Journal of Official Statistics, Vol. 15, No. 2, 1999, pp. 335–350

Field Substitution and Unit Nonresponse

Vasja Vehovar [1]

Field substitution is used to compensate for unit nonresponse in sample surveys. Unlike methods such as weighting or imputation, substitution preserves the designed (optimal) structure of the sample, and this can be a source of certain advantages over the alternative methods. The most important advantage is the potential gain in precision, particularly when two-stage cluster samples are used. However, due to specific properties of the increase in sampling variance, this gain is usually small, and practitioners who use substitution often overlook this fact. On the other hand, substitution introduces an extra bias into the estimates. Additionally, severe practical difficulties arise in the field work process, and these will generally outweigh the benefits of this approach. Thus, despite its relatively frequent use in probability samples, the substitution can rarely be justified. Nevertheless, specific circumstances in which this practice can be theoretically and practically advantageous do exist. We encounter such situations in samples with small take per cluster or stratum, where efficient field controls are provided, as with telephone surveys.

Key words: Missing data; survey methodology.

1. Introduction

Unit nonresponse is the commonly accepted term for an eligible unit which has been selected in a sample but which becomes missing in the field work stage of the survey. This can introduce distortions into statistical inference; consequently a variety of ways of minimising the problem have been developed in the design, field and processing stages of the survey.

Field substitution occurs when a nonresponding unit is replaced by a substitute (reserve) unit during the field work stage of the survey process. The substitution procedure is thus a specific tool for coping with unit nonresponse and, strictly speaking, it is a form of imputation; we impute a substitute unit instead of a non-responding one.

There are different ways of selecting substitutes; here we concentrate exclusively on substitute units that are selected with a probability mechanism. Attempts to select substitutes that will match the characteristics of nonrespondents are briefly discussed when practical aspects are considered. However, we will not discuss here any method that selects the substitute units purposely or with some other type of nonprobability mechanism.

[1] University of Ljubljana, Faculty of Social Sciences, 1109 Ljubljana, PO Box 2547, Slovenia. E-mail: vasja.vehovar@uni-lj.si

Acknowledgments: This research was supported by a research grant from the Ministry of Science and Technology of the Republic of Slovenia. It was also supported by the Fulbright fellowship grant in 1998 which made possible a period of study at the Institute for Social Research, University of Michigan, U.S.A. The author wishes to thank Jim Lepkowski for many constructive comments.

Textbooks on survey methodology, and survey sampling in particular, mention substitution very briefly, e.g., Kish (1965), Lessler and Kalsbeek (1991), or not at all, e.g., Cochran (1977), Groves (1989), Särndal, Swensson, and Wretman (1992). In general, the literature is not in favour of this option, at least not with probability samples.

A comprehensive overview of this procedure was made by Chapman (1983). He rejected the general criticism that substitution would not remove the nonresponse bias. He called such arguments ''unfair criticism,'' because all the methods for handling survey nonresponse suffered from this basic weakness and there was no research showing that other procedures would perform better. He reviewed the advantages and disadvantages of this procedure in four empirical studies and concluded that there was no clear theoretical or empirical evidence for either accepting or rejecting this practice as a whole. In his later work (Chapman and Roman 1985) it was found that substitution had potential as a viable procedure in Random Digit Dialing (RDD) surveys by telephone. Particularly in RDD telephone cluster samples, certain gains in sampling variance were observed as compared to the alternative weighting adjustments. On the other hand, an extra bias was observed with substitution; however, a detailed analysis of the mean squared error was not performed. Similar results were reported in other research done at the U.S. Bureau of the Census (Biemer, Chapman, and Alexander 1990).

Nathan (1980) discussed the specific role of substitution in achieving an exact sample size. He found that, with respect to the number of initial contacts and the sampling variance, the substitution performed in approximately the same manner as the fixed initial sample.

Marliani and Pacei (1993) discussed severe problems with this procedure in the Italian Family Expenditure Survey; a considerable bias was introduced with this practice. Similar findings were reported in the Slovenian Labour Force Survey (Vehovar 1993). An extensive evaluation of the Slovenian General Social Survey (GSS) showed that this practice cannot be justified from the bias-variance aspects of the estimates (Vehovar 1995).

Forsman and Berg (1992) treated the problem of substitution in the daily replicates of telephone samples. They found little effect of this procedure on the bias of the estimates; however, comparisons based on mean squared error were not performed.

From a practical point of view the EU *Labour Force Survey* and the *Family Budget Survey* are particularly interesting as they made use of substitution in some European countries. Conditionally – under specific circumstances – substitution was even recommended when the nonresponse rate exceeded 35% (Verma and Gabilondo 1993, p. 92).

Besides the above-described literature there exists a practice of substitution in many academic institutions and statistical offices. We encounter the use of substitution in academic surveys in Belgium[2], Spain[3], and Slovenia (Štebe 1995). The procedure is used in certain government surveys in Poland[4], Bulgaria[5], Spain, Portugal, Greece, and

[2] Personal communication with Jaak Billiet, Department of Sociology, Katholieke Universiteit Leuven, Belgium, September 1998.
[3] Personal communication with Juan Javier Sanches Carrion, Facultad de Ciencias Politicas y Sociologia, Universidad Complutense, Ciudad Universitaria, Madrid, July 1990.
[4] Personal communication with Malgorzata Zyra, Central Statistical Office, Poland, September 1998.
[5] Personal communication with Bogdan Bogdanov, Central Statistical Office, Bulgaria, July 1993.

Italy (Verma 1992, p. 14). Evidence of its use can be found also in statistical offices in developing countries, from South Africa[6] to the Philippines and Saudi Arabia[7]. It is true, however, that documentation about this practice is often very scarce. Additionally, in certain environments, particularly in academic and government surveys in the U.S. and in many European countries, the substitution procedure is strictly not used.

We can conclude that substitution is rarely recommended in textbooks. In addition, very limited research has been performed that would provide a basis for justifying or rejecting this practice. The lack of a more profound discussion of the bias-variance properties is particularly critical. On the other hand, we can observe a relatively widespread use of this procedure in many probability sample surveys. Of course, this practice is also extensively employed in market research, with quota samples just as a specific form of substitution. However, as already mentioned, we are concerned here only with the probability methods of sample selection.

In this article we question whether substitution can be justified, and if so, in what circumstances. We start with an outline of the problem (1). Next, the bias-variance issues are discussed (2), the practical aspects are evaluated (3) and the conclusions are summarised (4).

2. The Bias and the Variance

We restrict our analysis to the estimate $\bar{y}$ of the population mean. The estimate $\bar{y}_{SUB}$ based on a sample with substitution will be compared with the estimate $\bar{y}_r$ based on a sample of respondents where alternative adjustments (weighting, imputation) are applied.

When the variance is discussed we assume that the data is *missing at random* (MAR), which is often reduced to the assumption that the data is *missing completely at random* (MCAR) within the level (area, cluster, strata) where the substitution is performed (Little and Rubin 1987). Although this is a moderately strong assumption, it is nevertheless often accepted also with other methods (weighting, imputation) that compensate for unit nonresponse within certain adjustment cells (Kalton 1983). In the case of a two-stage cluster design with relatively small clusters – which will be our main concern here – such an assumption is reasonable, at least when variance is discussed. Of course, the issue of the bias will be treated separately and *without* such an assumption.

We also assume that the substitute units have the same field work costs per unit as the initial units, because the field operations are the same in both situations. The similarity in cost was also confirmed in Biemer et al. (1990) and in Vehovar (1995). There may be some differences in overhead (administrative, computing) expenditures, which will be discussed later in Section 3. With respect to the level and the structure of the field work costs we thus assume that they will not interfere with the (bias-variance) evaluations of the substitution.

Finally, we should repeat that at this point we are going to discuss only the procedures that strictly follow the principles of probability sample selection.

[6] Personal communication with Mick Couper, Institute for Social Research, University of Michigan, U.S.A., May 1998.
[7] E-mail communication with David Megill, U.S. Bureau of the Census, June 1998.

2.1. The bias

Following Cochran (1977, p. 359), we assume that the population consists of two strata: respondents and nonrespondents. The units from the first stratum respond if they are selected into the initial sample but the units from the other stratum do not. We will further split the (initial) respondents into *secondary respondents* and *secondary nonrespondents* according to their behaviour when contacted as substitute units. The secondary nonrespondents would respond if included in the initial sample, but they would not respond if selected as substitutes. Contrary to this, the secondary respondents would respond on both occasions. A typical source of variation between the two groups is the number of contacts. For example, the initial units may be contacted up to five times, whereas the substitute units are contacted only up to three times. Or, simply, less effort is put into the contact with substitute units, which results in a higher refusal and noncontact rate. Naturally, if substitute units are selected under exactly the same conditions as the initial units, there will be no extra bias arising from this procedure. However, in practice, this is almost never the case.

In a nonresponse situation, the sample obtained includes only the responding units. The expected value of the sample mean for respondents $\bar{y}_r$ equals the population mean of respondents $\bar{Y}_r$ and not the true population mean $\bar{Y}$. We can write the two components of the population mean as follows:

$$\bar{Y} = (1 - \bar{M})\bar{Y}_r + \bar{M}\bar{Y}_n \tag{1}$$

and we end up with the well-known expression (Cochran 1977, p. 361) for the *nonresponse bias*, which can be written in the following form:

$$\text{Bias}(\bar{y}_r) = \text{E}(\bar{y}_r) - \bar{Y} = \bar{Y}_r - \bar{Y} = \bar{M}(\bar{Y}_r - \bar{Y}_n) \tag{2}$$

where $\bar{Y}_n$ denotes the population mean for nonrespondents and $\bar{M}$ is the nonresponse rate (or, the proportion of nonresponding units).

Similarly we express the expected value of the estimate $\bar{y}_{SUB}$ of the population mean when substitution is used. Here, the estimate $\bar{y}_{SUB}$ also consists of two components: that of the initial respondents and that of the secondary respondents who replaced the nonresponding units. Following (1) we have the expression for the overall mean:

$$\text{E}(\bar{y}_{SUB}) = \bar{Y}_{SUB} = (1 - \bar{M})\bar{Y}_r + \bar{M}\bar{Y}_{sr} \tag{3}$$

where $\bar{Y}_{sr}$ denotes the population mean of secondary respondents. With some algebra, the above expression can be developed in two alternative ways:

$$\bar{Y}_{SUB} = \bar{Y} + \bar{M}(\bar{Y}_{sr} - \bar{Y}_n) = \bar{Y}_r + \bar{M}(\bar{Y}_{sr} - \bar{Y}_r) \tag{4}$$

Of course, by definition we have the relation $\bar{Y}_r = (1 - \bar{M}_{sn})\bar{Y}_{sr} + \bar{M}_{sn}\bar{Y}_{sn}$. We should also note that for substitute units the response rate can be expressed as $(1 - \bar{M})(1 - \bar{M}_{sn})$. The right part of Equation (3) can be thus further extended, so that we express the bias of the estimate $\bar{y}_{SUB}$ in the following form:

$$\text{Bias}(\bar{y}_{SUB}) = \bar{Y}_{SUB} - \bar{Y} = \bar{M}(\bar{Y}_r - \bar{Y}_n) + \bar{M}\bar{M}_{sn}(\bar{Y}_{sr} - \bar{Y}_{sn}) \tag{5}$$

where $\bar{M}_{sn}$ stands for the proportion of secondary nonrespondents among all initial respondents, and $\bar{Y}_{sn}$ denotes the population mean for secondary nonrespondents.

We will call the Bias($\bar{y}_{SUB}$) in expression (5) a *gross substitution bias*, and its last term $\bar{M}\bar{M}_{sn}(\bar{Y}_{sr} - \bar{Y}_{sn})$ will be called a *net substitution bias*, as this is an additional bias, added to the nonresponse bias (2) by the substitution procedure itself.

It is an empirical fact in almost all nonresponse research – including examples in standard texts such as Cochran (1977, p. 360) and Kish (1965, p. 544) – that the characteristics of late respondents (with more call-backs) are closer to those of nonrespondents than to those of the units which respond earlier. For example, males are more likely to be the nonrespondents than females. In addition, within the initial contacts, their proportion among respondents is usually smaller than within later contacts. However, the largest proportion of males can be found among the nonrespondents. According to our terminology, the late respondents are typical examples of secondary nonrespondents and their characteristics are thus between those of the initial respondents and those of the nonrespondents. These empirical characteristics of the secondary nonrespondents are the main reason why $\bar{Y}_{sn}$ generally lies between $\bar{Y}_{sr}$ and $\bar{Y}_n$. Since by the very definition $\bar{Y}_r$ is between $\bar{Y}_{sn}$ and $\bar{Y}_{sr}$, both terms on the right side in (5) are of the same sign. Due to the product $\bar{M}\bar{M}_{sn}$, the net substitution bias $\bar{M}\bar{M}_{sn}(\bar{Y}_{sr} - \bar{Y}_{sn})$ will be relatively small. However, it will almost always enlarge the initial nonresponse bias (2). No cancellation occurs here as is often the case with the components of survey errors.

The above derivations are typical for situations in which fewer contacts are performed with substitute units. However, in addition to this, the interviewers may also put less effort into contacts with the initially contacted units. The nonresponse rate $\bar{M}$ thus becomes higher and, as a consequence, the stratum of respondents shrinks and its mean $\bar{Y}_r$ may move further away from $\bar{Y}_n$. For example, the interviewers may omit the difficult-to-survey units more often in comparison to a situation in which they know they cannot select a substitute unit. In such cases, the nonresponse component $\bar{M}(\bar{Y}_r - \bar{Y}_n)$ in (5) may also increase.

In the absence of strict field work controls the net substitution bias can increase dramatically, because the factors $\bar{M}$ and $\bar{M}_{sn}$, as well as the differences $(\bar{Y}_{sr} - \bar{Y}_{sn})$ and $(\bar{Y}_r - \bar{Y}_n)$ are defined by the field work procedure itself.

In order to empirically evaluate the net substitution bias we must compare the estimates from the sample with substitutes, and the estimates from the sample of initial units which was adjusted with the alternative weighting (performed at the same level as the substitution). Another validation can be done when we know the individual population values. This is often the case with demographic characteristics when the register of population serves as a sampling frame.

Both approaches have confirmed that in highly controlled face-to-face surveys (General Social Surveys, Labour Force Survey, Family Expenditure Survey) with sampling from the register and with the same number of contacts (five) for initial and for substitute units, the net substitution bias was consistent, of the same sign as the nonresponse bias, and around 0.5% of the estimate for the variables sex, age (group 18–25), rural-urban component and education (Vehovar 1995). There was also a clear tendency towards additional under-representation of the units that are often inclined to be non-respondents (urban unit, uneducated, male).

With attitudinal variables the relative biases were higher. However, at least with surveys up to a few thousand units, the relative bias was much smaller than the coefficient of variation of the corresponding estimate.

Of course, the above-discussed results apply only to the bias within a cell (cluster) where the substitution is performed. If we have a simple random sample, it applies to the whole sample. The nonresponse bias arising from the different response rates across levels (adjustment cells) will be equally removed by substitution, as with other adjustments (weighting, imputation) performed at the same level (Kalton 1983). Of course, such a component of the nonresponse bias has a sign which is independent of the net substitution bias. However, with respect to the remaining within-cell component of the bias, result (5) applies.

2.2. *The variance*

We discuss the variance issues separately from the bias. When proper inclusion probabilities are used, we therefore assume that the estimates are unbiased, $\mathrm{E}(\bar{y}_r) = \mathrm{E}(\bar{y}_{SUB}) = \bar{Y}$.

A comparison of a simple random sample variance $\mathrm{Var}(\bar{y}_{SUB})$ where the substitution is applied and the variance $\mathrm{Var}(\bar{y})$ where there is no nonresponse shows that the results are almost the same. Extremely large differences in the elementary variation among secondary respondents and secondary nonrespondents are needed in order to have a significant difference in two sampling variances. As this is highly unlikely, we will not go into the details of this result here. Similarly, it can be shown that these differences remain negligible also in complex samples where strata and clusters are introduced. In general, we thus have $\mathrm{Var}(\bar{y}_{SUB}) \doteq \mathrm{Var}(\bar{y})$, and will refer to this result often in this section.

An example that is of primary concern here is a two-stage cluster sample. There are two reasons for this. *First*, considerable gains in precision may occur in this design when substitution is applied. *Second*, this is a design where substitution is most often used in survey practice.

We will compare the variance of the unweighted mean $\bar{y}_{SUB}$ where substitution is used with the variance of $\bar{y}_r$ in the sample of respondents where alternative adjustments (weighting, imputation) are performed at the same cluster level.

2.2.1. Two-stage cluster sample

First, let us consider the design (Kish 1965, p. 170; Cochran 1977, p. 277) with sampling without replacement, with clusters of equal size and with uniform sampling rates, and variance:

$$\mathrm{Var}(\bar{y}) = \left(1 - \frac{a}{A}\right)\frac{1}{a}\sum_{i=1}^{A}\frac{(\bar{Y}_i - \bar{Y})^2}{A-1} + \left(1 - \frac{b}{B}\right)\frac{1}{ab}\sum_{i=1}^{A}\sum_{j=1}^{B}\frac{(Y_{ij} - \bar{Y}_i)^2}{A(B-1)} \tag{6}$$

where a and A are the numbers of clusters in the sample and in the population, b is the sample size of the clusters, B is the population size of the clusters, Y_{ij} denotes the population value of the j-th unit in the i-th cluster, and $\bar{Y}_i$ stands for the population cluster mean.

More generally, in two-stage cluster designs the variances of the estimators of the population mean – including HT estimator, ratio estimator and PPZ estimator (Cochran 1977, ch. 10–11) – can be written in the following form:

$$\mathrm{Var}(\bar{y}_{SUB}) \doteq \mathrm{Var}(\bar{y}) = \left(1 - \frac{a}{A}\right)\frac{1}{a}U + \frac{1}{Aa}\sum_{i=1}^{A}\left(1 - \frac{b_i}{B_i}\right)\frac{1}{b_i}V_i \tag{7}$$

where U and V_i are the fixed population quantities (independent of the design parameters b and a), B_i stands for the population size of the i-th cluster and b_i for its sample size.

2.2.2. Nonresponse mechanism

Let us denote the initial sample size and the number of responding units by n and n_r respectively. The number of responding units in the i-th cluster is denoted by b_{ri}. We assume a uniform response rate $\bar{R} = (1 - \bar{M})$ with $\mathrm{E}(n_r) = \bar{R}n$. To obtain n respondents we have to start with a larger initial sample, $n^* = n/\bar{R}$. Similarly, we have an increased initial take per cluster, $b_i^* = b_i/\bar{R}$, so that we end up with b_{ri}^* respondents in the i-th cluster. Only then do we obtain $\mathrm{E}(n_r^*) = n$ and $\mathrm{E}(b_{ri}^*) = b_i$. The enlarged initial sample thus ensures that, after the nonresponse process, the proper comparisons with the sample based on substitution can be made.

First, let us consider the simplest case (6). We start with the initial sample size $n^* = n/\bar{R}$ and we apply the uniform nonresponse mechanism, so that we observe $\bar{y}_r^*|b_{ri}^*$, the sample mean in the increased initial sample being conditional on observed data b_{ri}^*. Of course, proper inclusion probabilities have to be used, so we have the estimate:

$$\bar{y}_r^*|b_{ri}^* = \frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{b_{ri}^*} y_{rij}\frac{b_i}{b_{ri}^*} \tag{8}$$

and the corresponding variance:

$$\mathrm{Var}(\bar{y}_r^*|b_{ri}^*) = \left(1 - \frac{a}{A}\right)\frac{1}{a}\sum_{i=1}^{A}\frac{(\bar{Y}_{ri} - \bar{Y}_r)^2}{A-1} + \frac{1}{Aa}\sum_{i=1}^{A}\sum_{j=1}^{B_{ri}}\frac{\left(1 - \dfrac{b_{ri}^*}{B_{ri}}\right)(Y_{rij} - \bar{Y}_{ri})^2}{(B_{ri}-1)b_{ri}^*} \tag{9}$$

Here, B_{ri} stands for the number of responding units in the i-th population cluster, Y_{rij} refers to the j-th responding unit in the i-th cluster, and $\bar{Y}_{ri}$ refers to the population cluster mean for respondents in the i-th cluster. Assuming the MCAR property for the nonresponse mechanism within each cluster, expression (9) differs from expression (6) only in the variable term b_{ri}^* which has moved into the summation symbol. We additionally assume that the terms b_{ri}^* are positive. We can now write the general expression that corresponds to (7) for the conditional variance in the case of nonresponse:

$$\mathrm{Var}(\bar{y}_r^*|b_{ri}^*) = \left(1 - \frac{a}{A}\right)\frac{1}{a}U + \frac{1}{Aa}\sum_{i=1}^{A}\left(1 - \frac{b_{ri}^*}{B_{ri}}\right)\frac{1}{b_{ri}^*}V_i \tag{10}$$

Due to the assumptions, the terms U and V_i from (7) remain unchanged in (10). The variances (9) and (10) are thus the proper expressions for population variances when correct inclusion probabilities (distorted by the nonresponse mechanism) are taken into account.

The estimation of the conditional sampling variance is not our primary concern here. In practice, the weights proportional to $w_i = b_i/b_{ri}^*$ will be attached to the respondents. Only with these weights will the variance estimation programs correctly estimate the corresponding population value (9) or (10). A similar result can also be achieved with imputations. However, the expected value of the estimate of the variance obtained with imputation procedures cannot be smaller than (10), so we will concentrate only on the weighting adjustments.

2.2.3. The comparisons

We have to calculate the unconditional variance $\mathrm{Var}(\bar{y}_r^*)$. As we assume $\bar{y}_r$ is unbiased, we also have $\mathrm{E}(\bar{y}_r^*|b_{ri}^*) = \bar{Y}$ and $\mathrm{VarE}(\bar{y}_r^*|b_{ri}^*) = 0$. Thus, only the expected value of (10) needs to be considered in the conditional variance formulae, i.e., $\mathrm{EVar}(\bar{y}_r^*|b_{ri}^*)$. However, the first term in (10) and in (7), the *between variance* component V_B, is a fixed quantity and only the second term varies from sample to sample. Its variation is based on the variability of the actual take b_{ri}^* per cluster.

In a simple, but realistic case, we assume that V_i and b_{ri}^* are independent, the nonresponse mechanism is a uniform mechanism with the parameter $\bar{R}$, and the population clusters are relatively large, so that the factor b_{ri}^*/B_{ri} is small or negligible, i.e., $(1 - b_{ri}^*/B_{ri}) = (1 - b_i/B_i) \approx 1$. We thus have the following expression:

$$\mathrm{Var}(\bar{y}_r^*) = \mathrm{EVar}(\bar{y}_r^*|b_{ri}^*) = \left(1 - \frac{a}{A}\right)\frac{1}{a}U + \frac{1}{Aa}\sum_{i=1}^{A}\left(1 - \frac{b_{ri}^*}{B_{ri}}\right)\left(\mathrm{E}\frac{1}{b_{ri}^*}\right)V_i \tag{11}$$

so that variances (11) and (7) differ only in the factors $\mathrm{E}(1/b_{ri}^*)$ and $1/b_i$. Thus, the increase in the second component of the variance (11) over the corresponding *within variance* component V_W in (7) is based on:

$$\mathrm{VIF}_W = \mathrm{E}\left(\frac{1}{b_{ri}^*}\right)\Big/\left(\frac{1}{b_i}\right) = b_i\mathrm{E}\left(\frac{1}{b_{ri}^*}\right) \tag{12}$$

In Table 1, the increase (12) is illustrated. The analytical calculations are based on a truncated hypergeometric distribution for the simplest design (6), with the population cluster size $B_i = B = 1,000$. The brackets in Table 1 indicate that more than 1% of the clusters were omitted (truncation) because no unit in the cluster responded. In the case of $B = 100$ the figures in Table 1 would be roughly 10% lower. For large b, the approximation with binomial distribution and Taylor linearisation can be used, as is done in Cochran (1977, p. 135).

Of course, the above increase refers only to the second component in (11). The proportion of the within variance component can be expressed as a function of the intracluster correlation ρ and the actual size of the cluster $\mathrm{E}(b_{ri}^*) = b_i$. In a special case (6) we can use well-known relations (Kish 1965, p. 171) to obtain the results in Table 2.

We can now express the overall increase of $\mathrm{Var}(\bar{y}_r^*)$ in (11) over $\mathrm{Var}_{SUB}(\bar{y})$ in (7):

$$\mathrm{VIF} = \frac{\mathrm{Var}(\bar{y}_r^*)}{\mathrm{Var}(\bar{y}_{SUB})} = \frac{V_B + \mathrm{VIF}_W * V_W}{V_B + V_W} = 1 + (\mathrm{VIF}_W - 1)\frac{V_W}{V_B + V_W} \tag{13}$$

Table 1. Increase ($\mathrm{VIF}_W - 1$) (%) at $B = 1{,}000$

b^*	$\bar{M} = (1 - \bar{R})$ – nonresponse rate				
	0.1	0.2	0.3	0.4	0.5
3	5.7	11.2	(16.0)	(18.2)	(18.2)
4	3.9	9.2	14.9	(19.4)	(21.7)
5	2.8	6.9	12.3	18.0	(22.4)
10	1.3	2.6	4.7	8.0	13.0
15	0.7	1.6	2.7	4.4	7.1
30	0.3	0.6	1.1	1.7	2.5

Table 2. Proportion of the within variance $V_W/(V_B + V_W)$

b	ρ – intracluster correlation					
	0.005	0.01	0.02	0.05	0.10	0.20
3	0.99	0.97	0.94	0.86	0.75	0.57
4	0.98	0.96	0.93	0.83	0.70	0.51
5	0.98	0.95	0.91	0.79	0.64	0.44
10	0.95	0.91	0.83	0.66	0.47	0.29
15	0.93	0.87	0.77	0.56	0.38	0.21
30	0.87	0.77	0.62	0.39	0.23	0.12

The increase VIF can be thus obtained by multiplying the corresponding cells in Tables 1 and 2. Obviously, the increase is relatively small. The within variance component V_W often takes only about a half of the sampling variance, so with a response rate at least $\bar{R} = 0.6$ and with a take $b = 5$ or larger, the increase will be below VIF = 1.05. However, there do exist some extreme situations – small b, small ρ, small $\bar{R}$ – with an increase close to VIF = 1.20.

As an approximation, the above results can be used also for other sampling strategies within a two-stage sampling scheme.

The analytical results above and the illustrations in Tables 1 and 2 are *essential* for an understanding of the gains in precision when substitution is used. It is obvious that we can benefit from substitution only in very specific circumstances which are not encountered very often.

2.2.4. Simulations

We will verify the above calculations with a simulation study. Let us observe the variable y with a normal distribution $y : N(\mu, \sigma)$, $\mu = 700$, $\sigma = 300$. The design used in the simulations partially followed the design for a Slovenian GSS with $n = 2{,}100$ and $a = 140$ primary sampling units, so we have a design (6). We assume $\rho = 0.07$ and $B_i = B = 1{,}000$, which – using the standard relations for this design (Kish 1965, p. 171) – gives the corresponding components (between, within) of the elementary variances σ_B^2 and σ_W^2. With substitution we have $n = 140 \times 15 = 2{,}100$, but with a uniform nonresponse mechanism $\bar{R} = 15/18 = 0.83$ we start with $n^* = 140 \times 18 = 2{,}520$. Also used was an extreme design with $\bar{R} = 3/5 = 0.60$ and $n^* = 700 \times 5$. The simulations were performed in the following steps:

1. The initial sample (size n or n^*) was generated in two successive steps, $\bar{y}_i : N(700, \sigma_B)$ and $y_{ij} : N(\bar{y}_i, \sigma_W)$.
2. The missing data was generated with $\bar{R} = 0.83$ or $\bar{R} = 0.60$.
3. The nonresponse weights $w_i = b_i/b_{ri}^*$ were attached to the respondents.

The above simulation was performed for substitution (Step 1) and for nonresponse weighting adjustments at the same cluster level (Steps 1–3). Within each simulation, $K = 5{,}000$ samples were generated and the corresponding estimates of the population mean were calculated, $\bar{y}_{SUB,k}$ or $\bar{y}_{r,k}^*$, $k = 1..K$. The mean and the variance were calculated as $\hat{\bar{Y}}_{SUB} = 1/K \sum_{k=1}^{K} \bar{y}_{SUB,k}$ and $\widehat{\mathrm{Var}}(\hat{\bar{Y}}_{SUB}) = 1/K \sum_{k=1}^{K} (\bar{y}_{SUB,k} - \hat{\bar{Y}}_{SUB})^2$, and similarly for the estimate $\bar{y}_r^*$. To control the stability of the results, three simulations were performed.

Table 3. Sampling variance for the design 140×15, $\bar{R} = 0.83$

Procedure	Simulations		
	1	2	3
$\widehat{\text{Var}}(\bar{y}_{SUB})$	84.6	85.2	84.1
$\widehat{\text{Var}}(\bar{y}_r^*)$	86.6	87.1	86.2
$\widehat{\text{VIF}}$	1.02	1.02	1.02

The sampling variance calculated analytically from (6) equals $\text{Var}(\bar{y}_{SUB}) = 84.8$ for the design $a = 140$, $b = 15$, $\bar{R} = 0.83$. From Table 3 we can observe the average increase (geometric mean) across the simulations $\widehat{\overline{\text{VIF}}} = 1.02$ which estimates the ratio (13). We can also extrapolate from Table 1 ($\bar{M} = 0.17$, $b^* = 18$) and Table 2 ($\rho = 0.07$, $b = 15$) the analytical calculation $\text{VIF} \doteq 1 + 1.2/100 * 0.49 = 1.01$.

There is a slight underestimation in the corresponding analytical expressions. The reason for this is a positive correlation among the factors $(1 - b_{ri}^*/B_{ri})$, $1/b_{ri}^*$ and V_i in the last term of (10). They all simultaneously increase in clusters with large nonresponse. However, we will not go into details of this effect here, since it does not change the key findings related to the gains in precision when a substitution is used.

With small clusters and a large nonresponse ($\bar{R} = 0.6, b = 3$), the differences between the procedures increase. The analytical result for the variance (6) gives $\text{Var}(\bar{y}_{SUB}) = 48.7$. The analytical approximation for the increase VIF is obtained by multiplying the corresponding cells in Table 1 ($\bar{M} = 0.4$, $b^* = 5$) and Table 2 ($\bar{M} = 0.4$, $b = 3$). From (13) we thus have $\text{VIF} = (1 + 18/100 * 0.82 \doteq 1.15)$. A similar result is obtained also with simulations $\widehat{\overline{\text{VIF}}} = 1.18$ (Table 4). We should repeat that this is an extreme design, where the underestimation of analytical result also reaches its extreme value.

The above simulations confirm that special situations with gains in precision close to 20% do exist, but in general the benefits tend to be small.

We should note that the above-discussed increase in variance due to weighting, substantially differs from the increase arising from oversampling strata (Kish 1965, p. 429) which is often used in survey practice.

2.3. *Empirical evaluation*

The above findings were verified (Vehovar 1995) by six recent Slovenian GSS surveys ($n = 2{,}100$). This is a face-to-face survey with 140 primary sampling units and 420 secondary sampling units, a completion rate of 20% and a nonresponse rate of 12% (Štebe 1995). Thirty target variables were selected from each survey, and the variances were compared (weighting versus substitution). The effect of the different sample sizes

Table 4. Sampling variance for the design n = 700×3, $\bar{R} = 0.60$

Procedure	Simulations		
	1	2	3
$\widehat{\text{Var}}(\bar{y}_{SUB})$	48.8	48.3	49.3
$\widehat{\text{Var}}(\bar{y}_r^*)$	57.9	57.5	58.1
$\widehat{\text{VIF}}$	1.19	1.19	1.18

was carefully removed using design effect and intracluster correlation. With clusters of $b = 15$, the median estimate of the increase VIF was 1.02 (2%) while with $b = 5$ the median was 10%. As with simulations, the increase was slightly above the theoretical results, and considerable variations around the median were observed.

The study of the mean squared error for these variables showed that with clusters of $b = 15$ the substitution has a clear disadvantage compared to the weighting adjustment, but with clusters of $b = 5$ there was a slight advantage for substitution. The differences were nevertheless small, so the bias-variance considerations alone could not justify either of the options.

The GSS survey is the only remaining survey among official and academic surveys in Slovenia that still uses substitution. One reason for this is tradition, as this same design has been used for more than 30 years, the other is the ease and the widespread use of such weight-free data. Many different users have extensively used the data and created their own series with simple tabulation packages.

It should be added that a major revision of the GSS sample design is planned for the year 1999, which also includes the abandoning of substitution. The important factor for this decision was the growing prolongation of the data collection period, which is in conflict with the increasing demand for a prompt release of results.

2.4. Generalisation of the results

The above results can be extended to a multistage cluster design with more than two stages. However, in such an event the potential gains in precision are even smaller. The substitution is usually performed within the last-stage clusters, consequently the corresponding component of the within cluster variability will represent an even smaller portion of the entire sampling variance. On the other hand, if the substitution is performed at the level of primary sampling units (clusters) the corresponding cluster size b will be far too large to allow for any gains in precision.

The results can be equally applied also to the case where the substitution is performed within the strata. There, we have only within stratum variance, so the increase VIF_W instead of VIF can be used. Thus, the gains from Table 1 apply directly, and they are similar to the advantage in precision of a proportionate stratified sample over a sample with post-stratification (Cochran 1977, p. 135). However, the strata are often relatively large, so the corresponding benefits of substitution tend to be small.

We can also use alternative substitution cells, for example some types of socio-demographic classes, and perform substitution within these adjustment cells. In household surveys, a substitute selection of the household of the same size as the nonresponding household is particularly typical. The comparisons of substitution with alternative procedures (population weights, sample weights, imputation) in these adjustment cells yield complex expressions, but the results are very close to those obtained with stratification.

When comparing the precision of one estimate with those of others we omitted the discussion of the imputations. We should repeat that with respect to precision, the estimates based on imputations – at least with the examples we discussed above – produce larger estimates for sampling variance than the corresponding weighting procedure.

2.5. *The adjustment cells and the substitution cells*

In practice, the substitution is usually performed within the clusters, and very rarely within strata or within other types of adjustment cells. In fact, in all reported research the substitution was applied at the level of the last-stage clusters. On the other hand, specific adjustment cells are often created for other types of nonresponse corrections (weighting, imputations). These adjustment cells typically differ from the clusters where the substitution is performed, and these cells generally crosscut the clusters. The relation between the two adjustments is very complicated due to the complexity of interaction between the two procedures.

However, the substitution procedure can be directly compared only with the methods that compensate for the distorted (due to nonresponse) inclusion probabilities at the same cluster (strata, cell) level. Basically, there are only two alternative methods for this: the weighting based on proper inclusion probabilities and the corresponding imputations (single or multiple). Other types of nonresponse corrections do not compare (or compete) with substitution, since we apply them regardless of the adjustments at the cluster level. The corrections in large adjustment cells are thus performed (or not performed) independently from the adjustments at the cluster level. These are complementary and consecutive (successive) steps for handling nonresponse, and are not alternatives.

Sometimes in the nonresponse adjustment process we even omit the compensation for missing data at the level where substitution is otherwise performed. In these situations the practice of substitution would be clearly redundant. Such an omission occurs automatically when we assume a simple random sample (even if the sample is complex). This is often the case when large socio-demographic classes are constructed as adjustment cells, or when we use models for handling nonresponse. Here, we can only repeat that, in the case of a simple random sampling, substitution has a clear disadvantage in the bias and no advantage in precision.

Leaving practical considerations aside, the only gain that substitution can provide is the improved precision arising from advantages over the alternative procedures that compensate at the same cluster level. The nonresponse corrections (weighting, imputation) based on adjustment cells that differ from the clusters where substitution was performed have little impact on the evaluation of this practice.

3. Practical Considerations

The practical aspects are based on extensive research and experiments with Slovenian national surveys – Labour Force Survey, Family Expenditure Survey, General Social Survey (GSS), Crime Victimisation Survey – during the years 1990–1996. Except for GSS, the substitution procedure has already been abandoned in all of the above surveys.

3.1. *Advantages*

a) Simplicity for the users. Substitution preserves the property of the self-weighted sample, and the merits of such samples are well-known.

b) Sample size controls. This is a minor advantage since there also exist variations in sample size arising from non-eligible units that cannot be controlled by substitution.

Furthermore, as for the interviewers' workload, the nonresponse itself creates certain differences in the number of visits assigned to each interviewer. Also, there are severe practical difficulties when it comes to obtaining an exact sample size by means of substitution, even for telephone surveys. And finally, the *supplement* sample (Kish 1965, p. 415) can provide the same control with fewer complications. The supplement sample differs from the substitution procedure in the sense that it is, in fact, a new sample. The pre-selected units are given to the interviewers, usually once in a survey process, when the management decides that additional units are needed. The interviewers do not participate in the decision whether to include an additional unit in the sample, as is the case with substitution.

c) Removal of the nonresponse bias. Compared to the situation without any nonresponse adjustments, an important improvement exists with substitution. For example, the urban-rural component of the bias is generally removed when substitution is performed at the cluster level. Of course, such a result can also be obtained with sample weighting adjustments, and with far less effort. However, the nonresponse bias within the level at which the substitution is performed generally remains unaffected.

In principle, we can select a substitute unit which is *similar* to the nonresponding one. The shortage of available covariates, together with their weak correlation with the nonresponse characteristics and with the target variables, makes the actual improvements much less successful. When combined with practical inconveniences in the selection of similar substitutes, the discouraging results reported in the empirical study (Vehovar 1994, p. 175) are not surprising. There, the substitute units were selected in the 1990 Slovenian Labour Force Survey ($n = 3{,}900$) together with initial units, from the register of population, where the variables location, sex and age group were available. Within each cluster the potential substitute units were pre-selected according to the anticipated nonresponse rate within population clusters and within age/sex demographic groups. However, no gains in the nonresponse bias were found in comparison to the alternative weighting adjustment at the same cluster level. There, the weighting additionally included the corrections based on the same age/sex information that was used for the selection of the substitutes.

There is no evidence of any other empirical research that would properly demonstrate the benefits of such an approach. However, as is often the case with substitution, the assumption of such an advantage does exist in survey practice.

d) Optimal structure of the sample. The substitution procedure provides the prescribed number of observations for each part of the sample. This is, of course, irrelevant in the case of simple random sampling, but in complex designs it becomes an issue, especially when small strata or small clusters are employed. In addition to the potential gains in precision, the avoidance of empty observations also preserves the basic features of the sampling design. For example, when selecting two units (schools or clusters) per stratum, it makes possible the calculation of the variance. However, to evaluate the benefits of substitution in such situations, the mean squared errors should be compared with the alternative procedures of collapsing clusters or strata.

The first three advantages above are not of great significance, although situations may occur where they can be beneficial. The issue of the optimal structure of the sample thus represents the key potential advantage of this procedure.

3.2. Disadvantages

a) Field work controls. This is the major deficiency of substitution, especially when area frames are used in face-to-face surveys. With telephone interviewing, however, the bulk of the problem disappears.

b) The illusion that nonresponse has been removed. The illusion that substitution has solved the nonresponse problem can be extremely strong, and the effort to deal with the nonresponse problem may be reduced.

c) Higher nonresponse rate. An interviewer's effort decreases if he or she knows that difficult-to-contact units can be declared non-interviews and then replaced by substitute units. A split-half experiment in the area sample for the 1991 Labour Force Survey ($n = 4{,}000$) showed an increase in the nonresponse rate from 9% to 10% (Vehovar 1994, p. 179).

d) Prolongation of the field work. According to the sequential nature of the substitution, the field work is substantially prolonged. Conservatively speaking, each wave can be treated as a separate survey with the same number of prescribed attempts (visits, calls). The field operation time is thus at least doubled. As a consequence, the administrative and the overhead costs also increase.

3.3. Practical guidelines

1. Field substitution is *not* an appropriate procedure for large probability samples where at least one of the following features is valid:
   - there is a short time available for field operations,
   - there is evidence of a strong net substitution bias,
   - there is weak (or expensive) control over field work procedures.
2. The following practical reasons *may* justify the use of substitution:
   - the need for a self-weighted sample is extremely strong; however, in this case the following conditions *must* additionally hold for the substitution procedure if it is to have any possible advantage:
     - there are no other theoretical reasons for weighting,
     - the substitution can remove the nonresponse bias, at least to the extent of the alternative procedures;
   - there is a danger that, due to nonresponse, a considerable number of clusters (or strata) would have no observations;
   - there exists a potential advantage of improved precision; however, this *must* be observed within the framework of the mean squared error.

4. Conclusions

The practice of substitution can be conditionally justified only in surveys with high nonresponse rates and a small take per cluster (or stratum). We encounter this situation

in surveys of institutions (stores, schools), and in specific household surveys, such as a Family Expenditure Survey, or media surveys where an exact number of units have to be surveyed each day. However, even with these surveys the alternative procedures of supplements or the increased initial sample (according to the anticipated nonresponse) often provide better results. There is, in fact, no empirical evidence that the alternatives would perform worse than substitution.

In surveys of up to a few thousand units with proper field work controls, the bias-variance issues become unimportant in most situations for the evaluation of the substitution. The prolongation of the data collection period remains – together with the inconveniences in the field work controls – the key disadvantage of this procedure. In large samples an additional drawback may be the small but consistent net substitution bias which will cancel out the negligible gains in precision.

The advantage of improved precision is generally small for the substitution procedure. However, with small clusters, high nonresponse rates and small intracluster correlation the gains may become considerable. When combined with sufficient field work controls substitution is potentially advantageous. A specific example of this is the Waksberg-Mitofsky two-stage procedure for telephone surveys (Waksberg 1978). The procedure is sequential and formally equal to substitution. The only difference is the fact that we are looking for a fixed number of eligible units instead of a fixed number of responding units. It has been shown that in certain circumstances this procedure has advantages over the modified one (Brick and Waksberg 1991), which is based on the weighting adjustments.

Here, we compared substitution and the alternative methods which compensate for nonresponse within the same clusters (or strata). Other adjustments, such as population weighting, may follow in all cases. Of course, if there is no need for adjustments at the cluster level, or if models are used to compensate for the nonresponse, as in Brehm (1993) or Little and Rubin (1987), substitution is simply redundant.

5. References

Biemer, P., Chapman, D., and Alexander, C. (1990). Some Research Issues in Random-Digit-Dialing Sampling and Estimation. 1990 Annual Research Conference. Washington: U.S. Bureau of the Census, 71–93.

Brehm, J. (1993). The Phantom Respondents. Ann Arbor: The University of Michigan Press.

Brick, M. and Waksberg, J. (1992). Avoiding Sequential Sampling with Random Digit Dialing. Survey Methodology, 17, 31–47.

Chapman, D. (1983). The Impact of Substitutions on Survey Estimates. In Incomplete Data in Sample Surveys, Vol. II, Theory and Bibliographies, eds. W. Madow, I. Olkin, and D. Rubin, New York: National Academy of Sciences, Academic Press, 45–61.

Chapman, D. and Roman, A. (1985). An Investigation of Substitution for an RDD Survey. Proceedings of the Survey Research Methods Section, American Statistical Association, 269–274.

Cochran, W. (1977). Sampling Techniques. New York: Wiley.

Forsman, G. and Berg, S. (1992). Telephone Interviewing and Data Quality, an Overview and Empirical Study. Linköping: Linköping University.

Groves, R. (1989). Survey Errors and Survey Costs. New York: Wiley.

Kalton, G. (1983). Compensating for Missing Survey Data. Ann Arbor: Institute for Social Research.

Kish, L. (1965). Survey Sampling. New York: Wiley.

Lessler, J. and Kalsbeek, W. (1991). Non-sampling Errors in Surveys. New York: Wiley.

Little, R. and Rubin, D. (1987). Statistical Analysis with Missing Data. New York: Wiley.

Marliani, G. and Pacei, S. (1993). Effects of Household Substitution in the Italian Consumer Expenditure Survey. Bulletin of the ISI, Contributed Paper, 49th ISI Session, Firenze 1993, 149–150.

Nathan, G. (1980). Substitution for Non-response as a Means to Control Sample Size. Sankhyā, C42, 1–2, 50–55.

Särndal, C., Swensson, B., and Wretman, J. (1992). Model-assisted Survey Sampling. New York: Springer-Verlag.

Štebe, J. (1995). Non-response in the Slovene Public Opinion Survey. Contributions to Methodology and Statistics, eds. A. Ferligoj and A. Kramberger, Ljubljana: Faculty of Social Sciences, 21–37.

Vehovar, V. (1993). The Field Substitution in the Slovene Labour Force Survey. Bulletin of the ISI, Contributed Paper, 49th ISI Session, Firenze 1993, 519–520.

Vehovar, V. (1994). The Field Substitution in Sample Surveys. Doctoral Dissertation (In Slovenian language with a summary in English). Ljubljana: University of Ljubljana.

Vehovar, V. (1995). The Field Substitution in the Slovene Public Opinion Survey. Contributions to Methodology and Statistics, eds. A. Ferligoj and A. Kramberger, Ljubljana: Faculty of Social Sciences, 38–66.

Verma, V. (1992). Household Surveys in Europe: Some Issues in Comparative Methodologies. Paper presented at the Seminar: International Comparisons of Survey Methodologies, Eurostat, Athens, April 1992.

Verma, V. and Gabilondo, L. (1993). Family Budget Surveys in the EC: Methodology and Recommendations for Harmonization. Eurostat, Statistical Document 3 E. Luxembourg: Eurostat.

Waksberg, J. (1978). Sampling Methods for Random Digit Dialing. Journal of the American Statistical Association, 73, 40–46.

Received May 1997
Revised February 1999