the key here is that GREG-type estimators (including most of the simpler cases such as ratio and regression estimation) are not supported. STATA and SUDAAN both work in a straightforward way with stratified designs, but WesVar needs clusters at the penultimate sampling stage in order to work effectively (mainly because of the BRR variance estimation method employed). This group is characterised by not having a weight calculation phase and requiring the (HT) weight to be input. In some cases the software can be made to produce valid or approximately valid results for estimators other than HT, but this is typically not easy and may require the package to be run more than once for each survey.

### **2.2.1 Sample designs**

CLAN and GES have the following designs built-in:

1. simple random sampling;
2. stratified designs;
3. probability proportional to size (with replacement) designs;
4. one stage cluster designs (optionally with the clusters in strata).

These cover the main probability designs used for business surveys in Member States, but do not extend to the more complex designs used in some social surveys. It is possible to force more complex designs through CLAN and GES by accepting some assumptions about variances at lower stages; one option is to set appropriate jackknife adjustment weights within GES for two-stage designs. All of these methods, however, are vanishingly rare in business surveys, and require considerable expertise and input from the user, so they are not considered further here. Statistics Canada have just begun to develop two-stage cluster sampling for inclusion in the next version of GES (version 5.0).

STATA and SUDAAN have the following designs built in:

1. simple random sampling;
2. stratified designs;
3. one stage cluster designs;
4. two- and multi-stage cluster designs.

These cover a wider range of designs, but the complex cluster designs are not typically used for business surveys, and we know of no examples of their current use in business surveys in member states. However, this does give some added flexibility in the use of the package for various surveys.

WesVar has the following two designs available:

1. simple random sampling;
2. two-stage cluster designs with exactly two primary sampling units in each cluster.

These designs are very restrictive in the context of business surveys where clusters are rarely used, and where treating a stratum as if it were a cluster typically gives more than two primary sampling units in each cluster. For this reason we will not concentrate much discussion on WesVar.

The finite population correction (fpc) can have a large effect on the variance estimates; within GES and CLAN it is included automatically (except for the jackknife estimator in GES). In STATA a specific command option must be used to get the fpc, and in SUDAAN it depends on the design whether the fpc is included or not. GES and SUDAAN alike include the fpc automatically in without-replacement designs, and exclude it in with-replacement designs. However, it can in some circumstances be reasonable to use with-replacement variance estimators as approximate variance estimators in without-replacement designs, when inclusion of the fpc can become important; inclusion of the fpc is unlikely, however, to solve all the difficulties of this approach .

### **2.2.2 Nonresponse models and outlier treatment**

CLAN is the only software package to include the specification of non-response models. This is done by defining response homogeneity groups, which can be defined differently from the stratification and model groups, and provide a flexible way of defining the weighting adjustment for non-response in line with equations (2.7)-(2.10). This additional option within CLAN is similar to the sort of methodology which would arise in a two-stage stratified design, with first stage selection being sampling from the frame and the second phase being “sampling” respondents from the selected sample. This means that the extra functionality can be used to make CLAN give appropriate answers in some complex designs if there is (or can be assumed to be) no non-response.

For the other software packages considered here, only two alternatives are available, either to assume that non-responding units were not sampled, which is equivalent to imputing their value with the mean under the estimation model for the stratum in which they were selected, or to fill in the missing values using some imputation procedure and then use the completed dataset. In both these cases (but particularly the second), it is very likely that the calculated variance underestimates the true variability. The only reasonable method of calculating variances with packages other than CLAN would be to use a stochastic imputation procedure to create multiple datasets (multiple imputation, Rubin 1987) and use the packages to produce a series of estimates which can then be suitably combined. This approach involves a lot of additional processing not available within the packages, and has not been attempted here.

Outlier treatment by moving outliers into a poststratum can be appropriately set up in most of the software described here (in GES and CLAN by setting up appropriate model groups, and in SUDAAN by using the poststratification options). Exact variance calculations for other methods, specifically winsorisation (Kokic & Smith 1999a, b), are not available in any package, but a good (first-order) approximation can be obtained by using the winsorised values as if they were the survey values.

### 2.2.3 Parameters

The parameters which can be estimated in GES are:

- (a) count (an estimate of domain size);
- (b) total (equations (2.1) and (2.2));
- (c) mean;
- (d) ratio (equations (2.3) and (2.4)).

Within CLAN, the user needs to construct several macros to specify the estimation to be undertaken, and at this stage it is possible to include arbitrary rational functions of totals, so that purpose-built estimands can be constructed and their sampling variances calculated explicitly within the package. GES allows only the four estimands described above, but in a similar way the variances of linear combinations can be found afterwards outside the package. In general however, this will require more expertise and effort than setting up the appropriate macros in CLAN. The PC-CARP documentation suggests that it estimates quantiles (with the appropriate variances) too, a facility not available in either GES or CLAN.

STATA and SUDAAN have:

- (a) count;
- (b) mean;
- (c) total;
- (d) ratio;
- (e) regression parameters;
- (f) Wald statistics;
- (g) logistic regression parameters;
- (h) quantiles;

and for STATA only

- (i) arbitrary linear combinations of parameters.

Some of these are not currently widely used in business surveys, but there seems to be some development in the field of estimating distributions, which will make the estimation of quantiles more important, and the facility to produce estimates and variance estimates for arbitrary linear combinations of parameters can be used to assist in the estimation of variances of “complex” population parameters such as changes, index numbers and so on (see chapter M3).

WesVar produces a similar range of statistics to STATA and SUDAAN, including arbitrary linear and non-linear combinations of statistics. The sampling variances of the non-linear statistics can be found because WesVar relies on replication methods.

Of particular interest in repeating business surveys are estimates of movement or change. Where the units are exactly common between two periods (almost never true even if the design is set up in this way because of differential non-response), then any of the packages here can be used to estimate the movement by including the responses for different periods as two survey variables. When the units are not the same, then it becomes very challenging to produce an appropriate estimate of change and its variance. Within CLAN this can be achieved by including the union of the two samples as the sample, and specifying the

response homogeneity groups in such a way that weighting adjustments are made for the units which were not sampled because of the sample rotation as well as those units which did not respond. Because the variance estimation reflects the additional uncertainty due to imputation, it gives an approximately correct variance for the estimate of change taking account of the substitution of units (if the non-response weighting completely adjusts for bias).

A similar imputation can be done to fill in the missing data for rotated (and non-responding) units before entry into the other packages, but because the packages do not appropriately account for imputation when estimating the sampling variance, it will typically be underestimated.

More complex statistics are also of interest, for example deflated index numbers. None of the software is currently able to tackle such combinations of information, and the only reasonable approaches are (i) linearisation of the target statistic and calculation of the appropriate components of the linear combination in CLAN or STATA or from results produced by any of the software packages, or (ii) a sensitivity-type analysis showing the effect of sampling errors on the overall statistic (see M3.4 and Kokic (1998)).

#### **2.2.4 Estimators**

A range of estimators is available for use in business surveys, depending on the range of auxiliary information available from the business register. The simplest estimation method is Horvitz-Thompson (HT) estimation (also called simple raising, expansion estimation and number raised estimation), which involves weighting each unit by the inverse of its selection probability. This estimator is available in CLAN, GES, STATA and SUDAAN, but is not given in WesVar which is designed purely for variance estimation and does not provide point estimates. This estimator is unusual in ONS business surveys, although there are some examples of its use in recent years; in other member states, for example at Statistics Sweden, it is widely used. The only information which is normally required is the number of units (although HT for  $\pi ps$  sampling has already used additional information in setting up the selection probabilities).

Where additional auxiliary information is available from the business register, more complex estimators are often used. In the ONS the ratio estimator (separate or combined, equation (2.13) and the simplification of it with a single stratum) is almost ubiquitous. The true ratio estimator is available only in CLAN and GES, where it is handled appropriately with the correct model used to calculate residuals to feed into the sampling variance calculation. In SUDAAN and STATA only the HT estimator is available. However, it is possible to obtain approximately correct variances for (one-variable) ratio estimation by (i) calculating the ratio of the survey variable to the auxiliary value, within strata (for separate ratio estimation), taking account of the selection probabilities; (ii) constructing an additional variable as the residual between the observed value and the ratio applied to the auxiliary value, and (iii) calculating the variance of this residual within strata again taking account of the selection weights. This involves two passes through the software with some additional manipulation and produces *only* the variance directly – there is no point estimate, and if this is required it

needs some additional processing after the ratios have been calculated to produce it. Cross-stratum ratio estimation can naturally be done in the same way by defining appropriate groups within which to calculate the ratio. The additional feature of choosing the variance function is not available; for estimation of ratios in SUDAAN only the “ratio of averages” ( $\hat{r} = \sum wy / \sum wx$ , with appropriate weights  $w$ ) method is supplied (that is, other ratios such as the “average ratio”  $\tilde{r} = \frac{1}{\sum w} \sum w \frac{y}{x}$  are not available). It is naturally also possible to supply different weights to the expansion estimator, such as those taken from ratio, regression and GREG estimators, but naïve application of these weights in the standard HT estimator does not give the correct variances (more detail is given in Chapter 4). Nevertheless the effects of using this scheme are investigated in the simulation in chapter 3.

Further complexity in the estimator can be introduced by using more variables within a regression estimation framework, although there are very few current examples of this sort of estimation in business surveys in the UK and Sweden (only the Annual Employment Survey uses this method in the UK). However, it seems likely that these methods will become more important in the future. As before, CLAN and GES cover these methods directly, whereas SUDAAN and STATA do not include the direct estimator, but can be used to estimate the regression parameters and hence calculate residuals to use in calculating the sampling variance. We have not attempted to verify that this works using classical regression estimation (that is, with the variance approximately constant with size). Getting an appropriate (non-constant) variance function in regression may be extremely involved (especially where there is more than one explanatory variable), but this is properly dealt with under full calibration in the next paragraph.

The most general estimator, the GREG estimator, which allows calibration to many auxiliary totals and provides a facility to add constraints to bound the weights, is available in only CLAN and GES, and cannot be incorporated into STATA or SUDAAN. We know of no business surveys in member states which rely on this technology at the moment. One side effect of the inclusion of the GREG estimator is that the variance function for the ratio and regression estimators can be defined by the user, by supplying suitable values to the software (normally  $\sigma_k^2 \propto x_k^\alpha$  where  $x$  is one of the auxiliary variables and  $\alpha = 1$ , see (2.12), or sometimes for some other value of  $\alpha$ ). By making the variance proportional to an extremely large number for any particular observation, its effect can be removed from estimation (that is, its  $g$ -weight will be  $\approx 1$ ), giving a rudimentary outlier treatment/robust estimation methodology.

### 2.2.5 Variance estimators

The use of BRR with business surveys is typically difficult, as described in section 0. WesVar relies almost entirely on the method of BRR, and so is not a serious contender for recommendation for business surveys. SUDAAN also has this method available as one option among several, but there seems to be little to commend it over the other methods in the current context.

The four main packages investigated (CLAN, GES, STATA, SUDAAN) all include the direct (“Taylor”) method of variance estimation (the SYG estimator). The implementation is basically a set of appropriate expressions for the variance of estimators, which has to be coded into the software. This is the way in which most business survey variances are calculated, and as such each of the four software packages fulfils our requirement for a basic design-based variance estimator. For the simpler cases of expansion and ratio estimation with model groups corresponding with strata where the full complexity is not needed, there can be little to be gained from the software; in these cases, purpose-written programmes may be perfectly adequate.

Most packages include the finite population correction automatically within their variance calculation for without-replacement designs, but STATA requires it to be specified explicitly as a command option if it is required.

Jackknife variance estimators are available in GES, SUDAAN and WesVar. It should be noted that the jackknife is only strictly applicable in with-replacement designs, and the documentation for the packages points this out. It can be used in without-replacement designs where the sampling fraction is “sufficiently small”, but in many business survey designs, the sampling fractions are high. A further adjustment can be made by including the *fpc*, but none of the packages do this automatically. In GES it is not obvious from the documentation that this is missing. The validity of the outputs is discussed as part of the results of the simulation exercise (chapter 3).

In GES the jackknife option requires the user to set up jackknife groups explicitly. The drop-one jackknife is the most efficient variance estimator, and the easiest and quickest set-up is to use this method, by making every element a jackknife group, and giving each group an equal jackknife adjustment weight. Although this is fairly intuitive, it is a shame that the software does not contain a default to allow it to happen automatically. If speed of processing is vital it is possible to set up jackknife groups containing several elements (faster, less efficient and less intuitive), in which case there are also several ways to form appropriate jackknife adjustment weights – usually the weight is equal to the number of elements in a group, but for multi-stage designs the weights can be set to the number of secondary sampling units to give a variance estimate under the complex design. This flexibility is useful in concept but unlikely to be applied in practice in business surveys.

SUDAAN provides a default jackknife method by simply choosing the keyword for jackknife variances; this is in fact the drop-one method. There is no facility for user-defined jackknife groups.

In WesVar two forms of the jackknife estimator are provided – one is dependent on the specific design with two elements in each final stage cluster, and the other is the drop-one jackknife, which is available only for simple random sampling. By processing strata separately and using the drop-one jackknife it is possible to force the software to deal with some business surveys, but it is not in general suited to them.

None of the software packages considered implements a bootstrap variance estimator.

### 2.2.6 Interfaces, documentation and help

In many NSIs it seems that SAS is becoming the main tool for survey analysis, and this is reflected in the software seen here. CLAN and GES are both written as a series of SAS macros, so that the SAS package is required to use them. CLAN uses only CORE and BASE SAS, whereas GES uses CORE, BASE, AF, FSP and IML. SUDAAN is available in two versions, one free-standing and one which can be called directly from SAS. WesVarPC is designed to look somewhat like SAS but otherwise has no connection with it. Following an agreement between the authors, SPSS versions will now include the WesVar software. STATA is stand-alone (but provides a complete statistical package), and only available for Windows 95, Windows NT or later operating systems.

There are two basic approaches to setting up the data and commands for the software, and these are not related to the groupings described at the head of section 2.2. The first is to provide appropriate commands and leave the user to construct a programme or script which is then submitted to the software, which returns with the completed calculations, and this is the basis for CLAN and SUDAAN. CLAN in fact goes a stage further and requires the user to construct several macros as well as putting together the code to produce the final outputs. CLAN is basically a series of macros, which accept data and other macros as input. Once the user-defined parts are written, the user calls the macros in the appropriate order and combination in order to get the results. Because the program is written in SAS, the entire interface is supplied by SAS. This method makes it relatively easy for the software to be flexible and to cope with cases where unusual estimates are required; it also, by dint of requiring the user to know a fair amount about the way in which the package is constructed in order to use it, prevents the mindless application of default methods in situations where they are not appropriate. By the same token, however, a reasonable amount of expertise in estimation theory and in SAS programming are required to use the package. Fortunately the recently produced CLAN manual (Andersson & Nordberg 1998) is very clearly written and shows in a very straightforward way how to set up the appropriate macros and data. There is no on-line help system available with CLAN. Output is sent only to a SAS dataset, which can then be printed, exported or further manipulated using SAS. There is no formal support system for CLAN, but informal support from Statistics Sweden is available on a case by case basis.

SUDAAN can be viewed in a similar way, except that the macros are called procedures, and in the SAS-callable version they behave like SAS procedures. In stand alone SUDAAN there are Program Editor and Output windows (the output here doubles as both Log window and Output window according to SAS's view of the world). All that is required is for the user to learn the appropriate syntax and to type in these commands. The submit button is then clicked, and the package processes the data as required, sending results to the output window (and/or an appropriate file). Most of the syntax is easily learnt, but there are a few oddities:

- two procedures have different names in the SAS version to avoid reserved keywords;
- the formatting statements in SUDAAN are notoriously long-winded and do not have short forms.

There is a two-volume manual which describes the syntax and the basic usage of SUDAAN, which is a very useful guide for the beginning user. However, it does not contain any explanation of the theory used in the software, and in several places there are bald statements from which it is almost impossible to work out exactly what the software is doing (for example, “a poststratified estimator” is mentioned for several procedures). There is in fact a methodology guide (Shah *et al.* 1995), but this wasn’t sent out as part of the documentation to accompany the license. With this guide to hand the package is well-documented. The on-line help system covers only the main “user-guide” part of the manual. SUDAAN has the advantage of reading and writing files in several formats, including SAS files (in the stand-alone and SAS-callable versions), text, and SPSS (in the stand-alone version only). The SAS-callable version is particularly useful when combining SUDAAN processing with other operations, for example in producing a ratio estimator or doing experiments or simulations where the procedure can be embedded in a macro or loop. There is support for SUDAAN, and an email support address, during office hours on the West Coast of the USA (approximately 1600 to 2400 GMT).

The second approach is one which provides an interface to lead the user through the stages of setting up the appropriate files, meanwhile writing the commands either in the foreground or behind the scenes. GES has an interface which leads the user through the stages. From v4.0 (the latest version) most of the information for a single run of the package is contained on one screenful; the catch is that a 17” (43cm) screen is required to be able to view all the appropriate buttons, and this does not seem to be mentioned in the documentation(!). At any stage the input files must first be defined to GES, so that they must be selected even if they already exist as SAS datasets, and otherwise imported to SAS; the import facility is built in to GES so that there is no need to exit and return. At the same time as a file is defined, the variables corresponding to certain key definitions (strata, etc) are chosen. All the identifiers are intended to be *text* variables, and although numerics can be used in their place in some (but not all) parts of the software, they can’t be chosen from lists of available variables unless they are text. This is frequently frustrating where, for example, the stratum is identified by a number in a numeric field, which must be converted to a string containing the number. Once GES is running it is also not possible to run any code from the program editor without exiting GES (the only way to amend a dataset without exiting is to use ASSIST). GES does contain facilities to generate input files in most of the cases which one would use in practice, normally using a SAS *By* statement. GES maintains its previous settings and data files from run to run, which can be convenient when several similar surveys are to be analysed, or several alternative models are compared for the same dataset. It also has a good system of survey organisation; each survey is individually labelled, and within each survey multiple periods can be held, with the files for each period stored in an individually named directory (the same directory can be reused for several periods as long as file names are not duplicated). This makes it very easy to produce results for repeating surveys when they are using the same definitions and procedures. It also means that by selecting a new survey the previous information for that survey is available on the definition screen. The SAS versions of output files are constructed to contain (meta-)information on the input files which have

been used to produce them, and this is displayed on screen when the outputs are to be viewed; this avoids the need to code in the information in an 8-character SAS name. An option also allows the results to be written to a text file; further output options can be obtained by manipulation from SAS, but not from GES directly. The outputs viewer and the definition screens include procedures to sort, browse and edit data without the need to exit GES. Because of the large amount of information which needs to be supplied to GES, the input screens are not really intuitive, but they do provide a common way of defining all the necessary input files.

The written documentation sent with GES is fairly basic – enough to get started and have an idea of what the package requires. The main documentation effort is on-line, where there are three types of help – the usual specific help for particular procedures, choices and actions, a list of GES error messages with their meanings and likely causes (quite a lot of the causes are not filled in), and GES methodology help, which explains the methods and gives the formulae in use within GES. This latter is very useful, and (if printed) would form a methodological guide to GES. The methodology has also been published in Estevao, Hidioglou & Särndal (1995). Support is available for GES, up to 30 days per year with a (compulsory) maintenance contract, and responses are normally available during Canadian east coast office hours (approximately 1400 to 2200 GMT). There is no dedicated support person or email address.

STATA is also command-driven, with commands entered in the command window. These must be learnt as there is no facility for selecting them from pull-down menus, but there is a review window which shows previously used commands, and these can be reselected. The syntax of commands is relatively straightforward, and there is a particular series beginning “svy” which are designed for survey analysis. There are two other windows in the STATA interface, an output window for results and messages, and a variable window which shows the names of the variables in the current dataset. The documentation available with STATA is copious, but the amount dealing with survey methods is relatively small, although the commands are clearly described. On-line help for specific commands (but not describing the background theory) is also available. Support is available by phone, fax and email, again during USA office hours (approximately 1600 to 2400 GMT). Uniquely among the packages considered here there is also a STATA listserver to which queries can be sent; the existence of this probably reflects the wider range of functions available in STATA.

WesVarPC also has an interface which leads the user through the setting-up stages, normally making choices from lists as to what should be next in the syntax statement. When the whole set of code has been constructed, it is submitted. The code is visible during the set-up process, and can be typed in directly for speed if the syntax is already known. There is a comprehensive user guide which is available for downloading with the software over the internet, which describes how to use the packages. Output is sent to text files, and inputs can be read from files in a range of formats including SAS (up to version 6.04; transfer files must be used for later versions), text, SPSS and dBase.

### 2.2.6.1 *Initial reactions of new users to the software*

During the study of these software products, the initial reactions of new users have been monitored, and these are summarised here:

CLAN	horror
GES	complicated
STATA	nice
SUDAAN	basically straightforward but a bit confusing in places
WesVar	nice interface but difficult to work out how to set up data to get the desired outputs.

### 2.2.7 **Correctness and speed**

CLAN, GES, STATA and SUDAAN all produce the same point and variance estimates for the Taylor-type variances of totals using number-raised estimation. CLAN and GES also agree on the ratio and regression estimators, with rounding error differences at only about the 10<sup>th</sup> decimal place. The artificial variance of a ratio estimator from SUDAAN as described in section 2.2.4 is rather more different from the CLAN/GES results, possibly from a double dose of rounding error, or possibly from some minor difference in the ultimate methodology used within SUDAAN in doing something it was not designed for. The simulation study gives no grounds to suggest that any of the software produces incorrect answers, and independent checking of GES at Statistics Sweden confirms this.

We did not try to run exact comparability trials, but instead give an overview of the speed of processing of these packages in the context of their use. SUDAAN and STATA are both relatively quick, taking a minute or less to produce estimates and variance estimates for a survey the size of the UK's Annual Business Inquiry (ABI), based on the simulation example, on a Pentium 166MHz PC with 128 Mb RAM (networked). Asking for jackknife estimates from SUDAAN increases the processing time slightly, but this is still the of the order of one minute. CLAN and GES take considerably longer; both have weight calculation and estimation phases; within CLAN the weight calculation phase is long (around half an hour), and then survey estimation proceeds in several minutes; for GES weight estimation takes about two minutes, but estimation takes around an hour. For GES the use of the jackknife variance estimator approximately doubles the processing time. In the context of producing survey results these times are broadly acceptable, since sampling errors are not normally a critical part of the production process. For very heavy processing or simulation work, both CLAN and GES (and GES in particular) are rather slow.

### 2.2.8 **Ease of integration with processing systems**

The ease with which the software can be integrated with processing depends very much on the actual processing system. SAS is becoming common as a tool for processing in NSIs, and where this is used the interfaces to CLAN, GES and SUDAAN are very straightforward. The ability of SAS to access databases directly for common database-platform combinations could be useful in this regard, but does not seem to be widely used in NSIs. However GES has client-server operation which allows this to be set up from within the software, and additionally allows processing on a larger machine away from the PC. Away from integrated

SAS-based systems, files must be transferred, and this usually requires some manual intervention. The range of file types supported for input and output within the packages reviewed here is sufficient that the data can be transferred fairly readily, although some reformatting may be required. In these cases there is no good automated procedure.

### 2.2.9 Costs

The costs of the various packages are given in the following table.

Package	Initial license	Annual maintenance	
CLAN	free	free	
GES	C\$30,000 for a site license (unlimited number of users) for one platform. Licenses for additional platforms cost C\$7,500 each	C\$3,000 for site license (unlimited number of users) Additional platforms cost C\$750 each	
STATA	US\$975	optional	
SUDAAN	stand-alone SAS-callable	US\$995 US\$800 (+US\$260 each additional user)	none, but upgrades must be purchased US\$400 (+US\$130 each additional user)
WesVar PC	free	free	

Table 2.1 The costs of a single license for the evaluated software packages (information correct at 1 January 1999).

## 2.3 Recommendations for variance estimation software for use in EU member states

The current position with variance estimation software is confusing. There are no clearly superior packages, and each has advantages and disadvantages which vary according to the situation in the particular survey to be processed. The group II software packages (STATA and SUDAAN) are only really appropriate when expansion estimation is used. In situations where this is the only (or perhaps predominant) method, they offer several advantages including fast processing, additional survey analysis features and a reasonably friendly interface.

Where survey estimators are more complex, from ratio estimators to GREG estimation, only CLAN and GES are really suitable in that they provide the correct variance estimators. They also produce the appropriate survey weights, which are not available from the other software. GES is very expensive and relatively slow, but has a reasonable user interface which leads through the set-up process in a logical way. CLAN is free and slightly quicker, but requires SAS programming experience and has no user interface beyond what SAS provides. This “user-unfriendliness” could be seen as a feature to prevent people with insufficient

knowledge from using the software in an inappropriate way, but is not helpful in a package designed for general use.

So as a general recommendation for all-purpose processing of the types of designs typical in business surveys, CLAN and GES are the main contenders, but in specific cases with expansion estimation, STATA and SUDAAN are equally acceptable.

### 3 Simulation study of alternative variance estimation methods

*Paul Smith, Susan Full & Ceri Underwood, Office for National Statistics  
Ray Chambers & David Holmes, University of Southampton*

The tender suggested a simulation study of the variance estimation methods available in the software, to assess the properties of the variance estimators – bias, coverage, variability, relation to the size of the estimate. As a bonus this has the effect of demonstrating that the various software packages do or do not produce the same solutions with the same model formulation and the same input data, as reported in section 2.2.7. The combinations of features (estimation method and variance calculation approach) which are available in the software considered are shown in Table 3.1.

	Taylor	Jackknife
Number raised estimation	CLAN, GES, STATA, SUDAAN	GES, SUDAAN
Ratio estimation	CLAN, GES, STATA*, SUDAAN* <sup>+</sup>	GES, SUDAAN* <sup>+</sup>
Regression estimation	CLAN, GES	GES
Constrained-weight regression estimation	STATA*	

Table 3.1 Combinations of estimators and variance estimation techniques used in the simulation study, with the packages which have been used. \* variance estimation uses weights in a manner which is not strictly valid (see section 2.2.4); <sup>+</sup> valid variances can be produced but only by using the software in a non-standard way (see section 2.2.4).

The simulation process has turned out to be a long one, and not all of the results obtained are presented here; instead we concentrate on the main messages to have emerged. Some of the results presented here seem to lack internal consistency, and on the whole it seems that the whole area will benefit from further detailed study in the future. It is hoped that the study will continue past the end of the present contract.

#### 3.1 The simulated population

##### 3.1.1 A model for data generation

For the purposes of the simulation study, data were taken from the UK's Annual Business Inquiry (ABI), which is a sample survey, cross-stratified by 5-digit industries of the SIC(92) (approximately four-digit NACE classes but slightly more detailed in places) and employment size (more detailed information on this survey is contained in the Model Quality Report, volume III chapter 3). The information on employment comes from the Inter-Departmental Business Register (IDBR) (Perry 1995), the UK's frame for business surveys. The survey data have been used to fit a model of the form

$$\log(y_i) = \beta_{h0} + \beta_{h1} \log(x_{i1} + 1) + \beta_{h2} \log(x_{i2})$$

where  $y_i$ ,  $x_{i1}$  and  $x_{i2}$  are respectively the survey value, register employment and register turnover (available from the IDBR) for unit  $i$ , and the  $\beta_{hj}$  are regression parameters to be estimated in stratum  $h$ . This model is then used to generate fitted values for the whole population of manufacturing businesses based on the values of  $x_{i1}$  and  $x_{i2}$  from the IDBR. The residuals from the model are hot-decked so that the survey outcomes are stochastic, and reflect the data which might be obtained if a real census of manufacturing industry could be undertaken. Any negative survey values which arise from this procedure are set to 0, which in fact slightly over-represents the proportion of zero responses in the simulated population. Then a collection of 1000 samples was made by repeatedly sampling from this population using the ABI design, and the information from these samples was used in the various software packages to produce estimates of totals and their corresponding sampling variance estimates.

### 3.1.2 Domains and estimators

The domains used were:

- the whole survey;
- the 2-digit industries of the SIC92, each of which corresponds to an amalgamation of strata/model groups (but see also 3.1.3 below);
- the standard statistical regions in the UK, of which there are 11. These cut completely across the stratification and the model groups, and so provide a good test of the ability of the packages to deal with domains whose size is unknown.

Estimation used three principal methods, number raised estimation, ratio estimation using register employment as the auxiliary variable, and regression estimation using one auxiliary variable (again register employment). In the last two cases the variance of the residuals was taken as proportional to the register employment value.

In this way the whole population is known, and hence the true population and domain totals are easily calculated, so that the overall error (with contributions from bias and variance) of the estimates from each of the simulated samples can be found. Using this to calculate the root mean square error, and comparing with the variance of these estimates allows us to deduce the bias in estimation. Also the variance of the estimates should correspond to the sampling variance, and the distribution of the point estimates of the sampling variance from the simulated samples can be compared with this; an additional useful piece of information here is the variability of the point estimates of sampling variance. Some further analysis looking at the relationship between the size of estimates and their estimated sampling errors may also be interesting, although we do not pursue this particular avenue of research any further in this report.

### 3.1.3 Data features

The simulated dataset has a number of features which are worth mentioning because they raise certain issues about the variance estimation process or the way in which the software

works. The two-digit industry domains are amalgamations of strata, and it is at stratum level that we are controlling for known auxiliary values in estimation, so we expect these to be relatively accurately estimated. There was a misspecification of the population data variables which resulted in one business in each two-digit industry having the wrong two-digit code, so the two-digit industries are not quite a strict amalgamation of the strata. The region-level domains are such that the regional totals of auxiliary variables are not controlled as a by-product of estimation, so the estimation includes an implicit estimation of the domain size, which adds an additional component of variability.

Some of the domains are very sparse. The misspecification described above resulted in only one business in the simulated population being in industry 29, so we have the most variable case possible (essentially binomial), with two possible estimates (one of which is zero).

The population dataset contains a few extreme values, which, in an ordinary survey situation would be adjusted or treated in some way. In this case they have been left without adjustment, which means, for example, that the variability in the total population is dominated by the variability in industry 07 and region 03. Taking examples from domains not so grossly affected allows us to see how well the different estimators perform in different situations.

## **3.2 Processing**

The speed of processing has been an issue in undertaking these simulations. STATA and SUDAAN run relatively quickly, but are quite restricted in the range of estimation models which can be used. CLAN and GES run slower, but have the weight calculation functions which are required to produce appropriate data for some of the “naïve” applications of STATA and SUDAAN methodology (see 3.3.2.1). In the case of GES the slow speed and the wide variety of estimator and variance estimator combinations has meant that the whole study has not been completed, and we will present only the preliminary results from less than the complete number of simulated samples.

## **3.3 Results**

### **3.3.1 Comparison of estimators**

The properties of the point estimators from the three different estimation models (expansion, ratio and regression) are summarised for three domains in Table 3.2. The estimates of the population total are affected by the extreme value in industry 07 and region 03. When this outlier is included in the sample it causes a huge overestimate, and when it is not included, a substantial underestimate. Because representation of this element in the simulations is not exactly in accord with its selection probability, but subject to random variation, this gives rise to some large biases in the point estimators. In other parts of the population where there are no such extreme outliers, such as in Industry 01 and region 01 in Table 3.2, the biases are very small and the mean square error is dominated by the variability of the estimates around their expectation. This is a more typical and much more expected pattern.

The main conclusion seems to be that there is very little difference in the bias and variance properties of the expansion, ratio and regression estimators for this population in strata where the population is “well-behaved” (in the mathematical sense); possibly the expansion estimator is slightly worse in general.

Estimator		Number of simulations	Average bias of point estimator (% of true total)	Standard deviation of point estimators (% of true total)	Mean squared error of point estimator (% of true total)
Total	SYG expansion	669	-6.1	171.1	171.2
	SYG ratio	663	-6.5	165.4	165.5
	SYG regression	510	-6.7	165.9	166.1
	JK expansion	317	8.5	247.7	247.8
	JK ratio	421	-14.8	97.4	98.5
	JK regression	299	-19.4	3.1	19.6
Industry 01	SYG expansion	669	-0.004	0.12	0.12
	SYG ratio	663	-0.009	0.11	0.11
	SYG regression	510	-0.044	0.11	0.12
	JK expansion	317	-0.009	0.12	0.12
	JK ratio	421	-0.012	0.11	0.11
	JK regression	299	-0.034	0.11	0.11
Region 01	SYG expansion	669	0.018	0.42	0.42
	SYG ratio	663	0.020	0.43	0.43
	SYG regression	510	0.017	0.42	0.42
	JK expansion	317	0.040	0.50	0.50
	JK ratio	421	0.013	0.43	0.43
	JK regression	299	0.039	0.53	0.53

Table 3.2 A comparison of the mean squared error characteristics of the point estimators from three different estimation models with two different estimators. Results are all taken from GES.

### 3.3.2 Comparison of variance estimators

Table 3.3, below, compares the standard deviation of the point estimates from the simulations with the average estimated standard error (the variances are averaged and then the root is taken). The SYG variance estimators are very close to the standard deviation of the point estimates, even in the samples which are affected by the extreme outlier. There are some configurations of sample data where the estimators work less well, for example with the regression estimator in industry 01 where the estimator underestimates the true variability. The biases in the SYG standard error estimators for total, industry and region domains combined for the three estimation schemes are shown in Figure 3.1. Note that Industry 29 has been omitted; this is the (spurious) industry with a single member, and the ratio estimator has a very large bias in this case (755%).

Estimator		Number of simulations	Standard deviation of point estimators (% of true total)	Average estimated standard error (% of true total)	Bias of standard error estimate (% of sd of point estimators)
Total	SYG expansion	669	171.3	171	-0.15
	SYG ratio	663	165.5	166	0.38
	SYG regression	510	166.1	166	0.26
	JK expansion	317	248.0	253	2.24
	JK ratio	421	97.5	575,715	591,259.78
	JK regression	299	31.5	21,365	678,263.93
Industry 1	SYG expansion	669	0.12	0.13	4.46
	SYG ratio	663	0.11	0.11	-0.70
	SYG regression	510	0.11	0.08	-29.70
	JK expansion	317	0.12	47.87	40,297.71
	JK ratio	421	0.11	38.36	34,778.76
	JK regression	299	0.11	40.98	38,222.72
Region 1	SYG expansion	669	0.42	0.42	0.27
	SYG ratio	663	0.43	0.44	1.79
	SYG regression	510	0.42	0.42	0.55
	JK expansion	317	0.50	69.4	13,774.04
	JK ratio	421	0.43	935,520.8	217,992,460.92
	JK regression	299	0.53	86.4	16,123.30

Table 3.3 Summary of the variance estimator properties from GES outputs, for Sen-Yates-Grundy (“Taylor”) variance estimators and (drop one) jackknife variance estimators for three example domains.

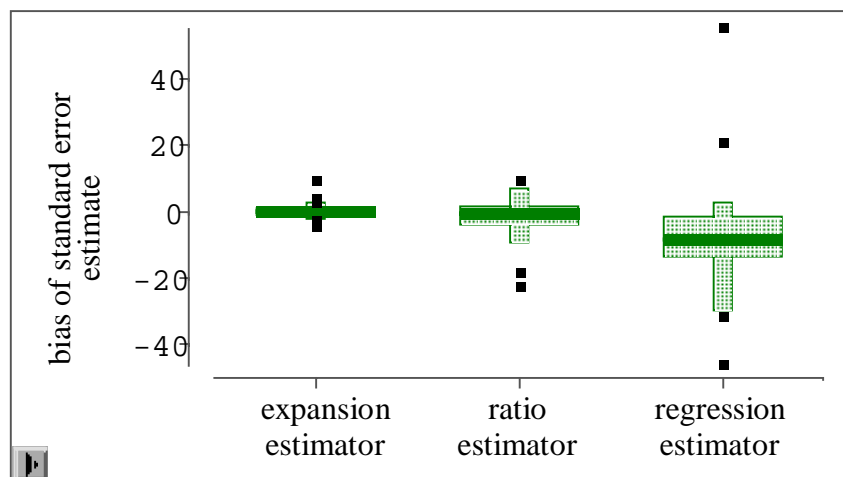


Figure 3.1 Boxplots of the bias of the SYG standard error estimators (expressed as a percentage of the standard error of the point estimates). Note that the biases for Industry 29 have been **omitted** as they swamp the rest of the information.

The jackknife estimators are extremely biased most of the time, as can be seen in Table 3.3. This is because the jackknife estimator is only strictly valid in with-replacement designs, and our design is stratified without replacement. It can be used as an approximation, and the approximation will be quite good where the sampling fraction is small (Wolter 1985, p168). In many business surveys the sampling fraction is, however, large, and in these cases the approximation can be dreadful. All is not lost, because a further approximate variance estimator can be obtained for this situation by introducing the finite population correction into the jackknife (Wolter 1985, p169). This is not an option within either GES or SUDAAN (the two packages with implementations of the stratified jackknife variance estimator), and must be included manually. An investigation of this estimator is still underway but preliminary results from 150 simulations are shown in Table 3.4.

Domain		Number of simulations	Standard deviation of point estimators (% of true total)	Average estimated standard error <sup>3</sup> (% of true total)	Bias of standard error estimate (% of sd of point estimators)
Industry 1	JK expansion	150	0.14	31.17	41,052,155.84
Industry 2		150	9.00	0.06	-100.00
Industry 3		150	0.68	49,265.27	111,542.80
Industry 6		150	1,456.78	86,465.70	-90.76
Industry 26		150	$6.7 \times 10^{-5}$	3,064.25	322,682.06

Table 3.4 Properties of jackknife sampling error estimates from runs of GES with the finite population correction included at the stratum level. Industries 6 and 25 are respectively the best and worst cases, and the sd of the point estimators in these domains may indicate a data problem.

Given Wolter’s assertion that this adjusted variance estimator is unbiased, the information on the biases of the standard error estimates in Table 3.4 doesn’t seem credible. It seems that some further work should be done to investigate whether this is driven by certain aspects of the data, or is an artifact of the (relatively) small number of replicates on which this table is based.

### 3.3.2.1 Naïve variance estimators

The variance estimators considered so far are basically the appropriate ones for the estimation methods and sample design under consideration. However, other possible combinations of inputs and the use of packages are possible, and we have called these “naïve variance estimators” because, although the inputs seem reasonable at first glance, the combination of weights and software gives an inappropriate estimator. However, under some circumstances this is the easiest approach, and it is worthwhile looking to see whether these estimators provide a sound approximation and hence whether they are practical alternatives.

First taking the ratio estimator weights from the calibration software and using those in STATA, we discover that the variance estimate is the same as for the *expansion* estimator.

<sup>3</sup> Includes the finite population correction in the calculation.

This is because the weights are still constant within strata. Using the same (ratio) weights in SUDAAN produces a variance estimator different from the number raised one, but not very different. Averaged over 50 samples, the ratio of the variance estimates (naïve estimator to true variance estimator) is from 0.53 to 1.23, but this translates into a ratio of cvs of only 0.88 to 1.08. Nevertheless this suggests that the naïve ratio variance is too close to the expansion variance, and not appropriate to ratio estimation.

Following the same approach with regression estimation is not possible because it produces negative weights in some strata, and neither STATA nor SUDAAN allow negative weights. An alternative is to use the constraining options within GES (in this case) to produce strictly positive weights wherever a solution to the calibration equations exists, and to replace the weights (for a whole stratum) with the expansion weights where such a solution does not exist. This guarantees no negative weights, but the extent of replacement and constraining can potentially cause a large increase in the variance. The results are shown in Table 3.5; note, however, that the standard deviation of the point estimates of total seem inconsistent with others presented here for the “regression” estimator. This may be due to the constraining process (see Hedlin, Falvey, Chambers & Kocic 1998).

The main indication from this table seems to be that the naïve “regression” variance severely underestimates the actual variability of the estimates obtained with regression weighting.

None of the naïve variance estimators considered here seems to offer a good approximation to the underlying variance of the estimates. This means that, where ratio or regression estimators are used in business surveys, appropriate software which explicitly takes account of the estimation model is necessary, and that software which uses only techniques for expansion weighting cannot be safely used.

domain		Number of simulations	Standard deviation of point estimators (% of true total)	Average estimated standard error (% of true total)	Bias of standard error estimate (% of sd of point estimators)
total	expansion	1000	184.66	185.13	0.26
	“regression”	733	14.26	1.59	-88.84
ind1	expansion	1000	0.08	0.13	53.97
	“regression”	732	6.84	0.03	-99.58
ind2	expansion	1000	9.56	9.42	-1.42
	“regression”	732	9.67	3.12	-67.78
reg1	expansion	1000	0.36	0.35	-3.04
	“regression”	733	23.06	0.02	-99.93

Table 3.5 Comparison of the properties of variance estimators obtained from STATA using the usual expansion weights, and using constrained or adjusted weights (see text for full description) to give a pseudo-regression estimator.

### 3.3.3 Comparison of software package outputs

*Taylor variances:* We will first look at the only estimation method common to all the packages, expansion estimation, and the “Taylor” (SYG) variance estimator. All the packages

produce identical solutions for the variance with the expansion estimator. The two packages which have appropriate processing for the ratio and regression estimators, CLAN and GES, also produce identical estimates for this estimator in standard cases, although the treatment of samples containing zero values of auxiliary variables can give rise to slight differences. Using SUDAAN twice as described in section 2.2.4 to produce a quasi-ratio estimator gives a theoretically correct variance estimator (but doesn't give a point estimate at all), but with a rather indirect implementation. However, the solution is not the same as the CLAN/GES one. The differences are shown in Table 3.6.

The general impression that the industry estimates are quite close whereas the region estimates are way out is broadly indicative of the trends in the remaining parts of the dataset. In general it seems that SUDAAN, even when apparently using a fix to give the correct form of the variance, is not appropriate software for calculating the variance of ratio estimates.

*Jackknife variances:* Only SUDAAN and GES allow the use of jackknife variance estimators in stratified designs. The default in SUDAAN is the drop-one jackknife estimator, and in GES the user must set up jackknife groups. Since the drop-one estimator is preferred (Wolter 1985 p164), this has been used here. In this case, the jackknife estimators of variance from SUDAAN and GES are identical for the number-raised estimator; neither includes the fpc, which must be added later if it is required (see section 3.3.2, above).

Domain		Number of simulations	Standard deviation of point estimators (% of true total)	Average estimated standard error (% of true total)	Bias of standard error estimate (% of sd of point estimators <sup>4</sup> )
Total	GES ratio	663	165.37	166.01	0.38
	SUDAAN ratio	150	na	314.28	90.04
Industry 1	GES ratio	663	0.11	0.11	-0.70
	SUDAAN ratio	150	na	0.11	6.33
Industry 2	GES ratio	663	9.27	8.46	-8.72
	SUDAAN ratio	150	na	6.10	-34.18
Region 1	GES ratio	663	0.43	0.44	1.79
	SUDAAN ratio	150	na	37.75	8,634.68

Table 3.6 Summary of the properties of SYG variance estimators for ratio estimation from GES and using an apparently correct fix in SUDAAN.

### 3.4 General conclusions

1. The variance estimators which are common to several packages do in fact produce the same results in each case, with rounding error contributing only after many significant figures.

<sup>4</sup> The SUDAAN figures have been compared with the variance of the GES point estimators; that is, the divisor for biases from both packages is the same in this column.

2. The Sen-Yates-Grundy variance estimators are generally very close to the actual variation in the estimates over repeated sampling.
3. The jackknife variance estimators are inappropriate in business surveys with high sampling fractions, and do not seem to be corrected by application of the finite population correction.
4. The use of software packages for estimators for which they are not designed, or the use of “naïve variance estimators” through using the right weights in the wrong formula both produce variance estimates which are very biased. These approaches are not recommended. Hence an appropriate package must be used when ratio, regression or more complex estimators are in use.

## 4 Variances in STATA/SUDAAN compared with analytical variances

David Holmes, University of Southampton

### 4.1 Expansion estimator

The usual estimator of a total in stratified sampling is

$$\hat{t}_y = \sum_{h=1}^H N_h \bar{y}_{s_h} \quad (4.1)$$

where  $\bar{y}_{s_h} = n_h^{-1} \sum_{s_h} y_k$ , and the variance is given by

$$V(\hat{t}_y) = \sum_h \frac{N_h^2 (1 - f_h)}{n_h} S_{yU_h}^2 \quad (4.2)$$

where  $S_{yU_h}^2$  is the stratum variance. This variance is estimated by

$$\hat{V}(\hat{t}_y) = \sum_h \frac{N_h^2 (1 - f_h)}{n_h} S_{ys_h}^2 \quad (4.3)$$

where  $S_{ys_h}^2$  is the sample variance in stratum  $h$ ,  $S_{ys_h}^2 = (n_h - 1)^{-1} \sum_s (y_k - \bar{y}_{s_h})^2$ .

Note: (4.1) can be written as  $\hat{t}_y = \sum_h \sum_{s_h} w_{hk} y_k$ , where  $w_{hk} = N_h / n_h$ .

### 4.2 Ratio estimator

The separate ratio estimator of a total in stratified sampling (used by ABI) is

$$\hat{t}_{y, \text{rat}} = \sum_h \frac{\bar{y}_{sh}}{\bar{x}_{sh}} t_{xh} \quad (4.4)$$

where  $\bar{x}_h = n_h^{-1} \sum_{s_h} x_k$  and  $t_{xh}$  is the stratum total of the  $x_k$ . The variance of the estimator is given by

$$V(\hat{t}_{y, \text{rat}}) = \sum_h \frac{N_h^2 (1 - f_h)}{n_h} S_{yU_h, \text{rat}}^2 \quad (4.5)$$

where

$$\begin{aligned} S_{yU_h, \text{rat}}^2 &= \frac{1}{N_h - 1} \sum_{U_h} (y_k - R_h x_k)^2 \\ &= S_{yU_h}^2 - 2R_h S_{xyU_h} + R_h^2 S_{xU_h}^2 \end{aligned} \quad (4.6)$$

is the stratum variance of the variable  $y_{hk} - R_h x_{hk}$  and  $R_h = t_{yh} / t_{xh}$ . See Cochran (1977), section 6-10. This variance can be estimated by

$$\hat{V}_1(\hat{t}_{y, rat}) = \sum_h \frac{N_h^2 (1-f_h)}{n_h} (S_{ys_h}^2 - 2\hat{R}_h S_{xys_h} + \hat{R}_h^2 S_{xs_h}^2) \quad (4.7)$$

where  $\hat{R}_h = \bar{y}_{sh} / \bar{x}_{sh}$ , (that is, the stratified version of equation 6.11 in Cochran). An alternative variance estimator (see equations 6.12 and 6.13) is

$$\hat{V}_2(\hat{t}_{y, rat}) = \sum_h \frac{N_h^2 (1-f_h)}{n_h} \frac{\bar{x}_{U_h}^2}{\bar{x}_{s_h}^2} (S_{ys_h}^2 - 2\hat{R}_h S_{xys_h} + \hat{R}_h^2 S_{xs_h}^2) \quad (4.8)$$

Note: (4.4) can be written  $\hat{t}_{y, rat} = \sum_h \sum_{s_h} w_h y_k$  where  $w_h = \frac{t_{xh}}{\sum_{s_h} x_k}$ .

### 4.3 What does SUDAAN do?

For stratified random sampling, the variance formula used is

$$\hat{V} = \sum_h (1-f_h) n_h S_{zs_h}^2 \quad (4.9)$$

where  $S_{zs_h}^2 = \frac{1}{n_h - 1} \sum_{s_h} (z_k - \bar{z}_{s_h})^2$ , and  $z_k$  is the ‘‘appropriate linearised value’’. Note that the variance formula corresponds to the design option DESIGN = STRWOR.

So, if we want to estimate the variance of the usual expansion estimator (see (4.1)), we use DESCRIPT. The ‘‘linearised value’’ is  $z_k = w_h y_k$ , and so long as  $w_h = N_h / n_h$  (that is, the sampling weight), the variance formula in (4.9) gives the **correct** variance estimator of (4.3).

What about the variance estimator for the ratio estimator defined in (4.4)? Can SUDAAN be tricked by defining the weight to be  $w_h = \frac{t_{xh}}{\sum_{s_h} x_k}$ ? The answer is **no**. The ratio estimator

itself will be correct, but the variance formula in (4.9) with  $z_k = w_h y_k = \frac{t_{xh}}{\sum_{s_h} x_k} y_k$  will give

$$\hat{V} = \sum_h \frac{N_h^2 (1-f_h)}{n_h} \frac{\bar{x}_{U_h}^2}{\bar{x}_{s_h}^2} S_{ys_h}^2 \quad (4.10)$$

This is **not** the variance given in (4.7) or (4.8).

The answer is to use the RATIO procedure. In general, we can estimate the ratio for any subgroup  $d$  as

$$\hat{R}_d = \frac{\sum_h \sum_{s_h} \delta_{hk}(d) w_h y_k}{\sum_h \sum_{s_h} \delta_{hk}(d) w_h x_k} \quad (4.11)$$

where  $\delta_{hi}(d) = \begin{cases} 1 & \text{if sample unit } k \text{ is in subgroup } d \\ 0 & \text{otherwise} \end{cases}$ . The “linearised value”

$$z_k(d) = \frac{\delta_{hk}(d)w_h(y_k - \hat{R}_d x_k)}{\sum_h \sum_i \delta_{hk}(d)w_h y_k} \quad (4.12)$$

is substituted into (4.9) to obtain the variance estimate. So, in the special case where the strata ( $h$ ) are defined as the subgroups ( $d$ ), we have from (11)  $\hat{R}_h = \bar{y}_{s_h} / \bar{x}_{s_h}$ . The “linearised value” in (4.12) becomes

$$z_k = \frac{N_h/n_h (y_k - \hat{R}_h x_k)}{N_h \bar{x}_{s_h}} = \frac{(y_k - \hat{R}_h x_k)}{n_h \bar{x}_{s_h}}$$

and substituting this in (4.9) we get

$$\hat{V}(\hat{R}_h) = \frac{(1-f_h)}{n_h \bar{x}_{s_h}^2} (S_{ys_h}^2 - 2\hat{R}_h S_{xys_h} + \hat{R}_h^2 S_{xs_h}^2) \quad (4.13)$$

If this is multiplied by  $(N_h \bar{x}_{U_h})^2$  and summed over  $h$ , we get the variance estimate obtained in (4.8). If, instead, this is multiplied by  $(N_h \bar{x}_{s_h})^2$  and summed over  $h$ , we get the variance estimate obtained in (4.7).

## 5 References

- ANDERSSON, C. & NORDBERG, L. (1998) *A user's guide to CLAN 97 – a SAS-program for computation of point- and standard error estimates in sample surveys*. Stockholm: Statistics Sweden.
- COCHRAN, W .G. (1977) *Sampling techniques*, third edition. New York: Wiley
- DEVILLE, J.-C. & SÄRNDAL, C.-E. (1994) Variance estimation for the regression imputed Horvitz-Thompson estimator. *Journal of Official Statistics*, **10**, 381-394.
- ESTEVAO, V., HIDIROGLOU, M.A. & SÄRNDAL, C.-E. (1995) Methodological principles for a generalized estimation system at Statistics Canada. *Journal of Official Statistics*, **11**, 181-204.
- HEDLIN, D., FALVEY, H., CHAMBERS, R. & KOKIC, P. (1998) The effective use of auxiliary information in a business survey. In *NTSS '98 – International Seminar on New Techniques and Technologies for Statistics, Contributed Papers*, pp235-240.
- KOKIC, P.N. (1998) Estimating the sampling variance of the UK Index of Production. *Journal of Official Statistics*, **14**, 163-179.
- KOKIC, P.N. & SMITH, P.A. (1999a) Winsorisation of outliers in business surveys. Submitted to *Journal of the Royal Statistical Society, Series D*.
- KOKIC, P.N. & SMITH, P.A. (1999b) Outlier-robust estimation in sample surveys using two-sided winsorisation. Submitted to *JASA*.
- NORDBERG, L. (1998) *On variance estimation for measures of change when samples are co-ordinated by a permanent random number technique*. R&D Report 1998:6, Statistics Sweden.
- OHLSSON, E. (1995) Coordination of samples using permanent random numbers. In *Business survey methods* (eds. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge & P.S. Kott), pp. 153-169. New York: Wiley.
- PERRY, J. (1995) The Inter-Departmental Business Register. *Economic Trends* **505**, November 1995.
- RAO, J.N.K. & SHAO, J. (1996) On balanced half-sample variance estimation in stratified sampling. *Journal of the American Statistical Association*, **68**, 612-614.
- RUBIN, D.B. (1986) Basic ideas of multiple imputation for nonresponse. *Survey Methodology*, **12**, 37-47.
- RUBIN, D.B. (1987) *Multiple imputation for nonresponse in surveys*. New York: Wiley.
- SÄRNDAL, C.-E. (1992) Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology*, **18**, 241-252.

- SÄRNDAL, C.-E. & SWENSSON, B. (1987) A general review of estimation for two phases of selection with applications to two-phase sampling and non-response. *International Statistical Review*, **55**, 279-294.
- SÄRNDAL, C.-E., SWENSSON, B. & WRETMAN, J. (1992) *Model-assisted survey sampling*. New York: Springer-Verlag.
- SEN, A.R. (1953) On the estimate of variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **5**, 119-127.
- SHAH, B.V., FOLSOM, R.E., LAVANGE, L.M., WHEELESS, S.C., BOYLE, K.E. & WILLIAMS, R.L. (1995) *Statistical Methods and mathematical algorithms used in SUDAAN*. North Carolina: Research Triangle Institute.
- SLOOTBEEK, G.T. (1998) Bias correction in the balanced half sample method if the number of sampled units in some strata is odd. *Journal of Official Statistics*, **14**, 181-188.
- WOLTER, K.M. (1985) *Introduction to variance estimation*. New York: Springer-Verlag.
- YATES, F. & GRUNDY, P.M. (1953) Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society, Series B*, **15**, 253-261.