On aligned composite estimates from overlapping samples for growth rates and totals

Paul Knottnerus

The views expressed in this paper are those of the author(s)
and do not necessarily reflect the policies of Statistics Netherlands

Discussion paper (201216)

Statistics Netherlands

The Hague/Heerlen, 2012

Explanation of symbols

.	data not available
*****	provisional figure
******	revised provisional figure (but not definite)
x	publication prohibited (confidential figure)
–	nil
–	(between two figures) inclusive
0 (0.0)	less than half of unit concerned
empty cell	not applicable
2011–2012	2011 to 2012 inclusive
2011/2012	average for 2011 up to and including 2012
2011/’12	crop year, financial year, school year etc. beginning in 2011 and ending in 2012
2009/’10– 2011/’12	crop year, financial year, etc. 2009/’10 to 2011/’12 inclusive

Due to rounding, some totals may not correspond with the sum of the separate figures.

Publisher
Statistics Netherlands
Henri Faasdreef 312
2492 JP The Hague

Prepress
Statistics Netherlands
Grafimedia

Cover
Teldesign, Rotterdam

Information
Telephone +31 88 570 70 70
Telefax +31 70 337 59 94
Via contact form:
www.cbs.nl/information

Where to order
E-mail: verkoop@cbs.nl
Telefax +31 45 570 62 68

Internet
www.cbs.nl

ISSN: 1572-0314

60083201216 X-10

On aligned composite estimates from overlapping samples for growth rates and totals

Paul Knottnerus

Summary: When monthly business surveys are not completely overlapping, there are two different estimators for the monthly growth rate of the turnover: (i) one that is based on the monthly estimated population totals and (ii) one that is purely based on enterprises observed on both occasions in the overlap of the corresponding surveys. The resulting estimates and variances might be quite different. This paper proposes an optimal composite estimator for the growth rate as well as the population totals.

Keywords: business surveys, coefficient of variation, general restriction estimator, Kalman equations, panels, variances.

1. Introduction

In many countries a monthly business survey is held for the major Standard Industrial Classification (SIC) codes to estimate the level of the monthly turnover and the change in that level compared to a month or a year ago. When repeatedly sampling a population, a complicating factor is that there are various methods for estimating the (relative) change from a panel with different outcomes especially when the samples on different occasions are not completely overlapping.

Kish (1965), Tam (1984), Laniel (1988), Hidiroglou, Särndal and Binder (1995), Nordberg (2000), Berger (2004), Qualité and Tillé (2008), Wood (2008) and Knottnerus and Van Delden (2012) examined various estimators for the parameter of change in different situations. The main aim of this paper is to derive estimators for a relative change as well as the corresponding population totals that are in line with each other and that have minimum variance property. The derivation of the aligned composite estimators is based on the general restriction (GR) estimator of Knottnerus (2003). Composite estimators for totals and (absolute) changes are also proposed by Särndal et al. (1992, pages 370-378) but in separate steps. Moreover, this paper focuses on estimators for growth rates because (i) users of figures from business surveys for a specific SIC code often are more interested in growth rates than in absolute changes, (ii) growth rates are needed for making an overall index for the (monthly) turnover for each of the major SIC codes, and (iii) in practice there might be model-assisted reasons to look at growth rates.

The outline of the paper is as follows. Section 2 briefly describes two methods for estimating a growth rate of the total turnover for enterprises with a certain SIC code. A small example illustrates the substantial differences between the two approaches. Section 3 discusses the question of which estimation method is to be preferred and explains as to why the difference between the variances of both estimators might be so large. For various situations sections 4 and 5 propose an optimal composite estimator.

2. Two estimators for the growth rate of the total turnover

Consider a population of *N* enterprises $U = \{1, ..., N\}$, and suppose there are no births and deaths in the population. Let Y_i denote the value of the turnover for the *i*th enterprise in a given month (say *t*) and X_i the value of the turnover of that enterprise in month *t*-12. Hence, the variables *y* and *x* concern the same variable on two different occasions. Denote their population totals by *Y* and *X,* and their population means by $\overline{Y}$ and $\overline{X}$, respectively. That is, $Y = \Sigma_{i \in U} Y_i$, $X = \Sigma_{i \in U} X_i$, $\overline{Y} = Y / N$ and $\overline{X} = X / N$. Let s_1, s_2 and s_3 denote three mutually disjoint simple random samples from *U* without replacement (SRS). Define s_{12} and s_{23} by $s_{12} = s_1 \cup s_2$ and $s_{23} = s_2 \cup s_3$, respectively. Denote the size of s_k by n_k $(k = 1, 2, 3, 12, 23)$ and the corresponding sample means by $\overline{y}_k$ and $\overline{x}_k$. Let the variable *x* be observed in s_{12} on the first occasion and the variable *y* in s_{23} on the second occasion. It is assumed that $n_{12} = n_{23} = n$. That is, all n_1 units in subsample s_1 on the first occasion are replaced by the n_3 $(= n_1)$ units of subsample s_3 on the second occasion 12 months later. Denote the overlap ratio by λ $(= n_2 / n)$ and the sampling fraction by *f* $(= n / N)$. The SRS estimators for the population totals *Y* and *X* are defined by $\hat{Y}_{SRS} = N\overline{y}_{23}$ and $\hat{X}_{SRS} = N\overline{x}_{12}$, respectively.

Define the growth rate *g* of the total turnover between the two occasions by $g = G - 1$ with *G*=*Y/X*. For estimating *G* there are two options. One of the standard (STN) options is based on the estimated totals on both occasions, that is

$$\hat{G}_{STN} = \frac{\hat{Y}_{SRS}}{\hat{X}_{SRS}} = \frac{\overline{y}_{23}}{\overline{x}_{12}} \quad ; \tag{1}$$

see Nordberg (2000), Qualité and Tillé (2008), and Knottnerus and Van Delden (2012). Note that the estimator $\hat{g}_{STN} = \hat{G}_{STN} - 1$ for *g* has the same variance as $\hat{G}_{STN}$. For sufficiently large *n* this variance can be approximated by using a first-order Taylor series expansion of $\hat{G}_{STN}$. That is,

$$\begin{aligned}\mathrm{var}(\hat{G}_{STN}) &\approx \frac{1}{\overline{X}^2}\mathrm{var}(\overline{y}_{23} - G\overline{x}_{12}) \\ &= \frac{1}{\overline{X}^2}\{\mathrm{var}(\overline{y}_{23}) + G^2\,\mathrm{var}(\overline{x}_{12}) - 2G\,\mathrm{cov}(\overline{y}_{23}, \overline{x}_{12})\} \\ &= \frac{1}{\overline{X}^2}\{(\frac{1}{n} - \frac{1}{N})(S_y^2 + G^2 S_x^2) - 2G(\frac{\lambda}{n} - \frac{1}{N})S_{xy}\}, \qquad (2)\end{aligned}$$

where $S_y^2 = \Sigma_U (Y_i - \overline{Y})^2 / (N-1)$ is the adjusted population variance of the Y_i and S_x^2 that of the X_i while $S_{xy} = \Sigma_U (X_i - \overline{X})(Y_i - \overline{Y}) / (N-1)$ is their adjusted population covariance. For (different) derivations of the expression for $\mathrm{cov}(\overline{y}_{23}, \overline{x}_{12})$ used in (2), see Tam (1984) and Knottnerus and Van Delden (2012). The adjusted population (co)variances can be estimated unbiasedly by the sample (co)variances; recall sample (co)variances s_{yk}^2 and s_{yxk} from sample s_k are defined by

$$\begin{aligned}s_{yk}^2 &= \frac{1}{n_k - 1}\sum_{i \in s_k}(Y_i - \overline{y}_k)^2 \qquad (k = 1, 2, 3, 12, 23) \\ s_{yxk} &= \frac{1}{n_k - 1}\sum_{i \in s_k}(Y_i - \overline{y}_k)(X_i - \overline{x}_k).\end{aligned}$$

An alternative option for estimating *G* and *g* is based on enterprises observed on both occasions in overlap s_2 (OLP). That is,

$$\hat{G}_{OLP} = \frac{\overline{y}_2}{\overline{x}_2}. \qquad (3)$$

For sufficiently large n_2 the well-known approximation for the variance of this estimator is

$$\begin{aligned}\mathrm{var}(\hat{G}_{OLP}) &\approx \frac{1}{\overline{X}^2}\mathrm{var}(\overline{y}_2 - G\overline{x}_2) \\ &= \frac{1}{\overline{X}^2}(\frac{1}{n_2} - \frac{1}{N})S_{y-Gx}^2, \qquad (4)\end{aligned}$$

where S_{y-Gx}^2 stands for $S_y^2 + G^2 S_x^2 - 2GS_{xy}$. In order to get some more insight into the merits of both $\hat{g}_{STN}$ and $\hat{g}_{OLP}$, consider the following example.

Example 2.1. The data in Table 1 are observations on the turnover of supermarkets and are taken from Table 2 in Knottnerus and Van Delden (2012). After some calculations we obtain

$$\bar{y}_{23} = 596\,,\ \bar{x}_{12} = 579\,,\ s^2_{y23} = 155505\ ,\ s^2_{x12} = 168666 \quad \text{and} \quad \hat{\rho}_{xy2} = 0.9997\ .$$

The population correlation coefficient ρ_{xy} $(= S_{xy} / S_x S_y)$ between the Y_i and the X_i is estimated from the overlap by $\hat{\rho}_{xy2} = s_{xy2} / s_{y2} s_{x2}$. To avoid negative variance estimates, the authors proposed estimating S_{xy} by $\hat{S}_{xy} = \hat{\rho}_{xy2} s_{x12} s_{y23} = 161901$.

Table 1. Panel data from a population with *N*=50

month	turnover per unit (in thousand euros)				
	1	2	3	4	5
t-12 (X_i)	380.0	493.9	264.3	1179.1	
t (Y_i)		472.0	267.0	1169.0	475.3

Substituting the above outcomes into (1) and (2), we obtain $\hat{g}_{STN} = 0.028$ (=2.8%) and $\hat{\text{var}}(\hat{g}_{STN}) = 0.063$. Assuming normality and using Student's *t*-value $t_3 = 3.18$, we get a fairly large (unrealistic) 95%-confidence interval I^{95}_{STN} = (-77%; 83%); such an interval can be called unrealistic because growth rates of more than 80% are never observed among supermarkets in that period. In contrast, from overlap s_2 we get the results

$$\bar{y}_2 = 636\,,\ \bar{x}_2 = 646\,,\ s^2_{y2} = 223573\ ,\ s^2_{x2} = 226512 \quad \text{and} \quad s_{xy2} = 224967\ .$$

Substituting these estimates into (3) and (4) yields $\hat{g}_{OLP} = -0.015$ $(= -1.5\%)$ and $\hat{\text{var}}(\hat{g}_{OLP}) = 0.000118$. Under the normality assumption this yields a much smaller 95%-confidence interval $I^{95}_{OLP} = (-6.2\%;\ 3.2\%)$ where we used $t_2 = 4.30$.

Although *n*=3 and *n*=4 are too small for applying the above variance formulas to an estimated ratio, Example 2.1 illustrates quite well the differences between the two approaches for estimating the growth rate. Moreover, it is easy to make a much larger artificial data set (say with *n*=40 and *N*=500) with the same features as the present example, including the same values for *f* and λ so that the (estimated) ratio $\hat{\mathrm{var}}(\hat{g}_{STN})/\hat{\mathrm{var}}(\hat{g}_{OLP})$ remains unchanged. In addition, this simple example may serve as a warning to be cautious when using sample means as $\bar{y}_{23}$ and $\bar{x}_{12}$ for estimating growth rates because these estimates may lead to unnecessarily large confidence intervals around a misleading estimate of the growth rate. In the next section we look more closely at the question of what kind of circumstances may lead to a large interval I_{STN}^{95} .

3. Reasons for a large interval I_{STN}^{95}

In order to get more insight into the factors determining the length of I_{STN}^{95} , consider the simple case that for all enterprises $Y_i = X_i$. Obviously, for this hypothetic situation using $\hat{g}_{OLP}$ leads to the correct results, that is a zero growth rate with zero variance. In contrast, for the variance of $\hat{g}_{STN}$ formula (2) now yields

$$\begin{aligned} \mathrm{var}(\hat{g}_{STN}) &\approx \frac{1}{\bar{X}^2}\{(\frac{1}{n}-\frac{1}{N})2S_y^2 - 2(\frac{\lambda}{n}-\frac{1}{N})S_y^2\} \\ &= 2(1-\lambda)\frac{CV_y^2}{n}, \end{aligned} \tag{5}$$

where the coefficient of variation CV_y is defined by $CV_y = S_y/\bar{Y}$. Formula (5) shows that even when all enterprises have a zero growth, the variance of $\hat{g}_{STN}$ might be quite large due to a high coefficient of variation CV_y among the Y_i, provided that $\lambda \neq 1$. Recall that economic variables such as turnover and salary often have a lognormal distribution with a large CV_y ; for the above observations, CV_y is about 0.7. Combining $CV_y = 0.7$ with, for instance, *n*=20 and $\lambda = 0.75$, formula (5) gives $\mathrm{var}(\hat{g}_{STN}) = 0.01225$. The corresponding confidence interval I_{STN}^{95} for *g* is still large. That is, for short, $I_{STN}^{95} = \hat{g}_{STN} \pm 0.23$ where we used $t_{19} = 2.09$; recall $I_{OLP}^{100} = \{0\}$ in this case. In addition, when a confidence interval $I_{STN}^{95} = \hat{g}_{STN} \pm 0.01$ (with margins of 1%) is required, it follows from (5) that one should take $n \geq 9412$; recall $CV_y = 0.7$, $\lambda = 0.75$, and $t_\infty = 1.96$. More generally, this case suggests that when the observations (Y_i, X_i) satisfy the model $Y_i / X_i = B + u_i$ with $E(u_i) = 0$,

$E(u_i^2) = \sigma_i^2$ and $E(u_i u_j) = 0 \quad (i \neq j)$, it may occur that $\mathrm{var}(\hat{g}_{STN})$ is much larger than $\mathrm{var}(\hat{g}_{OLP})$ especially when the σ_i^2 are small while S_y^2 and S_x^2 are large.

Furthermore, it is noteworthy that a decrease of λ, for instance, from 0.9 to 0.5 leads to a dramatic increase of $\mathrm{var}(\hat{g}_{STN})$ in (5) by 400%. This emphasizes once more the importance of avoiding panel attrition when using estimator $\hat{g}_{STN}$. Another striking feature of (5) is that the outcome in this particular case does not depend on *N* or the finite population correction.

Finally, it can be shown that $\mathrm{var}(\hat{g}_{OLP})$ may exceed $\mathrm{var}(\hat{g}_{STN})$ when λ is small enough. Assuming $G > 0$, this follows from subtracting (2) from (4), yielding

$$\begin{aligned}\mathrm{var}(\hat{g}_{OLP}) - \mathrm{var}(\hat{g}_{STN}) &\approx \frac{1}{\overline{X}^2}\{(\frac{1}{n_2} - \frac{1}{n})(S_y^2 + G^2 S_x^2) - 2G(\frac{1}{n_2} - \frac{\lambda}{n})S_{xy}\} \\ &= \frac{1}{n_2\overline{X}^2}\{(1-\lambda)(S_y^2 + G^2 S_x^2) - 2G(1-\lambda^2)S_{xy}\} \\ &= \frac{1-\lambda}{n_2\overline{X}^2}(S_{y-Gx}^2 - 2G\lambda S_{xy}). \qquad (6)\end{aligned}$$

In other words, $\mathrm{var}(\hat{g}_{OLP})$ is larger than $\mathrm{var}(\hat{g}_{STN})$ when $S_{xy} < 0$ or when $\lambda < S_{y-Gx}^2 / 2GS_{xy}$ provided $S_{xy} > 0$. Assuming $S_y^2 = S_x^2$, Qualité and Tillé (2008) derive a similar result for estimating the parameter of absolute change.

Here we follow a somewhat different approach. Suppose that the data satisfy the above model which can be rewritten as $Y_i = BX_i + \varepsilon_i$ with $\varepsilon_i = X_i u_i$, $E(\varepsilon_i) = 0$, $E(\varepsilon_i^2) = X_i^2 \sigma_i^2$ and $E(\varepsilon_i \varepsilon_j) = 0 \quad (i \neq j)$. Under this model, we make the (weak) assumptions (i) $G = S_{yx} / S_x^2$ and (ii) $S_{y-Gx}^2 = S_y^2(1-\rho_{xy}^2)$. Recall from regression theory that $\hat{B} = S_{yx} / S_x^2$ can be seen as the unbiased, consistent estimator for *B* in an ordinary least squares (OLS) regression of Y_i on X_i and a constant $(i = 1,..., N)$ while $G = Y / X$ can be seen as the unbiased, consistent estimator for *B* in an OLS regression of $Y_i / \sqrt{X_i}$ on $\sqrt{X_i}$. Hence, we get the somewhat counterintuitive result $G = Y / X = (S_{yx} / S_x^2)[1 + o(1)]$ as $N \to \infty$; also recall that for a regression model $y = Xb + \varepsilon$ (in vector notation), $\hat{b}_{OLS} = (X'X)^{-1} X'y$. Moreover, $S_y^2(1-\rho_{xy}^2)$ is the (unexplained) variance of the residuals in the first regression. However, these residuals are asymptotically equal to $Y_i - GX_i$ from which the *approximate* validity of (ii) follows. Furthermore, noting that $S_y^2 \rho_{xy}^2$ is the so-called *explained* variance of the first regression, it follows from assumption (i) that $S_y^2\rho_{xy}^2 = \hat{B}^2 S_x^2 = G^2 S_x^2$. Combining this with assumptions (i) and (ii), we can rewrite (6) as

$$\begin{aligned}\text{var}(\hat{g}_{OLP}) - \text{var}(\hat{g}_{STN}) &\approx \frac{1-\lambda}{n_2\overline{X}^2}\{(1-\rho_{xy}^2)S_y^2 - 2G^2\lambda S_x^2\} \\ &= \frac{(1-\lambda)S_y^2}{n_2\overline{X}^2}(1-\rho_{xy}^2 - 2\lambda\rho_{xy}^2).\end{aligned}$$

Hence, $\text{var}(\hat{g}_{OLP})$ is larger than $\text{var}(\hat{g}_{STN})$ when $\lambda < (1-\rho_{xy}^2)/2\rho_{xy}^2$. Thus for say $\rho_{xy} = 0.9$, $\text{var}(\hat{g}_{OLP})$ is under the above model for sufficiently large N larger than $\text{var}(\hat{g}_{STN})$ when $\lambda < 0.117$, and for say $\rho = 0.95$ when $\lambda < 0.054$.

4. Composite estimator for the growth rate

Examining a composite estimator of the form

$$\hat{g}_{COM} = k\hat{g}_{STN} + (1-k)\hat{g}_{OLP}, \qquad (7)$$

it follows from minimizing $\text{var}(\hat{g}_{COM})$ with respect to k that

$$k = \frac{\text{var}(\hat{g}_{OLP}) - \text{cov}(\hat{g}_{OLP}, \hat{g}_{STN})}{\text{var}(\hat{g}_{OLP}) + \text{var}(\hat{g}_{STN}) - 2\,\text{cov}(\hat{g}_{OLP}, \hat{g}_{STN})}; \qquad (8)$$

also see Särndal et al. (1992, page 372). Note that, by construction, $\text{var}(\hat{g}_{COM})$ can not exceed $\min\{\text{var}(\hat{g}_{STN}), \text{var}(\hat{g}_{OLP})\}$.

Using the linearized forms of the estimators $\hat{g}_{STN}$ and $\hat{g}_{OLP}$, we get for their covariance

$$\begin{aligned}\text{cov}(\hat{g}_{OLP}, \hat{g}_{STN}) &\approx \text{cov}(\frac{\overline{y}_2 - G\overline{x}_2}{\overline{X}}, \frac{\overline{y}_{23} - G\overline{x}_{12}}{\overline{X}}) \\ &= \frac{1}{\overline{X}^2}\{\text{cov}(\overline{y}_2, \overline{y}_{23}) - G\,\text{cov}(\overline{y}_2, \overline{x}_{12}) - G\,\text{cov}(\overline{x}_2, \overline{y}_{23}) + G^2\,\text{cov}(\overline{x}_2, \overline{x}_{12})\}.\end{aligned}$$

Now using some results from Knottnerus (2003, page 377)

$$\begin{aligned}&\text{cov}(\overline{y}_2, \overline{y}_{23}) = \text{var}(\overline{y}_{23}) \;\; [= (\frac{1}{n} - \frac{1}{N})S_y^2] \\ &\text{cov}(\overline{x}_2, \overline{y}_{23}) = \text{cov}(\overline{x}_{23}, \overline{y}_{23}) \;\; [= (\frac{1}{n} - \frac{1}{N})S_{xy}],\end{aligned}$$

we obtain

$$\mathrm{cov}(\hat{g}_{OLP}, \hat{g}_{STN}) \approx \frac{1}{n\bar{X}^2}(1-f)S^2_{y-Gx} \,. \tag{9}$$

In practice *k* can be estimated by replacing all (co)variances in (8) by their sample estimates, yielding

$$\hat{k} = \frac{\hat{\mathrm{var}}(\hat{g}_{OLP}) - \hat{\mathrm{cov}}(\hat{g}_{OLP}, \hat{g}_{STN})}{\hat{\mathrm{var}}(\hat{g}_{OLP}) + \hat{\mathrm{var}}(\hat{g}_{STN}) - 2\,\hat{\mathrm{cov}}(\hat{g}_{OLP}, \hat{g}_{STN})} \,. \tag{10}$$

As expected, applying (10) to Example 2.1 yields $\hat{k} \approx 0$ because $\mathrm{var}(\hat{g}_{STN})$ is much larger than $\mathrm{var}(\hat{g}_{OLP})$. In the following example we examine a less extreme case.

Example 4.1. The data are the same as for Example 2.1. Only for enterprise $i = 4$ we have changed the observations into $(X_4, Y_4) = (575, 400)$ corresponding with a (large) decrease of 30%. Applying formulas (1)-(4) and (9) to these new data yields

$\hat{g}_{STN} = -0.058\ (0.0147)$, $\hat{g}_{OLP} = -0{,}146\ (0.0099)$, and
$\hat{\mathrm{cov}}(\hat{g}_{STN}, \hat{g}_{OLP}) = 0.0073$.

The variances are mentioned between parentheses. Substituting these estimates into (10) yields $\hat{k} = 0.26$ and subsequently, $\hat{g}_{COM} = -0.123\ (0.0088)$. For the ease of exposition, all (co)variances in (10) are estimated from overlap s_2, including the estimates of *G* and $\bar{X}$ in (2), (4) and (9). Note that $\hat{g}_{COM}$ can be rewritten as

$$\begin{aligned}\hat{g}_{COM} &= \hat{g}_{OLP} + \hat{k}(\hat{g}_{STN} - \hat{g}_{OLP}) \\ &\approx \hat{g}_{OLP} + k(\hat{g}_{STN} - \hat{g}_{OLP}),\end{aligned}$$

where we used a first-order Taylor series approximation. Therefore, the random character of estimator $\hat{k}$ can be neglected for estimating $\mathrm{var}(\hat{g}_{COM})$. The error thus introduced is of order $1/n_2$ as $n_2 \to \infty$. Recall that this is in analogy with the variance estimation procedure for the regression estimator.

5. Aligned composite estimators for growth rates and totals

So far we only looked at growth rates because in practice the estimate $\hat{X}_{SRS}$ for the turnover of 12 months ago can be considered more or less as fixed (*i.e.,* nonrandom). When *X* refers to the total turnover in month *t*-1, it is likely that the figures for the preceding month can still be improved and modified. In such a situation the initial estimate $\hat{X}_{SRS}$ might be revised as well. In order to derive a multivariate composite estimator in this situation, define the initial vector estimator $\hat{\theta}_0$ by $\hat{\theta}_0 = (\hat{G}_{STN}, \hat{G}_{OLP}, \bar{y}_{23}, \bar{x}_{12})'$. Denote the underlying parameter vector to be estimated by $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)'$. Let V_0 denote the covariance matrix of $\hat{\theta}_0$. In terms of θ the problem is now to find an aligned composite estimator $\hat{\theta}_{AC}$ with elements satisfying the two prior restrictions (i) $\theta_1 - \theta_2 = 0$ and (ii) $\theta_3 - \theta_1\theta_4 = 0$ or, equivalently, $Y - GX = 0$.

If the restrictions were of the linear form $c - R\theta = 0$ where *R* is a $m \times 4$ matrix of rank *m* $(m \leq 4)$, the optimal composite estimator for θ would be equal to the general restriction (GR) estimator

$$\hat{\theta}_{GR} = \hat{\theta}_0 + K(c - R\hat{\theta}_0) \tag{11}$$

$$K = V_0 R'(RV_0 R')^{-1} \tag{12}$$

$$V_{GR} \equiv \mathrm{cov}(\hat{\theta}_{GR}) = (I_4 - KR)V_0,$$

where I_4 stands for the 4×4 identity matrix. The estimator $\hat{\theta}_{GR}$ is optimal in the sense that when $\hat{\theta}_0$ follows a multivariate normal distribution $N(\theta, V_0)$, the likelihood of $\hat{\theta}_0$ attains its maximum, under the constraint $c - R\theta = 0$, for $\theta_{\max} = \hat{\theta}_{GR}$. Moreover, given the form of (11), it can be shown that minimizing $\mathrm{tr}\{\mathrm{cov}(\hat{\theta}_K)\}$ with respect to the $4 \times m$ matrix *K* leads to (12). Recall that this means that for any other matrix *K* the corresponding covariance matrix $\mathrm{cov}(\hat{\theta}_K)$ of $\hat{\theta}_K = \hat{\theta}_0 + K(c - R\hat{\theta}_0)$ exceeds V_{GR} by a positive semidefinite matrix; see Magnus and Neudecker (1988, pages 255-256). For further details on the GR estimator, see Knottnerus (2003, pages 328-332).

In case of *m* nonlinear restrictions, say $c - R(\theta) = 0$, a first-order Taylor series approximation around $\theta = \hat{\theta}_0$ yields $c - R(\hat{\theta}_0) - D_R(\hat{\theta}_0)(\theta - \hat{\theta}_0) = 0$ where $D_R(\theta)$ stands for the $m \times 4$ matrix of partial derivatives of $R(\theta)$ $(i.e., D_R(\theta) = \partial R(\theta)/\partial\theta')$. Subsequently, an iterative procedure can be carried out by repeatedly applying (11) to the updated linearized versions of the nonlinear restrictions $c - R(\theta) = 0$. This yields

$$\left.\begin{aligned}
&\hat{\theta}_h = \hat{\theta}_0 + K_h \hat{e}_h;\\
&\hat{e}_h = c - R(\hat{\theta}_{h-1}) - D_h(\hat{\theta}_0 - \hat{\theta}_{h-1});\\
&K_h = V_0 D_h'(D_h V_0 D_h')^{-1};\\
&\mathrm{cov}(\hat{\theta}_h) = (I_4 - K_h D_h)V_0;\\
&D_h = D_R(\hat{\theta}_{h-1}) \qquad (h = 1,2,\ldots);
\end{aligned}\right\} \tag{13}$$

see Knottnerus (2003, pages 351-354). Note that the first equation can be seen as an update of $\hat{\theta}_0$ rather than of $\hat{\theta}_{h-1}$. This is an important difference with the celebrated Kalman equations; see Kalman (1960). The vectors $\hat{\theta}_{h-1}$ are only used for finding new (better) Taylor series approximations of $c - R(\theta) = 0$ around $\theta = \hat{\theta}_{h-1}$

$(h = 1,2,...)$ until convergence is reached. Furthermore, note that $\hat{e}_h$ can be seen as an *m*-vector of restriction errors when substituting $\theta = \hat{\theta}_0$ into the linearized restrictions around $\theta = \hat{\theta}_{h-1}$. To illustrate the use of the Kalman-like equations in (13) for deriving aligned composite estimators for growth rates and totals, consider the following example.

Example 5.1. We use the same data as in Example 4.1. The initial vector $\hat{\theta}_0$ defined above is given by $\hat{\theta}_0 = (0.9423\,, 0.8543\,, 403\,.6, 428\,.3)'$. These estimates do not satisfy the first restriction $\theta_1 - \theta_2 = 0$; recall the second restriction is $\theta_3 - \theta_1\theta_4 = 0$ so that $c = (0,\, 0)'$. Most elements of V_0 have already been discussed. Only a somewhat complicated covariance such as $\mathrm{cov}(\,\bar{y}_{23}, \hat{G}_{STN}\,)$ we did not discuss yet. Again using the above linearization of $\hat{G}_{STN}$, it is easily seen that

$$\begin{aligned}\mathrm{cov}(\,\bar{y}_{23}, \hat{G}_{STN}\,) &\approx \mathrm{cov}(\,\bar{y}_{23}, \frac{\bar{y}_{23} - G\bar{x}_{12}}{\bar{X}})\\ &= \frac{1}{\bar{X}}\{\mathrm{var}(\,\bar{y}_{23}) - G\,\mathrm{cov}(\,\bar{y}_{23}, \bar{x}_{12})\}. \qquad (14)\end{aligned}$$

Each term in (14) can be estimated from s_2 as described before. Three other covariances in V_0 have a similar form and can be estimated in the same manner. For the (*h*+1*)*th recursion the vector $R(\hat{\theta}_h)$ and the 2×4 matrix D_{h+1} are given by

$$R(\hat{\theta}_h) = \begin{pmatrix} \hat{\theta}_{h1} - \hat{\theta}_{h2} \\ \hat{\theta}_{h3} - \hat{\theta}_{h1}\hat{\theta}_{h4} \end{pmatrix}$$

$$D_{h+1} = \begin{pmatrix} 1 & -1 & 0 & 0 \\ -\hat{\theta}_{h4} & 0 & 1 & -\hat{\theta}_{h1} \end{pmatrix},$$

respectively; $\hat{\theta}_{hk}$ is the *k*th element of vector $\hat{\theta}_h$ $(1 \leq k \leq 4)$. Note that V_0 remains unchanged for all recursions. The first recursion from (13) yields

$$\hat{\theta}_1 = (0.90191\,,\ 0.90191\,,\ 379\,.32,\ 420\,.91)'\,.$$

The first restriction is satisfied and the second (nonlinear) restriction is almost satisfied, that is, $R(\hat{\theta}_1) = (0,\, -\,0.30\,)'$. The second recursion yields the following aligned composite (AC) estimates

$$\hat{G}_{AC} = 0.90915\ \ (0.0022\,), \quad \hat{\bar{Y}}_{AC} = 380\,.24\ (1663\,), \quad \text{and} \quad \hat{\bar{X}}_{AC} = 418\,,22\ (540\,).$$

Between parentheses the variances are mentioned. The (absolute) error of the second restriction further decreased, that is, $R(\hat{\theta}_2) = (0,\, 0.02\,)'$ and we stopped the recursions. Compared to $\hat{g}_{COM}$ discussed in the preceding section the variance of $\hat{g}_{AC}$ $(= \hat{G}_{AC} - 1 = -0.09)$ decreased from 0.0088 to 0.0022.

Finally, two remarks are in order. Firstly, $\hat{g}_{COM}$ discussed in the preceding section can be seen as a special case of the GR estimator. That is, choosing $\hat{\theta}_0 = (\hat{g}_{STN}, \hat{g}_{OLP})'$ with the restriction $\theta_1 - \theta_2 = 0$ leads to the same outcome as (7) and (8); see Knottnerus (2003, page 340). Secondly, when *p* auxiliary variables $z_1,..., z_p$ with known population means (say, the $p \times 1$ vector c_z) are observed in one or more (sub)samples, it is easy to extend the GR estimator as follows. Define $\hat{\theta}_{0+}$ by $\hat{\theta}_{0+} = (\hat{\theta}_0', \hat{\theta}_z')'$ where $\hat{\theta}_z$ stands for the initial estimator of the corresponding $p \times 1$ parameter vector θ_z for the *p* population means of the auxiliaries. Let V_{0+} denote the covariance matrix of $\hat{\theta}_{0+}$. The restrictions now become

$$R_+\theta_+ = \begin{pmatrix} R & 0 \\ 0 & I_p \end{pmatrix}\begin{pmatrix} \theta \\ \theta_z \end{pmatrix} = \begin{pmatrix} c \\ c_z \end{pmatrix} = c_+ \ ,$$

where I_p stands for the $p \times p$ identity matrix. Now applying (11) to $\hat{\theta}_{0+}$ with the extended restrictions yields an estimator that takes into account our prior knowledge that $\theta_z = c_z$ as well as the original restrictions $R\theta = c$. That is,

$$\begin{aligned} \hat{\theta}_{GR+} &= \hat{\theta}_{0+} + K_+(c_+ - R_+\hat{\theta}_{0+}) \\ K_+ &= V_{0+}R_+'(R_+V_{0+}R_+')^{-1} \\ V_{GR+} &\equiv \text{cov}(\hat{\theta}_{GR+}) = (I_{p+4} - K_+R_+)V_{0+}. \end{aligned}$$

For instance, when the variable *number of employees* is observed in the current month in s_{12} (say Z_{xi}) and in s_{23} (say Z_{yi}), then in obvious notation $\hat{\theta}_z = (\bar{z}_{x12}, \bar{z}_{y23})'$ and $c_z = (\bar{Z}_x, \bar{Z}_y)'$. Also see Knottnerus (2003, pages 335-339).

References

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

Hidiroglou, M.A., Särndal, C.E. and Binder, D.A. (1995). Weighting and estimation in business surveys. In *Business Survey Methods*, (Eds., B.G. Cox *et al*.). New York: John Wiley & Sons, Inc.

Kish, L. (1965). *Survey sampling.* New York: John Wiley & Sons, Inc.

Kalman, R.E. (1960). A new approach to linear filtering and prediction problems, *Transactions ASME, Journal of Basic Engineering*, 82, 35-45.

Knottnerus, P. (2003). *Sample Survey Theory: Some Pythagorean Perspectives.* New York: Springer-Verlag.

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

Laniel, N. (1987). Variances for a rotating sample from a changing population. *Proceedings of the Survey Research Methods Section,* American Statistical Association, 496–500.

Magnus, J.R. and Neudecker, H. (2001). *Matrix Differential Calculus with Applications in Statistics and Econometrics*, New York: John Wiley & Sons, Inc.

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.

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.

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

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

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