Abstract. Smith et al. (2003) recognised that assessing change is one of the most important challenges in survey statistics. 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. The estimation of covariances plays an important role in the estimation of the variance of a change. Berger and Priam (2010, 2013) proposed a simple approach based upon a multivariate linear regression model. The proposed estimator is not a model-based estimator. Indeed, the proposed estimator is design-consistent when the sampling fractions are negligible. The proposed estimator can accommodate stratified and two-stage sampling designs, and can be extended for complex estimators of change. 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. It allows to use a common approach for variance estimation for the different type of designs. The proposed approach does not rely on a specialised computer package and can be used with standard statistical software. The proposed approach can be used to produced variance estimates of change for the indicator of poverty and social exclusion (AROPE) based upon the European Union Statistics on Income and Living Conditions (EU-SILC) survey (Eurostat, 2012a). This poverty indicator is used to monitor change in poverty within the European Union (Eurostat, 2012a).


1 Introduction

The primary interest of many users is often in changes or trends from one time period to another. Suppose, we wish to estimate the absolute change

\[ \Delta = \tau_2 - \tau_1, \]
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 the variable of interest at wave 1 and 2. For simplicity, we assume that \( U \) is the same at wave 1 and 2. The estimator proposed in this paper can be also used when the population at wave 1 is different from the population at wave 2. In §5, we show that the proposed approach can be adapted for relative change or more complex measures of change. We adopt a design-based approach where the sampling distribution is specified by the sampling design. The change \( \Delta \) can be estimated by

\[
\hat{\Delta} = \hat{\tau}_2 - \hat{\tau}_1;
\]

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

\[
\hat{\tau}_1 = \sum_{i \in s_1} \frac{y_{1;i}}{\pi_{1;i}} \quad \text{and} \quad \hat{\tau}_2 = \sum_{i \in s_2} \frac{y_{2;i}}{\pi_{2;i}};
\]

where \( s_1 \) and \( s_2 \) denote respectively the first and second wave samples. 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. In this paper, \( s \) denotes the union of \( s_1 \) and \( s_2 \); that is, \( s = s_1 \cup s_2 \).

The design-based variance of the change \( \Delta \) is given by

\[
var(\Delta) = var(\hat{\tau}_1) + var(\hat{\tau}_2) - 2 \, cov(\hat{\tau}_1, \hat{\tau}_2)
\]

\[
= \nabla' \Sigma \nabla;
\]

where \( var(\hat{\tau}_1) \) and \( var(\hat{\tau}_2) \) denote respectively the design-based variances of \( \hat{\tau}_1 \) and \( \hat{\tau}_2 \). The quantity \( cov(\hat{\tau}_1, \hat{\tau}_2) \) denotes the covariance between \( \hat{\tau}_1 \) and \( \hat{\tau}_2 \) with respect to the sampling design. The matrix \( \Sigma \) is the design-based covariance matrix of the vector \( (\hat{\tau}_1, \hat{\tau}_2)' \) and \( \nabla = (-1, 1)' \).

Standard design-based estimators can be used to estimate the variances \( var(\hat{\tau}_1) \) and \( var(\hat{\tau}_2) \) (e.g. Wolter, 2007). The covariance \( cov(\hat{\tau}_1, \hat{\tau}_2) \) is the most difficult part to estimate because with overlapping samples, \( \hat{\tau}_1 \) and \( \hat{\tau}_2 \) are estimated from different samples. Several estimators have been proposed for the covariance in (2) (e.g. Kish, 1965; Tam, 1984; Nordberg, 2000; Holmes and Skinner, 2000; Berger, 2004b; 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 proposed by Berger (2004b) gives more accurate estimates than standard variance estimators (e.g. Tam, 1984; Qualité and Tillé, 2008) when we are interested in change within strata domains. The estimator proposed in this paper has the same property when the sampling fractions are small, because in §3.1, we show that the proposed estimator is approximately equivalent to the Berger (2004b) estimator, in this situation.

The proposed approach is based upon the residual matrix of a multivariate regression model. The multivariate regression is not a super-population approach, as it gives design-consistent covariance estimates. 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 (e.g. Wood, 2008; Muennich and Zins, 2011, p. 20).

In §2, we defined the class of rotating sampling designs considered. The proposed estimator for the covariance is defined in §3.1. In §4, we show how the proposed estimator can be extended to include stratification, multi-stage sampling and more complex measures of change. In §6, we support our result with a simulation study based upon the Italian EU-SILC survey. In §7, we show how the proposed approach can be used to estimate the variance of change of the EU-SILC indicator of poverty and social exclusion (AROPE).
2 Fixed sizes 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). The units sampled both on wave 1 and 2 usually represent a large fraction of the 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 sizes 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 $\pi_{1:i} = Pr\{i \in s_1\}$. 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 $\pi_{2:i} = E_1[\pi_{1:i|s_1}]$; where $E_1[\cdot]$ denotes the design expectation with respect to the first wave design. We consider that the sizes $n_1$, $n_2$ and $n_c$ are given quantities which are fixed (non random). 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 independent samples) or $g = 1$ (when we have two identical samples). The proposed approach is valid when $g = 0$ or 1, as it gives standards estimates in this situation.

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 (Muennich and Zins, 2011; Eurostat, 2012a) and the rotating design proposed by Tam (1984). Other examples can be found in Berger and Priam (2013) and in Christine and Rocher (2012).

3 Estimation of the covariance matrix

The estimation of the covariance matrix $\Sigma$ in (3) 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 ($g \leq 1$), because rotations are usually used in repeated surveys (e.g. Nordberg, 2000; Gambino and Silva, 2009; Kalton, 2009). Berger (2004b) showed that under the assumption of high entropy,

$$\hat{\Sigma} = \hat{\Sigma}_{\tau\tau} - \hat{\Sigma}_{\tau n} \hat{\Sigma}_{n n}^{-1} \hat{\Sigma}_{n \tau}^\prime,$$

is a consistent estimator for the covariance matrix $\Sigma$ (defined in (3)); with

$$\hat{\Sigma}_{\tau\tau} = \left( \begin{array}{cc} \sum_{i \in s} \bar{c}_{1:i} \bar{y}_{i:1}^2 & \sum_{i \in s} \bar{c}_{12:i} \bar{y}_{i:1} \bar{y}_{i:2} \\ \sum_{i \in s} \bar{c}_{12:i} \bar{y}_{i:1} \bar{y}_{i:2} & \sum_{i \in s} \bar{c}_{2:i} \bar{y}_{i:2}^2 \end{array} \right),$$

$$\hat{\Sigma}_{nn} = \left( \begin{array}{cc} \sum_{i \in s} \bar{c}_{1:i} z_{1:i} & \sum_{i \in s} \bar{c}_{12:i} z_{1:i} z_{2:i} \\ \sum_{i \in s} \bar{c}_{12:i} z_{1:i} z_{2:i} & \sum_{i \in s} \bar{c}_{2:i} z_{2:i}^2 \end{array} \right),$$

$$\hat{\Sigma}_{\tau n} = \left( \begin{array}{cc} \sum_{i \in s} \bar{c}_{1:i} \bar{y}_{i:1} z_{1:i} & \sum_{i \in s} \bar{c}_{12:i} \bar{y}_{i:1} z_{2:i} \\ \sum_{i \in s} \bar{c}_{12:i} \bar{y}_{i:1} z_{2:i} & \sum_{i \in s} \bar{c}_{2:i} \bar{y}_{i:2} z_{2:i} \end{array} \right),$$

where $s = s_1 \cup s_2$ denotes the overall sample. The quantities $\bar{c}_{\ell:i}$, $\bar{c}_i$ and $\bar{c}_{12:i}$ are finite population correction given by $\bar{c}_{\ell:i} = (1 - \pi_{\ell:i})$, $\bar{c}_i = (1 - \pi_{c:i})$ and $\bar{c}_{12:i} = (\pi_{c:i} - \pi_{1:i} \pi_{2:i})/\pi_{c:i}$; with $\pi_{c:i} = \cdots$
where \( pr\{i \in s_c\} \) denotes the probability with respect to the design. We consider that \( \bar{c}_{12;1} = 1 \) when \( \pi_{c_{11}} = 0 \). The variables \( z_{1;i} \) and \( z_{2;i} \) are design variables defined by

\[
z_{1;i} = \delta\{i \in s_1\}, \quad \text{and} \quad z_{2;i} = \delta\{i \in s_2\}.
\]  

The function \( \delta\{A\} \) is the indicator function which is equal to one when \( A \) is true and zero otherwise. The variables \( \bar{y}_{1;i} \) and \( \bar{y}_{2;i} \) are defined by

\[
\bar{y}_{i;1} = y_{1;i} \pi_{1;i}^{-1} \delta\{i \in s_1\} \quad \text{and} \quad \bar{y}_{i;2} = y_{2;i} \pi_{2;i}^{-1} \delta\{i \in s_2\};
\]  

with \( \bar{y}_{i;\ell} = 0 \) when \( i \notin s_\ell \).

The estimator (4) is not easy to implement, because it requires a specialised package (Berger, 2004a; Andersson et al., 2011a, b). Berger and Priam (2013) showed that when the sampling fractions are negligible, the estimator (4) can be easily approximated using a multivariate linear regression approach described in the following §.

### 3.1 Estimation of the covariance using a multivariate regression approach

Berger and Priam (2010) proposed to use the residual covariance of the following multivariate (or general) linear regression model (see also Berger and Priam, 2013).

\[
\begin{pmatrix}
\bar{y}_{1;i} \\
\bar{y}_{2;i}
\end{pmatrix} = \begin{pmatrix}
\beta_{1}(1) z_{1;i} + \beta_{2}(1) z_{2;i} + \beta_{12}(1) z_{1;i} z_{2;i} \\
\beta_{1}(2) z_{1;i} + \beta_{2}(2) z_{2;i} + \beta_{12}(2) z_{1;i} z_{2;i}
\end{pmatrix} + \epsilon_{i};
\]  

where \( i \in s = s_1 \cup s_2 \) and the residuals \( \epsilon_{i} \) have a bivariate distribution with mean 0 and an unknown residual variance-covariance matrix \( V \). It is important to point out that as the distribution of \( \epsilon_{i} \) 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 (10) are given by (9). The covariates \( z_{1;i} \) and \( z_{2;i} \) are design variables defined by (8). Note the absence of intercept and the presence of an interaction in the regression (10).

Matrix notations can be used to define the model (10) in a more convenient way. Let \( \beta = (\beta^{(1)}, \beta^{(2)}) \) be the \( 3 \times 2 \) matrix of parameters, where \( \beta^{(1)} = (\beta_{11}^{(1)}, \beta_{21}^{(1)}, \beta_{12}^{(1)}) \) and \( \beta^{(2)} = (\beta_{11}^{(2)}, \beta_{21}^{(2)}, \beta_{12}^{(2)}) \) are parameters of the model (10). The model (10) can be re-written as

\[
\bar{Y}_{s} = Z_{s} \beta + \epsilon;
\]  

where \( \epsilon = (\epsilon_{1}, \ldots, \epsilon_{n})' \). The quantities \( \bar{Y}_{s} \) and \( Z_{s} \) are respectively defined by the following \( n \times 2 \) matrix \( \bar{Y}_{s} \) and the \( n \times 3 \) matrix \( Z_{s} \).

\[
\bar{Y}_{s} = \begin{pmatrix}
\bar{y}_{1} \\
\bar{y}_{2}
\end{pmatrix},
\]  

\[
Z_{s} = \begin{pmatrix}
z_{1} \\
z_{2} \\
z_{c}
\end{pmatrix};
\]  

where

\[
\begin{align*}
\bar{y}_{\ell} & = (\bar{y}_{\ell;1}, \bar{y}_{\ell;2}, \ldots, \bar{y}_{\ell;n})', \\
z_{\ell} & = (z_{\ell;1}, z_{\ell;2}, \ldots, z_{\ell;n})', \\
z_{c} & = (z_{1;1}, z_{1;2}, z_{1;1} z_{2;2}, \ldots, z_{1;n} z_{2;n})'.
\end{align*}
\]
The model (11) is also a multivariate analysis of variance (MANOVA) model, as the covariates are all dummy variables.

The proposed estimator for the variance of change (2) is given by the following substitution estimator.

\[
\hat{\text{var}}(\hat{\Delta}) = \hat{\text{var}}(\hat{\tau}_1) + \hat{\text{var}}(\hat{\tau}_2) - 2 \left[ \hat{\text{var}}(\hat{\tau}_1)\hat{\text{var}}(\hat{\tau}_2) \right]^{1/2} \hat{\rho};
\]

where \(\hat{\text{var}}(\hat{\tau}_1)\) and \(\hat{\text{var}}(\hat{\tau}_2)\) denote respectively standard design-based variance estimators of \(\hat{\tau}_1\) and \(\hat{\tau}_2\) (e.g. Wolter, 2007) and

\[
\hat{\rho} = \hat{V}_{12} \left( \hat{V}_{11} \hat{V}_{22} \right)^{-1/2};
\]

where \(\hat{V}_{k\ell}\) is the component \((k,\ell)\) of the ordinary least squares residual matrix \(\hat{V}\) of the model (11). The matrix \(\hat{V}\) is an estimator of \(V\). Berger and Priam (2013) showed that \(\hat{V} \approx \alpha^{-1} \Sigma\), when \(\hat{c}_{\ell,i} \approx 1, \hat{c}_i \approx 1\) and \(\hat{c}_{12;i} \approx 1\),

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 \(Z_s\). This implies that (16) is design-consistent, because (4) is design-consistent. The assumptions (18) hold when sampling fractions are negligible (see Example 1 in Berger and Priam (2013)). Note that \(\hat{g}\) can be large even when the sampling fractions are negligible; that is, the assumptions (18) may hold even when \(\hat{g}\) is large.

Berger and Priam (2013) showed that \(\alpha \hat{V}_{12}\) is the standard estimator of covariance under pps with replacement sampling. Berger and Priam (2013) also showed that \(\alpha \hat{V}_{12}\) reduces to Tam (1984) estimator under equal probabilities and when the fraction \(n_1 n_2 / N n_c = \pi_2;g^{-1}\) is negligible and when \(n_c\) is sufficiently large.

It is preferable to have an estimator of covariance based upon a correlation (see (16) and (17)) to avoid negative variance estimates. Berger and Priam (2013) showed that (16) is always positive because the residual variance-covariance matrix is non-negative definit.

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, we are conditioning on variables which have their totals fixed by design. Note that there is a clear analogy between Birch (1963) 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 (13) which are used as covariates. The interactions (see (15)) take account of the rotation of the design. The weighted variables of interest (see (12)) measured at each waves are used as response variables. The proposed estimator takes into account of all the data, as it utilises the data of the units from the common sample and the units that rotate in and out.

The proposed estimator is easier to implement than Nordberg (2000), Wood (2008), Goga et al. (2009) and Muennich and Zins (2011, p. 20) estimators 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 matrix $\hat{V}$ can be easily calculated, as the multivariate regression (11) can be easily fitted by most statistical software. It is only necessary to create the variables $\hat{y}_{i;1}, \hat{y}_{i;2}, z_{1;i}$ and $z_{2;i}$. For example, the SAS procedure REG can be used to fit the multivariate regression. The multivariate regression can be also fitted using the GLM Multivariate procedure in SPSS. With Stata, the output $e(Sigma)$ of the function mvreg() gives the estimates $\hat{V}$. With the statistical software R (R Development Core Team, 2006), the command estVar(lm(formula=\text{Y}$~\text{-1+Z1*Z2}$)) gives the estimate $\hat{V}$; where $Z1$ and $Z2$ denote the $n \times 1$ vectors $z_1$ and $z_2$ (see (14)) and $Y$ is the matrix $Y_i$ given by (12). Note that Berger (2005) showed that $\hat{\text{var}}(\hat{\tau}_1)$ and $\hat{\text{var}}(\hat{\tau}_2)$ can also be calculated using a regression approach (see also Deville and Tillé, 2005).

4 Extensions to stratified and two-stage sampling design

The proposed estimator can be easily extended to accommodate stratification. Suppose that we have $H$ strata $U_1, U_2, \ldots, U_H$ 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 sizes rotating design (see §2) is implemented with 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 specifies the stratification.

The multivariate regression model is still given by (11) with the same response variables $\hat{Y}_i$ defined by (12) with $\hat{y}_{i;\ell}$ is given by (9). However, the matrix $Z_s$ is different as it now contains the stratification variables $z_{1h;i}$ and $z_{2h;i}$ and suitable interactions. 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 $Z_s$. The ordinary least squares estimate of residuals variance-covariance of model (11) is used to estimate the covariance between $\hat{\tau}_1$ and $\hat{\tau}_2$ (see (16)).

Berger and Priam (2013) showed that $\hat{S}_{12} = \sum_{h=1}^H \hat{S}_{12h}$ and $\hat{S}_{\ell\ell} = \sum_{h=1}^H \hat{S}_{\ell\ell h}$ where $\hat{S}_{12h}$ (resp. $\hat{S}_{\ell\ell h}$) denotes the within stratum covariance (resp. variance). Note that $\hat{S}_{12}$ and $\hat{S}_{\ell\ell}$ are natural estimators of covariance and variances under stratified designs. Consequently, the proposed estimator for the covariance is consistent when the assumptions (18) hold within each strata and when the number of strata $H$ is asymptotically bounded.

Note that the proposed estimator accommodates dynamic stratification. In other words, the strata at wave 1 can be different from the strata at wave 2; i.e. new strata can be created, and units can move between strata. For example, a unit may belong to stratum $U_1$ at wave 1 and to stratum $U_2$ at wave 2.

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. As we assume that the sampling fractions are negligible, we suggest using an ultimate cluster strategy to estimate the covariance; that is, we consider that the variance between PSUs captures most of the variability of the estimator of change. This usually holds in practice (e.g. Särndal et al., 1992, §4.3.2). Berger and Priam (2013) showed how the approach described in §3.1 can be adapted under an ultimate cluster strategy.

5 Extension to more complex measure of change

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

$$\hat{\Delta}_g = f(\hat{\tau}) \tag{19}$$

where $\hat{\tau} = (\hat{\tau}_1, \hat{\tau}_2, \ldots, \hat{\tau}_p)'$ and $P$ is the number of totals. The quantity $\hat{\tau}_p$ is the Horvitz and Thompson (1952) estimator $\hat{\tau}_p = \sum_{i \in s_i} y_{pi} / \pi_{\ell,i}$ of a variable $y_{pi}$; where $p = 1, \ldots, Q, Q + 1, \ldots, P$, $\ell = 1$ if $p \leq Q$ and $\ell = 2$ if $p > Q$. The constant $Q$ is the number of totals calculated from the first wave.

Suppose that $\hat{\Delta}_g$ is an approximately unbiased estimator of $\Delta_g = f(\tau)$; where $\tau = E(\hat{\tau})$. Using the delta method (Taylor linearisation), we have that an approximation of $\Delta_g$ in the neighbourhood of $\tau$ is given by $\hat{\Delta}_g - \Delta_g \approx \nabla(\tau)'(\hat{\tau} - \tilde{\tau})$; where $\nabla(\tau)$ is the gradient of $f(\tau)$ at $\tau$. Therefore, the linearisation estimator for the variance is

$$\hat{\text{var}}(\hat{\Delta}_g) = \nabla(\hat{\tau})' \hat{\text{var}}(\hat{\tau}) \nabla(\hat{\tau}). \tag{20}$$

Note that (20) is an estimator of the mean square error. However, this estimator is also a consistent estimator for the variance, because $\hat{\Delta}_g$ is asymptotically unbiased (Wolter, 2007, p. 232). Berger and Priam (2013) showed how the approach described in §3.1 can be used to estimate the covariance matrix $\hat{\text{cov}}(\hat{\tau})$.

Another approach consists in substituting $y_{\ell,i}$ by linearised variables in (1) (e.g. Deville, 1999; Demnati and Rao, 2004), and using the estimator of covariance (16) where $\hat{\tau}_\ell$ is the Horvitz and Thompson (1952) sum of the linearised variable at wave $\ell$. For example, this approach is recommended when $\tilde{\theta}_2$ and $\tilde{\theta}_1$ are not functions of totals. For example, when $\tilde{\theta}_2$ and $\tilde{\theta}_1$ are Gini coefficients. It can be shown that this approach is equivalent to (20) when $\hat{\theta}_\ell$ is a smooth function of totals computed only from $s_\ell$. This would not be the case if $\hat{\theta}_\ell$ depends on totals computed from $s_1$ and $s_2$. For example, if $\tilde{\theta}_1 = \hat{\tau}_1 / (\hat{\tau}_2 + \hat{\tau}_1)$ and $\tilde{\theta}_2 = \hat{\tau}_2 / (\hat{\tau}_2 + \hat{\tau}_1)$, where $\hat{\tau}_1$ and $\hat{\tau}_2$ are given by (1). In this case, the linearised variable approach gives a biased estimator for the variance and the estimator (20) is still approximately unbiased.

### 6 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 strata domains, the estimator proposed by Berger (2004b) is more accurate than standard estimators of variance of change (e.g. Tam, 1984; Qualité and Tillé, 2008). Therefore, based on Andersson et al. (2011b) simulation studies, the estimator proposed by Berger (2004b) is recommended when we are interested in change within strata domains. The estimator proposed in this paper has the same property, as it reduces to the Berger (2004b) estimator when the sampling fractions are small.

In this §, we report the results of a series of simulation based upon the Italian EU-SILC survey. For each simulation, 1000 samples were selected to compute the empirical relative bias (RB)

$$\text{RB} = \frac{E[\hat{\text{var}}(\hat{\Delta})] - \text{var}(\hat{\Delta})}{\text{var}(\hat{\Delta})} \%$$

and the empirical relative root mean square error (RRMSE)

$$\text{RRMSE} = \frac{\text{mse}[\hat{\text{var}}(\hat{\Delta})]^{1/2}}{\text{var}(\hat{\Delta})} \%.$$
Table 1: Observed RB and RRMSE for several variables of interest and several domain of interest. Italian EU-SILC data (2008, 2009). The column ”Prop.” gives the results for the proposed approach. The columns ”Corr. (23)” and ”Corr. (22)” give the results for the approaches based respectively on the correlations (23) and (22).

<table>
<thead>
<tr>
<th>Variables of interest</th>
<th>Domains of interest</th>
<th>RB (%)</th>
<th>RRMSE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afford holiday</td>
<td>Population</td>
<td>-4</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>Detached</td>
<td>-6</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Semi-detached</td>
<td>-4</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Home owner</td>
<td>-4</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>Males</td>
<td>-5</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>Females</td>
<td>-5</td>
<td>16</td>
</tr>
<tr>
<td>Own a car</td>
<td>Population</td>
<td>-9</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Detached</td>
<td>0</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>Semi-detached</td>
<td>-3</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Home owner</td>
<td>-6</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>Males</td>
<td>-5</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>Females</td>
<td>-10</td>
<td>9</td>
</tr>
<tr>
<td>Equivalised disposable income</td>
<td>Population</td>
<td>-3</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>Detached</td>
<td>-2</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Semi-detached</td>
<td>-5</td>
<td>-6</td>
</tr>
<tr>
<td></td>
<td>Home owner</td>
<td>-5</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Males</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>Females</td>
<td>-7</td>
<td>9</td>
</tr>
<tr>
<td>At risk of poverty</td>
<td>Population</td>
<td>-9</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>Detached</td>
<td>-3</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>Semi-detached</td>
<td>0</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>Home owner</td>
<td>-4</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>Males</td>
<td>-9</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>Females</td>
<td>-5</td>
<td>21</td>
</tr>
</tbody>
</table>

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

The Chao (1982) unequal probability design is used to select samples. For wave 1, a sample is selected with inclusion probabilities \( \pi_{1,i} \). For wave 2, we use the sampling design described in the Example 1 of Berger and Priam (2013). This rotating design belongs to the class of designs described in §2 and is such that \( \pi_{1,i} \approx \pi_{2,i} \). The statistical software \( \textbf{R} \) is used to fit the multivariate regression model. The variances \( \text{var}(\hat{\tau}_1) \) and \( \text{var}(\hat{\tau}_2) \) are estimated by the Hájek (1964) variance estimator.
The common sample of waves 2008 and 2009 of the Italian Statistics on Income and Living Conditions (EU-SILC) survey is treated as a population from which stratified samples are selected. This gives a population size \( N = 19644 \). Stratified samples of size \( n_1 = n_2 = 982 \) are selected using the uni-stage Chao (1982) sampling design. The strata are five geographical regions. We consider \( g = 75\% \).

We consider the change between the mean (or proportion) of several variables of interest. We consider three dummy variables of interest (afford holiday, own a car, at risk of poverty) and one quantitative variables (equivalised disposable income). The change between the means (or proportion) is estimated by

\[
\hat{\tau}_2 \frac{N}{N-2} - \hat{\tau}_1 \frac{N}{N-1}.
\]

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 (\( y_{\ell,i} = 0 \) if \( i \) does not belong to the domain of interest).

Berger and Priam (2013) showed that the proposed estimator is equivalent to the variance estimator proposed by Qualité (2009, p. 83) and Muennich and Zins (2011, p. 20). Thus, the proposed estimator will not be more accurate than these estimators. We propose to compare (16) with the estimator (21) which consists in using the same variance estimates as in (16) with a naïve correlation (see (22) below) based on Tam (1984) estimator for the covariance (under equal probability sampling).

\[
\hat{\text{var}}(\hat{\Delta})^* = \hat{\text{var}}(\hat{\tau}_1) + \hat{\text{var}}(\hat{\tau}_2) - 2 \hat{\rho}^* \left[ \hat{\text{var}}(\hat{\tau}_1) \hat{\text{var}}(\hat{\tau}_2) \right]^{1/2};
\]

with

\[
\hat{\rho}^* = \hat{\text{cov}}(\hat{\tau}_1, \hat{\tau}_2)_{Tam} \left[ \hat{\text{var}}(\hat{\tau}_1)_{SRS} \hat{\text{var}}(\hat{\tau}_2)_{SRS} \right]^{-1/2}
\]

and

\[
\hat{\text{var}}(\hat{\tau}_\ell)_{SRS} = N^2 \left( 1 - \frac{n_\ell}{N} \right) \hat{\sigma}_{\ell \ell},
\]

where \( \hat{\text{cov}}(\hat{\tau}_1, \hat{\tau}_2)_{Tam} \) is the covariance proposed by Tam (1984) and \( \hat{\sigma}_{\ell \ell} \) is defined by

\[
\hat{\sigma}_{\ell \ell} = \frac{1}{n_\ell} \sum_{i \in s_\ell} (\hat{y}_{\ell,i} - \bar{y}_\ell)^2 .
\]

We also compare the proposed estimator with another estimator which consists in using (21) with a correlation based upon a stratified equal probability sampling design (e.g. Qualité, 2009, p. 79, Muennich and Zins, 2011, p. 26). This correlation is given by

\[
\hat{\rho}_{st}^* = \hat{\text{cov}}(\hat{\tau}_1, \hat{\tau}_2)_{st} \left[ \hat{\text{var}}(\hat{\tau}_1)_{st} \hat{\text{var}}(\hat{\tau}_2)_{st} \right]^{-1/2};
\]

where

\[
\hat{\text{cov}}(\hat{\tau}_1, \hat{\tau}_2)_{st} = \sum_{h=1}^{H} \sum_{i \in s_{ch}} \left( 1 - \frac{n_{1h} n_{2h}}{N_h n_{ch}} \right) \frac{n_{ch}}{n_{ch} - 1} \frac{N_h^2 n_{ch} \hat{\sigma}_{ch}}{n_{1h} n_{2h} \hat{\sigma}_{ch}} ,
\]

\[
\hat{\text{var}}(\hat{\tau}_\ell)_{st} = \sum_{h=1}^{H} \sum_{i \in s_{ch}} \left( 1 - \frac{n_{\ell h}}{N_h} \right) N_h^2 \hat{\sigma}_{\ell \ell}^2 .
\]
\[
\hat{\sigma}_{ch} = \frac{1}{n_{ch}} \sum_{i \in s_{ch}} (y_{1;i} - \bar{y}_{1;ch}) (y_{2;i} - \bar{y}_{2;ch}),
\]
\[
\hat{\sigma}_{\ell h}^2 = \frac{1}{n_{\ell h}} \sum_{i \in s_{\ell h}} (y_{\ell;i} - \bar{y}_{\ell h})^2.
\]

The quantities \( \bar{y}_{1;ch}, \bar{y}_{2;ch} \) 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 \( \ell \).

The result of this simulation is given in Table 1. The observed bias of the proposed estimator is usually negligible. We observe that the estimator based upon (23) may have non negligible biases. This is due to the fact that this estimator overestimates the covariance, as it does not take the inclusion probabilities into account. However, the estimator based upon (22) may have negligible biases, but not for all the variables and domains considered. In most of the cases, we observed smaller RRMSEs for the proposed approach. The large values for the RRMSEs is due to the fact that the variances \( \text{var}(\Delta) \) can be very small.

7 AN APPLICATION TO THE EU-SILC HOUSEHOLD SURVEYS

Suppose we are interested in the change of the at-risk-of-poverty or social exclusion (AROPE) indicator between two consecutive years (2009 and 2010). The AROPE indicator is a key indicator used to monitor poverty within the European Union (Eurostat, 2012b; Atkinson and Marlier, 2010). 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. In this §, we show briefly how to estimate the variance of the net change of the AROPE indicator. The computations were made in SAS by Guillaume Osier (Statistics Luxembourg, STATEC and Luxembourg Income Study), Emilio Di Meglio (Eurostat Unit F4 Quality of Life) and Emanuela Di Falco (Eurostat Unit F4 Quality of Life). The EU-SILC production data base was used.

An ultimate cluster approach (see §4) 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

\[
\hat{y}_{\ell;i} = \delta\{i \in s_{\ell}\} \sum_{j \in PSU_i} w_{\ell;j} y_{\ell;j},
\]

where \( s_{\ell} \) is the sample of PSUs at wave \( \ell \), \( PSU_i \) 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 \( \ell \). The variable \( z_{\ell h;i} \) are dummy variable which specifies the stratification at PSU level. The variables \( \hat{y}_{\ell;i} \) and \( z_{\ell h;i} \) need to be defined for all \( i \in s = s_1 \cup s_2 \). A more detailed description of the computation can be found in Berger et al. (2012); Di Meglio et al. (2013).
Table 2: 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%.

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

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 \((\text{Osier}, 2009)\). Berger and Oguz Alper \((2013)\) 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 variances \((\text{Berger and Skinner}, 2003)\). Therefore, 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 (or exposure if we are interested in a domain). Hence \(\hat{\tau}\) in \((19)\) is a vector of four totals. The approach of \(\S 5\) is used. The effect of calibration can be taken into account, by replacing the response variables by residuals \((\text{Deville and Särndal}, 1992)\). 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 they 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. In this case, the proposed approach can still be used \((\text{see end of \(\S 4\)})\).

The estimates based on the proposed approach are given in Table 2. 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 quality of the design variables. These estimates are likely to overestimate the variance because the
Proceedings of the Septième Colloques Francophone sur les sondages (2012)

Effect of calibration adjustment was not taken into account. This effect may be more pronounced for Scandinavian countries. One of the aim of the Net-SILC2 project (Atkinson and Marlier, 2010) is to provide recommendations regarding the quality of the design variables. These recommendations can be found in Goedeme (2010).

8 Discussion

In this paper, we propose a novel variance estimator for change under complex rotating sampling designs. We show that the proposed variance estimator gives a design-consistent estimator for the variance when the finite population corrections are negligible, and always gives positive variance estimates.

The advantage of the proposed estimator is its generality, simplicity and flexibility. The proposed estimator can be used for a large class of complex sampling designs, involving unequal probabilities, stratification and two-stage sampling. It can be also extended for complex measure of changes. The proposed estimator is simple to implement, as it is based upon a multivariate (general) linear regression technique which can be easily implement with standard statistical software. This regression technique requires the creation of design variables which are used as covariates. Interactions take account of the rotation of the design. The weighted variables of interest measured at each waves are used as response variables. The proposed estimator takes into account of all the data, as it utilises the data of the units from the common sample and the units that rotates in and out.

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). I am also grateful to Guillaume Osier (Statistics Luxembourg, STATEC and Luxembourg Income Study), 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. I am also grateful to Dr. Rodolphe Priam, Dr. Emilio López Escobar (Instituto Tecnológico Autónomo de México, México), Melike Oguz Alper (University of Southampton) for helpful comments.

References


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


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


