Panels

Business Panels

10

Paul Knottnerus

Statistical Methods (201119)

Statistics Netherlands

The Hague/Heerlen, 2011

Explanation of symbols

.	= data not available
*	= provisional figure
**	= revised provisional figure
x	= publication prohibited (confidential figure)
–	= nil or less than half of unit concerned
–	= (between two figures) inclusive
0 (0,0)	= less than half of unit concerned
blank	= not applicable
2010–2011	= 2010 to 2011 inclusive
2010/2011	= average of 2010 up to and including 2011
2010/'11	= crop year, financial year, school year etc. beginning in 2010 and ending in 2011
2008/'09–2010/'11	= crop year, financial year, etc. 2008/'09 to 2010/'11 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
Facility Services

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: 1876-0333

60165201119 X-37

Table of contents

1. Introduction to the Business Panels subtheme

1.1 General description

Panels in which the same units of a population are observed in multiple periods have become increasingly common since the 1980s. Panels are an attractive option mainly when the aim is to measure changes in the population statistics of certain target variables.

This first part of this document is restricted to panels for estimating revenue changes in companies and changes in price indexes, such as the producer price index (PPI). However, the same methods, suitably modified, can also be applied to panels of persons and households, for instance to explore unemployment statistics. Because each statistic has its own characteristics, the modifications tend to be correspondingly specific, and therefore difficult to capture in a single general document.

Alongside the estimators concerned, there is also discussion of the variances of the different estimators. The document also discusses changes in a panel, replacements for companies that drop out or cease to exist, and the impact on a panel of company mergers.

This document also amply covers the derivation of the variances of estimated growth rates, since a recurring question is about the margin of uncertainty in growth figures published by Statistics Netherlands. External pressure is often involved. The subject receives scant treatment elsewhere in the literature. Only Nordberg (2000) goes into any detail on the subject, and then the orientation is entirely on the situation in Sweden. Another disadvantage of his approach is that simulations are still needed to estimate certain components of various covariances. Chapter 3 of this document, which deals with the migration of companies between strata, may strike some readers as particularly technical in nature.

Chapter 2 discusses details of panels of companies that are divided into multiple strata. The main focus is on estimating the relative change of a revenue total in a given month compared with 12 months previously, plus the associated margins of uncertainty. It is assumed in this chapter that the companies do not move from one stratum to another, although ‘births’ and ‘deaths’ may occur among the companies. Chapter 3 introduces panels in which companies may move from one size category to another because of a change in the number of active employees in the corresponding 12-month period. Chapter 4 discusses estimating indexes and changes in indexes relative to a base period. This chapter also discusses the weighting of companies involved in a merger or demerger.

1.2 Scope and relationship with other themes

It will be clear that there is a strong connection between the theory for panels and the subthemes of *Sample design* and *Weighting methods*. The reader is deemed to be

conversant with the material presented in these subthemes, in particular the stratified estimator, the ratio estimator and the related estimator of a ratio, including the associated variance formulas. There is also a relationship with the *Weighting as an adjustment for nonresponse* theme and with the *Duration models* subtheme. This document nonetheless tacitly assumes, unless stated to the contrary, that nonresponse does not lead to systematic bias. There is also a relationship with the *Indexes* theme.

1.3 Place in the statistical process

Panels have an important role in many places in the statistical process. The design of panels and the associated sampling demand sound preparation at the start of the process, whereas the calculations for the corresponding estimates with the associated margins of uncertainty occur more towards the end of the statistical process.

1.4 Definitions

Term	Description
Simple random sample	A sample design in which the elements are selected with equal probabilities from the sampling frame. Sampling without replacement requires each subset of sample size to have the same probability of realization. Sampling with replacement requires the sampling of the successive elements to be mutually independent.
Index	The ratio between a value and a statistic that is taken as a reference; usually expressed as a percentage.
Inclusion probability	The probability that an element will be selected in sampling without replacement; depending on the sample design, the probability may differ from element to element.
Panel survey	A survey in which the same observation units are approached more than once at different times in order to track a (micro or macro) trend in time.
Population parameter	A value to be estimated of some characteristic of a population, e.g. the mean score in the population for a certain variable.
Ratio estimator	An estimator for a population total that uses a quantitative auxiliary variable. The accuracy of this estimator increases with decreasing variation in the ratios of the scores on the target and auxiliary variables. The term ‘ratio estimator’ should not be confused with the estimator of a ratio, which is referred to in Chapter 4.

Sample allocation	The distribution of the total sample size over the various strata in stratified sampling.
Systematic sampling	A sample design in which elements are selected from a sampling frame by proceeding through the frame in a systematic manner. Only the starting point is chosen arbitrarily, after which a fixed step length is used.
Bias (of an estimator)	The expected difference between an estimator and the actual value of a (population) parameter, also referred to as ‘systematic error’. If the difference is zero, the estimator is said to be unbiased, and otherwise to be biased.

2. Panels from business populations without migration

2.1 Short description

Consider a population of companies divided into strata according to size category and standard industrial classification (SBI) code. It is assumed that company births and deaths can occur in the population, but companies cannot move from one stratum to another.

The most important objective of a panel in this context is to estimate for each month t (t = 1, 2, …) both the total revenue of the target population and the relative change in the revenue total in the month concerned compared with 12 months ago. The above includes estimating the associated margins of uncertainty. Estimating the relative change in a revenue total starts from the estimates of the revenue totals in the corresponding months. The ratio estimator therefore has an important role in the methodology for panels.

The monthly panel that we consider in this chapter involves simple random sampling without replacement (SRSWOR) from the various strata. The sampling fractions are fixed, but may differ between strata. As a rule, the higher the size category, the higher the sampling fraction. The largest size categories are usually observed in full. The companies in the panel may differ from month to month.

2.2 Applicability

Panels are used in both economic and socioeconomic statistics. The reason for their use is that, compared with completely independent sampling, panels can often produce more accurate estimates of changes in a wide variety of population totals between two months. This chapter is applicable to the case when hardly any units move between different strata.

2.3 Detailed description

2.3.1 Introduction

This chapter derives formulas in particular for estimators of 12-month growth figures of the monthly revenue of companies with a given SBI code. First we consider populations without births and deaths. Section 2.3.6 considers populations in which births and deaths occur. We also assume that 10 per cent of the sample is refreshed in January each year in every stratum.

The effects of correction, imputation and other calculations are initially disregarded in deriving the formulas for the margins of uncertainty, as is nonresponse.

2.3.2 Estimator for the revenue total of the population

Let $O^t \quad (t = 1,2,...)$ be defined as the revenue total in month t of all companies in the target population with a certain SBI code, or $O^t = \Sigma_{i=1}^{N} O_i^t$ where O_i^t represents the

revenue of company i in month t (i=1,…, N^t). As stated in Banning et al. (2010) and elsewhere, O^t can be estimated with the following stratification estimator based on H size categories

$$\hat{O}^t = \sum_{h=1}^{H} N_h^t \bar{o}_h^t ,$$

where

N_h^t: number of companies in stratum/size category h $(h = 1, \cdots, H)$ in month t;

$\bar{o}_h^t$: mean revenue in sample s_h^t in month t in stratum h $(\bar{o}_h^t = \frac{1}{n_h^t} \sum_{i=1}^{n_h^t} o_{hi}^t)$;

o_{hi}^t: revenue of i^{th} company in the sample from stratum h;

n_h^t: number of companies in the sample from stratum h in month t.

Define further

$n_{hh}^{t-12,t}$: number of companies in the overlap of the samples from stratum h in months t-12 and t;

$\bar{O}_h^t$: mean revenue per company in stratum h in month t $(\bar{O}_h^t = \frac{1}{N_h^t} \sum_{i=1}^{N_h^t} O_{hi}^t)$;

O_{hi}^t: revenue of i^{th} company from stratum h;

f_h: sampling fraction in stratum h.

Because the assumption of no births and deaths means that N_h^t and n_h^t do not depend on t, it would be acceptable to omit the superscript t. However, because the formulas with superscript t or t-12 are also of use in the more general situation in which births and deaths occur in the population, the superscripts t are retained in this section. The variance of the estimator $\hat{O}^t$ is

$$\begin{aligned} \mathrm{var}(\hat{O}^t) &= \sum_{h=1}^{H} (N_h^t)^2 (1 - f_h) \frac{(S_h^t)^2}{n_h^t} \\ (S_h^t)^2 &= \frac{1}{N_h^t - 1} \sum_{i=1}^{N_h^t} (O_{hi}^t - \bar{O}_h^t)^2 . \end{aligned} \qquad (2.1)$$

The variance of $\hat{O}^t$ can be estimated with

$$\begin{aligned} \hat{\mathrm{var}}(\hat{O}^t) &= \sum_{h=1}^{H} (N_h^t)^2 (1 - f_h) \frac{(s_h^t)^2}{n_h^t} \\ (s_h^t)^2 &= \frac{1}{n_h^t - 1} \sum_{i=1}^{n_h^t} (o_{hi}^t - \bar{o}_h^t)^2 . \end{aligned}$$

Comparable formulas are applicable to $\hat{O}^{t-12}$.

2.3.3 Estimator for a 12-month growth percentage

Let $g^{t,s}$ be defined as the relative revenue growth between the months *s* and *t*, or

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

The statistic $g^{t,s}$ can be estimated with

$$\hat{g}^{t,s} = \frac{\hat{O}^t}{\hat{O}^s} - 1 \tag{2.2}$$

We will now follow the steps of the derivation of an expression for $\text{var}(\hat{g}^{t,t-12})$. The variance of the right-hand side of (2.2) can be approximated with a first-order Taylor series expansion of a ratio. This gives for *s*=*t*-12

$$\begin{aligned}
\text{var}(\hat{g}^{t,t-12}) &= \text{var}\left\{\frac{\hat{O}^t}{\hat{O}^{t-12}}\right\} = \frac{1}{(O^{t-12})^2}\,\text{var}(\hat{O}^t - G^{t,t-12}\hat{O}^{t-12}) \\
&= \frac{1}{(O^{t-12})^2}\{\text{var}(\hat{O}^t) + (G^{t,t-12})^2\,\text{var}(\hat{O}^{t-12}) \\
&\quad - 2G^{t,t-12}\,\text{cov}(\hat{O}^{t-12}, \hat{O}^t)\} \qquad (2.3) \\
G^{t,t-12} &\equiv \frac{O^t}{O^{t-12}} = 1 + g^{t,t-12}.
\end{aligned}$$

The two variance terms in (2.3) can be estimated in the standard way as given in the previous subsection. The term $G^{t,t-12}$ can be estimated with

$$\hat{G}^{t,t-12} \equiv \frac{\hat{O}^t}{\hat{O}^{t-12}}.$$

The covariance term $\text{cov}(\hat{O}^{t-12}, \hat{O}^t)$ in (2.3) is

$$\begin{aligned}
\text{cov}(\hat{O}^{t-12}, \hat{O}^t) &= \sum_{h=1}^{H} \text{cov}(\hat{O}_h^{t-12}, \hat{O}_h^t) \\
&= \sum_{h=1}^{H} \text{cov}(N_h^{\text{t-12}}\bar{o}_h^{t-12}, N_h^t\bar{o}_h^t) \qquad (2.4) \\
&= \sum_{h=1}^{H} N_h^{t-12}N_h^t\,\text{cov}(\bar{o}_h^{t-12}, \bar{o}_h^t),
\end{aligned}$$

which relies on the mutual independence of the strata samples. The derivation of the expression for the covariances in the last line of (2.4) relies on the fixed value of $n_{hh}^{t-12,t} = 0.9n_h^{t-12}$, which follows from the above assumption that 10% of the sample in every stratum is replaced in January by different companies from outside the sample. It follows from (2.12) in the appendix (Section 2.5) that

$$\mathrm{cov}(\bar{o}_h^{t-12}, \bar{o}_h^t) = \left(\frac{n_{hh}^{t-12,t}}{n_h^{t-12} n_h^t} - \frac{1}{N_{hh}^{t-12,t}}\right) S_{hh}^{t-12,t} \tag{2.5}$$

$$S_{hh}^{t-12,t} = \frac{1}{N_{hh}^{t-12,t} - 1} \sum_{i=1}^{N_{hh}^{t-12,t}} (O_{hi}^{t-12} - \bar{O}_h^{t-12})(O_{hi}^t - \bar{O}_h^t).$$

$N_{hh}^{t-12,t}$ represents the number of companies that were included in stratum *h* in both month *t*-12 and month *t*. Under the assumption of no births and deaths in the population, then $N_{hh}^{t-12,t} = N_h^{t-12} = N_h^t$. It follows from (2.5) that $\mathrm{cov}(\hat{O}_h^{t-12}, \hat{O}_h^t)$ from the first line of (2.4) can be estimated without bias with

$$\hat{\mathrm{cov}}(\hat{O}_h^{t-12}, \hat{O}_h^t) = N_h^{t-12} N_h^t \left(\frac{n_{hh}^{t-12,t}}{n_h^{t-12} n_h^t} - \frac{1}{N_{hh}^{t-12,t}}\right) s_{hh}^{t-12,t}$$

$$s_{hh}^{t-12,t} = \frac{1}{n_{hh}^{t-12,t} - 1} \sum_{i=1}^{n_{hh}^{t-12,t}} (o_{hi}^{t-12} - \bar{o}_h^{t-12})(o_{hi}^t - \bar{o}_h^t).$$

2.3.4 The importance of having a substantial overlap between panels

In order to clarify the importance of having a large overlap between panels, we introduce the simplifying assumptions that $S_h^{t-12} = S_h^t = S_h$, $n_h^{t-12} = n_h^t = n_h$ and $G^{t,t-12} = 1$. Define further the overlap ratio λ_h as $\lambda_h = n_{hh}^{t-12,t} / n_h$. Because of the above assumption $N_{hh}^{t-12,t} = N_h^{t-12} = N_h^t = N_h$ an upper limit for the covariances in the first line of (2.4) follows from (2.5)

$$\mathrm{cov}(\hat{O}_h^{t-12}, \hat{O}_h^t) \le N_h^2 (\lambda_h - f_h) \frac{S_h^2}{n_h}.$$

In combination with (2.1) and (2.3) this gives the following lower limit for $\mathrm{var}(\hat{g}^{t,t-12})$

$$\mathrm{var}(\hat{g}^{t,t-12}) \ge \frac{2}{(O^{t-12})^2} \sum_{h=1}^{H} N_h^2 (1 - \lambda_h) \frac{S_h^2}{n_h}. \tag{2.6}$$

This formula indicates that if the overlap ratios λ_h decrease from e.g. 0.9 to 0.5, the lower limit for the variance of estimator $\hat{g}^{t,t-12}$ will increase by a factor of 5. The lower limit in the above formula is reached when, in addition to the above-mentioned assumptions, $corr(O_{hi}^{t-12}, O_{hi}^t) = 1$ also applies, or when the correlation coefficient between the O_{hi}^{t-12} and the O_{hi}^t is 1. As an extension to this reasoning, it is observed that the size of the overlap is entirely irrelevant in the extreme situation that $S_{hh}^{t-12,t} = 0$.

2.3.5 The importance and extent of annual panel refreshing

In revenue statistics in particular, it is important for changes in the population to be incorporated rapidly in the panel, because the revenue itself is now the target variable . There is somewhat less urgency in this respect for panels for price indexes

and suchlike, where revenue serves only as a weight, and the correlation between revenue weight and price changes tends to be modest. As well as in connection with changes in the population, the refreshing of a panel, or a rotating panel, is also needed as a way of limiting a panel's bias. A panel may also exhibit bias as a result of panel attrition, which may occur in the event of death or refusal to continue as a panel member. A third and final objective of refreshing is to achieve a fairer distribution of the burden on the companies involved.

The first factor examined in exploring the effect of the magnitude of annual refreshing $(1-\lambda_h)$ on the margins of uncertainty surrounding the estimated revenue growth is the *relative* margin of uncertainty surrounding the estimated revenue total under the same assumptions as in the previous subsection. It follows from (2.1) that the relative 95% margin of uncertainty surrounding the estimated revenue total $\hat{O}^{t-k}$ $(k = 0, 12)$, say $RM_{95}(\hat{O}^{t-k})$, can be expressed as

$$RM_{95}(\hat{O}^{t-k}) = \frac{1{,}96}{O^{t-k}} \sqrt{\sum_{h=1}^{H} N_h^2 (1- f_h) \frac{S_h^2}{n_h}}. \tag{2.7}$$

Under the additional assumption that $corr(O_{hi}^{t-12}, O_{hi}^{t}) \approx 1$ it also follows from (2.6) that the *absolute* margin of uncertainty surrounding the estimated revenue growth $\hat{g}^{t,t-12}$ is approximately

$$M_{95}(\hat{g}^{t,t-12}) = \frac{1{,}96}{O^{t-12}} \sqrt{2\sum_{h=1}^{H} N_h^2 (1-\lambda_h) \frac{S_h^2}{n_h}}. \tag{2.8}$$

It follows from these two formulas that $RM_{95}(\hat{O}^{t-12}) = M_{95}(\hat{g}^{t,t-12})$ when $\lambda_h = (1+ f_h)/2$. In other words, when the relative margin of uncertainty is 1%, say, then under the condition $\lambda_h = (1+ f_h)/2$, the absolute margin of uncertainty surrounding the estimated revenue growth is 1 percentage point. The overall formula (2.8) for the margin also shows clearly that as the degree of refreshing increases, $M_{95}(g)$ increases. However, if the assumptions made above are unrealistic, it may be necessary to resort to one of the more complex formulas for $\text{cov}(\hat{O}^{t-12}, \hat{O}^{t})$, which are discussed elsewhere in this document. This is mainly the case when $corr(O_{hi}^{t-12}, O_{hi}^{t})$ is not close to 1.

2.3.6 Births and deaths in the population

In the event of births and deaths among the companies in the target population, almost the same formulas can be used as those in Sections 2.3.2 and 2.3.3. Only the formula for $\text{cov}(\hat{O}^{t-12}, \hat{O}^{t})$ in (2.3) becomes a little more complicated, which has to do with the monthly update procedure of the sample, where account has to be taken of the monthly births and the deaths in the population. The monthly update procedure is described below.

Define $U_{0h}^{t-1,t}$ as the set of births in stratum h in month t-1 and define $N_{0h}^{t-1,t}$ as its size. The number of companies selected from $U_{0h}^{t-1,t}$ in month t is $n_{0h}^{t-1,t} = f_h N_{0h}^{t-1,t}$. The remainder of the sample from stratum h is then adjusted accordingly as follows. Define $n_{h,RST}^{t-1,t}$ as the necessary number of companies for the remainder of the sample, in other words $n_{h,RST}^{t-1,t} = n_h^t - n_{0h}^{t-1,t}$. Note that $n_h^t = f_h N_h^t$ is fixed. Also define the presample $s_{h,PRE}^t$ in month t-1 as $s_{h,PRE}^t = s_h^{t-1} \cap U_h^t$, or, in other words, the set of companies in s_h^{t-1} that still exist in month t. Denote the size of $s_{h,PRE}^t$ with $n_{h,PRE}^t$. If $n_{h,PRE}^t > n_{h,RST}^{t-1,t}$, randomly remove the excess from the sample, and if $n_{h,PRE}^t \leq n_{h,RST}^{t-1,t}$, randomly select the shortfall from $(U_h^t \setminus U_{0h}^{t-1,t}) \setminus s_{h,PRE}^t$. Note that it is possible for companies that are removed in the update process in this way to be reinstated in the sample later.

Based on the monthly update procedure, stratum h in month t-12 can be divided into two substrata: (*i*) substratum $U_{h,H+1}^{t-12,t}$ of companies in stratum h that die in the months t-12, …, t-1 and (*ii*) substratum $U_{hh}^{t-12,t}$ of the other companies in stratum h in month t-12. By analogy, stratum h in month t can also be divided into two substrata: (*i*) substratum $U_{0h}^{t-12,t}$ of all companies born in stratum h in the months t-12, …, t-1 and (*ii*) stratum $U_{hh}^{t-12,t}$ of the other companies in month t. The sizes of the samples from $U_{h,H+1}^{t-12,t}$ and $U_{hh}^{t-12,t}$ in month t-12 are denoted by $n_{h,H+1}^{t-12}$ and n_{hh}^{t-12}. The corresponding sample means of the revenue are denoted with $\bar{o}_{h,H+1}^{t-12}$ and $\bar{o}_{hh}^{t-12}$. For the samples from $U_{0h}^{t-12,t}$ and $U_{hh}^{t-12,t}$ in month t these four statistics are denoted with n_{0h}^t, n_{hh}^t, $\bar{o}_{0h}^t$ and $\bar{o}_{hh}^t$. It is also assumed that deaths do not coincide with births and that, after their first month in the population, births are no longer involved in the monthly updates for the rest of the study period. The advantage of these assumptions is that n_{0h}^t and n_{hh}^t are fixed. Section 3.5 shows that any error introduced in this way is usually small.

$\bar{o}_h^{t-12}$ and $\bar{o}_h^t$ can now be expressed as

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

$$\bar{o}_h^t = \frac{n_{0h}^t}{n_h^t} \bar{o}_{0h}^t + \frac{n_{hh}^t}{n_h^t} \bar{o}_{hh}^t.$$

The covariance in the last line of (2.4) can then be rewritten as

$$\operatorname{cov}(\bar{o}_h^{t-12}, \bar{o}_h^t) = \operatorname{cov}\left(\sum_{g \in \{h, H+1\}} \frac{n_{hg}^{t-12}}{n_h^{t-12}} \bar{o}_{hg}^{t-12}, \sum_{k \in \{0, h\}} \frac{n_{kh}^t}{n_h^t} \bar{o}_{kh}^t \right) \tag{2.9}$$

$$= \frac{1}{n_h^{t-12} n_h^t} \operatorname{cov}(n_{hh}^{t-12} \bar{o}_{hh}^{t-12}, n_{hh}^t \bar{o}_{hh}^t). \tag{2.10}$$

(2.10) assumes that n_{kh}^{t} is fixed and that $\mathrm{cov}(\bar{o}_{hg}^{t-12}, \bar{o}_{kh}^{t}) = 0$ for $g \neq h$ or $k \neq h$ because the corresponding samples were from different substrata.

Finding an expression for the covariance term in (2.10) involves using the formula for conditional covariances with conditioning on $\nu_h = (n_{hh}^{t-12}, n_{hh}^{t-12,t})$

$$\begin{aligned}
\mathrm{cov}(n_{hh}^{t-12}\bar{o}_{hh}^{t-12}, n_{hh}^{t}\bar{o}_{hh}^{t}) &= \mathrm{cov}\{E(n_{hh}^{t-12}\bar{o}_{hh}^{t-12} \mid \nu_h), E(n_{hh}^{t}\bar{o}_{hh}^{t} \mid \nu_h)\} \\
&\quad + E\{\mathrm{cov}(n_{hh}^{t-12}\bar{o}_{hh}^{t-12}, n_{hh}^{t}\bar{o}_{hh}^{t} \mid \nu_h)\} \\
&= \bar{O}_{hh}^{t-12}\bar{O}_{hh}^{t}\,\mathrm{cov}(n_{hh}^{t-12}, n_{hh}^{t}) + n_{hh}^{t}E\{n_{hh}^{t-12}\,\mathrm{cov}(\bar{o}_{hh}^{t-12}, \bar{o}_{hh}^{t} \mid \nu_h)\} \\
&= 0 + E(n_{hh}^{t-12,t} - \frac{n_{hh}^{t-12}n_{hh}^{t}}{N_{hh}^{t-12,t}})S_{hh}^{t-12,t}\,. \qquad (2.11)
\end{aligned}$$

The last line again uses (2.12) in the appendix and that n_{hh}^{t} is fixed. It follows from (2.10) that the covariance in (2.4) can be estimated without bias with

$$\hat{\mathrm{cov}}(\hat{O}_h^{t-12}, \hat{O}_h^{t}) = \frac{N_h^{t-12}N_h^{t}}{n_h^{t-12}n_h^{t}}\, n_{hh}^{t-12,t}(1 - \frac{n_{hh}^{t-12}n_{hh}^{t}}{n_{hh}^{t-12,t}N_{hh}^{t-12,t}})s_{hh}^{t-12,t}.$$

Note that the estimator is unbiased because $E(s_{hh}^{t-12,t} \mid \nu_h) = S_{hh}^{t-12,t}$.

2.4 Quality indicators

Quality indicators for panels closely resemble those of the direct estimator for SRSWOR, i.e.

- the margins of uncertainty of the corresponding estimators for a panel;
- the size of nonresponse;
- the size of the overlap of the panels in months *t* and *t*-12.

Nonresponse can severely affect the quality of the results if (i) the nonresponse is large and (ii) the nonresponse is selective. Adjustment of some of the bias from selective nonresponse can sometimes be achieved by using auxiliary variables that correspond with both the probability of response and the target variable. Something else that may happen with a panel in due course is that a member wishes to withdraw. Drop-out from a panel for this reason is also referred to as sample attrition, or panel attrition.

Another important aspect of panel quality is the size of the overlap. If nonresponse or substantial migration between strata significantly reduce the overlap between the panels in months *t* and *t*-12, the margin of uncertainty of a revenue growth estimate may rise considerably.

2.5 Appendix. Covariance formulas for overlapping samples

Let s_{123} be a master sample comprising three (disjunct) subsamples produced by SRSWOR s_1, s_2 and s_3 (SRSWOR: *simple random sampling without replacement*).

Let variable x be observed in s_{12} and variable y in s_{23}. The associated sample means are denoted $\bar{x}_{12}$ and $\bar{y}_{23}$, respectively The size of s_k is denoted with n_k ($k = 1,2,3,12,23$). Define also $\lambda = n_2 / n_{12}$, $\mu = n_2 / n_{23}$ and $f_k = n_k / N$. The covariance of $\bar{x}_{12}$ and $\bar{y}_{23}$ is then

$$\begin{aligned}\operatorname{cov}(\bar{x}_{12}, \bar{y}_{23}) &= (\frac{\lambda\mu}{n_2} - \frac{1}{N})S_{xy} = (\frac{\mu}{n_{12}} - \frac{1}{N})S_{xy} \\ &= (\frac{\lambda}{n_{23}} - \frac{1}{N})S_{xy} = (\frac{n_2}{n_{12}n_{23}} - \frac{1}{N})S_{xy} \qquad (2.12)\\ S_{xy} &= \frac{1}{N-1}\sum_{j=1}^{N}(X_j - \bar{X}_p)(Y_j - \bar{Y}_p).\end{aligned}$$

In the context of (2.5) s_{12} represents the sample in month t-12 and s_{23} represents the sample in month t, while s_2 represents the overlap between the two samples. N is the number of companies in the target population in both month t-12 and month t. The proof of (2.12) follows from

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

It is assumed in the third line that $\operatorname{cov}(\bar{x}_2, \bar{y}_3) = -S_{xy} / N$ for two sample means from two *disjunct* samples produced by SRSWOR, irrespective of their size. This follows from

$$\begin{aligned}\operatorname{cov}(\bar{x}_2, \bar{y}_3) &= E\{\operatorname{cov}(\bar{x}_2, \bar{y}_3 | \, s_2)\} + \operatorname{cov}\{E(\bar{x}_2 | \, s_2), E(\bar{y}_3 | \, s_2)\}) \\ &= 0 + \operatorname{cov}\{\bar{x}_2, \frac{\bar{Y}_p - f_2\bar{y}_2}{1-f_2}\} \\ &= -\frac{f_2}{1-f_2}\operatorname{cov}(\bar{x}_2, \bar{y}_2) = -\frac{f_2}{1-f_2}(1-f_2)\frac{S_{xy}}{n_2} = -\frac{S_{xy}}{N}.\end{aligned}$$

3. Panels with migration between strata

3.1 Short description

Like the previous chapter, this chapter assumes a population of companies that is divided into strata in accordance with SBI code and size category. The most significant difference with the previous chapter is that account is now taken of the phenomenon that companies may move from one size category to another in response to a change in the number of active employees.

3.2 Applicability

Panels of the kind discussed in this chapter are often used in practice for calculating a certain revenue growth rate for companies in a given SBI category, where companies change size category in the period under consideration. It is then often also relevant to have some indication of the margins of uncertainty surrounding the estimated growth rates.

3.3 Detailed description

3.3.1 Introduction

As mentioned in the previous chapter, there are various methods for refreshing panels. The monthly update for births and deaths is the same as in Section 2.3.6. Furthermore, every year, usually in January, adjustments are made for companies that are no longer in the correct stratum for their size. It is also usual in January each year to refresh 10 per cent of the sample.

3.3.2 Annual adjustment and refreshing in January

The sample is adjusted in January, taking account of the new stratum structure of January and the 10% refreshing of the panel.

All companies that were in the December sample and are still in business in January are again divided into strata in accordance with their current numbers of active employees and SBI codes in January. As a result, the sample obtained from a given stratum may include companies with different inclusion probabilities, because companies may have moved to a different stratum with a different sampling fraction from the stratum that they came from.

In order to adjust the strata for the presence of companies with different inclusion probabilities, a substratum $U_{h\ell}^{dec,jan}$ is introduced, which consists of companies that were in stratum *h* in December and in stratum ℓ in January. The size of this stratum is denoted with $N_{h\ell}^{dec,jan}$ $(h,\ell = 1,...,H)$. By analogy with the monthly update procedure given in Subsection 2.3.6, $s_{h\ell,PRE}^{jan}$ is defined by $s_{h\ell,PRE}^{jan} = s_h^{dec} \cap U_{h\ell}^{dec,jan}$. Let $n_{h\ell,PRE}^{jan}$ be the size of $s_{h\ell,PRE}^{jan}$. Since the required size of $s_{h\ell,RST}^{dec,jan}$ from $U_{h\ell}^{dec,jan}$ in

January should be $n^{dec,jan}_{h\ell,RST} = f_\ell N^{dec,jan}_{h\ell}$, the annual refreshing of the sample $s^{jan}_{h\ell,PRE}$ in January proceeds as follows.

If $n^{jan}_{h\ell,PRE} > n^{dec,jan}_{h\ell,RST}$, randomly remove the excess from $s^{jan}_{h\ell,PRE}$. 10% of the remaining companies are then replaced in the remaining sample by companies from $U^{dec,jan}_{h\ell} \setminus s^{jan}_{h\ell,PRE}$. An assumption is that there are enough companies in this set. If not, only $N^{dec,jan}_{h\ell} - n^{jan}_{h\ell,PRE}$ companies are replaced in the sample. Furthermore, if $n^{jan}_{h\ell,PRE} \le n^{dec,jan}_{h\ell,RST}$ the corresponding shortfall is added from $U^{dec,jan}_{h\ell} \setminus s^{jan}_{h\ell,PRE}$. Subsequently in this case another $n^{jan}_{h\ell,PRE} - .9n^{dec,jan}_{h\ell,RST}$ companies in $s^{dec,jan}_{h\ell,RST}$ are replaced if (*i*) this difference is positive and (*ii*) sufficient new companies are available. This procedure is performed for all substrata $h\ell$, including $h = \ell$. Finally, as with the monthly update procedure, the number of companies to be selected in January from substratum $U^{dec,jan}_{0\ell}$ of new births in stratum ℓ, is $n^{dec,jan}_{0\ell} = f_\ell N^{dec,jan}_{0\ell}$.

Comment. The above procedure may also be performed when the stratum limits of the size categories are changed in January, or when the sampling fractions in strata change.

3.3.3 The covariance term

The monthly revenue totals can be estimated in the manner described in Chapter 2 even if companies change size category in January. The same is true of the corresponding variances and margins of uncertainty. This chapter is therefore restricted to the covariance term $\mathrm{cov}(\hat{O}^{t-12}, \hat{O}^t)$ and its estimate.

In order to accommodate the migration of companies between size categories, the covariance is now expressed somewhat differently from (2.4), as follows

$$\begin{aligned}\mathrm{cov}(\hat{O}^{t-12}, \hat{O}^t) &= \mathrm{cov}\Big(\sum_{h=1}^{H} N_h^{t-12}\bar{o}_h^{t-12}, \sum_{\ell=1}^{H} N_\ell^t \bar{o}_\ell^t\Big) \\ &= \sum_{h=1}^{H}\sum_{\ell=1}^{H} N_h^{t-12} N_\ell^t \,\mathrm{cov}(\bar{o}_h^{t-12}, \bar{o}_\ell^t). \qquad (3.1)\end{aligned}$$

In order to accommodate the migration of companies between strata, the following statistics are defined:

$N^{t-12,t}_{h\ell}$: size of substratum $U^{t-12,t}_{h\ell}$, or the number of companies that were in stratum h in month $t-12$ and in stratum ℓ $(h, \ell = 0, 1, ..., H+1)$ in month *t*;

$\bar{O}^{t-m}_{h\ell}$: the substratum mean of the revenue in $U^{t-12,t}_{h\ell}$ in month $t-m$ $(m = 0, 12)$;

$n_{h\ell}^{t-m}$: size of sample $s_{h\ell}^{t-m}$, i.e., the actual sample from $U_{h\ell}^{t-12,t}$ in month $t-m$ $(0 \le m \le 12)$;

$\overline{o}_{h\ell}^{t-m}$: the sample mean of the revenue in $s_{h\ell}^{t-m}$ $(m = 0, 12)$;

$n_{h\ell}^{t-12,t}$: number of companies in the overlap $s_{h\ell}^{t-12,t} \equiv s_{h\ell}^{t-12} \cap s_{h\ell}^{t}$;

$\overline{o}_{h\ell,OLP}^{t-m}$: the sample mean of the revenue in the overlap $s_{h\ell}^{t-12,t}$ in month $t-m$ $(m = 0, 12)$.

As in the previous chapter, stratum 0 is reserved for births in the months *t*-12, …, *t*-1 and stratum *H*+1 for the deaths in the same period. Now $\overline{o}_{h}^{t-12}$ and $\overline{o}_{\ell}^{t}$ can be expressed as

$$\overline{o}_{h}^{t-12} = \sum_{g=1}^{H+1} \frac{n_{hg}^{t-12}}{n_{h}^{t-12}} \overline{o}_{hg}^{t-12}$$

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

The covariances in (3.1) can then be rewritten as

$$\operatorname{cov}(\overline{o}_{h}^{t-12}, \overline{o}_{\ell}^{t}) = \operatorname{cov}\left(\sum_{g=1}^{H+1} \frac{n_{hg}^{t-12}}{n_{h}^{t-12}} \overline{o}_{hg}^{t-12}, \sum_{k=0}^{H} \frac{n_{k\ell}^{t}}{n_{\ell}^{t}} \overline{o}_{k\ell}^{t} \right) \tag{3.2}$$

$$= \frac{1}{n_{h}^{t-12} n_{\ell}^{t}} \operatorname{cov}(n_{h\ell}^{t-12} \overline{o}_{h\ell}^{t-12}, n_{h\ell}^{t} \overline{o}_{h\ell}^{t}) \qquad (1 \le h, \ell \le H). \tag{3.3}$$

The second line used that

$$\operatorname{cov}(n_{hg}^{t-12} \overline{o}_{hg}^{t-12}, n_{k\ell}^{t} \overline{o}_{k\ell}^{t}) = 0 \tag{3.4}$$

for $k \neq h$. Furthermore, $\operatorname{cov}(n_{hg}^{t-12} \overline{o}_{hg}^{t-12}, n_{h\ell}^{t} \overline{o}_{h\ell}^{t}) = 0$ for $g \neq \ell$ (g=1,…,H+1), because

$$\begin{aligned} \operatorname{cov}(n_{hg}^{t-12} \overline{o}_{hg}^{t-12}, n_{h\ell}^{t} \overline{o}_{h\ell}^{t}) &= E \operatorname{cov}(n_{hg}^{t-12} \overline{o}_{hg}^{t-12}, n_{h\ell}^{t} \overline{o}_{h\ell}^{t} \,|\, n_{hg}^{t-12}, n_{h\ell}^{t}) \\ &\quad + \operatorname{cov}\{E(n_{hg}^{t-12} \overline{o}_{hg}^{t-12} \,|\, n_{hg}^{t-12}, n_{h\ell}^{t}), E(n_{h\ell}^{t} \overline{o}_{h\ell}^{t} \,|\, n_{hg}^{t-12}, n_{h\ell}^{t})\} \\ &= 0 + \overline{O}_{hg}^{t-12} \, \overline{O}_{h\ell}^{t} \operatorname{cov}(n_{hg}^{t-12}, n_{h\ell}^{t}) = 0. \end{aligned} \tag{3.5}$$

The last line used that

$$\operatorname{cov}(n_{hg}^{t-12}, n_{h\ell}^{t}) = 0 \qquad (g = 1, ..., H+1) \tag{3.6}$$

See the appendix (Section 3.5) for a justification of (3.6) and the associated assumptions. The appendix also derives an alternative estimating method that is applicable with a non-negligible covariance. Like (3.5) the covariance in (3.3) can be rewritten as

$$\begin{aligned}\operatorname{cov}(n_{h\ell}^{t-12}\overline{o}_{h\ell}^{t-12}, n_{h\ell}^{t}\overline{o}_{h\ell}^{t}) &= E\{\operatorname{cov}(n_{h\ell}^{t-12}\overline{o}_{h\ell}^{t-12}, n_{h\ell}^{t}\overline{o}_{h\ell}^{t} \mid \nu_{h\ell})\} \\ &\quad + \operatorname{cov}\{E(n_{h\ell}^{t-12}\overline{o}_{h\ell}^{t-12}|\nu_{h\ell}), E(n_{h\ell}^{t}\overline{o}_{h\ell}^{t} \mid \nu_{h\ell})\}.\end{aligned} \tag{3.7}$$

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

$$\begin{aligned}E\{\operatorname{cov}(n_{h\ell}^{t-12}\overline{o}_{h\ell}^{t-12}, n_{h\ell}^{t}\overline{o}_{h\ell}^{t} \mid \nu_{h\ell})\} &= E\{n_{h\ell}^{t-12} n_{h\ell}^{t} \operatorname{cov}(\overline{o}_{h\ell}^{t-12}, \overline{o}_{h\ell}^{t} \mid \nu_{h\ell})\} \\ &= E\{n_{h\ell}^{t-12} n_{h\ell}^{t} (\frac{n_{h\ell}^{t-12,t} / n_{h\ell}^{t-12}}{n_{h\ell}^{t}} - \frac{1}{N_{h\ell}^{t-12,t}}) S_{h\ell}^{t-12,t}\}. \end{aligned} \tag{3.8}$$

The last line again assumes (2.12). Furthermore, $S_{h\ell}^{t-12,t}$ is defined as

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

The second component on the right-hand side of (3.7) because of (3.6) equals

$$\overline{O}_{h\ell}^{t-12}\overline{O}_{h\ell}^{t} \operatorname{cov}(n_{h\ell}^{t-12}, n_{h\ell}^{t}) = 0.$$

From (3.3), (3.7) and (3.8) it follows that

$$\operatorname{cov}(\overline{o}_{h}^{t-12}, \overline{o}_{\ell}^{t}) = E\{\frac{n_{h\ell}^{t-12} n_{h\ell}^{t}}{n_{h}^{t-12} n_{\ell}^{t}} (\frac{n_{h\ell}^{t-12,t}}{n_{h\ell}^{t-12} n_{h\ell}^{t}} - \frac{1}{N_{h\ell}^{t-12,t}}) S_{h\ell}^{t-12,t}\}. \tag{3.9}$$

3.3.4 Estimate of the covariance term

Expression (3.9) can be estimated from the overlap $s_{h\ell}^{t-12,t}$ with

$$\begin{aligned}\hat{\operatorname{cov}}(\overline{o}_{h}^{t-12}, \overline{o}_{\ell}^{t}) &= \frac{n_{h\ell}^{t-12} n_{h\ell}^{t}}{n_{h}^{t-12} n_{\ell}^{t}} (\frac{n_{h\ell}^{t-12,t}}{n_{h\ell}^{t-12} n_{h\ell}^{t}} - \frac{1}{N_{h\ell}^{t-12,t}}) \hat{S}_{h\ell,OLP}^{t-12,t} \\ \hat{S}_{h\ell,OLP}^{t-12,t} &= \frac{1}{n_{h\ell}^{t-12,t} - 1} \sum_{i=1}^{n_{h\ell}^{t-12,t}} (o_{h\ell i}^{t-12} - \overline{o}_{h\ell,OLP}^{t-12})(o_{h\ell i}^{t} - \overline{o}_{h\ell,OLP}^{t}).\end{aligned} \tag{3.10}$$

Note that (3.10) is an unbiased estimator of (3.9) because

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

A disadvantage of estimator $\hat{S}_{h\ell,OLP}^{t-12,t}$ in (3.10) is nonetheless that it can give rise to a negative variance estimator

$$\begin{aligned}\hat{\operatorname{var}}(\hat{O}^{t} - G^{t,t-12}\hat{O}^{t-12}) &= \hat{\operatorname{var}}(\hat{O}^{t}) + (\hat{G}^{t,t-12})^{2} \hat{\operatorname{var}}(\hat{O}^{t-12}) \\ &\quad - 2\hat{G}^{t,t-12} \hat{\operatorname{cov}}(\hat{O}^{t}, \hat{O}^{t-12}).\end{aligned} \tag{3.11}$$

Knottnerus and Van Delden (2006) propose a refinement to (3.10). Define the standard deviations

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

Then the new, modified estimator for $S_{h\ell}^{t-12,t}$ is

$$\hat{S}_{h\ell}^{t-12,t} = \hat{\rho}_{h\ell,OLP}^{t-12,t}\hat{S}_{h\ell}^{t-12}\hat{S}_{h\ell}^{t}, \tag{3.12}$$

where $\rho_{h\ell}^{t-12,t}$ represents the correlation between the variables o^t and o^{t-12} in $U_{h\ell}^{t-12,t}$ and $\hat{\rho}_{h\ell,OLP}^{t-12,t}$ its estimate based on $s_{h\ell}^{t-12,t}$. In accordance with (3.10) and (3.12), the covariance in (3.1) can be estimated with

$$\hat{\text{cov}}(\hat{O}^{t-12},\hat{O}^{t}) = \sum_{h=1}^{H}\sum_{\ell=1}^{H}\frac{N_h^{t-12}N_\ell^{t}}{n_h^{t-12}n_\ell^{t}}\, n_{h\ell}^{t-12,t}\left(1-\frac{n_{h\ell}^{t-12}n_{h\ell}^{t}}{n_{h\ell}^{t-12,t}N_{h\ell}^{t-12,t}}\right)\hat{S}_{h\ell}^{t-12,t}. \tag{3.13}$$

For the estimate $\hat{\rho}_{h\ell,OLP}^{t-12,t}$ defined in this way, $\left|\hat{\rho}_{h\ell,OLP}^{t-12,t}\right| \le 1$ holds with probability 1, while use of (3.10) may lead implicitly to an estimated correlation greater than 1, and therefore to a possible negative result of (3.11). It is mentioned for the sake of completeness that in special circumstances use of (3.12) can also lead to a negative result of (3.11).

When using (3.12), a special problem may arise when $n_{h\ell}^{t}=1$ or $n_{h\ell}^{t-12}=1$. The first option for calculating the necessary variances is to borrow the sample variance from a related substratum, or from the same substratum in a previous month. A variance can also be imputed when there is evidence in the data of a relationship of the form $S_{h\ell}^2 \approx \sigma^2\overline{O}_{h\ell}^{\beta}$; see Särndal et al. (1992, page 461). Furthermore the corresponding covariance term can be ignored if it is expected to make only a small contribution to the total variance. This will often be the case for strata *h* and ℓ with relatively small sampling fractions and correspondingly relatively small variances compared with larger companies in the strata with larger sampling fractions. Similar remarks apply to the imputed values for $\rho_{h\ell}^{t-12,t}$ when $n_{h\ell}^{t-12,t} \le 2$ and $n_{h\ell}^{t-m} \ge 2$ $(m=0,12)$. This would appear to be an effective approach in view of the often high value of $\rho_{h\ell}^{t-12,t}$. It is observed moreover that when $n_{h\ell}^{t-m}=0$ $(m=0,12)$, the corresponding covariance term in (3.13) can be disregarded without detriment to the unbiasedness of (3.13), provided the other $S_{h\ell}^{t-12,t}$ can be estimated without bias. Under this assumption a term of this kind with $n_{h\ell}^{t-m}=0$ $(m=0,12)$ can be disregarded because the expected value

$$n_{h\ell}^{t-12,t}\left(1-\frac{n_{h\ell}^{t-12}n_{h\ell}^{t}}{n_{h\ell}^{t-12,t}N_{h\ell}^{t-12,t}}\right)\hat{S}_{h\ell}^{t-12,t} \tag{3.14}$$

from (3.10) is

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

and the expected value of the right-hand side is the parameter to be estimated. Furthermore when $n_{h\ell}^{t-m}=0$ $(m=0,12)$, then also $n_{h\ell}^{t-12,t}=0$ and therefore the result of (3.14) is zero, so that the estimator $\hat{S}_{h\ell}^{t-12,t}$ of $S_{h\ell}^{t-12,t}$ is irrelevant. Therefore

disregarding these kinds of terms with $n_{h\ell}^{t-m} = 0$ $(m = 0, 12)$ adds no bias to the estimators (3.13) and (3.14), provided the other $S_{h\ell}^{t-12,t}$ are estimated without bias.

3.4 Quality indicators

The quality indicators for panels in which companies migrate between strata or size categories are:

- the margins of uncertainty of the corresponding estimators;
- the size of nonresponse;
- the size of the overlap of the panels in months *t* and *t*-12.

Nonresponse can severely affect the quality of the results if (i) the nonresponse is large and (ii) the nonresponse is selective. Adjustment of some of the bias from selective nonresponse can sometimes be achieved by using auxiliary variables that correspond with both the probability of response and the target variable.

Another important aspect of panel quality is the size of the overlap. If nonresponse or substantial migration between strata significantly reduce the overlap between the panels in months *t* and *t*-12, the margin of uncertainty of a revenue growth estimate may rise considerably.

3.5 Appendix. Justification of (3.6)

The simple case with strata without births and deaths is considered first. There are now no monthly updates other than the annual update in January. Therefore, $n_{h\ell}^{t} = n_{h\ell,RST}^{dec,\,jan}$ is fixed, from which (3.6) follows. This is the situation that applies, for example, to supermarkets, which have formed a fairly stable population over the years.

If births and deaths do occur in a population, then express $n_{h\ell}^{t}$ as

$$n_{h\ell}^{t} = n_{\ell}^{t} - n_{0\ell}^{t-12,t} - \sum_{k \neq h}^{H} n_{k\ell}^{t}, \tag{3.15}$$

where $n_{0\ell}^{t-12,t}$, or more concisely $n_{0\ell}^{t}$, represents the number of births that occurred in months *t*-12, …, *t*-1 that are within sample s_{ℓ}^{t} for month *t*. Because the samples under the new births after month *t*-12 are independent of the n_{hg}^{t-12}, the random variables $n_{0\ell}^{t-12,t}$ and n_{hg}^{t-12} are uncorrelated. Because also $\mathrm{cov}(n_{hg}^{t-12}, n_{k\ell}^{t}) = 0$ for $k \neq h$, it follows from (3.15) that $\mathrm{cov}(n_{hg}^{t-12}, n_{h\ell}^{t}) = 0$ $(h = 1, \cdots, H)$.

The actual assumption so far has been that the $n_{k\ell}^{t}$ $(k \neq h)$ can be said to have a nearly 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}. The same comment applies to $n_{0\ell}^{t-12,t}$. However, it can be demonstrated that these assumptions can give rise in practice to a small second-

order error in the variance formulas. The following four simplifying assumptions are made in order to shed light on this error: (*i*) births and deaths do not migrate between strata, (*ii*) there are no deaths among the births, (*iii*) after their first month in the population births are no longer involved in the updates for the rest of the study period, including the refreshing in January, and (*iv*) no deaths are selected into or removed from the sample in the monthly updates. With these assumptions, a third-order error is still disregarded, and the covariance in (3.2) is scrutinized more closely. By analogy with (3.7) it can be divided into two components for $\ell = h$. The second component, say $C_{hh,\text{sec}}$, can be expressed as

$$\begin{aligned} C_{hh,\text{sec}} &\equiv \frac{1}{n_h^{t-12} n_h^t} \text{cov}\{E(\sum_{g=1}^{H+1} n_{hg}^{t-12} \bar{o}_{hg}^{t-12} | \nu_h), E(\sum_{k=0}^{H} n_{kh}^t \bar{o}_{kh}^t | \nu_h)\} \\ &= \frac{1}{n_h^{t-12} n_h^t} \sum_{g=1}^{H+1} \sum_{k=1}^{H} \bar{O}_{hg}^{t-12} \bar{O}_{kh}^t \text{ cov}(n_{hg}^{t-12}, n_{kh}^t) \qquad (3.16) \\ \nu_h &= (n_{h,1}^{t-12}, \cdots, n_{h,H+1}^{t-12}, n_{1h}^t, \cdots, n_{Hh}^t). \end{aligned}$$

Note that under the above assumptions $C_{h\ell,\text{sec}} = 0$ for $\ell \neq h$. Before estimating the covariances in (3.16), the formula is investigated for the conditional expectation of *y* given $x = x_0$ when *y* and *x* have a bivariate normal distribution. In other words, in standard notation,

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

For a given change Δx_0 in *x*, the conditional expectation of the change in *y* is $E(\Delta y|\Delta x_0) = \sigma_{yx}\Delta x_0 / \sigma_x^2$, or

$$\sigma_{yx} = \frac{E(\Delta y|\Delta x_0)}{\Delta x_0}\sigma_x^2 \qquad (3.17)$$

Therefore it is sufficient in estimating e.g. $\text{cov}(n_{h,H+1}^{t-12}, n_{kh}^t)$ in (3.16) under normality to investigate the expected effect on $y = n_{kh}^t$ caused by a change in future deaths $x = n_{h,H+1}^{t-12}$ in s_h^{t-12}. Let $\Delta n_{h,H+1}^{t-12}$ be an additional (positive) change in the number of deaths in s_h^{t-12}. Define $p_{h,H+1}^{jan,t}$ as $p_{h,H+1}^{jan,t} = N_{h,H+1}^{jan,t} / N_{h,H+1}^{t-12}$, where $N_{h,H+1}^{jan,t}$ represents the number of deaths in stratum *h* between January and month *t*. Likewise $p_{hg}^{t-12} = N_{hg}^{t-12,t} / N_h^{t-12}$ $(g = 1, \cdots, H+1)$. Based on assumption (*iv*) it is now possible to estimate the expected number of additional deaths in the January sample before refreshing at $p_{h,H+1}^{jan,t}\Delta n_{h,H+1}^{t-12}$. The expected number of additional deaths in the sample after refreshing can then be estimated with

$$\begin{aligned} &\gamma_{red}^{jan}\, p_{h,H+1}^{jan,t}\Delta n_{h,H+1}^{t-12} \\ &\gamma_{red}^{jan} = (0.9 - f_h)/(1 - f_h), \qquad (3.18) \end{aligned}$$

where γ_{red}^{jan} represents the reduction factor because of the refreshing in January. See the end of this appendix for the derivation of (3.18).

The corresponding monthly updates in stratum *h* between January and month *t* because of these additional deaths in the sample suggest the following estimate of the expected increase in incoming n_{kh}^{t} from stratum *k* $(k \neq h)$ in the sample of month *t*

$$\begin{aligned} &E(\Delta n_{kh}^{t} \,|\, \Delta n_{h,H+1}^{t-12}) = \gamma_{red}^{jan}\, p_{h,H+1}^{jan,t} \Delta n_{h,H+1}^{t-12}\, p_{kh}^{t} \\ &p_{kh}^{t} = N_{kh}^{t-12,t} / (N_{h}^{t} - N_{0h}^{t}). \end{aligned} \tag{3.19}$$

As was seen in Subsection 2.3.6, there is an update in month *s* only when $d_h^{s-1} \neq f_h D_h^{s-1}$, where $D_h^s\,(d_h^s)$ represents the number of deaths in $U_h^s\,(s_h^s)$ and $n_{kh}^{t} = f_h N_{kh}^{t-12,t}$ is fixed when $N_{h,H+1}^{jan,t} = 0$ $(k \neq h)$. Note also that births are excluded in the definition given in (3.19) because of assumption (*iii*).

Then define for $m = 0, 12$

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

Using (3.17) and (3.19) leads for $k \neq h$ to the following approximation for the covariance

$$\begin{aligned} \operatorname{acov}(n_{h,H+1}^{t-12}, n_{kh}^{t}) &= \frac{E(\Delta n_{kh}^{t} \,|\, \Delta n_{h,H+1}^{t-12})}{\Delta n_{h,H+1}^{t-12}} \operatorname{var}(n_{h,H+1}^{t-12}) \\ &\approx \gamma_{red}^{jan}\, p_{h,H+1}^{jan,t} p_{kh}^{t} n_h^{t-12} p_{h,H+1}^{t-12} (1 - p_{h,H+1}^{t-12})(1 - f_h) \\ &= n_h^{t-12} p_{kh}^{t} A_h / p_{in,h}^{t} \\ A_h &= \gamma_{red}^{jan}\, p_{in,h}^{t} p_{h,H+1}^{jan,t} p_{h,H+1}^{t-12} (1 - p_{h,H+1}^{t-12})(1 - f_h), \end{aligned} \tag{3.20}$$

where for the sake of simplicity the term $N_h^{t-12} / (N_h^{t-12} - 1)$ is omitted from the second line. Because n_h^{t-12} is fixed,

$$\operatorname{cov}(n_{h,H+1}^{t-12}, n_{kh}^{t}) = -\operatorname{cov}(n_{h1}^{t-12} + \cdots + n_{hH}^{t-12}, n_{kh}^{t}).$$

By analogy with the multihypergeometric distribution, the following relationship can be used for $1 \le g \le H$ and $k \neq h$ for an approximation of $\operatorname{cov}(n_{hg}^{t-12}, n_{kh}^{t})$

$$\operatorname{acov}(n_{hg}^{t-12}, n_{kh}^{t}) = -\frac{p_{hg}^{t-12}}{p_{h,\le H}^{t-12}} \operatorname{acov}(n_{h,H+1}^{t-12}, n_{kh}^{t}) = -n_h^{t-12} \frac{p_{hg}^{t-12}}{p_{h,\le H}^{t-12}} \frac{p_{kh}^{t}}{p_{in,h}^{t}} A_h, \tag{3.21}$$

where use is also made of (3.20). As an extension to the above,

$$\mathrm{cov}(n_{h,H+1}^{t-12}, n_{kh}^{t}) = -\,\mathrm{cov}(n_{h,\pounds H}^{t-12}, n_{kh}^{t}) = -\sum_{g\pounds H}\sum_{i\hat{\imath}\,U_{hg}^{t-12,t}} \mathrm{cov}(\delta_{hgi}^{t-12}, n_{kh}^{t})$$

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

Therefore, by symmetry, we have

$$\mathrm{cov}(\delta_{hgi}^{t-12}, n_{kh}^{t}) = -\,\mathrm{cov}(n_{h,H+1}^{t-12}, n_{kh}^{t}) / N_{h,\pounds H}^{t-12,t},$$

from which (3.21) follows $(1 \pounds g \pounds H)$. Likewise for $k = h$ based on (3.20) and (3.21)

$$\begin{aligned} &\mathrm{acov}(n_{h,H+1}^{t-12}, n_{hh}^{t}) = -\,n_h^{t-12} A_h \\ &\mathrm{acov}(n_{hg}^{t-12}, n_{hh}^{t}) = n_h^{t-12} p_{hg}^{t-12} A_h / p_{h,\pounds H}^{t} \qquad (1 \pounds g \pounds H). \end{aligned} \tag{3.22}$$

Substituting (3.20)-(3.22) in (3.16) then gives

$$C_{hh,\mathrm{sec}} = A_h(\bar{O}_{h,H+1}^{t-12} - \bar{O}_{h,\pounds H}^{t-12})(\bar{O}_{in,h}^{t} - \bar{O}_{hh}^{t}) / n_h^{t} \tag{3.23}$$

Under the assumption that the two bracketed terms in (3.23) are absolutely smaller than S_h^t, it follows from (3.23) that

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

It follows from this that when $p_{in,h}^{t}, p_{h,H+1}^{t-12} \pounds 0.1$, it must be concluded under the above assumptions that the contribution of the second covariance component is less than 1% of $\mathrm{var}(\bar{o}_h^t)$. In other words, (3.6) can be used without undue adverse effects on the outcomes. When $C_{hh,\mathrm{sec}}$ is not negligible in practice, it can be estimated from the data with

$$\hat{C}_{hh,\mathrm{sec}} = A_h(\bar{o}_{h,H+1}^{t-12} - \bar{o}_{h,\pounds H}^{t-12})(\bar{o}_{in,h}^{t} - \bar{o}_{hh}^{t}) / n_h^{t}. \tag{3.24}$$

This appendix concludes with the derivation of (3.18). In January the expected number of additional deaths remaining during the refreshing of the sample is $0.9\, p_{h,H+1}^{jan,t} \Delta n_{h,H+1}^{t-12}$. An estimate of the number of deaths outside the sample immediately prior to the refreshing is $N_{h,H+1}^{jan,t} - n_{h,H+1}^{jan} - p_{h,H+1}^{jan,t} \Delta n_{h,H+1}^{t-12}$. The number of new deaths in the sample because of the refreshing operations in all substrata $U_{hg}^{t-12,t}$ $(g = 1, \cdots, H)$ in January is therefore estimated at

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

Because $n_{0h}^{jan} = f_h N_{0h}^{t-12,jan}$ the above assumptions lead to

$$\gamma_{red}^{jan} = 0.9 - \frac{0.1(n_h^{jan} - n_{0h}^{jan})}{N_h^{jan} - N_{0h}^{t-12,\,jan} - (n_h^{jan} - n_{0h}^{jan})} = \frac{0.9 - f_h}{1 - f_h}.$$

4. Panels for estimating indicators and indexes

4.1 Short description

An index can actually be viewed as a ratio of two population totals, which can be estimated using the estimator of a ratio[1] based on SRSWOR of the population or the corresponding strata. Another way of estimating an index is the Horvitz-Thompson estimator based on probability-proportional-to-size (PPS) sampling. The related inclusion probability of a company is proportional to its corresponding revenue. We elaborate the corresponding formulas for indexes in this chapter. The effects of company mergers and demergers are also discussed. Formulas from this chapter are also applicable to price movements and economic indicators that are based on panels of companies. Unless stated to the contrary, this chapter assumes consistently that the target parameter is a weighted average of price movements for a target group of companies, where the weights are set equal to the revenue shares of the corresponding companies. Readers are referred to Chapter 3 for index changes, updates and refreshing of panels.

4.2 Applicability

Statistics Netherlands also uses panels to calculate special weighted population means, such as the PPI, DPI, the Business Survey of the Netherlands (COEN) and the Business Test (Conjunctuurtest). The first two are price indexes and the last two are economic indicators.

4.3 Detailed description

4.3.1 The estimator of a ratio for a general index

A composite price index I of price movements among N different companies can for our purposes also be expressed as

$$I = \frac{\sum_{h=1}^{H}\sum_{i=1}^{N_h} O_{hi} I_{hi}}{\sum_{h=1}^{H}\sum_{i=1}^{N_h} O_{hi}} = \sum_{h=1}^{H}\sum_{i=1}^{N_h} O_{hi} I_{hi} \Big/ O, \qquad \left(O = \sum_{h=1}^{H}\sum_{i=1}^{N_h} O_{hi}\right) \tag{4.1}$$

where

N_h : number of companies in stratum h $(h = 1,\ldots,H;\ \sum_{h=1}^{H} N_h = N)$

I_{hi} : price change for company i in stratum h

O_{hi} : revenue of company i in stratum h in a given period.

[1] Note that the term *estimator of a ratio* is used in the context of indexes in this chapter in its literal sense, and is not to be confused with the estimator more commonly mentioned in the literature of a population total based on an estimated ratio, which is referred to as a 'ratio estimator'.

It is convenient for estimating purposes to rewrite (4.1) as

$$\begin{aligned} I &= \left[\sum_{h=1}^{H} O_h\right]^{-1} \sum_{h=1}^{H} O_h I_h = \sum_{h=1}^{H} \frac{O_h}{O} I_h \qquad (O = \sum_{h=1}^{H} O_h) \\ I_h &= \sum_{i=1}^{N_h} \frac{O_{hi}}{O_h} I_{hi} \qquad (O_h = \sum_{i=1}^{N_h} O_{hi}). \end{aligned} \tag{4.2}$$

Because the revenue figures may generally be assumed to be known, an obvious estimator of (4.1) or (4.2) is

$$\hat{I} = \sum_{h=1}^{H} \frac{O_h}{O} \hat{I}_h \, . \tag{4.3}$$

$\hat{I}_h$ represents the standard estimator of a ratio of the mean price movement in stratum *h*

$$\hat{I}_h = \frac{\sum_{i=1}^{n_h} o_{hi} I_{hi}}{\sum_{i=1}^{n_h} o_{hi}}, \tag{4.4}$$

where o_{hi} and I_{hi} represent the revenue and price movement, respectively, of company *i* in the sample obtained through SRSWOR of size n_h from stratum *h* (i=1,…,n_h). Note that $\hat{I}_h$ can be viewed as a standard estimator of a ratio for I_h in (4.2).

The variance of $\hat{I}$ in (4.3) is

$$\mathrm{var}(\hat{I}) = \sum_{h=1}^{H} W_h^2 \, \mathrm{var}(\hat{I}_h), \qquad (W_h = O_h/O) \tag{4.5}$$

where

$$\begin{aligned} \mathrm{var}(\hat{I}_h) &= \frac{1 - f_h}{n_h \overline{O}_h^2} S_{he}^2 \\ f_h &= n_h / N_h \\ \overline{O}_h &= O_h / N_h \\ S_{he}^2 &= \frac{1}{(N_h - 1)} \sum_{i=1}^{N_h} O_{hi}^2 (I_{hi} - I_h)^2 \end{aligned} \tag{4.6}$$

see Banning et al. (2010, p. 76). Combination of (4.5) and (4.6) gives

$$\mathrm{var}(\hat{I}) = \sum_{h=1}^{H} \frac{W_h^2}{\overline{O}_h^2} (1 - f_h) \frac{S_{he}^2}{n_h} . \tag{4.7}$$

The variance in (4.7) can be estimated with

$$\hat{\text{var}}(\hat{I}) = \sum_{h=1}^{H} \frac{W_h^2}{\overline{O}_h^2}(1 - f_h)\frac{s_{he}^2}{n_h}$$

$$s_{he}^2 = \frac{1}{(n_h - 1)}\sum_{i=1}^{n_h} o_{hi}^2 (I_{hi} - \hat{I}_h)^2;$$

Analogous to the Neyman allocation for the standard stratification estimator, it is possible to prove that the variance in (4.5) is a minimum for the following allocation of the n_h

$$n_h = \frac{W_h S_{he} / \overline{O}_h}{\sum_{h=1}^{H} W_h S_{he} / \overline{O}_h} n_h = \frac{N_h S_{he}}{\sum_{h=1}^{H} N_h S_{he}} n_h . \tag{4.8}$$

If it can be assumed that the I_{hi} can be described well with the following model $I_{hi} = I_h + e_{hi}$ with $E(e_{hi}) = 0$, $E(e_{hi}^2) = \sigma^2$ and $E(e_{hi} e_{hj}) = 0$ $(i \neq j)$, then $S_{he}^2 / \overline{O}_h^2$ can be simplified as

$$\frac{S_{he}^2}{\overline{O}_h^2} \approx \frac{\sigma^2 \sum_{i=1}^{N_h} O_{hi}^2 / N_h}{\overline{O}_h^2} \approx \frac{\sigma^2 (S_{ho}^2 + \overline{O}_h^2)}{\overline{O}_h^2} = \sigma^2 (1 + CV_{ho}^2)$$

$$S_{ho}^2 = \frac{1}{(N_h - 1)}\sum_{i=1}^{N_h} (O_{hi} - \overline{O}_h)^2$$

$$CV_{ho} = S_{ho} / \overline{O}_h .$$

where it is assumed that $N_h \approx N_h - 1$. The allocation in (4.8) therefore becomes

$$n_h = \frac{W_h \sqrt{1 + CV_{ho}^2}}{\sum_{h=1}^{H} W_h \sqrt{1 + CV_{ho}^2}} n_h = \frac{O_h \sqrt{1 + CV_{ho}^2}}{\sum_{h=1}^{H} O_h \sqrt{1 + CV_{ho}^2}} n_h .$$

It is observed finally that I_h in (4.2) can be expressed as a normal stratum total $I_h = \sum_{i=1}^{N_h} Y_{hi}$ with $Y_{hi} = O_{hi} I_{hi} / O_h$. Therefore I_h can also be estimated with the direct estimator

$$\hat{I}_{h,DIR} = \frac{N_h}{n_h}\sum_{i=1}^{n_h} y_{hi} = \frac{N_h}{n_h}\sum_{i=1}^{n_h} \frac{o_{hi} I_{hi}}{O_h} = \frac{1}{n_h}\sum_{i=1}^{n_h} \frac{o_{hi} I_{hi}}{\overline{O}_h}.$$

Under the above model assumptions, a regression of Y_{hi} on O_{hi} and a constant $(i = 1,..., N_h)$ nonetheless yields a regression coefficient b_h approximately equal to $b_h = I_h$. Because of $b_h > I_h / 2$, the estimator of a ratio is preferable to the direct estimator; see Banning et al. (2010, p. 77). Note that b_h can be expressed as $b_h = \rho_{hyo} S_{hy} / S_{ho}$ where ρ_{hyo} represents the correlation coefficient between the Y_{hi} and the O_{hi}.

4.3.2 The PPS estimator for a general index

As shown in Chapter 5 of Banning et al. (2010) a (composite) price index can also be estimated based on systematic PPS sampling where the inclusion probabilities

$\pi_i = nW_i$ $(W_i = O_i / O)$ are proportional to the corresponding revenue figures. Furthermore, the companies are deemed to be randomly ordered in the corresponding list. The PPS estimator of the parameter $I = \sum_{i=1}^{N} W_i I_i$ now takes the following form in accordance with equation (5.5) of Banning et al. (2010)

$$\hat{I}_{PPS} = \sum_{i=1}^{N} a_i \frac{W_i I_i}{\pi_i} = \frac{1}{n} \sum_{i=1}^{n} I_i = \bar{I}_s$$
$$a_i = \begin{cases} 1 & \text{if company } i \text{ is included in the sample} \\ 0 & \text{otherwise,} \end{cases} \tag{4.9}$$

where $\bar{I}_s$ represents the mean of the observed price movements in the sample. In order to keep the notation simple, this subsection assumes that the population consists of 1 stratum. If the population comprises more than one stratum, then the same approach can be followed for estimating the indexes corresponding with the separate strata I_h.

Assuming that the sequence of population elements has no effect on the PPS sampling and that the I_i can be described well with the model $I_i = I + \varepsilon_i$ with $E(\varepsilon_i) = 0,\ E(\varepsilon_i^2) = \gamma^2 O_i^{\beta}$ and $E(\varepsilon_i \varepsilon_j) = 0$, $(i \neq j),$ can be used as an approximation of the variance of $\hat{I}_{PPS}$ in (4.9)

$$\begin{aligned} \text{var}\left(\hat{I}_{PPS}\right) &= \frac{1}{n^2} \sum_{i=1}^{N} \pi_i (1 - \pi_i)(I_i - I)^2 \\ &= \frac{1}{n} \sum_{i=1}^{N} W_i (1 - nW_i)(I_i - I)^2. \end{aligned} \tag{4.10}$$

It can be demonstrated under the above model assumptions with $\beta > -1$ that for $N \to \infty$ $\hat{I}_{PPS}$ has a smaller variance than the estimator of a ratio $\hat{I}_{EAZT}$ based on SRSWOR, or

$$\hat{I}_{EAZT} = \frac{\sum_{i=1}^{n} o_i I_i}{\sum_{i=1}^{n} o_i};$$

see Knottnerus (2011). An asymptotically unbiased estimator of the variance in (4.10) under the above assumptions is

$$\hat{\text{var}}\left(\hat{I}_{PPS}\right) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(1 - \frac{no_i}{O}\right)(I_i - \bar{I}_s)^2. \tag{4.11}$$

4.3.3 Replacement or replenishment with systematic PPS sampling

So far a stable sample that does not change in time has been assumed. Unfortunately, it is common for units in the sample to disappear because of nonresponse or death and other reasons. One method of replenishing a PPS sample is the following. Assume that the (presumed) attrition in a period is 10% of the sample, in which case gross sampling s_b could be performed in advance, producing

a sample of size $n_b = 1{,}1n$. Then performing SRSWOR on s_b to produce a sample of size *n* will yield on balance a PPS sample of the desired size *n*. Any companies that drop out of the panel later can be replaced at random from the remaining s_b. This approach can also be applied in practice with SRSWOR.

Another approach is that of 'circular systematic PPS sampling' in which the companies are sequenced cyclically rather than linearly, so that company *N* is between companies *N*-1 and 1 (disregarding the random sequence). If a company drops out, it is simple to continue the systematic sampling procedure with the same step length *L*. If necessary, a new random starting number can be chosen when the circle has been traversed. This new random starting number would be needed if the population had remained stable with no births or deaths. After the first full round of the circle, the company associated with the first randomly selected number would automatically reappear. A disadvantage of this cyclic approach is that the precise net inclusion probabilities are less than completely clear when the inclusion probabilities are unequal and the circle is traversed for the second time.

It is advisable to treat births as a separate group from which a separate random sample must be drawn. In the case of a systematic sample, the step length *L* must be the same as in the associated population of pre-existing units. The same method can be used to accommodate later attrition from the sample of this group.

4.3.4 Company mergers

Consider a population {1, 2, …,*N*} of *N* companies. Define π_i as the first order inclusion probability of company *i*. The first situation considered below is that in which Y_i is the value of an arbitrary variable *y* for company *i* (*i*=1,…,*N*). The modified Horvitz-Thompson (HT) estimator for the population total *Y* in the event of a merger of companies 1 and 2 is then as follows

$$\hat{Y}_{HT}^{\text{mod}} = (a_1 + a_2)\frac{Y_1 + Y_2}{\pi_1 + \pi_2} + \sum_{i=3}^{N} a_i \frac{Y_i}{\pi_i} \ . \tag{4.12}$$

$E(a_1 + a_2) = \pi_1 + \pi_2$ means that (4.12) is an unbiased estimator for *Y*, like the standard HT estimator in the situation without mergers. We can also express (4.12) as

$$\begin{aligned} \hat{Y}_{HT}^{\text{mod}} &= \sum_{i=1}^{N} a_i \frac{Y_i^*}{\pi_i} \\ Y_i^* &= \begin{cases} \pi_i (Y_1 + Y_2)/(\pi_1 + \pi_2) & \text{als } i \le 2 \\ Y_i & \text{als } i \ge 3. \end{cases} \end{aligned} \tag{4.13}$$

In other words, $Y_1 + Y_2$ is distributed over companies 1 and 2 with weight ratio of π_1 / π_2. Formula (4.13) is again the standard form of an HT estimator and can therefore serve as the basis for further variance calculations. For example, for simple random sampling without replacement (SRSWOR) with $\pi_i = n/N$ (*i*=1,..,*N*), then

$$\text{var}(\hat{Y}_{HT}^{\text{mod}}) = \frac{N^2}{n}(1 - \frac{n}{N})\frac{1}{N-1}\sum_{i=1}^{N}(Y_i^* - \bar{Y})^2 \ .$$

Note in this case that $Y_1^* = Y_2^* = (Y_1 + Y_2)/2$ and that the population means are equal: $\bar{Y}^* = \bar{Y}$. Also with systematic PPS sampling, (4.13) can be used to approximate and estimate the variance, as discussed in Subsection 4.3.2, provided the elements in the population are more or less randomly sequenced.

The approach described here can be used when it is desired to maximize the retention of companies in the panel. Relevant examples would be the Price indices for producers (PPI) and services (DPI). An alternative approach is to view the individual companies in the merger as deaths, and the new merged company as a birth. This alternative approach would not normally give a systematic bias, but may lead to larger margins of uncertainty surrounding the estimated changes.

4.3.4.1 Mergers with PPS sampling

Consider very generally a price index $I = \sum_{i=1}^{N} W_i I_i \quad (\sum_{i=1}^{N} W_i = 1)$ of N companies where W_i represents the weight or relevance of company i. Companies in the full observation are disregarded. Consider now a PPS sample of size n where the inclusion probabilities p_i are proportional to the weights W_i, or $p_i = nW_i$. If there are no mergers, we saw above that the standard HT estimator of I is

$$\hat{I}_{PPS} = \bar{I}_s = \frac{1}{n}\sum_{i=1}^{n} I_i \ .$$

If companies 1 and 2 merged, it will be more convenient from now on to denote the companies in the sample with their sequence numbers $i_1,...,i_n$. For the three different possible situations, the modified HT estimators are as follows. First, when $a_1 = a_2 = 0$, (4.12) equals the standard HT estimator.

$$\hat{I}_{PPS}^{\text{mod}} = \frac{1}{n}\sum_{b=1}^{n} I_{i_b} \ .$$

Second, when exactly one of the two companies in the merger was in the original sample, or $a_1 + a_2 = 1$, (4.12) takes the following form

$$\begin{aligned} \hat{I}_{PPS}^{\text{mod}} &= \frac{W_1 I_1 + W_2 I_2}{n(W_1 + W_2)} + \frac{I_{i_2} + ... + I_{i_n}}{n} \\ &= \frac{I_{12} + I_{i_2} + ... + I_{i_n}}{n} \qquad (I_{12} = \frac{W_1 I_1 + W_2 I_2}{W_1 + W_2}). \end{aligned} \tag{4.14}$$

In practice I_{12} represents the price observation of the new company 12 that was created in the merger of companies 1 and 2. Finally, when companies 1 and 2 were both in the original sample, then by analogy with (4.14), (4.12), becomes

$$\hat{I}_{PPS}^{\text{mod}} = (2I_{12} + I_{k_3} + ... + I_{k_n})/n\,.$$

In summary the modified HT estimator of the PPI with PPS sampling is

$$\hat{I}_{PPS}^{\text{mod}} = \begin{cases} (I_{i_1} + ... + I_{i_n})/n & \text{if } a_1 = a_2 = 0 \\ (I_{12} + I_{i_2} + ... + I_{i_n})/n & \text{if } a_1 + a_2 = 1 \\ (2I_{12} + I_{i_3} + ... + I_{i_n})/n & \text{if } a_1 = a_2 = 1. \end{cases}$$

The final formula shows that assuming an unbiased Laspeyres price index, a break in the trend in the price index may be visible relative to the base year when I_1 and I_2 differ greatly and *n* is not extremely large. In the hypothetical situation that, for example, $a_1 = 1$ and $a_2 = 0,$ the new I_{12} can in principle differ considerably from the old I_1 observed in the sample. Because the index is calculated based on a chain index, no visible break in the trend occurs in practice.

In summary, it can be stated that application of the modified HT estimator to a PPS sample for the PPI with a subsequent merger of companies 1 and 2 means that if company *i* (*i*=1.2) was already in the sample before the merger, its price index I_i after the merger is replaced by the price index I_{12} of the new, merged company. $\hat{I}_{PPS}$ remains, also after the merger, equal to the unweighted mean of the thereby observed price movements relative to the base year.

4.3.4.2 Mergers with SRSWOR

This subsection briefly comments on a simple random sample of size *n* in which all companies have the same inclusion probability $(\pi_i = n/N)$. As in the previous section, it is assumed that companies 1 and 2 (of the population) merge to create company 12. As long as neither company 1 nor company 2 are in the sample $(a_1 = a_2 = 0),$ the usual estimator of a ratio of the index *I* is

$$\hat{I}_{SRSWOR} = \frac{\sum_{i=1}^{n} W_i I_i}{\sum_{i=1}^{n} W_i}.$$

If company 1 is in the original sample but not company 2, or vice versa $(a_1 + a_2 = 1),$ in accordance with (4.12) the estimator of a ratio is

$$\hat{I}_{SRSWOR}^{\text{mod}} = \frac{\frac{W_1 I_1 + W_2 I_2}{2n/N} + \frac{W_{i_2} I_{i_2}}{n/N} + ... + \frac{W_{i_n} I_{i_n}}{n/N}}{\frac{W_{12}}{2n/N} + \frac{W_{i_2}}{n/N} + ... + \frac{W_{i_n}}{n/N}}$$

$$= \frac{\overline{W}_{12} I_{12} + W_{i_2} I_{i_2} + ... + W_{i_n} I_{i_n}}{\overline{W}_{12} + W_{i_2} + ... + W_{i_n}} \qquad (\overline{W}_{12} = \frac{W_{12}}{2} = \frac{W_1 + W_2}{2}).$$

This means that the total weight of the merged company must be halved in the later estimate of the price index after the merger. The following modification is proposed in order to avoid this effect

$$\hat{I}^{\text{mod}}_{SRSWOR} = \frac{W_i I_{12} + W_{i_2} I_{k_2} + \ldots + W_{i_n} I_{k_n}}{W_i + W_{i_2} + \ldots + W_{i_n}} \qquad \text{with} \quad i = \begin{cases} 1 & \text{if } a_1 = 1 \\ 2 & \text{if } a_2 = 1. \end{cases}$$

The advantage, as with PPS sampling, is that the weights of the companies in the sample after the merger do not have to be adjusted. Furthermore, the numerator and denominator continue to be unbiased estimators because $E(a_1 W_1 + a_2 W_2 | a_1 + a_2 = 1) = \overline{W_{12}}$.

When company 1 and company 2 are both in the sample $(a_1 = a_2 = 1)$, (4.12) yields the following estimator of a ratio

$$\begin{aligned} \hat{I}^{\text{mod}}_{SRSWOR} &= \frac{(W_1 I_1 + W_2 I_2 + W_{i_3} I_{i_3} + \ldots + W_{i_n} I_{i_n})/(n/N)}{(W_1 + W_2 + W_{i_3} + \ldots + W_{i_n})/(n/N)} \\ &= \frac{(W_1 + W_2) I_{12} + W_{i_3} I_{i_3} + \ldots + W_{i_n} I_{i_n}}{W_1 + W_2 + W_{i_3} + \ldots + W_{i_n}}. \end{aligned}$$

In other words, the merged company 12, like the other companies, acquires its associated weight $(W_1 + W_2)$ in accordance with the revenue share of the newly merged company.

In summary, this means that in the event of sampling for the PPI with equal inclusion probabilities, use of the modified HT estimator with a merger of companies 1 and 2 implies that when one of the two companies, say company i (i=1.2), was already in the sample before the merger, the price movement I_i after the merger must be replaced by the price movement I_{12} of the new, merged company, while its weight w_i remains the same. If both companies were already in the sample before the merger, then I_{12} in the sample acquires a weight of $w_1 + w_2$.

4.3.5 Company demergers

This subsection briefly discusses the effect on the weights of the company observations when a company in the sample splits into two. The first situation considered is that in which an index is estimated. It is assumed that an estimated index for some population or stratum can be expressed as

$$\hat{I}_s = \sum_{i=1}^{n} w_i I_i = w_1 I_1 + \sum_{i=2}^{n} w_i I_i. \qquad (4.15)$$

Suppose now that company 1 in the sample splits into two companies 1a and 1b. Assume furthermore that I_1 can be expressed as a weighted average of I_{1a} and I_{1b}, in other words

$$I_1 = w_{1a} I_{1a} + w_{1b} I_{1b}. \qquad (4.16)$$

Substituting (4.16) into (4.15) gives

$$\begin{aligned}\hat{I}_s &= \sum_{i=1}^{n} w_i I_i = w_1(w_{1a}I_{1a} + w_{1b}I_{1b}) + \sum_{i=2}^{n} w_i I_i \\ &= w_1 w_{1a} I_{1a} + w_1 w_{1b} I_{1b} + \sum_{i=2}^{n} w_i I_i .\end{aligned} \tag{4.17}$$

In other words, the weights of companies 1a and 1b resulting from the demerger are $w_1 w_{1a}$ and $w_1 w_{1b}$, respectively. It is assumed that both companies remain in the sample after the demerger.

The situation is somewhat different when e.g. the total revenue O has to be estimated for a population or stratum of companies. Assuming that the estimated revenue total can be expressed as

$$\hat{O}_s = w_1 O_1 + \sum_{i=2}^{n} w_i O_i \ , \tag{4.18}$$

the estimator after the demerger of company 1 becomes

$$\hat{O}_s = w_1 O_{1a} + w_1 O_{1b} + \sum_{i=2}^{n} w_i O_i . \tag{4.19}$$

In other words, the companies resulting from the demerger retain the same weight as the parent company before the demerger.

4.4 Quality indicators

Quality indicators for panels for estimating indexes are:

- the margins of uncertainty of the corresponding estimators;
- the size of nonresponse;

Nonresponse can severely affect the quality of the results if (i) the nonresponse is large and (ii) the nonresponse is selective. Adjustment of some of the bias from selective nonresponse can sometimes be achieved by using auxiliary variables that correspond with both the probability of response and the target variable.

5. References

Banning, R., Camstra, A. and Knottnerus, P. (2010). *Theme: Sampling theory, Subthemes: Sample design and Weighting methods*. Methods Series document, Statistics Netherlands, The Hague.

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

Knottnerus, P. (2011). On the efficiency of randomized PPS sampling, *Survey Methodology*, 37, 95-102.

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.

Version history

Version	Date	Description	Authors	Reviewers
Dutch version: Panels / Bedrijvenpanels				
1.0	23-03-2011	First Dutch version	Paul Knottnerus	Piet Daas Eric Schulte Nordholt
English version: Panels / Business panels				
1.0E	05-10-2011	First English version	Paul Knottnerus	