Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

On variances of changes estimated from rotating panels and dynamic strata

by Paul Knottnerus and Arnout van Delden

June 2012

Statistics Canada Statistique Canada

Canada

Survey Methodology, June 2012
Vol. 38, No. 1, pp. 43-52
Statistics Canada, Catalogue No. 12-001-X

On variances of changes estimated from rotating panels and dynamic strata

Paul Knottnerus and Arnout van Delden [1]

Abstract

Many business surveys provide estimates for the monthly turnover for the major Standard Industrial Classification codes. This includes estimates for the change in the level of the monthly turnover compared to 12 months ago. Because business surveys often use overlapping samples, the turnover estimates in consecutive months are correlated. This makes the variance calculations for a change less straightforward. This article describes a general variance estimation procedure. The procedure allows for yearly stratum corrections when establishments move into other strata according to their actual sizes. The procedure also takes into account sample refreshments, births and deaths. The paper concludes with an example of the variance for the estimated yearly growth rate of the monthly turnover of Dutch Supermarkets.

Key Words: Births; Business surveys; Conditional covariances; Deaths; Overlapping samples; Stratum corrections.

1. Introduction

In many surveys a changing population is repeatedly sampled so that the level and the change in the level of a characteristic between two occasions can be estimated. For example, in many countries a monthly business survey is held to estimate the level of the monthly turnover and the change in that level compared to a month or a year ago; see Konschnik, Monsour and Detlefsen (1985). Another example is the labour force survey in which the population is sampled on a monthly basis to estimate the number of unemployed persons and the unemployment rate. Variance estimation is needed to judge whether the observed changes are statistically significant. Variance estimation is also needed in the design stage of the survey, to determine the optimal sample size and allocation or to determine the optimal estimator.

In repeated surveys, changes are often estimated by using a stratification of the population. Businesses are extremely heterogeneous in terms of size and type of economic activity. Therefore, business surveys are usually designed as a stratified simple random sample selected without replacement (STSRS); see Smith, Pont and Jones (2003). In surveys for households or individuals the sample is usually not stratified because households are less heterogeneous. Some social surveys, such as labour force surveys, however, use poststratification to reduce the variance and bias of the estimator.

In deriving formulas for the variance of an estimated change in a population with dynamic strata, one has to pay attention to three complicating factors. Firstly, the change in a level is the result of two components. One component is due to the change in the population mean of units that remain in the same stratum on both occasions. The other component is caused by the change in the stratum composition between two occasions resulting from births and deaths in the population and from population units that migrate between strata; see Holt and Skinner (1989). Secondly, due to the migration of population units between strata, the estimated mean of stratum h at occasion t may be correlated with the mean of stratum ℓ at occasion $t+1$. Thirdly, another complicating factor is that the population is repeatedly sampled, resulting in partially overlapping samples between two occasions. Different rotating panel designs may be used in business surveys.

Various authors have derived formulas for design-based variance estimators for the estimation of changes. Assuming a large population without births and deaths, Kish (1965) derived an expression for the variance of an estimated change based on overlapping samples. Tam (1984) removed the assumption of a large population. Elaborating on Tam's results, Qualité and Tillé (2008) compare several variance estimators of an estimated change. Wood (2008) generalizes Tam's results for surveys with unequal probabilities. Lowerre (1979) and Laniel (1987) deal with the variance estimation of a change in dynamic populations, but they do not take stratification into account. Hidiroglou, Särndal and Binder (1995) deal with dynamic populations and stratification, but not with changing strata. Nordberg (2000) and Berger (2004) derived formulas for the more complicated situation of a dynamic population with units that move between strata. For the Swedish sampling design Nordberg (2000) derives formulas using inclusion indicators which requires some algebra. Assuming that the size of the overlap of two samples at two different occasions is fixed, Berger (2004) derives his formulas based on Poisson sampling conditional on the sample size per stratum which requires some matrix algebra.

1. Paul Knottnerus and Arnout van Delden, Statistics Netherlands, PO Box 24500, 2490 HA The Hague, The Netherlands. E-mail: pkts@cbs.nl and adln@cbs.nl.

In this paper, we derive the expressions for STSRS sampling in a more straightforward manner without assuming that sizes of overlaps are fixed. Furthermore, unlike the Swedish design, the Dutch one doesn't require time-consuming calculations for estimating one of the variance components for a change. In addition, we propose an alternative estimation method for sampling designs with such a non-zero component. In order to clarify the variance estimation procedure, we describe its application to the yearly growth rates of the turnover of Dutch Supermarkets of 4-week periods.

The outline of the paper is as follows. Section 2 briefly describes the Dutch business survey for monthly turnover, including the sampling design. The variance formulas for the estimator of a change are derived in section 3. Section 4 illustrates the variance estimation procedure by comparing the variances of two different estimators for the yearly growth rate of the monthly turnover of Dutch Supermarkets in the period 2003-2004. Section 5 summarizes the main results and conclusions.

2. The sampling design of the Dutch business surveys

Every month Statistics Netherlands estimates the monthly turnover for some of the major SIC codes. The publication includes the 12-month growth rates of the monthly turnover, *i.e.*, the relative change in the monthly level of turnover compared to 12 months ago. Throughout this paper we will refer to this growth rate as the yearly growth rate.

All statistical units or *establishments* are listed in the General Business Register (GBR) that is maintained by Statistics Netherlands. The register is updated each month for births and deaths from administrative sources, while once a year, on December 31, the size category and the type of economic activity (SIC code) are updated. Note that the registration in the GBR may lag behind the changes in the population (births, deaths, size class changes *etc.*). Moreover, the (unknown) deaths in the frame may lead to a biased estimate of the level of the turnover. In order to avoid this kind of bias, it is important to quickly detect and remove deaths from the frame. Deaths detected in the sample may play a role here. However, a further analysis and correction of these errors are beyond the scope of this paper on variance estimation for growth rates. For estimating these variances, we assume that the population units and their characteristics in the register are correct. Likewise, we assume that there is zero non-response among the surveys.

Every first day of the month an STSRS-like sample from the GBR is conducted to estimate the turnover of the current month. In fact, a rotating sample is used. The sample is stratified by size and by type of economic activity. The actual probability of selection depends on size and economic activity. The probability of selection increases with the size of establishment, with the largest establishments being included in the sample with probability 1. For some SICs there are not only survey data available but also data from administrative sources. The units already present in the administrative files are considered as a separate stratum. The estimates from this stratum have a zero variance.

The sample is updated in two ways. Every month the sample is updated to correct for births and deaths in the population. Once a year, in January, 10% of the sample units are replaced and stratum corrections are carried out. We will discuss the monthly and yearly updates in more detail.

2.1 Monthly update

Each month t ($t = 1, 2, ...$) a fixed proportion f_h of the N_h^t units in stratum U_h^t is sampled ($h = 1, ..., H$). This results in a sample s_h^t of size $n_h^t = f_h N_h^t$. Hence, the actual number of units in the sample may change from month to month due to births and deaths in the population. Note that apart from minor round-off errors the actual sampling fraction f_h does not depend on month t. In fact, the update procedure for s_h^t in month t is as follows. Define $U_{0h}^{t-1,t}$ as the set of births in stratum h in month $t-1$ and denote its size by $N_{0h}^{t-1,t}$. The number of sampled units from $U_{0h}^{t-1,t}$ in month t is $n_{0h}^{t-1,t} = f_h N_{0h}^{t-1,t}$. In addition, denote the further required difference $n_h^t - n_{0h}^{t-1,t}$ by $n_{h,\mathrm{REQ}}^{t-1,t}$ and define $s_{h,\mathrm{PRE}}^t$ by $s_{h,\mathrm{PRE}}^t = s_h^{t-1} \cap U_h^t$, that is the set of units in s_h^{t-1} that still exist in month t. Let $n_{h,\mathrm{PRE}}^t$ denote the size of $s_{h,\mathrm{PRE}}^t$. When $n_{h,\mathrm{PRE}}^t \geq n_{h,\mathrm{REQ}}^{t-1,t}$, randomly drop the difference, otherwise select the difference from $U_h^t \setminus U_{0h}^{t-1,t} \setminus s_{h,\mathrm{PRE}}^t$. Note that units dropped from the sample in month $t-1$ or earlier may be re-selected in month t.

2.2 Yearly update

Each January, the sample is updated to account for both a re-stratification of the units and a sample replacement of 10%. All sample units of December that still exist in January are stratified according to their actual size, *i.e.*, the number of employees and the SIC-code of January. The size class boundaries themselves remain unchanged. Consequently, the resulting sample from a stratum according to the new January stratification may consist of units with different inclusion probabilities because units move between strata with different sampling fractions.

In order to correct for possibly different inclusion probabilities in stratum ℓ, denote the substratum consisting of units that belonged to stratum h in December and in January to stratum ℓ by $U_{h\ell}^{\mathrm{dec,\,jan}}$ and denote its size by $N_{h\ell}^{\mathrm{dec,\,jan}}$ ($h, \ell = 1, ..., H$). In analogy with the monthly update procedure define $s_{h\ell,\mathrm{PRE}}^{\mathrm{jan}}$ by $s_{h\ell,\mathrm{PRE}}^{\mathrm{jan}} = s_h^{\mathrm{dec}} \cap U_{h\ell}^{\mathrm{dec,\,jan}}$ and let $n_{h\ell,\mathrm{PRE}}^{\mathrm{jan}}$ denote the size of $s_{h\ell,\mathrm{PRE}}^{\mathrm{jan}}$. Since the required size of

sample $s_{h\ell,\mathrm{REQ}}^{\mathrm{dec,\,jan}}$ from $U_{h\ell}^{\mathrm{dec,\,jan}}$ in January is $n_{h\ell,\mathrm{REQ}}^{\mathrm{dec,\,jan}} = f_\ell N_{h\ell}^{\mathrm{dec,\,jan}}$, the yearly update of sample $s_{h\ell,\mathrm{PRE}}^{\mathrm{jan}}$ is carried out as follows.

Firstly, when $n_{h\ell,\mathrm{PRE}}^{\mathrm{jan}} \geq n_{h\ell,\mathrm{REQ}}^{\mathrm{dec,\,jan}}$, randomly drop the difference from $s_{h\ell,\mathrm{PRE}}^{\mathrm{jan}}$. In addition, 10% of the $n_{h\ell,\mathrm{REQ}}^{\mathrm{dec,\,jan}}$ remaining units in $s_{h\ell,\mathrm{PRE}}^{\mathrm{jan}}$ is replaced by units from $U_{h\ell}^{\mathrm{dec,\,jan}} \setminus s_{h\ell,\mathrm{PRE}}^{\mathrm{jan}}$ provided that the latter set contains enough units. When there are not enough units available, the number of replaced units is only $N_{h\ell}^{\mathrm{dec,\,jan}} - n_{h\ell,\mathrm{PRE}}^{\mathrm{jan}}$. Secondly, when $n_{h\ell,\mathrm{PRE}}^{\mathrm{jan}} < n_{h\ell,\mathrm{REQ}}^{\mathrm{dec,\,jan}}$, select the difference from $U_{h\ell}^{\mathrm{dec,\,jan}} \setminus s_{h\ell,\mathrm{PRE}}^{\mathrm{jan}}$. Subsequently, an additional replacement of $n_{h\ell,\mathrm{PRE}}^{\mathrm{jan}} - 0.9 n_{h\ell,\mathrm{REQ}}^{\mathrm{dec,\,jan}}$ units in $s_{h\ell,\mathrm{REQ}}^{\mathrm{dec,\,jan}}$ takes place when this difference is positive and enough new units are available. This procedure is done for all substrata $h\ell$, including $h = \ell$. Thirdly, similar to the monthly update procedure the number of sampled units in January from substratum $U_{0\ell}^{\mathrm{dec,\,jan}}$ of new births in stratum ℓ is $n_{0\ell}^{\mathrm{dec,\,jan}} = f_\ell N_{0\ell}^{\mathrm{dec,\,jan}}$. In addition, note that this approach can also be followed when class size boundaries or sampling fractions are changed in January.

Apart from the stratum corrections in January, the resulting sample in month t can be considered more or less as a set of SRS samples from the strata U_h^t. When the population and the strata h are stable over the years, the procedure described so far amounts to a standard STSRS sampling design for month t. Therefore, Statistics Netherlands uses the familiar variance formulas for the STSRS sampling design for estimating the variance of the level of the monthly turnover. In the next section we show how the variance for a change of the level can be estimated under such an STSRS assumption.

3. Variance of the yearly growth rate of monthly turnover

3.1 Variance of the yearly growth rate

Let O^t denote the total turnover of all establishments in the population in month t and $g^{t,s}$ the relative change in the level of turnover between months t and s, *i.e.*,

$$g^{t,\,s} = \frac{O^t}{O^s} - 1 \quad (t > s).$$

For the corresponding estimates it holds by definition that

$$\hat{g}^{\,t,\,s} = \frac{\hat{O}^t}{\hat{O}^s} - 1, \qquad (1)$$

where a "hat" indicates an estimate; for an estimator we use the same notation. Furthermore, define

$$G^{t,\,t-12} \equiv \frac{O^t}{O^{t-12}} = 1 + g^{t,\,t-12}.$$

In order to estimate the variance of the yearly growth rate of the monthly turnover, we use the first-order Taylor series expansion of a ratio of two estimators. That is,

$$\begin{aligned}
&\mathrm{var}(\hat{g}^{\,t,\,t-12}) \\
&= \mathrm{var}\left\{\frac{\hat{O}^t}{\hat{O}^{t-12}}\right\} \\
&\approx \frac{\mathrm{var}(\hat{O}^t - G^{t,\,t-12}\hat{O}^{t-12})}{(O^{t-12})^2} \\
&= \frac{\mathrm{var}(\hat{O}^t) + (G^{t,\,t-12})^2\mathrm{var}(\hat{O}^{t-12}) - 2G^{t,\,t-12}\mathrm{cov}(\hat{O}^{t-12}, \hat{O}^t)}{(O^{t-12})^2}. \qquad (2)
\end{aligned}$$

The major problem is the estimation of $\mathrm{cov}(\hat{O}^{t-12}, \hat{O}^t)$. In the next sections we examine this term and its estimation.

3.2 The covariance term of the yearly growth rate

Using the stratified sampling design, we can write $\mathrm{cov}(\hat{O}^{t-12}, \hat{O}^t)$ from (2) as

$$\begin{aligned}
\mathrm{cov}(\hat{O}^{t-12}, \hat{O}^t) &= \mathrm{cov}\left(\sum_{h=1}^{H} N_h^{t-12}\,\bar{o}_h^{\,t-12}, \sum_{\ell=1}^{H} N_\ell^t\,\bar{o}_\ell^{\,t}\right) \\
&= \sum_{h=1}^{H}\sum_{\ell=1}^{H} N_h^{t-12} N_\ell^t\,\mathrm{cov}(\bar{o}_h^{\,t-12}, \bar{o}_\ell^{\,t}), \qquad (3)
\end{aligned}$$

where $\bar{o}_h^{\,t-m}$ stands for the sample mean of the turnover in stratum h in month $t-m$ $(m = 0, 12)$. Note that the stratification of the units in month $t-12$ may differ from that in month t. As we have seen in section 2.2, the standard refreshment of the panel takes place in January. Furthermore, each establishment is allocated to the correct stratum h according to its actual number of employees in January $(h = 1, ..., H)$. To take these design features into account, define

- $N_{h\ell}^{t-12,t}$: size of substratum $U_{h\ell}^{t-12,t}$, *i.e.*, the set of units that in month $t-12$ belonged to stratum h and in month t to stratum ℓ $(h, \ell = 1, ..., H)$;
- $O_{h\ell}^{t-m}$: the substratum population total of the turnover in $U_{h\ell}^{t-12,\,t}$ in month $t-m$ $(m = 0, 12)$;
- $\bar{O}_{h\ell}^{t-m}$: the substratum population mean of the turnover in $U_{h\ell}^{t-12,t}$ in month $t-m$ [*i.e.*, $\bar{O}_{h\ell}^{t-m} = O_{h\ell}^{t-m}/N_{h\ell}^{t-12,t}$ $(m = 0, 12)$];
- $n_{h\ell}^{t-m}$: size of sample $s_{h\ell}^{t-m}$, *i.e.*, the actual sample from $U_{h\ell}^{t-12,t}$ in month $t-m$ $(0 \leq m \leq 12)$;
- $o_{h\ell}^{t-m}$: the sample total of the turnover in $s_{h\ell}^{t-m}$ $(m = 0, 12)$;
- $\bar{o}_{h\ell}^{t-m}$: the sample mean of the turnover in $s_{h\ell}^{t-m}$ [*i.e.*, $\bar{o}_{h\ell}^{t-m} = o_{h\ell}^{t-m}/n_{h\ell}^{t-m}$ $(m = 0, 12)$];
- $n_{h\ell}^{t-12,t}$: number of units in the overlap $s_{h\ell}^{t-12,\,t} \equiv s_{h\ell}^{t-12} \cap s_{h\ell}^{t}$;

$\bar{o}_{h\ell,\mathrm{OLP}}^{\,t-m}$: the sample mean of the turnover in the overlap $s_{h\ell}^{t-12,\,t}$ in month $t-m$. [*i.e.*, $\bar{o}_{h\ell,\mathrm{OLP}}^{\,t-m} = o_{h\ell,\mathrm{OLP}}^{\,t-m} / n_{h\ell}^{t-12,t}$ $(m = 0,12)$].

In addition to the notation in section 2, define the auxiliary stratum 0 for the *births* in months $t-12, ..., t-1$ and likewise, stratum $H+1$ for the *deaths* in that period. Then $\bar{o}_h^{\,t-12}$ and $\bar{o}_\ell^{\,t}$ can be written as

$$\bar{o}_h^{\,t-12} = \sum_{g=1}^{H+1} \frac{n_{hg}^{t-12}}{n_h^{t-12}} \bar{o}_{hg}^{\,t-12}$$

$$\bar{o}_\ell^{\,t} = \sum_{k=0}^{H} \frac{n_{k\ell}^{t}}{n_\ell^{t}} \bar{o}_{k\ell}^{\,t},$$

respectively $(1 \le h, \ell \le H)$. Consequently, the covariances in (3) can be rewritten as

$$\mathrm{cov}(\bar{o}_h^{\,t-12}, \bar{o}_\ell^{\,t}) = \mathrm{cov}\left(\sum_{g=1}^{H+1} \frac{n_{hg}^{t-12}}{n_h^{t-12}} \bar{o}_{hg}^{\,t-12}, \sum_{k=0}^{H} \frac{n_{k\ell}^{t}}{n_\ell^{t}} \bar{o}_{k\ell}^{\,t} \right) \quad (4a)$$

$$= \frac{1}{n_h^{t-12} n_\ell^{t}} \mathrm{cov}(n_{h\ell}^{t-12} \bar{o}_{h\ell}^{\,t-12}, n_{h\ell}^{t} \bar{o}_{h\ell}^{\,t}) \,, \quad (4b)$$

where we used $\mathrm{cov}(n_{hg}^{t-12} \bar{o}_{hg}^{\,t-12}, n_{k\ell}^{t} \bar{o}_{k\ell}^{\,t}) = 0$ $(k \ne h)$ and $\mathrm{cov}(n_{hg}^{t-12} \bar{o}_{hg}^{\,t-12}, n_{h\ell}^{t} \bar{o}_{h\ell}^{\,t}) = 0$ $(g \ne \ell)$. The latter covariance is zero because

$$\begin{aligned} &\mathrm{cov}(n_{hg}^{t-12} \bar{o}_{hg}^{\,t-12}, n_{h\ell}^{t} \bar{o}_{h\ell}^{\,t}) \\ &\quad = E\,\mathrm{cov}(n_{hg}^{t-12} \bar{o}_{hg}^{\,t-12}, n_{h\ell}^{t} \bar{o}_{h\ell}^{\,t} \,|\, n_{hg}^{t-12}, n_{h\ell}^{t}) \\ &\quad\quad + \mathrm{cov}\{E(n_{hg}^{t-12} \bar{o}_{hg}^{\,t-12} \,|\, n_{hg}^{t-12}, n_{h\ell}^{t}), E(n_{h\ell}^{t} \bar{o}_{h\ell}^{\,t} \,|\, n_{hg}^{t-12}, n_{h\ell}^{t})\} \\ &\quad = 0 + \bar{O}_{hg}^{\,t-12}\, \bar{O}_{h\ell}^{\,t}\, \mathrm{cov}(n_{hg}^{t-12}, n_{h\ell}^{t}) = 0. \end{aligned}$$

In the last line we also used that for $1 \le g \le H+1$

$$\mathrm{cov}(n_{hg}^{t-12}, n_{h\ell}^{t}) = 0. \quad (5)$$

For a justification and the underlying assumptions of (5), see Appendix A. Moreover, in Appendix A we propose an alternative estimation method when this covariance is non-negligible. The covariance in (4b) can be expressed as

$$\begin{aligned} &\mathrm{cov}(n_{h\ell}^{t-12} \bar{o}_{h\ell}^{\,t-12}, n_{h\ell}^{t} \bar{o}_{h\ell}^{\,t}) \\ &\quad = E\{\mathrm{cov}(n_{h\ell}^{t-12} \bar{o}_{h\ell}^{\,t-12}, n_{h\ell}^{t} \bar{o}_{h\ell}^{\,t} \,|\, \nu_{h\ell})\} \\ &\quad\quad + \mathrm{cov}\{E(n_{h\ell}^{t-12} \bar{o}_{h\ell}^{\,t-12} \,|\, \nu_{h\ell}), E(n_{h\ell}^{t} \bar{o}_{h\ell}^{\,t} \,|\, \nu_{h\ell})\} \quad (6) \end{aligned}$$

where $\nu_{h\ell} = (n_{h\ell}^{t-12}, n_{h\ell}^{t-12,t}, n_{h\ell}^{t})$. The first component on the right-hand side is

$$\begin{aligned} &E\{\mathrm{cov}(n_{h\ell}^{t-12} \bar{o}_{h\ell}^{\,t-12}, n_{h\ell}^{t} \bar{o}_{h\ell}^{\,t} \,|\, \nu_{h\ell})\} \\ &\quad = E\{n_{h\ell}^{t-12} n_{h\ell}^{t}\, \mathrm{cov}(\bar{o}_{h\ell}^{\,t-12}, \bar{o}_{h\ell}^{\,t} \,|\, \nu_{h\ell})\} \\ &\quad = E\left\{ n_{h\ell}^{t-12} n_{h\ell}^{t} \left(\frac{n_{h\ell}^{t-12,t} / n_{h\ell}^{t-12}}{n_{h\ell}^{t}} - \frac{1}{N_{h\ell}^{t-12,t}} \right) S_{h\ell}^{t-12,t} \right\}. \quad (7) \end{aligned}$$

In the last line we used (26) in Appendix B. Furthermore,

$$S_{h\ell}^{t-12,t} = \frac{1}{N_{h\ell}^{t-12,t} - 1} \sum_{i=1}^{N_{h\ell}^{t-12,t}} (O_{h\ell i}^{t-12} - \bar{O}_{h\ell}^{\,t-12})(O_{h\ell i}^{t} - \bar{O}_{h\ell}^{\,t}). \quad (8)$$

The second component on the right-hand side of (6) is equal to $\bar{O}_{h\ell}^{\,t-12}\, \bar{O}_{h\ell}^{\,t}\, \mathrm{cov}(n_{h\ell}^{t-12}, n_{h\ell}^{t}) = 0$ on account of (5). It therefore follows from (4) and (6) that

$$\mathrm{cov}(\bar{o}_h^{\,t-12}, \bar{o}_\ell^{\,t}) = E\left\{ \frac{n_{h\ell}^{t-12} n_{h\ell}^{t}}{n_h^{t-12} n_\ell^{t}} \left(\frac{n_{h\ell}^{t-12,t}}{n_{h\ell}^{t-12} n_{h\ell}^{t}} - \frac{1}{N_{h\ell}^{t-12,t}} \right) S_{h\ell}^{t-12,t} \right\}. \quad (9)$$

3.3 Estimation of the covariance term of the yearly growth rate

Expression (9) can be estimated from the overlapping sample $s_{h\ell}^{t-12,t}$ by

$$\hat{\mathrm{cov}}(\bar{o}_h^{\,t-12}, \bar{o}_\ell^{\,t}) = \frac{n_{h\ell}^{t-12} n_{h\ell}^{t}}{n_h^{t-12} n_\ell^{t}} \left(\frac{n_{h\ell}^{t-12,t}}{n_{h\ell}^{t-12} n_{h\ell}^{t}} - \frac{1}{N_{h\ell}^{t-12,t}} \right) \hat{S}_{h\ell,\mathrm{OLP}}^{\,t-12,t}, \quad (10)$$

where

$$\hat{S}_{h\ell,\mathrm{OLP}}^{\,t-12,t} = \frac{1}{n_{h\ell}^{t-12,t} - 1} \sum_{i=1}^{n_{h\ell}^{t-12,t}} (o_{h\ell i}^{t-12} - \bar{o}_{h\ell,\mathrm{OLP}}^{\,t-12})(o_{h\ell i}^{t} - \bar{o}_{h\ell,\mathrm{OLP}}^{\,t}).$$

Note that (10) is unbiased for estimating (9) because

$$E(\hat{S}_{h\ell,\mathrm{OLP}}^{\,t-12,t} \,|\, \nu_{h\ell}) = S_{h\ell}^{t-12,t}.$$

Although (10) results in reasonable estimates for sufficiently large $n_{h\ell}^{t-12,t}$, a disadvantage of the covariance estimator $\hat{S}_{h\ell,\mathrm{OLP}}^{\,t-12,t}$ in (10) is that for small $n_{h\ell}^{t-12,t}$ it may lead to a negative estimate of $\mathrm{var}(\hat{O}^{\,t} - G^{\,t,t-12} \hat{O}^{\,t-12})$ in the numerator of (2). Recall that this variance is estimated by

$$\begin{aligned} \hat{\mathrm{var}}(\hat{O}^{\,t} - G^{\,t,t-12} \hat{O}^{\,t-12}) &= \hat{\mathrm{var}}(\hat{O}^{\,t}) + (\hat{G}^{\,t,t-12})^2\, \hat{\mathrm{var}}(\hat{O}^{\,t-12}) \\ &\quad - 2\hat{G}^{\,t,t-12}\, \hat{\mathrm{cov}}(\hat{O}^{\,t}, \hat{O}^{\,t-12}). \quad (11) \end{aligned}$$

Therefore, we propose an alternative estimator to $\hat{S}_{h\ell,\mathrm{OLP}}^{\,t-12,t}$ in (10). Define the standard deviations

$$\hat{S}_{h\ell}^{\,t-m} = \sqrt{\frac{1}{n_{h\ell}^{t-m} - 1} \sum_{i=1}^{n_{h\ell}^{t-m}} (o_{h\ell i}^{t-m} - \bar{o}_{h\ell}^{\,t-m})^2} \quad (m = 0, 12).$$

We propose the following modified estimator for $S_{h\ell}^{t-12,t}$

$$\hat{S}_{h\ell}^{\,t-12,t} = \hat{\rho}_{h\ell,\mathrm{OLP}}^{\,t-12,t}\, \hat{S}_{h\ell}^{\,t-12}\, \hat{S}_{h\ell}^{\,t}, \quad (12)$$

where $\rho_{h\ell}^{t-12,t}$ is the correlation between the variables o^t and o^{t-12} in $U_{h\ell}^{t-12,t}$ and $\hat{\rho}_{h\ell,\mathrm{OLP}}^{\,t-12,t}$ is its estimate from $s_{h\ell}^{t-12,t}$. According to (10) and (12) covariance (3) can be estimated by

$$\hat{\mathrm{cov}}(\hat{O}^{\,t-12}, \hat{O}^{\,t}) = \sum_{h=1}^{H} \sum_{\ell=1}^{H} \frac{N_h^{t-12} N_\ell^{t}}{n_h^{t-12} n_\ell^{t}}\, n_{h\ell}^{t-12,t} \left(1 - \frac{n_{h\ell}^{t-12} n_{h\ell}^{t}}{n_{h\ell}^{t-12,t} N_{h\ell}^{t-12,t}} \right) \hat{S}_{h\ell}^{\,t-12,t}. \quad (13)$$

For the estimate $\hat{\rho}_{h\ell,\mathrm{OLP}}^{t-12,t}$, $\hat{\rho}_{h\ell,\mathrm{OLP}}^{t-12,t} \leq 1$ always holds whereas using (10) may lead implicitly to an estimated correlation larger than 1 and a possibly negative outcome of (11). See the next section for an example. In all applications met so far, negative outcomes of (11) could be explained by the fact that unlike (12) use of (10) leads implicitly to an estimated correlation larger than 1. This is in line with the findings of Berger (2004, page 462) that an overestimation of the correlation between $\hat{O}^{t-12}$ and $\hat{O}^{t}$ may lead to a serious underestimation of the variance of a change. Nevertheless, in some extraordinary circumstances, the use of (12) might lead to a negative outcome of (11) as well. Sufficient conditions that the use of (12) leads to a nonnegative variance estimator with probability 1 are available from the authors upon request. For a general review of variance estimation methods in business surveys, see Brodie (2003).

Applying (12), a special problem may arise when $n_{h\ell}^{t} = 1$ or $n_{h\ell}^{t-12} = 1$. In order to evaluate the required sample variances, one may borrow the sample variance from a related substratum or from the same substratum in an earlier month. Alternatively, one may impute a variance when it emerges from the data that there is a relationship of the form $S_{h\ell}^2 \approx \sigma^2 \bar{O}_{h\ell}^{\beta}$; see Särndal, Swensson and Wretman (1992, page 461). In addition, the corresponding covariance term might be ignored when its (expected) contribution to the total variance is small. This is often the case when the sampling fractions in strata h and ℓ are small, that is in strata with relatively small units and, consequently, with small variances compared to the strata with larger units. Similar remarks apply to the imputed $\rho_{h\ell}^{t-12,t}$ when $n_{h\ell}^{t-12,t} \leq 2$ and $n_{h\ell}^{t-m} \geq 2$ $(m = 0, 12)$. Since the $\rho_{h\ell}^{t-12,t}$ are often fairly high, this seems to be a viable way. In the example given in section 4 the $\rho_{h\ell}^{t-12,t}$ have an overall mean of 0.90 and a variance of 0.0074 so that the impact of the imputed $\rho_{h\ell}^{t-12,t}$ on the final results is likely to be moderate.

Furthermore, note that when $n_{h\ell}^{t-m} = 0$ $(m = 0$ or $m = 12)$, the corresponding covariance term in (13) can be neglected without affecting its unbiasedness, provided that the remaining $S_{h\ell}^{t-12,t}$ are estimated in an unbiased way. Under this assumption such a term with $n_{h\ell}^{t-m} = 0$ $(m = 0$ or $m = 12)$ can be neglected because the expectation of

$$n_{h\ell}^{t-12,t}\left(1 - \frac{n_{h\ell}^{t-12} n_{h\ell}^{t}}{n_{h\ell}^{t-12,t} N_{h\ell}^{t-12,t}}\right)\hat{S}_{h\ell}^{t-12,t} \qquad (14)$$

from (13) is equal to

$$E\left[E\left\{n_{h\ell}^{t-12,t}\left(1 - \frac{n_{h\ell}^{t-12} n_{h\ell}^{t}}{n_{h\ell}^{t-12,t} N_{h\ell}^{t-12,t}}\right)\hat{S}_{h\ell}^{t-12,t} \middle| \nu_{h\ell}\right\}\right] = E\left\{n_{h\ell}^{t-12,t}\left(1 - \frac{n_{h\ell}^{t-12} n_{h\ell}^{t}}{n_{h\ell}^{t-12,t} N_{h\ell}^{t-12,t}}\right)S_{h\ell}^{t-12,t}\right\},$$

and the expectation on the right-hand side is the parameter to be estimated. Moreover, when $n_{h\ell}^{t-m} = 0$ $(m = 0$ or $m = 12)$ and consequently $n_{h\ell}^{t-12,t} = 0$, the outcome of (14) is zero and the estimator $\hat{S}_{h\ell}^{t-12,t}$ for $S_{h\ell}^{t-12,t}$ becomes irrelevant. Therefore, ignoring such a term when $n_{h\ell}^{t-m} = 0$ $(m = 0$ or $m = 12)$ does not affect the expectations of (13) and (14).

3.4 A comparison with Nordberg's results

Using the standard formalism of inclusion indicators δ_{hi}^{t} for each stratum, Nordberg (2000) derives a different expression for the first component in (6). However, it can be shown after some algebra that our expression (9) is equivalent to Nordberg's (3.4); a proof is available from the authors upon request. In addition, Nordberg derives a nonzero expression for the second component in (6), *i.e.*, the covariance between the two corresponding conditional expectations. Note that the Swedish sampling design is somewhat different from ours.

According to Nordberg (2000, page 370) the estimation of the second component for the Swedish sampling design requires a computer-intensive procedure which includes simulation of the sampling mechanism. However, since all $n_{h\ell}^{t}$, $n_{h\ell}^{t-12}$ and $n_{h\ell}^{t-12,t}$ are ancillary statistics, an alternative might be to condition on these statistics so that the second component can be ignored. Recall that a statistic is called ancillary when its marginal distribution doesn't depend on the target parameters to be estimated; see Cox and Hinkley (1974, pages 31-35). Such an alternative approach without the second component is to be recommended especially when $\hat{g}_{\mathrm{STSRS}}^{t-12,t} \approx \hat{g}_{\mathrm{PS,sub}}^{t-12,t}$ where $\hat{g}_{\mathrm{PS,sub}}^{t-12,t}$ is the poststratified estimator based on the substrata $h\ell$. However, when the difference between $\hat{g}_{\mathrm{STSRS}}^{t-12,t}$ and $\hat{g}_{\mathrm{PS,sub}}^{t-12,t}$ is non-negligible, the calculation of the unconditional variance seems to be indispensable, including the estimation of the second component according to Nordberg. For a different approach to the estimation problem of the second component, see Appendix A.

For a justification of the use of a conditional (co)variance, see Holt and Smith (1979). An important advantage of the conditional (co)variance is that the corresponding confidence interval has better coverage properties than the one based on the unconditional variance. Denote the standard conditional 95% confidence interval for an arbitrary parameter θ by $(\hat{\theta}_l, \hat{\theta}_u | \nu)$ where ν denotes the vector consisting of all (ancillary) statistics involved in the conditional (co)variances. Then under the normality assumption and some mild conditions it holds that the actual 95% confidence level (*CL*) equals the nominal confidence level because

$$\begin{aligned} \text{CL} &= \sum_{\nu\in\Omega_\nu} P(\nu)\, P(\hat{\theta}_l < \theta < \hat{\theta}_u \mid \nu) \\ &= 0.95 \sum_{\nu\in\Omega_\nu} P(\nu) = 0.95, \end{aligned}$$

where Ω_ν stands for the set of all possible outcomes of the random vector ν. When unconditional (co)variances are used, the confidence intervals thus obtained may be quite inaccurate for a given sample allocation. Moreover, when averaged over all allocations *CL* may differ from 0.95; for an example, see Knottnerus (2003, pages 133-135). Note that in the planning stage before the sample is drawn, unconditional variances are always useful for examining Kish's design effect for a comparison of different sampling designs. In addition, note that for evaluating a conditional confidence interval for $g^{t-12,t}$ the underlying variances of $\hat{O}^{t-m}_{\text{PS,sub}}$ should also be taken conditional on the $\nu_{h\ell}$ ($m = 0, 12$).

Finally, the unbiased estimator proposed by Nordberg [2000, Equation (3.9)] for the first component in (6) is quite different than those described in the previous subsection. In fact, his estimator is based on the following procedure for estimating the covariance term $S^{t-12,t}_{h\ell}$. Firstly, estimate the underlying quantity $\sum_{i=1}^{N^{t-12,t}_{h\ell}} O^{t-12}_{h\ell i} O^{t}_{h\ell i}$ from the overlap $s^{t-12,t}_{h\ell}$. Secondly, estimate the corresponding turnover means from $s^{t-12}_{h\ell}$ and $s^{t}_{h\ell}$, respectively. Since the components thus estimated stem from different samples, a negative outcome of (11) cannot always be avoided. For a small example with real data, see the following section. In the remainder Nordberg's underlying estimator for $S^{t-12,t}_{h\ell}$ is denoted by $\hat{S}^{t-12,t}_{h\ell\text{NBG}}$. A derivation of the explicit expression for $\hat{S}^{t-12,t}_{h\ell\text{NBG}}$ is available from the authors upon request.

4. An application to the change of turnover in Dutch Supermarkets

4.1 Two estimators for the yearly change of turnover

For the impact on the variance estimators it is important to know that in January the turnover is estimated twice. The first estimate, denoted by $\hat{O}^{\text{janO}}$ (with *O* for *old*), is made before the yearly sample update and is used to estimate the monthly change of the turnover in January compared to that in December. The second estimate, denoted by $\hat{O}^{\text{janN}}$ (with *N* for *new*), is made after the yearly sample update and is used to estimate the monthly change of the turnover in February compared to January. This procedure implies that units of the old sample as well as those of the new sample receive a questionnaire in January.

Unlike estimator (1) the actual estimator used by Statistics Netherlands for the yearly change in the monthly turnover is based on a chain of 12 monthly changes in turnover

$$\begin{aligned} \hat{G}^{t,t-12}_{\text{act}} &= 1 + \hat{g}^{t,t-12}_{\text{act}} = \prod_{j=0}^{11} (1 + \hat{g}^{t-j,t-j-1}) \\ &= \frac{\hat{O}^{t}}{\hat{O}^{t-1}} \times \frac{\hat{O}^{t-1}}{\hat{O}^{t-2}} \times \ \ldots \ \times \ \frac{\hat{O}^{\text{feb}}}{\hat{O}^{\text{janN}}} \\ &\quad \times \frac{\hat{O}^{\text{janO}}}{\hat{O}^{\text{dec}}} \times \ \ldots \ \times \ \frac{\hat{O}^{t-11}}{\hat{O}^{t-12}} \\ &= \frac{\hat{O}^{t}}{\hat{O}^{t-12}} \times \frac{\hat{O}^{\text{janO}}}{\hat{O}^{\text{janN}}} \quad (t \neq \text{jan}). \end{aligned} \tag{15}$$

In this section we will compare the variances of estimators (1) and (15). Similar to (2) the variance formulas for $\hat{g}^{t,t-12}_{\text{act}}$ can be derived by a first-order Taylor series expansion.

4.2 Description of the data

The calculations for the variances and confidence intervals in this example are based on turnover data of Dutch Supermarkets of 4-week periods in 2003 and 2004 (*i.e.*, $t = 1, \ldots, 26$). Hence, there are 13 observations in one year and, consequently, we use slightly adjusted symbols such as $g^{t,t-13}$ in the remainder of this section.

The population consists of about 3,500 establishments. The turnover data stem from a stratified sample and administrative files. A gross STSRS sample of about 900 units stratified by size is drawn from the full list of population units of the GBR that includes the units of the administrative files as well. Establishments with 50 or more *employees* are included with probability 1. The other establishments are sampled with decreasing inclusion probability from 1:2 (20-49 employees per establishment) to 1:40 in the smallest size (1 employee per establishment). The administrative files contain about 950 units, present in all size classes. About 500 of the 900 units in the gross sample were already present in the administrative files, but they do not receive a questionnaire. Thus, the net sample contains about 400 units. In fact, the sample size for each stratum in this specific example is random. However, as explained in subsection 3.4, we estimate all (co)variances conditional on the n_h in such a case. Data from units within the administrative files are put into a separate stratum with the sampling fraction being unity.

4.3 Results

Table 1 gives the yearly growth rates and their 95% margins for $t = 16, \ldots, 24$. It emerges that the 95% margins for the estimated growth rates $\hat{g}^{t,t-13}_{\text{act}}$, currently used by Statistics Netherlands, vary between 0.8 and 1.0 (per cent point). For example, in the first period ($t = 16$) the 95% confidence interval for the yearly growth rate is -1.3 to 0.7 per cent. As expected, the 95% margins for the more complicated estimator $\hat{g}^{t,t-13}_{\text{act}}$ are close to those for the simpler $\hat{g}^{t,t-13}$ from (1). The 95% margins of $\hat{g}^{t,t-13}$ vary

between 0.7 and 1.0 (per cent point). The estimator for the growth rate to be preferred is $\hat{g}_{\text{act}}^{t,\,t-13}$ as it corrects for the yearly sample update in January. The estimation of its variance, however, can be simplified by using the variance estimator described in section 3 rather than the more laborious expression for $\hat{\text{var}}(\hat{g}_{\text{act}}^{t,\,t-13})$.

Table 1
Estimated growth rates with 95% margins

t	$\hat{g}_{\text{act}}^{t,\,t-13} \times 100\%$	$\hat{g}^{t,\,t-13} \times 100\%$
16	-0.3 (± 1.0) [1]	-0.4 (± 1.0)
17	-3.7 (± 1.0)	-3.8 (± 0.9)
18	1.6 (± 1.0)	1.5 (± 0.9)
19	-2.2 (± 0.9)	-2.3 (± 0.9)
20	0.5 (± 0.8)	0.4 (± 0.7)
21	-1.7 (± 0.8)	-1.8 (± 0.7)
22	-2.2 (± 0.8)	-2.3 (± 0.7)
23	0.0 (± 0.8)	-0.1 (± 0.7)
24	-2.3 (± 0.9)	-2.4 (± 0.9)

[1] The 95% margins are given between parentheses.

As described in section 3, we have used the estimated correlation $\hat{\rho}_{h\ell,\text{OLP}}^{t-13,t}$ from the overlap $s_{h\ell}^{t-13,t}$ to estimate covariance $S_{h\ell}^{t-13,t}$ in order to avoid negative outcomes of (11). Knottnerus and Van Delden (2006) evaluated the bias of $\hat{\rho}_{h\ell,\text{OLP}}^{t-13,t}$ for the Dutch Supermarket data and found a small underestimation of $\hat{\rho}_{h\ell,\text{OLP}}^{t-13,t}$ resulting in a minor, less than 5%, overestimation of $\text{var}(\hat{g}^{t,t-13})$.

The use of estimator $\hat{S}_{h\ell,\text{OLP}}^{t-13,t}$ in (10) may give a negative outcome of (11) and an estimated correlation $\hat{\rho}_{h\ell}^{t-13,t}$ larger than 1. For example, consider the specific population with $N = 50$ and $H = 1$ consisting of the units of substratum $h\ell = 65$. From the panel data for this population, given in Table 2 for $t = 3$ and $t = 16$, we obtain after some calculations $\hat{S}^{t-13} = 410.7$, $\hat{S}^{t} = 394.3$ and $\hat{G}^{t,t-13} = 1.028$. Note that in the remainder of this section the subscript $h\ell = 11$ is omitted in the symbols because there is only one stratum. Table 3 gives, for three different approaches, some additional estimates for the panel data in Table 2. For example, using $\hat{S}_{\text{OLP}}^{t-13,t}$ in (10) results in an estimated correlation $\hat{\rho}^{t-13,t} = 1.39$. This then yields a negative variance estimate from (11) of minus 2.2 million. Likewise, for the same data the alternative estimator $\hat{S}_{\text{NBG}}^{t-13,t}$ of $S^{t-13,t}$ based on Nordberg (2000) results in minus 36.9 million as outcome of (11) because the corresponding estimate $\hat{\rho}_{\text{NBG}}^{t-13,t}$ becomes 1.64. In contrast, using the correlation estimated from the overlapping sample $s^{t-13,t}$ according to (12) yields $\hat{\rho}_{\text{OLP}}^{t-13,t} = 0.9997$ and the positive variance estimate from (11) becomes 52.1 million. In addition, for the panel data in Table 2 the outcome of Nordberg's estimator (3.9) for the covariance between $\hat{O}^{t-13}$ and $\hat{O}^{t}$ is 111.1 million whereas covariance estimator (13) proposed here yields 67.8 million.

Table 2
Panel data[1] from a population with $N = 50$ and $H = 1$

period	turnover per unit (in thousand euros)				
	1	2	3	4	5
$t = 3$		493.9	264.3	1,179.1	380.0
$t = 16$	475.3	472.0	267.0	1,169.0	

[1] Actually, the panel data belonged to substratum $h\ell = 65$.

5. Conclusions

The variance formulas obtained in this paper are useful for calculating the variance of an estimated yearly growth rate of monthly turnover. The use of (13) as an estimator for $\text{cov}(\hat{O}^{t-12}, \hat{O}^{t})$ results in reasonable estimates of the covariance of change in particular. The variance estimation procedure allows for rotating panels, births, deaths, and units that migrate between strata.

Furthermore, we recommend estimating a population covariance according to (12) based on the corresponding correlation estimated from the overlap and on the corresponding variances estimated from the larger separate samples. This may help to avoid a serious underestimation or a negative outcome of the variance estimator for the yearly growth rate. The resulting estimated covariances are only slightly biased.

Table 3
Estimates from three different approaches

approach		parameters to be estimated		
		$S^{t-13,t}$	$\rho^{t-13,t}$	$\text{var}(\hat{O}^{t} - G^{t,t-13}\hat{O}^{t-13})$
Nordberg (2000)	estimator	$\hat{S}_{\text{NBG}}^{t-13,t}$	$\dfrac{\hat{S}_{\text{NBG}}^{t-13,t}}{\hat{S}^{t-13}\hat{S}^{t}}$	Eq. (11)
	result	265.2×10^3	1.64	-36.9×10^6
Eq. (10)	estimator	$\hat{S}_{\text{OLP}}^{t-13,t}$	$\dfrac{\hat{S}_{\text{OLP}}^{t-13,t}}{\hat{S}^{t-13}\hat{S}^{t}}$	Eq. (11)
	result	225.0×10^3	1.39	-2.2×10^6
Eq. (12)	estimator	$\hat{\rho}_{\text{OLP}}^{t-13,t}\,\hat{S}^{t-13}\hat{S}^{t}$	$\hat{\rho}_{\text{OLP}}^{t-13,t}$	Eq. (11)
	result	161.9×10^3	1.00[1]	52.1×10^6

[1] In fact, 0.9997.

For the sampling design of the Dutch Supermarkets the second covariance term in (6) is negligible due to the fact that $n_{h\ell,\mathrm{REQ}}^{\mathrm{dec,\,jan}}$ is fixed. In contrast, for the SAMU design in Sweden this term is non-negligible and its estimation is time-consuming; the word SAMU (SAMordnade Urval) is a Swedish acronyme for coordinated samples. In Appendix A we propose an alternative method for estimating this covariance. However, under the condition that $\hat{g}^{\,t,t-12} \approx \hat{g}^{\,t,t-12}_{\mathrm{PS,sub}}$ it suffices in our opinion to only use the first covariance. This simplifies the estimation procedure considerably. Moreover, under the normality assumption the conditional confidence interval has better coverage properties compared to the unconditional interval.

The example of the Dutch Supermarkets shows one of the practical applications of the variance formulas: determining which estimator has the smallest variance. The results confirm that the variance of the simple estimator $\hat{g}^{\,t,t-13}$ is close to that of $\hat{g}^{\,t,t-13}_{\mathrm{act}}$ from section 4 which corrects for the sample refreshment in January. Hence, for the Dutch Supermarkets $\hat{\mathrm{var}}(\hat{g}^{\,t,t-13})$ might be used for estimating $\mathrm{var}(\hat{g}^{\,t,t-13}_{\mathrm{act}})$. For branches with another SIC code it needs to be checked whether $\mathrm{var}(\hat{g}^{\,t,t-13}_{\mathrm{act}}) \approx \mathrm{var}(\hat{g}^{\,t,t-13})$ since the impact of the refreshment in January need not be negligible.

Acknowledgements

The views expressed in the paper are those of the authors and do not necessarily reflect the policy of Statistics Netherlands. The authors would like to thank the Associate Editor and two anonymous referees for their useful comments and suggestions, which have led to a significant improvement of this paper.

Appendix A

Justification of (5)

Firstly, consider the case of strata without births and deaths. Apart from the yearly update in January, there are now no monthly updates. Hence, $n_{h\ell}^{t} = n_{h\ell,\mathrm{REQ}}^{\mathrm{dec,\,jan}}$ is fixed from which (5) follows. This case applies to the Dutch Supermarkets because that population has been quite stable over the years. Secondly, in case of births and deaths among the strata we can write $n_{h\ell}^{t}$ as

$$n_{h\ell}^{t} = n_{\ell}^{t} - n_{0\ell}^{t-12,\,t} - \sum_{k\neq h}^{H} n_{k\ell}^{t}, \qquad (16)$$

where $n_{0\ell}^{t-12,\,t}$ or, for short, $n_{0\ell}^{t}$ stands for the number of births in months $t-12, ..., t-1$ among s_{ℓ}^{t}. Because the sampling procedure among the new births after month $t-12$ is independent of the n_{hg}^{t-12}, the random variables $n_{0\ell}^{t-12,\,t}$ and n_{hg}^{t-12} have a zero covariance. Furthermore, using $\mathrm{cov}(n_{hg}^{t-12}, n_{k\ell}^{t}) = 0$ for $k \neq h$, it is seen from (16) that $\mathrm{cov}(n_{hg}^{t-12}, n_{h\ell}^{t}) = 0$ $(h = 1, ..., H)$.

In fact, it is assumed so far that the distribution of $n_{k\ell}^{t}$ $(k \neq h)$ can be described by a hypergeometric distribution with parameters $(N_{\ell}^{t}, N_{k\ell}^{t-12,\,t}, n_{\ell}^{t})$ irrespective of the values of the n_{hg}^{t-12}. A similar remark applies to $n_{0\ell}^{t-12,\,t}$. However, it can be argued that in practice these assumptions lead to a minor, second-order error in the variance formulas. In order to trace this error, we assume for simplicity's sake and without loss of generality that (*i*) births and deaths do not migrate between strata, (*ii*) there are no deaths among the births, (*iii*) $n_{0h}^{t} = f_h N_{0h}^{t}$ is fixed, (*iv*) after their first month in the population births are irrelevant for the monthly updates during the rest of the study period and (*v*) deaths are not selected in or removed from the sample by the monthly updates; so a third-order error is still ignored. Under these assumptions we now look more closely at the second covariance component for $\ell = h$, say $C_{hh,\,\mathrm{sec}}$, from (4a). In analogy with (6) $C_{hh,\,\mathrm{sec}}$ can be written as

$$\begin{aligned} C_{hh,\,\mathrm{sec}} &\equiv \frac{1}{n_h^{t-12} n_h^{t}} \\ &\quad \mathrm{cov}\left\{ E\left(\sum_{g=1}^{H+1} n_{hg}^{t-12}\, \bar{o}_{hg}^{\,t-12} | \nu_h \right), E\left(\sum_{k=0}^{H} n_{kh}^{t}\, \bar{o}_{kh}^{\,t} | \nu_h \right) \right\} \\ &= \frac{1}{n_h^{t-12} n_h^{t}} \sum_{g=1}^{H+1} \sum_{k=1}^{H} \bar{O}_{hg}^{\,t-12}\, \bar{O}_{kh}^{\,t}\, \mathrm{cov}(n_{hg}^{t-12}, n_{kh}^{t}), \end{aligned} \qquad (17)$$

where $\nu_h = (n_{h1}^{t-12}, ..., n_{h,H+1}^{t-12}, n_{1h}^{t}, ..., n_{Hh}^{t})$. Note that under the above assumptions $C_{h\ell,\mathrm{sec}} = 0$ for $\ell \neq h$.

To estimate the covariances in (17), consider the formula for the conditional expectation of y given $x = x_0$ when y and x follow a bivariate normal distribution. That is, in standard notation,

$$E(y | x_0) = \mu_y + \frac{\sigma_{yx}}{\sigma_x^2}(x_0 - \mu_x).$$

In addition, for a given change Δx_0 of x the conditional expectation of the change of y is equal to $E(\Delta y \,|\, \Delta x_0) = \sigma_{yx} \Delta x_0 / \sigma_x^2$ or, equivalently,

$$\sigma_{yx} = \frac{E(\Delta y \,|\, \Delta x_0)}{\Delta x_0} \sigma_x^2. \qquad (18)$$

So for estimating, for instance, $\mathrm{cov}(n_{h,H+1}^{t-12}, n_{kh}^{t})$ in (17) under normality it suffices to evaluate the expected effect on $y = n_{kh}^{t}$ caused by a change of the future deaths $x = n_{h,H+1}^{t-12}$ in s_h^{t-12}.

Let $\Delta n_{h,H+1}^{t-12}$ denote an additional (positive) change of these deaths in s_h^{t-12}. Define $p_{h,H+1}^{\mathrm{jan},t}$ by $p_{h,H+1}^{\mathrm{jan},t} = N_{h,H+1}^{\mathrm{jan},t} / N_{h,H+1}^{t-12}$ where $N_{h,H+1}^{\mathrm{jan},t}$ is the number of deaths in stratum h between January and month t. Also, $p_{hg}^{t-12} = N_{hg}^{t-12,\,t} / N_h^{t-12}$ $(g = 1, ..., H+1)$. Using assumption (*v*), the expected number of additional deaths in the sample of January before

the refreshment can be estimated by $p_{h,\,H+1}^{\text{jan},t}\Delta n_{h,\,H+1}^{t-12}$. Subsequently, the expected number of additional deaths in the sample after the refreshment can be estimated by

$$\gamma_{\text{red}}^{\text{jan}}\; p_{h,H+1}^{\text{jan},t}\,\Delta n_{h,H+1}^{t-12};$$
$$\gamma_{\text{red}}^{\text{jan}} = (0.9 - f_h)/(1 - f_h), \qquad (19)$$

where $\gamma_{\text{red}}^{\text{jan}}$ is the reduction factor due to the refreshment in January. For the derivation of (19), see the end of this appendix. The corresponding monthly updates between January and month t due to these additional deaths in the sample from stratum h lead to the following estimate of the expected increase of *incoming* units n_{kh}^t from stratum k $(k \neq h)$ in the sample of month t

$$E(\Delta n_{kh}^t \,|\, \Delta n_{h,H+1}^{t-12}) = \gamma_{\text{red}}^{\text{jan}}\; p_{h,H+1}^{\text{jan},t}\,\Delta n_{h,H+1}^{t-12}\, p_{kh}^t, \qquad (20)$$

where $p_{kh}^t = N_{kh}^{t-12,t}/(N_h^t - N_{0h}^t)$. Recall from subsection 2.1 that an update in month s occurs only when $d_h^{s-1} \neq f_h D_h^{s-1}$, where D_h^s (d_h^s) stands for the number of deaths in U_h^s (s_h^s), and that $n_{kh}^t = f_h N_{kh}^{t-12,t}$ is fixed when $N_{h,H+1}^{\text{jan},t} = 0$ $(k \neq h)$. Furthermore, note that births are excluded in the definition of p_{kh}^t in (20) because of assumption (*iv*).

Next, define for $m = 0, 12$

$$\bar{O}_h^{t-m} = \frac{1}{N_h^{t-m}} \sum_{i=1}^{N_h^{t-m}} O_{hi}^{t-m};$$
$$(S_h^{t-m})^2 = \frac{1}{N_h^{t-m} - 1} \sum_{i=1}^{N_h^{t-m}} (O_{hi}^{t-m} - \bar{O}_h^{t-m})^2;$$
$$p_{h,\le H}^{t-12} = 1 - p_{h,H+1}^{t-12};$$
$$p_{in,\,h}^t = 1 - p_{hh}^t;$$
$$\bar{O}_{h,\le H}^{t-12} = \sum_{g=1}^{H} \frac{p_{hg}^{t-12}}{p_{h,\le H}^{t-12}}\, \bar{O}_{hg}^{t-12};$$
$$\bar{O}_{in,\,h}^t = \sum_{\substack{k=1 \\ k\neq h}}^{H} \frac{p_{kh}^t}{p_{in,\,h}^t}\, \bar{O}_{kh}^t.$$

Now using (18) and (20), we obtain for $k \neq h$ the following covariance approximation

$$\begin{aligned}
&\text{acov}(n_{h,\,H+1}^{t-12},\, n_{kh}^t) \\
&\quad = \frac{E(\Delta n_{kh}^t \,|\, \Delta n_{h,H+1}^{t-12})}{\Delta n_{h,\,H+1}^{t-12}} \text{var}(n_{h,\,H+1}^{t-12}) \\
&\quad \approx \gamma_{\text{red}}^{\text{jan}}\; p_{h,H+1}^{\text{jan},t}\, p_{kh}^t\, n_h^{t-12}\, p_{h,H+1}^{t-12}(1 - p_{h,H+1}^{t-12})(1 - f_h) \\
&\quad = n_h^{t-12}\, p_{kh}^t A_h / p_{in,\,h}^t; \qquad (21)
\end{aligned}$$

$$A_h = \gamma_{\text{red}}^{\text{jan}}\; p_{in,\,h}^t\; p_{h,H+1}^{\text{jan},t}\; p_{h,H+1}^{t-12}(1 - p_{h,H+1}^{t-12})(1 - f_h),$$

where, for simplicity, we omitted the term $N_h^{t-12}/(N_h^{t-12} - 1)$ in the second line. Because n_h^{t-12} is fixed, it holds that $\text{cov}(n_{h,\,H+1}^{t-12},\, n_{kh}^t) = -\text{cov}(n_{h1}^{t-12} + \ldots + n_{hH}^{t-12},\, n_{kh}^t)$. Hence, in analogy with the multihypergeometric distribution we can use for $1 \le g \le H$ and $k \neq h$ the following relationship for an approximation of $\text{cov}(n_{hg}^{t-12},\, n_{kh}^t)$

$$\text{acov}(n_{hg}^{t-12},\, n_{kh}^t) = -\frac{p_{hg}^{t-12}}{p_{h,\le H}^{t-12}}\, \text{acov}(n_{h,\,H+1}^{t-12},\, n_{kh}^t) \qquad (22a)$$

$$= -\, n_h^{t-12}\, \frac{p_{hg}^{t-12}}{p_{h,\le H}^{t-12}}\, \frac{p_{kh}^t}{p_{in,\,h}^t} A_h, \qquad (22b)$$

where (21) is used as well. Alternatively, note that

$$\begin{aligned}
\text{cov}(n_{h,\,H+1}^{t-12},\, n_{kh}^t) &= -\text{cov}(n_{h,\le H}^{t-12},\, n_{kh}^t) \\
&= -\sum_{g\le H} \sum_{i \in U_{hg}^{t-12,\,t}} \text{cov}(\delta_{hgi}^{t-12},\, n_{kh}^t),
\end{aligned}$$

where

$$\delta_{hgi}^{t-12} = \begin{cases} 1 & \text{if } i^{\text{th}} \text{ unit in } U_{hg}^{t-12,\,t} \text{ is included in sample } s_h^{t-12} \\ 0 & \text{otherwise.} \end{cases}$$

Hence, by symmetry, $\text{cov}(\delta_{hgi}^{t-12},\, n_{kh}^t) = -\text{cov}(n_{h,\,H+1}^{t-12},\, n_{kh}^t)/N_{h,\le H}^{t-12,\,t}$ from which (22a) follows $(1 \le g \le H)$. Likewise, for $k = h$ we obtain from (21) and (22b)

$$\begin{aligned}
\text{acov}(n_{h,\,H+1}^{t-12},\, n_{hh}^t) &= -n_h^{t-12} A_h; \\
\text{acov}(n_{hg}^{t-12},\, n_{hh}^t) &= n_h^{t-12}\, p_{hg}^{t-12} A_h / p_{h,\le H}^t,
\end{aligned} \qquad (23)$$

respectively $(1 \le g \le H)$. Now substituting (21)-(23) into (17), we get the approximation

$$C_{hh,\,\text{sec}} = A_h(\bar{O}_{h,\,H+1}^{t-12} - \bar{O}_{h,\le H}^{t-12})(\bar{O}_{in,\,h}^t - \bar{O}_{hh}^t)/n_h^t. \qquad (24)$$

Assuming that the two terms between parentheses in (24) are absolutely smaller than S_h^t, it follows from (24) that

$$\left|C_{hh,\,\text{sec}}\right| \le \frac{\gamma_{\text{red}}^{\text{jan}}\; p_{h,H+1}^{\text{jan},\,t}\; p_{in,\,h}^t\; p_{h,H+1}^{t-12}(1 - p_{h,H+1}^{t-12})(1 - f_h)}{n_h^t} (S_h^t)^2.$$

Hence, when $p_{in,\,h}^t,\; p_{h,\,H+1}^{t-12} \le 0.1$, we may conclude that under the above assumptions the contribution of the second covariance component is less than 1% of $\text{var}(\bar{o}_h^t)$ so that (5) can be used without severely affecting the results. When $C_{hh,\text{sec}}$ is non-negligible, it can be estimated from the sample according to (24) by

$$\hat{C}_{hh,\text{sec}} = A_h\{(\bar{o}_{h,\,H+1}^{t-12} - \bar{o}_{h,\le H}^{t-12})(\bar{o}_{in,\,h}^t - \bar{o}_{hh}^t) - \hat{\text{cov}}(\bar{o}_{h,\le H}^{t-12}, \bar{o}_{hh}^t)\}/n_h^t, \qquad (25)$$

where in analogy with (10) and (12) $\hat{\text{cov}}(\bar{o}_{h,\le H}^{t-12}, \bar{o}_{hh}^t)$ is defined by

$$\hat{\text{cov}}(\bar{o}_{h,\le H}^{t-12}, \bar{o}_{hh}^t) = \frac{n_{hh}^{t-12}}{n_{h,\le H}^{t-12}} \left(\frac{n_{hh}^{t-12,\,t}}{n_{hh}^{t-12} n_{hh}^t} - \frac{1}{N_{hh}^{t-12,\,t}} \right) \hat{\rho}_{hh,\,\text{OLP}}^{\,t-12,\,t}\, \hat{S}_{hh}^{\,t-12}\, \hat{S}_{hh}^{\,t}.$$

We used in (25) that (*i*) for two arbitrary (unbiased) estimators $\hat{a}$ and $\hat{b}$, $E(\hat{a}\hat{b}) = ab + \text{cov}(\hat{a}, \hat{b})$ and (*ii*) $\text{cov}(\bar{o}_{hg}^{t-12}, \bar{o}_{kh}^t) = 0$ $(g \neq h$ or $k \neq h)$.

We conclude this appendix with the derivation of (19). The expected number of additional deaths remaining in the sample of January during the refreshment is $0.9\, p_{h,\,H+1}^{\text{jan},\,t}\, \Delta n_{h,\,H+1}^{t-12}$. The number of deaths outside the sample just

before the refreshment can be estimated by $N_{h,H+1}^{\text{jan},t} - n_{h,H+1}^{\text{jan}} - p_{h,H+1}^{\text{jan},t}\Delta n_{h,H+1}^{t-12}$. Hence, the number of *new* deaths in the sample due to the refreshments in all substrata $U_{hg}^{t-12,t}$ $(1 \le g \le H)$ in January can be estimated by

$$0.1(n_h^{\text{jan}} - n_{0h}^{\text{jan}})\frac{N_{h,H+1}^{\text{jan},t} - n_{h,H+1}^{\text{jan}} - p_{h,H+1}^{\text{jan},t}\Delta n_{h,H+1}^{t-12}}{N_h^{\text{jan}} - N_{0h}^{t-12,\,\text{jan}} - (n_h^{\text{jan}} - n_{0h}^{\text{jan}})}.$$

Now using $n_{0h}^{\text{jan}} = f_h N_{0h}^{t-12,\,\text{jan}}$ according to the above assumptions, it is seen that after the refreshments the final number of additional deaths in the sample due to $\Delta n_{h,H+1}^{t-12}$ can be estimated by

$$\left\{0.9 - \frac{0.1(n_h^{\text{jan}} - n_{0h}^{\text{jan}})}{N_h^{\text{jan}} - N_{0h}^{t-12,\,\text{jan}} - (n_h^{\text{jan}} - n_{0h}^{\text{jan}})}\right\} p_{h,H+1}^{\text{jan},t}\,\Delta n_{h,H+1}^{t-12}$$
$$= \frac{0.9 - f_h}{1 - f_h}\, p_{h,H+1}^{\text{jan},t}\,\Delta n_{h,H+1}^{t-12} = \gamma_{\text{red}}^{\text{jan}}\, p_{h,H+1}^{\text{jan},t}\,\Delta n_{h,H+1}^{t-12}.$$

Appendix B

Some useful covariance formulas for overlapping samples

Let s_{123} denote a mother sample consisting of three mutually disjoint SRS subsamples s_1, s_2 and s_3. Let the variable x be observed in s_{12} and the variable y in s_{23}. The corresponding sample means are denoted by $\bar{x}_{12}$ and $\bar{y}_{23}$, respectively. Denote the size of s_k by n_k ($k = 1, 2, 3, 12, 23$). Define $\lambda = n_2/n_{12}$, $\mu = n_2/n_{23}$ and $f_k = n_k/N$. Furthermore, define S_{xy} by

$$S_{xy} = \frac{1}{N-1}\sum_{j=1}^{N}(X_j - \bar{X}_p)(Y_j - \bar{Y}_p)\,.$$

Then the covariance between $\bar{x}_{12}$ and $\bar{y}_{23}$ is equal to

$$\text{cov}(\bar{x}_{12}, \bar{y}_{23}) = \left(\frac{\lambda\mu}{n_2} - \frac{1}{N}\right)S_{xy} = \left(\frac{\mu}{n_{12}} - \frac{1}{N}\right)S_{xy} = \left(\frac{\lambda}{n_{23}} - \frac{1}{N}\right)S_{xy}. \quad (26)$$

This can be shown as follows

$$\begin{aligned}
&\text{cov}(\bar{x}_{12}, \bar{y}_{23}) \\
&= \text{cov}\{(1-\lambda)\bar{x}_1 + \lambda\bar{x}_2, \mu\bar{y}_2 + (1-\mu)\bar{y}_3\} \\
&= (1-\lambda)\text{cov}(\bar{x}_1, \bar{y}_{23}) + \lambda\mu\,\text{cov}(\bar{x}_2, \bar{y}_2) \\
&\quad + \lambda(1-\mu)\text{cov}(\bar{x}_2, \bar{y}_3) \\
&= -(1-\lambda)\frac{S_{xy}}{N} + \lambda\mu\left(\frac{1}{n_2} - \frac{1}{N}\right)S_{xy} - \lambda(1-\mu)\frac{S_{xy}}{N} \\
&= \left(\frac{\lambda\mu}{n_2} - \frac{1}{N}\right)S_{xy} = \left(\frac{\mu}{n_{12}} - \frac{1}{N}\right)S_{xy} = \left(\frac{\lambda}{n_{23}} - \frac{1}{N}\right)S_{xy}.
\end{aligned}$$

In the third line we used that $\text{cov}(\bar{x}_1, \bar{y}_{23}) = \text{cov}(\bar{x}_2, \bar{y}_3) = -S_{xy}/N$. This follows from the conditional covariance formula

$$\begin{aligned}
\text{cov}(\bar{x}_2, \bar{y}_3) &= E\{\text{cov}(\bar{x}_2, \bar{y}_3|s_2)\} + \text{cov}\{E(\bar{x}_2|s_2), E(\bar{y}_3|s_2)\} \\
&= 0 + \text{cov}\left\{\bar{x}_2, \frac{\bar{Y}_p - f_2\bar{y}_2}{1 - f_2}\right\} \\
&= -\frac{f_2}{1 - f_2}\,\text{cov}(\bar{x}_2, \bar{y}_2) = -\frac{S_{xy}}{N}.
\end{aligned}$$

For an alternative proof based on the sampling autocorrelation coefficient, see Knottnerus (2003, page 375).

References

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

Brodie, P. (2003). Review of recent work on variance estimation methods in business surveys. Unpublished report, Office for National Statistics, London.

Cox, D.R., and Hinkley, D.V. (1974). *Theoretical Statistics.* London: Chapman and Hall.

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.

Holt, D., and Skinner, C.J. (1989). Components of change in repeated surveys. *International Statistical Review*, 57, 1-18.

Holt, D., and Smith, T.M.F. (1979). Poststratification. *Journal of the Royal Statistical Society*, A, 142, 33-46.

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

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

Knottnerus, P., and Van Delden, A. (2006). Estimation of changes in repeated surveys and their significance, http://www.iser.essex.ac.uk/ulsc/mols2006/programme/data/paper/Knottnerus.doc.

Konschnik, C.A., Monsour, N.J. and Detlefsen, R.E. (1985). Constructing and maintaining frames and samples for business surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 113-122.

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

Lowerre, J.M. (1979). Sampling for change. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 343-347.

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.

Smith, P., Pont, M. and Jones, T. (2003). Developments in business survey methodology in the Office for National Statistics, 1994-2000. *The Statistician*, 52, 257-295.

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.