SR1_2 Analytical expressions for the accuracy of growth rates as affected by classification errors[1]

Sander Scholtus[2], Arnout van Delden[2] and Joep Burger[3]

[2] Statistics Netherlands, Henri Faasdreef 312, 2492 JP The Hague, The Netherlands.

[3] Statistics Netherlands, CBS-weg 11, P.O. Box 4481, 6401 CZ Heerlen, The Netherlands.

Key words: configuration 1 (complementary microdata sources), classification errors, analytical formulas, bias and variance, growth rates.

1 Introduction

National Statistical Institutes (NSIs) often publish business statistics for a number of economic industries. These industries are determined by a classification of economic activity. In European countries, since 2008, the NACE-code rev. 2 classification is used (Eurostat, 2008). Determining a single NACE code is not an easy task, since statistical units often have multiple activities, some of which are ancillary (such as holding activities). Eurostat (2008, chapter 3) provides rules on how to derive a single NACE code for a statistical unit given certain input information, such as the set of economic activities and their relative importance. We define the true NACE code of a statistical unit as the code that is obtained when these rules are correctly applied, using error-free input data. An additional requirement for the true NACE code will be given in the third paragraph.

In practice, European countries have a general business register (GBR) with an enumeration of all statistical units. For each unit the values for a few background variables are given, such as the NACE code. This GBR is in turn compiled from one or more administrative sources that usually contain legal units, their NACE codes and ownership relations between those legal units. The observed NACE codes within the GBR often deviate from the true ones for a number of reasons. Firstly, the statistical unit may be wrongly derived from the underlying legal units due to erroneous or missing information. Secondly, some error may have occurred when the legal unit was registered. In the Netherlands, this registration is held by the Chamber of Commerce (CoC). During registration, the legal unit or the person at the registration desk of the CoC may make an error. Thirdly, one or more of the legal units underlying the statistical unit may have changed their activities but failed to report this to the CoC. Fourthly, the rules to derive the economic activity might not have been applied correctly, for instance because information on the relative importance of the multiple activities of a statistical unit was not accurate enough.

At Statistics Netherlands and in a number of other countries, the NACE codes that are used for statistical purposes, are frozen (kept fixed) in the GBR during the year. The advantage of this approach is that outcomes of monthly and quarterly statistics can be more easily compared with those of yearly statistics within the same economic sector, because the effect of NACE code changes is eliminated. In the current paper, the true NACE code refers to this frozen NACE code, i.e., the true and up-to-date NACE code on January 1st of a given year is

[1] This work has been carried out as part of the ESSnet on Quality in Multisource Statistics, and has been funded by the European Commission

also defined to be the true code up to December 31st of that same year. Thus, the true code only changes once a year.

In the present paper, NACE code errors are applied to a case study : evaluating the accuracy of turnover growth rates for the Dutch short-term statistics in the presence of classification errors in the NACE codes in the Dutch GBR. This case study is introduced in van Delden et al. (2016a). They also describe an audit sample that was drawn to estimate the NACE code error probabilities. Two experts were asked to determine the true NACE code for the sampled units. These experts had access to the composition of the statistical unit into the various legal units, they had access to information of the CoC registration on the activities of the units and they checked this on the internet (if websites were available). In case of doubt, they contacted the business and asked about their activities. The experts determined the code independently of each other, and then discussed their results to come to a final conclusion.

In van Delden et al. (2016a), a bootstrap approach was used to estimate the bias and variance of turnover growth rates as affected by classification errors in the Dutch GBR. We refer to that paper for details on the bootstrap approach, as well as an introduction to the case study. A disadvantage of the use of bootstrap estimates is that they can be quite computationally intensive. An alternative approach could be to use analytical approximations to the bias and variance of growth rates under classification errors. The present paper continues the work of van Delden et al. (2016a) by deriving such analytical expressions; it is not completely self-contained.

In Section 2, we start by deriving generic analytical approximations to the bias and variance of a growth rate based on a Taylor series. Specific expressions are then worked out for quarter-on-quarter growth rates in Section 3 and for year-on-year growth rates in Section 4. The resulting expressions are rather complicated, involving many unknown parameters, and are therefore not very suitable for practical use at national statistical institutes. We will therefore introduce several simplifying assumptions to derive approximate expressions that can be applied more easily in practical situations. These assumptions do not necessarily simplify the mathematical content of these expressions, but they do lead to expressions that are easier to compute. We will also discuss how to estimate these analytical approximations in practice. We start with a full treatment of quarter-on-quarter growth rates within the same year (Section 3), before proceeding to the more complicated case of year-on-year growth rates (Section 4). Finally, in Section 5, we will examine how well these approximations work in practice, by applying them to the case study of van Delden et al. (2016a).

2 The bias and variance of a growth rate

2.1 A generic result on the bias and variance of a ratio

Although we are mainly interested here in the accuracy of quarterly and yearly growth rates under classification errors, it is useful to first discuss a slightly more general situation. Let U denote a target population of units. Suppose that we have observed two data sets, where the first data set contains a variable y^r for all units of a subpopulation $U^r \subset U$ and the second data set contains a variable y^q for all units of a subpopulation $U^q \subset U$. For units in the intersection $U^{r,q} = U^r \cap U^q$, both variables y^r and y^q are available. In what follows, we tacitly assume that this intersection is relatively large, i.e., that the two subpopulations have a large overlap.

Suppose that both subpopulations are divided into strata, where the set of possible stratum codes is denoted by $\{1, \dots, M\}$. Let s_i^r be the true stratum of unit $i \in U^r$ in the first data set, and s_i^q the true stratum of unit

$i \in U^q$ in the second data set (which may be different from s_i^r, for instance because the two data sets refer to different points in time). Let $a_{hi}^r = 1$ if $s_i^r = h$ and 0 otherwise, and similarly let $a_{hi}^q = 1$ if $s_i^q = h$ and 0 otherwise. Let Y_h^r be the total of variable y^r in stratum h and Y_h^q the stratum total for variable y^q, with $Y_h^r = \sum_{i \in U^r} a_{hi}^r y_i^r$ and $Y_h^q = \sum_{i \in U^q} a_{hi}^q y_i^q$. Our statistic of interest is the ratio $R_h^{q,r} = Y_h^q / Y_h^r$.

We consider the situation that the classification of units into these strata is prone to errors, further referred to as classification errors. That is to say, instead of s_i^r and s_i^q we observe $\hat{s}_i^r$ and $\hat{s}_i^q$ which may contain errors. Let $\hat{a}_{hi}^r = 1$ if $\hat{s}_i^r = h$ and 0 otherwise and similarly let $\hat{a}_{hi}^q$ be the observed version of a_{hi}^q. We estimate the stratum totals and their ratio by $\hat{Y}_h^r = \sum_{i \in U^r} \hat{a}_{hi}^r y_i^r$, $\hat{Y}_h^q = \sum_{i \in U^q} \hat{a}_{hi}^q y_i^q$, and $\hat{R}_h^{q,r} = \hat{Y}_h^q / \hat{Y}_h^r$, respectively. Note that, for simplicity, we assume that no other errors occur besides classification errors. In particular, we assume that the variables y^r and y^q are measured without error.

We assume that the classification errors *within each data set* are independent across units. For a given unit $i \in U^{r,q}$, the classification errors in $\hat{s}_i^r$ and $\hat{s}_i^q$ may be dependent. We are now interested in the bias and variance of $\hat{R}_h^{q,r}$ as an estimator for the true ratio $R_h^{q,r}$. Under the assumptions made here, the following approximate expressions may be derived (see the appendix):

$$
\begin{aligned}
B\left(\hat{R}_h^{q,r}\right) &\approx \frac{1}{\left[E\left(\hat{Y}_h^r\right)\right]^2}\left[\breve{R}_h^{q,r} \sum_{i \in U^r} (y_i^r)^2 V(\hat{a}_{hi}^r) - \sum_{i \in U^{r,q}} y_i^r y_i^q C\left(\hat{a}_{hi}^r, \hat{a}_{hi}^q\right)\right] + \left(\breve{R}_h^{q,r} - R_h^{q,r}\right), \\
V\left(\hat{R}_h^{q,r}\right) &\approx \frac{1}{\left[E\left(\hat{Y}_h^r\right)\right]^2}\left[\sum_{i \in U^q} \left(y_i^q\right)^2 V\left(\hat{a}_{hi}^q\right) + \left(\breve{R}_h^{q,r}\right)^2 \sum_{i \in U^r} (y_i^r)^2 V(\hat{a}_{hi}^r)\right. \\
&\qquad \left. - 2\breve{R}_h^{q,r} \sum_{i \in U^{r,q}} y_i^r y_i^q C\left(\hat{a}_{hi}^r, \hat{a}_{hi}^q\right)\right], \qquad (1) \\
E\left(\hat{Y}_h^r\right) &= \sum_{i \in U^r} y_i^r E(\hat{a}_{hi}^r), \\
\breve{R}_h^{q,r} &= \frac{E\left(\hat{Y}_h^q\right)}{E\left(\hat{Y}_h^r\right)} = \frac{\sum_{i \in U^q} y_i^q E\left(\hat{a}_{hi}^q\right)}{\sum_{i \in U^r} y_i^r E(\hat{a}_{hi}^r)},
\end{aligned}
$$

where $B(.)$, $V(.)$ and $C(.,.)$ denote a bias, variance and covariance, respectively. Thus, the bias $B\left(\hat{R}_h^{q,r}\right)$ and the variance $V\left(\hat{R}_h^{q,r}\right)$ are approximated as functions of $V(\hat{a}_{hi}^r)$, $V\left(\hat{a}_{hi}^q\right)$, $C\left(\hat{a}_{hi}^r, \hat{a}_{hi}^q\right)$, $E(\hat{a}_{hi}^r)$ and $E\left(\hat{a}_{hi}^q\right)$. The precise form of these components will depend on the way the classification errors are modelled and on the specific application. The specific application determines, for instance, if and how the classification errors for overlapping units between the two data sets are correlated.

2.2 Application to growth rates

We will now apply the generic formulae in (1) to find expressions for the approximate bias and variance of quarterly and yearly growth rates in the presence of random classification errors in a GBR. The following table connects the notation that will be used in the remainder of this paper to the notation that was used in Section 2.1. Some additional notation that is specific to this application will be introduced below as needed.

We will denote the bias and variance approximation to the order used in (1) by $AB(.)$ and $AV(.)$, respectively. It follows directly from expression (1) that the bias of $\hat{G}_h^{q,q-u} = \hat{Y}_h^q / \hat{Y}_h^{q-u}$ can be approximated by

$$
\begin{aligned}
AB\left(\hat{G}_h^{q,q-u}\right) = &\frac{1}{\left[E\left(\hat{Y}_h^{q-u}\right)\right]^2}\left[\breve{G}_h^{q,q-u} \sum_{i \in U^{q-u}} \left(y_i^{q-u}\right)^2 V\left(\hat{a}_{hi}^{q-u}\right) - \sum_{i \in U^{q-u,q}} y_i^{q-u} y_i^q C\left(\hat{a}_{hi}^{q-u}, \hat{a}_{hi}^q\right)\right] \\
&+ \left(\breve{G}_h^{q,q-u} - G_h^{q,q-u}\right). \qquad (2)
\end{aligned}
$$

Similarly, it follows from expression (1) that

$$AV\left(\hat{G}_h^{q,q-u}\right) = \frac{1}{\left[E\left(\hat{Y}_h^{q-u}\right)\right]^2}\Big[\sum_{i\in U^q}\left(y_i^q\right)^2 V\left(\hat{a}_{hi}^q\right) + \left(\breve{G}_h^{q,q-u}\right)^2 \sum_{i\in U^{q-u}}\left(y_i^{q-u}\right)^2 V\left(\hat{a}_{hi}^{q-u}\right) - 2\breve{G}_h^{q,q-u}\sum_{i\in U^{q-u,q}} y_i^{q-u} y_i^q C\left(\hat{a}_{hi}^{q-u}, \hat{a}_{hi}^q\right)\Big]. \quad (3)$$

Table 1. Explanation of frequently used symbols ($u = 1$ for quarterly growth rates, $u = 4$ for yearly ones)

Symbol	Meaning	Corresponding symbol (Sec. 2.1)
U^{q-u}, U^q	Population in quarters $q-u$ and q	U^r, U^q
$U^{q-u,q}$	Continuing units in population ($U^{q-u,q} = U^{q-u} \cap U^q$)	$U^{r,q}$
y_i^{q-u}, y_i^q	Turnover of unit i in quarters $q-u$ and q	y_i^r, y_i^q
$s_i^{q-u}, s_i^q, \hat{s}_i^{q-u}, \hat{s}_i^q$	True and observed stratum of unit i in quarters $q-u$ and q	$s_i^r, s_i^q, \hat{s}_i^r, \hat{s}_i^q$
$a_{hi}^{q-u}, a_{hi}^q, \hat{a}_{hi}^{q-u}, \hat{a}_{hi}^q$	True and observed stratum indicator in quarters $q-u$ and q	$a_{hi}^r, a_{hi}^q, \hat{a}_{hi}^r, \hat{a}_{hi}^q$
$Y_h^{q-u}, Y_h^q, \hat{Y}_h^{q-u}, \hat{Y}_h^q$	True and observed quarterly total turnover in stratum h	$Y_h^r, Y_h^q, \hat{Y}_h^r, \hat{Y}_h^q$
$G_h^{q,q-u}$	True growth rate in stratum h ($G_h^{q,q-u} = Y_h^q / Y_h^{q-u}$)	$R_h^{q,r}$
$\hat{G}_h^{q,q-u}$	Observed growth rate in stratum h ($\hat{G}_h^{q,q-u} = \hat{Y}_h^q / \hat{Y}_h^{q-u}$)	$\hat{R}_h^{q,r}$
$\breve{G}_h^{q,q-u}$	Ratio of expected observed turnover [$\breve{G}_h^{q,q-u} = E\left(\hat{Y}_h^q\right)/E\left(\hat{Y}_h^{q-u}\right)$]	$\breve{R}_h^{q,r}$

Regarding the component $\breve{G}_h^{q,q-u} - G_h^{q,q-u}$ in (2), it is interesting to note that

$$\breve{G}_h^{q,q-u} = \frac{E\left(\hat{Y}_h^q\right)}{E\left(\hat{Y}_h^{q-u}\right)} = \frac{Y_h^q + B\left(\hat{Y}_h^q\right)}{Y_h^{q-u} + B\left(\hat{Y}_h^{q-u}\right)} = \frac{Y_h^q}{Y_h^{q-u}} \frac{1 + RB\left(\hat{Y}_h^q\right)}{1 + RB\left(\hat{Y}_h^{q-u}\right)} = G_h^{q,q-u} \frac{1 + RB\left(\hat{Y}_h^q\right)}{1 + RB\left(\hat{Y}_h^{q-u}\right)},$$

where $RB\left(\hat{Y}_h^q\right) = B\left(\hat{Y}_h^q\right)/Y_h^q$ denotes the relative bias of $\hat{Y}_h^q$ (and similarly for $\hat{Y}_h^{q-u}$). If it is reasonable to assume that $RB\left(\hat{Y}_h^q\right) = RB\left(\hat{Y}_h^{q-u}\right)$, then it follows that $\breve{G}_h^{q,q-u} - G_h^{q,q-u} = 0$. In other words, if the relative bias of the estimated turnover levels does not vary much between quarters, the bias of the growth rates will be dominated by the first component in expression (2).

In the next two sections, we will derive explicit expressions for $AB\left(\hat{G}_h^{q,q-u}\right)$ and $AV\left(\hat{G}_h^{q,q-u}\right)$ under an assumed probability model for the classification errors that occur in $\hat{s}_i^{q-u}$ and $\hat{s}_i^q$. We use models that were developed previously in Burger et al. (2015), van Delden et al. (2016b, 2016a).

3 Bias and variance approximations for quarter-on-quarter growth rates

3.1 Introduction

In this section, we will consider the relatively simple case of a quarter-on-quarter growth rate ($u = 1$) within the *same* year. This case is made simpler by the fact that, as mentioned in the introduction, the Dutch GBR uses coordinated industry codes: the industry codes of units in the GBR are only changed between 31 December and 1 January and are subsequently kept fixed during the year. Because of this convention of coordinated industry codes, we may assume that $s_i^q = s_i^{q-1}$ and $\hat{s}_i^q = \hat{s}_i^{q-1}$ for all units that exist in both quarters ($i \in U^{q-1,q}$).

In Section 3.2 we will derive explicit expressions for $AB\left(\hat{G}_h^{q,q-1}\right)$ and $AV\left(\hat{G}_h^{q,q-1}\right)$ based on (2) and (3). The resulting expressions (7) and (8) below are exact to the order of approximation in the Taylor series, but too complicated for practical use at NSIs. Subsequently, we will therefore derive approximations to these expressions by introducing some simplifying assumptions about the nature of the classification errors. In Section 3.3, we start with a stable population and a very simple model that may be unrealistic, but which leads to very simple expressions for the bias and variance that can be interpreted easily. We then proceed to a more realistic version of this model in Section 3.4. Finally, births and deaths are re-introduced into the model in Section 3.5. Note also that in practice, for a given case study with real data, we cannot directly compute the expressions for the expectations, bias and variance, since we do not know the true stratum that the units belong to in the different quarters. The estimation of the bias and variance will be treated in Section 3.6.

3.2 Bias and variance for quarter-on-quarter growth rates

In what follows, the classification errors in $\hat{s}_i^{q-1}$ (and in $\hat{s}_i^q$ for units that occur only in U^q) are described by the so-called level matrix $\mathbf{P}_i^{OL} = (p_{ghi}^{OL})$, with $p_{ghi}^{OL} = P\left(\hat{s}_i^{q-1} = h \middle| s_i^{q-1} = g\right)$. Note that, for the moment, we allow that each unit i has its own level matrix; hence, the introduction of such a matrix to describe classification errors involves no real loss of generalisation.

Let $\boldsymbol{a}_i^{q-u} = (a_{1i}^{q-u}, \dots, a_{Mi}^{q-u})^T$ and $\widehat{\boldsymbol{a}}_i^{q-u} = (\hat{a}_{1i}^{q-u}, \dots, \hat{a}_{Mi}^{q-u})^T$ for $u \in \{0,1,4\}$. By analogy to Burger et al. (2015), the following properties can be derived for units $i \in U^{q-1}$:

$$\begin{aligned} E\left(\widehat{\boldsymbol{a}}_i^{q-1}\right) &= (\mathbf{P}_i^{OL})^T \boldsymbol{a}_i^{q-1}, \\ V\left(\widehat{\boldsymbol{a}}_i^{q-1}\right) &= \operatorname{diag}\left[(\mathbf{P}_i^{OL})^T \boldsymbol{a}_i^{q-1}\right] - (\mathbf{P}_i^{OL})^T \operatorname{diag}(\boldsymbol{a}_i^{q-1})\, \mathbf{P}_i^{OL}. \end{aligned} \tag{4}$$

In particular, it holds that

$$\begin{aligned} E\left(\hat{a}_{hi}^{q-1}\right) &= \sum_{g=1}^{M} a_{gi}^{q-1} p_{ghi}^{OL}, \\ V\left(\hat{a}_{hi}^{q-1}\right) &= \sum_{g=1}^{M} a_{gi}^{q-1} p_{ghi}^{OL} - \sum_{g=1}^{M} a_{gi}^{q-1} \left(p_{ghi}^{OL}\right)^2 = \sum_{g=1}^{M} a_{gi}^{q-1} p_{ghi}^{OL} \left(1 - p_{ghi}^{OL}\right). \end{aligned} \tag{5}$$

For units that exist in both quarters ($i \in U^{q-1,q}$), $\widehat{\boldsymbol{a}}_i^q$ has the exact same expectation and variance as $\widehat{\boldsymbol{a}}_i^{q-1}$. Furthermore, for these units $C\left(\hat{a}_{hi}^{q-1}, \hat{a}_{hi}^q\right) = V\left(\hat{a}_{hi}^{q-1}\right)$. For units that occur only in U^q, the expressions for $E\left(\widehat{\boldsymbol{a}}_i^q\right)$ and $V\left(\widehat{\boldsymbol{a}}_i^q\right)$ are similar to (4), but with $\boldsymbol{a}_i^{q-1}$ replaced by $\boldsymbol{a}_i^q$. From (4) and (5), it follows that

$$\begin{aligned} E\left(\hat{Y}_h^{q-1}\right) &= \sum_{i \in U^{q-1}} \left(y_i^{q-1} \sum_{g=1}^{M} E\left(\hat{a}_{ghi}^{q-1}\right) \right) \\ &= \sum_{i \in U^{q-1}} \left(y_i^{q-1} \sum_{g=1}^{M} a_{gi}^{q-1} p_{ghi}^{OL} \right), \\ \breve{G}_h^{q,q-1} &= \frac{\sum_{i \in U^q} \left(y_i^q \sum_{g=1}^{M} a_{gi}^q p_{ghi}^{OL} \right)}{\sum_{i \in U^{q-1}} \left(y_i^{q-1} \sum_{g=1}^{M} a_{gi}^{q-1} p_{ghi}^{OL} \right)}, \\ \sum_{i \in U^{q-1}} \left(y_i^{q-1}\right)^2 V\left(\hat{a}_{hi}^{q-1}\right) &= \sum_{i \in U^{q-1}} \left[\left(y_i^{q-1}\right)^2 \sum_{g=1}^{M} a_{gi}^{q-1} p_{ghi}^{OL} \left(1 - p_{ghi}^{OL}\right) \right], \\ \sum_{i \in U^{q-1,q}} y_i^{q-1} y_i^q C\left(\hat{a}_{hi}^{q-1}, \hat{a}_{hi}^q\right) &= \sum_{i \in U^{q-1,q}} \left[G_i^{q,q-1} \left(y_i^{q-1}\right)^2 \sum_{g=1}^{M} a_{gi}^{q-1} p_{ghi}^{OL} \left(1 - p_{ghi}^{OL}\right) \right], \end{aligned} \tag{6}$$

where $G_i^{q,q-1} = y_i^q / y_i^{q-1}$ denotes the individual growth rate of unit $i \in U^{q-1,q}$.

Applying these results to (2), we obtain the following approximate expression for the bias of $\hat{G}_h^{q,q-1}$:

$$
\begin{aligned}
AB\left(\hat{G}_h^{q,q-1}\right) &= \frac{1}{\left[E\left(\hat{Y}_h^{q-1}\right)\right]^2}\left\{\sum_{i\in U^{q-1,q}}\left[\left(\breve{G}_h^{q,q-1}-G_i^{q,q-1}\right)\left(y_i^{q-1}\right)^2\sum_{g=1}^{M}a_{gi}^{q-1}p_{ghi}^{OL}\left(1-p_{ghi}^{OL}\right)\right]\right. \\
&\quad \left. + \breve{G}_h^{q,q-1}\sum_{i\in U^{q-1}\setminus U^{q-1,q}}\left[\left(y_i^{q-1}\right)^2\sum_{g=1}^{M}a_{gi}^{q-1}p_{ghi}^{OL}\left(1-p_{ghi}^{OL}\right)\right]\right\} \\
&\quad + \left(\breve{G}_h^{q,q-1}-G_h^{q,q-1}\right) \\
&= \frac{1}{\left[E\left(\hat{Y}_h^{q-1}\right)\right]^2}\left(B_{hO}^{q,q-1}+B_{hD}^{q,q-1}\right)+\left(\breve{G}_h^{q,q-1}-G_h^{q,q-1}\right),
\end{aligned}
\tag{7}
$$

with $E\left(\hat{Y}_h^{q-1}\right)$ and $\breve{G}_h^{q,q-1}$ as given in (6). Notice that the bias consists of three components: the continuing or overlapping units (subscript O), the dead units (subscript D) and a correction-term involving all units.

Similarly, for the approximate variance of $\hat{G}_h^{q,q-1}$ we obtain from (3):

$$
\begin{aligned}
AV\left(\hat{G}_h^{q,q-1}\right) &= \frac{1}{\left[E\left(\hat{Y}_h^{q-1}\right)\right]^2}\left\{\sum_{i\in U^{q-1,q}}\left[\left(\breve{G}_h^{q,q-1}-G_i^{q,q-1}\right)^2\left(y_i^{q-1}\right)^2\sum_{g=1}^{M}a_{gi}^{q-1}p_{ghi}^{OL}\left(1-p_{ghi}^{OL}\right)\right]\right. \\
&\quad + \left(\breve{G}_h^{q,q-1}\right)^2\sum_{i\in U^{q-1}\setminus U^{q-1,q}}\left[\left(y_i^{q-1}\right)^2\sum_{g=1}^{M}a_{gi}^{q-1}p_{ghi}^{OL}\left(1-p_{ghi}^{OL}\right)\right] \\
&\quad \left. + \sum_{i\in U^{q}\setminus U^{q-1,q}}\left[\left(y_i^{q}\right)^2\sum_{g=1}^{M}a_{gi}^{q}p_{ghi}^{OL}\left(1-p_{ghi}^{OL}\right)\right]\right\} \\
&= \frac{1}{\left[E\left(\hat{Y}_h^{q-1}\right)\right]^2}\left(V_{hO}^{q,q-1}+V_{hD}^{q,q-1}+V_{hB}^{q,q-1}\right).
\end{aligned}
\tag{8}
$$

In the derivation of the first component of this expression, we used the fact that continuing units $i \in U^{q-1,q}$ occur in all three parts of expression (3) with $V\left(\hat{a}_{hi}^{q}\right)=V\left(\hat{a}_{hi}^{q-1}\right)=C\left(\hat{a}_{hi}^{q-1},\hat{a}_{hi}^{q}\right)$, and that furthermore

$$
\begin{aligned}
\left(y_i^q\right)^2+\left(\breve{G}_h^{q,q-1}\right)^2\left(y_i^{q-1}\right)^2-2\breve{G}_h^{q,q-1}y_i^{q-1}y_i^q &= \left[\left(G_i^{q,q-1}\right)^2+\left(\breve{G}_h^{q,q-1}\right)^2-2\breve{G}_h^{q,q-1}G_i^{q,q-1}\right]\left(y_i^{q-1}\right)^2 \\
&= \left(\breve{G}_h^{q,q-1}-G_i^{q,q-1}\right)^2\left(y_i^{q-1}\right)^2.
\end{aligned}
$$

Notice that the variance consists of three components: the continuing or overlapping units (subscript O), the dead units (subscript D) and the new-born units (subscripts B).

3.3 Stable population and a single p value

3.3.1 Assumptions and notation

In order to derive simpler bias and variance approximations, we begin by making two strong assumptions:

1. The population for the quarters $q-1$ and q is stable, i.e., there are no births and deaths, so $U^{q-1}=U^{q}=U^{q-1,q}$.
2. All units in the population are correctly classified with probability p, and all misclassified units are divided uniformly over the remaining industries. Thus, all units have the same level matrix $\mathbf{P}_i^{OL}$ with elements given by

$$
p_{ghi}^{OL}=\begin{cases} p & \text{if } g=h \\ \dfrac{1-p}{M-1} & \text{if } g\neq h \end{cases}
$$

In the literature on linkage errors, a model with a matrix of this form is known as an *exchangeable linkage errors* model (Neter et al., 1965; Chambers, 2009). In the context of classification errors, Burger et al. (2015) studied the accuracy of estimated turnover levels under a model of this form. We can immediately reproduce one of their results by applying assumption 2 to our expression for $E\left(\hat{Y}_h^{q-1}\right)$ in (6):

$$
\begin{aligned}
E\left(\hat{Y}_h^{q-1}\right) &= \sum\nolimits_{i\in U^{q-1}} y_i^{q-1}\left[a_{hi}^{q-1}p + \left(1-a_{hi}^{q-1}\right)\frac{1-p}{M-1}\right] \\
&= p\sum\nolimits_{i\in U_h^{q-1}} y_i^{q-1} + \frac{1-p}{M-1}\sum\nolimits_{i\in U^{q-1}\setminus U_h^{q-1}} y_i^{q-1} \\
&= pY_h^{q-1} + (1-p)\bar{Y}_{(-h)}^{q-1},
\end{aligned}
\tag{9}
$$

with

$$
\bar{Y}_{(-h)}^{q-1} = \frac{Y^{q-1} - Y_h^{q-1}}{M-1},
$$

where $Y^{q-u} = \sum_{g=1}^{M} Y_g^{q-u}$ denotes the total turnover in quarter $q-u$ ($u \in \{0,1,4\}$) for all units in the population. Note that $\bar{Y}_{(-h)}^{q-1}$ denotes the true average quarterly turnover across all strata in the population except h.

Similarly, we find that

$$
\begin{aligned}
E\left(\hat{Y}_h^{q}\right) &= pY_h^{q} + (1-p)\bar{Y}_{(-h)}^{q}, \\
\bar{Y}_{(-h)}^{q} &= \frac{Y^{q} - Y_h^{q}}{M-1}.
\end{aligned}
\tag{10}
$$

From (9) and (10) it follows immediately that

$$
\breve{G}_h^{q,q-1} = \frac{E(\hat{Y}_h^{q})}{E(\hat{Y}_h^{q-1})} = \frac{pY_h^{q} + (1-p)\bar{Y}_{(-h)}^{q}}{pY_h^{q-1} + (1-p)\bar{Y}_{(-h)}^{q-1}}.
\tag{11}
$$

To shorten the notation in the remainder of this section, it is useful to introduce a pseudo-residual $e_{hi}^{q,q-1}$:

$$
e_{hi}^{q,q-1} = \left(G_i^{q,q-1} - \breve{G}_h^{q,q-1}\right)y_i^{q-1} = y_i^{q} - \breve{G}_h^{q,q-1}y_i^{q-1}.
\tag{12}
$$

(Recall that $G_i^{q,q-1} = y_i^{q}/y_i^{q-1}$.)

It will also be useful to have shorthand expressions for the sum and the sum of squares of a generic variable z for all units in (true) stratum g:

$$
\begin{aligned}
S_g(z) &= \sum\nolimits_{i\in U^{q-1}} a_{gi}^{q-1} z_i\,, \quad g = 1,\dots,M, \\
SS_g(z) &= \sum\nolimits_{i\in U^{q-1}} a_{gi}^{q-1} z_i^2\,, \quad g = 1,\dots,M.
\end{aligned}
\tag{13}
$$

We also define a sum and a sum of squares over all units in the population:

$$
S(z) = \sum\nolimits_{i\in U^{q-1}} z_i = \sum\nolimits_{g=1}^{M} S_g(z).
$$

$$
SS(z) = \sum\nolimits_{i\in U^{q-1}} z_i^2 = \sum\nolimits_{g=1}^{M} SS_g(z).
$$

Finally, analogous to $\bar{Y}_{(-h)}^{q-1}$ and $\bar{Y}_{(-h)}^{q}$, we define:

$$
\begin{aligned}
\bar{S}_{(-h)}(z) &= \frac{S(z) - S_h(z)}{M-1}, \\
\overline{SS}_{(-h)}(z) &= \frac{SS(z) - SS_h(z)}{M-1}.
\end{aligned}
\tag{14}
$$

Note that $\bar{Y}_{(-h)}^{q-1} = \bar{S}_{(-h)}(y^{q-1})$ and $\bar{Y}_{(-h)}^{q} = \bar{S}_{(-h)}(y^{q})$.

3.3.2 Bias

Under assumption 1 from Section 3.3.1, the approximate bias $AB\big(\hat{G}_h^{q,q-1}\big)$ in formula (7) reduces to:

$$AB\big(\hat{G}_h^{q,q-1}\big) = \frac{1}{\big[E\big(\hat{Y}_h^{q-1}\big)\big]^2}\left\{\sum\nolimits_{i\in U^{q-1}}\left[\big(\breve{G}_h^{q,q-1} - G_i^{q,q-1}\big)\big(y_i^{q-1}\big)^2\sum\nolimits_{g=1}^{M} a_{gi}^{q-1}p_{ghi}^{OL}\big(1-p_{ghi}^{OL}\big)\right]\right\} + \big(\breve{G}_h^{q,q-1} - G_h^{q,q-1}\big). \tag{15}$$

Next, using assumption 2 and the notation introduced above, we find:

$$\begin{aligned} AB\big(\hat{G}_h^{q,q-1}\big) &= \big(\breve{G}_h^{q,q-1} - G_h^{q,q-1}\big) - \frac{\sum_{i\in U^{q-1}} e_{hi}^{q,q-1}y_i^{q-1}\Big[a_{hi}^{q-1}p(1-p) + \big(1-a_{hi}^{q-1}\big)\frac{1-p}{M-1}\Big(1-\frac{1-p}{M-1}\Big)\Big]}{\big[E\big(\hat{Y}_h^{q-1}\big)\big]^2} \\ &= \big(\breve{G}_h^{q,q-1} - G_h^{q,q-1}\big) \\ &\quad - \frac{p(1-p)S_h\big(e_h^{q,q-1}y^{q-1}\big) + \frac{1-p}{M-1}\Big(1-\frac{1-p}{M-1}\Big)\big[S\big(e_h^{q,q-1}y^{q-1}\big) - S_h\big(e_h^{q,q-1}y^{q-1}\big)\big]}{\big[E\big(\hat{Y}_h^{q-1}\big)\big]^2} \\ &= \big(\breve{G}_h^{q,q-1} - G_h^{q,q-1}\big) - \frac{p(1-p)S_h\big(e_h^{q,q-1}y^{q-1}\big) + (1-p)\Big(1-\frac{1-p}{M-1}\Big)\bar{S}_{(-h)}\big(e_h^{q,q-1}y^{q-1}\big)}{\big[E\big(\hat{Y}_h^{q-1}\big)\big]^2} \end{aligned}$$

and thus

$$AB\big(\hat{G}_h^{q,q-1}\big) = \big(\breve{G}_h^{q,q-1} - G_h^{q,q-1}\big) - \frac{(1-p)\Big[pS_h\big(e_h^{q,q-1}y^{q-1}\big) + \Big(1-\frac{1-p}{M-1}\Big)\bar{S}_{(-h)}\big(e_h^{q,q-1}y^{q-1}\big)\Big]}{\big[E\big(\hat{Y}_h^{q-1}\big)\big]^2}, \tag{16}$$

with $E\big(\hat{Y}_h^{q-1}\big)$ given by (9) and

$$\breve{G}_h^{q,q-1} - G_h^{q,q-1} = \frac{pY_h^q + (1-p)\bar{Y}_{(-h)}^q}{pY_h^{q-1} + (1-p)\bar{Y}_{(-h)}^{q-1}} - \frac{Y_h^q}{Y_h^{q-1}}$$

according to (11). We also used the shorthand notation from (13) and (14). Note that

$$S_g\big(e_h^{q,q-1}y^{q-1}\big) = \sum\nolimits_{i\in U^{q-1}} a_{gi}^{q-1}e_{hi}^{q,q-1}y_i^{q-1} = \sum\nolimits_{i\in U^{q-1}} a_{gi}^{q-1}\big(G_i^{q,q-1} - \breve{G}_h^{q,q-1}\big)\big(y_i^{q-1}\big)^2.$$

Since the number of strata M is large in practice (for the Dutch GBR, $M \approx 300$), the above expression can be simplified further by noting that

$$\begin{aligned} \bar{Y}_{(-h)}^{q} &\approx \bar{Y}^q = \frac{Y^q}{M}, \\ \breve{G}_h^{q,q-1} &\approx \frac{pY_h^q + (1-p)\bar{Y}^q}{pY_h^{q-1} + (1-p)\bar{Y}^{q-1}} = \frac{\bar{Y}^q + p\big(Y_h^q - \bar{Y}^q\big)}{\bar{Y}^{q-1} + p\big(Y_h^{q-1} - \bar{Y}^{q-1}\big)} \equiv \breve{G}_{h0}^{q,q-1}, \\ e_{hi}^{q,q-1} &\approx y_i^q - \breve{G}_{h0}^{q,q-1}y_i^{q-1} \equiv e_{h0}^{q,q-1}, \\ S_g\big(e_h^{q,q-1}y^{q-1}\big) &\approx S_g\big(e_{h0}^{q,q-1}y^{q-1}\big) = S_g\big(y^{q-1}y^q\big) - \breve{G}_{h0}^{q,q-1}SS_g\big(y^{q-1}\big), \\ \bar{S}_{(-h)}\big(e_h^{q,q-1}y^{q-1}\big) &\approx \bar{S}\big(e_{h0}^{q,q-1}y^{q-1}\big) = \frac{1}{M}\sum\nolimits_{g=1}^{M} S_g\big(e_{h0}^{q,q-1}y^{q-1}\big). \end{aligned}$$

Hence, we obtain the following further approximation to (16) that may be slightly easier to compute:

$$AB\left(\hat{G}_h^{q,q-1}\right) \approx \frac{\bar{Y}^q + p\left(Y_h^q - \bar{Y}^q\right)}{\bar{Y}^{q-1} + p\left(Y_h^{q-1} - \bar{Y}^{q-1}\right)} - \frac{Y_h^q}{Y_h^{q-1}} - \frac{(1-p)\left[pS_h\left(e_{h0}^{q,q-1}y^{q-1}\right) + \left(1 - \frac{1-p}{M-1}\right)\bar{S}\left(e_{h0}^{q,q-1}y^{q-1}\right)\right]}{\left[\bar{Y}^{q-1} + p\left(Y_h^{q-1} - \bar{Y}^{q-1}\right)\right]^2}. \quad (17)$$

Results (16) and (17) show that the approximate bias $AB\left(\hat{G}_h^{q,q-1}\right)$ varies as a function of the key characteristics p, $S_h\left(e_h^{q,q-1}y^{q-1}\right)$ and $\bar{S}_{(-h)}\left(e_h^{q,q-1}y^{q-1}\right)$. It can be shown that when p – the probability of correct classification – goes to 1, $AB\left(\hat{G}_h^{q,q-1}\right)$ goes to zero. Furthermore, when the growth rates of the individual units within the target stratum h vary more, the pseudo-residuals $e_{hi}^{q,q-1}$ and $S_h\left(e_h^{q,q-1}y^{q-1}\right)$ will become larger and so, when all other characteristics remain equal, the absolute bias will also increase. Likewise, when units outside the target stratum vary more, then $\bar{S}_{(-h)}\left(e_h^{q,q-1}y^{q-1}\right)$ becomes larger and (all else being equal) so will the absolute bias. Finally, we remark that expression (16) for $AB\left(\hat{G}_h^{q,q-1}\right)$ consists of two components: a term $\breve{G}_h^{q,q-1} - G_h^{q,q-1}$ that follows directly from the bias in turnover levels $\hat{Y}_h^{q-1}$ and $\hat{Y}_h^q$ and an additional bias component due to variations in individual turnover growth rates. We already noted in Section 2.2 that the second of these components will dominate the bias in the growth rate if the relative bias in turnover levels does not vary much across quarters.

3.3.3 Variance

For the approximate variance $AV\left(\hat{G}_h^{q,q-1}\right)$ we find that, again under assumption 1 from Section 3.3.1, formula (8) reduces to

$$AV\left(\hat{G}_h^{q,q-1}\right) = \frac{1}{\left[E\left(\hat{Y}_h^{q-1}\right)\right]^2}\left\{\sum\nolimits_{i\in U^{q-1},q}\left[\left(\breve{G}_h^{q,q-1} - G_i^{q,q-1}\right)^2\left(y_i^{q-1}\right)^2\sum\nolimits_{g=1}^{M} a_{gi}^{q-1}p_{ghi}^{OL}\left(1 - p_{ghi}^{OL}\right)\right]\right\}. \quad (18)$$

Next we use assumption 2 and proceed along similar lines as for the bias to obtain:

$$\begin{aligned} AV\left(\hat{G}_h^{q,q-1}\right) &= \frac{\sum_{i\in U^{q-1}}\left(e_{hi}^{q,q-1}\right)^2\left[a_{hi}^{q-1}p(1-p) + \left(1 - a_{hi}^{q-1}\right)\frac{1-p}{M-1}\left(1 - \frac{1-p}{M-1}\right)\right]}{\left[E\left(\hat{Y}_h^{q-1}\right)\right]^2} \\ &= \frac{p(1-p)SS_h\left(e_h^{q,q-1}\right) + \frac{1-p}{M-1}\left(1 - \frac{1-p}{M-1}\right)\left[SS\left(e_h^{q,q-1}\right) - SS_h\left(e_h^{q,q-1}\right)\right]}{\left[E\left(\hat{Y}_h^{q-1}\right)\right]^2} \\ &= \frac{(1-p)\left[pSS_h\left(e_h^{q,q-1}\right) + \left(1 - \frac{1-p}{M-1}\right)\overline{SS}_{(-h)}\left(e_h^{q,q-1}\right)\right]}{\left[E\left(\hat{Y}_h^{q-1}\right)\right]^2}, \end{aligned} \quad (19)$$

with $E\left(\hat{Y}_h^{q-1}\right)$ given by (9). Here, we used the shorthand notation for sums of squares defined in (13) and (14). Note that

$$SS_g\left(e_h^{q,q-1}\right) = \sum\nolimits_{i\in U^{q-1}} a_{gi}^{q-1}\left(e_{hi}^{q,q-1}\right)^2 = \sum\nolimits_{i\in U^{q-1}} a_{gi}^{q-1}\left(G_i^{q,q-1} - \breve{G}_h^{q,q-1}\right)^2\left(y_i^{q-1}\right)^2, \quad g = 1, \ldots, M.$$

It is instructive to write the numerator of the last line of (19) in a slightly different form:

$$\left[p(1-p)SS_h\left(e_h^{q,q-1}\right)\right] + (M-1)\left[\frac{1-p}{M-1}\left(1 - \frac{1-p}{M-1}\right)\overline{SS}_{(-h)}\left(e_h^{q,q-1}\right)\right]. \quad (20)$$

This shows that the numerator of the above variance approximation consists of M terms, corresponding to the contributions of the M industries: the target industry h has a contribution of $p(1-p)SS_h\left(e_h^{q,q-1}\right)$ and each of the $(M-1)$ remaining industries has the comparable contribution $\frac{1-p}{M-1}\left(1-\frac{1-p}{M-1}\right)\overline{SS}_{(-h)}\left(e_h^{q,q-1}\right)$.

As with the bias, we can make some further approximations when the number of strata M is large, to obtain a variance approximation that is more amenable to easy computation in practice:

$$AV\left(\hat{G}_h^{q,q-1}\right) \approx \frac{(1-p)\left[pSS_h\left(e_{h0}^{q,q-1}\right)+\left(1-\frac{1-p}{M-1}\right)\overline{SS}\left(e_{h0}^{q,q-1}\right)\right]}{\left[\bar{Y}^{q-1}+p\left(Y_h^{q-1}-\bar{Y}^{q-1}\right)\right]^2}, \tag{21}$$

where we re-used some of the notation from Section 3.3.2. Also, $\overline{SS}\left(e_{h0}^{q,q-1}\right)=SS\left(e_{h0}^{q,q-1}\right)/M$.

According to (19), the variance approximation varies as a function of p, $SS_h\left(e_h^{q,q-1}\right)$ and $\overline{SS}_{(-h)}\left(e_h^{q,q-1}\right)$. It is easy to see that when p approaches 1, $AV\left(\hat{G}_h^{q,q-1}\right)$ goes to zero, which stands for the situation that there are no classification errors. In addition, when the growth rates of the different units in the target stratum h varies more, $SS_h\left(e_h^{q,q-1}\right)$ increases and so will the variance. The same holds for the individual growth rates of units outside the target stratum.

3.4 Stable population and multiple p values

3.4.1 Assumptions and notation

The above assumption 2 that all units in the population have the same level matrix of classification error probabilities and that, moreover, the probabilities in this matrix can be described by a single parameter p, is not realistic for the Dutch GBR. We will now relax this assumption in two steps. (For the moment, we retain assumption 1 that the population is stable; this assumption will be relaxed in Section 3.5.)

The first step acknowledges that some units are more likely to be classified in their correct stratum than others, for instance because of the amount of attention paid to these units during the construction of the GBR. For instance, the large and most complex units are checked more carefully than the smaller ones. For this refinement, van Delden et al. (2016b) introduced the concept of a *probability class*. The population U^{q-1} is partitioned into probability classes $U_1^{q-1},\ldots,U_C^{q-1}$, with $U_1^{q-1}\cup\cdots\cup U_C^{q-1}=U^{q-1}$ and $U_c^{q-1}\cap U_d^{q-1}=\emptyset$ for all $c\neq d$. All units $i\in U_c^{q-u}$ are supposed to have the same level matrix $\mathbf{P}_i^{OL}=\mathbf{P}_c^{OL}$, that is the probability classes are homogeneous with respect to the classification error probabilities. We now obtain the following version of assumption 2:

2'. The population consists of probability classes $U_1^{q-1},\ldots,U_C^{q-1}$. All units in probability class U_c^{q-1} are correctly classified with probability p_c, and all misclassified units in probability class U_c^{q-1} are divided uniformly over the remaining industries. Thus, all units in probability class U_c^{q-1} have the same level matrix $\mathbf{P}_c^{OL}$ with elements given by

$$p_{ghc}^{OL}=\begin{cases} p_c & \text{if } i\in U_c^{q-1} \text{ and } g=h \\ \dfrac{1-p_c}{M-1} & \text{if } i\in U_c^{q-1} \text{ and } g\neq h \end{cases}$$

It is not difficult to show that, with assumption 2 replaced by assumption 2', the following bias and variance approximations are obtained instead of (16) and (19):

$$AB\left(\hat{G}_h^{q,q-1}\right) = \frac{\sum_{c=1}^{\mathcal{C}}\left[p_c Y_{hc}^q + (1-p_c)\bar{Y}_{(-h)c}^q\right]}{\sum_{c=1}^{\mathcal{C}}\left[p_c Y_{hc}^{q-1} + (1-p_c)\bar{Y}_{(-h)c}^{q-1}\right]} - \frac{Y_h^q}{Y_h^{q-1}}$$
$$- \frac{\sum_{c=1}^{\mathcal{C}}(1-p_c)\left[p_c S_{hc}\left(e_h^{q,q-1}y^{q-1}\right) + \left(1-\frac{1-p_c}{M-1}\right)\bar{S}_{(-h)c}\left(e_h^{q,q-1}y^{q-1}\right)\right]}{\left\{\sum_{c=1}^{\mathcal{C}}\left[p_c Y_{hc}^{q-1} + (1-p_c)\bar{Y}_{(-h)c}^{q-1}\right]\right\}^2}, \qquad (22)$$
$$AV\left(\hat{G}_h^{q,q-1}\right) = \frac{\sum_{c=1}^{\mathcal{C}}(1-p_c)\left[p_c SS_{hc}\left(e_h^{q,q-1}\right) + \left(1-\frac{1-p_c}{M-1}\right)\overline{SS}_{(-h)c}\left(e_h^{q,q-1}\right)\right]}{\left\{\sum_{c=1}^{\mathcal{C}}\left[p_c Y_{hc}^{q-1} + (1-p_c)\bar{Y}_{(-h)c}^{q-1}\right]\right\}^2}.$$

Here, $Y_{hc}^{q-1} = \sum_{i\in U_c^{q-1}} a_{hi}^{q-1} y_i^{q-1}$ and $Y_{hc}^q = \sum_{i\in U_c^q} a_{hi}^{q-1} y_i^q$ denote the stratum turnover levels within probability class U_c^{q-1}. We have also generalised the notation from Section 3.3.1 to operate within probability classes:

$$\bar{Y}_{(-h)c}^{q-u} = \frac{Y_{+c}^{q-u} - Y_{hc}^{q-u}}{M-1} = \frac{\sum_{g=1}^M Y_{gc}^{q-u} - Y_{hc}^{q-u}}{M-1}, \quad u = 0,1,$$
$$S_{gc}(z) = \sum\nolimits_{i\in U_c^{q-1}} a_{gi}^{q-1} z_i, \quad g = 1,\dots,M,$$
$$\bar{S}_{(-h)c}(z) = \frac{S_{+c}(z) - S_{hc}(z)}{M-1} = \frac{\sum_{g=1}^M S_{gc}(z) - S_{hc}(z)}{M-1}, \qquad (23)$$
$$SS_{gc}(z) = \sum\nolimits_{i\in U_c^{q-1}} a_{gi}^{q-1} z_i^2, \quad g = 1,\dots,M,$$
$$\overline{SS}_{(-h)c}(z) = \frac{SS_{+c}(z) - SS_{hc}(z)}{M-1} = \frac{\sum_{g=1}^M SS_{gc}(z) - SS_{hc}(z)}{M-1}.$$

A second step towards a more realistic classification error model is to acknowledge that, within each level matrix $\mathbf{P}_c^{OL}$, not all diagonal probabilities need to be equal to a single value p_c, and also the off-diagonal probabilities need not all be equal to each other. To keep the bias and variance expressions manageable in practice, we do want to limit as much as possible the number of parameters needed to describe each matrix $\mathbf{P}_c^{OL}$. [If all elements of $\mathbf{P}_c^{OL}$ were left free, then hardly any simplification of expressions (7) and (8) would be possible.] As a compromise between simplicity and accuracy, we propose the following. For each true industry g there exists a subset of possible observed industry codes $\mathcal{H}_g \subset \{1,\dots,M\}$ for which specific classification probabilities p_{ghc}^{OL} apply (with $\sum_{h\in\mathcal{H}_g} p_{ghc}^{OL} \leq 1$). It is assumed that $g \in \mathcal{H}_g$, so the diagonal probabilities are always included in this subset. The subset $\mathcal{H}_g$ is supposed to be chosen as small as possible (in particular: $|\mathcal{H}_g|$ is much smaller than M) but at the same time large enough to capture all 'significant' cases, so that each probability p_{ghc}^{OL} with $h \notin \mathcal{H}_g$ is small. For ease of notation, we will assume here that the same subsets $\mathcal{H}_g$ apply to all probability classes, but this assumption is not strictly necessary and it is not difficult to generalise the results below to probability-class-specific subsets $\mathcal{H}_{gc}$.

We can define an indicator $I_{gh} = 1$ if $h \in \mathcal{H}_g$ and $I_{gh} = 0$ otherwise. From the point of view of an observed industry code h, these indicators define a subset of the true industries: $\mathcal{G}_h = \{g|I_{gh} = 1\} = \{g|h \in \mathcal{H}_g\}$. This subset consists of all true industries g for which the classification probability p_{ghc}^{OL} is 'significant'. The remaining probabilities p_{ghc}^{OL} with $g \notin \mathcal{G}_h$ for a given observed industry code h are not necessarily equal to each other, but they are all close to zero. This suggests that they may be approximated by their mean value, which is $(p_{+hc}^{OL} - \sum_{g\in\mathcal{G}_h} p_{ghc}^{OL})/(M - |\mathcal{G}_h|)$, with $p_{+hc}^{OL} = \sum_{g=1}^M p_{ghc}^{OL}$. Note that $p_{+hc}^{OL} \neq 1$ in general. (The exchangeable errors model is an exception.)

This leads to the following, final version of assumption 2:

2''. The population consists of probability classes $U_1^{q-1}, \ldots, U_C^{q-1}$. All units in probability class U_c^{q-1} have the same level matrix $\mathbf{P}_c^{OL}$ that can be approximated well by a matrix $\widetilde{\mathbf{P}}_c^{OL}$ with elements given by

$$\tilde{p}_{ghc}^{OL} = \begin{cases} p_{ghc}^{OL} & \text{if } i \in U_c^{q-1} \text{ and } g \in \mathcal{G}_h \\ \dfrac{p_{+hc}^{OL} - \sum_{g \in \mathcal{G}_h} p_{ghc}^{OL}}{M - |\mathcal{G}_h|} & \text{if } i \in U_c^{q-1} \text{ and } g \notin \mathcal{G}_h \end{cases}$$

Note that, by definition, $\widetilde{\mathbf{P}}_c^{OL}$ can only be an approximation to the real level matrix $\mathbf{P}_c^{OL}$, because the rows of $\widetilde{\mathbf{P}}_c^{OL}$ do not necessarily sum to 1.

In what follows, we will use the following shorthand notation which partly generalises (23):

$$\begin{aligned} \tilde{p}_{(-\mathcal{G}_h)hc}^{OL} &= p_{+hc}^{OL} - \sum\nolimits_{g \in \mathcal{G}_h} p_{ghc}^{OL}, \\ \bar{Y}_{(-\mathcal{G}_h)c}^{q-u} &= \frac{Y_{+c}^{q-u} - \sum_{g \in \mathcal{G}_h} Y_{gc}^{q-u}}{M - |\mathcal{G}_h|}, \quad u = 0,1, \\ \bar{S}_{(-\mathcal{G}_h)c}(z) &= \frac{S_{+c}(z) - \sum_{g \in \mathcal{G}_h} S_{gc}(z)}{M - |\mathcal{G}_h|}, \\ \overline{SS}_{(-\mathcal{G}_h)c}(z) &= \frac{SS_{+c}(z) - \sum_{g \in \mathcal{G}_h} SS_{gc}(z)}{M - |\mathcal{G}_h|}. \end{aligned} \tag{24}$$

3.4.2 Bias

Starting again with expression (6) and proceeding along similar lines as in (9), we obtain the following approximation to $E\left(\hat{Y}_h^{q-1}\right)$ under assumption 2'':

$$\begin{aligned} E\left(\hat{Y}_h^{q-1}\right) &= \sum_{c=1}^{\mathcal{C}} \sum\nolimits_{i \in U_c^{q-1}} y_i^{q-1} \sum\nolimits_{g=1}^{M} a_{gi}^{q-1} p_{ghc}^{OL} \\ &\approx \sum_{c=1}^{\mathcal{C}} \sum\nolimits_{i \in U_c^{q-1}} y_i^{q-1} \left[\sum_{g \in \mathcal{G}_h} a_{gi}^{q-1} p_{ghc}^{OL} + \left(1 - \sum_{g \in \mathcal{G}_h} a_{gi}^{q-1} \right) \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}}{M - |\mathcal{G}_h|} \right] \\ &= \sum_{c=1}^{\mathcal{C}} \left[\sum_{g \in \mathcal{G}_h} p_{ghc}^{OL} Y_{gc}^{q-1} + \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}}{M - |\mathcal{G}_h|} \left(Y_{+c}^{q-1} - \sum_{g \in \mathcal{G}_h} Y_{gc}^{q-1} \right) \right] \\ &= \sum_{c=1}^{\mathcal{C}} \left[\sum_{g \in \mathcal{G}_h} p_{ghc}^{OL} Y_{gc}^{q-1} + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL} \bar{Y}_{(-\mathcal{G}_h)c}^{q-1} \right], \end{aligned} \tag{25}$$

where we used notation defined in (24). Note that in the second line of (25) we used $\sum_{g \in \mathcal{G}_h} a_{gi}^{q-1} + \sum_{g \notin \mathcal{G}_h} a_{gi}^{q-1} = 1$ thus $\sum_{g \notin \mathcal{G}_h} a_{gi}^{q-1} = 1 - \sum_{g \in \mathcal{G}_h} a_{gi}^{q-1}$ since each unit observed in h belongs to one true industry code g. Analogously, we also obtain:

$$\begin{aligned} E\left(\hat{Y}_h^{q}\right) &\approx \sum_{c=1}^{\mathcal{C}} \left[\sum_{g \in \mathcal{G}_h} p_{ghc}^{OL} Y_{gc}^{q} + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL} \bar{Y}_{(-\mathcal{G}_h)c}^{q} \right], \\ \breve{G}_h^{q,q-1} &= \frac{E\left(\hat{Y}_h^{q}\right)}{E\left(\hat{Y}_h^{q-1}\right)} \approx \frac{\sum_{c=1}^{\mathcal{C}} \left[\sum_{g \in \mathcal{G}_h} p_{ghc}^{OL} Y_{gc}^{q} + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL} \bar{Y}_{(-\mathcal{G}_h)c}^{q} \right]}{\sum_{c=1}^{\mathcal{C}} \left[\sum_{g \in \mathcal{G}_h} p_{ghc}^{OL} Y_{gc}^{q-1} + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL} \bar{Y}_{(-\mathcal{G}_h)c}^{q-1} \right]}. \end{aligned} \tag{26}$$

To approximate the bias, we can still use (15) as a starting point. We can proceed analogously to Section 3.3.2 to approximate the numerator in this expression:

$$
\begin{aligned}
&\sum_{c=1}^{\mathcal{C}} \sum\nolimits_{i \in U_c^{q-1}} e_{hi}^{q,q-1} y_i^{q-1} \left[\sum\nolimits_{g=1}^{M} a_{gi}^{q-1} p_{ghi}^{OL} (1 - p_{ghi}^{OL}) \right] \\
&\approx \sum_{c=1}^{\mathcal{C}} \sum\nolimits_{i \in U_c^{q-1}} e_{hi}^{q,q-1} y_i^{q-1} \left[\sum_{g \in \mathcal{G}_h} a_{gi}^{q-1} p_{ghc}^{OL} (1 - p_{ghc}^{OL}) + \left(1 - \sum_{g \in \mathcal{G}_h} a_{gi}^{q-1} \right) \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}}{M - |\mathcal{G}_h|} \left(1 - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}}{M - |\mathcal{G}_h|} \right) \right] \\
&= \sum_{c=1}^{\mathcal{C}} \left\{ \sum_{g \in \mathcal{G}_h} p_{ghc}^{OL} (1 - p_{ghc}^{OL}) S_{gc} \left(e_h^{q,q-1} y^{q-1} \right) \right. \\
&\qquad \left. + \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}}{M - |\mathcal{G}_h|} \left(1 - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}}{M - |\mathcal{G}_h|} \right) \left[S_{+c} \left(e_h^{q,q-1} y^{q-1} \right) - \sum_{g \in \mathcal{G}_h} S_{gc} \left(e_h^{q,q-1} y^{q-1} \right) \right] \right\} \\
&= \sum_{c=1}^{\mathcal{C}} \left[\sum_{g \in \mathcal{G}_h} p_{ghc}^{OL} (1 - p_{ghc}^{OL}) S_{gc} \left(e_h^{q,q-1} y^{q-1} \right) + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL} \left(1 - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}}{M - |\mathcal{G}_h|} \right) \bar{S}_{(-\mathcal{G}_h)c} \left(e_h^{q,q-1} y^{q-1} \right) \right].
\end{aligned}
$$

Combining this with (15), (25) and (26), we obtain:

$$
\begin{aligned}
&AB\left(\hat{G}_h^{q,q-1} \right) \\
&\approx \frac{\sum_{c=1}^{\mathcal{C}} \left[\sum_{g \in \mathcal{G}_h} p_{ghc}^{OL} Y_{gc}^{q} + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL} \bar{Y}_{(-\mathcal{G}_h)c}^{q} \right]}{\sum_{c=1}^{\mathcal{C}} \left[\sum_{g \in \mathcal{G}_h} p_{ghc}^{OL} Y_{gc}^{q-1} + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL} \bar{Y}_{(-\mathcal{G}_h)c}^{q-1} \right]} - \frac{Y_h^q}{Y_h^{q-1}} \\
&- \frac{\sum_{c=1}^{\mathcal{C}} \left[\sum_{g \in \mathcal{G}_h} p_{ghc}^{OL} (1 - p_{ghc}^{OL}) S_{gc} \left(e_h^{q,q-1} y^{q-1} \right) + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL} \left(1 - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}}{M - |\mathcal{G}_h|} \right) \bar{S}_{(-\mathcal{G}_h)c} \left(e_h^{q,q-1} y^{q-1} \right) \right]}{\left\{ \sum_{c=1}^{\mathcal{C}} \left[\sum_{g \in \mathcal{G}_h} p_{ghc}^{OL} Y_{gc}^{q-1} + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL} \bar{Y}_{(-\mathcal{G}_h)c}^{q-1} \right] \right\}^2}.
\end{aligned}
\tag{27}
$$

3.4.3 Variance

To approximate the variance, we can use (18) as a starting point. For the numerator in this expression, we find:

$$
\begin{aligned}
&\sum_{c=1}^{\mathcal{C}} \sum\nolimits_{i \in U_c^{q-1}} \left(e_{hi}^{q,q-1} \right)^2 \left[\sum\nolimits_{g=1}^{M} a_{gi}^{q-1} p_{ghi}^{OL} (1 - p_{ghi}^{OL}) \right] \\
&\approx \sum_{c=1}^{\mathcal{C}} \left\{ \sum_{g \in \mathcal{G}_h} p_{ghc}^{OL} (1 - p_{ghc}^{OL}) SS_{gc} \left(e_h^{q,q-1} \right) + \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}}{M - |\mathcal{G}_h|} \left(1 - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}}{M - |\mathcal{G}_h|} \right) \left[SS_{+c} \left(e_h^{q,q-1} \right) - \sum_{g \in \mathcal{G}_h} SS_{gc} \left(e_h^{q,q-1} \right) \right] \right\} \\
&= \sum_{c=1}^{\mathcal{C}} \left[\sum_{g \in \mathcal{G}_h} p_{ghc}^{OL} (1 - p_{ghc}^{OL}) SS_{gc} \left(e_h^{q,q-1} \right) + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL} \left(1 - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}}{M - |\mathcal{G}_h|} \right) \overline{SS}_{(-\mathcal{G}_h)c} \left(e_h^{q,q-1} \right) \right],
\end{aligned}
$$

by proceeding along similar lines as before. Combining this with approximation (25) for the denominator in (18), we obtain the following variance approximation:

$$
\begin{aligned}
&AV\left(\hat{G}_h^{q,q-1} \right) \\
&\approx \frac{\sum_{c=1}^{\mathcal{C}} \left[\sum_{g \in \mathcal{G}_h} p_{ghc}^{OL} (1 - p_{ghc}^{OL}) SS_{gc} \left(e_h^{q,q-1} \right) + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL} \left(1 - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}}{M - |\mathcal{G}_h|} \right) \overline{SS}_{(-\mathcal{G}_h)c} \left(e_h^{q,q-1} \right) \right]}{\left\{ \sum_{c=1}^{\mathcal{C}} \left[\sum_{g \in \mathcal{G}_h} p_{ghc}^{OL} Y_{gc}^{q-1} + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL} \bar{Y}_{(-\mathcal{G}_h)c}^{q-1} \right] \right\}^2}.
\end{aligned}
\tag{28}
$$

Analogous to (20), each term in the above numerator can also be written as

$$\sum_{g\in\mathcal{G}_h}\left[p_{ghc}^{OL}\left(1-p_{ghc}^{OL}\right)SS_{gc}\left(e_h^{q,q-1}\right)\right]+\left(M-|\mathcal{G}_h|\right)\left[\frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}}{M-|\mathcal{G}_h|}\left(1-\frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}}{M-|\mathcal{G}_h|}\right)\overline{SS}_{(-\mathcal{G}_h)c}\left(e_h^{q,q-1}\right)\right].$$

Thus, it is still true that – within each probability class – each industry has a similar contribution, namely $p_{ghc}^{OL}\left(1-p_{ghc}^{OL}\right)SS_{gc}\left(e_h^{q,q-1}\right)$ for the 'significant' industries $g\in\mathcal{G}_h$ and $\frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}}{M-|\mathcal{G}_h|}\left(1-\frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}}{M-|\mathcal{G}_h|}\right)\overline{SS}_{(-\mathcal{G}_h)c}\left(e_h^{q,q-1}\right)$ for each of the other industries $g\notin\mathcal{G}_h$.

3.5 Dynamic population

3.5.1 Assumptions and notation

We will now drop assumption 1 from Section 3.3.1, that is, we will allow for changes in the population between the quarters $q-1$ and q. This means that we can no longer ignore the terms in the bias and variance approximations in Section 3.2 that concern dead units ($i\in U^{q-1}\backslash U^{q-1,q}$) and new-born units ($i\in U^{q}\backslash U^{q-1,q}$). In this section, we will retain assumption 2''. So far, we have defined the probability classes only for units in the population at time $q-1$. We now assume that new-born units at time q are also assigned to a probability class. By extension of assumption 2'', it is supposed that the level matrices for the new-born units can be approximated well by the same matrices that apply to units in U^{q-1}. We also assume that units do not switch probability classes between $q-1$ and q. This is a realistic assumption for the Dutch GBR.

In what follows, we will re-use much of the notation developed in the previous subsections. We will use the convention that turnover levels, sums and sums of squares that were defined previously for a stable population now refer to all relevant units that exist at a given point in time. Thus, for instance, Y_{hc}^{q} includes the turnover of continuing *and* new-born units in quarter q that belong to probability class c and industry h. If we want to restrict a population quantity to the subset of continuing units, deaths or births, we add a subscript O, D or B, respectively, to that quantity. Thus, for instance, $Y_{hc}^{q}=Y_{hcO}^{q}+Y_{hcB}^{q}$ and $Y_{hc}^{q-1}=Y_{hcO}^{q-1}+Y_{hcD}^{q-1}$.

3.5.2 Bias

If the population is not stable between quarters $q-1$ and q, all terms in the bias approximation in formula (7) are relevant:

$$AB\left(\hat{G}_h^{q,q-1}\right)=\frac{1}{\left[E\left(\hat{Y}_h^{q-1}\right)\right]^2}\left(B_{hO}^{q,q-1}+B_{hD}^{q,q-1}\right)+\left(\breve{G}_h^{q,q-1}-G_h^{q,q-1}\right).$$

The component $E\left(\hat{Y}_h^{q-1}\right)$ can be approximated by expression (25) as before, and $\left(\breve{G}_h^{q,q-1}-G_h^{q,q-1}\right)$ can still be approximated as in expression (26):

$$\breve{G}_h^{q,q-1}-G_h^{q,q-1}=\frac{E\left(\hat{Y}_h^{q}\right)}{E\left(\hat{Y}_h^{q-1}\right)}-\frac{Y_h^q}{Y_h^{q-1}}\approx\frac{\sum_{c=1}^{\mathcal{C}}\left[\sum_{g\in\mathcal{G}_h}p_{ghc}^{OL}Y_{gc}^{q}+\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}\bar{Y}_{(-\mathcal{G}_h)c}^{q}\right]}{\sum_{c=1}^{\mathcal{C}}\left[\sum_{g\in\mathcal{G}_h}p_{ghc}^{OL}Y_{gc}^{q-1}+\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}\bar{Y}_{(-\mathcal{G}_h)c}^{q-1}\right]}-\frac{Y_h^q}{Y_h^{q-1}},$$

using the above-defined notation convention. It remains to evaluate the two components $B_{hO}^{q,q-1}$ and $B_{hD}^{q,q-1}$.

The component $B_{hO}^{q,q-1}$ refers to continuing units. To approximate this term, we can adapt expression (27):

$$B_{hO}^{q,q-1}\approx-\sum_{c=1}^{\mathcal{C}}\left[\sum_{g\in\mathcal{G}_h}p_{ghc}^{OL}\left(1-p_{ghc}^{OL}\right)S_{gcO}\left(e_h^{q,q-1}y^{q-1}\right)+\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}\left(1-\frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}}{M-|\mathcal{G}_h|}\right)\bar{S}_{(-\mathcal{G}_h)cO}\left(e_h^{q,q-1}y^{q-1}\right)\right],$$

where $S_{gcO}\left(e_h^{q,q-1}y^{q-1}\right)$ and $\bar{S}_{(-\mathcal{G}_h)cO}\left(e_h^{q,q-1}y^{q-1}\right)$ are now computed using only the units in $U^{q-1,q}$. It is important to note that the pseudo-residual $e_h^{q,q-1}$ is still given by (12), using the overall $\breve{G}_h^{q,q-1}$.

Finally, the component $B_{hD}^{q,q-1}$ for units in $U^{q-1}\backslash U^{q-1,q}$ in (7) has a similar form to $B_{hO}^{q,q-1}$. It is not difficult to see that we may therefore write:

$$B_{hD}^{q,q-1} \approx \breve{G}_h^{q,q-1}\sum_{c=1}^{\mathcal{C}}\left[\sum_{g\in\mathcal{G}_h} p_{ghc}^{OL}\left(1-p_{ghc}^{OL}\right)SS_{gcD}(y^{q-1}) + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL}\left(1-\frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}}{M-|\mathcal{G}_h|}\right)\overline{SS}_{(-\mathcal{G}_h)cD}(y^{q-1})\right].$$

Combining these three expressions, we find the following approximation for the total bias:

$$\begin{aligned} AB\left(\hat{G}_h^{q,q-1}\right) \approx{}& \frac{\sum_{c=1}^{\mathcal{C}}\left[\sum_{g\in\mathcal{G}_h} p_{ghc}^{OL}Y_{gc}^{q} + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL}\bar{Y}_{(-\mathcal{G}_h)c}^{q}\right]}{\sum_{c=1}^{\mathcal{C}}\left[\sum_{g\in\mathcal{G}_h} p_{ghc}^{OL}Y_{gc}^{q-1} + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL}\bar{Y}_{(-\mathcal{G}_h)c}^{q-1}\right]} - \frac{Y_h^q}{Y_h^{q-1}} \\ &- \frac{\sum_{c=1}^{\mathcal{C}}\left[\sum_{g\in\mathcal{G}_h} p_{ghc}^{OL}\left(1-p_{ghc}^{OL}\right)BS_{gc}^{q,q-1} + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL}\left(1-\frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}}{M-|\mathcal{G}_h|}\right)\overline{BS}_{(-\mathcal{G}_h)c}^{q,q-1}\right]}{\left\{\sum_{c=1}^{\mathcal{C}}\left[\sum_{g\in\mathcal{G}_h} p_{ghc}^{OL}Y_{gc}^{q-1} + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL}\bar{Y}_{(-\mathcal{G}_h)c}^{q-1}\right]\right\}^2}, \qquad (29) \\ BS_{gc}^{q,q-1} ={}& S_{gcO}\left(e_h^{q,q-1}y^{q-1}\right) - \breve{G}_h^{q,q-1}SS_{gcD}(y^{q-1}), \\ \overline{BS}_{(-\mathcal{G}_h)c}^{q,q-1} ={}& \bar{S}_{(-\mathcal{G}_h)cO}\left(e_h^{q,q-1}y^{q-1}\right) - \breve{G}_h^{q,q-1}\overline{SS}_{(-\mathcal{G}_h)cD}(y^{q-1}). \end{aligned}$$

Note that we can combine the contributions of the continuing and dead units into two single terms, because these contributions are linear in $S_{gcO}\left(e_h^{q,q-1}y^{q-1}\right)$ and $SS_{gcD}(y^{q-1})$, and in $\bar{S}_{(-\mathcal{G}_h)cO}\left(e_h^{q,q-1}y^{q-1}\right)$ and $\overline{SS}_{(-\mathcal{G}_h)cD}(y^{q-1})$, respectively.

3.5.3 Variance

If the population is not stable between quarters $q-1$ and q, the variance approximation in formula (8) consists of three components:

$$AV\left(\hat{G}_h^{q,q-1}\right) = \frac{1}{\left[E\left(\hat{Y}_h^{q-1}\right)\right]^2}\left(V_{hO}^{q,q-1} + V_{hD}^{q,q-1} + V_{hB}^{q,q-1}\right).$$

The component $V_{hO}^{q,q-1}$ refers to continuing units. To approximate this term, we can adapt expression (28):

$$V_{hO}^{q,q-1} \approx \sum_{c=1}^{\mathcal{C}}\left[\sum_{g\in\mathcal{G}_h} p_{ghc}^{OL}(1-p_{ghc}^{OL})SS_{gcO}\left(e_h^{q,q-1}\right) + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL}\left(1-\frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}}{M-|\mathcal{G}_h|}\right)\overline{SS}_{(-\mathcal{G}_h)cO}\left(e_h^{q,q-1}\right)\right],$$

again using the convention that $SS_{gcO}\left(e_h^{q,q-1}\right)$ and $\overline{SS}_{(-\mathcal{G}_h)cO}\left(e_h^{q,q-1}\right)$ are computed using only the units in $U^{q-1,q}$.

For the remaining two components, we can again use that they are similar in form to the corresponding component for continuing units which we have just evaluated. The component for dead units $V_{hD}^{q,q-1}$ may therefore be approximated by

$$V_{hD}^{q,q-1} \approx \left(\breve{G}_h^{q,q-1}\right)^2\sum_{c=1}^{\mathcal{C}}\left[\sum_{g\in\mathcal{G}_h} p_{ghc}^{OL}\left(1-p_{ghc}^{OL}\right)SS_{gcD}(y^{q-1}) + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL}\left(1-\frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}}{M-|\mathcal{G}_h|}\right)\overline{SS}_{(-\mathcal{G}_h)cD}(y^{q-1})\right],$$

and the component for new-born units $V_{hB}^{q,q-1}$ may be approximated by

$$V_{hB}^{q,q-1} \approx \sum_{c=1}^{\mathcal{C}} \left[\sum_{g \in \mathcal{G}_h} p_{ghc}^{OL}(1-p_{ghc}^{OL}) SS_{gcB}(y^q) + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL} \left(1 - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}}{M - |\mathcal{G}_h|}\right) \overline{SS}_{(-\mathcal{G}_h)cB}(y^q) \right].$$

Combining these three expressions, we find the following approximation for the total variance:

$$\begin{aligned} AV\left(\hat{G}_h^{q,q-1}\right) &\approx \frac{\sum_{c=1}^{\mathcal{C}} \left[\sum_{g \in \mathcal{G}_h} p_{ghc}^{OL}(1-p_{ghc}^{OL}) VS_{gc}^{q,q-1} + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL} \left(1 - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}}{M - |\mathcal{G}_h|}\right) \overline{VS}_{(-\mathcal{G}_h)c}^{q,q-1} \right]}{\left\{ \sum_{c=1}^{\mathcal{C}} \left[\sum_{g \in \mathcal{G}_h} p_{ghc}^{OL} Y_{gc}^{q-1} + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL} \bar{Y}_{(-\mathcal{G}_h)c}^{q-1} \right] \right\}^2} \\ VS_{gc}^{q,q-1} &= SS_{gcO}\left(e_h^{q,q-1}\right) + SS_{gcD}\left(\check{G}_h^{q,q-1} y^{q-1}\right) + SS_{gcB}(y^q), \\ \overline{VS}_{(-\mathcal{G}_h)c}^{q,q-1} &= \overline{SS}_{(-\mathcal{G}_h)cO}\left(e_h^{q,q-1}\right) + \overline{SS}_{(-\mathcal{G}_h)cD}\left(\check{G}_h^{q,q-1} y^{q-1}\right) + \overline{SS}_{(-\mathcal{G}_h)cB}(y^q). \end{aligned} \tag{30}$$

Note that, again, we can combine the contributions of the different subsets of units into two single terms, because the above expressions are linear in the sums of squares.

3.6 Estimating the bias and variance

The bias and variance approximations for $\hat{G}_h^{q,q-1}$ that have been derived above depend on the true turnover totals per industry. Of course, in practice, these will be unknown. We will now discuss how to estimate these expressions for the bias and variance of $\hat{G}_h^{q,q-1}$.

The bias and variance approximations also depend on the classification error probabilities p_{ghc}^{OL}. Here, we will treat these as known parameters. In practice, they have to be estimated as well, for instance from an audit sample of units for which the true industry codes are known. Although this introduces uncertainty in the estimated bias and variance of the observed growth rates, it does not affect the bias and variance values themselves, since $\hat{G}_h^{q,q-1}$ does not depend on the estimated classification error probabilities.

3.6.1 Ratio of the expectations

Recall that the ratio of the expectations, $\check{G}_h^{q,q-1}$, is given in Section 3.5.2 by

$$\check{G}_h^{q,q-1} \approx \frac{\sum_{c=1}^{\mathcal{C}} \left[\sum_{g \in \mathcal{G}_h} p_{ghc}^{OL} Y_{gc}^{q} + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL} \bar{Y}_{(-\mathcal{G}_h)c}^{q} \right]}{\sum_{c=1}^{\mathcal{C}} \left[\sum_{g \in \mathcal{G}_h} p_{ghc}^{OL} Y_{gc}^{q-1} + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL} \bar{Y}_{(-\mathcal{G}_h)c}^{q-1} \right]}.$$

We can estimate $\check{G}_h^{q,q-1}$by:

$$\hat{\check{G}}_h^{q,q-1} = \frac{\sum_{c=1}^{\mathcal{C}} \left[\sum_{g \in \mathcal{G}_h} p_{ghc}^{OL} \hat{Y}_{gc}^{q} + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL} \hat{\bar{Y}}_{(-\mathcal{G}_h)c}^{q} \right]}{\sum_{c=1}^{\mathcal{C}} \left[\sum_{g \in \mathcal{G}_h} p_{ghc}^{OL} \hat{Y}_{gc}^{q-1} + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL} \hat{\bar{Y}}_{(-\mathcal{G}_h)c}^{q-1} \right]}, \tag{31}$$

with, for $u \in \{0,1\}$, $\hat{Y}_{gc}^{q-u} = \sum_{i \in U_c^{q-u}} \hat{a}_{gi}^{q-u} y_i^{q-u}$ and

$$\hat{\bar{Y}}_{(-\mathcal{G}_h)c}^{q-u} = \frac{Y_{+c}^{q-u} - \sum_{g \in \mathcal{G}_h} \hat{Y}_{gc}^{q-u}}{M - |\mathcal{G}_h|}.$$

Note that, since we have assumed that classification errors are the only errors that occur, the total turnover in each probability class, Y_{+c}^{q-u}, is observed without error.

3.6.2 Bias

To estimate the bias approximation (29), we replace all unknown quantities by their observed values:

$$\widehat{AB}\left(\hat{G}_h^{q,q-1}\right) = \frac{\sum_{c=1}^{\mathcal{C}}\left[\sum_{g\in\mathcal{G}_h} p_{ghc}^{OL}\hat{Y}_{gc}^{q} + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL}\hat{\tilde{Y}}_{(-\mathcal{G}_h)c}^{q}\right]}{\sum_{c=1}^{\mathcal{C}}\left[\sum_{g\in\mathcal{G}_h} p_{ghc}^{OL}\hat{Y}_{gc}^{q-1} + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL}\hat{\tilde{Y}}_{(-\mathcal{G}_h)c}^{q-1}\right]} - \frac{\hat{Y}_h^{q}}{\hat{Y}_h^{q-1}}$$
$$- \frac{\sum_{c=1}^{\mathcal{C}}\left[\sum_{g\in\mathcal{G}_h} p_{ghc}^{OL}(1-p_{ghc}^{OL})\widehat{BS}_{gc}^{q,q-1} + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL}\left(1-\frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}}{M-|\mathcal{G}_h|}\right)\widehat{\widetilde{BS}}_{(-\mathcal{G}_h)c}^{q,q-1}\right]}{\left\{\sum_{c=1}^{\mathcal{C}}\left[\sum_{g\in\mathcal{G}_h} p_{ghc}^{OL}\hat{Y}_{gc}^{q-1} + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL}\hat{\tilde{Y}}_{(-\mathcal{G}_h)c}^{q-1}\right]\right\}^2}, \tag{32}$$
$$\widehat{BS}_{gc}^{q,q-1} = \hat{S}_{gcO}\left(\hat{e}_h^{q,q-1}y^{q-1}\right) - \hat{G}_h^{q,q-1}\widehat{SS}_{gcD}(y^{q-1}),$$
$$\widehat{\widetilde{BS}}_{(-\mathcal{G}_h)c}^{q,q-1} = \hat{\tilde{S}}_{(-\mathcal{G}_h)cO}\left(\hat{e}_h^{q,q-1}y^{q-1}\right) - \hat{\tilde{G}}_h^{q,q-1}\widehat{\widetilde{SS}}_{(-\mathcal{G}_h)cD}(y^{q-1}).$$

In this expression, $\hat{G}_h^{q,q-1}$ is given by (31) and the observed residual $\hat{e}_{hi}^{q,q-1}$ is defined as

$$\hat{e}_{hi}^{q,q-1} = \left(G_i^{q,q-1} - \hat{G}_h^{q,q-1}\right)y_i^{q-1} = y_i^{q} - \hat{G}_h^{q,q-1}y_i^{q-1}. \tag{33}$$

For continuing units, the estimated sums and sums of squares are defined for an arbitrary variable z^{q-u} with respect to quarter $q-u$ ($u \in \{0,1\}$) by:

$$\hat{S}_{gcO}(z^{q-u}) = \sum\nolimits_{i\in U_c^{q-1,q}} \hat{a}_{gi}^{q-u} z_i^{q-u}, \quad g = 1,\dots,M,$$
$$\hat{\tilde{S}}_{(-\mathcal{G}_h)cO}(z^{q-u}) = \frac{S_{+cO}(z^{q-u}) - \sum_{g\in\mathcal{G}_h}\hat{S}_{gcO}(z^{q-u})}{M-|\mathcal{G}_h|} = \frac{\sum_{g=1}^{M}\hat{S}_{gcO}(z^{q-u}) - \sum_{g\in\mathcal{G}_h}\hat{S}_{gcO}(z^{q-u})}{M-|\mathcal{G}_h|},$$
$$\widehat{SS}_{gcO}(z^{q-u}) = \sum\nolimits_{i\in U_c^{q-1,q}} \hat{a}_{gi}^{q-u}(z_i^{q-u})^2, \quad g = 1,\dots,M, \tag{34}$$
$$\widehat{\widetilde{SS}}_{(-\mathcal{G}_h)cO}(z^{q-u}) = \frac{SS_{+cO}(z^{q-u}) - \sum_{g\in\mathcal{G}_h}\widehat{SS}_{gcO}(z^{q-u})}{M-|\mathcal{G}_h|} = \frac{\sum_{g=1}^{M}\widehat{SS}_{gcO}(z^{q-u}) - \sum_{g\in\mathcal{G}_h}\widehat{SS}_{gcO}(z^{q-u})}{M-|\mathcal{G}_h|}.$$

The definitions for new-born units and dead units are analogous, with O replaced by B and D, respectively. Note that, analogous to Y_{+c}^{q-u}, the total sums $S_{+cO}(z^{q-u})$ and $SS_{+cO}(z^{q-u})$ in (34) are over fixed sets of units. Whether the outcome of these sums is also error-free depends on the variable z^{q-u}; for instance $SS_{+cO}(y^{q-1})$ is known but $S_{+cO}\left(\hat{e}_h^{q,q-1}y^{q-1}\right)$ is estimated.

3.6.3 Variance

Variance approximation (30) can be estimated by:

$$\widehat{AV}\left(\hat{G}_h^{q,q-1}\right) = \frac{\sum_{c=1}^{\mathcal{C}}\left[\sum_{g\in\mathcal{G}_h} p_{ghc}^{OL}(1-p_{ghc}^{OL})\widehat{VS}_{gc}^{q,q-1} + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL}\left(1-\frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}}{M-|\mathcal{G}_h|}\right)\widehat{\widetilde{VS}}_{(-\mathcal{G}_h)c}^{q,q-1}\right]}{\left\{\sum_{c=1}^{\mathcal{C}}\left[\sum_{g\in\mathcal{G}_h} p_{ghc}^{OL}\hat{Y}_{gc}^{q-1} + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL}\hat{\tilde{Y}}_{(-\mathcal{G}_h)c}^{q-1}\right]\right\}^2} \tag{35}$$
$$\widehat{VS}_{gc}^{q,q-1} = \widehat{SS}_{gcO}\left(\hat{e}_h^{q,q-1}\right) + \widehat{SS}_{gcD}\left(\hat{G}_h^{q,q-1}y^{q-1}\right) + \widehat{SS}_{gcB}(y^{q}),$$
$$\widehat{\widetilde{VS}}_{(-\mathcal{G}_h)c}^{q,q-1} = \widehat{\widetilde{SS}}_{(-\mathcal{G}_h)cO}\left(\hat{e}_h^{q,q-1}\right) + \widehat{\widetilde{SS}}_{(-\mathcal{G}_h)cD}\left(\hat{G}_h^{q,q-1}y^{q-1}\right) + \widehat{\widetilde{SS}}_{(-\mathcal{G}_h)cB}(y^{q}),$$

where $\hat{e}_h^{q,q-1}$ is given by (33) and the estimated sums of squares are defined as in (34).

3.6.4 Final remarks

In the above formulas, we have estimated the turnover totals Y_{hc}^{q-u} per industry h and probability class by their observed counterparts $\hat{Y}_{hc}^{q-u}$. In practice, these are biased due to the presence of classification errors. Since the expectation of $\hat{Y}_{hc}^{q-u}$ is approximated by

$$E\left(\hat{Y}_{hc}^{q-u}\right) \approx \sum_{g\in\mathcal{G}_h} p_{ghc}^{OL}Y_{gc}^{q-u} + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL}\bar{Y}_{(-\mathcal{G}_h)c}^{q-u},$$

its bias is approximately given by

$$B\left(\hat{Y}_{hc}^{q-u}\right) \approx \left(p_{hhc}^{OL}-1\right)Y_{hc}^{q-u} + \sum_{g\in\mathcal{G}_h\setminus\{h\}} p_{ghc}^{OL}Y_{gc}^{q-u} + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL}\bar{Y}_{(-\mathcal{G}_h)c}^{q-u}. \tag{36}$$

From this, it follows that $\hat{G}_h^{q,q-1}$ in (31) may also be biased. As a first-order approximation, we find:

$$\begin{aligned} E\left(\hat{G}_h^{q,q-1}\right) &= E\left\{\frac{\sum_{c=1}^{\mathcal{C}}\left[\sum_{g\in\mathcal{G}_h} p_{ghc}^{OL}\hat{Y}_{gc}^{q} + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL}\hat{\bar{Y}}_{(-\mathcal{G}_h)c}^{q}\right]}{\sum_{c=1}^{\mathcal{C}}\left[\sum_{g\in\mathcal{G}_h} p_{ghc}^{OL}\hat{Y}_{gc}^{q-1} + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL}\hat{\bar{Y}}_{(-\mathcal{G}_h)c}^{q-1}\right]}\right\} \\ &\approx \frac{\sum_{c=1}^{\mathcal{C}}\left[\sum_{g\in\mathcal{G}_h} p_{ghc}^{OL}E\left(\hat{Y}_{gc}^{q}\right) + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL}E\left(\hat{\bar{Y}}_{(-\mathcal{G}_h)c}^{q}\right)\right]}{\sum_{c=1}^{\mathcal{C}}\left[\sum_{g\in\mathcal{G}_h} p_{ghc}^{OL}E\left(\hat{Y}_{gc}^{q-1}\right) + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL}E\left(\hat{\bar{Y}}_{(-\mathcal{G}_h)c}^{q-1}\right)\right]}. \end{aligned}$$

Now using that

$$E\left(\hat{Y}_{gc}^{q-u}\right) = Y_{gc}^{q-u} + B\left(\hat{Y}_{gc}^{q-u}\right)$$

and

$$E\left(\hat{\bar{Y}}_{(-\mathcal{G}_h)c}^{q-u}\right) = \frac{Y_{+c}^{q-u} - \sum_{g\in\mathcal{G}_h} E\left(\hat{Y}_{gc}^{q-u}\right)}{M-|\mathcal{G}_h|} = \frac{Y_{+c}^{q-u} - \sum_{g\in\mathcal{G}_h} Y_{gc}^{q-u} - \sum_{g\in\mathcal{G}_h} B\left(\hat{Y}_{gc}^{q-u}\right)}{M-|\mathcal{G}_h|} = \bar{Y}_{(-\mathcal{G}_h)c}^{q-u} - \frac{\sum_{g\in\mathcal{G}_h} B\left(\hat{Y}_{gc}^{q-u}\right)}{M-|\mathcal{G}_h|},$$

we obtain:

$$\begin{aligned} E\left(\hat{G}_h^{q,q-1}\right) &\approx \frac{\sum_{c=1}^{\mathcal{C}}\left[\sum_{g\in\mathcal{G}_h} p_{ghc}^{OL}Y_{gc}^{q} + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL}\bar{Y}_{(-\mathcal{G}_h)c}^{q} + \sum_{g\in\mathcal{G}_h}\left(p_{ghc}^{OL} - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}}{M-|\mathcal{G}_h|}\right)B\left(\hat{Y}_{gc}^{q}\right)\right]}{\sum_{c=1}^{\mathcal{C}}\left[\sum_{g\in\mathcal{G}_h} p_{ghc}^{OL}Y_{gc}^{q-1} + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL}\bar{Y}_{(-\mathcal{G}_h)c}^{q-1} + \sum_{g\in\mathcal{G}_h}\left(p_{ghc}^{OL} - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}}{M-|\mathcal{G}_h|}\right)B\left(\hat{Y}_{gc}^{q-1}\right)\right]} \\ &\approx \frac{E\left(\hat{Y}_h^{q}\right) + \sum_{c=1}^{\mathcal{C}}\left[\sum_{g\in\mathcal{G}_h}\left(p_{ghc}^{OL} - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}}{M-|\mathcal{G}_h|}\right)B\left(\hat{Y}_{gc}^{q}\right)\right]}{E\left(\hat{Y}_h^{q-1}\right) + \sum_{c=1}^{\mathcal{C}}\left[\sum_{g\in\mathcal{G}_h}\left(p_{ghc}^{OL} - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}}{M-|\mathcal{G}_h|}\right)B\left(\hat{Y}_{gc}^{q-1}\right)\right]} \\ &= \frac{E\left(\hat{Y}_h^{q}\right)\left\{1 + \sum_{c=1}^{\mathcal{C}}\left[\sum_{g\in\mathcal{G}_h}\left(p_{ghc}^{OL} - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}}{M-|\mathcal{G}_h|}\right)\frac{B\left(\hat{Y}_{gc}^{q}\right)}{E\left(\hat{Y}_h^{q}\right)}\right]\right\}}{E\left(\hat{Y}_h^{q-1}\right)\left\{1 + \sum_{c=1}^{\mathcal{C}}\left[\sum_{g\in\mathcal{G}_h}\left(p_{ghc}^{OL} - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}}{M-|\mathcal{G}_h|}\right)\frac{B\left(\hat{Y}_{gc}^{q-1}\right)}{E\left(\hat{Y}_h^{q-1}\right)}\right]\right\}} \\ &= \breve{G}_h^{q,q-1}\frac{1 + \sum_{c=1}^{\mathcal{C}}\left[\sum_{g\in\mathcal{G}_h}\left(p_{ghc}^{OL} - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}}{M-|\mathcal{G}_h|}\right)\frac{B\left(\hat{Y}_{gc}^{q}\right)}{E\left(\hat{Y}_h^{q}\right)}\right]}{1 + \sum_{c=1}^{\mathcal{C}}\left[\sum_{g\in\mathcal{G}_h}\left(p_{ghc}^{OL} - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}}{M-|\mathcal{G}_h|}\right)\frac{B\left(\hat{Y}_{gc}^{q-1}\right)}{E\left(\hat{Y}_h^{q-1}\right)}\right]}. \end{aligned}$$

Thus, if for all $g \in \mathcal{G}_h$ it holds that

$$\frac{B\left(\hat{Y}_{gc}^{q}\right)}{E\left(\hat{Y}_h^{q}\right)} \approx \frac{B\left(\hat{Y}_{gc}^{q-1}\right)}{E\left(\hat{Y}_h^{q-1}\right)}$$

then $E\left(\hat{G}_h^{q,q-1}\right) \approx \breve{G}_h^{q,q-1}$ and so $\hat{G}_h^{q,q-1}$ is an approximately unbiased estimator for $\breve{G}_h^{q,q-1}$. If the industries $g \in \mathcal{G}_h$ all have a similar quarterly growth rates, then this relation may be expected to hold. If these industries have widely different growth rates, then the above relation need not hold and the bias in $\hat{G}_h^{q,q-1}$ could be substantial.

In addition, the above bias and variance estimators may be affected by the fact that the observed residual $\hat{e}_{hi}^{q,q-1} = y_i^q - \hat{G}_h^{q,q-1} y_i^{q-1}$ depends on the estimated $\hat{G}_h^{q,q-1}$, and that the sums and sums of squares are computed over groups of units that may be misclassified. In theory, the resulting bias in the estimated bias and variance could be substantial. In practice, we believe that this bias may be limited because in the probability classes that contain the largest shares of turnover, p_{hhc}^{OL} is close to 1 for all target industries and the remaining p_{ghc}^{OL} ($g \neq h$) are small. In this case, $B\left(\hat{Y}_{hc}^{q-1}\right) \approx 0$ by (36).

4 Bias and variance approximations for year-on-year growth rates

4.1 Introduction

In this section, we will consider the case of a growth rate between a quarter and the corresponding quarter of the previous year ($u = 4$). The results in this section also apply to a growth rate between the first quarter of a year and the fourth quarter that directly precedes it (but in that case all instances of $q - 4$ should be read as $q - 1$). In both these cases, it does not necessarily hold that $s_i^q = s_i^{q-4}$ and $\hat{s}_i^q = \hat{s}_i^{q-4}$ for continuing units, and we therefore have to consider the additional effects of changes in true and observed strata.

Just as in Section 3, we begin by deriving explicit expressions for $AB\left(\hat{G}_h^{q,q-4}\right)$ and $AV\left(\hat{G}_h^{q,q-4}\right)$ based on (2) and (3), in Section 4.2. The resulting expressions (51) and (53) below are rather complicated. Next, we will therefore try to approximate these expressions. We begin by describing the simplified model for changes in observed classification errors that was used by van Delden et al. (2016a), in Section 4.3. In Section 4.4 we will introduce some additional assumptions on the classification error probabilities. The resulting bias and variance approximations are derived in Section 4.5. Estimation of the resulting expressions is discussed in Section 4.6.

4.2 Bias and variance for yearly growth rates

In the case of a growth rate that involves a yearly transition, expressions (4) and (5) from Section 3.2 still apply – with some trivial modifications – to the classification errors in $\hat{s}_i^{q-4}$. Expressions of the same form also apply to errors in $\hat{s}_i^q$ for units that do not occur in U^{q-4}. For continuing units $i \in U^{q-4,q}$, the quantities $E\left(\hat{\boldsymbol{a}}_i^q\right)$, $V\left(\hat{\boldsymbol{a}}_i^q\right)$ and $C\left(\hat{\boldsymbol{a}}_i^{q-4}, \hat{\boldsymbol{a}}_i^q\right)$ also depend on the effects of changes in classification errors between $q - 4$ and q.

The classification errors in $\hat{s}_i^q$ for continuing units are described by the so-called change matrix $\mathbf{P}_i^{OC} = (p_{jklhi}^{OC})$, with $p_{jklhi}^{OC} = P\left(\hat{s}_i^q = h \middle| s_i^{q-4} = j, s_i^q = k, \hat{s}_i^{q-4} = l\right)$. To write $\mathbf{P}_i^{OC}$ as a two-dimensional matrix, we let the rows denote all possible combinations of (j, k, l) and the columns denote all possible h; this makes $\mathbf{P}_i^{OC}$ a $M^3 \times M$ matrix. The $M \times M$ sub-matrix of $\mathbf{P}_i^{OC}$ that applies to units with $s_i^{q-4} = j$ and $s_i^q = k$ is denoted as $\widetilde{\mathbf{P}}_{i|jk}^{OC}$. [Note that the two remaining dimensions in this sub-matrix refer to $\hat{s}_i^{q-4} = l$ (rows) and $\hat{s}_i^q = h$ (columns).] The matrix $\mathbf{P}_i^{OC}$ consists of these blocks of rows, and we suppose that they are ordered lexicographically, so starting with $\widetilde{\mathbf{P}}_{i|11}^{OC}$, $\widetilde{\mathbf{P}}_{i|12}^{OC}$, etc., and ending with $\widetilde{\mathbf{P}}_{i|MM}^{OC}$.

Since the true strata s_i^{q-4} and s_i^q are considered fixed, it will be useful below to have a short-hand expression for the matrix that selects the $M \times M$ block $\widetilde{\mathbf{P}}_{i|jk}^{OC}$ from $\mathbf{P}_i^{OC}$ that actually applies to unit i, given the values of s_i^{q-4} and s_i^q. It is not difficult to see that this sub-matrix is given by

$$\boldsymbol{\Pi}_i^{OC} = \boldsymbol{\Lambda}_i \mathbf{P}_i^{OC}, \text{with } \boldsymbol{\Lambda}_i = \left(\boldsymbol{a}_i^{q-4} \otimes \boldsymbol{a}_i^q\right)^T \otimes \mathbf{I}_M, \tag{37}$$

where $\mathbf{I}_M$ denotes the $M \times M$ identity matrix and $\otimes$ denotes a Kronecker product. The operation of pre-multiplying $\mathbf{P}_i^{OC}$ by the $M \times M^3$ matrix $\mathbf{\Lambda}_i$ selects out the M rows of $\mathbf{P}_i^{OC}$ that correspond to (j,k,l) with $j = s_i^{q-4}$ and $k = s_i^q$, i.e., exactly those rows that apply to unit i. It follows from formula (37) that

$$\mathbf{\Pi}_i^{OC} = \sum\nolimits_{j=1}^{M} \sum\nolimits_{k=1}^{M} a_{ji}^{q-4} a_{ki}^{q} \widetilde{\mathbf{P}}_{i|jk}^{OC}. \tag{38}$$

The matrix $\mathbf{\Pi}_i^{OC} = \mathbf{\Lambda}_i \mathbf{P}_i^{OC}$ contains the relevant transition probabilities from $\hat{s}_i^{q-4}$ to $\hat{s}_i^q$, just like $\mathbf{P}_i^{OL}$ contains transition probabilities from s_i^{q-4} to $\hat{s}_i^{q-4}$. By analogy to (4), this implies that

$$\begin{aligned} E\big(\widehat{\boldsymbol{a}}_i^q \big| \hat{s}_i^{q-4}\big) &= (\mathbf{\Pi}_i^{OC})^T \widehat{\boldsymbol{a}}_i^{q-4}, \\ V\big(\widehat{\boldsymbol{a}}_i^q \big| \hat{s}_i^{q-4}\big) &= \mathrm{diag}\big[(\mathbf{\Pi}_i^{OC})^T \widehat{\boldsymbol{a}}_i^{q-4}\big] - (\mathbf{\Pi}_i^{OC})^T \,\mathrm{diag}(\widehat{\boldsymbol{a}}_i^{q-4})\, \mathbf{\Pi}_i^{OC}. \end{aligned} \tag{39}$$

Using standard expansion rules for conditional expectations and (co)variances, we obtain from (4) and (39):

$$E\big(\widehat{\boldsymbol{a}}_i^q\big) = E\big[E\big(\widehat{\boldsymbol{a}}_i^q \big| \hat{s}_i^{q-4}\big)\big] = (\mathbf{\Pi}_i^{OC})^T (\mathbf{P}_i^{OL})^T \boldsymbol{a}_i^{q-4} \tag{40}$$

and

$$\begin{aligned} V\big(\widehat{\boldsymbol{a}}_i^q\big) &= E\big[V\big(\widehat{\boldsymbol{a}}_i^q \big| \hat{s}_i^{q-4}\big)\big] + V\big[E\big(\widehat{\boldsymbol{a}}_i^q \big| \hat{s}_i^{q-4}\big)\big] \\ &= E\big\{\mathrm{diag}\big[(\mathbf{\Pi}_i^{OC})^T \widehat{\boldsymbol{a}}_i^{q-4}\big] - (\mathbf{\Pi}_i^{OC})^T \,\mathrm{diag}(\widehat{\boldsymbol{a}}_i^{q-4})\, \mathbf{\Pi}_i^{OC}\big\} + V\big[(\mathbf{\Pi}_i^{OC})^T \widehat{\boldsymbol{a}}_i^{q-4}\big] \\ &= \mathrm{diag}\big[(\mathbf{\Pi}_i^{OC})^T (\mathbf{P}_i^{OL})^T \boldsymbol{a}_i^{q-4}\big] - (\mathbf{\Pi}_i^{OC})^T \,\mathrm{diag}\big[(\mathbf{P}_i^{OL})^T \boldsymbol{a}_i^{q-4}\big]\, \mathbf{\Pi}_i^{OC} \\ &\qquad + (\mathbf{\Pi}_i^{OC})^T \big\{\mathrm{diag}\big[(\mathbf{P}_i^{OL})^T \boldsymbol{a}_i^{q-4}\big] - (\mathbf{P}_i^{OL})^T \,\mathrm{diag}(\boldsymbol{a}_i^{q-4})\, \mathbf{P}_i^{OL}\big\} \mathbf{\Pi}_i^{OC} \\ &= \mathrm{diag}\big[(\mathbf{\Pi}_i^{OC})^T (\mathbf{P}_i^{OL})^T \boldsymbol{a}_i^{q-4}\big] - (\mathbf{\Pi}_i^{OC})^T (\mathbf{P}_i^{OL})^T \,\mathrm{diag}(\boldsymbol{a}_i^{q-4})\, \mathbf{P}_i^{OL} \mathbf{\Pi}_i^{OC} \end{aligned} \tag{41}$$

and

$$\begin{aligned} C\big(\widehat{\boldsymbol{a}}_i^{q-4}, \widehat{\boldsymbol{a}}_i^q\big) &= E\big[C\big(\widehat{\boldsymbol{a}}_i^{q-4}, \widehat{\boldsymbol{a}}_i^q \big| \hat{s}_i^{q-4}\big)\big] + C\big[E\big(\widehat{\boldsymbol{a}}_i^{q-4} \big| \hat{s}_i^{q-4}\big), E\big(\widehat{\boldsymbol{a}}_i^q \big| \hat{s}_i^{q-4}\big)\big] \\ &= 0 + C\big[\widehat{\boldsymbol{a}}_i^{q-4}, (\mathbf{\Pi}_i^{OC})^T \widehat{\boldsymbol{a}}_i^{q-4}\big] \\ &= V\big(\widehat{\boldsymbol{a}}_i^{q-4}\big)\, \mathbf{\Pi}_i^{OC} \\ &= \big\{\mathrm{diag}\big[(\mathbf{P}_i^{OL})^T \boldsymbol{a}_i^{q-4}\big] - (\mathbf{P}_i^{OL})^T \,\mathrm{diag}\big(\boldsymbol{a}_i^{q-4}\big)\, \mathbf{P}_i^{OL}\big\} \mathbf{\Pi}_i^{OC}. \end{aligned} \tag{42}$$

Since $\boldsymbol{a}_i^{q-4}$ is a vector with one element equal to 1 and all other elements equal to 0, it holds that $\mathrm{diag}\big(\boldsymbol{a}_i^{q-4}\big) = \boldsymbol{a}_i^{q-4}\big(\boldsymbol{a}_i^{q-4}\big)^T$. For the derivation below, it is useful to re-write (41) and (42) as follows:

$$\begin{aligned} V\big(\widehat{\boldsymbol{a}}_i^q\big) &= \mathrm{diag}\big[(\mathbf{\Pi}_i^{OC})^T (\mathbf{P}_i^{OL})^T \boldsymbol{a}_i^{q-4}\big] - (\mathbf{\Pi}_i^{OC})^T (\mathbf{P}_i^{OL})^T \boldsymbol{a}_i^{q-4} \big(\boldsymbol{a}_i^{q-4}\big)^T \mathbf{P}_i^{OL} \mathbf{\Pi}_i^{OC}, \\ C\big(\widehat{\boldsymbol{a}}_i^{q-4}, \widehat{\boldsymbol{a}}_i^q\big) &= \mathrm{diag}\big[(\mathbf{P}_i^{OL})^T \boldsymbol{a}_i^{q-4}\big]\, \mathbf{\Pi}_i^{OC} - (\mathbf{P}_i^{OL})^T \boldsymbol{a}_i^{q-4} \big(\boldsymbol{a}_i^{q-4}\big)^T \mathbf{P}_i^{OL} \mathbf{\Pi}_i^{OC}. \end{aligned} \tag{43}$$

Thus, $E\big(\widehat{\boldsymbol{a}}_i^q\big)$ is equal to $(\mathbf{\Pi}_i^{OC})^T (\mathbf{P}_i^{OL})^T \boldsymbol{a}_i^{q-4}$ according to (40) and this quantity (as well as its transpose $\big(\boldsymbol{a}_i^{q-4}\big)^T \mathbf{P}_i^{OL} \mathbf{\Pi}_i^{OC}$) also occurs in $V\big(\widehat{\boldsymbol{a}}_i^q\big)$ and $C\big(\widehat{\boldsymbol{a}}_i^{q-4}, \widehat{\boldsymbol{a}}_i^q\big)$ according to (43). In fact, $(\mathbf{\Pi}_i^{OC})^T (\mathbf{P}_i^{OL})^T \boldsymbol{a}_i^{q-4}$ is a column vector of length M of which the elements are given by [cf. (38)]:

$$\begin{aligned} \big[(\mathbf{\Pi}_i^{OC})^T (\mathbf{P}_i^{OL})^T \boldsymbol{a}_i^{q-4}\big]_h &= \sum\nolimits_{g=1}^{M} a_{gi}^{q-4} (\mathbf{P}_i^{OL} \mathbf{\Pi}_i^{OC})_{gh} \\ &= \sum\nolimits_{g=1}^{M} a_{gi}^{q-4} \left[\sum\nolimits_{l=1}^{M} p_{gli}^{OL} \sum\nolimits_{j=1}^{M} \sum\nolimits_{k=1}^{M} a_{ji}^{q-4} a_{ki}^{q} p_{jklhi}^{OC}\right] \\ &= \sum\nolimits_{g=1}^{M} \sum\nolimits_{k=1}^{M} a_{gi}^{q-4} a_{ki}^{q} \sum\nolimits_{l=1}^{M} p_{gli}^{OL} p_{gklhi}^{OC}. \end{aligned} \tag{44}$$

In the last line, we used the fact that $a_{gi}^{q-4} a_{ji}^{q-4} = 0$ when $j \neq g$ and $\big(a_{gi}^{q-4}\big)^2 = a_{gi}^{q-4}$.

We define $\mathbb{p}^{OLC}_{gklhi} = p^{OL}_{gli}\, p^{OC}_{gklhi}$ and $p^{OLC}_{gkhi} = \sum_{l=1}^{M} \mathbb{p}^{OLC}_{gklhi}$. For continuing units, the quantity p^{OLC}_{gkhi} can be interpreted as a transition probability between the true industry code g in quarter $q-4$ and the observed industry code h in quarter q, given that the true industry code in quarter q is k. To see this, note first of all that

$$\begin{aligned}
\mathbb{p}^{OLC}_{gklhi} &= p^{OL}_{gli}\, p^{OC}_{gklhi} \\
&= P\left(\hat{s}_i^{q-4} = l \middle| s_i^{q-4} = g\right) P\left(\hat{s}_i^{q} = h \middle| s_i^{q-4} = g, s_i^{q} = k, \hat{s}_i^{q-4} = l\right) \\
&= P\left(\hat{s}_i^{q-4} = l \middle| s_i^{q-4} = g, s_i^{q} = k\right) P\left(\hat{s}_i^{q} = h \middle| s_i^{q-4} = g, s_i^{q} = k, \hat{s}_i^{q-4} = l\right) \\
&= P\left(\hat{s}_i^{q-4} = l, \hat{s}_i^{q} = h \middle| s_i^{q-4} = g, s_i^{q} = k\right).
\end{aligned}$$

In the third line, we used that $P\left(\hat{s}_i^{q-4} = h \middle| s_i^{q-4} = g\right) = P\left(\hat{s}_i^{q-4} = h \middle| s_i^{q-4} = g, s_i^{q} = k\right)$ since errors in $q-4$ are not supposed to be affected by true events in q. Thus, $\mathbb{p}^{OLC}_{gklhi}$ represents the joint probability of observing an industry code l in quarter $q-4$ and an industry code h in quarter q, given that the true industry code is g in quarter $q-4$ and k in quarter q. Hence, it follows that

$$p^{OLC}_{gkhi} = \sum_{l=1}^{M} \mathbb{p}^{OLC}_{gklhi} = P\left(\hat{s}_i^{q} = h \middle| s_i^{q-4} = g, s_i^{q} = k\right), \tag{45}$$

the probability of observing an industry code h in quarter q, given that the true industry code is g in quarter $q-4$ and k in quarter q. Note that, for continuing units, it must hold that $\sum_{h=1}^{M} p^{OLC}_{gkhi} = 1$.

It follows directly from (40) and (44) that

$$E\left(\hat{a}^{q}_{hi}\right) = \left[(\mathbf{\Pi}^{OC}_{i})^{T} (\mathbf{P}^{OL}_{i})^{T} \boldsymbol{a}_i^{q-4}\right]_h = \sum_{g=1}^{M} \sum_{k=1}^{M} a^{q-4}_{gi} a^{q}_{ki} p^{OLC}_{gkhi}. \tag{46}$$

Similarly, it follows from (43) and (44) that

$$V\left(\hat{a}^{q}_{hi}\right) = \sum_{g=1}^{M} \sum_{k=1}^{M} a^{q-4}_{gi} a^{q}_{ki} p^{OLC}_{gkhi} - \left(\sum_{g=1}^{M} \sum_{k=1}^{M} a^{q-4}_{gi} a^{q}_{ki} p^{OLC}_{gkhi}\right)^2.$$

The second term can be re-arranged as:

$$\left(\sum_{g=1}^{M} \sum_{k=1}^{M} a^{q-4}_{gi} a^{q}_{ki} p^{OLC}_{gkhi}\right)^2 = \sum_{g=1}^{M} \left(a^{q-4}_{gi} \sum_{k=1}^{M} a^{q}_{ki} p^{OLC}_{gkhi}\right)^2 = \sum_{g=1}^{M} \sum_{k=1}^{M} a^{q-4}_{gi} a^{q}_{ki} \left(p^{OLC}_{gkhi}\right)^2,$$

where we again used that $\boldsymbol{a}_i^{q-4}$ and $\boldsymbol{a}_i^{q}$ are vectors with one element equal to 1 and all other elements equal to 0. Thus, we find that

$$V\left(\hat{a}^{q}_{hi}\right) = \sum_{g=1}^{M} \sum_{k=1}^{M} a^{q-4}_{gi} a^{q}_{ki} p^{OLC}_{gkhi} \left(1 - p^{OLC}_{gkhi}\right). \tag{47}$$

Finally, for the covariance between $\hat{a}^{q-4}_{hi}$ and $\hat{a}^{q}_{hi}$, it follows from (43), (44) and (38) that

$$\begin{aligned}
C\left(\hat{a}^{q-4}_{hi}, \hat{a}^{q}_{hi}\right) &= \sum_{g=1}^{M} a^{q-4}_{gi} p^{OL}_{ghi} (\mathbf{\Pi}^{OC}_{i})_{hh} - \left(\sum_{g=1}^{M} a^{q-4}_{gi} p^{OL}_{ghi}\right)\left[\sum_{g=1}^{M} \sum_{k=1}^{M} a^{q-4}_{gi} a^{q}_{ki} p^{OLC}_{gkhi}\right] \\
&= \sum_{g=1}^{M} a^{q-4}_{gi} p^{OL}_{ghi} \sum_{k=1}^{M} a^{q}_{ki} p^{OC}_{gkhhi} - \sum_{g=1}^{M} a^{q-4}_{gi} p^{OL}_{ghi} \sum_{k=1}^{M} a^{q}_{ki} p^{OLC}_{gkhi} \\
&= \sum_{g=1}^{M} \sum_{k=1}^{M} a^{q-4}_{gi} a^{q}_{ki} \mathbb{p}^{OLC}_{gkhhi} - \sum_{g=1}^{M} \sum_{k=1}^{M} a^{q-4}_{gi} a^{q}_{ki} p^{OL}_{ghi} p^{OLC}_{gkhi},
\end{aligned}$$

or

$$C\left(\hat{a}^{q-4}_{hi}, \hat{a}^{q}_{hi}\right) = \sum_{g=1}^{M} \sum_{k=1}^{M} a^{q-4}_{gi} a^{q}_{ki} \left(\mathbb{p}^{OLC}_{gkhhi} - p^{OL}_{ghi} p^{OLC}_{gkhi}\right). \tag{48}$$

Note that, in the special case that $\hat{s}_i^{q-4}$ and $\hat{s}_i^q$ are independent conditional on the true industry codes in both quarters (i.e., independent classification errors across years), it holds that

$$\begin{aligned}\mathbb{p}_{gkhhi}^{OLC} &= P\left(\hat{s}_i^{q-4}=h,\hat{s}_i^q=h\middle|s_i^{q-4}=g,s_i^q=k\right)\\ &= P\left(\hat{s}_i^{q-4}=h\middle|s_i^{q-4}=g,s_i^q=k\right)P\left(\hat{s}_i^q=h\middle|s_i^{q-4}=g,s_i^q=k\right)\\ &= p_{ghi}^{OL}\,p_{gkhi}^{OLC}\end{aligned}$$

and hence by (48) that $C\left(\hat{a}_{hi}^{q-4},\hat{a}_{hi}^q\right)=0$. However, this special case does not arise in practice for industry codes in the Dutch GBR.

Using the expressions (46)–(48), we can again evaluate the different components in (2) and (3). For the bias, the following results are obtained analogously to (6):

$$\begin{aligned}E\left(\hat{Y}_h^{q-4}\right) &= \sum\nolimits_{i\in U^{q-4}}\left(y_i^{q-4}\sum\nolimits_{g=1}^{M}a_{gi}^{q-4}p_{ghi}^{OL}\right),\\ E\left(\hat{Y}_h^{q}\right) &= \sum\nolimits_{i\in U^{q-4,q}}\left(y_i^{q}\sum\nolimits_{g=1}^{M}\sum\nolimits_{k=1}^{M}a_{gi}^{q-4}a_{ki}^{q}p_{gkhi}^{OLC}\right)+\sum\nolimits_{i\in U^{q}\setminus U^{q-4,q}}\left(y_i^{q}\sum\nolimits_{g=1}^{M}a_{gi}^{q}p_{ghi}^{OL}\right),\\ \breve{G}_h^{q,q-4} &= \frac{E\left(\hat{Y}_h^{q}\right)}{E\left(\hat{Y}_h^{q-4}\right)}.\end{aligned} \tag{49}$$

Note that in the expression for $E\left(\hat{Y}_h^q\right)$, a distinction is made between continuing units and units that are new in quarter q. For the latter group of units, the classification errors in quarter q are described again by a level matrix. (Note: At this point, this can be done without loss of generalisation, because the probabilities p_{ghi}^{OL} are unit-specific, which leaves open the possibility that the level matrix for a new unit in quarter q differs from the level matrix in quarter $q-4$ for a continuing unit. However, we will later introduce the assumption that the level matrix does not vary between two subsequent years.)

In addition, we find from (6) and (48):

$$\begin{aligned}\sum\nolimits_{i\in U^{q-4}}\left(y_i^{q-4}\right)^2V\left(\hat{a}_{hi}^{q-4}\right) &= \sum\nolimits_{i\in U^{q-4}}\left[\left(y_i^{q-4}\right)^2\sum\nolimits_{g=1}^{M}a_{gi}^{q-4}p_{ghi}^{OL}\left(1-p_{ghi}^{OL}\right)\right],\\ \sum\nolimits_{i\in U^{q-4,q}}y_i^{q-4}y_i^{q}C\left(\hat{a}_{hi}^{q-4},\hat{a}_{hi}^{q}\right) &= \sum\nolimits_{i\in U^{q-4,q}}\left[G_i^{q,q-4}\left(y_i^{q-4}\right)^2\sum\nolimits_{g=1}^{M}\sum\nolimits_{k=1}^{M}a_{gi}^{q-4}a_{ki}^{q}\left(\mathbb{p}_{gkhhi}^{OLC}-p_{ghi}^{OL}p_{gkhi}^{OLC}\right)\right],\end{aligned} \tag{50}$$

where $G_i^{q,q-4}=y_i^q/y_i^{q-4}$ denotes an individual growth rate, as before. Thus, expression (2) becomes:

$$\begin{aligned}AB\left(\hat{G}_h^{q,q-4}\right) &= \frac{1}{\left[E\left(\hat{Y}_h^{q-4}\right)\right]^2}\left(B_{hO}^{q,q-4}+B_{hD}^{q,q-4}\right)+\left(\breve{G}_h^{q,q-4}-G_h^{q,q-4}\right),\\ B_{hO}^{q,q-4} &= \sum\nolimits_{i\in U^{q-4,q}}\left(y_i^{q-4}\right)^2\sum\nolimits_{g=1}^{M}a_{gi}^{q-4}\left[\breve{G}_h^{q,q-4}p_{ghi}^{OL}\left(1-p_{ghi}^{OL}\right)\right.\\ &\qquad\left.-G_i^{q,q-4}\sum\nolimits_{k=1}^{M}a_{ki}^{q}\left(\mathbb{p}_{gkhhi}^{OLC}-p_{ghi}^{OL}p_{gkhi}^{OLC}\right)\right],\\ B_{hD}^{q,q-4} &= \breve{G}_h^{q,q-4}\sum\nolimits_{i\in U^{q-4}\setminus U^{q-4,q}}\left[\left(y_i^{q-4}\right)^2\sum\nolimits_{g=1}^{M}a_{gi}^{q-4}p_{ghi}^{OL}\left(1-p_{ghi}^{OL}\right)\right].\end{aligned} \tag{51}$$

For the variance, we need the following expression in addition to (49) and (50), which follows from (47):

$$\sum_{i\in U^q}(y_i^q)^2 V(\hat{a}_{hi}^q)$$
$$= \sum_{i\in U^{q-4,q}}\left[\left(G_i^{q,q-4}y_i^{q-4}\right)^2\sum_{g=1}^{M}\sum_{k=1}^{M}a_{gi}^{q-4}a_{ki}^{q}p_{gkhi}^{OLC}\left(1-p_{gkhi}^{OLC}\right)\right] \quad (52)$$
$$+\sum_{i\in U^q\setminus U^{q-4,q}}\left[\left(y_i^q\right)^2\sum_{g=1}^{M}a_{gi}^{q}p_{ghi}^{OL}\left(1-p_{ghi}^{OL}\right)\right].$$

In (52), again, a distinction is made between continuing units and new-born units. New-born units in quarter q are treated the same way as in Section 3.2. Expression (3) now yields:

$$AV\left(\hat{G}_h^{q,q-4}\right) = \frac{1}{\left[E\left(\hat{Y}_h^{q-4}\right)\right]^2}\left(V_{hO}^{q,q-4}+V_{hD}^{q,q-4}+V_{hB}^{q,q-4}\right),$$
$$V_{hO}^{q,q-4} = \sum_{i\in U^{q-4,q}}\left(y_i^{q-4}\right)^2\sum_{g=1}^{M}a_{gi}^{q-4}\left\{\left(\breve{G}_h^{q,q-4}\right)^2 p_{ghi}^{OL}\left(1-p_{ghi}^{OL}\right)\right.$$
$$\left.+\sum_{k=1}^{M}a_{ki}^{q}\left[\left(G_i^{q,q-4}\right)^2 p_{gkhi}^{OLC}\left(1-p_{gkhi}^{OLC}\right)-2\breve{G}_h^{q,q-4}G_i^{q,q-4}\left(\mathbb{p}_{gkhhi}^{OLC}-p_{ghi}^{OL}p_{gkhi}^{OLC}\right)\right]\right\}, \quad (53)$$
$$V_{hD}^{q,q-4} = \left(\breve{G}_h^{q,q-4}\right)^2\sum_{i\in U^{q-4}\setminus U^{q-4,q}}\left[\left(y_i^{q-4}\right)^2\sum_{g=1}^{M}a_{gi}^{q-4}p_{ghi}^{OL}\left(1-p_{ghi}^{OL}\right)\right],$$
$$V_{hB}^{q,q-4} = \sum_{i\in U^q\setminus U^{q-4,q}}\left[\left(y_i^q\right)^2\sum_{g=1}^{M}a_{gi}^{q}p_{ghi}^{OL}\left(1-p_{ghi}^{OL}\right)\right].$$

Thus, to evaluate the bias and variance of a growth rate that involves a yearly transition we need to estimate two types of probabilities. Firstly we need the level-probabilities p_{ghi}^{OL} that are also involved in the bias and variance of a turnover level or a quarterly growth rate within a year. Secondly we need the transition probabilities p_{gkhi}^{OLC} as well as the underlying 'diagonal' probabilities $\mathbb{p}_{gkhhi}^{OLC}$. (Recall that $p_{gkhi}^{OLC}=\sum_{l=1}^{M}\mathbb{p}_{gklhi}^{OLC}$.) Estimates for these additional probabilities can either be derived from estimated level and change matrices, or they can be estimated directly. We will return to this issue in Section 4.6.

4.3 An explicit expression for $\boldsymbol{\Pi}_i^{OC}$

So far, we have not placed any restrictions on the elements of the change matrix $\mathbf{P}_i^{OC}=(p_{jklhi}^{OC})$. In particular, the probabilities p_{jklhi}^{OC} might differ per unit i. In this and the next subsection, we will introduce rather restrictive assumptions on these probabilities (as well as the derived probabilities $\mathbb{p}_{gkhhi}^{OLC}$ and p_{gkhi}^{OLC}) to reduce the number of parameters. This is important to obtain a classification error model that is estimable in practice.

van Delden et al. (2016a) introduced a model for $\mathbf{P}_i^{OC}$ that describes the changes in classification errors between quarters $q-4$ and q in terms of three key parameters: p_R, p_N and p_S. Here, p_R denotes the probability that an existing classification error (i.e., an error that was already present in $q-4$) is corrected (R = restore); p_N denotes the probability that a true change in classification between $q-4$ and q is correctly implemented in the observed industry code (N = notice); and p_S denotes the probability of a change in the observed industry code between $q-4$ and q that cannot be explained from the underlying true industry codes in either period (S = spurious). The model also makes use of transition fractions between observed industry codes in the GBR (see below).

For the model $p_{jklhi}^{OC}=P\left(\hat{s}_i^q=h\middle|s_i^{q-4}=j,s_i^q=k,\hat{s}_i^{q-4}=l\right)$ defined in Table 2 of van Delden et al. (2016a), the matrix $\boldsymbol{\Pi}_i^{OC}$ that was introduced in Section 4.2 is given by:

$$\boldsymbol{\Pi}_i^{OC}=\boldsymbol{\Pi}^{OC}\left(\boldsymbol{a}_i^{q-4},\boldsymbol{a}_i^{q}\right) \quad (54)$$

$$
\begin{aligned}
&= \left(\boldsymbol{a}_i^{q-4}\right)^T \boldsymbol{a}_i^q \{\mathrm{diag}(\boldsymbol{a}_i^q)\,\mathbf{A} + [\mathbf{I}_M - \mathrm{diag}(\boldsymbol{a}_i^q)]\mathbf{B}(\boldsymbol{a}_i^q)\} \\
&\qquad + \left[1 - \left(\boldsymbol{a}_i^{q-4}\right)^T \boldsymbol{a}_i^q\right] \{\mathrm{diag}(\boldsymbol{a}_i^q)\,\mathbf{A} + [\mathbf{I}_M - \mathrm{diag}(\boldsymbol{a}_i^q)]\mathbf{C}(\boldsymbol{a}_i^q)\} \\
&= \mathrm{diag}(\boldsymbol{a}_i^q)\,\mathbf{A} + [\mathbf{I}_M - \mathrm{diag}(\boldsymbol{a}_i^q)] \left\{\left[\left(\boldsymbol{a}_i^{q-4}\right)^T \boldsymbol{a}_i^q\right] \mathbf{B}(\boldsymbol{a}_i^q) + \left[1 - \left(\boldsymbol{a}_i^{q-4}\right)^T \boldsymbol{a}_i^q\right] \mathbf{C}(\boldsymbol{a}_i^q)\right\},
\end{aligned}
$$

with

$$
\begin{aligned}
\mathbf{A} &= \mathbf{I}_M + p_S(\mathbf{R} - \mathbf{I}_M), \\
\mathbf{B}(\boldsymbol{a}_i^q) &= \frac{1}{1 - p_R p_S}\left[(1 - p_R)(1 - p_S)\mathbf{I}_M + (1 - p_R)p_S\mathbf{R} + p_R(1 - p_S)\mathbf{1}_M\left(\boldsymbol{a}_i^q\right)^T\right], \\
\mathbf{C}(\boldsymbol{a}_i^q) &= \frac{1}{1 - p_N p_S}\left[(1 - p_N)(1 - p_S)\mathbf{I}_M + (1 - p_N)p_S\mathbf{R} + p_N(1 - p_S)\mathbf{1}_M\left(\boldsymbol{a}_i^q\right)^T\right],
\end{aligned}
\tag{55}
$$

where furthermore $\mathbf{R}$ is an $M \times M$ matrix which contains the observed transition fractions ρ_{jk} from the GBR, with $\rho_{jj} = 0$ on the diagonal, and $\mathbf{1}_M$ is an M vector with all elements equal to 1.

Note that the inner product $\left(\boldsymbol{a}_i^{q-4}\right)^T \boldsymbol{a}_i^q$ is equal to 1 if $s_i^{q-4} = s_i^q$ and equal to 0 otherwise. The above expression therefore follows from the description in van Delden et al. (2016a) by observing that $\mathrm{diag}(\boldsymbol{a}_i^q)\,\mathbf{A} + [\mathbf{I}_M - \mathrm{diag}(\boldsymbol{a}_i^q)]\mathbf{B}(\boldsymbol{a}_i^q)$ is the form of $\boldsymbol{\Pi}_i^{OC}$ when $s_i^{q-4} = s_i^q$ and $\mathrm{diag}(\boldsymbol{a}_i^q)\,\mathbf{A} + [\mathbf{I}_M - \mathrm{diag}(\boldsymbol{a}_i^q)]\mathbf{C}(\boldsymbol{a}_i^q)$ is the form of $\boldsymbol{\Pi}_i^{OC}$ when $s_i^{q-4} \neq s_i^q$. It may also be noted that the matrix $\mathbf{A}$ contains the probabilities that hold (for any unit) under Situations A and D in van Delden et al. (2016a), and the matrices $\mathbf{B}(\boldsymbol{a}_i^q)$ and $\mathbf{C}(\boldsymbol{a}_i^q)$ contain the probabilities that hold for unit i under Situations B and C, respectively.

The notation $\boldsymbol{\Pi}_i^{OC} = \boldsymbol{\Pi}^{OC}(\boldsymbol{a}_i^{q-4}, \boldsymbol{a}_i^q)$ in (54) highlights that the change matrix depends on i only as a function of $\boldsymbol{a}_i^{q-4}$ and $\boldsymbol{a}_i^q$. In other words, units with the same values of s_i^{q-4} and s_i^q have the same probabilities p_{jklhi}^{OC}. Since $s_i^{q-4} = j$ and $s_i^q = k$ are already included as indices in the notation of these probabilities, we can remove the index i and write $p_{jklhi}^{OC} = p_{jklh}^{OC}$.

The above model assumes a single set of parameters (p_R, p_N, p_S) that applies to all units in the population. A more realistic model is obtained by dividing the population into probability classes $U_1^{q-u}, \ldots, U_C^{q-u}$ (with $u \in \{0,1,4\}$) such that all units $i \in U_c^{q-u}$ have the same parameters (p_{Rc}, p_{Nc}, p_{Sc}) and hence the same probabilities $p_{jklhi}^{OC} = p_{jklhc}^{OC}$. van Delden et al. (2016a) considered three such homogeneous probability classes, consisting of Simple, Complex and Most complex units, respectively. In this paper, we have already introduced probability classes for the probabilities in the level matrix in Section 3.4. For simplicity of notation, we assume that the *same* division of units into probability classes applies to both the level and change matrices.

In Section 3, we could assume that continuing units remain in the same probability class between $q-1$ and q within the same year. The same assumption cannot be made as easily for continuing units between $q-4$ and q. In practice, there will be some units in the GBR that change between probability classes at the yearly transition. It is not desirable to introduce separate terms in the expressions below to handle all possible changes between probability classes. We will therefore assume here that all continuing units can be assigned to a single probability class for both periods $q-4$ and q. We denote these classes as $U_1^{q-4,q}, \ldots, U_C^{q-4,q}$.

In practical applications, a procedure is needed to handle units that change between probability classes. Two simple, approximate solutions are:

- assigning each unit that changes between $q-4$ and q to the probability class $U_c^{q-4,q}$ that has the least favourable classification error probabilities;
- assigning each unit that changes between $q-4$ and q to one of the two probability classes at random.

With the first solution, the uncertainty due to classification errors is more likely to be overestimated than underestimated, which may be advantageous for some applications. However, in practice it may sometimes be unclear which of the two probability classes contains the least favourable probabilities. The second solution has the advantage that it is possible to test the robustness of the estimated uncertainty to this approximation by repeating the random assignment multiple times and comparing the resulting bias and variance estimates. We will therefore use the second solution in our application.

4.4 Further approximation of the probabilities

The bias and variance approximations derived in Section 4.2 involve sums with a separate contribution for each unit in the population. In Section 3, we were able to reduce similar expressions for the case of a within-year growth rate to expressions that involve separate contributions for a limited number of groups of units, by introducing assumptions about the structure of the level matrix $\mathbf{P}_i^{OL}$. In this section, we would like to do the same for the bias and variance of a yearly growth rate. This requires assumptions about the structure of both the level matrix $\mathbf{P}_i^{OL}$ and the change matrix $\mathbf{P}_i^{OC}$. Note that these assumptions do not simplify the mathematical form of these expressions *per se* but they do reduce the computations that are needed to evaluate the bias and the variance in practice.

We have already introduced the assumption that all probabilities are constant within each probability class, so that $\mathbf{P}_i^{OL} = \mathbf{P}_c^{OL}$ and $\mathbf{P}_i^{OC} = \mathbf{P}_c^{OC}$. From this, it also follows that $\mathbb{p}_{gkhhi}^{OLC} = \mathbb{p}_{gkhhc}^{OLC} = p_{ghc}^{OL} p_{gkhhc}^{OC}$ and $p_{gkhi}^{OLC} = p_{gkhc}^{OLC} = \sum_{l=1}^{M} p_{glc}^{OL} p_{gklhc}^{OC}$ for all units in probability class c. For the probabilities p_{ghc}^{OL}, we can apply assumption 2'' from Section 3.4 to reduce the number of parameters within each probability class. We will now introduce a similar assumption for p_{gkhc}^{OLC} and $\mathbb{p}_{gkhhc}^{OLC}$. [Note that this is sufficient, because the probabilities in $\mathbf{P}_c^{OC}$ themselves do not occur separately in expressions (51) and (53).]

The probabilities p_{gkhc}^{OLC} and $\mathbb{p}_{gkhhc}^{OLC}$ refer to a stratification of the units in each probability class in terms of $\left(s_i^{q-4}, s_i^q, \hat{s}_i^q\right) = (g, k, h)$. Potentially, this concerns M^3 strata for each probability class, i.e., an extremely large number. (Recall that, for the Dutch GBR, $M \approx 300$.) Moreover, most of these strata contain only a limited number of units; in fact many of them may be empty. The approximations that follow are therefore important to increase the stability of the bias and variance estimates, to reduce the amount of computational work, and to improve the interpretability of the probabilities.

As we will be evaluating the bias and variance of a growth rate for a given observed industry $\hat{s}_i^q = h$, we consider h fixed in what follows. For a given h, we consider four types of combinations of $\left(s_i^{q-4}, s_i^q, \hat{s}_i^q\right)$:

- <u>diagonal cases</u>: $\left(s_i^{q-4}, s_i^q, \hat{s}_i^q\right) = (g, g, h)$, for $g \in \{1, \dots, M\}$;
  (these are cases where the true industry code does not change)
- <u>cases within column h</u>: $\left(s_i^{q-4}, s_i^q, \hat{s}_i^q\right) = (g, h, h)$ with $g \neq h$;
  (these are cases where the true industry code changes and the observed code in q is correct)
- <u>cases within row h</u>: $\left(s_i^{q-4}, s_i^q, \hat{s}_i^q\right) = (h, k, h)$ with $k \neq h$;
  (these are cases where the true industry code changes and there is an error in the observed code in q, but the same code would have been correct in $q-4$)
- <u>all other cases</u>: $\left(s_i^{q-4}, s_i^q, \hat{s}_i^q\right) = (g, k, h)$ with $g \neq h$ and $k \neq h$ and $g \neq k$.
  (these are cases where the true industry code changes and there is an error in the observed code in q, and the same code would also have been erroneous in $q-4$)

Within each type, we can distinguish *special* combinations, i.e., combinations of $\left(s_i^{q-4}, s_i^q, \hat{s}_i^q\right)$ that occur relatively often. For these special combinations, we will reserve separate terms in the bias and variance

approximations. For all other combinations, we will approximate the actual probabilities p_{gkhc}^{OLC} and $\mathbb{p}_{gkhhc}^{OLC}$ by average values. This is similar to assumption 2'' for the level matrix.

In Section 3.4, we introduced the subsets $\mathcal{H}_g \subset \{1, \dots, M\}$ of industry codes that are observed relatively often when the true industry code is g. We will re-use these subsets here. In addition, we introduce the subsets $\mathcal{K}_g \subset \{1, \dots, M\}$ of true industry codes that occur relatively often in quarter q for units with true industry code g in quarter $q-4$, given that the industry code changes between $q-4$ and q. Note that the latter condition implies that $g \notin \mathcal{K}_g$.

We now define the special combinations for each of the four above-mentioned types as follows:

- <u>diagonal cases</u>: $\left(s_i^{q-4}, s_i^q, \hat{s}_i^q\right) = (g, g, h)$ is special when $h \in \mathcal{H}_g$;
  (the combination of true and observed codes in q has to occur relatively often)
- <u>cases within column h</u>: $\left(s_i^{q-4}, s_i^q, \hat{s}_i^q\right) = (g, h, h)$ with $g \neq h$ is special when $h \in \mathcal{K}_g$;
  (the combination of true codes in $q-4$ and q has to occur relatively often)
- <u>cases within row h</u>: $\left(s_i^{q-4}, s_i^q, \hat{s}_i^q\right) = (h, k, h)$ with $k \neq h$ is special when $k \in \mathcal{K}_h$;
  (the combination of true codes in $q-4$ and q has to occur relatively often)
- <u>all other cases</u>: $\left(s_i^{q-4}, s_i^q, \hat{s}_i^q\right) = (g, k, h)$ with $g \neq h$ and $k \neq h$ and $g \neq k$ is special when both $k \in \mathcal{K}_g$ and $h \in \mathcal{H}_k$.
  (the combination of true codes in $q-4$ and q, as well as the combination of true and observed codes in q both have to occur relatively often)

All combinations that do not satisfy the relevant condition do not count as special combinations.

For a given h, we define $I_{gkh} = 1$ if (g, k, h) satisfies the relevant condition to count as a special combination, and $I_{gkh} = 0$ otherwise. Next, we define $\mathcal{T}_h = \{(g, k) : I_{gkh} = 1\}$, the subset of all pairs (g, k) that constitute special combinations for the observed stratum h. Note that this is similar to the way we defined $\mathcal{G}_h$ for the level matrix in Section 3.4. Again, we suppose that the number of special combinations $|\mathcal{T}_h|$ is quite small in comparison to the total number of pairs M^2.

In the bias and variance formulas below, we will encounter probabilities p_{gkhi}^{OLC} and $\mathbb{p}_{gkhhi}^{OLC}$ for a fixed observed industry code $\hat{s}_i^q = h$. To simplify the computation of these formulas, we propose to approximate these probabilities as follows. For all pairs $(g, k) \in \mathcal{T}_h$, the probabilities p_{gkhc}^{OLC} and $\mathbb{p}_{gkhhc}^{OLC}$ are derived from the expressions in Section 4.3, when unit i is in probability class c. For the remaining pairs $(g, k) \notin \mathcal{T}_h$, where these probabilities are supposed to be small, we compute the total remaining probability:

$$\begin{aligned} \tilde{p}_{(-\mathcal{T}_h)hc}^{OLC} &= p_{++hc}^{OLC} - \sum_{(g,k)\in\mathcal{T}_h} p_{gkhc}^{OLC} = \sum_{g=1}^{M}\sum_{k=1}^{M} p_{gkhc}^{OLC} - \sum_{(g,k)\in\mathcal{T}_h} p_{gkhc}^{OLC}, \\ \tilde{\mathbb{p}}_{(-\mathcal{T}_h)hhc}^{OLC} &= \mathbb{p}_{++hhc}^{OLC} - \sum_{(g,k)\in\mathcal{T}_h} \mathbb{p}_{gkhhc}^{OLC} = \sum_{g=1}^{M}\sum_{k=1}^{M} \mathbb{p}_{gkhhc}^{OLC} - \sum_{(g,k)\in\mathcal{T}_h} \mathbb{p}_{gkhhc}^{OLC}. \end{aligned} \tag{56}$$

We then divide this total remaining probability mass uniformly over the remaining pairs, to obtain an approximate probability parameter:

$$\tilde{p}_{gkhc}^{OLC} = \begin{cases} p_{gkhc}^{OLC} & \text{if } (g,k) \in \mathcal{T}_h \\ \dfrac{\tilde{p}_{(-\mathcal{T}_h)hc}^{OLC}}{M^2 - |\mathcal{T}_h|} & \text{if } (g,k) \notin \mathcal{T}_h \end{cases} \tag{57}$$

$$\widetilde{\mathbb{p}}^{OLC}_{gkhhc} = \begin{cases} \mathbb{p}^{OLC}_{gkhhc} & \text{if } (g,k) \in \mathcal{T}_h \\ \dfrac{\widetilde{\mathbb{p}}^{OLC}_{(-\mathcal{T}_h)hhc}}{M^2 - |\mathcal{T}_h|} & \text{if } (g,k) \notin \mathcal{T}_h \end{cases}$$

Note that for all $(g,k) \notin \mathcal{T}_h$ it holds that:

$$\widetilde{\mathbb{p}}^{OLC}_{gkhhc} = \frac{\widetilde{\mathbb{p}}^{OLC}_{(-\mathcal{T}_h)hhc}}{M^2 - |\mathcal{T}_h|} = \frac{\sum_{(g,k)\notin\mathcal{T}_h} \mathbb{p}^{OLC}_{gkhhc}}{M^2 - |\mathcal{T}_h|} \leq \frac{\sum_{(g,k)\notin\mathcal{T}_h} p^{OLC}_{gkhc}}{M^2 - |\mathcal{T}_h|} = \frac{\tilde{p}^{OLC}_{(-\mathcal{T}_h)hc}}{M^2 - |\mathcal{T}_h|} = \tilde{p}^{OLC}_{gkhc};$$

in other words, the two approximations are in line with each other [cf. (45)].

We thus obtain the following additional assumption for continuing units between $q-4$ and q:

3. The population of continuing units consists of probability classes $U_1^{q-4,q}, \dots, U_C^{q-4,q}$. All units in probability class $U_c^{q-4,q}$ have the same probabilities p^{OLC}_{gkhc} and $\mathbb{p}^{OLC}_{gkhhc}$ that can be approximated well by $\tilde{p}^{OLC}_{gkhc}$ and $\widetilde{\mathbb{p}}^{OLC}_{gkhhc}$ in (57).

4.5 Approximate bias and variance formulae

4.5.1 Preliminary results

We begin by deriving approximations to $E\left(\hat{Y}_h^{q-4}\right)$, $E\left(\hat{Y}_h^{q}\right)$ and $\breve{G}_h^{q,q-4} = E\left(\hat{Y}_h^{q}\right)/E\left(\hat{Y}_h^{q-4}\right)$ under assumptions 2'' and 3, based on the exact expressions in (49). The expression for $E\left(\hat{Y}_h^{q-4}\right)$ involves only probabilities from the level matrix. We can therefore proceed analogously to the derivation for $E\left(\hat{Y}_h^{q-1}\right)$ in (25) to obtain:

$$E\left(\hat{Y}_h^{q-4}\right) \approx \sum_{c=1}^{\mathcal{C}} \left[\sum_{g\in\mathcal{G}_h} p^{OL}_{ghc} Y^{q-4}_{gc} + \tilde{p}^{OL}_{(-\mathcal{G}_h)hc} \bar{Y}^{q-4}_{(-\mathcal{G}_h)c} \right]. \tag{58}$$

Here and elsewhere in this section, we will re-use the notation that was introduced in Section 3 for quarterly growth rates, with minor variations if necessary.

For $E\left(\hat{Y}_h^{q}\right)$, the expression in (49) consists of two terms, the first of which is an expectation over continuing units and the second an expectation over new-born units:

$$E\left(\hat{Y}_h^{q}\right) = E\left(\hat{Y}_{hO}^{q}\right) + E\left(\hat{Y}_{hB}^{q}\right).$$

The term $E\left(\hat{Y}_{hB}^{q}\right)$ again involves only the level matrix, and it follows analogously to (58) that

$$E\left(\hat{Y}_{hB}^{q}\right) \approx \sum_{c=1}^{\mathcal{C}} \left[\sum_{g\in\mathcal{G}_h} p^{OL}_{ghc} Y^{q}_{gcB} + \tilde{p}^{OL}_{(-\mathcal{G}_h)hc} \bar{Y}^{q}_{(-\mathcal{G}_h)cB} \right], \tag{59}$$

in obvious notation.

The term $E\left(\hat{Y}_{hO}^{q}\right)$ involves probabilities p^{OLC}_{gkhi}. Using assumption 3, we find:

$$
\begin{aligned}
E\left(\hat{Y}_{hO}^{q}\right) &= \sum_{c=1}^{\mathcal{C}} \sum\nolimits_{i \in U_c^{q-4,q}} y_i^q \sum\nolimits_{g=1}^{M} \sum\nolimits_{k=1}^{M} a_{gi}^{q-4} a_{ki}^{q} p_{gkhc}^{OLC} \\
&\approx \sum_{c=1}^{\mathcal{C}} \sum\nolimits_{i \in U_c^{q-4,q}} y_i^q \left[\sum_{(g,k) \in \mathcal{T}_h} a_{gi}^{q-4} a_{ki}^{q} p_{gkhc}^{OLC} + \left(1 - \sum_{(g,k) \in \mathcal{T}_h} a_{gi}^{q-4} a_{ki}^{q} \right) \frac{\tilde{p}_{(-\mathcal{T}_h)hc}^{OLC}}{M^2 - |\mathcal{T}_h|} \right] \\
&= \sum_{c=1}^{\mathcal{C}} \left[\sum_{(g,k) \in \mathcal{T}_h} p_{gkhc}^{OLC} Y_{gkcO}^{q} + \frac{\tilde{p}_{(-\mathcal{T}_h)hc}^{OLC}}{M^2 - |\mathcal{T}_h|} \left(Y_{++cO}^{q} - \sum_{(g,k) \in \mathcal{T}_h} Y_{gkcO}^{q} \right) \right] \\
&= \sum_{c=1}^{\mathcal{C}} \left[\sum_{(g,k) \in \mathcal{T}_h} p_{gkhc}^{OLC} Y_{gkcO}^{q} + \tilde{p}_{(-\mathcal{T}_h)hc}^{OLC} \bar{Y}_{(-\mathcal{T}_h)cO}^{q} \right].
\end{aligned}
\tag{60}
$$

In the third line, $Y_{gkcO}^{q} = \sum_{i \in U_c^{q-4,q}} a_{gi}^{q-4} a_{ki}^{q} y_i^q$ denotes the total quarterly turnover in q for all continuing units in probability class c that belong to stratum g in quarter $q-4$ and to stratum k in quarter q. These turnovers are computed only for the combinations (g,k) that are listed in $\mathcal{T}_h$. In the last line,

$$
\bar{Y}_{(-\mathcal{T}_h)cO}^{q} = \frac{Y_{++cO}^{q} - \sum_{(g,k) \in \mathcal{T}_h} Y_{gkcO}^{q}}{M^2 - |\mathcal{T}_h|}
$$

denotes the average value of the total quarterly turnover Y_{gkcO}^{q} across all subsets of continuing units in probability class c for combinations (g,k) that are not listed in $\mathcal{T}_h$. As noted before, there are $M^2 - |\mathcal{T}_h|$ such combinations.

Combining (59) and (60), we obtain:

$$
E\left(\hat{Y}_{h}^{q}\right) \approx \sum_{c=1}^{\mathcal{C}} \left[\sum_{(g,k) \in \mathcal{T}_h} p_{gkhc}^{OLC} Y_{gkcO}^{q} + \tilde{p}_{(-\mathcal{T}_h)hc}^{OLC} \bar{Y}_{(-\mathcal{T}_h)cO}^{q} + \sum_{g \in \mathcal{G}_h} p_{ghc}^{OL} Y_{gcB}^{q} + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL} \bar{Y}_{(-\mathcal{G}_h)cB}^{q} \right]. \tag{61}
$$

Furthermore, it follows from (58) and (61) that

$$
\breve{G}_{h}^{q,q-4} \approx \frac{\sum_{c=1}^{\mathcal{C}} \left[\sum_{(g,k) \in \mathcal{T}_h} p_{gkhc}^{OLC} Y_{gkcO}^{q} + \tilde{p}_{(-\mathcal{T}_h)hc}^{OLC} \bar{Y}_{(-\mathcal{T}_h)cO}^{q} + \sum_{g \in \mathcal{G}_h} p_{ghc}^{OL} Y_{gcB}^{q} + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL} \bar{Y}_{(-\mathcal{G}_h)cB}^{q} \right]}{\sum_{c=1}^{\mathcal{C}} \left[\sum_{g \in \mathcal{G}_h} p_{ghc}^{OL} Y_{gc}^{q-4} + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL} \bar{Y}_{(-\mathcal{G}_h)c}^{q-4} \right]}. \tag{62}
$$

4.5.2 Bias

The bias approximation in formula (51) consists of three components:

$$
AB\left(\hat{G}_{h}^{q,q-4}\right) = \frac{1}{\left[E\left(\hat{Y}_{h}^{q-4}\right)\right]^2} \left(B_{hO}^{q,q-4} + B_{hD}^{q,q-4}\right) + \left(\breve{G}_{h}^{q,q-4} - G_{h}^{q,q-4}\right).
$$

An approximation for the term $\breve{G}_{h}^{q,q-4} - G_{h}^{q,q-4}$ follows directly from (62). The component $B_{hO}^{q,q-4}$ refers to continuing units ($U^{q-4,q}$). In the derivation of $B\left(\hat{G}_{h}^{q,q-1}\right)$, it was possible to write the corresponding component $B_{hO}^{q,q-1}$ as a single expression in terms of a pseudo-residual. For $B_{hO}^{q,q-4}$, this is unfortunately not possible, and we have to consider two separate sub-terms:

$$
\begin{aligned}
B_{hO}^{q,q-4} &= B_{hO1}^{q,q-4} + B_{hO2}^{q,q-4}, \\
B_{hO1}^{q,q-4} &= \breve{G}_{h}^{q,q-4} \sum\nolimits_{i \in U^{q-4,q}} \left(y_i^{q-4}\right)^2 \sum\nolimits_{g=1}^{M} a_{gi}^{q-4} p_{ghi}^{OL} \left(1 - p_{ghi}^{OL}\right),
\end{aligned}
\tag{63}
$$

$$B_{hO2}^{q,q-4} = -\sum\nolimits_{i\in U^{q-4,q}} y_i^{q-4} y_i^q \sum\nolimits_{g=1}^{M} \sum\nolimits_{k=1}^{M} a_{gi}^{q-4} a_{ki}^{q} \left(\mathbb{p}_{gkhhi}^{OLC} - p_{ghi}^{OL} p_{gkhi}^{OLC}\right).$$

The first sub-term only involves the level matrix, and we can proceed in a similar fashion as for $E\left(\hat{Y}_h^{q-4}\right)$ in (58) to obtain:

$$B_{hO1}^{q,q-4} \approx \breve{G}_h^{q,q-4} \sum_{c=1}^{\mathcal{C}} \left[\sum_{g\in\mathcal{G}_h} p_{ghc}^{OL}\left(1 - p_{ghc}^{OL}\right) SS_{gcO}\left(y^{q-4}\right) + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL} \left(1 - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}}{M - |\mathcal{G}_h|}\right) \overline{SS}_{(-\mathcal{G}_h)cO}\left(y^{q-4}\right) \right].$$

Here, $SS_{gcO}\left(y^{q-4}\right)$ and $\overline{SS}_{(-\mathcal{G}_h)cO}\left(y^{q-4}\right)$ denote sums of squares as defined previously in Section 3.5.

The second sub-term $B_{hO2}^{q,q-4}$ in (63) is slightly more complicated, because it involves on the one hand probabilities p_{ghi}^{OL} and on the other hand probabilities p_{gkhi}^{OLC} and $\mathbb{p}_{gkhhi}^{OLC}$. For p_{ghi}^{OL}, we have defined a partition of the industries into $\mathcal{G}_h$ and its complement, while for the other two probabilities we have defined a partition of pairs of industries into $\mathcal{I}_h$ and its complement. Here, both partitions are relevant. We therefore need to define a more elaborate partition of pairs of industries into four subsets:

$$\begin{aligned}
\mathcal{A}_{11h} &= \{(g,k): g \in \mathcal{G}_h, (g,k) \in \mathcal{I}_h\};\\
\mathcal{A}_{10h} &= \{(g,k): g \in \mathcal{G}_h, (g,k) \notin \mathcal{I}_h\};\\
\mathcal{A}_{01h} &= \{(g,k): g \notin \mathcal{G}_h, (g,k) \in \mathcal{I}_h\};\\
\mathcal{A}_{00h} &= \{(g,k): g \notin \mathcal{G}_h, (g,k) \notin \mathcal{I}_h\}.
\end{aligned} \tag{64}$$

Note that every pair (g,k) belongs to exactly one of these subsets.

Using this partition in combination with assumptions 2'' and 3, we can approximate $B_{hO2}^{q,q-4}$ as follows:

$$\begin{aligned}
B_{hO2}^{q,q-4} \approx -\sum_{c=1}^{\mathcal{C}} \sum\nolimits_{i\in U_c^{q-4,q}} y_i^{q-4} y_i^q \Bigg[& \sum_{(g,k)\in\mathcal{A}_{11h}} a_{gi}^{q-4} a_{ki}^{q} \left(\mathbb{p}_{gkhhc}^{OLC} - p_{ghc}^{OL} p_{gkhc}^{OLC}\right)\\
& + \sum_{(g,k)\in\mathcal{A}_{10h}} a_{gi}^{q-4} a_{ki}^{q} \left(\widetilde{\mathbb{p}}_{gkhhc}^{OLC} - p_{ghc}^{OL} \tilde{p}_{gkhc}^{OLC}\right) + \sum_{(g,k)\in\mathcal{A}_{01h}} a_{gi}^{q-4} a_{ki}^{q} \left(\mathbb{p}_{gkhhc}^{OLC} - \tilde{p}_{ghc}^{OL} p_{gkhc}^{OLC}\right)\\
& + \sum_{(g,k)\in\mathcal{A}_{00h}} a_{gi}^{q-4} a_{ki}^{q} \left(\widetilde{\mathbb{p}}_{gkhhc}^{OLC} - \tilde{p}_{ghc}^{OL} \tilde{p}_{gkhc}^{OLC}\right) \Bigg]\\
= -\sum_{c=1}^{\mathcal{C}} \Bigg[& \sum_{(g,k)\in\mathcal{A}_{11h}} \left(\mathbb{p}_{gkhhc}^{OLC} - p_{ghc}^{OL} p_{gkhc}^{OLC}\right) S_{gkcO}\left(y^{q-4} y^q\right)\\
& + \sum_{(g,k)\in\mathcal{A}_{10h}} \left(\frac{\widetilde{\mathbb{p}}_{(-\mathcal{I}_h)hhc}^{OLC}}{M^2 - |\mathcal{I}_h|} - p_{ghc}^{OL} \frac{\tilde{p}_{(-\mathcal{I}_h)hc}^{OLC}}{M^2 - |\mathcal{I}_h|}\right) S_{gkcO}\left(y^{q-4} y^q\right)\\
& + \sum_{(g,k)\in\mathcal{A}_{01h}} \left(\mathbb{p}_{gkhhc}^{OLC} - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}}{M - |\mathcal{G}_h|} p_{gkhc}^{OLC}\right) S_{gkcO}\left(y^{q-4} y^q\right)\\
& + \sum_{(g,k)\in\mathcal{A}_{00h}} \left(\frac{\widetilde{\mathbb{p}}_{(-\mathcal{I}_h)hhc}^{OLC}}{M^2 - |\mathcal{I}_h|} - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}}{M - |\mathcal{G}_h|} \frac{\tilde{p}_{(-\mathcal{I}_h)hc}^{OLC}}{M^2 - |\mathcal{I}_h|}\right) S_{gkcO}\left(y^{q-4} y^q\right) \Bigg]
\end{aligned}$$

$$= -\sum_{c=1}^{\mathcal{C}} \Bigg[\sum_{(g,k)\in\mathcal{A}_{11h}} \left(\mathbb{p}^{OLC}_{gkhhc} - p^{OL}_{ghc} p^{OLC}_{gkhc}\right) S_{gkcO}(y^{q-4} y^{q}) + \sum_{(g,k)\in\mathcal{A}_{10h}} \left(\widetilde{\mathbb{p}}^{OLC}_{(-\mathcal{T}_h)hhc} - p^{OL}_{ghc} \tilde{p}^{OLC}_{(-\mathcal{T}_h)hc}\right) \frac{1}{M^2 - |\mathcal{T}_h|} S_{gkcO}(y^{q-4} y^{q}) + \sum_{(g,k)\in\mathcal{A}_{01h}} \left(\mathbb{p}^{OLC}_{gkhhc} - \frac{\tilde{p}^{OL}_{(-\mathcal{G}_h)hc}}{M - |\mathcal{G}_h|} p^{OLC}_{gkhc}\right) S_{gkcO}(y^{q-4} y^{q}) + \left(\widetilde{\mathbb{p}}^{OLC}_{(-\mathcal{T}_h)hhc} - \frac{\tilde{p}^{OL}_{(-\mathcal{G}_h)hc}}{M - |\mathcal{G}_h|} \tilde{p}^{OLC}_{(-\mathcal{T}_h)hc}\right) \frac{|\mathcal{A}_{00h}|}{M^2 - |\mathcal{T}_h|} \bar{S}_{(\mathcal{A}_{00h})cO}(y^{q-4} y^{q}) \Bigg].$$

Here, we made use of the following extension of the shorthand notation defined in (24):

$$S_{gkcO}(z) = \sum\nolimits_{i\in U_c^{q-4,q}} a_{gi}^{q-4} a_{ki}^{q} z\,,$$

$$\bar{S}_{(\mathcal{A}_{00h})cO}(z) = \frac{\sum_{(g,k)\in\mathcal{A}_{00h}} S_{gkcO}(z)}{|\mathcal{A}_{00h}|} = \frac{S_{++cO}(z) - \sum_{(g,k)\in(\mathcal{A}_{11h}\cup\mathcal{A}_{10h}\cup\mathcal{A}_{01h})} S_{gkcO}(z)}{M^2 - |\mathcal{A}_{11h}| - |\mathcal{A}_{10h}| - |\mathcal{A}_{01h}|},$$

with z denoting an arbitrary variable.

Combining the two sub-terms, we find the following approximation to $B_{hO}^{q,q-4}$:

$$\begin{aligned} B_{hO}^{q,q-4} \approx \sum_{c=1}^{\mathcal{C}} \Bigg\{ & \breve{G}_h^{q,q-4} \left[\sum_{g\in\mathcal{G}_h} p^{OL}_{ghc}(1 - p^{OL}_{ghc}) SS_{gcO}(y^{q-4}) + \tilde{p}^{OL}_{(-\mathcal{G}_h)hc} \left(1 - \frac{\tilde{p}^{OL}_{(-\mathcal{G}_h)hc}}{M - |\mathcal{G}_h|}\right) \overline{SS}_{(-\mathcal{G}_h)cO}(y^{q-4}) \right] \\ & - \sum_{(g,k)\in\mathcal{A}_{11h}} \left(\mathbb{p}^{OLC}_{gkhhc} - p^{OL}_{ghc} p^{OLC}_{gkhc}\right) S_{gkcO}(y^{q-4} y^{q}) \\ & - \sum_{(g,k)\in\mathcal{A}_{10h}} \left(\widetilde{\mathbb{p}}^{OLC}_{(-\mathcal{T}_h)hhc} - p^{OL}_{ghc} \tilde{p}^{OLC}_{(-\mathcal{T}_h)hc}\right) \frac{1}{M^2 - |\mathcal{T}_h|} S_{gkcO}(y^{q-4} y^{q}) \\ & - \sum_{(g,k)\in\mathcal{A}_{01h}} \left(\mathbb{p}^{OLC}_{gkhhc} - \frac{\tilde{p}^{OL}_{(-\mathcal{G}_h)hc}}{M - |\mathcal{G}_h|} p^{OLC}_{gkhc}\right) S_{gkcO}(y^{q-4} y^{q}) \\ & - \left(\widetilde{\mathbb{p}}^{OLC}_{(-\mathcal{T}_h)hhc} - \frac{\tilde{p}^{OL}_{(-\mathcal{G}_h)hc}}{M - |\mathcal{G}_h|} \tilde{p}^{OLC}_{(-\mathcal{T}_h)hc}\right) \frac{|\mathcal{A}_{00h}|}{M^2 - |\mathcal{T}_h|} \bar{S}_{(\mathcal{A}_{00h})cO}(y^{q-4} y^{q}) \Bigg\}. \end{aligned} \tag{65}$$

It is interesting to note that $|\mathcal{A}_{00h}| = M^2 - |\mathcal{A}_{11h}| - |\mathcal{A}_{10h}| - |\mathcal{A}_{01h}| = M^2 - |\mathcal{T}_h| - |\mathcal{A}_{10h}|$ by definition (64). In practice, the fraction $\frac{|\mathcal{A}_{00h}|}{M^2 - |\mathcal{T}_h|} = 1 - \frac{|\mathcal{A}_{10h}|}{M^2 - |\mathcal{T}_h|}$ is supposed to be close to 1.

The term $B_{hD}^{q,q-4}$ in (51) refers to units that no longer exist in the new quarter ($U^{q-4} \backslash U^{q-4,q}$). An approximation to this term can be derived completely analogously to the above derivation for $B_{hO1}^{q,q-4}$. We obtain:

$$B_{hD}^{q,q-4} \approx \breve{G}_h^{q,q-4} \sum_{c=1}^{\mathcal{C}} \left[\sum_{g\in\mathcal{G}_h} p^{OL}_{ghc}(1 - p^{OL}_{ghc}) SS_{gcD}(y^{q-4}) + \tilde{p}^{OL}_{(-\mathcal{G}_h)hc} \left(1 - \frac{\tilde{p}^{OL}_{(-\mathcal{G}_h)hc}}{M - |\mathcal{G}_h|}\right) \overline{SS}_{(-\mathcal{G}_h)cD}(y^{q-4}) \right]. \tag{66}$$

Combining all terms from (62), (65) and (66), we obtain the following approximation to $AB\left(\hat{G}_h^{q,q-4}\right)$:

$$AB\left(\hat{G}_h^{q,q-4}\right) \approx \breve{G}_h^{q,q-4} - G_h^{q,q-4} + \frac{\breve{G}_h^{q,q-4} \sum_{c=1}^{\mathcal{C}} B_{hc}^{q,q-4} - \sum_{c=1}^{\mathcal{C}} C_{hcO}^{q,q-4}}{\left\{\sum_{c=1}^{\mathcal{C}} \left[\sum_{g\in\mathcal{G}_h} p^{OL}_{ghc} Y_{gc}^{q-4} + \tilde{p}^{OL}_{(-\mathcal{G}_h)hc} \bar{Y}_{(-\mathcal{G}_h)c}^{q-4}\right]\right\}^2}, \tag{67}$$

$$\breve{G}_h^{q,q-4} \approx \frac{\sum_{c=1}^{\mathcal{C}} \left[\sum_{(g,k)\in\mathcal{T}_h} p_{gkhc}^{OLC} Y_{gkcO}^{q} + \tilde{p}_{(-\mathcal{T}_h)hc}^{OLC} \bar{Y}_{(-\mathcal{T}_h)cO}^{q} + \sum_{g\in\mathcal{G}_h} p_{ghc}^{OL} Y_{gcB}^{q} + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL} \bar{Y}_{(-\mathcal{G}_h)cB}^{q}\right]}{\sum_{c=1}^{\mathcal{C}} \left[\sum_{g\in\mathcal{G}_h} p_{ghc}^{OL} Y_{gc}^{q-4} + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL} \bar{Y}_{(-\mathcal{G}_h)c}^{q-4}\right]},$$

$$B_{hc}^{q,q-4} = \sum_{g\in\mathcal{G}_h} p_{ghc}^{OL}\left(1-p_{ghc}^{OL}\right) BS_{gc}^{q,q-4} + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL}\left(1-\frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}}{M-|\mathcal{G}_h|}\right)\overline{BS}_{(-\mathcal{G}_h)c}^{q,q-4},$$

$$BS_{gc}^{q,q-4} = SS_{gcO}(y^{q-4}) + SS_{gcD}(y^{q-4}),$$

$$\overline{BS}_{(-\mathcal{G}_h)c}^{q,q-4} = \overline{SS}_{(-\mathcal{G}_h)cO}(y^{q-4}) + \overline{SS}_{(-\mathcal{G}_h)cD}(y^{q-4}),$$

$$\begin{aligned} C_{hcO}^{q,q-4} = & \sum_{(g,k)\in\mathcal{A}_{11h}} \left(\mathbb{p}_{gkhhc}^{OLC} - p_{ghc}^{OL} p_{gkhc}^{OLC}\right) S_{gkcO}(y^{q-4}y^{q}) \\ & + \sum_{(g,k)\in\mathcal{A}_{10h}} \left(\tilde{\mathbb{p}}_{(-\mathcal{T}_h)hhc}^{OLC} - p_{ghc}^{OL}\tilde{p}_{(-\mathcal{T}_h)hc}^{OLC}\right)\frac{1}{M^2-|\mathcal{T}_h|} S_{gkcO}(y^{q-4}y^{q}) \\ & + \sum_{(g,k)\in\mathcal{A}_{01h}} \left(\mathbb{p}_{gkhhc}^{OLC} - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}}{M-|\mathcal{G}_h|} p_{gkhc}^{OLC}\right) S_{gkcO}(y^{q-4}y^{q}) \\ & + \left(\tilde{\mathbb{p}}_{(-\mathcal{T}_h)hhc}^{OLC} - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}}{M-|\mathcal{G}_h|}\tilde{p}_{(-\mathcal{T}_h)hc}^{OLC}\right)\frac{|\mathcal{A}_{00h}|}{M^2-|\mathcal{T}_h|}\bar{S}_{(\mathcal{A}_{00h})cO}(y^{q-4}y^{q}). \end{aligned}$$

Note that, similar to (29), we have combined some of the contributions to $AB\left(\hat{G}_h^{q,q-4}\right)$ from continuing and dead units into two single components $BS_{gc}^{q,q-4}$ and $\overline{BS}_{(-\mathcal{G}_h)c}^{q,q-4}$. In (67) the terms $B_{hc}^{q,q-4}$ involve the contribution of the level matrix, concerning quarter $q-4$ for dead and continuing units, and the terms $C_{hcO}^{q,q-4}$ involve the contribution of the level-change matrix, concerning both quarters for the overlapping units.

4.5.3 Variance

The variance approximation in formula (53) consists of three components:

$$AV\left(\hat{G}_h^{q,q-4}\right) = \frac{1}{\left[E\left(\hat{Y}_h^{q-4}\right)\right]^2}\left(V_{hO}^{q,q-4} + V_{hD}^{q,q-4} + V_{hB}^{q,q-4}\right).$$

The first term refers to the continuing units ($U^{q-4,q}$). We can write $V_{hO}^{q,q-4} = V_{hO1}^{q,q-4} + V_{hO2}^{q,q-4} + V_{hO3}^{q,q-4}$, with

$$\begin{aligned} V_{hO1}^{q,q-4} &= \left(\breve{G}_h^{q,q-4}\right)^2 \sum\nolimits_{i\in U^{q-4,q}} \left(y_i^{q-4}\right)^2 \sum\nolimits_{g=1}^{M} a_{gi}^{q-4} p_{ghi}^{OL}\left(1-p_{ghi}^{OL}\right), \\ V_{hO2}^{q,q-4} &= \sum\nolimits_{i\in U^{q-4,q}} \left(y_i^{q}\right)^2 \sum\nolimits_{g=1}^{M}\sum\nolimits_{k=1}^{M} a_{gi}^{q-4} a_{ki}^{q} p_{gkhi}^{OLC}\left(1-p_{gkhi}^{OLC}\right), \\ V_{hO3}^{q,q-4} &= -2\breve{G}_h^{q,q-4} \sum\nolimits_{i\in U^{q-4,q}} y_i^{q-4} y_i^{q} \sum\nolimits_{g=1}^{M}\sum\nolimits_{k=1}^{M} a_{gi}^{q-4} a_{ki}^{q} \left(\mathbb{p}_{gkhhi}^{OLC} - p_{ghi}^{OL} p_{gkhi}^{OLC}\right). \end{aligned} \tag{68}$$

The first of these sub-terms involves only the level matrix. By a derivation that is almost identical to the one for $B_{hO1}^{q,q-4}$ in Section 4.5.2, we obtain:

$$V_{hO1}^{q,q-4} \approx \left(\breve{G}_h^{q,q-4}\right)^2 \sum_{c=1}^{\mathcal{C}} \left[\sum_{g\in\mathcal{G}_h} p_{ghc}^{OL}\left(1-p_{ghc}^{OL}\right) SS_{gcO}(y^{q-4}) + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL}\left(1-\frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}}{M-|\mathcal{G}_h|}\right)\overline{SS}_{(-\mathcal{G}_h)cO}(y^{q-4})\right].$$

The second sub-term $V_{hO2}^{q,q-4}$ in (68) involves probabilities p_{gkhi}^{OLC}. Here, we can proceed in a similar way as for $E\left(\hat{Y}_{hO}^{q}\right)$ in (60) to obtain:

$$V_{hO2}^{q,q-4} \approx \sum_{c=1}^{\mathcal{C}} \left[\sum_{(g,k)\in\mathcal{T}_h} p_{gkhc}^{OLC}\left(1-p_{gkhc}^{OLC}\right) SS_{gkcO}(y^{q}) + \tilde{p}_{(-\mathcal{T}_h)hc}^{OLC}\left(1-\frac{\tilde{p}_{(-\mathcal{T}_h)hc}^{OLC}}{M^2-|\mathcal{T}_h|}\right)\overline{SS}_{(-\mathcal{T}_h)cO}(y^{q})\right],$$

in notation that should by now be obvious.

The third and final sub-term $V_{hO3}^{q,q-4}$ in (68) is, again, slightly more complicated, because it involves all probabilities p_{ghi}^{OL}, p_{gkhi}^{OLC} and $\mathbb{p}_{gkhhi}^{OLC}$. Using the same partition of pairs of industries into four subsets (64) that was used in Section 4.5.2, we can derive analogously to the approximation for $B_{hO2}^{q,q-4}$ that:

$$\begin{aligned} V_{hO3}^{q,q-4} \approx -2\breve{G}_h^{q,q-4} \sum_{c=1}^{\mathcal{C}} \Bigg[& \sum_{(g,k)\in\mathcal{A}_{11h}} \left(\mathbb{p}_{gkhhc}^{OLC} - p_{ghc}^{OL} p_{gkhc}^{OLC}\right) S_{gkcO}(y^{q-4}y^{q}) \\ & + \sum_{(g,k)\in\mathcal{A}_{10h}} \left(\widetilde{\mathbb{p}}_{(-\mathcal{T}_h)hhc}^{OLC} - p_{ghc}^{OL} \tilde{p}_{(-\mathcal{T}_h)hc}^{OLC}\right) \frac{1}{M^2 - |\mathcal{T}_h|} S_{gkcO}(y^{q-4}y^{q}) \\ & + \sum_{(g,k)\in\mathcal{A}_{01h}} \left(\mathbb{p}_{gkhhc}^{OLC} - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}}{M - |\mathcal{G}_h|} p_{gkhc}^{OLC}\right) S_{gkcO}(y^{q-4}y^{q}) \\ & + \left(\widetilde{\mathbb{p}}_{(-\mathcal{T}_h)hhc}^{OLC} - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}}{M - |\mathcal{G}_h|} \tilde{p}_{(-\mathcal{T}_h)hc}^{OLC}\right) \frac{|\mathcal{A}_{00h}|}{M^2 - |\mathcal{T}_h|} \bar{S}_{(\mathcal{A}_{00h})cO}(y^{q-4}y^{q}) \Bigg] \\ = -2\breve{G}_h^{q,q-4} & \sum_{c=1}^{\mathcal{C}} C_{hcO}^{q,q-4}, \end{aligned}$$

with $C_{hcO}^{q,q-4}$ as defined in (67).

Combining the three sub-terms, we find the following approximation to $V_{hO}^{q,q-4}$:

$$\begin{aligned} V_{hO}^{q,q-4} \approx \sum_{c=1}^{\mathcal{C}} \Bigg\{ & \left(\breve{G}_h^{q,q-4}\right)^2 \Bigg[\sum_{g\in\mathcal{G}_h} p_{ghc}^{OL}\left(1 - p_{ghc}^{OL}\right) SS_{gcO}(y^{q-4}) \\ & + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL} \left(1 - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}}{M - |\mathcal{G}_h|}\right) \overline{SS}_{(-\mathcal{G}_h)cO}(y^{q-4}) \Bigg] \\ & + \sum_{(g,k)\in\mathcal{T}_h} p_{gkhc}^{OLC}\left(1 - p_{gkhc}^{OLC}\right) SS_{gkcO}(y^{q}) + \tilde{p}_{(-\mathcal{T}_h)hc}^{OLC} \left(1 - \frac{\tilde{p}_{(-\mathcal{T}_h)hc}^{OLC}}{M^2 - |\mathcal{T}_h|}\right) \overline{SS}_{(-\mathcal{T}_h)cO}(y^{q}) \\ & - 2\breve{G}_h^{q,q-4} C_{hcO}^{q,q-4} \Bigg\}. \end{aligned} \tag{69}$$

The remaining two terms in (53) are now straightforward. The term $V_{hD}^{q,q-4}$ refers to units that no longer exist in the new quarter ($U^{q-4}\backslash U^{q-4,q}$). Analogously to $V_{hO1}^{q,q-4}$, we find that

$$\begin{aligned} V_{hD}^{q,q-4} \approx \left(\breve{G}_h^{q,q-4}\right)^2 \sum_{c=1}^{\mathcal{C}} \Bigg[& \sum_{g\in\mathcal{G}_h} p_{ghc}^{OL}\left(1 - p_{ghc}^{OL}\right) SS_{gcD}(y^{q-4}) \\ & + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL} \left(1 - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}}{M - |\mathcal{G}_h|}\right) \overline{SS}_{(-\mathcal{G}_h)cD}(y^{q-4}) \Bigg] \end{aligned} \tag{70}$$

The term $V_{hB}^{q,q-4}$ refers to new-born units in the new quarter ($U^{q}\backslash U^{q-4,q}$). For this term, we find that

$$V_{hB}^{q,q-4} \approx \sum_{c=1}^{\mathcal{C}} \left[\sum_{g \in \mathcal{G}_h} p_{ghc}^{OL}\left(1 - p_{ghc}^{OL}\right) SS_{gcB}(y^q) + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL} \left(1 - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}}{M - |\mathcal{G}_h|}\right) \overline{SS}_{(-\mathcal{G}_h)cB}(y^q) \right]. \tag{71}$$

Upon combining the expressions in (69), (70) and (71), we obtain the following variance approximation:

$$\begin{aligned}
AV\left(\hat{G}_h^{q,q-4}\right) &\approx \frac{\sum_{c=1}^{\mathcal{C}} \left(V_{1hc}^{q,q-4} + V_{2hcO}^{q,q-4} - 2\check{G}_h^{q,q-4} C_{hOc}^{q,q-4}\right)}{\left\{\sum_{c=1}^{\mathcal{C}} \left[\sum_{g \in \mathcal{G}_h} p_{ghc}^{OL} Y_{gc}^{q-4} + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL} \bar{Y}_{(-\mathcal{G}_h)c}^{q-4}\right]\right\}^2}, \\
V_{1hc}^{q,q-4} &= \sum_{g \in \mathcal{G}_h} p_{ghc}^{OL}\left(1 - p_{ghc}^{OL}\right) VS_{gc}^{q,q-4} + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL} \left(1 - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}}{M - |\mathcal{G}_h|}\right) \overline{VS}_{(-\mathcal{G}_h)c}^{q,q-4}, \\
V_{2hcO}^{q,q-4} &= \sum_{(g,k) \in \mathcal{T}_h} p_{gkhc}^{OLC}\left(1 - p_{gkhc}^{OLC}\right) SS_{gkcO}(y^q) + \tilde{p}_{(-\mathcal{T}_h)hc}^{OLC} \left(1 - \frac{\tilde{p}_{(-\mathcal{T}_h)hc}^{OLC}}{M^2 - |\mathcal{T}_h|}\right) \overline{SS}_{(-\mathcal{T}_h)cO}(y^q), \\
VS_{gc}^{q,q-4} &= SS_{gcO}\left(\check{G}_h^{q,q-4} y^{q-4}\right) + SS_{gcD}\left(\check{G}_h^{q,q-4} y^{q-4}\right) + SS_{gcB}(y^q), \\
\overline{VS}_{(-\mathcal{G}_h)c}^{q,q-4} &= \overline{SS}_{(-\mathcal{G}_h)cO}\left(\check{G}_h^{q,q-4} y^{q-4}\right) + \overline{SS}_{(-\mathcal{G}_h)cD}\left(\check{G}_h^{q,q-4} y^{q-4}\right) + \overline{SS}_{(-\mathcal{G}_h)cB}(y^q),
\end{aligned} \tag{72}$$

with $C_{hcO}^{q,q-4}$ as defined in (67).

4.6 Estimating the bias and variance

To estimate the bias and variance formulas in Section 4.5, we can proceed in a similar way as in Section 3.6. Define:

$$\begin{aligned}
\hat{B}_{hc}^{q,q-4} &= \sum_{g \in \mathcal{G}_h} p_{ghc}^{OL}\left(1 - p_{ghc}^{OL}\right) \widehat{BS}_{gc}^{q,q-4} + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL} \left(1 - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}}{M - |\mathcal{G}_h|}\right) \widehat{\overline{BS}}_{(-\mathcal{G}_h)c}^{q,q-4}, \\
\widehat{BS}_{gc}^{q,q-4} &= \widehat{SS}_{gcO}(y^{q-4}) + \widehat{SS}_{gcD}(y^{q-4}), \\
\widehat{\overline{BS}}_{(-\mathcal{G}_h)c}^{q,q-4} &= \widehat{\overline{SS}}_{(-\mathcal{G}_h)cO}(y^{q-4}) + \widehat{\overline{SS}}_{(-\mathcal{G}_h)cD}(y^{q-4}), \\
\hat{C}_{hcO}^{q,q-4} &= \sum_{(g,k) \in \mathcal{A}_{11h}} \left(\mathbb{p}_{gkhhc}^{OLC} - p_{ghc}^{OL} p_{gkhc}^{OLC}\right) \hat{S}_{gkcO}(y^{q-4} y^q) \\
&\quad + \sum_{(g,k) \in \mathcal{A}_{10h}} \left(\tilde{\mathbb{p}}_{(-\mathcal{T}_h)hhc}^{OLC} - p_{ghc}^{OL} \tilde{p}_{(-\mathcal{T}_h)hc}^{OLC}\right) \frac{1}{M^2 - |\mathcal{T}_h|} \hat{S}_{gkcO}(y^{q-4} y^q) \\
&\quad + \sum_{(g,k) \in \mathcal{A}_{01h}} \left(\mathbb{p}_{gkhhc}^{OLC} - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}}{M - |\mathcal{G}_h|} p_{gkhc}^{OLC}\right) \hat{S}_{gkcO}(y^{q-4} y^q) \\
&\quad + \left(\tilde{\mathbb{p}}_{(-\mathcal{T}_h)hhc}^{OLC} - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}}{M - |\mathcal{G}_h|} \tilde{p}_{(-\mathcal{T}_h)hc}^{OLC}\right) \frac{|\mathcal{A}_{00h}|}{M^2 - |\mathcal{T}_h|} \hat{\tilde{S}}_{(\mathcal{A}_{00h})cO}(y^{q-4} y^q).
\end{aligned}$$

Then the bias in (67) can be estimated by:

$$\begin{aligned}
\widehat{AB}\left(\hat{G}_h^{q,q-4}\right) &= \hat{\tilde{G}}_h^{q,q-4} - \frac{\hat{Y}_h^q}{\hat{Y}_h^{q-4}} + \frac{\hat{\tilde{G}}_h^{q,q-4} \sum_{c=1}^{\mathcal{C}} \hat{B}_{hc}^{q,q-4} - \sum_{c=1}^{\mathcal{C}} \hat{C}_{hcO}^{q,q-4}}{\left\{\sum_{c=1}^{\mathcal{C}} \left[\sum_{g \in \mathcal{G}_h} p_{ghc}^{OL} \hat{Y}_{gc}^{q-4} + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL} \hat{\bar{Y}}_{(-\mathcal{G}_h)c}^{q-4}\right]\right\}^2}, \\
\hat{\tilde{G}}_h^{q,q-4} &= \frac{\sum_{c=1}^{\mathcal{C}} \left[\sum_{(g,k) \in \mathcal{T}_h} p_{gkhc}^{OLC} \hat{Y}_{gkcO}^q + \tilde{p}_{(-\mathcal{T}_h)hc}^{OLC} \hat{\bar{Y}}_{(-\mathcal{T}_h)cO}^q + \sum_{g \in \mathcal{G}_h} p_{ghc}^{OL} \hat{Y}_{gcB}^q + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL} \hat{\bar{Y}}_{(-\mathcal{G}_h)cB}^q\right]}{\sum_{c=1}^{\mathcal{C}} \left[\sum_{g \in \mathcal{G}_h} p_{ghc}^{OL} \hat{Y}_{gc}^{q-4} + \tilde{p}_{(-\mathcal{G}_h)hc}^{OL} \hat{\bar{Y}}_{(-\mathcal{G}_h)c}^{q-4}\right]}.
\end{aligned} \tag{73}$$

Here, $\hat{\tilde{G}}_h^{q,q-4}$ is defined analogously to $\hat{\tilde{G}}_h^{q,q-1}$ in (31), and the estimated sums and sums of squares are defined analogously to (34).

Similarly, the variance in (72) can be estimated by:

$$
\begin{aligned}
\widehat{AV}\left(\hat{G}_h^{q,q-4}\right) &= \frac{\sum_{c=1}^{\mathcal{C}}\left(\hat{V}_{1hc}^{q,q-4}+\hat{V}_{2hcO}^{q,q-4}-2\hat{\tilde{G}}_h^{q,q-4}\hat{C}_{hOc}^{q,q-4}\right)}{\left\{\sum_{c=1}^{\mathcal{C}}\left[\sum_{g\in\mathcal{G}_h}p_{ghc}^{OL}\hat{Y}_{gc}^{q-4}+\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}\hat{\tilde{Y}}_{(-\mathcal{G}_h)c}^{q-4}\right]\right\}^2},\\
\hat{V}_{1hc}^{q,q-4} &= \sum_{g\in\mathcal{G}_h}p_{ghc}^{OL}\left(1-p_{ghc}^{OL}\right)\widehat{VS}_{gc}^{q,q-4}+\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}\left(1-\frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}}{M-|\mathcal{G}_h|}\right)\widehat{\widetilde{VS}}_{(-\mathcal{G}_h)c}^{q,q-4},\\
\hat{V}_{2hcO}^{q,q-4} &= \sum_{(g,k)\in\mathcal{T}_h}p_{gkhc}^{OLC}\left(1-p_{gkhc}^{OLC}\right)\widehat{SS}_{gkcO}(y^q)+\tilde{p}_{(-\mathcal{T}_h)hc}^{OLC}\left(1-\frac{\tilde{p}_{(-\mathcal{T}_h)hc}^{OLC}}{M^2-|\mathcal{T}_h|}\right)\widehat{\widetilde{SS}}_{(-\mathcal{T}_h)cO}(y^q),\\
\widehat{VS}_{gc}^{q,q-4} &= \widehat{SS}_{gcO}\left(\hat{\tilde{G}}_h^{q,q-4}y^{q-4}\right)+\widehat{SS}_{gcD}\left(\hat{\tilde{G}}_h^{q,q-4}y^{q-4}\right)+\widehat{SS}_{gcB}(y^q),\\
\widehat{\widetilde{VS}}_{(-\mathcal{G}_h)c}^{q,q-4} &= \widehat{\widetilde{SS}}_{(-\mathcal{G}_h)cO}\left(\hat{\tilde{G}}_h^{q,q-4}y^{q-4}\right)+\widehat{\widetilde{SS}}_{(-\mathcal{G}_h)cD}\left(\hat{\tilde{G}}_h^{q,q-4}y^{q-4}\right)+\widehat{\widetilde{SS}}_{(-\mathcal{G}_h)cB}(y^q).
\end{aligned}
\tag{74}
$$

5 Application

5.1 Data

We applied the expressions concerning the accuracy to a case study on the short-term statistics for a small economic sector as an example: car trade (NACE G45). Car trade consists of nine underlying industries. In terms of total quarterly turnover, the largest industry within car trade is industry code 45112 (sale and repair cars and light motor vehicles) and the smallest industry is code 45194 (sale and repair of caravans); see Table 2.

Table 2. Average quarterly turnover and average number of units per car trade industry (Q2 2014 – Q4 2015)

Industry	Number of units	Turnover (mln. Euros)
45112	18 618	7 961
45111	157	2 910
45310	1 961	2 194
45191X	1 329	1 202
45200	6 022	687
45401	446	374
45402	1 246	150
45320	841	89
45194	363	79

The Dutch short-term statistics on turnover are derived from two data sources. Value Added Tax (VAT) data are used for all fiscal units that can be linked uniquely to an enterprise in the Dutch GBR (“simple units”); in practice, these are mostly smaller units. For the remaining enterprises (“complex units”), Statistics Netherlands conducts a census survey on a monthly or quarterly basis. Thus, the two sources are complementary and together cover the entire car trade population in the GBR.

In practice, classification errors with respect to NACE code in the GBR can be detected as part of the editing process during regular statistical production. The focus of editing is usually on the largest units. As a rule of thumb, subject-matter experts ‘know’ the largest 25 units in each industry well enough so that we may assume that the NACE codes of these units are correct. (Below, the production-editing of these largest units will be referred to as “supplemental editing”.) In addition, a special team has been set up at Statistics Netherlands to ensure consistency between statistical outputs for units that belong to an enterprise group with a complicated (international) structure (“most complex units”). We also assume that the NACE codes of these most complex units are correctly observed.

For the present application, we used turnover data for eight quarters in 2014 and 2015.

5.2 Input parameters for quarter-on-quarter growth rates

In the current paper we will limit the computations of the analytical expressions to the quarter-on-quarter growth rates within a single year as given in Section 3.

To apply formula (32) for the estimated bias and formula (35) for the estimated variance of a quarterly growth rate, the following input is needed:

- the division of units into probability classes $c = 1, \dots, \mathcal{C}$;
- for each observed industry code h within car trade, the subset $\mathcal{G}_h \subset \{1, \dots, M\}$;
- the classification error probabilities p_{ghc}^{OL} (for $g \in \mathcal{G}_h$) and $\tilde{p}_{(-\mathcal{G}_h)hc}^{OL}$.

For the specification and estimation of these input parameters, we re-used some previous results of van Delden et al. (2016b). In particular, we re-used the audit data that were collected for that study. These audit data were obtained for a small sample of 25 units drawn at 1 July 2014 from each of the (nine) observed car trade industries. For these audited units, the true industry code was determined by two experts. We assumed that the results of this audit could be applied to all quarters in 2014 and 2015.

We selected a special audit sample, because Statistics Netherlands no longer has a regular GBR quality study. Some other countries (for instance, Switzerland and Croatia) do conduct such quality studies or collect data on the quality of the GBR as part of their regular statistical production. These countries may be able to use the resulting data to specify and estimate the input parameters of the bias and variance formulas, without the need for collecting additional audit data. In the future, we aim to investigate the possibilities of using data from regular statistical production instead of an audit sample also at Statistics Netherlands.

5.2.1 Probability classes and their probabilities

In van Delden et al. (2016b), eleven probability classes were distinguished. However, some of these classes actually had the same level matrix $\mathbf{P}_c^{OL}$. By merging these cases, we obtained five distinct probability classes:

1. simple units with less than 10 employees (EMP) that consist of at most two legal units;
2. simple units with less than 10 EMP that consist of at least three legal units;
3. simple units with 10–19 EMP, complex units with less than 20 EMP, and most complex units with less than 10 EMP;
4. simple units with at least 20 EMP, complex units with 20–49 EMP, and most complex units with 10–19 EMP;
5. complex units with at least 50 EMP, most complex units with at least 20 EMP, and all units in supplemental editing.

Note that a unit that is part of supplemental editing always belongs to probability class 5, even when it satisfies the definition of one of the other classes.

In estimating the level matrices $\mathbf{P}_c^{OL}$, van Delden et al. (2016b) distinguished between rows and columns within car trade and misclassifications between car trade and other economic sectors:

$$\frac{true \setminus observed}{\begin{matrix} car\ trade\ industries \\ other\ industries \end{matrix}} \quad \left(\begin{matrix} car\ trade\ industries \\ within\ car\ trade \\ erroneously\ included \end{matrix} \middle| \begin{matrix} other\ industries \\ missed \\ outside\ car\ trade \end{matrix} \right)$$

For probability classes 1 and 2, the classification errors probabilities “within car trade” were estimated from the audit data, using a logistic regression model for the diagonal elements and a log-linear model for the off-diagonal elements. The level matrix for probability class 5 was assumed to be an identity matrix; that is, it was

assumed that no classification errors occur for units in this class. The level matrices for probability classes 3 and 4 were obtained by interpolating the matrices of classes 2 and 5; see van Delden et al. (2016b) for details.

For easy reference, the estimated probabilities for the five probability classes are given in Table 3 and Figure 1. The table contains the diagonal elements (probability of correct classification) which differ by probability class. The figure contains the *conditional* probabilities of misclassification, given that a classification error occurs; these probabilities are the same for each probability class. From the information in the table and figure, one can compute all elements of the level matrices $\mathbf{P}_c^{OL}$ for the components "within car trade" and "missed" in the above diagram. For instance, the probability that a unit in probability class 2 with true industry code 45112 is correctly classified is 0.88 and the probability that this unit is misclassified as industry code 45200 is (1 – 0.88) × 0.36 = 0.0432.

Figure 1 also contains "overall" conditional probabilities for the "erroneously included" units (i.e., with the non-car trade industries collapsed into one industry "Other"). In the previous bootstrap approach, the components "erroneously included" and "outside car trade" of the matrix $\mathbf{P}_c^{OL}$ were not computed explicitly at a more detailed level; see van Delden et al. (2016a, 2016b). Here, we estimated these detailed probabilities by a slightly different approach, while still re-using some results from the previous study. In what follows, we use $\mathcal{M} = \{1, \dots, M\}$ to denote the total set of industries and $\mathcal{M}_T$ to denote the set of target industries (in this application: the nine car trade industries).

Firstly, according to van Delden et al. (2015, Appendix A.2) the probability that a unit with true industry code outside car trade is misclassified into a car trade industry is estimated as 0.000625. By multiplying this probability with the conditional probabilities in the last row of Figure 1, we obtained estimates of the unconditional probabilities that a unit with true industry code outside car trade is erroneously classified into a specific car trade industry.

Table 3. Estimated probabilities for the diagonal elements of the level matrix within car trade (values taken from van Delden et al., 2016b).

Probability class	True industry code								
	45112	45111	45310	45191X	45200	45401	45402	45320	45194
1	0.97	0.10	0.93	0.84	0.93	0.44	0.92	0.48	0.83
2	0.88	0.02	0.75	0.53	0.74	0.15	0.73	0.16	0.52
3	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
4	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

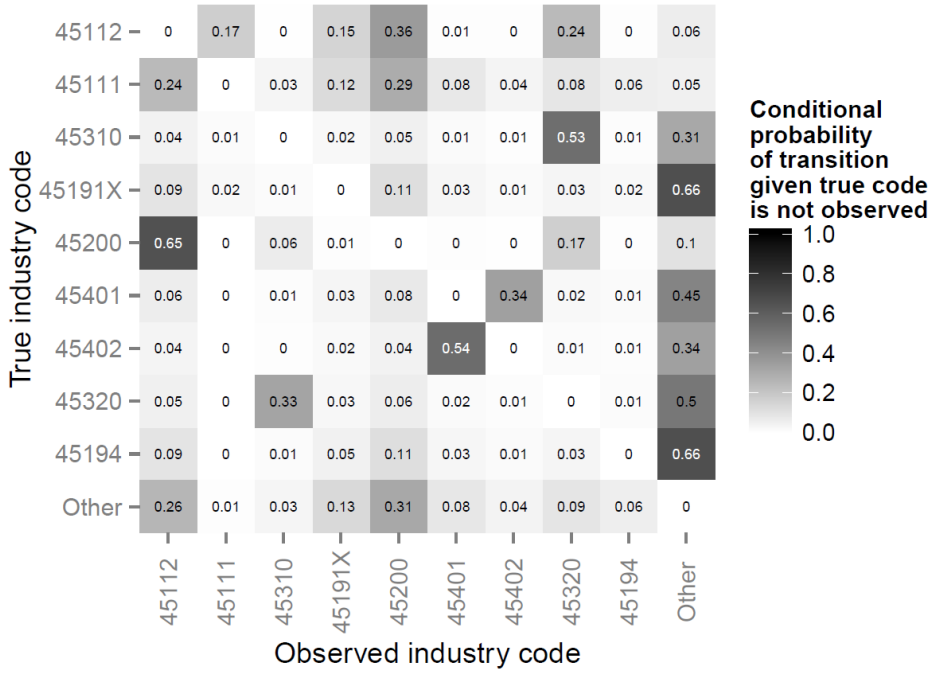

Figure 1. Estimated conditional probabilities for the off-diagonal elements of the level matrix (taken from van Delden et al., 2016b). Each row adds up to 1.

Table 4. Estimated unconditional probabilities and expected numbers of units with true industry code outside car trade that are observed in a car trade industry.

Observed industry code	45112	45111	45310	45191X	45200	45401	45402	45320	45194	Total
Probability	$1.6 \cdot 10^{-4}$	$3.7 \cdot 10^{-6}$	$2.1 \cdot 10^{-5}$	$8.1 \cdot 10^{-5}$	$1.9 \cdot 10^{-4}$	$5.1 \cdot 10^{-5}$	$2.5 \cdot 10^{-5}$	$5.5 \cdot 10^{-5}$	$3.7 \cdot 10^{-5}$	$6.3 \cdot 10^{-4}$
Expected number	168.2	3.9	22.3	85.3	202.5	54.0	25.8	58.5	38.5	659.1

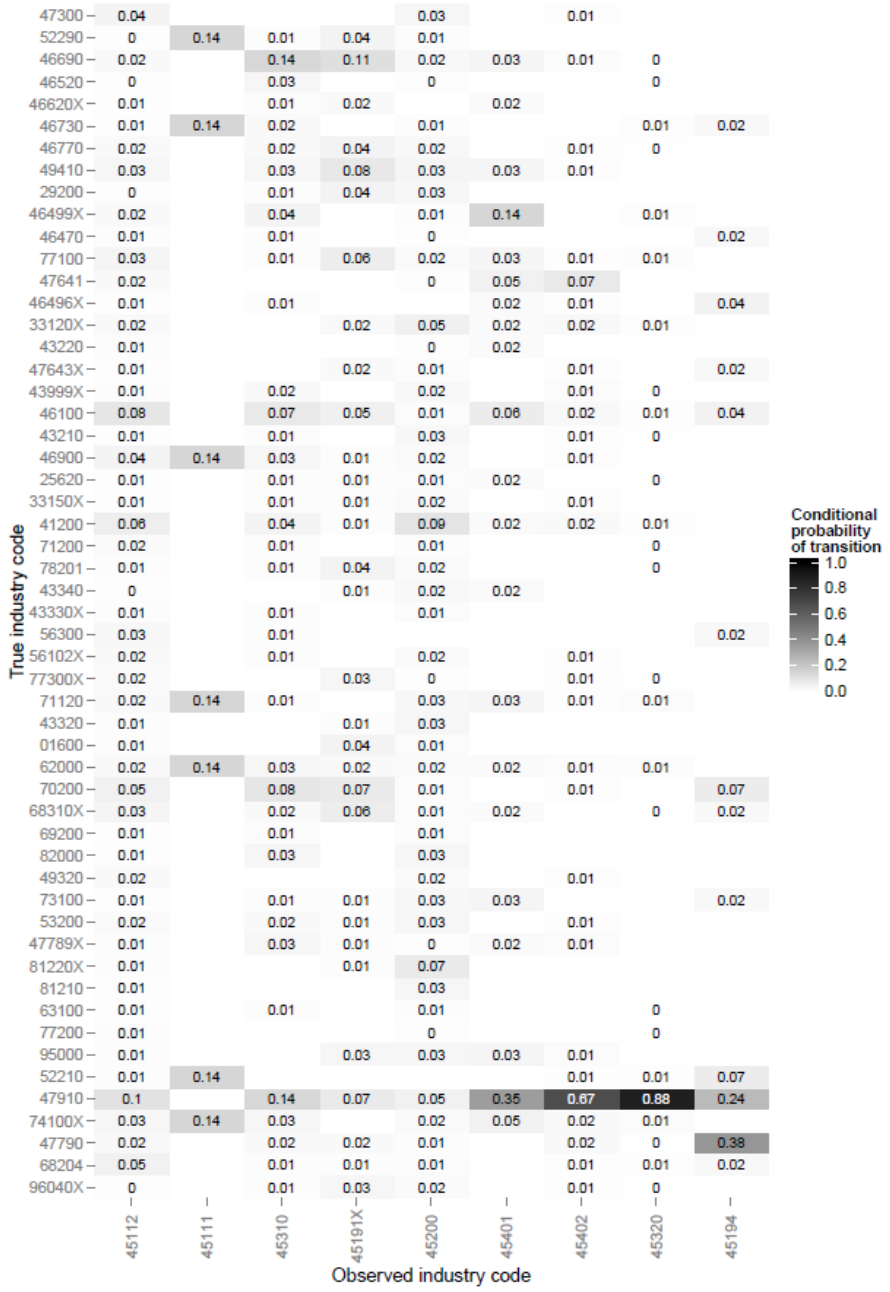

Figure 2. Relative frequencies of the "erroneously included" units observed in car trade by true industry code (taken from van Delden et al., 2015). Each column adds up to 1. The industries outside car trade are in descending order of average turnover.

Secondly, we computed the average number of units in each industry as observed in the GBR, across the eight quarters in 2014 and 2015; we denote these numbers as N_g. In particular, the average number of units outside car trade ($\sum_{g \in \mathcal{M} \setminus \mathcal{M}_T} N_g$) was 1,053,863. Multiplying this number of units by the above unconditional probabilities in the last row of the collapsed level matrix, we obtained expected numbers of non-car trade units that are erroneously included in each car trade industry. The results are shown in Table 4.

Finally, we need to compute, for *specific* true industry codes outside car trade, the probabilities that they are observed in a car trade industry. In theory we should compute these probabilities for each of the industry

codes outside car trade. Based on the data of the yearly transitions in the GBR, van Delden et al. (2015) found a set of 54 industries that covered, for each of the car trade industries, at least 70 per cent of the total number of yearly outflowing units to industries outside car trade. For ease of computation, we restrict the attention to this subset of non-target industries.

The relative contributions among the 54 industries to the total number of erroneously included units in car trade industries are given in Figure 2. The numbers in this figure were estimated from the observed yearly transitions in the GBR; see van Delden et al. (2015) for more details. To illustrate the use of the figure, consider non-car trade units that are misclassified in industry 45112. According to Table 4, we expect 168.2 units with true industry code outside car trade to be misclassified in this car trade industry. The first column of Figure 2 specifies the relative distribution of these 168.2 units across the non-car trade industries. Thus, about 0.04 × 168.2 ≈ 7 units in true industry 47300 are expected to be misclassified in industry 45112, etc.

Finally, these industry-specific expected numbers of misclassifications (say n_{gh}, with $g \in \mathcal{M} \backslash \mathcal{M}_T$ and $h \in \mathcal{M}_T$) can be translated into classification error probabilities by dividing them by the total number of units in each industry (average over 2014 and 2015): $p_{gh}^{OL} = n_{gh}/N_g$, for $g \in \mathcal{M} \backslash \mathcal{M}_T$ and $h \in \mathcal{M}_T$. Note that, for the classification error probabilities in the "erroneously included" part of the level matrix, we did not distinguish between different probability classes.

5.2.2 Defining the $\mathcal{G}_h$ groups

For each of the five probability classes we have obtained a level matrix $\mathbf{P}_c^{OL}$. This matrix contains the probabilities $p_{ghc}^{OL} = P\left(\hat{s}_i^q = h \middle| s_i^q = g\right)$ for units i in probability class c, for all combinations of g and h. To simplify the bias and variance computations in practice, we now select a limited number of special cases p_{ghc}^{OL} within $\mathbf{P}_c^{OL}$ that we want to consider separately. For all other cases we work with an average probability; cf. assumption 2'' in Section 3.4. In terms of the notation in Section 3.4, for each observed industry code $h \in \mathcal{M}_T$ we select a group $\mathcal{G}_h$ that contains the 'special' true industry codes. Recall that for simplicity we use the same groups $\mathcal{G}_h$ for all probability classes. Therefore we include industries as special cases if they appear to be important in at least one of the probability classes.

Let $h \in \mathcal{M}_T$ be one of the car trade industry codes. To select $\mathcal{G}_h$, we focus on two properties:

- the magnitude of a probability p_{ghc}^{OL};
- the expected number of units with true industry code g that are misclassified in industry h.

We define criteria to find the cases with the largest probabilities and the largest expected numbers of misclassified units (see below). An element of the level matrix is selected as a special case if it satisfies *at least* one of these criteria (for at least one probability class).

We apply these criteria to the estimated probabilities from the previous subsection. For the first property, we use $p_{gh,max}^{OL} = \max_c p_{ghc}^{OL}$ and select as special cases those combinations with $p_{gh,max}^{OL} > 0.05$:

$$\mathcal{G}_{h1} = \left\{ g \middle| p_{gh,max}^{OL} > 0.05 \right\}.$$

The criterion for the second property requires some more explanation. To obtain an expected number of units with true industry code g that are observed in industry h, we multiply each p_{ghc}^{OL} by the number of units in industry g. Let f_{ghc} denote a normalised version of these expected numbers, with $\sum_{g=1}^{M} f_{ghc} = 1$ for each h and each probability class c. Finally, we compute $f_{gh,max} = \max_c f_{ghc}$ for each $h \in \mathcal{M}_T$. For a given $h \in \mathcal{M}_T$, large values of $f_{gh,max}$ correspond to true industry codes g for which relatively many units are expected to be (mis)classified in target industry h (for at least one of the probability classes). Therefore, we define

$$\mathcal{G}_{h2} = \{g | f_{gh,max} > \alpha\},$$

for some chosen cut-off value α. By decreasing α, more elements of the level matrix will be selected a special cases. For the results to be discussed below, we have tested the values $\alpha = 5\%$, $\alpha = 2\%$ and $\alpha = 1\%$. Finally, we combine the two criteria to obtain $\mathcal{G}_h = \mathcal{G}_{h1} \cup \mathcal{G}_{h2}$.

Table 5 displays results of the application of the two criteria to our case study on car trade. We have omitted the industry codes g for which all $f_{gh,max}$ were smaller than 1%. The first nine rows in the table correspond to industry codes within car trade. In addition, there are four industry codes outside car trade where units have a relatively large probability to be misclassified into (at least one of) the car trade industries.

The results in this table illustrate that the two criteria are complementary to some extent. For instance, for target industry 45112 the second criterion selects two industries for all above choices of α (45112 and 45200), and the first criterion selects two additional industries (45111 and 45401). The probabilities in the level matrix for the latter two cases are relatively large, but these industries contain few units, so their contributions to 45112 in terms of $f_{gh,max}$ are relatively small.

It is further interesting to note that there are many combinations (g, h) with g and h both within car trade that are not considered 'special' by the above selection criteria. Moreover, the number of 'special' cases to be included in $\mathcal{G}_h$ varies considerably between target industries: from three special cases for target industries 45310 and 45402 to twelve for target industry 45401 (with $\alpha = 1\%$).

Table 5. Selection of groups based on the two criteria. Rows indicate true industry codes g; columns indicate observed industry codes h. Cells with $p^{OL}_{gh,max} > 0.05$ are highlighted in red (criterion 1). Cell values denote $f_{gh,max}$ (criterion 2), with value ranges indicated by cell colours: less than 1% (no colour); 1–2% (faint blue); 2–5% (light blue); 5% or more (dark blue). Only rows with at least one value of $f_{gh,max}$ above 1% are shown.

Industry code	45111	45112	45191X	45194	45200	45310	45320	45401	45402
45111	0.644	0.002	0.015	0.030	0.008	0.003	0.016	0.033	0.005
45112	0.932	0.985	0.284	0.037	0.144	0.003	0.424	0.039	0.006
45191X	0.042	0.003	0.890	0.046	0.012	0.004	0.015	0.049	0.008
45194	0.000	0.001	0.007	0.891	0.003	0.001	0.004	0.014	0.002
45200	0.000	0.058	0.009	0.017	0.939	0.054	0.195	0.018	0.003
45310	0.013	0.001	0.009	0.017	0.004	0.979	0.196	0.018	0.003
45320	0.001	0.002	0.015	0.029	0.008	0.124	0.784	0.031	0.005
45401	0.001	0.001	0.010	0.019	0.005	0.002	0.007	0.849	0.118
45402	0.000	0.001	0.005	0.010	0.002	0.001	0.003	0.485	0.974
46100	0.000	0.001	0.003	0.006	0.000	0.001	0.000	0.010	0.001
46499X	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.023	0.000
47790	0.000	0.000	0.001	0.051	0.000	0.000	0.000	0.000	0.000
47910	0.000	0.001	0.005	0.033	0.002	0.002	0.068	0.056	0.016

Having selected the groups $\mathcal{G}_h$, the classification error probabilities p^{OL}_{ghc} for $g \in \mathcal{G}_h$ follow directly from the estimated probabilities in Section 5.2.1. We also compute $\tilde{p}^{OL}_{(-\mathcal{G}_h)hc}$ from these estimated probabilities by means of expression (24). In this application, we restrict the latter computation to the nine car trade industries and the 54 non-car trade industries shown in Figure 2.

As an illustration, Table 6 shows the original level matrix $\mathbf{P}^{OL}_c$ for the second probability class (simple units with less than 10 EMP that consist of at least three legal units) and Table 7 shows the corresponding matrix $\widetilde{\mathbf{P}}^{OL}_c$ for

the largest set $\mathcal{G}_h$ (with $\alpha = 1\%$). The cells that contain approximate probabilities based on $\tilde{p}^{OL}_{(-\mathcal{G}_h)hc}$ are shaded in grey.

Table 6. Original estimated level matrix for simple units with less than 10 EMP that consist of at least three legal units.

Industry code	45111	45112	45191X	45194	45200	45310	45320	45401	45402
45111	0.024642	0.238084	0.120697	0.054457	0.286656	0.031524	0.082754	0.076471	0.036567
45112	0.021085	0.875761	0.018719	0.000567	0.044457	0.000328	0.030438	0.000797	0.000381
45191X	0.011103	0.043434	0.531617	0.009935	0.052296	0.005751	0.015097	0.013951	0.006671
45194	0.000322	0.044498	0.022558	0.518824	0.053576	0.005892	0.015467	0.014292	0.006834
45200	0.000026	0.169676	0.001792	0.000809	0.740866	0.016527	0.043385	0.001135	0.000543
45310	0.002787	0.010903	0.005527	0.002494	0.013127	0.748324	0.133810	0.003502	0.001675
45320	0.000310	0.042890	0.021743	0.009810	0.051640	0.272587	0.164332	0.013776	0.006588
45401	0.000387	0.053426	0.027084	0.012220	0.064325	0.007074	0.018570	0.145688	0.289738
45402	0.000069	0.009538	0.004835	0.002182	0.011484	0.001263	0.003315	0.147042	0.727693
46100	0.000000	0.001105	0.000335	0.000134	0.000201	0.000129	0.000026	0.000269	0.000046
46499X	0.000000	0.000635	0.000000	0.000000	0.000340	0.000164	0.000065	0.001537	0.000000
47790	0.000000	0.001210	0.000518	0.004411	0.000519	0.000125	0.000050	0.000000	0.000134
47910	0.000000	0.000541	0.000196	0.000309	0.000308	0.000101	0.001693	0.000619	0.000568

Table 7. Approximated level matrix for simple units with less than 10 EMP that consist of at least three legal units ($\mathcal{G}_h$ based on $\alpha = 1\%$). Grey cells indicate non-special cases.

Industry code	45111	45112	45191X	45194	45200	45310	45320	45401	45402
45111	0.024642	0.238084	0.120697	0.054457	0.286656	0.000929	0.082754	0.076471	0.001027
45112	0.021085	0.875761	0.018719	0.000567	0.044457	0.000929	0.030438	0.000797	0.001027
45191X	0.011103	0.003225	0.531617	0.009935	0.052296	0.000929	0.015097	0.013951	0.001027
45194	0.000034	0.003225	0.001510	0.518824	0.053576	0.000929	0.000697	0.014292	0.001027
45200	0.000034	0.169676	0.001510	0.000809	0.740866	0.016527	0.043385	0.001135	0.001027
45310	0.002787	0.003225	0.001510	0.002494	0.001358	0.748324	0.133810	0.003502	0.001027
45320	0.000034	0.003225	0.021743	0.009810	0.051640	0.272587	0.164332	0.013776	0.001027
45401	0.000034	0.053426	0.001510	0.012220	0.064325	0.000929	0.000697	0.145688	0.289738
45402	0.000034	0.003225	0.001510	0.000124	0.001358	0.000929	0.000697	0.147042	0.727693
46100	0.000034	0.003225	0.001510	0.000124	0.001358	0.000929	0.000697	0.000269	0.001027
46499X	0.000034	0.003225	0.001510	0.000124	0.001358	0.000929	0.000697	0.001537	0.001027
47790	0.000034	0.003225	0.001510	0.004411	0.001358	0.000929	0.000697	0.000112	0.001027
47910	0.000034	0.003225	0.001510	0.000309	0.001358	0.000929	0.001693	0.000619	0.000568
other (50)	0.000034	0.003225	0.001510	0.000124	0.001358	0.000929	0.000697	0.000112	0.001027

5.3 Input parameters for year-on-year growth rates

In Section 5.2 we described how we obtained all input parameters to evaluate the analytical expressions for the quarter-on-quarter growth rates. For the analytical expressions for the year-on-year growth rates, we will limit ourselves here to describing which additional input parameters will be needed and how we are planning to estimate them.

To apply formula (73) for the estimated bias and formula (74) for the estimated variance of a year-on-year growth rate, the following input is needed in addition to the parameters in Section 5.2:

- for each true industry code $g \in \{1, \dots, M\}$, the subset $\mathcal{K}_g \subset \{1, \dots, M\}$ of industry codes such that industry g has relatively many yearly transitions to an industry in $\mathcal{K}_g$ (in reality);
- the probabilities (p_{Rc}, p_{Nc}, p_{Sc}) for each probability class c;

- the matrix $\mathbf{R}$ of yearly transition probabilities as observed in the GBR.

Recall from Section 4 that the subsets $\mathcal{T}_h$ that occur in these bias and variance formulas follow directly from the definitions of $\mathcal{G}_h$ and $\mathcal{K}_g$.

The estimation of the probabilities p_{Rc}, p_{Nc}, p_{Sc} and the transition matrix $\mathbf{R}$ has been described in van Delden et al. (2016a). To select the subsets $\mathcal{K}_g$, we propose to use the relative numbers of *observed* transitions between industries in the GBR as a proxy for the *true* transitions between industries. That is, we assume that pairs of industries with many observed transitions are likely to also have many true transitions.

5.4 Results

Results on bias and variance estimates for quarter-on-quarter growth rates will be added later (within SGA 1).

6 Discussion

In this paper, analytical approximations have been presented for the bias and variance of industry growth rates when the classification of businesses by industry in a business register is affected by errors. These formulas were based on a model for random classification errors with respect to the true industry code of a unit. The model describes the occurrence of classification errors at a single point in time (in terms of a "level matrix") and their development over time (in terms of a "change matrix"). The model is generic in the sense that it can accommodate classification errors between all domains, with possibly different error probabilities for each pair of domains. More restrictive classification error models, such as the exchangeable-errors model, can be obtained as special cases. Several types of classification error probabilities occur as unknown parameters in the model, which have to be estimated in practice.

NSIs have some flexibility in applying the analytical approach of this paper. The formulas that have been derived above are actually a family of bias and variance approximations, in which the number of parameters can be varied. By increasing the number of distinct parameters, the bias and variance approximations should become more accurate in theory, but it may be difficult or expensive to estimate these parameters in practice. Moreover, the resulting bias and variance estimates may become unstable due to uncertainty in the estimated parameters. Conversely, if the number of distinct parameters is kept small, there may be some loss of theoretical accuracy, but it may be easier to obtain accurate estimates of the required parameters and the resulting bias and variance estimates may be more stable.

In the application of Section 5, the required parameter estimates were derived from estimated classification error probabilities in previous studies. These estimates were based on an audit sample of businesses, historical data on observed transitions between industries in the GBR, and interviews with experts. A question for future research is whether the audit sample could be avoided, for instance, by collecting paradata on misclassifications as part of the editing process in regular production. So far, we have focussed on a small subset of the NACE domain (car trade, consisting of nine target industries). In order to be able to extend our approach to larger sections of the NACE domain, audit samples should be avoided when possible.

The above-mentioned choice on the number of required model parameters depends in particular on the choice of probability classes and of groups $\mathcal{G}_h$ and $\mathcal{K}_g$ that occur in the bias and variance formulas. In our application, we re-used the probability classes from a previous study. The groups were defined afterwards, based on the initially estimated probabilities. An alternative approach could be to define the groups beforehand – e.g., using expert knowledge and/or historical GBR data – and to incorporate these groups explicitly in the estimation procedure for classification error probabilities. For instance, the groups $\mathcal{G}_h$ could be used in a log-linear model

for the level matrix (cf. van Delden et al., 2016b). The latter approach seems to be more natural for new applications.

In practice, the bias and variance estimators in Sections 3.6 and 4.6 involve replacing unknown domain totals by their observed versions which are known to be biased due to classification errors. Therefore, in general, these estimators of the bias and variance are biased themselves. It is not clear yet whether this bias is large enough to be problematic for practical applications. Furthermore, the accuracy of the estimated bias and variance also depends on the accuracy of the estimated classification error probabilities. It is not currently known how sensitive the bias and variance estimates are to small changes in these parameters. Both issues could be investigated in a simulation study.

The focus of this paper has been on growth rates by industry code, but the same analytical approach could be applied to other domain statistics that are affected by errors in the assignment of units to domains. The only restriction is that the target parameter can be written as a function of domain totals that can be approximated by a Taylor series. Consider a target parameter of the general form $\theta_h = f(X_{1h}, \dots, X_{Qh})$ where X_{qh} denotes the total of a variable x_q for domain h and f is a function with first and second-order partial derivatives. The target parameter is estimated by $\hat{\theta}_h = f(\hat{X}_{1h}, \dots, \hat{X}_{Qh})$, with $\hat{X}_{qh}$ the observed version of X_{qh}. Then, in analogy with the derivation in the appendix, it can be shown that

$$\begin{aligned} B(\hat{\theta}_h) &\approx \frac{1}{2}\sum_{q=1}^{Q}\sum_{r=1}^{Q}\frac{\partial^2 f}{\partial x_q \partial x_r}(\mu_{1h}, \dots, \mu_{Qh})\, C(\hat{X}_{qh}, \hat{X}_{rh}) + f(\mu_{1h}, \dots, \mu_{Qh}) - \theta_h, \\ V(\hat{\theta}_h) &\approx \sum_{q=1}^{Q}\sum_{r=1}^{Q}\frac{\partial f}{\partial x_q}(\mu_{1h}, \dots, \mu_{Qh})\frac{\partial f}{\partial x_r}(\mu_{1h}, \dots, \mu_{Qh})\, C(\hat{X}_{qh}, \hat{X}_{rh}), \end{aligned} \tag{75}$$

with $\mu_{qh} = E(\hat{X}_{qh})$. The covariances $C(\hat{X}_{qh}, \hat{X}_{rh})$ and expectations $E(\hat{X}_{qh})$ can be evaluated under a classification error model. The precise form of the resulting bias and variance approximations depends on the assumptions in this model. A possible example is that one is interested in statistics on persons classified by education level. An interesting target parameter could be the monthly income per number of hours worked, for each education level. Suppose that administrative data on education, income (x_1) and number of hours worked (x_2) are available for all persons, but that the observed education levels are prone to classification errors. The bias and variance of the monthly income per number of hours worked for persons with observed education level h follow from (75) with $\theta_h = X_{1h}/X_{2h}$. An extension of this target parameter could be to consider the difference in income per number of hours worked between males and females.

To estimate the bias and variance of growth rates, one could also use the bootstrap approach of van Delden et al. (2016a), with the same set of parameters of the classification error model as input. In comparison to the analytical approach, the bootstrap has the disadvantage that it requires more effort in terms of programming and computational work. Furthermore, it is more difficult to disentangle how different bias and variance components contribute to the overall accuracy. On the other hand, the bootstrap has the advantage that the same approach can be applied to many other statistics than growth rates with hardly any modification, whereas the corresponding analytical approximations would have to be derived separately from (75) for each statistic. Concerning the bias in the accuracy estimates, this issue occurs in the same way in both the bootstrap and the analytical approach (van Delden et al., 2016b).

Having estimated the accuracy of estimated growth rates due to classification errors in actual statistical production, one may find that the accuracy is insufficient for some industries. It is then natural to wonder how the accuracy could be improved, i.e., how the effect of classification errors on those industries could be reduced. This problem was examined for turnover levels by van Delden et al. (2015, 2016b). They simulated the

effect of an additional editing effort to correct individual classification errors, where the number of units to be edited was either distributed evenly across target industries or assigned proportionally to the estimated mean squared error per industry. It was found that this additional editing effort did not always improve the accuracy of estimated turnover levels for each industry. In particular, a proportional assignment of units does not necessarily work well, because the error in a domain total is also affected by units that belong to that domain but are misclassified in other domains. More research is needed to develop an efficient and effective strategy for improving the accuracy of domain estimates under classification errors.

References

J. Burger, A. van Delden, and S. Scholtus (2015). Sensitivity of Mixed-Source Statistics to Classification Errors. *Journal of Official Statistics* **31**, 489–506.

R. Chambers (2009). Regression Analysis of Probability-Linked Data. Report, Official Statistics Research Series, Volume 4, Statistics New Zealand, Wellington.

A. van Delden, S. Scholtus, and J. Burger (2015). Quantifying the Effect of Classification Errors on the Accuracy of Mixed-Source Statistics. Discussion Paper 2015-10, Statistics Netherlands, The Hague and Heerlen. Available at www.cbs.nl (retrieved: 7 April 2017).

A. van Delden, S. Scholtus, and J. Burger (2016a). Exploring the Effect of Time-Related Classification Errors on the Accuracy of Growth Rates in Business Statistics. Paper presented at the ICES V conference, 21–24 June 2016, Geneva.

A. van Delden, S. Scholtus, and J. Burger (2016b). Accuracy of Mixed-Source Statistics as Affected by Classification Errors. *Journal of Official Statistics* **32**, 619–642.

Eurostat (2008). NACE rev. 2. Statistical Classification of Economic Activities in the European Community. Eurostat Methodologies and Working Papers. Available at http://ec.europa.eu/eurostat (retrieved: 13 April 2017).

J. Neter, E.S. Maynes, and R. Ramanathan (1965). The Effect of Mismatching on the Measurement of Response Error. *Journal of the American Statistical Association* **60**, 1005–1027.

Appendix: Derivation of Expression (1)

In this appendix to Section 2.1, we will derive the approximations to the bias and variance of a generic estimated ratio under classification errors in expression (1).

Bias

To obtain a formula for the approximate bias of the estimated ratio $\hat{R}_h^{q,r}$, we will make use of a second-order Taylor expansion. (At least a second-order Taylor expansion is needed to evaluate the bias of any estimator, because the contributions of the first-order terms are zero; cf. expressions (A2) and (A3) below.) A second-order Taylor expansion of the function $f(u,v) = u/v$ in a neighbourhood of the point (u_0, v_0) yields:

$$\frac{u}{v} \approx \frac{u_0}{v_0}\left\{1 + \frac{1}{u_0}(u-u_0) - \frac{1}{v_0}(v-v_0) + \frac{1}{2}\left[0 - \frac{2}{v_0 u_0}(v-v_0)(u-u_0) + \frac{2}{v_0^2}(v-v_0)^2\right]\right\} \qquad \text{(A1)}$$

$$= \frac{u_0}{v_0}\left\{1 + \frac{u-u_0}{u_0} - \frac{v-v_0}{v_0} + \frac{(v-v_0)^2}{v_0^2} - \frac{(v-v_0)(u-u_0)}{v_0 u_0}\right\}.$$

Using (A1), the second-order Taylor expansion of $\hat{R}_h^{q,r} = \hat{Y}_h^q / \hat{Y}_h^r$ in a neighbourhood of the point $\left(E(\hat{Y}_h^q), E(\hat{Y}_h^r)\right)$ is given by:

$$\hat{R}_h^{q,r} \approx \frac{E(\hat{Y}_h^q)}{E(\hat{Y}_h^r)}\left\{1 + \frac{\hat{Y}_h^q - E(\hat{Y}_h^q)}{E(\hat{Y}_h^q)} - \frac{\hat{Y}_h^r - E(\hat{Y}_h^r)}{E(\hat{Y}_h^r)} + \frac{\left[\hat{Y}_h^r - E(\hat{Y}_h^r)\right]^2}{\left[E(\hat{Y}_h^r)\right]^2} - \frac{\left[\hat{Y}_h^r - E(\hat{Y}_h^r)\right]\left[\hat{Y}_h^q - E(\hat{Y}_h^q)\right]}{E(\hat{Y}_h^r)E(\hat{Y}_h^q)}\right\}. \quad \text{(A2)}$$

From (A2), we obtain the following second-order approximation to $E(\hat{R}_h^{q,r})$:

$$\begin{aligned} E(\hat{R}_h^{q,r}) &\approx \frac{E(\hat{Y}_h^q)}{E(\hat{Y}_h^r)}\left\{1 + 0 - 0 + \frac{E\left[\hat{Y}_h^r - E(\hat{Y}_h^r)\right]^2}{\left[E(\hat{Y}_h^r)\right]^2} - \frac{E\left[\hat{Y}_h^r - E(\hat{Y}_h^r)\right]\left[\hat{Y}_h^q - E(\hat{Y}_h^q)\right]}{E(\hat{Y}_h^r)E(\hat{Y}_h^q)}\right\} \\ &= \frac{E(\hat{Y}_h^q)}{E(\hat{Y}_h^r)}\left\{1 + \frac{V(\hat{Y}_h^r)}{\left[E(\hat{Y}_h^r)\right]^2} - \frac{C(\hat{Y}_h^r, \hat{Y}_h^q)}{E(\hat{Y}_h^r)E(\hat{Y}_h^q)}\right\}, \end{aligned} \quad \text{(A3)}$$

where $V(.)$ denotes a variance and $C(.,.)$ denotes a covariance.

To simplify the notation, let $\check{R}_h^{q,r} = E(\hat{Y}_h^q)/E(\hat{Y}_h^r)$. The bias $B(\hat{R}_h^{q,r})$ can now be written as:

$$\begin{aligned} B(\hat{R}_h^{q,r}) &= E(\hat{R}_h^{q,r}) - R_h^{q,r} \\ &\approx \frac{E(\hat{Y}_h^q)}{E(\hat{Y}_h^r)}\left\{\frac{V(\hat{Y}_h^r)}{\left[E(\hat{Y}_h^r)\right]^2} - \frac{C(\hat{Y}_h^r, \hat{Y}_h^q)}{E(\hat{Y}_h^r)E(\hat{Y}_h^q)}\right\} + \frac{E(\hat{Y}_h^q)}{E(\hat{Y}_h^r)} - R_h^{q,r} \\ &= \frac{1}{\left[E(\hat{Y}_h^r)\right]^2}\left\{\check{R}_h^{q,r} V(\hat{Y}_h^r) - C(\hat{Y}_h^r, \hat{Y}_h^q)\right\} + (\check{R}_h^{q,r} - R_h^{q,r}). \end{aligned} \quad \text{(A4)}$$

An interesting alternative formulation of the last line in (A4) in terms of residuals $\hat{Y}_h^q - \check{R}_h^{q,r}\hat{Y}_h^r$ can be obtained by observing that $V(\hat{Y}_h^r) = C(\hat{Y}_h^r, \hat{Y}_h^r)$:

$$B(\hat{R}_h^{q,r}) \approx -\frac{C(\hat{Y}_h^r, \hat{Y}_h^q - \check{R}_h^{q,r}\hat{Y}_h^r)}{\left[E(\hat{Y}_h^r)\right]^2} + (\check{R}_h^{q,r} - R_h^{q,r}). \quad \text{(A5)}$$

Expression (A4) can be refined further by making use of the assumption that classification errors are independent across units (but not for the same units in the different data sets). Recall that $\hat{Y}_h^r, \hat{Y}_h^q$ partly refer to different sets of units. We can write: $\hat{Y}_h^r = \hat{Y}_{hSEP}^r + \hat{Y}_{hOLP}^r$ where $\hat{Y}_{hSEP}^r = \sum_{i \in U^r \setminus U^{r,q}} \hat{a}_{hi}^r y_i^r$ and $\hat{Y}_{hOLP}^r = \sum_{U^{r,q}} \hat{a}_{hi}^r y_i^r$. Likewise, $\hat{Y}_h^q = \hat{Y}_{hSEP}^q + \hat{Y}_{hOLP}^q$ where $\hat{Y}_{hSEP}^q = \sum_{i \in U^q \setminus U^{r,q}} \hat{a}_{hi}^q y_i^q$ and $\hat{Y}_{hOLP}^q = \sum_{U^{r,q}} \hat{a}_{hi}^q y_i^q$.

The variance $V(\hat{Y}_h^r)$ can be written as:

$$V(\hat{Y}_h^r) = V\left(\sum_{i \in U^r} \hat{a}_{hi}^r y_i^r\right) = \sum_{i \in U^r} (y_i^r)^2 V(\hat{a}_{hi}^r), \quad \text{(A6)}$$

where we used the assumption that the classification errors are independent across units. Similarly, the covariance $C(\hat{Y}_h^r, \hat{Y}_h^q)$ can be written as:

$$\begin{aligned} C(\hat{Y}_h^r, \hat{Y}_h^q) &= C(\hat{Y}_{hSEP}^r + \hat{Y}_{hOLP}^r, \hat{Y}_{hSEP}^q + \hat{Y}_{hOLP}^q) \\ &= C(\hat{Y}_{hOLP}^r, \hat{Y}_{hOLP}^q) \\ &= C\left(\sum_{i \in U^{r,q}} \hat{a}_{hi}^r y_i^r, \sum_{i \in U^{r,q}} \hat{a}_{hi}^q y_i^q\right) \end{aligned} \quad \text{(A7)}$$

$$= \sum\nolimits_{i \in U^{r,q}} y_i^r y_i^q C\left(\hat{a}_{hi}^r, \hat{a}_{hi}^q\right).$$

Combining (A4), (A6) and (A7), we obtain:

$$B\left(\hat{R}_h^{q,r}\right) \approx \frac{1}{\left[E\left(\hat{Y}_h^r\right)\right]^2}\left\{\check{R}_h^{q,r} \sum\nolimits_{i \in U^r} (y_i^r)^2 V(\hat{a}_{hi}^r) - \sum\nolimits_{i \in U^{r,q}} y_i^r y_i^q C\left(\hat{a}_{hi}^r, \hat{a}_{hi}^q\right)\right\} + \left(\check{R}_h^{q,r} - R_h^{q,r}\right), \tag{A8}$$

where furthermore $E\left(\hat{Y}_h^r\right) = \sum_{i \in U^r} y_i^r E(\hat{a}_{hi}^r)$ and $\check{R}_h^{q,r} = \sum_{i \in U^q} y_i^q E\left(\hat{a}_{hi}^q\right) / \sum_{i \in U^r} y_i^r E(\hat{a}_{hi}^r)$. This yields the bias approximation in (1).

Variance

For the variance we make use of the first-order Taylor expansion of a ratio. The first-order Taylor expansion of the function $f = u/v$ in a neighbourhood of the point (u_0, v_0) yields:

$$\frac{u}{v} \approx \frac{u_0}{v_0}\left\{1 + \frac{u}{u_0} - \frac{v}{v_0}\right\} = \frac{u_0}{v_0} + \frac{1}{v_0}\left(u - \frac{u_0}{v_0} v\right) \tag{A9}$$

where in the first line we used expression (A1). Using (A9), $V\left(\hat{R}_h^{q,r}\right)$ can be approximated as:

$$\begin{aligned}
V\left(\hat{R}_h^{q,r}\right) &\approx V\left\{\frac{E\left(\hat{Y}_h^q\right)}{E\left(\hat{Y}_h^r\right)} + \frac{1}{E\left(\hat{Y}_h^r\right)}\left[\hat{Y}_h^q - \frac{E\left(\hat{Y}_h^q\right)}{E\left(\hat{Y}_h^r\right)}\hat{Y}_h^r\right]\right\} \\
&= \frac{1}{\left[E\left(\hat{Y}_h^r\right)\right]^2} V\left(\hat{Y}_h^q - \check{R}_h^{q,r}\hat{Y}_h^r\right) \\
&= \frac{1}{\left[E\left(\hat{Y}_h^r\right)\right]^2}\left\{V\left(\hat{Y}_h^q\right) + \left(\check{R}_h^{q,r}\right)^2 V\left(\hat{Y}_h^r\right) - 2\check{R}_h^{q,r} C\left(\hat{Y}_h^r, \hat{Y}_h^q\right)\right\} \\
&= \frac{1}{\left[E\left(\hat{Y}_h^r\right)\right]^2}\left\{V\left(\hat{Y}_h^q\right) + \left(\check{R}_h^{q,r}\right)^2 V\left(\hat{Y}_h^r\right) - 2\check{R}_h^{q,r} C\left(\hat{Y}_{hOLP}^r, \hat{Y}_{hOLP}^q\right)\right\} \\
&= \frac{1}{\left[E\left(\hat{Y}_h^r\right)\right]^2}\left\{\sum\nolimits_{i \in U^q} (y_i^q)^2 V\left(\hat{a}_{hi}^q\right) + \left(\check{R}_h^{q,r}\right)^2 \sum\nolimits_{i \in U^r} (y_i^r)^2 V(\hat{a}_{hi}^r)\right. \\
&\qquad \left. - 2\check{R}_h^{q,r} \sum\nolimits_{i \in U^{r,q}} y_i^r y_i^q C\left(\hat{a}_{hi}^r, \hat{a}_{hi}^q\right)\right\},
\end{aligned} \tag{A10}$$

where in the third and fourth line we used (A6) and (A7). This yields the variance approximation in (1).