A simple variance estimator of change for rotating repeated surveys: an application to the EU-SILC household surveys

Y.G. Berger and R. Priam

University of Southampton, UK

Summary. A common problem is to compare two cross-sectional estimates for the same study variable taken on two different waves or occasions, and to judge whether the observed change is statistically significant. This involves the estimation of the sampling variance of the estimator of change. The estimation of this variance would be relatively straightforward if cross-sectional estimates were based upon the same sample. Unfortunately, samples are not completely overlapping, because of rotations used in repeated surveys. We propose a simple approach based upon a multivariate (general) linear regression model. The proposed variance estimator is not a model-based estimator. We show that the proposed estimator is design-consistent when the sampling fractions are negligible. It can accommodate stratified and two-stage sampling designs. The main advantage of the proposed approach is its simplicity and flexibility. It can be applied to a wide class of sampling designs, and can be implemented with standard statistical regression techniques. Because of its flexibility, the proposed approach is well suited for the estimation of variance for the EU-SILC surveys (e.g. Di Meglio *et al.*, 2013). It allows to use a common approach for variance estimation for the different types of designs. The proposed approach is a useful tool, because it only involves modelling skills and requires a limited knowledge of survey sampling theory.

Keywords: Design-based approach, Linearisation, Multivariate regression, Stratification, Two-stage sampling, Unequal inclusion probabilities.

1. Introduction

Measuring changes over time is a central problem for many users of social, economic and demographic data and is of interest in many areas of economics and social sciences. For example, the European Union Statistics on Income and Living Conditions (EU-SILC) surveys are used to monitor change in poverty within the European Union (Eurostat, 2012a). Smith *et al.* (2003) recognised that assessing change is one of the most important challenges in survey statistics. Suppose we have two partially overlapping samples s_1 and s_2; where s_1 and s_2 denote respectively the samples from the first and second wave (or first and second time period). In this paper, s denotes the union of s_1 and s_2; that is, $s = s_1 \cup s_2$.

The primary interest of many users is often in changes or trends from one time period to another. We start by considering changes between totals. In § 4.3, we extend the proposed approach to more complex measures of change. Suppose, we wish to estimate the following change

$$\Delta = \tau_2 - \tau_1,$$

Address for correspondence: Y.G. Berger, University of Southampton, Southampton, SO17 1BJ, UK. E-mail: y.g.berger@soton.ac.uk

between two population totals $\tau_1 = \sum_{i \in U} y_{1;i}$ and $\tau_2 = \sum_{i \in U} y_{2;i}$, of wave 1 and 2; where U denotes the population of interest. The quantities $y_{1;i}$ and $y_{2;i}$ denote respectively the values of variables of interest at wave 1 and 2. For simplicity, we assume that U is the same at both waves. The estimator proposed in this paper can also be used when the population at wave 1 is different from the population at wave 2. We adopt a design-based approach where the sampling distribution is specified by the sampling design. The change Δ can be estimated by

$$\widehat{\Delta} = \widehat{\tau}_2 - \widehat{\tau}_1;$$

where $\widehat{\tau}_1$ and $\widehat{\tau}_2$ are two cross-sectional Horvitz and Thompson (1952) estimators given by

$$\widehat{\tau}_1 = \sum_{i \in s_1} \frac{y_{1;i}}{\pi_{1;i}} \quad \text{and} \quad \widehat{\tau}_2 = \sum_{i \in s_2} \frac{y_{2;i}}{\pi_{2;i}} \; ; \tag{1}$$

The quantities $\pi_{1;i}$ and $\pi_{2;i}$ are the first-order inclusion probabilities at wave 1 and 2. These probabilities are defined in §2. The design-based variance of the change $\widehat{\Delta}$ is given by

$$\begin{aligned} var(\widehat{\Delta}) &= var(\widehat{\tau}_1) + var(\widehat{\tau}_2) - 2\, cov(\widehat{\tau}_1, \widehat{\tau}_2) && (2) \\ &= var(\widehat{\tau}_1) + var(\widehat{\tau}_2) - 2\, [var(\widehat{\tau}_1) var(\widehat{\tau}_2)]^{\frac{1}{2}}\ \rho \\ &= \boldsymbol{\nabla}^\top \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\tau}}} \boldsymbol{\nabla}; && (3) \end{aligned}$$

where $var(\widehat{\tau}_1)$ and $var(\widehat{\tau}_2)$ denote respectively the design-based variances of $\widehat{\tau}_1$ and $\widehat{\tau}_2$. The quantities $cov(\widehat{\tau}_1, \widehat{\tau}_2)$ and ρ denote respectively the covariance and the correlation between $\widehat{\tau}_1$ and $\widehat{\tau}_2$ with respect to the sampling design. The matrix $\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\tau}}}$ is the design-based covariance matrix of the vector $(\widehat{\tau}_1, \widehat{\tau}_2)^\top$ and $\boldsymbol{\nabla} = (-1,\ 1)^\top$.

Any standard design-based estimators can be used to estimate the variances $var(\widehat{\tau}_1)$ and $var(\widehat{\tau}_2)$, such as direct or re-sampling estimators. We focus our attention on the correlation ρ between $\widehat{\tau}_1$ and $\widehat{\tau}_2$, which are estimated from different overlapping samples. Several estimators have been proposed for the covariance in (2) (e.g. Kish, 1965; Tam, 1984; Nordberg, 2000; Holmes and Skinner, 2000; Berger, 2004; Qualité and Tillé, 2008; Wood, 2008; Goga *et al.*, 2009; Muennich and Zins, 2011; Knottnerus and van Delden, 2012). In a series of simulations based on the Swedish Labour Force Survey, Andersson *et al.* (2011a,b) showed that the estimator for the covariance proposed by Berger (2004) gives accurate estimates when we are interested in change within domains defined by the strata. In §3, we show that the estimator proposed in this paper and the estimator proposed by Berger (2004) are approximately equal, when the sampling fractions are small.

The main contribution of the paper is to show that the correlation can be calculated using the covariance of the residuals of a multivariate regression model with suitable interactions. Using this fact, the proposed approach can tackle a large class of parameters. Any statistical software can be used to compute the covariance matrix of the multivariate regression model. The multivariate regression is not a super-population approach, as it gives design-consistent covariance estimates (see Appendix A). However, it relies on the assumption that the sampling fractions are negligible, which is usually the case for social surveys, such as the EU-SILC surveys (Eurostat, 2012a). The proposed approach has the advantage of not requiring joint-inclusion probabilities which can be unknown with rotating designs.

With small sampling fractions, the covariance can be estimated by the following standard Hansen and Hurwitz (1943) 'type' estimator (e.g. Qualité, 2009, p. 83) based on the common sample $s_c = s_1 \cap s_2$.

$$\widehat{cov}(\widehat{\tau}_1, \widehat{\tau}_2)_{HH} = \frac{n_c}{n_c - 1} \sum_{i \in s_c} \left(\breve{y}_{i;1} - \bar{\breve{y}}_{1;c}\right) \left(\breve{y}_{i;2} - \bar{\breve{y}}_{2;c}\right) \; , \tag{4}$$

where $s_c = s_1 \cap s_2$ and

$$\breve{\bar{y}}_{\ell;c} = \frac{1}{n_c} \sum_{i \in s_c} \breve{y}_{i;\ell}.$$

The variables $\breve{y}_{1;i}$ and $\breve{y}_{2;i}$ are defined by

$$\breve{y}_{i;1} = y_{1;i}\, \pi_{1;i}^{-1} \delta\{i \in s_1\} \quad \text{and} \quad \breve{y}_{i;2} = y_{2;i}\, \pi_{2;i}^{-1} \delta\{i \in s_2\}\ ; \tag{5}$$

with $\breve{y}_{\ell;i} = 0$ when $i \notin s_\ell$. The function $\delta\{A\}$ is the indicator function which is equal to one when A is true and zero otherwise. The Hansen and Hurwitz (1943) 'type' estimator for the correlation is given by

$$\widehat{\rho}_{HH} = \widehat{cov}(\widehat{\tau}_1, \widehat{\tau}_2)_{HH}\ [\widehat{var}(\widehat{\tau}_1)\widehat{var}(\widehat{\tau}_2)]^{-\frac{1}{2}}\ ; \tag{6}$$

where $\widehat{var}(\widehat{\tau}_1)$ and $\widehat{var}(\widehat{\tau}_2)$ denote respectively any standard design-based variance estimators of $\widehat{\tau}_1$ and $\widehat{\tau}_2$. In § 5, we show that (6) produces a variance estimator for change which may be less accurate than the estimators we propose in § 3. Furthermore, resulting variance estimates for change can be negative, as $\widehat{\rho}_{HH}$ could be larger than one, because the covariance and the variances are not estimated from the same sample. Note that (4) is a covariance between $\breve{y}_{i;1}$ and $\breve{y}_{i;2}$. The paper elaborates from the principle that a covariance can be estimated from a linear model.

In §2, we define the class of rotating sampling designs considered in this paper. The proposed estimator for the covariance is defined in §3. Alternative expressions for the proposed estimators are given in § 3.1. In §4, we show how the proposed estimator can be extended to account for stratification, multi-stage sampling and more complex measures of change. In §5, we support our result with a simulation study based on the British Labour Force Survey data and on the Italian EU-SILC survey data. In §6, we show how the proposed approach can be used to estimate the variance of change of the EU-SILC at risk of poverty and social exclusion (AROPE) indicator.

2. Fixed size rotating sampling designs

With panel surveys, it is common practice to select new units in order to replace old units that have been in the survey for a specified number of waves (e.g. Gambino and Silva, 2009; Kalton, 2009; Eurostat, 2012a). The units sampled on wave 1 and on wave 2 usually represent a large fraction of the first wave sample s_1. This fraction is called the *fraction of the common sample* and is denoted by g. For example, for the EU-SILC surveys, $g = 75\%$. For the Canadian labour force survey and the British labour force survey, $g = 80\%$. For the Finish labour force survey, $g = 60\%$.

The class of fixed size rotating sampling designs is defined as follows. Assume that s_1 is a probability sample of size n_1 selected without replacement with first-order inclusion probabilities $\pi_{1;i} = pr\{i \in s_1\}$, where $pr\{\cdot\}$ denotes the probability with respect to the design. Suppose that s_2 is a sample of size n_2 selected with conditional inclusion probabilities $\pi_{2;i}(s_1) = pr\{i \in s_2 | s_1\}$ such that s_2 contains n_c units from s_1; where $0 \leq n_c \leq n_1$. The wave 2 inclusion probabilities are given by $\pi_{2;i} = E_1[\pi_{2;i}(s_1)]$; where $E_1[\,\cdot\,]$ denotes the design expectation with respect to the first wave design. Note that the fraction of the common sample is given by $g = n_c/n_1$. The units from $s_1 \setminus s_2$ are the units that rotate out and the units from $s_2 \setminus s_1$ are the units that rotate in. In principle, we can have $g = 0$ (when we have two non-overlapping samples) or $g = 1$ (when we have two completely overlapping samples). The proposed approach is valid when $g = 0$ or 1. When $g = 0$ the covariance equals zero. We consider that the sizes n_1, n_2 and n_c are given quantities which are fixed (non-random). The variance estimators, proposed in § 3, are consistent in this case. These estimators may not be suitable when the sizes are random.

This class contains standard rotating sampling designs such as the rotating randomised systematic sampling design (e.g. Holmes and Skinner, 2000), the rotation groups sampling design (e.g. Kalton, 2009; Gambino and Silva, 2009, p. 415) used for the EU-SILC surveys (Eurostat, 2012a) and the rotating design proposed by Tam (1984).

EXAMPLE 1. *Suppose that the first wave sample s_1 is selected without replacement with inclusion probabilities $\pi_{1;i}$, and that the second wave sample s_2 is a sample of n_c units selected without replacement from s_1 with probabilities proportional to p_i combined with a sample of $n_{2|c} = n_2 - n_c$ units selected without replacement from $U \setminus s_1$ with probabilities proportional to q_i; where p_i and q_i being known positive quantities; where $U \setminus s_1$ denotes the set of units not selected at wave 1. Tam (1984) studied this design when $\pi_{1;i} = n_1/N$ and $p_i = q_i = 1$. The following equation gives the wave 2 conditional first-order inclusion probabilities given s_1.*

$$\pi_{2;i}(s_1) = p_i(s_1) z_{1;i} + q_i(s_1)(1 - z_{1;i}) \; ;$$

where $z_{1;i} = \delta\{i \in s_1\}$, $p_i(s_1) = n_c \, p_i / (\sum_{i \in s_1} p_i)$ and $q_i(s_1) = n_{2|c} \, q_i / (\sum_{i \notin s_1} q_i)$. We assume that the p_i and q_i are such that $p_i(s_1) \leq 1$ and $q_i(s_1) \leq 1$. An approximation for the wave 2 first-order inclusion probabilities is given by (e.g. Christine and Rocher, 2012)

$$\pi_{2;i} \doteq E_1[p_i(s_1)]\pi_{1;i} + E_1[q_i(s_1)](1 - \pi_{1;i}) \; ;$$

where $E_1[p_i(s_1)] \doteq n_c \, p_i / [\sum_{i \in U} p_i \pi_{1;i}]$ and $E_1[q_i(s_1)] \doteq n_{2|c} \, q_i / [\sum_{i \in U} q_i (1 - \pi_{1;i})]$. We have that $\pi_{2;i} \doteq n_2 n_1^{-1} \pi_{1;i}$, when $p_i = 1$ and $q_i = \pi_{1;i}/(1 - \pi_{1;i})$. We also have that $\pi_{2;i} \doteq \pi_{1;i}$, when $p_i = 1/\pi_{1;i}$, $q_i = \{\pi_{1;i} - n_c/N\}/(1 - \pi_{1;i})$ and $n_1 = n_2$. Note that $\pi_{2;i} = n_2/N$ when $p_i = q_i = 1$, $\pi_{1;i} = n_1/N$.

3. Proposed estimator of the variance for change

The estimation of the correlation would be relatively straightforward if s_1 and s_2 were the same sample ($g = 1$). Unfortunately, s_1 and s_2 are usually not completely overlapping sets of units ($g < 1$), because rotations are usually used in repeated surveys (see § 2).

We propose to estimate the variance of change (2) by

$$\widehat{var}(\widehat{\Delta})^{(\cdot)} = \widehat{var}(\widehat{\tau}_1) + \widehat{var}(\widehat{\tau}_2) - 2 \, \widehat{\rho}^{(\cdot)}_{prop} \, [\widehat{var}(\widehat{\tau}_1) \, \widehat{var}(\widehat{\tau}_2)]^{\frac{1}{2}} \; ; \tag{7}$$

where $\widehat{var}(\widehat{\tau}_1)$ and $\widehat{var}(\widehat{\tau}_2)$ denote respectively any design-based variance estimator of $\widehat{\tau}_1$ and $\widehat{\tau}_2$. The quantity $\widehat{\rho}^{(\cdot)}_{prop}$ is an estimator for the correlation. In this §, we propose two estimators for the correlation: (11) and (12).

We propose to estimate the correlation from the covariance of the residuals of the following multivariate (or general) linear regression model (see also Berger and Priam, 2010).

$$\begin{pmatrix} \breve{y}_{1;i} \\ \breve{y}_{2;i} \end{pmatrix} = \begin{pmatrix} \beta_1^{(1)} z_{1;i} + \beta_2^{(1)} z_{2;i} + \beta_{12}^{(1)} z_{1;i} \, z_{2;i} \\ \beta_1^{(2)} z_{1;i} + \beta_2^{(2)} z_{2;i} + \beta_{12}^{(2)} z_{1;i} \, z_{2;i} \end{pmatrix} + \boldsymbol{\epsilon}_i \; ; \tag{8}$$

where $i \in s = s_1 \cup s_2$ and the residuals $\boldsymbol{\epsilon}_i$ have a bivariate distribution with mean $\mathbf{0}$ and an unknown variance-covariance matrix $\boldsymbol{V}$. The distribution of $\boldsymbol{\epsilon}_i$ does not need to be specified and is not used for inference, as a least squares technique (or a projection in the space spanned by the design variables) will be used. The response variables in the regression (8) are given by (5). The covariates $z_{1;i}$ and $z_{2;i}$ are design variables defined by

$$z_{1;i} = \delta\{i \in s_1\}, \qquad \text{and} \qquad z_{2;i} = \delta\{i \in s_2\} \cdot \tag{9}$$

Note the absence of intercept and the presence of an interaction in the regression (8). When we have completely overlapping samples ($g = 1$), we propose to remove the interaction $z_{1;i}\ z_{2;i}$ and the covariate $z_{2;i}$ from the model (8), as $z_{1;i} = z_{2;i}$ for all $i \in s$ in this case.

Let $\widehat{\boldsymbol{V}}^{(A)}$ be the ordinary least squares estimate of the variance-covariance matrix $\boldsymbol{V}$. Let

$$\widehat{\boldsymbol{S}}^{(A)} = \alpha \widehat{\boldsymbol{V}}^{(A)}\ ; \tag{10}$$

where $\alpha = (n - r)$ is a constant scale factor, where $n = \#s$ is the number of units in the sample $s = s_1 \cup s_2$ and r is the number of linearly independent columns of $\boldsymbol{Z}_s$. In the Appendix A, we show that $\widehat{\boldsymbol{S}}^{(A)}$ is a design consistent estimator of $\boldsymbol{\Sigma}_{\widehat{\boldsymbol{T}}}$. Therefore, a consistent estimator for the correlation is given by

$$\widehat{\rho}_{prop}^{(A)} \quad = \quad \widehat{V}_{12}^{(A)} \left(\widehat{V}_{11}^{(A)}\ \widehat{V}_{22}^{(A)} \right)^{-\frac{1}{2}}; \tag{11}$$

where the quantity $\widehat{V}_{k\ell}^{(A)}$ is the component (k, ℓ) of the matrix $\widehat{\boldsymbol{V}}^{(A)}$.

For the second estimator, we propose to substitute $\breve{y}_{1;i}$ and $\breve{y}_{2;i}$ respectively by $\breve{y}_{1;i}^{(B)}$ and $\breve{y}_{2;i}^{(B)}$ into (8), where $\breve{y}_{1;i}^{(B)} = \breve{y}_{1;i}\ z_{2;i}$ and $\breve{y}_{2;i}^{(B)} = \breve{y}_{2;i}\ z_{1;i}$. The quantities $\breve{y}_{1;i}^{(B)}$ and $\breve{y}_{2;i}^{(B)}$ pick out the common sample elements, as $\breve{y}_{1;i}^{(B)} = 0$ and $\breve{y}_{2;i}^{(B)} = 0$ when $i \notin s_c$. Let $\widehat{\boldsymbol{V}}^{(B)}$ be the ordinary least squares estimate of $\boldsymbol{V}$. The second estimator for the correlation is given by

$$\widehat{\rho}_{prop}^{(B)} = \widehat{V}_{12}^{(B)} \left(\widehat{V}_{11}^{(B)}\ \widehat{V}_{22}^{(B)} \right)^{-\frac{1}{2}} g; \tag{12}$$

where the quantity $\widehat{V}_{k\ell}^{(B)}$ is the component (k, ℓ) of the matrix $\widehat{\boldsymbol{V}}^{(B)}$.

Note both variance estimators of change based on (11) and (12) are always positive, because $\widehat{\rho}_{prop}^{(A)} \leq 1$ and $\widehat{\rho}_{prop}^{(B)} \leq 1$.

We have the following fixed size constraints $\sum_{i \in s} z_{1;i} = n_1$, $\sum_{i \in s} z_{2;i} = n_2$ and $\sum_{i \in s} z_{1;i} z_{2;i} = n_c$, because only samples with these sample sizes can be selected. Thus, $\widehat{\boldsymbol{V}}^{(A)}$ and $\widehat{\boldsymbol{V}}^{(B)}$ are variance-covariance matrices conditional on these variables which have their totals fixed by design. Note that there is a clear analogy between Birch (1963)'s approach and the proposed conditioning approach. This regression includes interactions between the variable $z_{1;i}$ and $z_{2;i}$. These interactions capture the rotation of the sampling design which is represented by the constraint $\sum_{i \in s} z_{1;i} z_{2;i} = n_c$.

The proposed approach requires the creation of design variables (9) which are used as covariates. The interactions (in (8)) take the rotation of the design into account. The weighted variables of interest (5) measured at each wave are used as response variables. The proposed estimator (11) takes all the data into account, as it utilises the data of the units from the common sample and the units that rotate in and out. The estimator (12) only utilises the data from the common sample.

The proposed estimator is easy to implement because it does not rely on joint-inclusion probabilities. Furthermore, the proposed estimator is based on a multivariate regression approach which can be implemented with most statistical software, without the need of a specialised statistical package. The ordinary least squares estimate of the variance-covariance matrix can be easily calculated, as the multivariate regression (15) can be easily fitted by most statistical software. It is only necessary to create the variables $\breve{y}_{i;1}$, $\breve{y}_{i;2}$, $z_{1;i}$ and $z_{2;i}$. For example, the SAS procedure `REG` can be used to fit the multivariate regression. The multivariate regression can also be fitted using the `GLM Multivariate` procedure in SPSS. With Stata, the output `e(Sigma)` of the function `mvreg()` gives the variance-covariance matrix. With the statistical software R (R Development Core Team, 2014), the command `estVar(lm(formula=Y~-1+Z1*Z2))` gives the variance-covariance matrix; where `Z1` and `Z2` denote the $n \times 1$ vectors $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ (see (18)) and `Y` is the matrix $\breve{\boldsymbol{Y}}_s$ given by (16). Note that Berger

(2005) showed that $\widehat{var}(\widehat{\tau}_1)$ and $\widehat{var}(\widehat{\tau}_2)$ can also be calculated using a regression approach. Note that when we have non-overlapping samples ($g = 0$), the interaction term is always equal to zero, and therefore automatically removed by statistical software. With completely overlapping samples ($g = 1$), the interaction term and $z_{2;i}$ are also automatically removed.

The advantage of the proposed approach is the fact that (i) it gives an approximately unbiased estimator for the variance of change, (ii) it can be implemented with any standard statistical software (iii) and it can be easily generalised to function of totals (see § 4.3).

3.1. Alternative expressions for the proposed estimators

The proposed estimator for the variance of change (7) can be rewritten as

$$\widehat{var}(\widehat{\Delta})^{(\cdot)} = \boldsymbol{\nabla}^\top \widehat{\boldsymbol{\Sigma}}^{(\cdot)}_{\widehat{\boldsymbol{\tau}}} \boldsymbol{\nabla} \; ;$$

where $\boldsymbol{\nabla} = (-1,\ 1)^\top$ and $\widehat{\boldsymbol{\Sigma}}^{(\cdot)}_{\widehat{\boldsymbol{\tau}}}$ is the following 2×2 matrix

$$\widehat{\boldsymbol{\Sigma}}^{(\cdot)}_{\widehat{\boldsymbol{\tau}}} = \begin{pmatrix} \widehat{var}(\widehat{\tau}_1) & \widehat{cov}(\widehat{\tau}_1, \widehat{\tau}_2)^{(\cdot)}_{prop} \\ \widehat{cov}(\widehat{\tau}_1, \widehat{\tau}_2)^{(\cdot)}_{prop} & \widehat{var}(\widehat{\tau}_2) \end{pmatrix} ; \tag{13}$$

where

$$\widehat{cov}(\widehat{\tau}_1, \widehat{\tau}_2)^{(\cdot)}_{prop} \;=\; \widehat{\rho}^{(\cdot)}_{prop} \left[\widehat{var}(\widehat{\tau}_1)\ \widehat{var}(\widehat{\tau}_2)\right]^{\frac{1}{2}}, \tag{14}$$

with a correlation $\widehat{\rho}^{(\cdot)}_{prop}$ given by (11) or (12).

Matrix notations can be used to define the model (8) in a more convenient way. Let $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)})$ be the 3×2 matrix of parameters, where $\boldsymbol{\beta}^{(1)} = (\beta_1^{(1)}, \beta_2^{(1)}, \beta_{12}^{(1)})^\top$ and $\boldsymbol{\beta}^{(2)} = (\beta_1^{(2)}, \beta_2^{(2)}, \beta_{12}^{(2)})^\top$ are parameters of the model (8). The model (8) can be re-written as

$$\breve{\boldsymbol{Y}}_s = \boldsymbol{Z}_s \boldsymbol{\beta} + \boldsymbol{\epsilon} \; ; \tag{15}$$

where $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_n)^\top$. The quantities $\breve{\boldsymbol{Y}}_s$ and $\boldsymbol{Z}_s$ are respectively defined by

$$\breve{\boldsymbol{Y}}_s \;=\; (\breve{\boldsymbol{y}}_1, \breve{\boldsymbol{y}}_2) \; , \tag{16}$$

$$\boldsymbol{Z}_s \;=\; (\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_c) \; ; \tag{17}$$

where

$$\breve{\boldsymbol{y}}_\ell \;=\; (\breve{y}_{\ell;1}, \breve{y}_{\ell;2}, \cdots, \breve{y}_{\ell;n})^\top \; ,$$

$$\boldsymbol{z}_\ell \;=\; (z_{\ell;1}, z_{\ell;2}, \cdots, z_{\ell;n})^\top \; , \tag{18}$$

$$\boldsymbol{z}_c \;=\; (z_{1;1} z_{2;1}, z_{1;2} z_{2;2}, \cdots, z_{1;n} z_{2;n})^\top \; . \tag{19}$$

The model (15) is also a multivariate analysis of variance (MANOVA) model, as the covariates are all dummy variables. With completely overlapping samples ($g = 1$) and non-overlapping samples, we use $\boldsymbol{Z}_s = \boldsymbol{z}_1$ instead; that is, $\boldsymbol{z}_c$ and $\boldsymbol{z}_2$ are removed from $\boldsymbol{Z}_s$.

Let $\widehat{S}_{12}^{(A)}$ be the extra-diagonal element of $\widehat{\boldsymbol{S}}^{(A)}$ and $\widehat{S}_{12}^{(B)}$ be the extra diagonal element of $\widehat{\boldsymbol{S}}^{(B)} = \alpha \widehat{\boldsymbol{V}}^{(B)}$. In Appendix B, we show that

$$\widehat{S}_{12}^{(A)} = \widehat{S}_{12}^{(B)} = \frac{n_c - 1}{n_c} \widehat{cov}(\widehat{\tau}_1, \widehat{\tau}_2)_{HH} \; , \tag{20}$$

which is approximately the Hansen and Hurwitz (1943) type estimator (4), when n_c is sufficiently large. When the first-order inclusion probabilities are equal ($\pi_{1;i} = n_1/N$ and $\pi_{2;i} = n_2/N$), the

estimators $\widehat{S}_{12}^{(A)}$ and $\widehat{S}_{12}^{(B)}$ reduce to

$$\widehat{S}_{12}^{(A)} = \widehat{S}_{12}^{(B)} = \frac{N^2 n_c}{n_1 n_2}\widehat{\sigma}_c\,, \tag{21}$$

where $\widehat{\sigma}_c$ denotes the following covariance between the variables $y_{1;i}$ and $y_{2;i}$ calculated from the sample s_c.

$$\widehat{\sigma}_c = \frac{1}{n_c}\sum_{i\in s_c}\left(y_{1;i} - \bar{y}_{1;c}\right)\left(y_{2;i} - \bar{y}_{2;c}\right), \tag{22}$$

where $\bar{y}_{\ell;c} = n_c^{-1}\sum_{i\in s_c} y_{\ell;i}$.

The estimator of the covariance proposed by Tam (1984) (see also Qualité and Tillé (2008)) is given by

$$\widehat{cov}(\widehat{\tau}_1, \widehat{\tau}_2)_{Tam} = \left(1 - \frac{n_1 n_2}{N n_c}\right)\frac{n_c}{n_c - 1}\,\frac{N^2 n_c}{n_1 n_2}\widehat{\sigma}_c. \tag{23}$$

Note that $\widehat{S}_{12}^{(A)}$ and $\widehat{S}_{12}^{(B)}$ reduces to the estimator proposed by Tam (1984) when they meet all of the conditions of equal probabilities, the fraction $n_1 n_2/N n_c$ is negligible and n_c is sufficiently large (see (21) and (23)).

Note that the proposed estimators for the covariance given by (14) are different from $\widehat{S}_{12}^{(A)}$, $\widehat{S}_{12}^{(B)}$ and $\widehat{cov}(\widehat{\tau}_1, \widehat{\tau}_2)_{HH}$, because the variances used in the proposed correlations $\widehat{\rho}_{prop}^{(\cdot)}$ are not necessarily equal to $\widehat{var}(\widehat{\tau}_1)$ and $\widehat{var}(\widehat{\tau}_2)$ in (6). The Hansen-Hurwitz (HH)-type estimator for the correlation $\widehat{\rho}_{HH}$ in (6) is different from $\widehat{\rho}_{prop}^{(A)}$ in (11) and $\widehat{\rho}_{prop}^{(B)}$ in (12). Hence, the approach considered in this paper is not a reformulation of the Hansen and Hurwitz (1943) estimator (4). In Appendix B, we show that (12) is approximately equal to the estimator for the correlation proposed by Qualité (2009, p. 83) and defined in (37).

3.2. Design consistency

In Appendix A, we show that $\widehat{\boldsymbol{S}}^{(A)}$ is a design consistent estimator of $\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\tau}}}$ under a high entropy without sampling design and under the following conditions.

$$\max_{i\in s}\pi_{\ell;i} = o_p(1), \quad \text{for } \ell = 1,\ 2, \tag{24}$$

$$\max_{i\in s}\pi_{c;i} = o_p(1), \tag{25}$$

$$\max_{i\in s}\frac{\pi_{1;i}\pi_{2;i}}{\pi_{c;i}} = o_p(1), \text{ when } g \neq 0; \tag{26}$$

where $\pi_{c;i} = pr\{i \in s_c\}$. Thus, $\widehat{\rho}_{prop}^{(A)}$ is a consistent estimator for the correlation, because it is a smooth function of consistent estimators. This implies that (7) is a consistent estimator for the variance of change, as long as $\widehat{var}(\widehat{\tau}_1)$ and $\widehat{var}(\widehat{\tau}_2)$ are design-consistent.

Most sampling designs used in practice have large entropy. There are designs with low entropy, such as the non-randomized systematic sampling design. Under this sampling design, it is not possible to obtain an unbiased estimator for the variance.

The assumptions (24)-(26) hold when the sampling fraction is negligible. Note that g can be large even when the sampling fractions are negligible; that is, the assumptions (24)-(26) may hold even when g is large or when $g = 1$. With the rotating design described in Example 1 of §2, we have that $\pi_{c;i} = g\pi_{1;i}$ when $p_i = 1$ and $q_i = \pi_{1;i}/(1 - \pi_{1;i})$. Hence the assumptions (24)-(26) hold when $\pi_{\ell;i}$, $g\pi_{1;i}$ and $\pi_{2;i}/g$ are negligible. Note that with non-overlapping samples ($g = 0$), the condition

(26) does not need to hold and the covariance equals zero. The condition (26) may not hold when g tends to zero more quickly than $\pi_{2;i}$. This is a situation rarely found in practice. Furthermore, in this situation the covariance $cov(\widehat{\tau}_1, \widehat{\tau}_2)$ is negligible and it is reasonable to consider that $cov(\widehat{\tau}_1, \widehat{\tau}_2) = 0$.

4. Extensions

4.1. Stratified sampling design

The proposed estimator can be easily extended to accommodate stratification. Suppose that we have H strata $U_1, U_2, \cdots, U_H$ such that $\cup_{h=1}^{H} U_h = U$. Let s_{1h} and s_{2h} denote respectively the samples of U_h for wave 1 and 2. Let n_{1h}, n_{2h} and n_{ch} be respectively the sample sizes of s_{1h}, s_{2h} and $s_{ch} = s_{1h} \cap s_{2h}$. Suppose that a fixed size rotating design (see §2) is implemented within each stratum. We have the following covariates $z_{1h;i} = \delta\{i \in s_{1h}\}$ and $z_{2h;i} = \delta\{i \in s_{2h}\}$ which specify in which stratum the unit i belongs.

The multivariate regression model is still given by (15) with the same response variables $\breve{\boldsymbol{Y}}_s$ defined by (16), where $\breve{y}_{i;\ell}$ is defined by (5). However, the matrix $\boldsymbol{Z}_s$ is different and contains the stratification variables $z_{1h;i}$ and $z_{2h;i}$ and the interactions $z_{1h;i} \times z_{2h;i}$. As we have a rotation within each stratum, the sample sizes $n_{ch} = \#s_{ch}$ are fixed and we need to include the interactions $z_{1h;i} \times z_{2h;i}$ in $\boldsymbol{Z}_s$. The ordinary least squares estimate of the variance-covariance of the residuals of model (15) is used to estimate the correlation between $\widehat{\tau}_1$ and $\widehat{\tau}_2$.

With the statistical software R (R Development Core Team, 2014), the command `estVar(lm(formula=Y~as.factor(Str.1)*as.factor(Str.2)))` gives the matrix $\widehat{\boldsymbol{V}}^{(A)}$; where `Str.1` and `Str.2` denote the $n \times 1$ vectors of strata labels for wave 1 and 2. The object `Y` is the matrix $\breve{\boldsymbol{Y}}_s$.

In Appendix C, we show that $\widehat{S}_{12}^{(A)} = \sum_{h=1}^{H} \widehat{S}_{12h}^{(A)}$ and $\widehat{S}_{\ell\ell}^{(A)} = \sum_{h=1}^{H} \widehat{S}_{\ell\ell h}^{(A)}$ where $\widehat{S}_{12h}^{(A)}$ (resp. $\widehat{S}_{\ell\ell h}^{(A)}$) denotes the within stratum covariance (resp. variance). Note that $\widehat{S}_{12}^{(A)}$ and $\widehat{S}_{\ell\ell}^{(A)}$ are natural estimators of covariance and variances under stratified designs. Consequently, the proposed estimator for the covariance is consistent when the assumptions (24)-(26) hold within each stratum and when the number of strata H is asymptotically bounded. This excludes heavily stratified designs.

The same result can be obtained for $\widehat{S}_{12}^{(B)}$ and $\widehat{S}_{\ell\ell}^{(B)}$ when we use the response variables $\breve{y}_{1;i}^{(B)}$ and $\breve{y}_{2;i}^{(B)}$ (see(12)). This gives a consistent estimator when $n_{ch} n_{1h}^{-1} = n_{ch'} n_{1h'}^{-1}$ for all $h \neq h'$.

4.2. Two-stage sampling design

Suppose that we have overlapping stratified samples of primary sampling units (PSU), and that the rotation consists in rotating PSUs rather than secondary sampling units. We suggest using an ultimate cluster strategy (Hansen *et al.*, 1953) to estimate the covariance; because the sampling fraction is assumed negligible. This usually holds for social surveys. This is the approach used in § 6.

The two-stage Horvitz and Thompson (1952) estimators $\widehat{\tau}_1$ and $\widehat{\tau}_2$ are now given by

$$\widehat{\tau}_1 = \sum_{i \in s_1} \frac{\widehat{\tau}_{1;i}}{\pi_{1;i}} \quad \text{and} \quad \widehat{\tau}_2 = \sum_{i \in s_2} \frac{\widehat{\tau}_{2;i}}{\pi_{2;i}} \; ; \tag{27}$$

where s_1 and s_2 denote the first and the second wave sample of PSUs. The quantities $\pi_{1;i}$ and $\pi_{2;i}$ are the first-order inclusion probabilities of the i-th PSU for the first and the second wave. The quantities $\widehat{\tau}_{1;i}$ and $\widehat{\tau}_{2;i}$ denote the Horvitz and Thompson (1952) totals of the i-th PSU, for the first and the second wave.

Now, $\breve{\boldsymbol{Y}}_s$ contains the following response variables,

$$\breve{y}_{1;i} = \frac{\widehat{\tau}_{1;i}}{\pi_{1;i}}\delta\{i \in s_1\} \quad \text{and} \quad \breve{y}_{2;i} = \frac{\widehat{\tau}_{2;i}}{\pi_{2;i}}\delta\{i \in s_2\}\ ; \tag{28}$$

with $\breve{y}_{\ell;i} = 0$ when $i \notin s_\ell$. Let $z_{1h;i}$ and $z_{2h;i}$ be the variables that specify the stratification of the PSUs; that is, $z_{\ell h;i} = 1$ if the i-th PSU is selected in stratum U_h at wave $\ell = 1, 2$. The covariates $z_{1h;i}$ and $z_{2h;i}$ and then interactions $z_{1h;i} \times z_{2h;i}$ are included in the matrix $\boldsymbol{Z}_s$. The ordinary least square covariance matrix of the residuals can be used to estimate the covariance (see (7)). The second estimator (see (12)) can also be used by multiplying $\widehat{\tau}_{1;i}$ and $\widehat{\tau}_{2;i}$ respectively by $z_{2;i}$ and $z_{1;i}$.

If we have a rotation within PSUs, the proposed approach can still be used. In this case, $g = 1$ and the interactions and $z_{2;i}$ have to be removed from the model. This gives classical estimates of covariances, as the samples of PSUs are the same at wave 1 and 2.

4.3. Complex measures of change

Suppose that we are interested in the variance of the change $\widehat{\Delta}_\theta = \widehat{\theta}_2 - \widehat{\theta}_1$ or the relative change $\widehat{\Delta}_\theta = \widehat{\theta}_2/\widehat{\theta}_1$, where $\widehat{\theta}_2$, $\widehat{\theta}_1$ are two smooth (differentiable) functions of estimators of totals. Therefore, in both cases, $\widehat{\Delta}_\theta$ is a smooth function of totals; that is,

$$\widehat{\Delta}_\theta = f(\widehat{\boldsymbol{\tau}}); \tag{29}$$

where $\widehat{\boldsymbol{\tau}} = (\widehat{\tau}_1, \widehat{\tau}_2 \cdots, \widehat{\tau}_p, \cdots, \widehat{\tau}_P)^\top$ and P is the number of totals. We consider that the first Q totals of $\widehat{\boldsymbol{\tau}}$ are based on s_1. The other totals are based on s_2. The quantity $\widehat{\tau}_p$ is the following Horvitz and Thompson (1952) estimator.

$$\widehat{\tau}_p = \sum_{i \in s_\ell} \breve{y}_{p;i},$$

where

$$\breve{y}_{p;i} = \frac{y_{p;i}}{\pi_{\ell;i}}\delta\{i \in s_\ell\}\ ; \tag{30}$$

where $\ell = 1$ if $p \leq Q$ and $\ell = 2$ if $p > Q$. Note that we consider a general setting, where $\widehat{\theta}_\ell$ could depend on totals computed from s_1 and s_2. This can be the case in practice. For example, the price and volume indices are function of totals across several waves (e.g. Wood, 2009).

Suppose that $\widehat{\Delta}_\theta$ is an approximately unbiased estimator of $\Delta_\theta = f(\boldsymbol{\tau})$; where $\boldsymbol{\tau} = E(\widehat{\boldsymbol{\tau}})$. Using the delta method (Taylor linearisation), we have that an approximation of $\widehat{\Delta}_\theta$ in the neighbourhood of $\boldsymbol{\tau}$ is given by $\widehat{\Delta}_\theta - \Delta_\theta \doteq \boldsymbol{\nabla}(\boldsymbol{\tau})^\top(\widehat{\boldsymbol{\tau}} - \boldsymbol{\tau})$; where $\boldsymbol{\nabla}(\boldsymbol{\tau})$ is the gradient of $f(\boldsymbol{\tau})$ at $\boldsymbol{\tau}$. Therefore, the proposed estimator for the variance is

$$\widehat{var}(\widehat{\Delta}_\theta) = \boldsymbol{\nabla}(\widehat{\boldsymbol{\tau}})^\top \widehat{\boldsymbol{\Sigma}}^{(A)}_{\widehat{\boldsymbol{\tau}}} \boldsymbol{\nabla}(\widehat{\boldsymbol{\tau}}) \cdot \tag{31}$$

The covariance matrix $\widehat{\boldsymbol{\Sigma}}^{(A)}_{\widehat{\boldsymbol{\tau}}}$ in (31) can be estimated using the multivariate regression approach. Now, the matrix $\breve{\boldsymbol{Y}}_s$ contains P response variables $(\breve{y}_{1;i}, \ldots, \breve{y}_{P;i})^\top$ which are given by (30). The matrix $\boldsymbol{Z}_s$ specifies the stratification and suitable interactions which depends on the design used (see §§ 4.2 and 4.1). Let $\widehat{\boldsymbol{V}}^{(A)}$ be the ordinary least squares estimate of the $P \times P$ variance-covariance matrix of the residuals of the model (15). The estimator of the variance-covariance matrix is given by

$$\widehat{\boldsymbol{\Sigma}}^{(A)}_{\widehat{\boldsymbol{\tau}}} \equiv \left\{ \widehat{V}^{(A)}_{pp'} \left(\widehat{V}^{(A)}_{pp} \, \widehat{V}^{(A)}_{p'p'} \right)^{-\frac{1}{2}} [\widehat{var}(\widehat{\tau}_p)\, \widehat{var}(\widehat{\tau}_{p'})]^{\frac{1}{2}} \quad ; \quad p, p' = 1, \ldots, P \right\}; \tag{32}$$

where the right hand side of (32) denotes the element (p, p') of the matrix $\widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\tau}}}^{(A)}$. The quantity $\widehat{V}_{pp'}^{(A)}$ is the elements (p, p') of the matrix $\widehat{\boldsymbol{V}}^{(A)}$. The estimator of the variance of change is given by (31) with $\widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\tau}}}^{(A)}$ given by (32). Note that $\widehat{var}(\widehat{\Delta}_\theta) > 0$ because $\widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\tau}}}^{(A)}$ is positive definite. This is due to the fact that $\widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\tau}}}^{(A)} = \widehat{\boldsymbol{D}}^\top \widehat{\boldsymbol{V}}^{(A)} \widehat{\boldsymbol{D}}$ and $\widehat{\boldsymbol{V}}^{(A)} \propto \widehat{\boldsymbol{S}}^{(A)}$ (see (10)) which is positive definite (see Appendix A), where $\widehat{\boldsymbol{D}} = diag\{\widehat{var}(\widehat{\tau}_p)^{\frac{1}{2}} \widehat{V}_{pp}^{(A)-\frac{1}{2}} \; ; \; p = 1, \ldots, \; P\}$.

The second estimator (12) can be generalised for complex estimators of change by multiplying $\breve{y}_{p;i}$ by $z_{2;i}$ when $p \leq Q$ and by $z_{1;i}$ when $p > Q$. In this case, the covariance between estimators based on the same sample would be computed from the common sample. We recommend to use the estimator (32), because the covariance between estimators based on the same samples s_ℓ is computed from the whole sample s_ℓ, and the covariance between estimators based on different samples are computed from the common sample. This is another advantage of the proposed approach.

Another approach consists in substituting $y_{\ell;i}$ by linearised variables in (5) (e.g. Deville, 1999; Demnati and Rao, 2004), and using the estimator of covariance (7). This approach is recommended when $\widehat{\theta}_2$, $\widehat{\theta}_1$ are not functions of totals (e.g. Oguz-Alper and Berger, 2014). For example, when $\widehat{\theta}_2$ and $\widehat{\theta}_1$ are Gini coefficients.

5. Empirical simulation studies

In a series of simulations based on the Swedish Labour Force Survey, Andersson *et al.* (2011b) (see also Andersson *et al.* (2011a)) showed that for estimation of change within domains defined by the strata, the estimator proposed by Berger (2004) gives accurate variance estimates of change. The estimator proposed in this paper has the same property, as it reduces to the Berger (2004) estimator when the sampling fractions are small (see (43)).

In this §, we report the results of two series of simulations. The first one is based upon the British Labour Force Survey data and the second one is based upon the Italian EU-SILC survey data. For each simulation, 10,000 samples were selected to compute the empirical relative bias (RB)

$$\text{RB} = \frac{E[\widehat{var}(\widehat{\Delta})] - var(\widehat{\Delta})}{var(\widehat{\Delta})} \times 100\% \tag{33}$$

and the empirical relative root mean square error (RRMSE)

$$\text{RRMSE} = \frac{mse[\widehat{var}(\widehat{\Delta})]^{1/2}}{var(\widehat{\Delta})} \times 100\%\cdot \tag{34}$$

The quantity $var(\widehat{\Delta})$ denotes the empirical variance of $\widehat{\Delta}$. The quantities $E[\widehat{var}(\widehat{\Delta})]$ and $mse[\widehat{var}(\widehat{\Delta})]$ denote respectively the empirical expectation and the empirical mean square error of a variance estimator $\widehat{var}(\widehat{\Delta})$.

The Chao (1982) unequal probability design is used to select samples. The sample s_1 is selected with inclusion probabilities $\pi_{1;i}$. The sample s_2 is selected using the sampling design described in Example 1, with $p_i = 1$ and $q_i = \pi_{1;i}/(1 - \pi_{1;i})$. The statistical software R is used to fit the multivariate regression model. The variances $var(\widehat{\tau}_1)$ and $var(\widehat{\tau}_2)$ are estimated by the Hájek (1964) variance estimator.

5.1. British Labour Force Survey Data

We consider the common sample of two waves of the British labour force survey: October-December 2007 and October-December 2008. The dataset is replicated 10 times in order to create a large dataset,

Table 1. Observed RB and RRMSE for the change in unemployment and in mean income from an artificial population based upon the British Labour Force Survey data. Wave 1: October-December 2007. Wave 2: October-December 2008. The values reported in this Table do not reflect the actual estimates from the British Labour force Survey.

		RB (%)				RRMSE (%)			
	g	Proposed	Qualité and Tillé (2008)	Tam (1984)	Wood (2008)	Proposed	Qualité and Tillé (2008)	Tam (1984)	Wood (2008)
Unemployment rate	40%	-7	-7	-6	-7	47	47	46	48
	54%	-3	-2	-1	-2	49	49	47	50
	68%	-4	-3	-2	-2	52	52	48	53
Mean income	40%	-1	0	0	0	13	13	12	19
	54%	-1	0	0	0	14	14	12	22
	68%	-2	0	0	0	16	16	13	27

of size $N = 27{,}320$, which is treated as a population from which samples are selected. This population is stratified into 5 strata based upon the consecutive number of stints. We use a proportional allocation with equal inclusion probabilities within each stratum. We consider the change between the unemployment rates and between the means of income. The sample sizes are $n_1 = 250$ and $n_2 = 275$. We consider several fractions for the common sample: $g = n_c/n_1 = 40\%$, 54% and 68%.

We compare the proposed estimator (7) based on (11) with the variance of change proposed by Tam (1984) which is based on (23) and the estimators proposed by Qualité and Tillé (2008) and Wood (2008). The results of this simulation study are given in Table 1. The relative bias (RB) and the relative root mean squared errors (RRMSE) are those of the variance estimator relative to the empirical variance $var(\widehat{\Delta})$. As this is an illustrative example, the values reported in Table 1 do not reflect the actual estimates from the British Labour force Survey.

With the change in mean income, we observe slight negative biases for the proposed approach. This is due to the fact that the finite population corrections are not taken into account. This effect is more pronounced with skewed variables such as the income variable. All the estimators have similar RB and RRMSE. We notice that the Wood (2008) estimator is slightly more unstable, and Tam (1984) estimator is slightly more stable. The observed differences between the estimator proposed by Tam (1984) and the proposed estimator is due to the fact that the covariance due to Tam (1984) is defined by (23) and with the proposed approach the covariance is given by $[\widehat{var}(\widehat{\tau}_1)\widehat{var}(\widehat{\tau}_2)]^{1/2}\widehat{\rho}^{(A)}_{prop} \simeq [\widehat{var}(\widehat{\tau}_1)\widehat{S}^{-1}_{11}\widehat{var}(\widehat{\tau}_2)\widehat{S}^{-1}_{22}]^{1/2}\widehat{cov}(\widehat{\tau}_1,\widehat{\tau}_2)_{Tam}$ (see (7), (11), (21) and (23)).

5.2. Italian Survey on Income and Living Conditions Data

In this §, we give the results of another simulation study based upon the Italian Statistics on Income and Living Conditions (EU-SILC) survey (see §6 for a description of the EU-SILC surveys). For this simulation study, we consider unequal inclusion probabilities. The common sample of two consecutive years (2008 and 2009) is treated as a population from which stratified samples are selected. This gives a population size $N = 19{,}644$. Stratified samples of size $n_1 = n_2 = 982$ are selected using the uni-stage Chao (1982) sampling design. The strata are the five geographical regions. We consider that we have the same fraction of the common sample, $g = 75\%$, within each stratum.

We consider the change between means (or proportions) of several variables of interest $y_{1;i}$ and $y_{2;i}$. We consider three dummy variables of interest (afford holiday, own a car, at risk of poverty) and one quantitative variable (equivalised disposable income). We also create artificial log-normal

variables with different correlations between the variables of interest. The change between the means (or proportions) is estimated by $\widehat{\Delta} = \widehat{\tau}_2 N^{-1} - \widehat{\tau}_1 N^{-1}$. We consider that the inclusion probabilities $\pi_{1;i}$ are proportional to the inverse of the cross-sectional sampling weights at wave 1. We also consider several domains of interest given by the type of accommodation (detached, semi-detached), the population of home owners, the population of males and the population of females. The households are the units, and the quantities $y_{1;i}$ and $y_{2;i}$ denote the household totals of the variables of interest.

We propose to compare the estimators of the form (7) based on different estimator for the correlation. The following naïve estimator is based on the estimator for the covariance proposed by Tam (1984) (under equal probability sampling).

$$\widehat{\rho}_{SRS} = \widehat{cov}(\widehat{\tau}_1, \widehat{\tau}_2)_{Tam} \left[\widehat{var}(\widehat{\tau}_1)_{SRS} \quad \widehat{var}(\widehat{\tau}_2)_{SRS}\right]^{-\frac{1}{2}} \tag{35}$$

with

$$\widehat{var}(\widehat{\tau}_\ell)_{SRS} = N^2 \left(1 - \frac{n_\ell}{N}\right) \frac{\widehat{\sigma}^2_\ell}{n_\ell},$$

$\widehat{cov}(\widehat{\tau}_1, \widehat{\tau}_2)_{Tam}$ defined by (23) and $\widehat{\sigma}^2_\ell = n_\ell^{-1} \sum_{i \in s} (y_{\ell;i} - \bar{y}_\ell)^2$; where $\bar{y}_\ell$ is the sample mean of s_ℓ.

Another naïve estimator is based upon the stratified Tam (1984) estimator. This correlation is given by

$$\widehat{\rho}_{SSRS} = \widehat{cov}(\widehat{\tau}_1, \widehat{\tau}_2)_{SSRS} \left[\widehat{var}(\widehat{\tau}_1)_{SSRS} \quad \widehat{var}(\widehat{\tau}_2)_{SSRS}\right]^{-\frac{1}{2}}; \tag{36}$$

where

$$\begin{aligned}
\widehat{cov}(\widehat{\tau}_1, \widehat{\tau}_2)_{SSRS} &= \sum_{h=1}^{H} \sum_{i \in s_{ch}} \left(1 - \frac{n_{1h} n_{2h}}{N_h n_{ch}}\right) \frac{n_{ch}}{n_{ch} - 1} \frac{N_h^2 n_{ch}}{n_{1h} n_{2h}} \widehat{\sigma}_{ch}, \\
\widehat{var}(\widehat{\tau}_\ell)_{SSRS} &= \sum_{h=1}^{H} \sum_{i \in s_{\ell h}} \left(1 - \frac{n_{\ell h}}{N_h}\right) N_h^2 \frac{\widehat{\sigma}^2_{\ell h}}{n_{\ell h}}, \\
\widehat{\sigma}_{ch} &= \frac{1}{n_{ch}} \sum_{i \in s_{ch}} (y_{1;i} - \bar{y}_{1h;c}) (y_{2;i} - \bar{y}_{2h;c}) \ , \\
\widehat{\sigma}^2_{\ell h} &= \frac{1}{n_{\ell h}} \sum_{i \in s_{\ell h}} (y_{\ell;i} - \bar{y}_{\ell h})^2 \ .
\end{aligned}$$

The quantities $\bar{y}_{1h;c}$, $\bar{y}_{2h;c}$ are the sample means of the common sample of the stratum h, and $\bar{y}_{\ell h}$ is the sample mean of the stratum h at wave ℓ. We consider the estimators (35) and (36) because they are alternative straightforwardly calculated covariance and variance estimators.

Qualité (2009, p. 83) proposed an estimator for the correlation based on s_c which is treated as a second phase sample drawn randomly from s_1. This estimator is given by

$$\widehat{\rho}_Q = \frac{\sum_{h=1}^{H} n_{ch} (n_{ch} - 1)^{-1} \sum_{i \in s_c} \left(\breve{y}_{i;1} - \bar{\breve{y}}_{1h;c}\right) \left(\breve{y}_{i;2} - \bar{\breve{y}}_{2h;c}\right)}{\left[\widehat{var}(\widehat{\tau}_1)_{HH} \quad \widehat{var}(\widehat{\tau}_2)_{HH}\right]^{\frac{1}{2}}}, \tag{37}$$

where

$$\widehat{var}(\widehat{\tau}_\ell)_{HH} = \sum_{h=1}^{H} \frac{n_{ch}}{g_h (n_{ch} - 1)} \sum_{i \in s_c} \left(\breve{y}_{i;2} - \bar{\breve{y}}_{2h;c}\right)^2 . \tag{38}$$

where $g_h = n_{ch}/n_\ell$. Note that the variance estimator of change based on (6), (35), (36) are not necessarily positive because the correlation can be larger than one as the covariance and the variances

are based on different samples. The proposed variance estimator based on (11) is always positive.

The result of this simulation is given in Table 2. Note that we observe a large negative RB for the Hansen and Hurwitz (1943) 'type' estimator (6). The observed RB of the proposed estimators (based on $\widehat{\rho}_{prop}^{(A)}$ and $\widehat{\rho}_{prop}^{(B)}$) are usually negligible. We observe a slight negative bias for the proposed estimators which is probably due to the fact that the finite population correction is ignored in the correlations. The range of the RB of the proposed estimators across the variables in Table 2 is smaller than the range of the RB of the naïve estimators (based on $\widehat{\rho}_{SSRS}$ and $\widehat{\rho}_{SRS}$). This means that there is less chance of outlying estimates with the proposed approaches. The RRMSE of the naïve estimators are mostly larger than the RRMSE of the proposed estimators. For the log-normal distribution, the RB and RRMSE of the naïve estimators are significantly larger. We do not observe significant differences between the estimator based on $\widehat{\rho}_{prop}^{(B)}$ and $\widehat{\rho}_{Q}$ (Qualité, 2009, p. 83), in term of RB and RRMSE.

Finally, the proposed estimators and Qualité (2009, p. 83) estimator are a good compromise in term of RB and RRMSE. The advantages of the proposed estimator are the fact that they can be computed with any statistical software, they always giver positive variance estimates and they can be easily implemented for change between functions of totals (see § 4.3).

6. AN APPLICATION TO THE EU-SILC HOUSEHOLD SURVEYS

We consider a key poverty indicator: the *at-risk-of-poverty or social exclusion* (AROPE) indicator (Eurostat, 2012b; Atkinson and Marlier, 2010) which is used to monitor poverty within the European Union. This indicator is calculated from the EU-SILC surveys (Eurostat, 2012a) which collect yearly information on income, poverty, social exclusion and living conditions from approximately 300,000 households across Europe.

We consider the change of the AROPE indicator between two consecutive years (2009 and 2010). In this §, we show how to estimate the variance of the net change of the AROPE indicator. The computations of the estimator (31) were made in SAS by Guillaume Osier (European Central Bank), Emilio Di Meglio (Eurostat Unit F4 Quality of Life) and Emanuela Di Falco (Eurostat Unit F4 Quality of Life). The EU-SILC production data bases were used within the premises of Eurostat.

An ultimate cluster approach (see §4.2) was adopted, because the sampling fractions are small. The units are the primary sampling units (PSUs). For some countries, the PSUs are households (e.g. Austria, UK, Latvia). Scandinavian countries, use single stage design based on registers. In this case, the PSUs are sets containing one individual. The response variables of the multivariate model are given by (28) or equivalently

$$\breve{y}_{\ell;i} = \delta\{i \in s_\ell\} \sum_{j \in \text{PSU}_i} w_{\ell;j}\, y_{\ell;j} \ ,$$

where s_ℓ is the sample of PSUs at wave ℓ, PSU_i denotes the i-th PSU, $y_{\ell;j}$ is the value of the variable of interest for individuals j and $w_{\ell;j}$ is the survey cross-sectional weight of individuals j at wave ℓ. The variables $z_{\ell h;i}$ are dummy variables which specify the stratification at PSU level. The variables $\breve{y}_{\ell;i}$ and $z_{\ell h;i}$ need to be defined for all $i \in s = s_1 \cup s_2$.

The AROPE depends on a poverty threshold which is estimated. The estimation of the poverty threshold can be taken into account using a linearised variables technique (Osier, 2009). Oguz-Alper and Berger (2014) showed how the proposed approach can be used to take into account the variability of the threshold. For simplicity, we assume that the poverty threshold is fixed which ensures conservative cross-sectional variances (Preston, 1995; Berger and Skinner, 2003).

In this §, we consider that the AROPE indicator is a ratio of two totals: an estimate of the total number of individuals in poverty and social exclusion divided by an estimate of the population size

Table 2. Observed RB and RRMSE for several variables of interest and several domain of interest. Italian EU-SILC data (2008, 2009). These values are for illustrative purpose only, and are not part of any results officially released.

		RB (%) of (7) with						RRMSE (%) of (7) with					
Variables	Domains	$\widehat{\rho}_{prop}^{(A)}$	$\widehat{\rho}_{prop}^{(B)}$	$\widehat{\rho}_{HH}$	$\widehat{\rho}_{Q}$	$\widehat{\rho}_{SSRS}$	$\widehat{\rho}_{SRS}$	$\widehat{\rho}_{prop}^{(A)}$	$\widehat{\rho}_{prop}^{(B)}$	$\widehat{\rho}_{HH}$	$\widehat{\rho}_{Q}$	$\widehat{\rho}_{SSRS}$	$\widehat{\rho}_{SRS}$
Afford Holiday	Population	-3.6	-3.0	-9.6	-3.0	13.2	12.2	13.5	12.1	16.4	12.1	18.8	18.0
	Detached	-1.5	-0.6	-7.2	-0.5	4.2	5.0	27.3	23.9	28.5	23.9	23.0	23.3
	Semi-detached	-5.7	-5.0	-11.1	-5.0	-2.2	-1.5	27.6	23.7	29.5	23.7	22.9	22.9
	Home owner	-2.9	-2.5	-8.7	-2.4	10.8	10.5	13.5	11.8	16.0	11.8	17.2	16.9
	Males	-3.8	-3.2	-9.7	-3.1	7.9	7.3	16.2	14.2	18.7	14.2	17.5	17.1
	Females	-3.5	-3.0	-9.4	-2.9	9.7	9.3	15.7	14.0	18.1	14.0	17.5	17.2
Own Car	Population	-7.0	-6.2	-19.5	-6.1	0.2	1.4	16.4	12.2	24.8	12.2	12.2	12.4
	Detached	-6.3	-5.0	-15.3	-4.8	-2.8	-1.3	26.5	18.6	30.3	18.6	20.1	20.2
	Semi-detached	-6.4	-5.2	-15.2	-5.0	-2.7	-1.2	28.5	20.5	32.1	20.5	21.6	21.6
	Home owner	-4.4	-3.7	-13.9	-3.6	2.0	3.0	14.3	11.0	19.7	11.0	12.1	12.4
	Males	-6.9	-6.1	-18.3	-6.0	-1.5	-0.2	18.9	13.7	25.7	13.6	14.5	14.5
	Females	-6.2	-5.2	-17.8	-5.1	0.3	1.6	18.0	13.0	24.8	12.9	14.1	14.3
Equivalised	Population	-5.8	-4.7	-16.0	-4.6	4.1	4.5	29.4	22.6	33.3	22.6	29.3	29.3
Disposable	Detached	-3.6	-2.2	-11.6	-2.2	-0.6	0.5	40.4	32.6	42.1	32.6	37.4	37.6
Income	Semi-detached	-4.7	-3.2	-12.8	-3.1	0.7	1.8	38.7	29.4	40.8	29.3	35.3	35.6
	Home owner	-4.0	-2.9	-12.8	-2.8	5.6	6.0	29.0	22.0	31.7	22.0	28.9	28.9
	Males	-5.0	-3.7	-14.4	-3.6	4.1	4.8	33.9	25.2	36.8	25.2	31.2	31.2
	Females	-5.8	-5.1	-15.5	-5.0	-0.4	0.2	27.4	23.2	31.1	23.2	26.0	26.1
At Risk of	Population	-2.6	-2.1	-5.8	-2.0	6.0	2.3	26.7	24.1	27.5	24.1	27.1	25.7
Poverty	Detached	-4.0	-1.9	-7.3	-1.8	1.5	1.3	52.2	47.8	52.9	47.9	54.2	54.0
	Semi-detached	-1.1	0.5	-4.4	0.7	2.4	2.6	58.0	53.2	58.5	53.2	53.8	53.9
	Home owner	-0.9	-0.1	-3.9	0.0	6.6	4.7	29.8	27.5	30.3	27.5	32.0	31.1
	Males	-3.9	-3.5	-6.8	-3.4	3.0	0.5	35.2	32.8	35.9	32.8	30.9	30.0
	Females	-1.8	-0.9	-5.3	-0.8	5.3	2.4	26.8	23.1	27.6	23.1	29.3	28.2
LogNormal	$\mathrm{Corr}(y_1, y_2) = 0.90$	-4.8	-3.2	-14.1	-3.1	4.8	6.0	35.2	29.2	37.9	29.2	34.9	35.3
	$\mathrm{Corr}(y_1, y_2) = 0.80$	-3.9	-3.2	-11.1	-3.1	7.4	8.2	23.2	18.7	25.6	18.8	24.3	24.8
	$\mathrm{Corr}(y_1, y_2) = 0.70$	-1.9	-1.7	-7.6	-1.6	15.5	16.2	27.2	24.1	28.1	24.1	33.6	34.0
	$\mathrm{Corr}(y_1, y_2) = 0.50$	-1.2	-0.9	-4.8	-0.9	15.6	16.0	15.8	14.9	16.5	14.9	23.2	23.5
	$\mathrm{Corr}(y_1, y_2) = 0.30$	-1.5	-1.4	-3.7	-1.3	15.1	15.3	13.3	12.9	13.7	12.9	21.6	21.7
	$\mathrm{Corr}(y_1, y_2) = 0.20$	-2.9	-2.8	-4.7	-2.8	17.9	18.0	15.1	14.7	15.5	14.7	24.8	24.9
	$\mathrm{Corr}(y_1, y_2) = 0.10$	-2.3	-2.3	-3.6	-2.2	18.3	18.3	13.0	12.8	13.3	12.8	23.6	23.6

(or exposure if we are interested in a domain). Hence $\widehat{\tau}$ in (29) is a vector of four totals. The estimator (31) is used. The effect of calibration can be taken into account, by replacing the response variables by residuals (Deville and Särndal, 1992). However, the effect of calibration was ignored, because the calibration variables were not available. For multi-stage designs, the effect of re-weighting due to non-response adjustments does not need to be taken into account, because these adjustments are done within PSUs. For single stage designs, the effect of non-response adjustments is ignored. This is not crucial, because single stage designs are often based on registers (like in the Scandinavian countries) which usually have a small fraction of missing values. The effect of imputation was ignored. Note that some countries use a rotation within PSU (e.g. Belgium). In this case, the proposed approach can still be used (see end of §4.2).

The estimates based on the proposed approach are given in Table 3. We notice that the change can be significant for some countries (the values in bold face). However, these estimates need to be interpreted with caution. These estimates are for illustrative purpose only, and are not part of any results officially released by Eurostat. The quality of these estimates relies on the availability and

Table 3. Illustrative estimates of the AROPE indicator for 2009 and 2010 based on the EU-SILC surveys' data. The estimates of change in bold face are statistically significant at 5%. These estimates need to be interpreted with caution. These estimates are for illustrative purpose only, and are not part of any results officially released by Eurostat. The values in bold face are significantly different from zero (p-value$<$0.05)

Country	AROPE 2009 (%)	AROPE 2010 (%)	Change (in % point)	Standard Error	Country	AROPE 2009 (%)	AROPE 2010 (%)	Change (in % point)	Standard Error
Iceland	11.6	13.7	**2.09**	0.34	Malta	20.2	20.3	0.09	0.42
Czech Rep.	14.0	14.4	0.36	0.30	UK	22.0	23.1	**1.18**	0.25
Netherlands	15.1	15.1	-0.07	0.14	Cyprus	22.9	23.6	0.67	0.55
Norway	15.2	14.9	-0.34	0.28	Estonia	23.4	21.7	**-1.69**	0.38
Sweden	15.9	15.0	**-0.90**	0.29	Spain	23.4	25.5	**2.16**	0.02
Finland	16.9	16.9	-0.01	0.33	Italy	24.7	24.5	-0.16	0.32
Austria	17.0	16.6	-0.44	0.27	Portugal	24.9	25.3	**0.40**	0.10
Slovenia	17.1	18.3	**1.17**	0.22	Ireland	25.7	29.9	**4.18**	0.93
Switzerland	17.2	17.2	-0.08	0.39	Greece	27.6	27.7	0.11	0.30
Denmark	17.6	18.3	0.74	0.40	Poland	27.8	27.8	-0.07	0.27
Luxembourg	17.8	17.1	-0.72	0.43	Lithuania	29.5	33.4	**3.90**	0.48
France	18.5	19.2	0.71	0.53	Hungary	29.6	29.9	0.32	0.41
Slovakia	19.6	20.6	**1.01**	0.17	Latvia	37.4	38.1	0.64	0.34
Germany	20.0	19.7	-0.26	0.24	Romania	43.1	41.4	**-1.66**	0.11
Belgium	20.2	20.8	**0.66**	0.07	Bulgaria	46.2	41.6	**-4.57**	0.75

quality of the design variables. These estimates are likely to overestimate the variance because the effect of calibration adjustment was not taken into account. This effect may be more pronounced for Scandinavian countries.

7. Discussion

The proposed approach can be used for a large class of parameters which can be expressed as functions of totals (see § 4.3). The main contribution of the paper is to show that variance estimates can be calculated using the covariance of the residuals of a multivariate regression model with suitable interactions. It does not require the development of a specialised package, as any statistical software can be used to compute the covariance of the residuals of the multivariate regression model. The simplicity and flexibility of the proposed approach makes it a suitable tool for common variance estimation procedure across the EU-SILC surveys.

One of the advantage of the proposed approach is the fact that the variance-covariance matrix is estimated using a single regression model, even if we have several totals and several strata. Alternative approaches would involve calculating each component of the matrix separately by using (4), for each combination of variables and for each stratum. This approach may give a negative definite covariance matrices and possible negative variance estimates of change. The proposed approach always gives a positive definite covariance matrix and positive variance estimates.

For functions of totals, the linearised variable approach involves deriving linearised variables for each measure of change that can be considered by the users. Another advantage of the proposed approach is the fact that the same variance-covariance matrix can be used for several measures of change (function of the same set of totals). Only the gradient $\nabla(\widehat{\tau})$ differs. The proposed approach involves computing a single covariance matrix (32) which could be provided to the users. This matrix can be used for any differentiable function of the totals involved in (29). Only the gradient $\nabla(\widehat{\tau})$ has to be specified by the user. The proposed approach is more suitable when we do know which measure

of change will be considered by the user and when confidential information, such as design variables cannot be released. The users only need to know the covariance matrix. The EU-SILC user database does not contain all the design and auxiliary variables for confidentiality reasons. For example, the stratification is not available in the 2010 EU-SILC user database.

The proposed approach can be used under without replacement sampling with negligible sampling fractions which is a common feature of social surveys. Large sampling fractions (combined with sampling without replacement) are common practice in business surveys. The proposed approach is not suitable in this case. The approaches proposed by Nordberg (2000), Berger (2004), Wood (2008), Goga *et al.* (2009), Muennich and Zins (2011) and Knottnerus and van Delden (2012) can be used in this case.

With calibration within each wave, we propose to include the auxiliary variables in the regression model (8) or (15). This would give a suitable covariance estimator for single stage designs because the response variables will be projected within the space spanned by the auxiliary variables. For more complex situations (e.g. multi-stage designs), we recommend using the approach proposed in Section 4.3. Note that for regression estimators, the number of response variables in the multivariate regression model is equal to the number of totals involved in the function of totals. For example, if we have three auxiliary variables for each wave, the number of response variables is $2(1+3)=8$.

The proposed approach can also be extended for measures of poverty which are not functions of totals (Oguz-Alper and Berger, 2014). It can also be extended for measuring trends from more than two overlapping samples (Berger, 2011). Berger and Escobar (2013) extended the proposed approach under non-response. The effect of panel attrition has been ignored in the variance estimation, and is beyond the scope of this paper.

The proposed approach relies on a set of conditions which are met by a wide range of social survey designs used in practice. We give here a summary of these conditions. We assume that the number of strata is asymptotically bounded. This assumption might not be valid for heavily stratified designs. We assume that the sampling fraction is negligible. The proposed approach relies on the assumption that sample size of the overlapping sample is fixed. It cannot be used when this sample size is random. We also assume that the sampling design has a high entropy. This assumption is usually met in practice except with the non-randomised systematic sampling design.

Acknowledgements

This work was supported by the grant RES-000-22-3045 of the Economic and Social Research Council (UK) and by consulting work for the Net-SILC2 project (Atkinson and Marlier, 2010). We are also grateful to Guillaume Osier (European Central Bank), Emilio Di Meglio (Eurostat Unit F4 Quality of Life), Emanuela Di Falco (Eurostat Unit F4 Quality of Life) for testing the approach on the EU-SILC survey data. We are also grateful to Dr. Emilio López Escobar (Instituto Technológico Autónomo de México, México), Melike Oguz Alper (University of Southampton) and Tim Goedemé (University of Antwerp, Belgium) for helpful comments. We wishe to thank the anonymous reviewers for helpful comments and suggestions.

Appendix A (Proof of equation (43))

Berger (2004) showed that under the assumption of high entropy,

$$\widehat{\boldsymbol{S}}^{*}=\widehat{\boldsymbol{S}}_{\boldsymbol{\tau\tau}}-\widehat{\boldsymbol{S}}_{\boldsymbol{\tau n}}\widehat{\boldsymbol{S}}_{\boldsymbol{nn}}^{-1}\widehat{\boldsymbol{S}}_{\boldsymbol{\tau n}}^{\top}\,; \tag{39}$$

is a consistent estimator for the covariance matrix $\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\tau}}}$ (defined in (3)); with

$$\widehat{\boldsymbol{S}}_{\boldsymbol{\tau\tau}} = \begin{pmatrix} \sum_{i\in s} \breve{c}_{1;i}\ \breve{y}_{i;1}^2 & \sum_{i\in s} \breve{c}_{12;i}\ \breve{y}_{i;1}\ \breve{y}_{i;2} \\ \sum_{i\in s} \breve{c}_{12;i}\ \breve{y}_{i;1}\ \breve{y}_{i;2} & \sum_{i\in s} \breve{c}_{2;i}\ \breve{y}_{i;2}^2 \end{pmatrix}, \tag{40}$$

$$\widehat{\boldsymbol{S}}_{\boldsymbol{nn}} = \begin{pmatrix} \sum_{i\in s} \breve{c}_{1;i}\ z_{1;i} & \sum_{i\in s} \breve{c}_{12;i}\ z_{1;i}\ z_{2;i} & \sum_{i\in s} \breve{c}_{1;i}\ z_{1;i}\ z_{2;i} \\ \sum_{i\in s} \breve{c}_{12;i}\ z_{1;i}\ z_{2;i} & \sum_{i\in s} \breve{c}_{2;i}\ z_{2;i} & \sum_{i\in s} \breve{c}_{2;i}\ z_{1;i}\ z_{2;i} \\ \sum_{i\in s} \breve{c}_{1;i}\ z_{1;i}\ z_{2;i} & \sum_{i\in s} \breve{c}_{2;i}\ z_{1;i}\ z_{2;i} & \sum_{i\in s} \breve{c}_{c;i}\ z_{1;i}\ z_{2;i} \end{pmatrix}, \tag{41}$$

$$\widehat{\boldsymbol{S}}_{\boldsymbol{\tau n}} = \begin{pmatrix} \sum_{i\in s} \breve{c}_{1;i}\ \breve{y}_{i;1}\ z_{1;i} & \sum_{i\in s} \breve{c}_{12;i}\ \breve{y}_{i;1}\ z_{2;i} & \sum_{i\in s} \breve{c}_{1;i}\ \breve{y}_{i;1}\ z_{1;i}\ z_{2;i} \\ \sum_{i\in s} \breve{c}_{12;i}\ \breve{y}_{i;2}\ z_{1;i} & \sum_{i\in s} \breve{c}_{2;i}\ \breve{y}_{i;2}\ z_{2;i} & \sum_{i\in s} \breve{c}_{2;i}\ \breve{y}_{i;2}\ z_{1;i}\ z_{2;i} \end{pmatrix}; \tag{42}$$

where $s = s_1 \cup s_2$ denotes the overall sample. The quantities $\breve{c}_{\ell;i}$, $\breve{c}_{c;i}$ and $\breve{c}_{12;i}$ are finite population corrections given by $\breve{c}_{\ell;i} = (1 - \pi_{\ell;i})$, $\breve{c}_{c;i} = (1 - \pi_{c;i})$ and $\breve{c}_{12;i} = 1 - \pi_{1;i}\pi_{2;i}/\pi_{c;i}$. The variables $z_{1;i}$ and $z_{2;i}$ are design variables defined by (9). The variables $\breve{y}_{1;i}$ and $\breve{y}_{2;i}$ are defined by (5) with $\breve{y}_{\ell;i} = 0$ when $i \notin s_\ell$.

When we have non-overlapping samples, we may have $g = 0$ and $\pi_{c;i} = 0$ for all i. In this case, we consider that $\breve{c}_{12;i} = 1$ for all i by definition, as $\breve{c}_{12;i}\breve{y}_{i;1}\breve{y}_{i;2} = \breve{c}_{12;i}z_{1;i}\ z_{2;i} = \breve{c}_{12;i}\ \breve{y}_{i;1}z_{2;i} = 0$ for all i. In this case, the last row and column of (41) have to be removed, as well as the last column of (42). This way (39) reduces to an estimator of covariance for non-overlapping samples. Note that the extra-diagonal element of (39) equal zero in this case.

The ordinary least squares estimate of the variance-covariance matrix $\boldsymbol{V}$ of the residuals is given by

$$\widehat{\boldsymbol{V}}^{(A)} = \widehat{\boldsymbol{S}}^{(A)}/\alpha\ ;$$

where

$$\widehat{\boldsymbol{S}}^{(A)} = (\breve{\boldsymbol{Y}}_s - \boldsymbol{Z}_s\widehat{\boldsymbol{\beta}})^\top(\breve{\boldsymbol{Y}}_s - \boldsymbol{Z}_s\widehat{\boldsymbol{\beta}})\ ; \tag{43}$$

with

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{Z}_s^\top \boldsymbol{Z}_s)^{-1}\boldsymbol{Z}_s^\top \breve{\boldsymbol{Y}}_s.$$

The $n \times 2$ matrix $\breve{\boldsymbol{Y}}_s$ and the $n \times 3$ matrix $\boldsymbol{Z}_s$ are defined by (16) and (17). Note that (43) implies

$$\widehat{\boldsymbol{S}}^{(A)} = \breve{\boldsymbol{Y}}_s^\top \breve{\boldsymbol{Y}}_s - \breve{\boldsymbol{Y}}_s^\top \boldsymbol{Z}_s(\boldsymbol{Z}_s^\top \boldsymbol{Z}_s)^{-1}\boldsymbol{Z}_s^\top \breve{\boldsymbol{Y}}_s. \tag{44}$$

Assumptions (24)-(26) imply that (40), (41) and (42) reduce to

$$\widehat{\boldsymbol{S}}_{\boldsymbol{\tau\tau}} \eqsim \breve{\boldsymbol{Y}}_s^\top \breve{\boldsymbol{Y}}_s\ ,\quad \widehat{\boldsymbol{S}}_{\boldsymbol{nn}} \eqsim \boldsymbol{Z}_s^\top \boldsymbol{Z}_s \quad \text{and} \quad \widehat{\boldsymbol{S}}_{\boldsymbol{\tau n}} \eqsim \breve{\boldsymbol{Y}}_s^\top \boldsymbol{Z}_s. \tag{45}$$

Finally by substituting (45) into (39), we obtain

$$\begin{aligned} \widehat{\boldsymbol{S}}^* &\eqsim \breve{\boldsymbol{Y}}_s^\top \breve{\boldsymbol{Y}}_s - \breve{\boldsymbol{Y}}_s^\top \boldsymbol{Z}_s(\boldsymbol{Z}_s^\top \boldsymbol{Z}_s)^{-1}\boldsymbol{Z}_s^\top \breve{\boldsymbol{Y}}_s \\ &= \widehat{\boldsymbol{S}}^{(A)}\ , \end{aligned}$$

using (44). Thus assuming (24)-(26), we have that $\widehat{\boldsymbol{S}}^* \eqsim \widehat{\boldsymbol{S}}^{(A)}$. Note that $\widehat{\boldsymbol{S}}^{(A)}$ is positive definite because $\widehat{\boldsymbol{S}}^{(A)}$ is a Gram matrix.

For completely overlapping samples ($g = 1$), the interactions are removed from the model and (44) gives the standard estimator for the covariance, that is (4) with $s_c = s_1 = s_2$. This completes the proof.

Appendix B

When $g = 1$, the interaction $z_{1;i}\ z_{2;i}$ and $z_{2;i}$ are removed from the model (8). The proposed approach gives the standard estimators for the correlation from completely overlapping samples. In this Appendix, we consider that $g < 1$.

We have that

$$\boldsymbol{Z}_s^\top \boldsymbol{Z}_s = n_c \begin{pmatrix} m_1 & 1 & 1 \\ 1 & m_2 & 1 \\ 1 & 1 & 1 \end{pmatrix};$$

where $m_1 = n_1/n_c$ and $m_2 = n_2/n_c$. This implies that

$$(\boldsymbol{Z}_s^\top \boldsymbol{Z}_s)^{-1} = \begin{pmatrix} a & 0 & -a \\ 0 & b & -b \\ -a & -b & c \end{pmatrix}; \tag{46}$$

where $a = [n_c(m_1 - 1)]^{-1}$, $b = [n_c(m_2 - 1)]^{-1}$, $c = (m_1 m_2 - 1)[n_c(m_1 - 1)(m_2 - 1)]^{-1}$.

We also have that

$$\breve{\boldsymbol{Y}}_s^\top \boldsymbol{Z}_s = \begin{pmatrix} \widehat{t}_1 & \widehat{t}_{1;c} & \widehat{t}_{1;c} \\ \widehat{t}_{2;c} & \widehat{t}_2 & \widehat{t}_{2;c} \end{pmatrix}; \tag{47}$$

where $\widehat{t}_\ell = n_\ell \breve{\bar{y}}_\ell$, $\widehat{t}_{\ell;c} = n_c \breve{\bar{y}}_{\ell;c}$ and $\widehat{t}_{\ell|c} = n_{\ell|c} \breve{\bar{y}}_{\ell|c}$. By multiplying (46) by (47), we have that

$$\breve{\boldsymbol{Y}}_s^\top \boldsymbol{Z}_s (\boldsymbol{Z}_s^\top \boldsymbol{Z}_s)^{-1} = \begin{pmatrix} \widehat{q}_{11} & 0 & \widehat{q}_{13} \\ 0 & \widehat{q}_{22} & \widehat{q}_{23} \end{pmatrix};$$

where $\widehat{q}_{11} = a(\widehat{t}_1 - \widehat{t}_{1;c})$, $\widehat{q}_{13} = c\,\widehat{t}_{1;c} - a\,\widehat{t}_1 - b\,\widehat{t}_{1;c}$, $\widehat{q}_{22} = b(\widehat{t}_2 - \widehat{t}_{2;c})$ and $\widehat{q}_{23} = c\,\widehat{t}_{2;c} - a\,\widehat{t}_{2;c} - b\,\widehat{t}_2$. Thus the element $(1,2)$ of the matrix $\breve{\boldsymbol{Y}}_s^\top \boldsymbol{Z}_s (\boldsymbol{Z}_s^\top \boldsymbol{Z}_s)^{-1} \boldsymbol{Z}_s^\top \breve{\boldsymbol{Y}}_s$ is given by

$$\{\breve{\boldsymbol{Y}}_s^\top \boldsymbol{Z}_s (\boldsymbol{Z}_s^\top \boldsymbol{Z}_s)^{-1} \boldsymbol{Z}_s^\top \breve{\boldsymbol{Y}}_s\}_{(1,2)} = \widehat{q}_{11}\widehat{t}_{2;c} + \widehat{q}_{13}\widehat{t}_{2;c}. \tag{48}$$

The element$(1,2)$ of the matrix $\breve{\boldsymbol{Y}}_s^\top \breve{\boldsymbol{Y}}_s$ is given by

$$\{\breve{\boldsymbol{Y}}_s^\top \breve{\boldsymbol{Y}}_s\}_{(1,2)} = \sum_{i \in s_c} \breve{y}_{i;1} \breve{y}_{i;2}. \tag{49}$$

By subtracting (48) from (49), and by using (44), we have that the element $(1,2)$ of $\widehat{\boldsymbol{S}}^{(A)}$ is given by

$$\begin{aligned} \widehat{S}_{12}^{(A)} &= \sum_{i \in s_c} \breve{y}_{i;1} \breve{y}_{i;2} - \widehat{q}_{11}\widehat{t}_{2;c} - \widehat{q}_{13}\widehat{t}_{2;c} \\ &= \sum_{i \in s_c} \breve{y}_{i;1} \breve{y}_{i;2} - \frac{\widehat{t}_{1;c}\widehat{t}_{2;c}}{n_c} \\ &= \frac{n_c - 1}{n_c} \widehat{cov}(\widehat{\tau}_1, \widehat{\tau}_2)_{HH}. \end{aligned}$$

It can also be shown that the diagonal element (ℓ,ℓ) of $\widehat{\boldsymbol{S}}^{(A)}$ is given by

$$\widehat{S}_{\ell\ell}^{(A)} = \sum_{s_\ell} \breve{y}_{i;\ell}^2 - \frac{1}{n_c(m_\ell - 1)} \left\{ \left(\widehat{t}_\ell - \widehat{t}_{\ell;c}\right)^2 + (m_\ell - 1)\widehat{t}_{\ell;c}^2 \right\}$$

$$
\begin{aligned}
&= \sum_{s_c} \breve{y}^2_{i;\ell} - \bar{\breve{y}}^2_{\ell;c} + \frac{n_\ell - n_c}{n_{\ell|c}} \sum_{i \in s_\ell / s_c} \breve{y}^2_{i;\ell} - \bar{\breve{y}}^2_{\ell|c} \\
&= \sum_{i \in s_c} \left(\breve{y}_{i;\ell} - \bar{\breve{y}}_{\ell;c}\right)^2 + \sum_{i \in s_\ell / s_c} \left(\breve{y}_{i;\ell} - \bar{\breve{y}}_{\ell|c}\right)^2 \qquad (50) \\
&= \sum_{i \in s_\ell} \left(\breve{y}^2_{i;\ell} - \bar{\breve{y}}_{\ell}\right)^2 - \nu_\ell \cdot
\end{aligned}
$$

where $\nu_\ell = n_{\ell|c}(\bar{\breve{y}}_\ell - \bar{\breve{y}}_{\ell|c})^2 + n_c(\bar{\breve{y}}_\ell - \bar{\breve{y}}_{\ell;c})^2$, with $n_{\ell|c} = n_\ell - n_c$ and

$$
\bar{\breve{y}}_\ell = \frac{1}{n_\ell} \sum_{i \in s_\ell} \breve{y}_{i;\ell} \; , \qquad \bar{\breve{y}}_{\ell;c} = \frac{1}{n_c} \sum_{i \in s_c} \breve{y}_{i;\ell} \; , \qquad \bar{\breve{y}}_{\ell|c} = \frac{1}{n_{\ell|c}} \sum_{i \in s_{\ell|c}} \breve{y}_{i;\ell} \; ;
$$

where $s_{\ell|c} = s_\ell \setminus s_c$. Note that the first term on the right hand side of (50) is the Hansen and Hurwitz (1943) variance (see also Särndal *et al.*, 1992, pp 51 & 52). Thus $\widehat{S}^{(A)}_{\ell\ell}$ is equal to the standard with replacement variance estimator minus a negligible term ν_ℓ. This term is negligible, because $\sum_{i \in s_\ell} \left(\breve{y}_{i;\ell} - \bar{\breve{y}}_\ell\right)^2 = O_p(N^2 n^{-1})$ and $\nu_\ell = O_p(N^2 n^{-2})$. Thus, (11) is a consistent estimator for the correlation because of (10).

Now, we show that (12) is approximately equal to the estimator for the correlation proposed by Qualité (2009, p. 83) defined by (37). We have that $\breve{y}^{(B)}_{1;i} = \breve{y}^{(B)}_{2;i} = 0$ for $i \notin s_c$. Thus, $\widehat{S}^{(A)}_{12} = \widehat{S}^{(B)}_{12}$ and (50) implies

$$
\widehat{S}^{(B)}_{\ell\ell} = \sum_{i \in s_c} \left(\breve{y}_{i;\ell} - \bar{\breve{y}}_{\ell;c}\right)^2 .
$$

Hence

$$
\widehat{\rho}^{(B)}_{prop} \;=\; g \frac{n_c - 1}{n_c} \widehat{cov}(\widehat{\tau}_1, \widehat{\tau}_2)_{HH} \left[\sum_{i \in s_c} \left(\breve{y}_{i;1} - \bar{\breve{y}}_{1;c}\right)^2 \sum_{i \in s_c} \left(\breve{y}_{i;2} - \bar{\breve{y}}_{2;c}\right)^2 \right]^{-\frac{1}{2}}
$$

which is approximately equal to (37), when we have a single stratum and $(n_c - 1) n_c^{-1} \backsimeq 1$.

Appendix C

For a stratified design, we have that the design variables are given by the $n \times (2H - 1)$ matrix

$$
\boldsymbol{Z}_s = (\boldsymbol{Z}_{s1}, \boldsymbol{Z}_{s2}, \cdots, \boldsymbol{Z}_{sh}, \cdots, \boldsymbol{Z}_{sH}) \; ;
$$

where $\boldsymbol{Z}_{sh}$ are $n_h \times 3$ matrices given by $\boldsymbol{Z}_{sh} = (\boldsymbol{z}_{1h}, \boldsymbol{z}_{2h}, \boldsymbol{z}_{ch})$, with $\boldsymbol{z}_\ell = (z_{\ell;1}, z_{\ell;2}, \cdots, z_{\ell;n})^\top$ and $\boldsymbol{z}_{ch} = (z_{1;1h} z_{2;1h}, z_{1;2h} z_{2;2h}, \cdots, z_{1;n_h h} z_{2;n_h h})^\top$. Note that $n_h = \#\{s_{1h} \cup s_{2h}\}$.

The $n \times 2$ matrix $\breve{\boldsymbol{Y}}_s$ is given by

$$
\breve{\boldsymbol{Y}}_s = \left(\breve{\boldsymbol{Y}}^\top_{s1}, \breve{\boldsymbol{Y}}^\top_{s2}, \cdots, \breve{\boldsymbol{Y}}^\top_{sh}, \cdots, \breve{\boldsymbol{Y}}^\top_{sH}\right)^\top \; ;
$$

where $\breve{\boldsymbol{Y}}_{sh} = (\breve{\boldsymbol{y}}_{1h}, \breve{\boldsymbol{y}}_{2h})$ is a $n_h \times 2$ matrix with $\breve{\boldsymbol{y}}_{\ell h} = (\breve{y}_{\ell;1h}, \breve{y}_{\ell;2h}, \cdots, \breve{y}_{\ell;n_h h})^\top$.

The $(3H) \times (3H)$ block diagonal matrix $\boldsymbol{Z}^\top_s \boldsymbol{Z}_s$ is given by

$$
\boldsymbol{Z}^\top_s \boldsymbol{Z}_s = diag\left\{ \boldsymbol{Z}^\top_{s1} \boldsymbol{Z}_{s1} \; , \cdots \; , \; \boldsymbol{Z}^\top_{sH} \boldsymbol{Z}_{sH} \right\} . \qquad (51)
$$

The $2 \times (2H-1)$ matrix $\breve{\boldsymbol{Y}}_s^\top \boldsymbol{Z}_s$ is given by

$$\breve{\boldsymbol{Y}}_s^\top \boldsymbol{Z}_s = (\breve{\boldsymbol{Y}}_{s1}^\top \boldsymbol{Z}_{s1}, \breve{\boldsymbol{Y}}_{s2}^\top \boldsymbol{Z}_{s2}, \cdots, \breve{\boldsymbol{Y}}_{sH}^\top \boldsymbol{Z}_{sH}) \cdot \tag{52}$$

We also have that the 2×2 matrix $\breve{\boldsymbol{Y}}_s^\top \breve{\boldsymbol{Y}}_s$ is given by

$$\breve{\boldsymbol{Y}}_s^\top \breve{\boldsymbol{Y}}_s = \sum_{h=1}^{H} \breve{\boldsymbol{Y}}_{sh}^\top \breve{\boldsymbol{Y}}_{sh} \cdot \tag{53}$$

Finally, by substituting (51), (52) and (53) into (44), we obtain the ordinary least squares estimate $\widehat{\boldsymbol{S}}^{(A)}$ of the covariance of the residuals given by

$$\begin{aligned}
\widehat{\boldsymbol{S}}^{(A)} &= \sum_{h=1}^{H} \breve{\boldsymbol{Y}}_{sh}^\top \breve{\boldsymbol{Y}}_{sh} - \left(\breve{\boldsymbol{Y}}_{s1}^\top \boldsymbol{Z}_{s1}, \cdots, \breve{\boldsymbol{Y}}_{sH}^\top \boldsymbol{Z}_{sH} \right) diag\left\{ \boldsymbol{Z}_{s1}^\top \boldsymbol{Z}_{s1}, \cdots, \boldsymbol{Z}_{sH}^\top \boldsymbol{Z}_{sH} \right\}^{-1} \begin{pmatrix} \boldsymbol{Z}_{s1}^\top \breve{\boldsymbol{Y}}_{s1} \\ \vdots \\ \boldsymbol{Z}_{s1}^\top \breve{\boldsymbol{Y}}_{s1} \end{pmatrix} \\
&= \sum_{h=1}^{H} \left[\breve{\boldsymbol{Y}}_{sh}^\top \breve{\boldsymbol{Y}}_{sh} - \breve{\boldsymbol{Y}}_{sh}^\top \boldsymbol{Z}_{sh} (\boldsymbol{Z}_{sh}^\top \boldsymbol{Z}_{sh})^{-1} \boldsymbol{Z}_{sh}^\top \breve{\boldsymbol{Y}}_{sh} \right] \\
&= \sum_{h=1}^{H} \begin{pmatrix} \widehat{S}_{11h} & \widehat{S}_{ch} \\ \widehat{S}_{ch}^\top & \widehat{S}_{22h} \end{pmatrix};
\end{aligned}$$

where $\widehat{S}_{ch}$ (resp. $\widehat{S}_{\ell\ell h}$) denotes the within stratum covariances (resp. variances) of the residuals. The same result can be derived for $\widehat{\boldsymbol{S}}^{(B)}$. This completes the proof.

References

Andersson, C., Andersson, K. and Lundquist, P. (2011a) Estimation of change in a rotation panel design. *Proceeding of the 58th session of International Statistical Institute, Dublin.*

Andersson, C., Andersson, K. and Lundquist, P. (2011b) Variansskattningar avseende förändringsskattningar i panelundersökningar (variance estimation of change in panel surveys). *Methodology reports from Statistics Sweden (Statistiska centralbyrån).*

Atkinson, A. B. and Marlier, E. (2010) *Income and living conditions in Europe*. Luxembourg: Office for Official Publications `http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-31-10-555/EN/KS-31-10-555-EN.PDF`.

Berger, Y. G. (2004) Variance estimation for measures of change in probability sampling. *Canadian Journal of Statistics*, **32**, 451–467.

Berger, Y. G. (2005) Variance estimation with highly stratified sampling designs with unequal probabilities. *Australian and New Zealand Journal of Statistics*, **47**, 365–373.

Berger, Y. G. (2011) Variance estimation for measures of trends with rotated repeated surveys. *Proceeding of the Section on Survey Research Methods JSM 2011*.

Berger, Y. G. and Escobar, E. L. (2013) Variance estimation of hot-deck imputed estimators of change for repeated rotating surveys. Southampton Statistical Sciences Research Institute.

Berger, Y. G. and Priam, R. (2010) Estimation of correlations between cross-sectional estimates from repeated surveys - an application to the variance of change. *Proceeding of the 2010 Symposium of Statistics Canada*.

Berger, Y. G. and Skinner, C. J. (2003) Variance estimation of a low-income proportion. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **52**, 457–468.

Birch, M. W. (1963) Maximum likelihood in three-way contingency tables. *J. R. Stat. Soc.*, **B**, 220–233.

Chao, M. T. (1982) A general purpose unequal probability sampling plan. *Biometrika*, **69**, 653–656.

Christine, M. and Rocher, T. (2012) Construction d'échantillons astreints à des conditions de recouvrement par rapport un échantillon antérieur et à des conditions d'équilibrage par rapport à des variables courantes. *Proceeding of the 10th Journée de Méthodologie Statistique de l'INSEE (Paris, 24-26 January 2012)*.

Demnati, A. and Rao, J. N. K. (2004) Linearization variance estimators for survey data. *Survey Methodology*, **30**, 17–26.

Deville, J. C. (1999) Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey Methodology*, **25**, 193–203.

Deville, J. C. and Särndal, C. E. (1992) Calibration estimators in survey sampling. *Journal of the American Statistical Association*, **87**, 376–382.

Di Meglio, E., Osier, G., Goedemé, T., Berger, Y. G. and Di Falco, E. (2013) *Standard Error Estimation in EU-SILC - First Results of the Net-SILC2 Project*. Brussels: Proceeding of the conference on New Techniques and Technologies for Statistics, Brussels. `http://www.cros-portal.eu/sites/default/files/NTTS2013fullPaper_144.pdf`.

Eurostat (2012a) European union statistics on income and living conditions (EU-SILC). `http://epp.eurostat.ec.europa.eu/portal/page/portal/microdata/eu_silc`.

Eurostat (2012b) People at risk of poverty or social exclusion (EU-SILC). `http://epp.eurostat.ec.europa.eu/portal/page/portal/product_details/dataset?p_product_code=T2020_50`.

Gambino, J. G. and Silva, P. L. N. (2009) Sampling and estimation in household surveys. *Handbook of Statistics: Design, Method and Applications: D. Pfeffermann and C.R. Rao.(editors). Elsevier*, **29A**, 407–439.

Goga, C., Deville, J. C. and Ruiz-Gazen, A. (2009) Use of functionals in linearization and composite estimation with application to two-sample survey data. *Biometrika*, **96**, 691–709.

Hájek, J. (1964) Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, **35**, 1491–1523.

Hansen, M., Hurwitz, W. and Madow, W. (1953) *Sample Survey Methods and Theory, volume I*. New York: John Wiley and Sons.

Hansen, M. H. and Hurwitz, W. N. (1943) On the theory of sampling from finite populations. *The Annals of Mathematical Statistics*, **14**, pp. 333–362.

Holmes, D. J. and Skinner, C. J. (2000) Variance estimation for labour force survey estimates of level and change. *Government Statistical Service Methodology Series*.

Horvitz, D. G. and Thompson, D. J. (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.

Kalton, G. (2009) Design for surveys over time. *Handbook of Statistics: Design, Method and Applications: D. Pfeffermann and C.R. Rao.(editors). Elsevier*, **29A**, 89–108.

Kish, L. (1965) *Survey Sampling*. Wiley.

Knottnerus, P. and van Delden, A. (2012) On variances of changes estimated from rotating panels and dynamic strata. *Survey Methodology*, **38**, 43–52.

Muennich, R. and Zins, S. (2011) Variance estimation for indicators of poverty and social exclusion. Work-package of the European project on Advanced Methodology for European Laeken Indicators (AMELI) `http://www.uni-trier.de/index.php?id=24676`.

Nordberg, L. (2000) On variance estimation for measures of change when samples are coordinated by the use of permanent random numbers. *Journal of Official Statistics*, **16**, 363–378.

Oguz-Alper, M. and Berger, Y. G. (2014) Variance estimation of change of poverty based upon the turkish EU-SILC survey. To appear in the Journal of Official Statistics.

Osier, G. (2009) Variance estimation for complex indicators of poverty and inequality using linearization techniques. *Survey Research Method*, **3**, 167–195.

Preston, I. (1995) Sampling distributions of relative poverty statistics. *Appl. Statist.*, **44**, 91–99.

Qualité, L. (2009) Unequal probability sampling and repeated surveys. PhD Thesis, University of Neuchatel, Switzerland.

Qualité, L. and Tillé, Y. (2008) Variance estimation of changes in repeated surveys and its application to the Swiss survey of value added. *Survey Methodology*, **34**, 173–181.

R Development Core Team (2014) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. `http://www.R-project.org`, Vienna, Austria.

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

Smith, P., Pont, M. and Jones, T. (2003) Developments in business survey methodology in the office for national statistics, 1994-2000. *Journal of the Royal Statistical Society. Series D (The Statistician)*, **52**, 257–295.

Tam, S. M. (1984) On covariances from overlapping samples. *American Statistician*, **38**, 288–289.

Wood, J. (2008) On the covariance between related Horvitz-Thompson estimators. *Journal of Official Statistics*, **24**, 53–78.

Wood, J. (2009) Variance estimation for price and volume indices in the UK Office for National Statistics. *Proceeding of the 57th session of International Statistical Institute, Durban*.