Discussion paper

Evaluating the accuracy of growth rates in the presence of classification errors

Sander Scholtus
Arnout van Delden
Joep Burger

October 2019

Content

Summary

Producing reliable, undisputed statistical figures is the backbone of national statistical institutes. When administrative data are used in the production of statistics, the accuracy is often mainly determined by so-called non-sampling errors. For statistics that are published by domain, classification errors in the domain codes are an important example of a non-sampling error. Quantifying the effect of non-sampling errors on statistics is often difficult in practice.
In previous work we have developed approaches to evaluate the effect of classification errors on the bias and variance of estimated domain totals, both analytically and by means of a bootstrap method. In this paper, we extend both approaches to estimated growth rates. Here, a more complicated model for classification errors is needed to account for the relation between errors made at different points in time. An important practical question is how to estimate the unknown parameters of this classification error model. We illustrate this in detail for a case study on the effect of errors in the classification of businesses by industry (NACE code) on the bias and variance of quarterly turnover growth rates for the Dutch car trade sector.

Keywords

accuracy, non-sampling error, analytical approximation, bootstrap, NACE code

The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Statistics Netherlands.

Part of the work underlying this paper has been carried out as part of the ESSnet on Quality of Multisource Statistics (SGA 07112.2015.015-2015.705 and SGA 07112.2016.019-2017.144), funded by the European Commission.

The authors would like to thank Harm Jan Boonstra, Danutė Krapavickaitė, and Ton de Waal for their comments on earlier versions of this paper.

1. Introduction

National statistical institutes (NSIs) often publish statistics that involve a classification of the population units into some domains. As a running example in this paper we will consider business statistics on turnover that are published by industry. (Some other examples will be mentioned in Section 10.) In these statistics, each business is classified into one industry according to its main economic activity. In European countries, since 2008, the NACE code Rev. 2 classification is used (Eurostat, 2008). Assigning a single NACE code to each business 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. Throughout this report, we consider the true NACE code of a statistical unit to be the code that would be obtained when these rules are applied correctly, using error-free input data.

Many NSIs have a general business register (GBR) with an enumeration of all statistical business units. For each unit the values of a few background variables are given, including 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, i.e., classification errors occur. A number of reasons can be given for this. 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 change their activities but fail to report this to the CoC. Fourthly, the rules to derive the main economic activity might not have been applied correctly, for instance because the available information on the relative importance of the multiple activities of a statistical unit was not accurate enough.

A natural question is how classification errors in the observed data affect the accuracy of published statistics. Van Delden et al. (2015, 2016b) developed methods to evaluate the effect of classification errors on cross-sectional statistics. As an application, they investigated the effect of NACE code errors on estimated total turnover levels by industry. They considered both an analytical approach and a bootstrap approach to estimate the bias and variance due to classification errors. In this report, we will extend both approaches to evaluate the effect of classification errors on growth rates. For NSIs, quarterly turnover growth rates are an important short-term indicator for the economic business cycle.

In this report, we begin by developing an analytical approach. In Section 2, generic analytical approximations to the bias and variance of a growth rate are derived based on a Taylor series. Specific versions of these expressions are then worked out for quarter-on-quarter growth rates and for year-on-year growth rates; as will be explained there, the practical context for these two cases is different at Statistics Netherlands. The special case of levels and growth rates for a binary classification (two classes) is treated in Section 3, as relatively simple analytical results can be obtained for this case which arises often in practice. In the remainder of this report, we focus on problems where the classification has a (potentially much) larger number of classes. We will introduce several simplifying assumptions to derive approximate expressions that can be applied more easily in practical situations than the generic expressions of Section 2. These assumptions do not necessarily simplify the mathematical content of these expressions, but they do lead to expressions that are easier to compute. We 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 4), before proceeding to the more complicated case of year-on-year growth rates (Section 5). Readers who are only interested in the main ideas of the proposed method may wish to skip some of the detailed derivations in these two sections.

As an alternative to the analytical approach, a bootstrap approach is developed in Section 6 to estimate the bias and variance of growth rates as affected by classification errors. For practical applications, both the analytical and the bootstrap approach require the specification of a classification error model. A particular model that can be used in conjunction with both approaches is introduced in Section 7. In Section 8, a small simulation study is conducted to verify whether bias and variance estimates obtained by the analytical approach and bootstrap approach are in agreement. Subsequently, Section 9 discusses an application of both approaches to a case study on real data: evaluating the accuracy of quarterly turnover growth rates of Dutch car trade industries due to NACE code errors in the Dutch GBR. Finally, some conclusions and questions for further research are discussed in Section 10.

2. Accuracy of growth rates

2.1 Notation and set-up

Consider a population of units that is divided into strata, where the total set of strata is denoted by $\mathcal{H}_{\text{full}} = \{1, \dots, M\}$. In this paper, we will focus in particular on the case where a population of businesses is divided into industries based on economic activity (NACE code) as derived in a GBR. Moreover, we will consider dynamic populations that change over time, both in terms of births and deaths and in terms of units whose stratum changes over time. Let $U^{t,q}$ denote the population in quarter q of year t. The starting year of the computations is denoted by $t = T_0$ and the following years by $T_0 + 1, T_0 + 2$, etc.

For each quarter q, each active unit (business) i has an unknown true industry code $s_i^{t,q} = g$ and an observed industry code $\hat{s}_i^{t,q} = h$, where $g, h \in \mathcal{H}_{\text{full}}$. At Statistics Netherlands, the true and observed industry codes are kept constant during a year so that the same frame can be used for intra-annual and annual statistics. This is called the *coordinated industry code*. Between 31 December of year $t - 1$ and 1 January of year t, the true and observed industry codes are updated for all units that are present in both quarter 4 of year $t - 1$ and quarter 1 of year t (continuing units). We stress that, for quarters within the same year, $s_i^{t,q}$ and $\hat{s}_i^{t,q}$ do not change.

Due to classification errors, some of the observed industry codes may differ from the true ones. Those classification errors may affect the publication figures. In this paper, we consider the simple case where classification errors are the only errors that occur. In particular, we assume that the target variable (turnover) is observed for all units in the population. This can happen, e.g., when administrative data are used (cf. Section 9). In this paper, we do not consider the effect of measurement errors in the observed target variable.

We are interested in changes in quarterly turnover per industry. First denote the true total turnover for industry $h \in \mathcal{H}_{\text{full}}$ in quarter q of year t by $Y_h^{t,q} = \sum_{i \in U^{t,q}} a_{hi}^{t,q} y_i^{t,q}$, where $y_i^{t,q}$ denotes the turnover of unit i in quarter q of year t and (for each h)

$$a_{hi}^{t,q} = I(s_i^{t,q} = h) = \begin{cases} 1 & \text{if } s_i^{t,q} = h, \\ 0 & \text{if } s_i^{t,q} \neq h. \end{cases}$$

In practice, $Y_h^{t,q}$ is estimated by $\hat{Y}_h^{t,q} = \sum_{i \in U^{t,q}} \hat{a}_{hi}^{t,q} y_i^{t,q}$, with $\hat{a}_{hi}^{t,q} = I(\hat{s}_i^{t,q} = h)$. In the remainder of this paper we will drop t and use a single time index q in the notation unless both indices are needed to avoid confusion.

We denote the turnover ratio between quarter q and an earlier quarter $q - u$ as $G_h^{q,q-u} = Y_h^q / Y_h^{q-u}$, where $u = 1$ gives the ratio with respect to the previous quarter and $u = 4$ the ratio with respect to the same quarter in the previous year. $G_h^{q,q-u}$ is estimated as $\hat{G}_h^{q,q-u} = \hat{Y}_h^q / \hat{Y}_h^{q-u}$. The corresponding growth rates (relative changes) are expressed as $g_h^{q,q-u} = G_h^{q,q-u} - 1$ and $\hat{g}_h^{q,q-u} = \hat{G}_h^{q,q-u} - 1$.

We would like to assess the bias and variance of $\hat{g}_h^{q,q-u}$ as an estimator for $g_h^{q,q-u}$:

$$B\left(\hat{g}_h^{q,q-u}\right) = E\left(\hat{g}_h^{q,q-u}\right) - g_h^{q,q-u},$$
$$V\left(\hat{g}_h^{q,q-u}\right) = E\left[\left(\hat{g}_h^{q,q-u} - E\left(\hat{g}_h^{q,q-u}\right)\right)^2\right].$$

In this paper, we will pursue two different approaches to estimate this bias and variance: using analytical approximations and using a bootstrap method. Both approaches require the specification of a model for the classification errors that occur in the GBR over time. In the present section, we begin by deriving general formulae for the approximate bias and variance of $\hat{g}_h^{q,q-u}$, making only minimal assumptions about the error model. Specific approximations will be derived later for quarter-on-quarter and year-on-year growth rates under certain model assumptions.

2.2 Generic bias and variance approximations for growth rates

Throughout this paper we will consider the true industry codes as fixed and the observed industry codes as stochastic, in line with Kuha and Skinner (1997). Classification errors in the observed industry codes at a given time point are supposed to be independent across units. For a given unit i, the classification errors that occur in $\hat{s}_i^q$ over time may be dependent. In fact, it is likely that classification errors in the GBR are correlated strongly over time, since most units are not monitored actively and remain assigned to the same industry until new information arrives. The model that will be introduced in Section 7 reflects this dependency.

Let U^q and U^{q-u} denote the target populations in quarters q and $q-u$, and let $U_O^{q-u,q} = U^{q-u} \cap U^q$ denote the overlapping part of these populations, i.e., the units that exist in both quarters. Later, we will also use the notation $U_D^{q-u,q} = U^{q-u} \backslash U_O^{q-u,q}$ for units that died between $q-u$ and q and $U_B^{q-u,q} = U^q \backslash U_O^{q-u,q}$ for new-born units between $q-u$ and q. Note that $U^{q-u} = U_O^{q-u,q} \cup U_D^{q-u,q}$ and $U^q = U_O^{q-u,q} \cup U_B^{q-u,q}$.

Under the assumptions made so far, the following approximate expressions may be derived for $B\left(\hat{g}_h^{q,q-u}\right)$ and $V\left(\hat{g}_h^{q,q-u}\right)$:

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

where $C(.,.)$ denotes a covariance. These approximations are based on Taylor series expansions; a full derivation is given in Appendix A. In the remainder of this paper, we will denote the bias and variance approximations to the order used in (1) by $AB\left(\hat{g}_h^{q,q-u}\right)$ and $AV\left(\hat{g}_h^{q,q-u}\right)$, respectively, where the A denotes 'approximate'.

The approximate bias $AB\left(\hat{g}_h^{q,q-u}\right)$ and variance $AV\left(\hat{g}_h^{q,q-u}\right)$ are functions of $V\left(\hat{a}_{hi}^{q-u}\right)$, $V\left(\hat{a}_{hi}^{q}\right)$, $C\left(\hat{a}_{hi}^{q-u},\hat{a}_{hi}^{q}\right)$, $E\left(\hat{a}_{hi}^{q-u}\right)$ 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 quarters are correlated.

For the second term $\breve{G}_h^{q,q-u} - G_h^{q,q-u}$ in $AB\left(\hat{g}_h^{q,q-u}\right)$, it is interesting to note that

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

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 $\breve{G}_h^{q,q-u} = G_h^{q,q-u}$. 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 of $B\left(\hat{g}_h^{q,q-u}\right)$ — between square brackets — in expression (1).

In the next two subsections, we will derive explicit expressions for $AB\left(\hat{g}_h^{q,q-u}\right)$ and $AV\left(\hat{g}_h^{q,q-u}\right)$, based on (1), for the special cases $u = 1$ (quarter-on-quarter growth rates, Subsection 2.3) and $u = 4$ (year-on-year growth rates, Subsection 2.4). Here, we still do not make restrictive assumptions about the classsification error model, other than the above-mentioned convention of coordinated industry codes in the Dutch GBR. The resulting formulae are very general but too complicated for practical use at NSIs. Simplified expressions, based on particular model assumptions, will be derived in Sections 4 and 5.

2.3 Bias and variance approximations for quarter-on-quarter growth rates ($\boldsymbol{u} = 1$)

The case of a quarter-on-quarter growth rate *within a single year* is relatively simple when the industry codes are coordinated. We may then 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 ($i \in U_O^{q-1,q}$).

We suppose that 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 a transition matrix $\mathbf{P}_i^L = (p_{ghi}^L)$, with $p_{ghi}^L = P(\hat{s}_i^{q-1} = h | s_i^{q-1} = g)$. Here, the superscript L stands for *level*, since these probabilities determine in particular the accuracy of the estimated turnover levels $\hat{Y}_h^q$ (van Delden et al., 2016b). Table 1 illustrates the form of the level matrix $\mathbf{P}_i^L$. 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 (yet).

Table 1. Transition probabilities (subscript i omitted) for level matrix $\mathbf{P}_i^L$.

True industry	Observed industry			
	1	2		M
1	p_{11}^L	p_{12}^L	$\cdots$	p_{1M}^L
2	p_{21}^L	p_{22}^L	$\cdots$	p_{2M}^L
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
M	p_{M1}^L	p_{M2}^L	$\cdots$	p_{MM}^L

Define the vectors $\boldsymbol{a}_i^{q-u} = (a_{1i}^{q-u}, \dots, a_{Mi}^{q-u})^T$ and $\hat{\boldsymbol{a}}_i^{q-u} = (\hat{a}_{1i}^{q-u}, \dots, \hat{a}_{Mi}^{q-u})^T$ for $u \in \{0,1,4\}$. For each unit i, these vectors consist of $M - 1$ zeros and 1 one. By analogy to Burger et al. (2015, p. 502), the following properties can be derived for $i \in U^{q-1}$:

$$\begin{aligned} E(\hat{\boldsymbol{a}}_i^{q-1}) &= (\mathbf{P}_i^L)^T \boldsymbol{a}_i^{q-1}, \\ V(\hat{\boldsymbol{a}}_i^{q-1}) &= \mathrm{diag}\left[(\mathbf{P}_i^L)^T \boldsymbol{a}_i^{q-1}\right] - (\mathbf{P}_i^L)^T \,\mathrm{diag}(\boldsymbol{a}_i^{q-1})\, \mathbf{P}_i^L. \end{aligned} \tag{2}$$

In particular, it holds that

$$\begin{aligned} E(\hat{a}_{hi}^{q-1}) &= \sum_{g=1}^{M} a_{gi}^{q-1} p_{ghi}^L, \\ V(\hat{a}_{hi}^{q-1}) &= \sum_{g=1}^{M} a_{gi}^{q-1} p_{ghi}^L - \sum_{g=1}^{M} a_{gi}^{q-1} (p_{ghi}^L)^2 \\ &= \sum_{g=1}^{M} a_{gi}^{q-1} p_{ghi}^L (1 - p_{ghi}^L). \end{aligned} \tag{3}$$

For units that exist in both quarters ($i \in U_O^{q-1,q}$), $\hat{\boldsymbol{a}}_i^q$ has the exact same expectation and variance as $\hat{\boldsymbol{a}}_i^{q-1}$. Furthermore, for these units $C(\hat{a}_{hi}^{q-1}, \hat{a}_{hi}^q) = V(\hat{a}_{hi}^{q-1})$. For units that occur only in U^q, the expressions for $E(\hat{\boldsymbol{a}}_i^q)$ and $V(\hat{\boldsymbol{a}}_i^q)$ are similar to (2), but with $\boldsymbol{a}_i^{q-1}$ replaced by $\boldsymbol{a}_i^q$. From (2) and (3), it follows that

$$\begin{aligned} E\left(\hat{Y}_h^{q-1}\right) &= \sum\nolimits_{i\in U^{q-1}} \left(y_i^{q-1} \sum\nolimits_{g=1}^{M} E\left(\hat{a}_{ghi}^{q-1}\right) \right) \\ &= \sum\nolimits_{i\in U^{q-1}} \left(y_i^{q-1} \sum\nolimits_{g=1}^{M} a_{gi}^{q-1} p_{ghi}^{L} \right), \\ \breve{G}_h^{q,q-1} &= \frac{\sum_{i\in U^{q}}\left(y_i^{q} \sum_{g=1}^{M} a_{gi}^{q} p_{ghi}^{L}\right)}{\sum_{i\in U^{q-1}}\left(y_i^{q-1} \sum_{g=1}^{M} a_{gi}^{q-1} p_{ghi}^{L}\right)}, \end{aligned} \tag{4}$$

and also that

$$\sum_{i\in U^{q-1}} \left(y_i^{q-1}\right)^2 V\left(\hat{a}_{hi}^{q-1}\right) = \sum_{i\in U^{q-1}} \left[\left(y_i^{q-1}\right)^2 \sum_{g=1}^{M} a_{gi}^{q-1} p_{ghi}^{L}\left(1 - p_{ghi}^{L}\right) \right], \tag{5}$$

and

$$\sum_{i\in U_O^{q-1,q}} y_i^{q-1} y_i^{q} C\left(\hat{a}_{hi}^{q-1}, \hat{a}_{hi}^{q}\right) = \sum_{i\in U_O^{q-1,q}} \left[G_i^{q,q-1} \left(y_i^{q-1}\right)^2 \sum_{g=1}^{M} a_{gi}^{q-1} p_{ghi}^{L}\left(1 - p_{ghi}^{L}\right) \right],$$

where $G_i^{q,q-1} = y_i^{q}/y_i^{q-1}$ denotes the individual turnover ratio of unit $i \in U_O^{q-1,q}$.

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

$$\begin{aligned} AB\left(\hat{g}_h^{q,q-1}\right) &= \frac{1}{\left[E\left(\hat{Y}_h^{q-1}\right)\right]^2} \left(B_{hO}^{q,q-1} + B_{hD}^{q,q-1}\right) + \left(\breve{G}_h^{q,q-1} - G_h^{q,q-1}\right), \\ B_{hO}^{q,q-1} &= \sum_{i\in U_O^{q-1,q}} \left[\left(\breve{G}_h^{q,q-1} - G_i^{q,q-1}\right) \left(y_i^{q-1}\right)^2 \sum_{g=1}^{M} a_{gi}^{q-1} p_{ghi}^{L}\left(1 - p_{ghi}^{L}\right) \right], \\ B_{hD}^{q,q-1} &= \breve{G}_h^{q,q-1} \sum_{i\in U_D^{q-1,q}} \left[\left(y_i^{q-1}\right)^2 \sum_{g=1}^{M} a_{gi}^{q-1} p_{ghi}^{L}\left(1 - p_{ghi}^{L}\right) \right], \end{aligned} \tag{6}$$

with $E\left(\hat{Y}_h^{q-1}\right)$ and $\breve{G}_h^{q,q-1}$ as given in (4). Also, recall that $U_D^{q-1,q} = U^{q-1} \backslash U_O^{q-1,q}$. The bias approximation consists of three components: for continuing or overlapping units (subscript O), for 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 (1):

$$
\begin{aligned}
AV\left(\hat{g}_h^{q,q-1}\right) &= \frac{1}{\left[E\left(\hat{Y}_h^{q-1}\right)\right]^2}\left(V_{hO}^{q,q-1} + V_{hD}^{q,q-1} + V_{hB}^{q,q-1}\right),\\
V_{hO}^{q,q-1} &= \sum_{i\in U_O^{q-1,q}} \left[\left(\breve{G}_h^{q,q-1} - G_i^{q,q-1}\right)^2 \left(y_i^{q-1}\right)^2 \sum_{g=1}^{M} a_{gi}^{q-1} p_{ghi}^{L}\left(1 - p_{ghi}^{L}\right)\right]\\
V_{hD}^{q,q-1} &= \left(\breve{G}_h^{q,q-1}\right)^2 \sum_{i\in U_D^{q-1,q}} \left[\left(y_i^{q-1}\right)^2 \sum_{g=1}^{M} a_{gi}^{q-1} p_{ghi}^{L}\left(1 - p_{ghi}^{L}\right)\right],\\
V_{hB}^{q,q-1} &= \sum_{i\in U_B^{q-1,q}} \left[\left(y_i^{q}\right)^2 \sum_{g=1}^{M} a_{gi}^{q} p_{ghi}^{L}\left(1 - p_{ghi}^{L}\right)\right].
\end{aligned}
\tag{7}
$$

In the derivation of the component $V_{hO}^{q,q-1}$ in this expression, we used the fact that continuing units $i \in U_O^{q-1,q}$ occur in all three parts of expression (1) for $AV\left(\hat{g}_h^{q,q-1}\right)$ 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

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

The variance approximation consists of three components: for continuing or overlapping units (subscript O), for dead units (D) and for new-born units (B).

2.4 Bias and variance approximations for year-on-year growth rates ($u = 4$)

We will now consider the case of a growth rate between a quarter and the corresponding quarter of the previous year ($u = 4$). The results in this subsection also apply to a growth rate between the first quarter of a year and the fourth quarter that precedes it, but in that case all instances of $q - 4$ should be read as $q - 1$. In both situations, 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.

In the case of a growth rate that involves a yearly transition, expressions (2) and (3) from Subsection 2.3 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} (new-born units). For continuing units $i \in U_O^{q-4,q}$, the quantities $E\left(\hat{a}_{hi}^q\right)$, $V\left(\hat{a}_{hi}^q\right)$ and $C\left(\hat{a}_{hi}^{q-4}, \hat{a}_{hi}^q\right)$ also depend on the effects of changes in classification errors between $q - 4$ and q.

In addition to the level matrix of Subsection 2.3, we introduce a second transition matrix to describe the distribution of classification errors in $\hat{s}_i^q$ for continuing units: $\mathbf{P}_i^C = \left(p_{gklhi}^C\right)$, with $p_{gklhi}^C = P\left(\hat{s}_i^q = h \middle| s_i^{q-4} = g, s_i^q = k, \hat{s}_i^{q-4} = l\right)$. Here, the superscript C stands for *change*. This matrix describes how the probability of observing a particular code in q for unit i depends on its true and observed codes in

$q-4$ and on its true code in q. We assume that this probability does not depend on any other variables (for instance, true or observed codes at times before $q-4$). Apart from this assumption, the introduction of this matrix entails no loss of generalisation, since each unit i can have its own matrix. Further model assumptions on the change matrix will be introduced in Section 7.

It will be useful below to introduce a special notation for the following probabilities that are derived from a combination of $\mathbf{P}_i^L$ and $\mathbf{P}_i^C$: $\mathbb{p}_{gklhi}^{LC} = p_{gli}^L p_{gklhi}^C$ and $p_{gkhi}^{LC} = \sum_{l=1}^M \mathbb{p}_{gklhi}^{LC}$. For continuing units, the quantity p_{gkhi}^{LC} 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

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

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

$$p_{gkhi}^{LC} = \sum_{l=1}^M \mathbb{p}_{gklhi}^{LC} = P\left(\hat{s}_i^q = h \middle| s_i^{q-4} = g, s_i^q = k\right), \tag{8}$$

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 $\sum_{h=1}^M p_{gkhi}^{LC} = 1$.

With the help of some matrix algebra, the following expressions can be derived for continuing units $i \in U_O^{q-4,q}$ (see Appendix B for details):

$$\begin{aligned}E\left(\hat{a}_{hi}^q\right) &= \sum_{g=1}^M \sum_{k=1}^M a_{gi}^{q-4} a_{ki}^q p_{gkhi}^{LC}, \\ V\left(\hat{a}_{hi}^q\right) &= \sum_{g=1}^M \sum_{k=1}^M a_{gi}^{q-4} a_{ki}^q p_{gkhi}^{LC}\left(1 - p_{gkhi}^{LC}\right), \\ C\left(\hat{a}_{hi}^{q-4}, \hat{a}_{hi}^q\right) &= \sum_{g=1}^M \sum_{k=1}^M a_{gi}^{q-4} a_{ki}^q \left(\mathbb{p}_{gkhhi}^{LC} - p_{ghi}^L p_{gkhi}^{LC}\right).\end{aligned} \tag{9}$$

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

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

and hence by (9) that $C\left(\hat{a}_{hi}^{q-4}, \hat{a}_{hi}^{q}\right) = 0$. However, as noted above, this assumption is not realistic for industry codes in the Dutch GBR.

Using the expressions in (9), we can again evaluate the different components in (1). For the bias, the following results are obtained analogously to (4):

$$
\begin{aligned}
E\left(\hat{Y}_h^{q-4}\right) &= \sum\nolimits_{i\in U^{q-4}} \left(y_i^{q-4} \sum\nolimits_{g=1}^{M} a_{gi}^{q-4} p_{ghi}^{L} \right), \\
E\left(\hat{Y}_h^{q}\right) &= \sum\nolimits_{i\in U_O^{q-4,q}} \left(y_i^{q} \sum\nolimits_{g=1}^{M} \sum\nolimits_{k=1}^{M} a_{gi}^{q-4} a_{ki}^{q} p_{gkhi}^{LC} \right) \\
&\quad + \sum\nolimits_{i\in U_B^{q-4,q}} \left(y_i^{q} \sum\nolimits_{g=1}^{M} a_{gi}^{q} p_{ghi}^{L} \right), \\
\breve{G}_h^{q,q-4} &= \frac{E\left(\hat{Y}_h^{q}\right)}{E\left(\hat{Y}_h^{q-4}\right)}.
\end{aligned} \qquad (10)
$$

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}^{L} 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 (5) and (9):

$$
\begin{aligned}
&\sum_{i\in U^{q-4}} (y_i^{q-4})^2 V\left(\hat{a}_{hi}^{q-4}\right) \\
&\qquad = \sum_{i\in U^{q-4}} \left[\left(y_i^{q-4}\right)^2 \sum_{g=1}^{M} a_{gi}^{q-4} p_{ghi}^{L} \left(1 - p_{ghi}^{L}\right) \right], \\
&\sum\nolimits_{i\in U_O^{q-4,q}} y_i^{q-4} y_i^{q} C\left(\hat{a}_{hi}^{q-4}, \hat{a}_{hi}^{q}\right) \\
&= \sum\nolimits_{i\in U_O^{q-4,q}} \left[G_i^{q,q-4} \left(y_i^{q-4}\right)^2 \sum\nolimits_{g=1}^{M} \sum\nolimits_{k=1}^{M} a_{gi}^{q-4} a_{ki}^{q} \left(\mathbb{p}_{gkhhi}^{LC} - p_{ghi}^{L} p_{gkhi}^{LC} \right) \right],
\end{aligned} \qquad (11)
$$

where $G_i^{q,q-4} = y_i^{q} / y_i^{q-4}$ denotes an individual turnover ratio, as before. Thus, expression (1) for the approximate bias of $\hat{g}_h^{q,q-4}$ becomes:

$$\boxed{\begin{aligned}
AB\left(\hat{g}_h^{q,q-4}\right) &= \frac{1}{\left[E\left(\hat{Y}_h^{q-4}\right)\right]^2}\left(B_{hO}^{q,q-4} + B_{hD}^{q,q-4}\right) + \left(\breve{G}_h^{q,q-4} - G_h^{q,q-4}\right),\\
B_{hO}^{q,q-4} &= \sum_{i\in U_O^{q-4,q}} \left\{\left(y_i^{q-4}\right)^2 \sum_{g=1}^{M} a_{gi}^{q-4}\left[\breve{G}_h^{q,q-4} p_{ghi}^{L}\left(1-p_{ghi}^{L}\right)\right.\right.\\
&\qquad \left.\left. - G_i^{q,q-4}\sum_{k=1}^{M} a_{ki}^{q}\left(\mathbb{p}_{gkhhi}^{LC} - p_{ghi}^{L}p_{gkhi}^{LC}\right)\right]\right\},\\
B_{hD}^{q,q-4} &= \breve{G}_h^{q,q-4} \sum_{i\in U_D^{q-4,q}} \left[\left(y_i^{q-4}\right)^2 \sum_{g=1}^{M} a_{gi}^{q-4} p_{ghi}^{L}\left(1-p_{ghi}^{L}\right)\right],
\end{aligned}} \tag{12}$$

with $E\left(\hat{Y}_h^{q-4}\right)$ and $\breve{G}_h^{q,q-4}$ as given in (10).

For the variance, we need the following expression in addition to (10) and (11), which follows from (9):

$$\begin{aligned}
\sum_{i\in U^q}\left(y_i^q\right)^2 V\left(\hat{a}_{hi}^q\right) &= \sum_{i\in U_O^{q-4,q}}\left[\left(G_i^{q,q-4}y_i^{q-4}\right)^2 \sum_{g=1}^{M}\sum_{k=1}^{M} a_{gi}^{q-4}a_{ki}^{q} p_{gkhi}^{LC}\left(1-p_{gkhi}^{LC}\right)\right]\\
&\quad + \sum_{i\in U_B^{q-4,q}}\left[\left(y_i^q\right)^2 \sum_{g=1}^{M} a_{gi}^{q} p_{ghi}^{L}\left(1-p_{ghi}^{L}\right)\right].
\end{aligned}$$

Here, 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 Subsection 2.3.

Expression (1) for the approximate variance of $\hat{g}_h^{q,q-4}$ now yields:

$$\boxed{\begin{aligned}
AV\left(\hat{g}_h^{q,q-4}\right) &= \frac{1}{\left[E\left(\hat{Y}_h^{q-4}\right)\right]^2}\left(V_{hO}^{q,q-4} + V_{hD}^{q,q-4} + V_{hB}^{q,q-4}\right),\\
V_{hO}^{q,q-4} &= \sum_{i\in U_O^{q-4,q}} \left(y_i^{q-4}\right)^2 \sum_{g=1}^{M} a_{gi}^{q-4}\left\{\left(\breve{G}_h^{q,q-4}\right)^2 p_{ghi}^{L}\left(1-p_{ghi}^{L}\right)\right.\\
&\qquad + \sum_{k=1}^{M} a_{ki}^{q}\left[\left(G_i^{q,q-4}\right)^2 p_{gkhi}^{LC}\left(1-p_{gkhi}^{LC}\right)\right.\\
&\qquad \left.\left. - 2\breve{G}_h^{q,q-4}G_i^{q,q-4}\left(\mathbb{p}_{gkhhi}^{LC} - p_{ghi}^{L}p_{gkhi}^{LC}\right)\right]\right\},\\
V_{hD}^{q,q-4} &= \left(\breve{G}_h^{q,q-4}\right)^2 \sum_{i\in U_D^{q-4,q}}\left[\left(y_i^{q-4}\right)^2 \sum_{g=1}^{M} a_{gi}^{q-4} p_{ghi}^{L}\left(1-p_{ghi}^{L}\right)\right],\\
V_{hB}^{q,q-4} &= \sum_{i\in U_B^{q-4,q}}\left[\left(y_i^q\right)^2 \sum_{g=1}^{M} a_{gi}^{q} p_{ghi}^{L}\left(1-p_{ghi}^{L}\right)\right].
\end{aligned}} \tag{13}$$

3. Special case: two classes

In the next two sections, further approximations will be derived to the general expressions for the approximate bias and variance of growth rates $\hat{g}_h^{q,q-1}$ and $\hat{g}_h^{q,q-4}$ given in Section 2, under specific model assumptions about classification errors. Some of the resulting expressions are still rather complicated, although they can be implemented on a computer in a straightforward manner. This complexity is partly due to our focus on industry codes in the Dutch GBR, which take on many different values ($M \approx 300$). It is therefore instructive to first examine some results for a simpler case that involves a binary classification variable ($M = 2$).

This binary situation arises often in practice, whenever units are classified according to a single domain of interest. Some recent examples at Statistics Netherlands include: whether a business has a webshop or not (Meertens et al., 2018) and whether a business is innovative or not (Van der Doef et al., 2018). Both of these examples involve machine learning classifiers that are trained on a labeled set (supervised learning). In such applications, a direct estimate of the level matrix $\mathbf{P}^L$ is provided by the confusion matrix of the classifier applied to the test set.

Note: we will state all results in this section without proof, as they are special cases of more general results that will be derived in subsequent sections.

When $M = 2$, all true and observed codes s_i^q and $\hat{s}_i^q$ can take just two possible values: $\mathcal{H}_{\text{full}} = \{1,2\}$. We assume without loss of generality that the code 1 indicates that a unit belongs to the domain of interest. Hence, we are interested in the bias and variance of the observed growth rates $\hat{g}_1^{q,q-u} = (\hat{Y}_1^q / \hat{Y}_1^{q-u}) - 1$. We note that in this special case it always holds that $a_{2i}^q = 1 - a_{1i}^q$ and $\hat{a}_{2i}^q = 1 - \hat{a}_{1i}^q$. To keep the example as simple as possible we will also assume in this section that all units have the same level and change matrix: $\mathbf{P}_i^L = \mathbf{P}^L$ and $\mathbf{P}_i^C = \mathbf{P}^C$ for all i. In particular, we assume that the probabilities in these matrices do not change over time.

First consider a growth rate between two periods during which the true and observed codes do not change. To remain in line with the notation in the rest of the paper, we will denote this growth rate as $\hat{g}_1^{q,q-1}$. Similarly, we will use $\hat{g}_1^{q,q-4}$ for the case where the codes may have changed. The general formulas for the approximate bias and variance of $\hat{g}_1^{q,q-1}$ are given by (6) and (7). In this particular situation, the level matrix $\mathbf{P}^L$ is a 2×2 matrix which can be written in the format

$$\mathbf{P}^L = \begin{pmatrix} p_{11}^L & 1 - p_{11}^L \\ 1 - p_{22}^L & p_{22}^L \end{pmatrix},$$

since the probabilities in each row have to sum to 1.

Define the domain-specific sums of squares $K_g^{q-u} = \sum_{i \in U^{q-u}} a_{gi}^{q-u} \left(y_i^{q-u}\right)^2$ for $u \in \{0,1,4\}$ and define the domain-specific cross-products for continuing units $X_{gO}^{q,q-1} = \sum_{i \in U_O^{q-1,q}} a_{gi}^{q-1} y_i^{q-1} y_i^q$. For the special case considered here, it can be shown that

$$\breve{G}_1^{q,q-1} = \frac{E\left(\hat{Y}_1^q\right)}{E\left(\hat{Y}_1^{q-1}\right)} = \frac{p_{11}^L Y_1^q + (1 - p_{22}^L) Y_2^q}{p_{11}^L Y_1^{q-1} + (1 - p_{22}^L) Y_2^{q-1}}$$

and

$$\boxed{AB\left(\hat{g}_1^{q,q-1}\right) = \left(\breve{G}_1^{q,q-1} - \frac{Y_1^q}{Y_1^{q-1}}\right) + \frac{p_{11}^L(1 - p_{11}^L)\left[\breve{G}_1^{q,q-1} K_1^{q-1} - X_{1O}^{q,q-1}\right] + p_{22}^L(1 - p_{22}^L)\left[\breve{G}_1^{q,q-1} K_2^{q-1} - X_{2O}^{q,q-1}\right]}{\left[p_{11}^L Y_1^{q-1} + (1 - p_{22}^L) Y_2^{q-1}\right]^2} \quad (14)}$$

and

$$\boxed{AV\left(\hat{g}_1^{q,q-1}\right) = \frac{p_{11}^L(1 - p_{11}^L)\left[\left(\breve{G}_1^{q,q-1}\right)^2 K_1^{q-1} + K_1^q - 2\breve{G}_1^{q,q-1} X_{1O}^{q,q-1}\right]}{\left[p_{11}^L Y_1^{q-1} + (1 - p_{22}^L) Y_2^{q-1}\right]^2} + \frac{p_{22}^L(1 - p_{22}^L)\left[\left(\breve{G}_1^{q,q-1}\right)^2 K_2^{q-1} + K_2^q - 2\breve{G}_1^{q,q-1} X_{2O}^{q,q-1}\right]}{\left[p_{11}^L Y_1^{q-1} + (1 - p_{22}^L) Y_2^{q-1}\right]^2}. \quad (15)}$$

These expressions are special cases of the general formulas (39), (40) and (41) that will be derived below from (6) and (7).

For the growth rate $\hat{g}_1^{q,q-4}$ we also have to consider the change matrix $\mathbf{P}^C$. In the notation of Appendix B, these probabilities can be arranged in a $2^3 \times 2$ matrix as follows:

$$\mathbf{P}^C = \begin{pmatrix} p_{1111}^C & 1 - p_{1111}^C \\ p_{1121}^C & 1 - p_{1121}^C \\ p_{1211}^C & 1 - p_{1211}^C \\ p_{1221}^C & 1 - p_{1221}^C \\ p_{2111}^C & 1 - p_{2111}^C \\ p_{2121}^C & 1 - p_{2121}^C \\ p_{2211}^C & 1 - p_{2211}^C \\ p_{2221}^C & 1 - p_{2221}^C \end{pmatrix}.$$

Note that, again, the probabilities in each row have to sum to 1. In total, there are (at most) 10 distinct parameters involved in $\mathbf{P}^L$ and $\mathbf{P}^C$. We can easily derive the corresponding probabilities $\mathbb{p}_{gklh}^{LC} = p_{gl}^L p_{gklh}^C$ and $p_{gkh}^{LC} = \mathbb{p}_{gk1h}^{LC} + \mathbb{p}_{gk2h}^{LC}$ that were defined in Section 2. This yields:

$$\begin{pmatrix} \mathbb{P}^{LC}_{1111} & \mathbb{P}^{LC}_{1112} \\ \mathbb{P}^{LC}_{1121} & \mathbb{P}^{LC}_{1122} \\ \mathbb{P}^{LC}_{1211} & \mathbb{P}^{LC}_{1212} \\ \mathbb{P}^{LC}_{1221} & \mathbb{P}^{LC}_{1222} \\ \mathbb{P}^{LC}_{2111} & \mathbb{P}^{LC}_{2112} \\ \mathbb{P}^{LC}_{2121} & \mathbb{P}^{LC}_{2122} \\ \mathbb{P}^{LC}_{2211} & \mathbb{P}^{LC}_{2212} \\ \mathbb{P}^{LC}_{2221} & \mathbb{P}^{LC}_{2222} \end{pmatrix} = \begin{pmatrix} p^L_{11}p^C_{1111} & p^L_{11}(1-p^C_{1111}) \\ (1-p^L_{11})p^C_{1121} & (1-p^L_{11})(1-p^C_{1121}) \\ p^L_{11}p^C_{1211} & p^L_{11}(1-p^C_{1211}) \\ (1-p^L_{11})p^C_{1221} & (1-p^L_{11})(1-p^C_{1221}) \\ (1-p^L_{22})p^C_{2111} & (1-p^L_{22})(1-p^C_{2111}) \\ p^L_{22}p^C_{2121} & p^L_{22}(1-p^C_{2121}) \\ (1-p^L_{22})p^C_{2211} & (1-p^L_{22})(1-p^C_{2211}) \\ p^L_{22}p^C_{2221} & p^L_{22}(1-p^C_{2221}) \end{pmatrix}$$

and

$$\begin{aligned} &\begin{pmatrix} p^{LC}_{111} & p^{LC}_{112} \\ p^{LC}_{121} & p^{LC}_{122} \\ p^{LC}_{211} & p^{LC}_{212} \\ p^{LC}_{221} & p^{LC}_{222} \end{pmatrix} \\ &= \begin{pmatrix} p^L_{11}p^C_{1111} + (1-p^L_{11})p^C_{1121} & p^L_{11}(1-p^C_{1111}) + (1-p^L_{11})(1-p^C_{1121}) \\ p^L_{11}p^C_{1211} + (1-p^L_{11})p^C_{1221} & p^L_{11}(1-p^C_{1211}) + (1-p^L_{11})(1-p^C_{1221}) \\ (1-p^L_{22})p^C_{2111} + p^L_{22}p^C_{2121} & (1-p^L_{22})(1-p^C_{2111}) + p^L_{22}(1-p^C_{2121}) \\ (1-p^L_{22})p^C_{2211} + p^L_{22}p^C_{2221} & (1-p^L_{22})(1-p^C_{2211}) + p^L_{22}(1-p^C_{2221}) \end{pmatrix}. \end{aligned}$$

Note that the probabilities in the latter matrix again sum to one in each row; i.e., $p^{LC}_{gk2} = 1 - p^{LC}_{gk1}$ for all $g, k \in \{1,2\}$.

In addition, we define the following turnover sub-totals (for $g, k \in \{1,2\}$):

- Y^{q-u}_{gkO} (with $u \in \{0,4\}$) denotes the total turnover of continuing units $i \in U^{q-4,q}_O$ with $s^{q-4}_i = g$ and $s^q_i = k$;
- Y^q_{kB} denotes the total turnover of new-born units $i \in U^{q-4,q}_B$ with $s^q_i = k$;
- Y^{q-4}_{gD} denotes the total turnover of dead units $i \in U^{q-4,q}_D$ with $s^{q-4}_i = g$.

Note that $Y^{q-4}_g = Y^{q-4}_{g1O} + Y^{q-4}_{g2O} + Y^{q-4}_{gD}$ and $Y^q_k = Y^q_{1kO} + Y^q_{2kO} + Y^q_{kB}$. Similarly, we partition the sums of squares K^{q-4}_g and K^q_k as $K^{q-4}_g = K^{q-4}_{g1O} + K^{q-4}_{g2O} + K^{q-4}_{gD}$ and $K^q_k = K^q_{1kO} + K^q_{2kO} + K^q_{kB}$. Finally, as a variation on the cross-product $X^{q,q-1}_{gO}$ we define $X^{q,q-4}_{gkO} = \sum_{i \in U^{q-4,q}_O} a^{q-4}_{gi} a^q_{ki} y^{q-4}_i y^q_i$.

Using this notation, it can be shown that

$$\begin{aligned} \breve{G}^{q,q-4}_1 &= \frac{E(\hat{Y}^q_1)}{E(\hat{Y}^{q-4}_1)} \\ &= \frac{[p^{LC}_{111}Y^q_{11O} + p^{LC}_{121}Y^q_{12O} + p^{LC}_{211}Y^q_{21O} + p^{LC}_{221}Y^q_{22O}] + [p^L_{11}Y^q_{1B} + (1-p^L_{22})Y^q_{2B}]}{p^L_{11}Y^{q-4}_1 + (1-p^L_{22})Y^{q-4}_2}, \end{aligned}$$

$$\boxed{\begin{aligned} AB(\hat{g}^{q,q-4}_1) &= \left(\breve{G}^{q,q-4}_1 - \frac{Y^q_1}{Y^{q-4}_1}\right) + \frac{\breve{G}^{q,q-4}_1 B^{q,q-4}_1 - C^{q,q-4}_{1O}}{[p^L_{11}Y^{q-4}_1 + (1-p^L_{22})Y^{q-4}_2]^2}, \\ B^{q,q-4}_1 &= p^L_{11}(1-p^L_{11})K^{q-4}_1 + p^L_{22}(1-p^L_{22})K^{q-4}_2, \\ C^{q,q-4}_{1O} &= (\mathbb{P}^{LC}_{1111} - p^L_{11}p^{LC}_{111})X^{q,q-4}_{11O} + (\mathbb{P}^{LC}_{1211} - p^L_{11}p^{LC}_{121})X^{q,q-4}_{12O} \\ &\quad + [\mathbb{P}^{LC}_{2111} - (1-p^L_{22})p^{LC}_{211}]X^{q,q-4}_{21O} \\ &\quad + [\mathbb{P}^{LC}_{2211} - (1-p^L_{22})p^{LC}_{221}]X^{q,q-4}_{22O}, \end{aligned}} \quad (16)$$

and

$$\boxed{\begin{aligned} AV\left(\hat{g}_1^{q,q-4}\right) &= \frac{V_{11}^{q,q-4} + V_{21O}^{q,q-4} - 2\breve{G}_1^{q,q-4} C_{1O}^{q,q-4}}{\left[p_{11}^L Y_1^{q-4} + (1 - p_{22}^L) Y_2^{q-4}\right]^2}, \\ V_{11}^{q,q-4} &= p_{11}^L (1 - p_{11}^L) \left[\left(\breve{G}_1^{q,q-4}\right)^2 K_1^{q-4} + K_{1B}^q\right] \\ &\quad + p_{22}^L (1 - p_{22}^L) \left[\left(\breve{G}_1^{q,q-4}\right)^2 K_2^{q-4} + K_{2B}^q\right], \\ V_{21O}^{q,q-4} &= p_{111}^{LC}(1 - p_{111}^{LC}) K_{11O}^q + p_{121}^{LC}(1 - p_{121}^{LC}) K_{12O}^q \\ &\quad + p_{211}^{LC}(1 - p_{211}^{LC}) K_{21O}^q + p_{221}^{LC}(1 - p_{221}^{LC}) K_{22O}^q, \end{aligned} \qquad (17)}$$

with $C_{1O}^{q,q-4}$ as defined in (16). These expressions are special cases of the general formulas (53), (58) and (63) that will be derived below from (12) and (13).

The above formulas can account for any type of dependence between the classification errors that occur in periods $q - 4$ and q, as expressed by the probabilities $\mathbb{p}_{gkl1}^{LC}$ and p_{gk1}^{LC}. For the special case that errors at different time points can be considered independent, it was already noted in Section 2 that $\mathbb{p}_{gk11}^{LC} = p_{g1}^L p_{gk1}^{LC}$ and hence that $C_{1O}^{q,q-4} = 0$ in (16) and (17). Moreover, it then follows from (8) that

$$p_{gkh}^{LC} = P\left(\hat{s}_i^q = h \middle| s_i^{q-4} = g, s_i^q = k\right) = P\left(\hat{s}_i^q = h \middle| s_i^q = k\right) = p_{kh}^L.$$

Using this, it can be shown that the above expressions now reduce to:

$$\breve{G}_1^{q,q-4} = \frac{E\left(\hat{Y}_1^q\right)}{E\left(\hat{Y}_1^{q-4}\right)} = \frac{p_{11}^L Y_1^q + (1 - p_{22}^L) Y_2^q}{p_{11}^L Y_1^{q-4} + (1 - p_{22}^L) Y_2^{q-4}}$$

and

$$\boxed{\begin{aligned} AB\left(\hat{g}_1^{q,q-4}\right) &= \left(\breve{G}_1^{q,q-4} - \frac{Y_1^q}{Y_1^{q-4}}\right) \\ &\quad + \breve{G}_1^{q,q-4} \frac{p_{11}^L (1 - p_{11}^L) K_1^{q-4} + p_{22}^L (1 - p_{22}^L) K_2^{q-4}}{\left[p_{11}^L Y_1^{q-4} + (1 - p_{22}^L) Y_2^{q-4}\right]^2}, \end{aligned} \qquad (18)}$$

and

$$\boxed{\begin{aligned} AV\left(\hat{g}_1^{q,q-4}\right) &= \frac{p_{11}^L (1 - p_{11}^L) \left[\left(\breve{G}_1^{q,q-4}\right)^2 K_1^{q-4} + K_1^q\right]}{\left[p_{11}^L Y_1^{q-4} + (1 - p_{22}^L) Y_2^{q-4}\right]^2} \\ &\quad + \frac{p_{22}^L (1 - p_{22}^L) \left[\left(\breve{G}_1^{q,q-4}\right)^2 K_2^{q-4} + K_2^q\right]}{\left[p_{11}^L Y_1^{q-4} + (1 - p_{22}^L) Y_2^{q-4}\right]^2}. \end{aligned} \qquad (19)}$$

Comparing expressions (18) and (19) to expressions (14) and (15), if we ignore the difference in notation between $q - 4$ and $q - 1$, it is seen that the only difference is that the terms involving cross-products are absent in (18) and (19). For non-negative target variables y, such as turnover, it holds that $X_{1O}^{q,q-1} \geq 0$ and $X_{2O}^{q,q-1} \geq 0$. Hence, all else being equal, the variance in (19) (under an assumption of independent errors over time) is at least as large as the corresponding variance in (15) (under an assumption of no change in errors over time).

We conclude this section by remarking that the above expressions contain population parameters that are unknown in practice. Details on how to estimate these bias and variance formulas are discussed for the general case in Subsections 4.5 and 5.4.

4. Practical bias and variance approximations for quarter-on-quarter growth rates

4.1 Introduction

Expressions (6) and (7) in Subsection 2.3 for the approximate bias and variance of a quarter-on-quarter growth rate $\hat{g}_h^{q,q-1}$ are exact to the order of approximation in the Taylor series, but the number of unknown parameters p_{ghi}^L in these expressions is very large, because each unit i has its own level and change matrix. These expressions are therefore too complicated for practical use at NSIs. In this section, we will derive further approximations to (6) and (7) by introducing some simplifying assumptions about the nature of the classification errors. In Subsection 4.2, we start with a stable population and a very simple model that may be unrealistic, but which leads to simple expressions for the bias and variance that can be interpreted easily. We then proceed to a more realistic version of this model in Subsection 4.3. Finally, births and deaths are re-introduced into the model in Subsection 4.4. In practice, for a given application 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 in practice will be treated in Subsection 4.5.

4.2 Stable population and a single probability parameter

4.2.1 Assumptions and notation

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

A1. The population for the quarters $q-1$ and q is stable, i.e., there are no births and deaths, so $U^{q-1} = U^q = U_O^{q-1,q}$.
A2. 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 $\mathbf{P}_i^L$ with elements given by

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

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 A2 to our expression for $E\left(\hat{Y}_h^{q-1}\right)$ in (4):

$$\begin{aligned} E\left(\hat{Y}_h^{q-1}\right) &= \sum\nolimits_{i\in U^{q-1}} y_i^{q-1}\left[a_{hi}^{q-1}p + \left(1-a_{hi}^{q-1}\right)\frac{1-p}{M-1}\right] \\ &= p\sum\nolimits_{i\in U_h^{q-1}} y_i^{q-1} + \frac{1-p}{M-1}\sum\nolimits_{i\in U^{q-1}\setminus U_h^{q-1}} y_i^{q-1} \\ &= pY_h^{q-1} + (1-p)\bar{Y}_{(-h)}^{q-1}, \end{aligned} \tag{20}$$

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. Also, $U_h^{q-1} \subseteq U^{q-1}$ denotes the sub-population of units in stratum h according to the true classification. 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

$$\begin{aligned} E\left(\hat{Y}_h^{q}\right) &= pY_h^{q} + (1-p)\bar{Y}_{(-h)}^{q}, \\ \bar{Y}_{(-h)}^{q} &= \frac{Y^{q} - Y_h^{q}}{M-1}. \end{aligned} \tag{21}$$

and hence that

$$\breve{G}_h^{q,q-1} = \frac{E(\hat{Y}_h^{q})}{E(\hat{Y}_h^{q-1})} = \frac{pY_h^{q} + (1-p)\bar{Y}_{(-h)}^{q}}{pY_h^{q-1} + (1-p)\bar{Y}_{(-h)}^{q-1}}. \tag{22}$$

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

$$e_{hi}^{q,q-1} = \left(G_i^{q,q-1} - \breve{G}_h^{q,q-1}\right)y_i^{q-1} = y_i^{q} - \breve{G}_h^{q,q-1}y_i^{q-1}. \tag{23}$$

(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:

$$\begin{aligned} S_g(z) &= \sum\nolimits_{i\in U^{q-1}} a_{gi}^{q-1} z_i\,, \quad g = 1, \dots, M, \\ SS_g(z) &= \sum\nolimits_{i\in U^{q-1}} a_{gi}^{q-1} z_i^2\,, \quad g = 1, \dots, M. \end{aligned} \tag{24}$$

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

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

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

$$\begin{aligned} \bar{S}_{(-h)}(z) &= \frac{S(z) - S_h(z)}{M-1}, \\ \overline{SS}_{(-h)}(z) &= \frac{SS(z) - SS_h(z)}{M-1}. \end{aligned} \tag{25}$$

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

4.2.2 Bias

Under assumption A1 from Subsection 4.2.1, the approximate bias $AB\left(\hat{g}_h^{q,q-1}\right)$ in formula (6) reduces to:

$$\begin{aligned} & AB\left(\hat{g}_h^{q,q-1}\right) \\ & = \frac{\sum_{i \in U^{q-1}} \left[\left(\breve{G}_h^{q,q-1} - G_i^{q,q-1}\right)\left(y_i^{q-1}\right)^2 \sum_{g=1}^{M} a_{gi}^{q-1} p_{ghi}^{L}\left(1 - p_{ghi}^{L}\right)\right]}{\left[E\left(\hat{Y}_h^{q-1}\right)\right]^2} \\ & + \left(\breve{G}_h^{q,q-1} - G_h^{q,q-1}\right). \end{aligned} \tag{26}$$

Next, using assumption A2 and the notation introduced above, we find that the numerator of the first term in the above expression is equal to

$$\begin{aligned} & \sum\nolimits_{i \in U^{q-1}} \left(-e_{hi}^{q,q-1}\right) y_i^{q-1} \left[a_{hi}^{q-1} p(1-p) + \left(1 - a_{hi}^{q-1}\right) \frac{1-p}{M-1}\left(1 - \frac{1-p}{M-1}\right)\right] \\ & = p(1-p) S_h\left(-e_h^{q,q-1} y^{q-1}\right) \\ & \qquad\qquad + \frac{1-p}{M-1}\left(1 - \frac{1-p}{M-1}\right)\left[S\left(-e_h^{q,q-1} y^{q-1}\right) - S_h\left(-e_h^{q,q-1} y^{q-1}\right)\right] \\ & = -\left\{p(1-p) S_h\left(e_h^{q,q-1} y^{q-1}\right) + (1-p)\left(1 - \frac{1-p}{M-1}\right) \bar{S}_{(-h)}\left(e_h^{q,q-1} y^{q-1}\right)\right\} \end{aligned}$$

and thus that

$$\boxed{\begin{aligned} & AB\left(\hat{g}_h^{q,q-1}\right) \\ & = \left(\breve{G}_h^{q,q-1} - G_h^{q,q-1}\right) \\ & - \frac{(1-p)\left[p S_h\left(e_h^{q,q-1} y^{q-1}\right) + \left(1 - \frac{1-p}{M-1}\right) \bar{S}_{(-h)}\left(e_h^{q,q-1} y^{q-1}\right)\right]}{\left[E\left(\hat{Y}_h^{q-1}\right)\right]^2}, \end{aligned}} \tag{27}$$

with $E\left(\hat{Y}_h^{q-1}\right)$ given by (20) and

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

according to (22). We also used the shorthand notation from (24) and (25). Note that, for all $g = 1, \ldots, M$,

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

Expression (27) shows 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 is not difficult to see 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 (27) 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 Subsection 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.

4.2.3 Variance

For the approximate variance $AV\left(\hat{g}_h^{q,q-1}\right)$ we find that, again under assumption A1 from Subsection 4.2.1, formula (7) reduces to

$$AV\left(\hat{g}_h^{q,q-1}\right) = \frac{\sum_{i\in U^{q-1}}\left[\left(\breve{G}_h^{q,q-1} - G_i^{q,q-1}\right)^2\left(y_i^{q-1}\right)^2\sum_{g=1}^{M} a_{gi}^{q-1} p_{ghi}^{L}\left(1 - p_{ghi}^{L}\right)\right]}{\left[E\left(\hat{Y}_h^{q-1}\right)\right]^2}. \tag{28}$$

Next we use assumption A2 and proceed along similar lines as for the bias to obtain:

$$AV\left(\hat{g}_h^{q,q-1}\right) = \frac{p(1-p)SS_h\left(e_h^{q,q-1}\right) + \frac{1-p}{M-1}\left(1 - \frac{1-p}{M-1}\right)\left[SS\left(e_h^{q,q-1}\right) - SS_h\left(e_h^{q,q-1}\right)\right]}{\left[E\left(\hat{Y}_h^{q-1}\right)\right]^2}$$

and hence

$$\boxed{AV\left(\hat{g}_h^{q,q-1}\right) = \frac{(1-p)\left[pSS_h\left(e_h^{q,q-1}\right) + \left(1 - \frac{1-p}{M-1}\right)\overline{SS}_{(-h)}\left(e_h^{q,q-1}\right)\right]}{\left[E\left(\hat{Y}_h^{q-1}\right)\right]^2},} \tag{29}$$

with $E\left(\hat{Y}_h^{q-1}\right)$ given by (20). Here, we used the shorthand notation for sums of squares defined in (24) and (25). Note that, for all $g = 1, \dots, M$,

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

It is instructive to write the numerator of (29) in a slightly different form:

$$\left[p(1-p)SS_h\left(e_h^{q,q-1}\right)\right] + (M-1)\left[\frac{1-p}{M-1}\left(1 - \frac{1-p}{M-1}\right)\overline{SS}_{(-h)}\left(e_h^{q,q-1}\right)\right]. \tag{30}$$

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 $\frac{1-p}{M-1}\left(1-\frac{1-p}{M-1}\right)\overline{SS}_{(-h)}\left(e_h^{q,q-1}\right)$.

According to (29), 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 vary 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.

4.3 Stable population and multiple probability parameters

4.3.1 Assumptions and notation

The above assumption A2 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 A1 that the population is stable; this assumption will be relaxed in Subsection 4.4.)

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 and maintenance of the GBR. In practice, the largest 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 $U_1^{q-1}, \dots, U_C^{q-1}$, with $U_1^{q-1} \cup \dots \cup U_C^{q-1} = U^{q-1}$ 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 $\mathbf{P}_i^L = \mathbf{P}_c^L$, that is, the probability classes are homogeneous with respect to the classification error probabilities. We now obtain the following version of assumption A2:

A2’. The population consists of probability classes $U_1^{q-1}, \dots, U_C^{q-1}$. 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 $\mathbf{P}_c^L$ with elements given by

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

It is not difficult to show that, with assumption A2 replaced by assumption A2’, the following bias and variance approximations are obtained instead of (27) and (29):

$$
\begin{aligned}
&AB\left(\hat{g}_h^{q,q-1}\right)\\
&= \frac{\sum_{c=1}^{C}\left[p_c Y_{hc}^{q} + (1-p_c)\bar{Y}_{(-h)c}^{q}\right]}{\sum_{c=1}^{C}\left[p_c Y_{hc}^{q-1} + (1-p_c)\bar{Y}_{(-h)c}^{q-1}\right]} - \frac{Y_h^q}{Y_h^{q-1}}\\
&- \frac{\sum_{c=1}^{C}(1-p_c)\left[p_c S_{hc}\left(e_h^{q,q-1}y^{q-1}\right) + \left(1-\frac{1-p_c}{M-1}\right)\bar{S}_{(-h)c}\left(e_h^{q,q-1}y^{q-1}\right)\right]}{\left\{\sum_{c=1}^{C}\left[p_c Y_{hc}^{q-1} + (1-p_c)\bar{Y}_{(-h)c}^{q-1}\right]\right\}^2}
\end{aligned}
\qquad (31)
$$

and

$$
\begin{aligned}
&AV\left(\hat{g}_h^{q,q-1}\right)\\
&= \frac{\sum_{c=1}^{C}(1-p_c)\left[p_c SS_{hc}\left(e_h^{q,q-1}\right) + \left(1-\frac{1-p_c}{M-1}\right)\overline{SS}_{(-h)c}\left(e_h^{q,q-1}\right)\right]}{\left\{\sum_{c=1}^{C}\left[p_c Y_{hc}^{q-1} + (1-p_c)\bar{Y}_{(-h)c}^{q-1}\right]\right\}^2}.
\end{aligned}
\qquad (32)
$$

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 Subsection 4.2.1 to operate within probability classes:

$$
\begin{aligned}
\bar{Y}_{(-h)c}^{q-u} &= \frac{Y_{+c}^{q-u} - Y_{hc}^{q-u}}{M-1} = \frac{\sum_{g=1}^{M} Y_{gc}^{q-u} - Y_{hc}^{q-u}}{M-1}, \quad u = 0{,}1,\\
S_{gc}(z) &= \sum\nolimits_{i\in U_c^{q-1}} a_{gi}^{q-1} z_i\,, \quad g = 1, \dots, M,\\
\bar{S}_{(-h)c}(z) &= \frac{S_{+c}(z) - S_{hc}(z)}{M-1} = \frac{\sum_{g=1}^{M} S_{gc}(z) - S_{hc}(z)}{M-1},\\
SS_{gc}(z) &= \sum\nolimits_{i\in U_c^{q-1}} a_{gi}^{q-1} z_i^2\,, \quad g = 1, \dots, M,\\
\overline{SS}_{(-h)c}(z) &= \frac{SS_{+c}(z) - SS_{hc}(z)}{M-1} = \frac{\sum_{g=1}^{M} SS_{gc}(z) - SS_{hc}(z)}{M-1}.
\end{aligned}
\qquad (33)
$$

A second step towards a more realistic classification error model is to acknowledge that, within each level matrix $\mathbf{P}_c^L$, 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. In the context of the GBR, the diagonal probabilities can differ because in some industries it is more difficult to classify units correctly than in other industries. Also, the off-diagonal probabilities can differ because certain types of misclassification are more likely to occur than others; e.g., wholesale traders are more likely to be misclassified as retail traders than as banks. Hence, in principle all elements of $\mathbf{P}_c^L$ could be different.

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 $\mathbf{P}_c^L$. [If all elements of $\mathbf{P}_c^L$ were left free, then hardly any simplification of expressions (6) and (7) 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}^L apply. It is assumed that $g \in \mathcal{H}_g$, so the diagonal probabilities are always included in this subset. The subsets $\mathcal{H}_g$ are supposed to capture all 'significant' cases, i.e., all industry codes h that have a relatively large probability of being observed when the true industry code is g. In other words, it is assumed that each probability p_{ghc}^{L} with $h \notin \mathcal{H}_g$ is small, so that $0 \ll \sum_{h \in \mathcal{H}_g} p_{ghc}^{L} \leq 1$. In addition, the subsets $\mathcal{H}_g$ should be chosen as small as possible; in particular, each $\left|\mathcal{H}_g\right|$ should be much smaller than M. 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}^{L} is 'significant'. The remaining probabilities p_{ghc}^{L} 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 (from the perspective of the observed stratum h), which is $(p_{+hc}^{L} - \sum_{g \in \mathcal{G}_h} p_{ghc}^{L})/(M - |\mathcal{G}_h|)$, with $p_{+hc}^{L} = \sum_{g=1}^{M} p_{ghc}^{L}$. Note that $p_{+hc}^{L} \neq 1$ in general. (The exchangeable errors model of assumption A2 is an exception.)

This leads to the following, final version of assumption A2:

A2''. 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 $\mathbf{P}_c^L$ that can be approximated well by a matrix $\widetilde{\mathbf{P}}_c^L$ with elements given by

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

Note that, by definition, $\widetilde{\mathbf{P}}_c^L$ can only be an *approximation* to the real level matrix $\mathbf{P}_c^L$, because the rows of $\widetilde{\mathbf{P}}_c^L$ do not necessarily sum to 1. In what follows, we will use more shorthand notation that partly generalises (33):

$$\begin{aligned} \tilde{p}_{(-\mathcal{G}_h)hc}^{L} &= p_{+hc}^{L} - \sum\nolimits_{g \in \mathcal{G}_h} p_{ghc}^{L}, \\ \bar{Y}_{(-\mathcal{G}_h)c}^{q-u} &= \frac{Y_{+c}^{q-u} - \sum_{g \in \mathcal{G}_h} Y_{gc}^{q-u}}{M - |\mathcal{G}_h|}, \quad u = 0{,}1, \\ \bar{S}_{(-\mathcal{G}_h)c}(z) &= \frac{S_{+c}(z) - \sum_{g \in \mathcal{G}_h} S_{gc}(z)}{M - |\mathcal{G}_h|}, \\ \overline{SS}_{(-\mathcal{G}_h)c}(z) &= \frac{SS_{+c}(z) - \sum_{g \in \mathcal{G}_h} SS_{gc}(z)}{M - |\mathcal{G}_h|}. \end{aligned} \tag{34}$$

4.3.2 Bias

Starting again with expression (4) and proceeding along similar lines as in (20), we obtain the following approximation to $E(\hat{Y}_h^{q-1})$ under assumption A2'':

$$
\begin{aligned}
E\left(\hat{Y}_h^{q-1}\right) &= \sum_{c=1}^{\mathcal{C}} \sum\nolimits_{i \in U_c^{q-1}} y_i^{q-1} \sum\nolimits_{g=1}^{M} a_{gi}^{q-1} p_{ghc}^{L} \\
&\approx \sum_{c=1}^{\mathcal{C}} \sum\nolimits_{i \in U_c^{q-1}} y_i^{q-1} \left[\sum_{g \in \mathcal{G}_h} a_{gi}^{q-1} p_{ghc}^{L} \right. \\
&\qquad\qquad \left. + \left(1 - \sum_{g \in \mathcal{G}_h} a_{gi}^{q-1}\right) \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{L}}{M - |\mathcal{G}_h|} \right] \\
&= \sum_{c=1}^{\mathcal{C}} \left[\sum_{g \in \mathcal{G}_h} p_{ghc}^{L} Y_{gc}^{q-1} + \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{L}}{M - |\mathcal{G}_h|} \left(Y_{+c}^{q-1} - \sum_{g \in \mathcal{G}_h} Y_{gc}^{q-1} \right) \right] \\
&= \sum_{c=1}^{\mathcal{C}} \left[\sum_{g \in \mathcal{G}_h} p_{ghc}^{L} Y_{gc}^{q-1} + \tilde{p}_{(-\mathcal{G}_h)hc}^{L} \bar{Y}_{(-\mathcal{G}_h)c}^{q-1} \right],
\end{aligned}
\tag{35}
$$

where we used notation defined in (34). Note that in the second line of (35) we used that $\sum_{g \notin \mathcal{G}_h} a_{gi}^{q-1} = 1 - \sum_{g \in \mathcal{G}_h} a_{gi}^{q-1}$ since each unit observed in h belongs to one true industry code g. Analogously, we also obtain:

$$
\begin{aligned}
E\left(\hat{Y}_h^{q}\right) &\approx \sum_{c=1}^{\mathcal{C}} \left[\sum_{g \in \mathcal{G}_h} p_{ghc}^{L} Y_{gc}^{q} + \tilde{p}_{(-\mathcal{G}_h)hc}^{L} \bar{Y}_{(-\mathcal{G}_h)c}^{q} \right], \\
\breve{G}_h^{q,q-1} &= \frac{E\left(\hat{Y}_h^{q}\right)}{E\left(\hat{Y}_h^{q-1}\right)} \approx \frac{\sum_{c=1}^{\mathcal{C}} \left[\sum_{g \in \mathcal{G}_h} p_{ghc}^{L} Y_{gc}^{q} + \tilde{p}_{(-\mathcal{G}_h)hc}^{L} \bar{Y}_{(-\mathcal{G}_h)c}^{q} \right]}{\sum_{c=1}^{\mathcal{C}} \left[\sum_{g \in \mathcal{G}_h} p_{ghc}^{L} Y_{gc}^{q-1} + \tilde{p}_{(-\mathcal{G}_h)hc}^{L} \bar{Y}_{(-\mathcal{G}_h)c}^{q-1} \right]}.
\end{aligned}
\tag{36}
$$

To approximate the bias, we can still use (26) as a starting point. We can proceed analogously to Subsection 4.2.2 and the derivation of (35) to obtain eventually:

$$
\boxed{
\begin{aligned}
AB\left(\hat{g}_h^{q,q-1}\right) &\approx \frac{\sum_{c=1}^{\mathcal{C}} \left[\sum_{g \in \mathcal{G}_h} p_{ghc}^{L} Y_{gc}^{q} + \tilde{p}_{(-\mathcal{G}_h)hc}^{L} \bar{Y}_{(-\mathcal{G}_h)c}^{q} \right]}{\sum_{c=1}^{\mathcal{C}} \left[\sum_{g \in \mathcal{G}_h} p_{ghc}^{L} Y_{gc}^{q-1} + \tilde{p}_{(-\mathcal{G}_h)hc}^{L} \bar{Y}_{(-\mathcal{G}_h)c}^{q-1} \right]} - \frac{Y_h^{q}}{Y_h^{q-1}} \\
&\qquad - \frac{\sum_{c=1}^{\mathcal{C}} B_{hc}^{q,q-1}}{\left\{ \sum_{c=1}^{\mathcal{C}} \left[\sum_{g \in \mathcal{G}_h} p_{ghc}^{L} Y_{gc}^{q-1} + \tilde{p}_{(-\mathcal{G}_h)hc}^{L} \bar{Y}_{(-\mathcal{G}_h)c}^{q-1} \right] \right\}^2}, \\
B_{hc}^{q,q-1} &= \sum_{g \in \mathcal{G}_h} p_{ghc}^{L} \left(1 - p_{ghc}^{L}\right) S_{gc}\left(e_h^{q,q-1} y^{q-1}\right) \\
&\qquad + \tilde{p}_{(-\mathcal{G}_h)hc}^{L} \left(1 - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{L}}{M - |\mathcal{G}_h|}\right) \bar{S}_{(-\mathcal{G}_h)c}\left(e_h^{q,q-1} y^{q-1}\right).
\end{aligned}
}
\tag{37}
$$

4.3.3 Variance

To approximate the variance, we can use (28) as a starting point. Again by proceeding along similar lines as in Subsection 4.2.3 and using approximation (35), we eventually find the following approximation formula:

$$
\begin{aligned}
AV\left(\hat{g}_h^{q,q-1}\right) &\approx \frac{\sum_{c=1}^{\mathcal{C}} V_{hc}^{q,q-1}}{\left\{\sum_{c=1}^{\mathcal{C}}\left[\sum_{g\in\mathcal{G}_h} p_{ghc}^L Y_{gc}^{q-1} + \tilde{p}_{(-\mathcal{G}_h)hc}^L \bar{Y}_{(-\mathcal{G}_h)c}^{q-1}\right]\right\}^2}, \\
V_{hc}^{q,q-1} &= \sum_{g\in\mathcal{G}_h} p_{ghc}^L\left(1 - p_{ghc}^L\right) SS_{gc}\left(e_h^{q,q-1}\right) \\
&\quad + \tilde{p}_{(-\mathcal{G}_h)hc}^L\left(1 - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^L}{M - |\mathcal{G}_h|}\right)\overline{SS}_{(-\mathcal{G}_h)c}\left(e_h^{q,q-1}\right).
\end{aligned}
\tag{38}
$$

Analogous to (30), each term $V_{hc}^{q,q-1}$ can also be written as

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

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

4.4 Dynamic population and multiple probability parameters

4.4.1 Assumptions and notation

We will now drop assumption A1 from Subsection 4.2.1, 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 Subsection 2.3 that concern dead units ($i \in U_D^{q-1,q}$) and new-born units ($i \in U_B^{q-1,q}$). In this section, we will retain assumption A2''. 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 A2'', 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 for quarters within the same year.

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 $Y_{hc}^{q-1} = Y_{hcO}^{q-1} + Y_{hcD}^{q-1}$.

4.4.2 Bias

If the population is not stable between quarters $q - 1$ and q, all terms in the bias approximation in formula (6) 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 (35) as before, and $\left(\breve{G}_h^{q,q-1} - G_h^{q,q-1}\right)$ can still be approximated as in expression (36):

$$\breve{G}_h^{q,q-1} - G_h^{q,q-1} \approx \frac{\sum_{c=1}^{\mathcal{C}}\left[\sum_{g\in\mathcal{G}_h} p_{ghc}^L Y_{gc}^q + \tilde{p}_{(-\mathcal{G}_h)hc}^L \bar{Y}_{(-\mathcal{G}_h)c}^q\right]}{\sum_{c=1}^{\mathcal{C}}\left[\sum_{g\in\mathcal{G}_h} p_{ghc}^L Y_{gc}^{q-1} + \tilde{p}_{(-\mathcal{G}_h)hc}^L \bar{Y}_{(-\mathcal{G}_h)c}^{q-1}\right]} - \frac{Y_h^q}{Y_h^{q-1}}, \qquad (39)$$

using the notation convention introduced above. 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 $B_{hc}^{q,q-1}$ from expression (37):

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

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 $U_O^{q-1,q}$. It is important to note that the pseudo-residual $e_h^{q,q-1}$ is still given by (23), using the overall $\breve{G}_h^{q,q-1}$.

Finally, the component $B_{hD}^{q,q-1}$ for units in $U_D^{q-1,q}$ in (6) has a similar form to $B_{hO}^{q,q-1}$. It is not difficult to see that we may therefore write:

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

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

$$\boxed{\begin{aligned} AB\left(\hat{g}_h^{q,q-1}\right) &\approx \frac{\sum_{c=1}^{\mathcal{C}}\left[\sum_{g\in\mathcal{G}_h} p_{ghc}^L Y_{gc}^q + \tilde{p}_{(-\mathcal{G}_h)hc}^L \bar{Y}_{(-\mathcal{G}_h)c}^q\right]}{\sum_{c=1}^{\mathcal{C}}\left[\sum_{g\in\mathcal{G}_h} p_{ghc}^L Y_{gc}^{q-1} + \tilde{p}_{(-\mathcal{G}_h)hc}^L \bar{Y}_{(-\mathcal{G}_h)c}^{q-1}\right]} - \frac{Y_h^q}{Y_h^{q-1}} \\ &\quad - \frac{\sum_{c=1}^{\mathcal{C}} B_{hc}^{q,q-1}}{\left\{\sum_{c=1}^{\mathcal{C}}\left[\sum_{g\in\mathcal{G}_h} p_{ghc}^L Y_{gc}^{q-1} + \tilde{p}_{(-\mathcal{G}_h)hc}^L \bar{Y}_{(-\mathcal{G}_h)c}^{q-1}\right]\right\}^2}, \\ B_{hc}^{q,q-1} &= \sum_{g\in\mathcal{G}_h} p_{ghc}^L\left(1-p_{ghc}^L\right) BS_{gc}^{q,q-1} \\ &\quad + \tilde{p}_{(-\mathcal{G}_h)hc}^L\left(1-\frac{\tilde{p}_{(-\mathcal{G}_h)hc}^L}{M-|\mathcal{G}_h|}\right)\overline{BS}_{(-\mathcal{G}_h)c}^{q,q-1}, \\ BS_{gc}^{q,q-1} &= S_{gcO}\left(e_h^{q,q-1}y^{q-1}\right) - \breve{G}_h^{q,q-1} SS_{gcD}\left(y^{q-1}\right), \\ \overline{BS}_{(-\mathcal{G}_h)c}^{q,q-1} &= \bar{S}_{(-\mathcal{G}_h)cO}\left(e_h^{q,q-1}y^{q-1}\right) - \breve{G}_h^{q,q-1}\overline{SS}_{(-\mathcal{G}_h)cD}\left(y^{q-1}\right). \end{aligned} \qquad (40)}$$

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 $SS_{gcD}\left(y^{q-1}\right)$, and in $\bar{S}_{(-\mathcal{G}_h)cO}\left(e_h^{q,q-1}y^{q-1}\right)$ and $\overline{SS}_{(-\mathcal{G}_h)cD}\left(y^{q-1}\right)$, respectively.

4.4.3 Variance

If the population is not stable between quarters $q-1$ and q, the variance approximation in formula (7) 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 $V_{hc}^{q,q-1}$ from expression (38):

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

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 $U_O^{q-1,q}$.

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

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

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

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

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

$$\boxed{\begin{aligned} AV\left(\hat{g}_h^{q,q-1}\right) &\approx \frac{\sum_{c=1}^{C} V_{hc}^{q,q-1}}{\left\{\sum_{c=1}^{C} \left[\sum_{g \in \mathcal{G}_h} p_{ghc}^{L} Y_{gc}^{q-1} + \tilde{p}_{(-\mathcal{G}_h)hc}^{L} \bar{Y}_{(-\mathcal{G}_h)c}^{q-1}\right]\right\}^2}, \\ V_{hc}^{q,q-1} &= \sum_{g \in \mathcal{G}_h} p_{ghc}^{L} \left(1 - p_{ghc}^{L}\right) VS_{gc}^{q,q-1} \\ &\quad + \tilde{p}_{(-\mathcal{G}_h)hc}^{L} \left(1 - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{L}}{M - |\mathcal{G}_h|}\right) \overline{VS}_{(-\mathcal{G}_h)c}^{q,q-1}, \\ VS_{gc}^{q,q-1} &= SS_{gcO}\left(e_h^{q,q-1}\right) + SS_{gcD}\left(\breve{G}_h^{q,q-1} y^{q-1}\right) + SS_{gcB}(y^q), \\ \overline{VS}_{(-\mathcal{G}_h)c}^{q,q-1} &= \overline{SS}_{(-\mathcal{G}_h)cO}\left(e_h^{q,q-1}\right) + \overline{SS}_{(-\mathcal{G}_h)cD}\left(\breve{G}_h^{q,q-1} y^{q-1}\right) \\ &\quad + \overline{SS}_{(-\mathcal{G}_h)cB}(y^q). \end{aligned}} \quad (41)$$

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

4.5 Estimating the bias and variance

4.5.1 Introduction

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}^{L}. 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 (see Section 7 on classification error models and Section 9 on a case study). 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 these estimated parameters. In this paper, we will ignore the additional uncertainty due to estimating the parameters of the classification error model. See Meertens et al. (2019a) for a Bayesian approach that accounts for parameter uncertainty.

A natural starting point in the estimation procedure is to calculate an estimate of the ratio of the expectations, $\breve{G}_h^{q,q-1}$, based on expression (39):

$$\hat{\hat{G}}_h^{q,q-1} = \frac{\sum_{c=1}^{\mathcal{C}} \left[\sum_{g \in \mathcal{G}_h} p_{ghc}^L \hat{Y}_{gc}^q + \tilde{p}_{(-\mathcal{G}_h)hc}^L \hat{\hat{Y}}_{(-\mathcal{G}_h)c}^q \right]}{\sum_{c=1}^{\mathcal{C}} \left[\sum_{g \in \mathcal{G}_h} p_{ghc}^L \hat{Y}_{gc}^{q-1} + \tilde{p}_{(-\mathcal{G}_h)hc}^L \hat{\hat{Y}}_{(-\mathcal{G}_h)c}^{q-1} \right]}, \tag{42}$$

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

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

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.

4.5.2 Bias

To estimate bias approximation (40), we replace all unknown quantities by their observed values:

$$\begin{aligned}
\widehat{AB}\left(\hat{g}_h^{q,q-1}\right) \approx{}& \frac{\sum_{c=1}^{\mathcal{C}} \left[\sum_{g \in \mathcal{G}_h} p_{ghc}^L \hat{Y}_{gc}^q + \tilde{p}_{(-\mathcal{G}_h)hc}^L \hat{\hat{Y}}_{(-\mathcal{G}_h)c}^q \right]}{\sum_{c=1}^{\mathcal{C}} \left[\sum_{g \in \mathcal{G}_h} p_{ghc}^L \hat{Y}_{gc}^{q-1} + \tilde{p}_{(-\mathcal{G}_h)hc}^L \hat{\hat{Y}}_{(-\mathcal{G}_h)c}^{q-1} \right]} - \frac{\hat{Y}_h^q}{\hat{Y}_h^{q-1}} \\
&- \frac{\sum_{c=1}^{\mathcal{C}} \hat{B}_{hc}^{q,q-1}}{\left\{ \sum_{c=1}^{\mathcal{C}} \left[\sum_{g \in \mathcal{G}_h} p_{ghc}^L \hat{Y}_{gc}^{q-1} + \tilde{p}_{(-\mathcal{G}_h)hc}^L \hat{\hat{Y}}_{(-\mathcal{G}_h)c}^{q-1} \right] \right\}^2}, \\
\hat{B}_{hc}^{q,q-1} ={}& \sum_{g \in \mathcal{G}_h} p_{ghc}^L \left(1 - p_{ghc}^L\right) \widehat{BS}_{gc}^{q,q-1} \\
&+ \tilde{p}_{(-\mathcal{G}_h)hc}^L \left(1 - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^L}{M - |\mathcal{G}_h|}\right) \widehat{\widehat{BS}}_{(-\mathcal{G}_h)c}^{q,q-1}, \\
\widehat{BS}_{gc}^{q,q-1} ={}& \hat{S}_{gcO}\left(\hat{e}_h^{q,q-1} y^{q-1}\right) - \hat{\hat{G}}_h^{q,q-1} \widehat{SS}_{gcD}\left(y^{q-1}\right), \\
\widehat{\widehat{BS}}_{(-\mathcal{G}_h)c}^{q,q-1} ={}& \hat{\hat{S}}_{(-\mathcal{G}_h)cO}\left(\hat{e}_h^{q,q-1} y^{q-1}\right) - \hat{\hat{G}}_h^{q,q-1} \widehat{\widehat{SS}}_{(-\mathcal{G}_h)cD}\left(y^{q-1}\right).
\end{aligned} \tag{43}$$

Here, $\hat{\hat{G}}_h^{q,q-1}$ is given by (42) and the observed pseudo-residual $\hat{e}_{hi}^{q,q-1}$ is defined as

$$\hat{e}_{hi}^{q,q-1} = \left(G_i^{q,q-1} - \hat{\hat{G}}_h^{q,q-1}\right) y_i^{q-1} = y_i^q - \hat{\hat{G}}_h^{q,q-1} y_i^{q-1}. \tag{44}$$

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:

$$\begin{aligned}
\hat{S}_{gcO}\left(z^{q-u}\right) &= \sum\nolimits_{i \in U_{cO}^{q-1,q}} \hat{a}_{gi}^{q-u} z_i^{q-u}, \quad g = 1, \dots, M, \\
\hat{\hat{S}}_{(-\mathcal{G}_h)cO}\left(z^{q-u}\right) &= \frac{S_{+cO}\left(z^{q-u}\right) - \sum_{g \in \mathcal{G}_h} \hat{S}_{gcO}\left(z^{q-u}\right)}{M - |\mathcal{G}_h|} \\
&= \frac{\sum_{g=1}^{M} \hat{S}_{gcO}\left(z^{q-u}\right) - \sum_{g \in \mathcal{G}_h} \hat{S}_{gcO}\left(z^{q-u}\right)}{M - |\mathcal{G}_h|}, \\
\widehat{SS}_{gcO}\left(z^{q-u}\right) &= \sum\nolimits_{i \in U_{cO}^{q-1,q}} \hat{a}_{gi}^{q-u} \left(z_i^{q-u}\right)^2, \quad g = 1, \dots, M, \\
\widehat{\widehat{SS}}_{(-\mathcal{G}_h)cO}\left(z^{q-u}\right) &= \frac{SS_{+cO}\left(z^{q-u}\right) - \sum_{g \in \mathcal{G}_h} \widehat{SS}_{gcO}\left(z^{q-u}\right)}{M - |\mathcal{G}_h|} \\
&= \frac{\sum_{g=1}^{M} \widehat{SS}_{gcO}\left(z^{q-u}\right) - \sum_{g \in \mathcal{G}_h} \widehat{SS}_{gcO}\left(z^{q-u}\right)}{M - |\mathcal{G}_h|}.
\end{aligned} \tag{45}$$

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 (45) 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.

4.5.3 Variance

Variance approximation (41) can be estimated by:

$$\begin{aligned}
\widehat{AV}\left(\hat{g}_h^{q,q-1}\right) &\approx \frac{\sum_{c=1}^{\mathcal{C}} \hat{V}_{hc}^{q,q-1}}{\left\{\sum_{c=1}^{\mathcal{C}}\left[\sum_{g\in\mathcal{G}_h} p_{ghc}^L \hat{Y}_{gc}^{q-1} + \tilde{p}_{(-\mathcal{G}_h)hc}^L \hat{\hat{Y}}_{(-\mathcal{G}_h)c}^{q-1}\right]\right\}^2}, \\
\hat{V}_{hc}^{q,q-1} &= \sum_{g\in\mathcal{G}_h} p_{ghc}^L\left(1-p_{ghc}^L\right)\widehat{VS}_{gc}^{q,q-1} \\
&\quad + \tilde{p}_{(-\mathcal{G}_h)hc}^L\left(1-\frac{\tilde{p}_{(-\mathcal{G}_h)hc}^L}{M-|\mathcal{G}_h|}\right)\widehat{\widehat{VS}}_{(-\mathcal{G}_h)c}^{q,q-1}, \\
\widehat{VS}_{gc}^{q,q-1} &= \widehat{SS}_{gcO}\left(\hat{e}_h^{q,q-1}\right) + \widehat{SS}_{gcD}\left(\hat{\hat{G}}_h^{q,q-1}y^{q-1}\right) + \widehat{SS}_{gcB}(y^q), \\
\widehat{\widehat{VS}}_{(-\mathcal{G}_h)c}^{q,q-1} &= \widehat{\widehat{SS}}_{(-\mathcal{G}_h)cO}\left(\hat{e}_h^{q,q-1}\right) + \widehat{\widehat{SS}}_{(-\mathcal{G}_h)cD}\left(\hat{\hat{G}}_h^{q,q-1}y^{q-1}\right) \\
&\quad + \widehat{\widehat{SS}}_{(-\mathcal{G}_h)cB}(y^q).
\end{aligned} \tag{46}$$

where $\hat{e}_h^{q,q-1}$ is given by (44) and we use the notation defined in (45).

4.5.4 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

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

its bias is approximately given by

$$B\left(\hat{Y}_{hc}^{q-u}\right) \approx (p_{hhc}^L - 1)Y_{hc}^{q-u} + \sum_{g\in\mathcal{G}_h\backslash\{h\}} p_{ghc}^L Y_{gc}^{q-u} + \tilde{p}_{(-\mathcal{G}_h)hc}^L \bar{Y}_{(-\mathcal{G}_h)c}^{q-u}. \tag{47}$$

From this, it follows that $\hat{\hat{G}}_h^{q,q-1}$ in (42) may also be biased. As a first-order approximation, we find:

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

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

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

we obtain:

$$
\begin{aligned}
&E\left(\hat{\breve{G}}_h^{q,q-1}\right) \\
&\approx \frac{\sum_{c=1}^{\mathcal{C}}\left[\sum_{g\in\mathcal{G}_h} p_{ghc}^L Y_{gc}^q + \tilde{p}_{(-\mathcal{G}_h)hc}^L \bar{Y}_{(-\mathcal{G}_h)c}^q + \sum_{g\in\mathcal{G}_h}\left(p_{ghc}^L - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^L}{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}^L Y_{gc}^{q-1} + \tilde{p}_{(-\mathcal{G}_h)hc}^L \bar{Y}_{(-\mathcal{G}_h)c}^{q-1} + \sum_{g\in\mathcal{G}_h}\left(p_{ghc}^L - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^L}{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}^L - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^L}{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}^L - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^L}{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}^L - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^L}{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}^L - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^L}{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}^L - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^L}{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}^L - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^L}{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 $E\left(\hat{\breve{G}}_h^{q,q-1}\right) \approx \breve{G}_h^{q,q-1}$ and so $\hat{\breve{G}}_h^{q,q-1}$ is an approximately unbiased estimator for $\breve{G}_h^{q,q-1}$. If the industries $g \in \mathcal{G}_h$ all have similar quarterly growth rates, then this relation may be expected to hold. If these industries have widely different growth rates, then this relation need not hold and the bias in $\hat{\breve{G}}_h^{q,q-1}$ could be substantial.

In addition, the above bias and variance estimators may be affected by the fact that the observed residual $\hat{e}_{hi}^{q,q-1} = y_i^q - \hat{\breve{G}}_h^{q,q-1} y_i^{q-1}$ depends on the estimated $\breve{G}_h^{q,q-1}$, 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}^L is close to 1 for all target industries and the remaining p_{ghc}^L $(g \neq h)$ are small. In this case, $B\left(\hat{Y}_{hc}^{q-1}\right) \approx 0$ by (47).

5. Practical bias and variance approximations for year-on-year growth rates

5.1 Introduction

The bias and variance approximations (12) and (13) derived in Subsection 2.4 for a year-on-year growth rate $\hat{g}_h^{q,q-4}$ (or a quarter-on-quarter growth rate between the fourth and first quarters of two different years) involve sums with a separate contribution for each unit in the population. In Section 4, we were able to reduce similar expressions for the case of a within-year quarter-on-quarter growth rate to expressions that involve separate contributions for a limited number of groups of units, by introducing assumptions about the level probabilities p_{ghi}^{L}. In this section, we would like to do the same for the bias and variance of a yearly growth rate. This requires additional assumptions about the transition probabilities p_{gkhi}^{LC} as well as the underlying 'diagonal' probabilities $\mathbb{p}_{gkhhi}^{LC}$. [Recall that $p_{gkhi}^{LC} = \sum_{l=1}^{M} \mathbb{p}_{gklhi}^{LC}$. We note that the probabilities $\mathbb{p}_{gklhi}^{LC}$ with $l \neq h$ do not occur separately in expressions (12) and (13).] As will be seen below, these assumptions do not necessarily simplify the mathematical form of the bias and variance expressions, but they do reduce the number of parameters that need to be estimated as well as the amount of computation that is needed to evaluate the bias and the variance in practice.

In Subsection 5.2, we begin by introducing some simplifying assumptions about p_{gkhi}^{LC} and $\mathbb{p}_{gkhhi}^{LC}$. The resulting bias and variance approximations are derived in Subsection 5.3. Estimation of the resulting expressions is discussed in Subsection 5.4.

5.2 Assumptions and notation

Assumption A2'' of Section 4 is also relevant for year-on-year growth rates. Recall that this assumption involves so-called probability classes, where units in the same class are homogeneous with respect to the level matrix $\mathbf{P}_i^L$. We will now extend the definition of these classes to the change matrix $\mathbf{P}_i^C$ of Subsection 2.4. For simplicity of notation, we will assume that the *same* division of units into probability classes applies to both the level and change matrices.

In Section 4, we could realistically 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 are assigned to a single probability class for both periods $q-4$ and q. We denote these classes as $U_{1O}^{q-4,q}, \dots, U_{CO}^{q-4,q}$.

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 $U_{cO}^{q-4,q}$ 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 evaluate the effect of this approximation on the estimated uncertainty 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.

From the assumption that all probabilities are constant within a probability class, so that $\mathbf{P}_i^L = \mathbf{P}_c^L$ and $\mathbf{P}_i^C = \mathbf{P}_c^C$ for all $i \in U_{cO}^{q-4,q}$, it follows immediately that $\mathbb{p}_{gkhhi}^{LC} = \mathbb{p}_{gkhhc}^{LC} = p_{ghc}^{L} p_{gkhhc}^{C}$ and $p_{gkhi}^{LC} = p_{gkhc}^{LC} = \sum_{l=1}^{M} p_{glc}^{L} p_{gklhc}^{C}$ for all units in probability class $U_{cO}^{q-4,q}$. For the probabilities p_{ghc}^{L}, we can apply assumption A2'' to reduce the number of parameters within each probability class. We will now introduce a similar assumption for p_{gkhc}^{LC} and $\mathbb{p}_{gkhhc}^{LC}$. [Note that this is sufficient, because the probabilities in $\mathbf{P}_c^C$ themselves do not occur separately in expressions (12) and (13).]

The probabilities p_{gkhc}^{LC} and $\mathbb{p}_{gkhhc}^{LC}$ 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, which could be a very large number. (Recall that for the Dutch GBR, $M \approx 300$.) Moreover, most of these strata contain a limited number of units; in fact many of them may be empty. The approximations that follow are important to increase the stability of the bias and variance estimates, to reduce the computational workload, 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 $\left(s_i^{q-4}, s_i^{q}, \hat{s}_i^{q}\right)$:

- **diagonal cases**: $\left(s_i^{q-4}, s_i^{q}, \hat{s}_i^{q}\right) = (g, g, h)$, for $g \in \{1, \dots, M\}$;
  (these are cases where the true industry code does not change)
- **cases within column** h: $\left(s_i^{q-4}, s_i^{q}, \hat{s}_i^{q}\right) = (g, h, h)$ with $g \neq h$;
  (these are cases where the true industry code changes and the observed code in q *is correct)*
- **cases within row** h: $\left(s_i^{q-4}, s_i^{q}, \hat{s}_i^{q}\right) = (h, k, h)$ 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* $\mathrm{q}-4$*)*
- **all other cases**: $\left(s_i^{q-4}, s_i^{q}, \hat{s}_i^{q}\right) = (g, k, h)$ 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* $\mathrm{q}-4$*)*

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}^{LC} and $\mathbb{p}_{gkhhc}^{LC}$ by average values. This is similar to assumption A2'' for the level matrix.

In Subsection 4.3, 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, for a given observed code $\hat{s}_i^q = h$:

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

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 $\mathcal{T}_h = \left\{(g, k): I_{gkh} = 1\right\}$, 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 Subsection 4.3. Again, we suppose that the number of special combinations $|\mathcal{T}_h|$ is small in comparison to the total number of ordered pairs M^2.

In the bias and variance formulas below, we will encounter probabilities p_{gkhi}^{LC} and $\mathbb{p}_{gkhhi}^{LC}$ 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 $(g, k) \in \mathcal{T}_h$, separate probabilities p_{gkhc}^{LC} and $\mathbb{p}_{gkhhc}^{LC}$ are used for all units 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:

$$\begin{aligned}\tilde{p}^{LC}_{(-\mathcal{T}_h)hc} &= p^{LC}_{++hc} - \sum_{(g,k)\in\mathcal{T}_h} p^{LC}_{gkhc} = \sum_{g=1}^{M}\sum_{k=1}^{M} p^{LC}_{gkhc} - \sum_{(g,k)\in\mathcal{T}_h} p^{LC}_{gkhc}\,,\\ \tilde{\mathbb{p}}^{LC}_{(-\mathcal{T}_h)hhc} &= \mathbb{p}^{LC}_{++hhc} - \sum_{(g,k)\in\mathcal{T}_h} \mathbb{p}^{LC}_{gkhhc} = \sum_{g=1}^{M}\sum_{k=1}^{M} \mathbb{p}^{LC}_{gkhhc} - \sum_{(g,k)\in\mathcal{T}_h} \mathbb{p}^{LC}_{gkhhc}.\end{aligned} \tag{48}$$

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

$$\begin{aligned}\tilde{p}^{LC}_{gkhc} &= \begin{cases} p^{LC}_{gkhc} & \text{if } (g,k)\in\mathcal{T}_h \\ \dfrac{\tilde{p}^{LC}_{(-\mathcal{T}_h)hc}}{M^2-|\mathcal{T}_h|} & \text{if } (g,k)\notin\mathcal{T}_h \end{cases}\\ \tilde{\mathbb{p}}^{LC}_{gkhhc} &= \begin{cases} \mathbb{p}^{LC}_{gkhhc} & \text{if } (g,k)\in\mathcal{T}_h \\ \dfrac{\tilde{\mathbb{p}}^{LC}_{(-\mathcal{T}_h)hhc}}{M^2-|\mathcal{T}_h|} & \text{if } (g,k)\notin\mathcal{T}_h \end{cases}\end{aligned} \tag{49}$$

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

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

in this sense, the two approximations are in line with each other [cf. (8)].

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

A3. The population of continuing units consists of probability classes $U^{q-4,q}_{1O},\ldots,U^{q-4,q}_{CO}$. All units in probability class $U^{q-4,q}_{cO}$ have the same probabilities p^{LC}_{gkhc} and $\mathbb{p}^{LC}_{gkhhc}$ that can be approximated well by $\tilde{p}^{LC}_{gkhc}$ and $\tilde{\mathbb{p}}^{LC}_{gkhhc}$ in (49).

5.3 Approximate bias and variance formulae

5.3.1 Preliminary results

We begin by deriving approximations to $E\left(\hat{Y}^{q-4}_h\right)$, $E\left(\hat{Y}^{q}_h\right)$ and $\tilde{G}^{q,q-4}_h = E\left(\hat{Y}^{q}_h\right)/E\left(\hat{Y}^{q-4}_h\right)$ under assumptions A2'' and A3, based on the exact expressions in (10). The expression for $E\left(\hat{Y}^{q-4}_h\right)$ involves only probabilities from the level matrix. We can therefore proceed analogously to the derivation for $E\left(\hat{Y}^{q-1}_h\right)$ in (35) to obtain:

$$E\left(\hat{Y}^{q-4}_h\right) \approx \sum_{c=1}^{C}\left[\sum_{g\in\mathcal{G}_h} p^{L}_{ghc}Y^{q-4}_{gc} + \tilde{p}^{L}_{(-\mathcal{G}_h)hc}\bar{Y}^{q-4}_{(-\mathcal{G}_h)c}\right]. \tag{50}$$

Here and elsewhere in this section, we will re-use the notation that was introduced in Section 4 for quarterly growth rates, with minor variations if necessary.

For $E\left(\hat{Y}_h^q\right)$, the expression in (10) 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 (50) that

$$E\left(\hat{Y}_{hB}^q\right) \approx \sum_{c=1}^{\mathcal{C}} \left[\sum_{g \in \mathcal{G}_h} p_{ghc}^L Y_{gcB}^q + \tilde{p}_{(-\mathcal{G}_h)hc}^L \bar{Y}_{(-\mathcal{G}_h)cB}^q \right], \tag{51}$$

in obvious notation.

The term $E\left(\hat{Y}_{hO}^q\right)$ involves probabilities p_{gkhi}^{LC}. Using assumption A3, we find:

$$\begin{aligned}
E\left(\hat{Y}_{hO}^q\right) &= \sum_{c=1}^{\mathcal{C}} \sum\nolimits_{i \in U_{cO}^{q-4,q}} y_i^q \sum\nolimits_{g=1}^{M} \sum\nolimits_{k=1}^{M} a_{gi}^{q-4} a_{ki}^q p_{gkhc}^{LC} \\
&\approx \sum_{c=1}^{\mathcal{C}} \sum\nolimits_{i \in U_{cO}^{q-4,q}} y_i^q \left[\sum_{(g,k) \in \mathcal{T}_h} a_{gi}^{q-4} a_{ki}^q p_{gkhc}^{LC} \right. \\
&\qquad \left. + \left(1 - \sum_{(g,k) \in \mathcal{T}_h} a_{gi}^{q-4} a_{ki}^q \right) \frac{\tilde{p}_{(-\mathcal{T}_h)hc}^{LC}}{M^2 - |\mathcal{T}_h|} \right] \\
&= \sum_{c=1}^{\mathcal{C}} \left[\sum_{(g,k) \in \mathcal{T}_h} p_{gkhc}^{LC} Y_{gkcO}^q + \frac{\tilde{p}_{(-\mathcal{T}_h)hc}^{LC}}{M^2 - |\mathcal{T}_h|} \left(Y_{++cO}^q - \sum_{(g,k) \in \mathcal{T}_h} Y_{gkcO}^q \right) \right] \\
&= \sum_{c=1}^{\mathcal{C}} \left[\sum_{(g,k) \in \mathcal{T}_h} p_{gkhc}^{LC} Y_{gkcO}^q + \tilde{p}_{(-\mathcal{T}_h)hc}^{LC} \bar{Y}_{(-\mathcal{T}_h)cO}^q \right].
\end{aligned} \tag{52}$$

In the third line, $Y_{gkcO}^q = \sum_{i \in U_{cO}^{q-4,q}} a_{gi}^{q-4} a_{ki}^q y_i^q$ denotes the total quarterly turnover in q for all continuing units in probability class c that belong to the stratum g in quarter $q-4$ and to the 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 (51) and (52), we obtain:

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

Furthermore, it follows from this expression and (50) that

$$\breve{G}_h^{q,q-4} \approx \frac{1}{\sum_{c=1}^{\mathcal{C}} \left[\sum_{g\in\mathcal{G}_h} p_{ghc}^{L} Y_{gc}^{q-4} + \tilde{p}_{(-\mathcal{G}_h)hc}^{L} \bar{Y}_{(-\mathcal{G}_h)c}^{q-4}\right]} \times \left\{\sum_{c=1}^{\mathcal{C}} \left[\sum_{(g,k)\in\mathcal{J}_h} p_{gkhc}^{LC} Y_{gkcO}^{q} + \tilde{p}_{(-\mathcal{J}_h)hc}^{LC} \bar{Y}_{(-\mathcal{J}_h)cO}^{q} + \sum_{g\in\mathcal{G}_h} p_{ghc}^{L} Y_{gcB}^{q} + \tilde{p}_{(-\mathcal{G}_h)hc}^{L} \bar{Y}_{(-\mathcal{G}_h)cB}^{q}\right]\right\}. \tag{53}$$

5.3.2 Bias

The bias approximation in formula (12) 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 (53). The component $B_{hO}^{q,q-4}$ refers to continuing units ($U_O^{q-4,q}$). In the derivation of $AB\left(\hat{g}_h^{q,q-1}\right)$, 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:

$$\begin{aligned} B_{hO}^{q,q-4} &= B_{hO1}^{q,q-4} + B_{hO2}^{q,q-4}, \\ B_{hO1}^{q,q-4} &= \breve{G}_h^{q,q-4} \sum_{i\in U_O^{q-4,q}} \left(y_i^{q-4}\right)^2 \sum_{g=1}^{M} a_{gi}^{q-4} p_{ghi}^{L}\left(1 - p_{ghi}^{L}\right), \\ B_{hO2}^{q,q-4} &= -\sum_{i\in U_O^{q-4,q}} y_i^{q-4} y_i^{q} \sum_{g=1}^{M}\sum_{k=1}^{M} a_{gi}^{q-4} a_{ki}^{q}\left(\mathbb{p}_{gkhhi}^{LC} - p_{ghi}^{L} p_{gkhi}^{LC}\right). \end{aligned} \tag{54}$$

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 (50) to obtain:

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

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 Subsection 4.4.

The second sub-term $B_{hO2}^{q,q-4}$ in (54) is slightly more complicated, because it involves on the one hand probabilities p_{ghi}^{L} and on the other hand probabilities p_{gkhi}^{LC} and $\mathbb{p}_{gkhhi}^{LC}$. For p_{ghi}^{L}, 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 $\mathcal{T}_h$ and its complement. Here, both partitions are relevant. We therefore need to define a more elaborate partition of pairs of industries into four subsets:

$$\begin{aligned}
\mathcal{A}_{11h} &= \{(g,k): g \in \mathcal{G}_h, (g,k) \in \mathcal{T}_h\}; \\
\mathcal{A}_{10h} &= \{(g,k): g \in \mathcal{G}_h, (g,k) \notin \mathcal{T}_h\}; \\
\mathcal{A}_{01h} &= \{(g,k): g \notin \mathcal{G}_h, (g,k) \in \mathcal{T}_h\}; \\
\mathcal{A}_{00h} &= \{(g,k): g \notin \mathcal{G}_h, (g,k) \notin \mathcal{T}_h\}.
\end{aligned} \tag{55}$$

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

Using this partition in combination with assumptions A2'' and A3, we can approximate $B_{hO2}^{q,q-4}$ as follows:

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

Here, we used the following extension of the shorthand notation defined in (34):

$$S_{gkcO}(z) = \sum\nolimits_{i \in U_{cO}^{q-4,q}} a_{gi}^{q-4} a_{ki}^{q} z \,,$$
$$\bar{S}_{(\mathcal{A}_{00h})cO}(z) = \frac{\sum_{(g,k)\in\mathcal{A}_{00h}} S_{gkcO}(z)}{|\mathcal{A}_{00h}|} = \frac{S_{++cO}(z) - \sum_{(g,k)\in(\mathcal{A}_{11h}\cup\mathcal{A}_{10h}\cup\mathcal{A}_{01h})} S_{gkcO}(z)}{|\mathcal{A}_{00h}|},$$

with z denoting an arbitrary variable. Note that, by (55),

$$|\mathcal{A}_{00h}| = M^2 - |\mathcal{A}_{11h}| - |\mathcal{A}_{10h}| - |\mathcal{A}_{01h}| = M^2 - |\mathcal{J}_h| - |\mathcal{A}_{10h}|.$$

In practice, the factor $\gamma_h = \frac{|\mathcal{A}_{00h}|}{M^2 - |\mathcal{J}_h|} = 1 - \frac{|\mathcal{A}_{10h}|}{M^2 - |\mathcal{J}_h|}$ is supposed to be close to 1.

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

$$\begin{aligned} B_{hO}^{q,q-4} \approx \sum_{c=1}^{\mathcal{C}} \Bigg\{ & \breve{G}_h^{q,q-4} \Bigg[\sum_{g\in\mathcal{G}_h} p_{ghc}^{L}\big(1 - p_{ghc}^{L}\big) SS_{gcO}(y^{q-4}) \\ & + \tilde{p}_{(-\mathcal{G}_h)hc}^{L} \left(1 - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{L}}{M - |\mathcal{G}_h|}\right) \overline{SS}_{(-\mathcal{G}_h)cO}(y^{q-4}) \Bigg] \\ & - \sum_{(g,k)\in\mathcal{A}_{11h}} \big(\mathbb{p}_{gkhhc}^{LC} - p_{ghc}^{L} p_{gkhc}^{LC}\big) S_{gkcO}(y^{q-4}y^{q}) \\ & - \sum_{(g,k)\in\mathcal{A}_{10h}} \big(\tilde{\mathbb{p}}_{(-\mathcal{J}_h)hhc}^{LC} - p_{ghc}^{L} \tilde{p}_{(-\mathcal{J}_h)hc}^{LC}\big) \frac{S_{gkcO}(y^{q-4}y^{q})}{M^2 - |\mathcal{J}_h|} \\ & - \sum_{(g,k)\in\mathcal{A}_{01h}} \left(\mathbb{p}_{gkhhc}^{LC} - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{L}}{M - |\mathcal{G}_h|} p_{gkhc}^{LC}\right) S_{gkcO}(y^{q-4}y^{q}) \\ & - \left(\tilde{\mathbb{p}}_{(-\mathcal{J}_h)hhc}^{LC} - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{L}}{M - |\mathcal{G}_h|} \tilde{p}_{(-\mathcal{J}_h)hc}^{LC}\right) \gamma_h \bar{S}_{(\mathcal{A}_{00h})cO}(y^{q-4}y^{q}) \Bigg\}. \end{aligned} \tag{56}$$

The term $B_{hD}^{q,q-4}$ in (12) refers to units that no longer exist in the new quarter ($U_D^{q-4,q}$). An approximation to this term can be derived completely analogously to $B_{hO1}^{q,q-4}$. We obtain:

$$\begin{aligned} B_{hD}^{q,q-4} \approx \breve{G}_h^{q,q-4} \sum_{c=1}^{\mathcal{C}} \Bigg[& \sum_{g\in\mathcal{G}_h} p_{ghc}^{L}\big(1 - p_{ghc}^{L}\big) SS_{gcD}(y^{q-4}) \\ & + \tilde{p}_{(-\mathcal{G}_h)hc}^{L} \left(1 - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{L}}{M - |\mathcal{G}_h|}\right) \overline{SS}_{(-\mathcal{G}_h)cD}(y^{q-4}) \Bigg]. \end{aligned} \tag{57}$$

Combining all terms from (56) and (57), we obtain the following approximation to $AB\big(\hat{g}_h^{q,q-4}\big)$:

$$\boxed{\begin{aligned}
AB\left(\hat{g}_h^{q,q-4}\right) &\approx \breve{G}_h^{q,q-4} - G_h^{q,q-4} \\
&\qquad + \frac{\breve{G}_h^{q,q-4}\sum_{c=1}^{\mathcal{C}} B_{hc}^{q,q-4} - \sum_{c=1}^{\mathcal{C}} C_{hcO}^{q,q-4}}{\left\{\sum_{c=1}^{\mathcal{C}}\left[\sum_{g\in\mathcal{G}_h} p_{ghc}^{L} Y_{gc}^{q-4} + \tilde{p}_{(-\mathcal{G}_h)hc}^{L}\bar{Y}_{(-\mathcal{G}_h)c}^{q-4}\right]\right\}^2}, \\
B_{hc}^{q,q-4} &= \sum_{g\in\mathcal{G}_h} p_{ghc}^{L}\left(1-p_{ghc}^{L}\right) BS_{gc}^{q,q-4} \\
&\qquad + \tilde{p}_{(-\mathcal{G}_h)hc}^{L}\left(1-\frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{L}}{M-|\mathcal{G}_h|}\right)\overline{BS}_{(-\mathcal{G}_h)c}^{q,q-4}, \\
BS_{gc}^{q,q-4} &= SS_{gcO}\left(y^{q-4}\right) + SS_{gcD}\left(y^{q-4}\right), \\
\overline{BS}_{(-\mathcal{G}_h)c}^{q,q-4} &= \overline{SS}_{(-\mathcal{G}_h)cO}\left(y^{q-4}\right) + \overline{SS}_{(-\mathcal{G}_h)cD}\left(y^{q-4}\right), \\
C_{hcO}^{q,q-4} &= \sum_{(g,k)\in\mathcal{A}_{11h}} \left(\mathbb{p}_{gkhhc}^{LC} - p_{ghc}^{L} p_{gkhc}^{LC}\right) S_{gkcO}\left(y^{q-4}y^{q}\right) \\
&\qquad + \sum_{(g,k)\in\mathcal{A}_{10h}} \left(\tilde{\mathbb{p}}_{(-\mathcal{J}_h)hhc}^{LC} - p_{ghc}^{L}\tilde{p}_{(-\mathcal{J}_h)hc}^{LC}\right)\frac{S_{gkcO}\left(y^{q-4}y^{q}\right)}{M^2-|\mathcal{J}_h|} \\
&\qquad + \sum_{(g,k)\in\mathcal{A}_{01h}} \left(\mathbb{p}_{gkhhc}^{LC} - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{L}}{M-|\mathcal{G}_h|} p_{gkhc}^{LC}\right) S_{gkcO}\left(y^{q-4}y^{q}\right) \\
&\qquad + \left(\tilde{\mathbb{p}}_{(-\mathcal{J}_h)hhc}^{LC} - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^{L}}{M-|\mathcal{G}_h|}\tilde{p}_{(-\mathcal{J}_h)hc}^{LC}\right)\gamma_h \bar{S}_{(\mathcal{A}_{00h})cO}\left(y^{q-4}y^{q}\right),
\end{aligned}} \tag{58}$$

with $\breve{G}_h^{q,q-4}$ approximated by (53). Note that, similar to (40), 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 (58) 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.

5.3.3 Variance

The variance approximation in formula (13) 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 ($U_O^{q-4,q}$). We can write $V_{hO}^{q,q-4} = V_{hO1}^{q,q-4} + V_{hO2}^{q,q-4} + V_{hO3}^{q,q-4}$, with

$$\begin{aligned}
V_{hO1}^{q,q-4} &= \left(\breve{G}_h^{q,q-4}\right)^2 \sum_{i\in U_O^{q-4,q}} \left(y_i^{q-4}\right)^2 \sum_{g=1}^{M} a_{gi}^{q-4} p_{ghi}^{L}\left(1-p_{ghi}^{L}\right), \\
V_{hO2}^{q,q-4} &= \sum_{i\in U_O^{q-4,q}} \left(y_i^{q}\right)^2 \sum_{g=1}^{M}\sum_{k=1}^{M} a_{gi}^{q-4} a_{ki}^{q} p_{gkhi}^{LC}\left(1-p_{gkhi}^{LC}\right), \\
V_{hO3}^{q,q-4} &= -2\breve{G}_h^{q,q-4} \sum_{i\in U_O^{q-4,q}} y_i^{q-4} y_i^{q} \sum_{g=1}^{M}\sum_{k=1}^{M} a_{gi}^{q-4} a_{ki}^{q}\left(\mathbb{p}_{gkhhi}^{LC} - p_{ghi}^{L} p_{gkhi}^{LC}\right).
\end{aligned} \tag{59}$$

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 Subsection 5.3.2, we obtain:

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

The second sub-term $V_{hO2}^{q,q-4}$ in (59) involves probabilities p_{gkhi}^{LC}. Here, we can proceed in a similar way as for $E\left(\hat{Y}_{hO}^q\right)$ in (52) to obtain (in by now obvious notation):

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

The third and final sub-term $V_{hO3}^{q,q-4}$ in (59) is, again, slightly more complicated, because it involves all probabilities p_{ghi}^L, p_{gkhi}^{LC} and $\mathbb{p}_{gkhhi}^{LC}$. Using the same partition of pairs of industries into four subsets (55) that was used in Subsection 5.3.2, we can derive analogously to the approximation for $B_{hO2}^{q,q-4}$ that:

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

with $C_{hcO}^{q,q-4}$ as defined in (58).

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

$$
\begin{aligned}
V_{hO}^{q,q-4} \approx \sum_{c=1}^{\mathcal{C}} \Bigg\{ & \left(\breve{G}_h^{q,q-4}\right)^2 \Bigg[\sum_{g \in \mathcal{G}_h} p_{ghc}^L \left(1 - p_{ghc}^L\right) SS_{gcO}\left(y^{q-4}\right) \\
& + \tilde{p}_{(-\mathcal{G}_h)hc}^L \left(1 - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^L}{M - |\mathcal{G}_h|}\right) \overline{SS}_{(-\mathcal{G}_h)cO}\left(y^{q-4}\right) \Bigg] \\
& + \sum_{(g,k) \in \mathcal{T}_h} p_{gkhc}^{LC} \left(1 - p_{gkhc}^{LC}\right) SS_{gkcO}\left(y^q\right) - 2\breve{G}_h^{q,q-4} C_{hcO}^{q,q-4} \\
& + \tilde{p}_{(-\mathcal{T}_h)hc}^{LC} \left(1 - \frac{\tilde{p}_{(-\mathcal{T}_h)hc}^{LC}}{M^2 - |\mathcal{T}_h|}\right) \overline{SS}_{(-\mathcal{T}_h)cO}\left(y^q\right) \Bigg\}.
\end{aligned} \tag{60}
$$

The remaining two terms in (13) are now straightforward. The term $V_{hD}^{q,q-4}$ refers to units that no longer exist in quarter q ($U_D^{q-4,q}$). Analogously to $V_{hO1}^{q,q-4}$, we find:

$$
\begin{aligned}
V_{hD}^{q,q-4} \approx \left(\breve{G}_h^{q,q-4}\right)^2 \sum_{c=1}^{\mathcal{C}} \Bigg[& \sum_{g \in \mathcal{G}_h} p_{ghc}^L \left(1 - p_{ghc}^L\right) SS_{gcD}\left(y^{q-4}\right) \\
& + \tilde{p}_{(-\mathcal{G}_h)hc}^L \left(1 - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^L}{M - |\mathcal{G}_h|}\right) \overline{SS}_{(-\mathcal{G}_h)cD}\left(y^{q-4}\right) \Bigg].
\end{aligned} \tag{61}
$$

The term $V_{hB}^{q,q-4}$ refers to new-born units ($U_B^{q-4,q}$). For this term, we find:

$$
\begin{aligned}
V_{hB}^{q,q-4} \approx \sum_{c=1}^{\mathcal{C}} \Bigg[& \sum_{g \in \mathcal{G}_h} p_{ghc}^L \left(1 - p_{ghc}^L\right) SS_{gcB}\left(y^q\right) \\
& + \tilde{p}_{(-\mathcal{G}_h)hc}^L \left(1 - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^L}{M - |\mathcal{G}_h|}\right) \overline{SS}_{(-\mathcal{G}_h)cB}\left(y^q\right) \Bigg].
\end{aligned} \tag{62}
$$

Upon combining (60), (61), and (62), we obtain the following variance approximation:

$$
\boxed{
\begin{aligned}
AV\left(\hat{g}_h^{q,q-4}\right) &\approx \frac{\sum_{c=1}^{\mathcal{C}} \left(V_{1hc}^{q,q-4} + V_{2hcO}^{q,q-4} - 2\breve{G}_h^{q,q-4} C_{hOc}^{q,q-4}\right)}{\left\{\sum_{c=1}^{\mathcal{C}} \left[\sum_{g \in \mathcal{G}_h} p_{ghc}^L Y_{gc}^{q-4} + \tilde{p}_{(-\mathcal{G}_h)hc}^L \bar{Y}_{(-\mathcal{G}_h)c}^{q-4}\right]\right\}^2}, \\
V_{1hc}^{q,q-4} &= \sum_{g \in \mathcal{G}_h} p_{ghc}^L \left(1 - p_{ghc}^L\right) VS_{gc}^{q,q-4} \\
&\qquad + \tilde{p}_{(-\mathcal{G}_h)hc}^L \left(1 - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^L}{M - |\mathcal{G}_h|}\right) \overline{VS}_{(-\mathcal{G}_h)c}^{q,q-4}, \\
V_{2hcO}^{q,q-4} &= \sum_{(g,k) \in \mathcal{T}_h} p_{gkhc}^{LC} \left(1 - p_{gkhc}^{LC}\right) SS_{gkcO}\left(y^q\right) \\
&\qquad + \tilde{p}_{(-\mathcal{T}_h)hc}^{LC} \left(1 - \frac{\tilde{p}_{(-\mathcal{T}_h)hc}^{LC}}{M^2 - |\mathcal{T}_h|}\right) \overline{SS}_{(-\mathcal{T}_h)cO}\left(y^q\right), \\
VS_{gc}^{q,q-4} &= SS_{gcO}\left(\breve{G}_h^{q,q-4} y^{q-4}\right) + SS_{gcD}\left(\breve{G}_h^{q,q-4} y^{q-4}\right) + SS_{gcB}\left(y^q\right), \\
\overline{VS}_{(-\mathcal{G}_h)c}^{q,q-4} &= \overline{SS}_{(-\mathcal{G}_h)cO}\left(\breve{G}_h^{q,q-4} y^{q-4}\right) + \overline{SS}_{(-\mathcal{G}_h)cD}\left(\breve{G}_h^{q,q-4} y^{q-4}\right) \\
&\qquad + \overline{SS}_{(-\mathcal{G}_h)cB}\left(y^q\right),
\end{aligned}
} \tag{63}
$$

with $C_{hcO}^{q,q-4}$ as defined in (58).

5.4 Estimating the bias and variance

To estimate the bias and variance formulas in Subsection 5.3, we can proceed in a similar way as in Subsection 4.5. Define:

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

Then the bias in (58) can be estimated by:

$$
\begin{aligned}
\widehat{AB}\left(\hat{g}_h^{q,q-4}\right) &= \hat{\hat{G}}_h^{q,q-4} - \frac{\hat{Y}_h^{q}}{\hat{Y}_h^{q-4}}\\
&\quad + \frac{\hat{\hat{G}}_h^{q,q-4}\sum_{c=1}^{\mathcal{C}}\hat{B}_{hc}^{q,q-4} - \sum_{c=1}^{\mathcal{C}}\hat{C}_{hcO}^{q,q-4}}{\left\{\sum_{c=1}^{\mathcal{C}}\left[\sum_{g\in\mathcal{G}_h}p_{ghc}^{L}\hat{Y}_{gc}^{q-4} + \tilde{p}_{(-\mathcal{G}_h)hc}^{L}\hat{\hat{Y}}_{(-\mathcal{G}_h)c}^{q-4}\right]\right\}^2},\\
\hat{\hat{G}}_h^{q,q-4} &= \frac{1}{\left\{\sum_{c=1}^{\mathcal{C}}\left[\sum_{g\in\mathcal{G}_h}p_{ghc}^{L}\hat{Y}_{gc}^{q-4} + \tilde{p}_{(-\mathcal{G}_h)hc}^{L}\hat{\hat{Y}}_{(-\mathcal{G}_h)c}^{q-4}\right]\right\}^2}\\
&\quad \times\left\{\sum_{c=1}^{\mathcal{C}}\left[\sum_{(g,k)\in\mathcal{I}_h}p_{gkhc}^{LC}\hat{Y}_{gkcO}^{q} + \tilde{p}_{(-\mathcal{I}_h)hc}^{LC}\hat{\hat{Y}}_{(-\mathcal{I}_h)cO}^{q}\right.\right.\\
&\quad \left.\left. + \sum_{g\in\mathcal{G}_h}p_{ghc}^{L}\hat{Y}_{gcB}^{q} + \tilde{p}_{(-\mathcal{G}_h)hc}^{L}\hat{\hat{Y}}_{(-\mathcal{G}_h)cB}^{q}\right]\right\}.
\end{aligned} \tag{64}
$$

Here, $\hat{\hat{G}}_h^{q,q-4}$ is defined analogously to $\hat{\hat{G}}_h^{q,q-1}$ in (42), and the estimated sums and sums of squares are defined analogously to (45).

Similarly, the variance in (63) can be estimated by:

$$
\begin{aligned}
\widehat{AV}\left(\hat{g}_h^{q,q-4}\right) &= \frac{\sum_{c=1}^{\mathcal{C}}\left(\hat{V}_{1hc}^{q,q-4} + \hat{V}_{2hcO}^{q,q-4} - 2\hat{\hat{G}}_h^{q,q-4}\hat{C}_{hcO}^{q,q-4}\right)}{\left\{\sum_{c=1}^{\mathcal{C}}\left[\sum_{g\in\mathcal{G}_h} p_{ghc}^L \hat{Y}_{gc}^{q-4} + \tilde{p}_{(-\mathcal{G}_h)hc}^L \hat{\hat{Y}}_{(-\mathcal{G}_h)c}^{q-4}\right]\right\}^2}, \\
\hat{V}_{1hc}^{q,q-4} &= \sum_{g\in\mathcal{G}_h} p_{ghc}^L\left(1 - p_{ghc}^L\right)\widehat{VS}_{gc}^{q,q-4} \\
&\quad + \tilde{p}_{(-\mathcal{G}_h)hc}^L\left(1 - \frac{\tilde{p}_{(-\mathcal{G}_h)hc}^L}{M - |\mathcal{G}_h|}\right)\widehat{\widehat{VS}}_{(-\mathcal{G}_h)c}^{q,q-4}, \\
\hat{V}_{2hcO}^{q,q-4} &= \sum_{(g,k)\in\mathcal{T}_h} p_{gkhc}^{LC}\left(1 - p_{gkhc}^{LC}\right)\widehat{SS}_{gkcO}(y^q) \\
&\quad + \tilde{p}_{(-\mathcal{T}_h)hc}^{LC}\left(1 - \frac{\tilde{p}_{(-\mathcal{T}_h)hc}^{LC}}{M^2 - |\mathcal{T}_h|}\right)\widehat{\widehat{SS}}_{(-\mathcal{T}_h)cO}(y^q), \\
\widehat{VS}_{gc}^{q,q-4} &= \widehat{SS}_{gcO}\left(\hat{\hat{G}}_h^{q,q-4} y^{q-4}\right) + \widehat{SS}_{gcD}\left(\hat{\hat{G}}_h^{q,q-4} y^{q-4}\right) + \widehat{SS}_{gcB}(y^q), \\
\widehat{\widehat{VS}}_{(-\mathcal{G}_h)c}^{q,q-4} &= \widehat{\widehat{SS}}_{(-\mathcal{G}_h)cO}\left(\hat{\hat{G}}_h^{q,q-4} y^{q-4}\right) + \widehat{\widehat{SS}}_{(-\mathcal{G}_h)cD}\left(\hat{\hat{G}}_h^{q,q-4} y^{q-4}\right) \\
&\quad + \widehat{\widehat{SS}}_{(-\mathcal{G}_h)cB}(y^q),
\end{aligned}
\tag{65}
$$

with $\hat{\hat{G}}_h^{q,q-4}$ as defined in (64).

6. The bootstrap approach

In this section, we will consider an alternative approach to estimate the bias $B\left(\hat{g}_h^{q,q-u}\right)$ and variance $V\left(\hat{g}_h^{q,q-u}\right)$ due to classification errors, based on a parametric bootstrap method. We extend the bootstrap method developed in Burger et al. (2015) and van Delden et al. (2016b) for turnover levels to growth rates.

To model the occurrence of classification errors in the observed industry codes, including their dependence over time for continuing units, in Section 2 we introduced two transition matrices: the level matrix $\mathbf{P}_i^L = (p_{ghi}^L)$, with $p_{ghi}^L = P\left(\hat{s}_i^{q-u} = h \middle| s_i^{q-u} = g\right)$, and the change matrix $\mathbf{P}_i^C = (p_{gklhi}^C)$, with $p_{gklhi}^C = P\left(\hat{s}_i^{q} = h \middle| s_i^{q-4} = g, s_i^{q} = k, \hat{s}_i^{q-4} = l\right)$. The relation between the true and observed industry codes for the first two years is illustrated in the left part of Figure 1.

In the bootstrap approach, we start by simulating the situation of the first quarter $q = 1$ in year $t = T_0$. (For clarity, we revert to using the double time index here.) For $q = 1$ we apply the transition matrix $\mathbf{P}_i^L$, as in van Delden et al. (2016b), to the observed $\hat{s}_i^{t,q}$, which results in a new industry assignment variable denoted by $\hat{s}_i^{*t,q}$. That is to say, we consider realisations of the alternative classification error model given by:

$$P\left(\hat{s}_i^{*t,q} = h \middle| \hat{s}_i^{t,q} = g\right) \equiv P\left(\hat{s}_i^{t,q} = h \middle| s_i^{t,q} = g\right) = p_{ghi}^L \qquad (t = T_0).$$

For the other quarters in year T_0, each unit retains the same industry code, so $\hat{s}_i^{*t,q} = h$. For any new-born units again transition matrix $\mathbf{P}_i^L$ is applied to derive $\hat{s}_i^{*t,q}$. Next, for continuing units in the first quarter of the next year ($t = T_0 + 1$, $q = 1$) we apply transition matrix $\mathbf{P}_i^C$ to obtain $\hat{s}_i^{*t,q}$ (with $t = T_0 + 1$) given the values of $\hat{s}_i^{t-1,q}$, $\hat{s}_i^{t,q}$ and $\hat{s}_i^{*t-1,q}$. Thus, likewise to $\mathbf{P}_i^L$, we consider realisations of the alternative classification error model

$$\begin{aligned} &P\left(\hat{s}_i^{*t,q} = h \middle| \hat{s}_i^{t-1,q} = g, \hat{s}_i^{t,q} = k, \hat{s}_i^{*t-1,q} = l\right) \\ &\qquad\qquad \equiv P\left(\hat{s}_i^{t,q} = h \middle| s_i^{t-1,q} = g, s_i^{t,q} = k, \hat{s}_i^{t-1,q} = l\right) = p_{gklhi}^C. \end{aligned}$$

These new codes $\hat{s}_i^{*t,q}$ are again kept fixed for the remaining quarters in year $t = T_0 + 1$. The right part of Figure 1 illustrates the bootstrap for the first two years.

reality

$$\begin{tikzcd}
 & s_i^{T_0,q} \arrow[r] \arrow[d] \arrow[rd] & s_i^{T_0+1,q} \arrow[d] & \\
\mathbf{P}_i^L & \hat{s}_i^{T_0,q} \arrow[r] & \hat{s}_i^{T_0+1,q} & \mathbf{P}_i^C
\end{tikzcd}$$

bootstrap

$$\begin{tikzcd}
 & \hat{s}_i^{T_0,q} \arrow[r] \arrow[d] \arrow[rd] & \hat{s}_i^{T_0+1,q} \arrow[d] & \\
\mathbf{P}_i^L & \hat{s}_i^{*T_0,q} \arrow[r] & \hat{s}_i^{*T_0+1,q} & \mathbf{P}_i^C
\end{tikzcd}$$

Figure 1. Classification errors over time: reality versus bootstrap.

We continue this procedure for $t = T_0, T_0 + 1, \dots$ as a Markov chain, in the sense that estimates for the quarters within the current year t depend on values of the previous year $t - 1$, but not of earlier years. Next, we define: $\hat{a}_{hi}^{*t,q} = I(\hat{s}_i^{*t,q} = h)$ for all t and q in the period under consideration.

The above bootstrap procedure is repeated a large number of times, say R. Denote the bootstrapped stratum indicators in the r^{th} round by $\hat{a}_{hir}^{*q}$ (where the time index t is now suppressed again). We obtain the sequence of bootstrapped turnover levels in industry h: $\hat{Y}_{hr}^{*q} = \sum_{i \in U^q} \hat{a}_{hir}^{*q} y_i^q$. From this, we derive the sequence of growth rates $\hat{g}_{hr}^{*q,q-u} = \hat{Y}_{hr}^{*q} / \hat{Y}_{hr}^{*q-u} - 1$ $(u = 1{,}4)$. Based on these replicated growth rates, the bias and variance of the original estimated growth rates $\hat{g}_h^{q,q-u}$ are then estimated as follows (Efron and Tibshirani, 1993):

$$\hat{B}_R^*\left(\hat{g}_h^{q,q-u}\right) = m_R\left(\hat{g}_h^{*q,q-u}\right) - \hat{g}_h^{q,q-u}; \tag{66}$$

$$\hat{V}_R^*\left(\hat{g}_h^{q,q-u}\right) = \frac{1}{R-1} \sum_{r=1}^{R} \left\{\hat{g}_{hr}^{*q,q-u} - m_R\left(\hat{g}_h^{*q,q-u}\right)\right\}^2; \tag{67}$$

with $m_R\left(\hat{g}_h^{*q,q-u}\right) = R^{-1} \sum_{r=1}^{R} \hat{g}_{hr}^{*q,q-u}$.

With the bootstrap approach, there is no appreciable difference in complexity between estimating the bias of a growth rate with respect to the previous quarter and with respect to the same quarter in the previous year. This is clearly different for the analytical approach (compare Sections 4 and 5). Furthermore, the bootstrap can also be used to estimate the bias and variance of other statistics than growth rates, completely analogously to (66) and (67). With the analytical approach, bias and variance approximations have to be derived separately for any new type of statistic. Thus, the bootstrap may offer some practical advantages for bias and variance estimation. The main drawback of the bootstrap in comparison to the analytical approach is that it requires much more computational effort and memory. For large populations, dedicated hardware may be required to execute the bootstrap in a reasonable amount of time. Note that, in principle, parallel processing of the bootstrap resamples can be used to reduce the overall computation time.

7. Classification error models

7.1 Introduction

In this section, we introduce specific model assumptions about classification error probabilities. These will be used in particular in the case study of Section 9, about the effect of errors in NACE codes in the Dutch GBR on estimated growth rates of turnover per industry. In this application, the total number of industries in $\mathcal{H}_{\text{full}}$ is large ($M \approx 300$). We restrict ourselves to estimating the accuracy of growth rates for a subset of nine target industries. We do take the effect of misclassifications between target and non-target industries into account. In what follows, we use $\mathcal{H} = \{1, \dots, H\}$ to denote the set of target industries, for which we want to estimate $B\left(\hat{g}_h^{q,q-u}\right)$ and $V\left(\hat{g}_h^{q,q-u}\right)$, and $\mathcal{H}_{\text{full}} \backslash \mathcal{H}$ to denote the other industries.

We will introduce assumptions to further restrict the number of unknown parameters in the level matrix $\mathbf{P}_i^L$ and the change matrix $\mathbf{P}_i^C$, in Subsection 7.2 and Subsection 7.3, respectively. This is relevant for both the analytical approach of Sections 2–5 and the bootstrap approach of Section 6. In fact, to estimate the probabilities in these matrices, we propose to use audit samples of units for which both $\hat{s}_i^q$ and s_i^q are observed (see Section 9). As such an audit sample is typically small, parsimonious models are needed for the classification error probabilities.

7.2 Model assumptions for the level matrix

An approach to model and estimate the transition matrix $\mathbf{P}_i^L$ has been described previously in van Delden et al. (2016b), so we only discuss this briefly here. The idea is to divide the level matrix into three parts: (1) the diagonal elements within $\mathcal{H}$ ($h = 1, \dots, H$); (2) the off-diagonal elements within $\mathcal{H}$; and (3) the rows and columns that belong to $\mathcal{H}_{\text{full}} \backslash \mathcal{H}$. The latter set of strata is initially collapsed to one stratum, denoted as stratum $H + 1$; see Table 2 (and compare with Table 1).

Table 2. Transition probabilities for level matrix $\mathbf{P}_i^L$, collapsed version.

True industry	Observed industry				
	1	2		H	$H+1$
1	p_{11}^L	p_{12}^L	$\cdots$	p_{1H}^L	$p_{1,H+1}^L$
2	p_{21}^L	p_{22}^L	$\cdots$	p_{2H}^L	$p_{2,H+1}^L$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
H	p_{H1}^L	p_{H2}^L	$\cdots$	p_{HH}^L	$p_{H,H+1}^L$
$H+1$	$p_{H+1,1}^L$	$p_{H+1,2}^L$	$\cdots$	$p_{H+1,H}^L$	$p_{H+1,H+1}^L$

7.2.1 The diagonal elements

For the diagonal elements within $\mathcal{H}$, we propose to estimate the probability π_i of unit i to be classified correctly, $\pi_i = P\left(\hat{s}_i^q = g \middle| s_i^q = g\right)$, by means of a logistic regression on a number of independent variables (McCullagh and Nelder, 1989):

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}. \tag{68}$$

In the application to industry codes in van Delden et al. (2016b), three independent variables were used for each business: its size class (based on the number of employees), the number of legal units of which it consists (as registered at the Chamber of Commerce), and its observed industry code in the GBR. The probabilities π_i were estimated using an audit sample of units with observed industry codes in the target set $\mathcal{H}$ that was drawn on 1 July 2014, further referred to as the '2014 audit sample'. For the '2014 audit sample', those three variables described the π_i sufficiently well (van Delden et al., 2016b).

7.2.2 The off-diagonal elements

For the off-diagonal elements, the starting point is $1 - \pi_i$, which stands for the probability that the observed industry code is misclassified. Given that a unit is misclassified, we assume that the probability distribution over the other observed industry codes can be written as (van Delden et al., 2016b):

$$P\left(\hat{s}_i^q = h \middle| s_i^q = g, \hat{s}_i^q \neq g\right) = \frac{P\left(\hat{s}_i^q = h \middle| s_i^q = g\right)}{1-\pi_i} \equiv \psi(g,h), \quad (g \neq h), \tag{69}$$

where the conditional probabilities $\psi(g,h)$ are the same for all units. So in contrast to the diagonal elements, we assume that the off-diagonal elements do not depend on any property of the unit, such as its size class, other than its true industry code. A log-linear model may then be used to describe the numbers of misclassified units in the off-diagonal cells. To further reduce the number of parameters, van Delden et al. (2016b) suggested to group the off-diagonal cells into a limited number of clusters $q = 1, \dots Q$, where cells within the same cluster are supposed to have a comparable probability of misclassification. This leads to a non-saturated log-linear model:

$$\log m_{gh} = u + u_{2(h)} + \sum_{q=1}^{Q} \delta_q(g,h) u_{3(q)}, \quad (g \neq h), \tag{70}$$

where m_{gh} denotes the expected number of units with $\left(s_i^q = g, \hat{s}_i^q = h\right)$ and $\delta_q(g,h)$ is a 0-1 indicator that equals 1 for those cells that belong to cluster q. Having estimated this log-linear model, we can estimate $\psi(g,h)$ by $\hat{\psi}(g,h) = \hat{m}_{gh} / \sum_{h' \neq g} \hat{m}_{gh'}$, with $\hat{m}_{gh} = \exp\{\hat{u} + \hat{u}_{2(h)} + \sum_{q=1}^{Q} \delta_q(g,h) \hat{u}_{3(q)}\}$ $(g \neq h)$. See van Delden et al. (2016b) for more details.

7.2.3 The last row and column (non-target strata)

Next, we consider the overall probabilities $p_{H+1,h}^L$ and $p_{g,H+1}^L$ in the last row and column of Table 2. Units in the last row are classified within the target set of industries $\mathcal{H}$ but in reality belong outside this set. Such units can occur in an audit sample that is drawn from $\mathcal{H}$ as observed in the GBR, so the corresponding probabilities $p_{H+1,h}^{L}$ can be estimated from this sample, e.g., using the log-linear model (70). (Recall that the probability $p_{H+1,h}^{L}$ only stands for the total probability that a unit which in reality does not belong to the target population is observed in stratum h of the target population. How to divide this probability over a range of specific industries is treated in the next subsection.) Units in the last column, on the other hand, belong to one of the target industries in reality but are observed in one of the (many) non-target industries. Direct estimation of the probabilities $p_{g,H+1}^{L}$ in the last column would therefore require an additional, infeasibly large audit sample from $\mathcal{H}_{\text{full}} \backslash \mathcal{H}$ as observed in the GBR. Instead we propose an indirect approach to approximate the probabilities in the last column, given that we have estimated the probabilities in the last row.

Let $B = \sum_{g=1}^{H} N_{g,H+1} / \sum_{h=1}^{H} N_{H+1,h}$ denote the (unknown) ratio between the total number of "missed units" in the true industries $\{1, \ldots, H\}$ and the number of "excess units" in the observed industries $\{1, \ldots, H\}$. As noted above, the sum $\sum_{h=1}^{H} N_{H+1,h}$ can be estimated from the '2014 audit sample'. For a given value of B, we may then estimate the sum $\sum_{g=1}^{H} N_{g,H+1}$ as $B \sum_{h=1}^{H} N_{H+1,h}$. We propose to approximate B by the corresponding ratio of observed yearly transitions in the GBR, which is the ratio between the numbers of units that enter and leave the target industries. Given an estimate for $\sum_{g=1}^{H} N_{g,H+1}$, we can estimate the probabilities $p_{g,H+1}^{L}$ using the log-linear model (70); see van Delden et al. (2016b) for details. In that paper it was assumed that $B = 1$.

7.2.4 Consequences for bootstrapping

For the bootstrap simulations, the probability $p_{g,H+1}^{L} = P(\hat{s}_i \in \mathcal{H}_{\text{full}} \backslash \mathcal{H} | s_i = g)$ refers to the event that a unit from a given target industry g is observed in an unspecified industry outside the target set ("missed turnover"). There is no need to specify the observed industry code for these units in more detail, as they do not contribute to the bootstrap turnover replicates $\hat{Y}_{hr}^{*q}$ for the target industries. Likewise, the probability $p_{H+1,h}^{L} = P(\hat{s}_i = h | s_i \in \mathcal{H}_{\text{full}} \backslash \mathcal{H})$ refers to the event that a unit from an unspecified industry outside the target set is observed in a given target industry h. For this "excess turnover" we do require a further specification, because these units contribute to the bootstrap turnover replicates $\hat{Y}_{hr}^{*q}$ for the target industries. Moreover, the properties of such a unit may depend on its actual industry.

For the case study to be discussed below in Section 9, van Delden et al. (2016b) found that erroneously included units originated from a wide range of non-target industries with different turnover distributions. In that paper, it was assumed that the relative number of units from each non-target industry $g \in \mathcal{H}_{\text{full}} \backslash \mathcal{H}$ that are erroneously observed in a given target industry $h \in \mathcal{H}$ is proportional to the number of corresponding yearly transitions observed in the GBR. The associated excess turnover values were obtained simply by drawing from a log-normal distribution. The latter step is not easily extended to time-related classification errors because in the current study we need a time series of turnover values for each unit. Therefore, we introduce an alternative approach here.

We propose to make use of the actual units in a given year that according to the GBR were observed outside the target set. These units encompass the empirical distribution of (potentially) erroneously observed units within the target industries. Within each bootstrap iteration, we extend our observed population of units by drawing a bootstrap sample (with replacement) from this empirical distribution. We do this in such a way that the number of erroneously observed units is $1/B$ times the number of missed units of that industry in the current bootstrap resample. Note that for these units we use a non-parametric bootstrap in this step.

The procedure can in principle be repeated for multiple years. However, for practical reasons, we will limit the procedure to sets of two subsequent years. The reason is that the simplifications we used to handle the missed units and the erroneously included units become less realistic over time.

For the analytical approximations, we also need to specify a detailed set of probabilities p_{gh}^{L} for $g \in \mathcal{H}_{\text{full}} \backslash \mathcal{H}$ and $h \in \mathcal{H}$. The solution we used for the case study will be discussed in Section 9.

7.3 Model assumptions for the change matrix

7.3.1 The model

The probabilities $p_{gklhi}^{C} = P\left(\hat{s}_i^q = h \middle| s_i^{q-4} = g, s_i^q = k, \hat{s}_i^{q-4} = l\right)$ in the matrix $\mathbf{P}_i^C$ can be grouped into four situations (A–D), given the values for s_i^{q-4}, s_i^q and $\hat{s}_i^{q-4}$. We take the true industry code in the current year, so s_i^q, as the starting point. Next we consider whether the true industry code is the same as the one in the previous year ($s_i^q = s_i^{q-4}$, situations A and B) or not ($s_i^q \neq s_i^{q-4}$, situations C and D). Further, we regard whether last year's observed industry code is now correct $\hat{s}_i^{q-4} = s_i^q$ (situations A and D) or not $\hat{s}_i^{q-4} \neq s_i^q$ (situations B and C). The logic behind this approach is that the GBR aims to obtain the currently correct industry code when the codes are updated between December and January. Thus when the previously observed industry code is now correct ($\hat{s}_i^{q-4} = s_i^q$), there is in fact no need to change the observed industry code, so the correct transition would be $\hat{s}_i^q = \hat{s}_i^{q-4}$. However, when $\hat{s}_i^{q-4} \neq s_i^q$ the GBR should change its observed industry code into the true value: $\hat{s}_i^q = s_i^q$ and $\hat{s}_i^q \neq \hat{s}_i^{q-4}$.

In summary, we have the following four situations (see also Table 3):

A. $s_i^q = s_i^{q-4}$ and $\hat{s}_i^{q-4} = s_i^q$: "no change in true industry, the previously observed code is now correct";
B. $s_i^q = s_i^{q-4}$ and $\hat{s}_i^{q-4} \neq s_i^q$: "no change in true industry, the previously observed code is now incorrect";
C. $s_i^q \neq s_i^{q-4}$ and $\hat{s}_i^{q-4} \neq s_i^q$: "change in true industry, the previously observed code is now incorrect";
D. $s_i^q \neq s_i^{q-4}$ and $\hat{s}_i^{q-4} = s_i^q$: "change in true industry, the previously observed code is now correct".

Within these four situations different events may occur. When $\hat{s}^q = \hat{s}^{q-4}$ we say that the observed industry is UNCHANGING (event U). In case the observed industry code changes ($\hat{s}^q \neq \hat{s}^{q-4}$) there are several possibilities for the transition $\hat{s}^{q-4} \rightarrow \hat{s}^q$. Each of these has a certain probability of occurrence, depending on the situation. We model the probability for $\hat{s}^q \neq \hat{s}^{q-4}$ for three events:

- NOTICE (N) a true change in industry. We denote the probability of this event — given that $s^q \neq s^{q-4}$ — by p_N. When this event occurs then $\hat{s}^q = s^q$.
- RESTORE (R) an industry error that was present in $q-4$ (for instance when the true industry code had changed in the past but that change was not noticed in the GBR at the time). The probability that this event occurs — given that $s^q = s^{q-4}$ *and* $\hat{s}^{q-4} \neq s^{q-4}$ — is p_R. When the event occurs then $\hat{s}^q = s^q$.
- SPURIOUS CHANGE (S) of the observed industry. This event can occur under any condition, with probability p_S. The newly observed industry code $\hat{s}^q$ is drawn from a transition matrix with elements $\rho(j,k) = P(\hat{s}_i^q = k | \hat{s}_i^{q-4} = j, \text{S occurs})$ where $\rho(j,j) = 0$. This event covers all changes without a clear explanation.

Table 3. Four situations for s_i^{q-4}, s_i^q and $\hat{s}_i^{q-4}$.

Sit.	s^{q-4}	s^q	$\hat{s}^{q-4}$	Possible events			Probability
				Nr.	$\hat{s}^q$	Event	
A	j	j	j	1	j	U	$1-p_S$
				2	k	S	$p_S\rho(j,k)$
B	j	j	k	1	k	U	$\frac{(1-p_R)(1-p_S)}{1-p_Rp_S}$
				2	j	R or S	$\frac{p_R(1-p_S)+(1-p_R)p_S\rho(k,j)}{1-p_Rp_S}$
				3	l	S	$\frac{(1-p_R)p_S\rho(k,l)}{1-p_Rp_S}$
C	j	l	k or j	1	k or j	U	$\frac{(1-p_N)(1-p_S)}{1-p_Np_S}$
				2	l	N or S	$\frac{p_N(1-p_S)+(1-p_N)p_S\rho(k,l)}{1-p_Np_S}$
				3	r	S	$\frac{(1-p_N)p_S\rho(k,r)}{1-p_Np_S}$
D	j	k	k	1	k	U	$1-p_S$
				2	l	S	$p_S\rho(k,l)$

Legend: grey = no change in true industry, blue = change in true industry, green = correct observed industry, red = incorrect observed industry. The colouring of $\hat{s}^{q-4}$ and $\hat{s}^q$ is relative to the value of s^q.

For each of the four situations, the different events that may occur are shown in Table 3. For instance, in situation A, after the transition $\hat{s}^{q-4} \rightarrow \hat{s}^q$ the observed industry code may be correct ($\hat{s}^q = s^q$), corresponding to event U, or incorrect ($\hat{s}^q \neq s^q$), corresponding to event S. Notice that event R can occur only for units in situation

B. Event N only applies to units in situation C.[1] Events S and U can occur for any unit. We further assume that events N, R and S occur independently, with the restriction that at most one event can occur in the transition from $q-4$ to q. The probability that two (or more) events occur simultaneously is small anyway, but the model becomes easier to analyse if we simply exclude that option.

We have worked out the probabilities of different possible events within the situations A–D, in terms of the three parameters p_N, p_R and p_S. The result is shown in the final column of Table 3. Note that within each of the situations A, B, C, and D, the probabilities sum to 1. From these probabilities, all corresponding elements of the matrix $\mathbf{P}_i^C$ can be derived. A single formula for all elements of $\mathbf{P}_i^C$ under the model defined in Table 3 is provided in Appendix C, using matrix notation.

In the description of the model so far, we have assumed a single set of parameters (p_R, p_N, p_S) and $\rho(j,k)$ that applies to all units in the population. A more realistic version of the model is obtained by dividing the population into probability classes $U_{cO}^{q-4,q}$ as defined in Subsection 5.2, and assuming that all units $i \in U_{cO}^{q-4,q}$ have the same parameters (p_{Rc}, p_{Nc}, p_{Sc}). This form of the model will be used in the case study in Section 9.

7.3.2 Estimating the parameters

The model for $\mathbf{P}_i^C$ of Subsection 7.3.1 can be estimated from an audit sample where for each sampled unit the values for $\hat{s}_i^q$, $\hat{s}_i^{q-4}$ and s_i^q, s_i^{q-4} are available. We propose to estimate only the parameters p_N, p_R and p_S from the audit sample. Since the events are rare, a small audit sample is insufficient to estimate all of the conditional transition probabilities $\rho(j,k)$. Instead, we approximate all $\rho(j,k)$ by the relative frequencies of observed changes within the GBR. So we assume that the $\rho(j,k)$ for the true industry codes are close to those of the observed industry codes.

The parameters p_N, p_R and p_S can be estimated by maximum likelihood (ML) under a multinomial distribution for the number of observed events of each type for each situation in Table 3. Maximising the likelihood function of the observed data directly is complicated, because there are two cases for which it is not clear from the observed values which event occurred (see Table 3): event R or S for “Situation B – Possible event 2” and event N or S for “Situation C – Possible event 2”. By introducing two latent binary variables that indicate which event occurred in these situations, we obtain a complete-data likelihood function that is easy to maximise. An Expectation-Maximisation (EM) algorithm (Little and Rubin, 2002) can then be used to obtain ML estimates for p_N, p_R and p_S.

In this case, the EM algorithm works as follows. For notational simplicity, we assume that the probabilities are estimated for a single probability class. Denote the number of sampled units (within the probability class) for situation $X \in \{A, B, C, D\}$ as n_X.

[1] For units in situation D, doing nothing actually has the same effect as event N. Somewhat arbitrarily, we restrict the event N to apply only to units in situation C. It turns out that this greatly simplifies the estimation of the model.

Further, let $n = n_A + n_B + n_C + n_D$ denote the total sample size and let n_0 be the number of sampled units with $\hat{s}_i^{q-4} \neq \hat{s}_i^q$. The estimated population equivalents (after multiplying by the sampling weights w_i) are $\widehat{N} = \widehat{N}_A + \widehat{N}_B + \widehat{N}_C + \widehat{N}_D$ and $\widehat{N}_0$. As a further specification, let $\widehat{N}_{Ax}$, $\widehat{N}_{Bx}$, $\widehat{N}_{Cx}$ and $\widehat{N}_{Dx}$ denote the weighted number of sampled units for which possible event x within a given situation is observed, according to the numbering of events in Table 3; hence, e.g., $\widehat{N}_A = \sum_{x=1}^{2} \widehat{N}_{Ax}$.

To initialise the EM algorithm we use starting values $p_N^{(0)} = \widehat{N}_{C2}/\widehat{N}_C$, $p_R^{(0)} = \widehat{N}_{B2}/\widehat{N}_B$ and $p_S^{(0)} = \left(\widehat{N}_{A2} + \widehat{N}_{D2}\right)/\left(\widehat{N}_A + \widehat{N}_D\right)$. Next, the following E and M steps are repeated in turn until the parameter estimates for p_N, p_R and p_S have converged.

E step. Given the current parameter estimates $p_N^{(\tau)}$, $p_R^{(\tau)}$ and $p_S^{(\tau)}$, compute

$$\begin{aligned} M_B^{(\tau)} &\equiv M_B\left(p_R^{(\tau)}, p_S^{(\tau)}\right) \\ &= \sum_{\substack{i \in s_B: \\ \hat{s}_i^q = s_i^q}} w_i \frac{p_R^{(\tau)}\left[1 - p_S^{(\tau)}\right]}{p_R^{(\tau)}\left[1 - p_S^{(\tau)}\right] + \left[1 - p_R^{(\tau)}\right] p_S^{(\tau)} \rho\left(\hat{s}_i^{q-4}, \hat{s}_i^q\right)} \end{aligned} \tag{71}$$

and

$$\begin{aligned} M_C^{(\tau)} &\equiv M_C\left(p_N^{(\tau)}, p_S^{(\tau)}\right) \\ &= \sum_{\substack{i \in s_C: \\ \hat{s}_i^q = s_i^q}} w_i \frac{p_N^{(\tau)}\left[1 - p_S^{(\tau)}\right]}{p_N^{(\tau)}\left[1 - p_S^{(\tau)}\right] + \left[1 - p_N^{(\tau)}\right] p_S^{(\tau)} \rho\left(\hat{s}_i^{q-4}, \hat{s}_i^q\right)}, \end{aligned} \tag{72}$$

where $M_B^{(\tau)}$ and $M_C^{(\tau)}$ denote the expected numbers within situation B2 and C2 (given the current parameter estimates) where the first event occurs (R or N).

M step. Given the expected numbers $M_B^{(\tau)}$ and $M_C^{(\tau)}$ from the previous E step, compute

$$\begin{aligned} p_N^{(\tau+1)} &= \frac{M_C^{(\tau)}\left[\widehat{N} - M_B^{(\tau)} - M_C^{(\tau)}\right]}{M_C^{(\tau)}\left[\widehat{N} - M_B^{(\tau)} - M_C^{(\tau)}\right] + \left[\widehat{N}_C - M_C^{(\tau)}\right]\left(\widehat{N} - \widehat{N}_0\right)}, \\ p_R^{(\tau+1)} &= \frac{M_B^{(\tau)}\left[\widehat{N} - M_B^{(\tau)} - M_C^{(\tau)}\right]}{M_B^{(\tau)}\left[\widehat{N} - M_B^{(\tau)} - M_C^{(\tau)}\right] + \left[\widehat{N}_B - M_B^{(\tau)}\right]\left(\widehat{N} - \widehat{N}_0\right)}, \\ p_S^{(\tau+1)} &= \frac{\widehat{N}_0 - M_B^{(\tau)} - M_C^{(\tau)}}{\widehat{N} - M_B^{(\tau)} - M_C^{(\tau)}}. \end{aligned} \tag{73}$$

8. Simulation study

8.1 Introduction and set-up

To test the validity of the analytical formulas derived in Sections 4 and 5 in a realistic setting, we conducted a small simulation study. The target population for this study consisted of three strata ($M = 3$) and two probability classes ($\mathcal{C} = 2$).

The population was constructed by sampling from the Dutch GBR. In particular:

- The two probability classes were constructed by assigning all units with fewer than 20 employees to the first class and all other units to the second class.
- The three strata were constructed by selecting 50 units at random from each probability class in each of three particular industries in the Dutch GBR. Thus, the total number of units in the population was 3 × 2 × 50 = 300. The selection was restricted to units that existed and remained in the same industry and probability class for all eight quarters in 2014 and 2015.
- By construction, the above population does not contain any units that change between strata. To simulate change, the original industry code in 2015 was changed arbitrarily to one of the other codes for 10 randomly selected units in the population. Thus, these units change from one stratum to another between the fourth quarter of 2014 and the first quarter of 2015.

We examined three different set-ups to generate turnover values for the units in the target population:

a. Use the actual observed turnover values of the selected units for eight quarters in 2014 and 2015.
b. Use the actual observed turnover values of the selected units for eight quarters in 2014 and 2015, but construct the population using only "simple" units (see Section 9 for details on the distinction between “simple”, “complex” and “most complex” units in the Dutch GBR).
c. Draw turnover values from a normal distribution.

For set-up (c), the eight quarterly turnover values for each unit were drawn from a multivariate normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with different parameters for each stratum and probability class. For probability class 1, the mean values $\boldsymbol{\mu}$ in the three strata are shown in Table 4 and the standard deviations for stratum 1, 2 and 3 were chosen as 5, 7 and 10, respectively. The corresponding parameters for probability class 2 were found by multiplying $\boldsymbol{\mu}$ by 2 and multiplying the standard deviations by $\sqrt{2}$. The remaining elements of $\boldsymbol{\Sigma}$ were constructed by setting the correlation between the turnover of the same unit in quarters $q_1, q_2 \in \{1,2, \dots ,8\}$ equal to $(0.90)^{|q_1 - q_2|}$.

Table 4. Mean values of normal distributions for probability class 1 under set-up (c).

Quarter \ Stratum	1	2	3
2014Q1	50	70	100
2014Q2	52	68	105
2014Q3	50	70	110
2014Q4	52	68	115
2015Q1	55	75	120
2015Q2	57	71	125
2015Q3	55	75	130
2015Q4	57	71	135

Figure 2 shows boxplots of the resulting turnover distributions under the three set-ups, for each stratum and probability class, with the values for all eight quarters put together. It is seen that the original turnover distribution (set-up (a)) is extremely skewed, so that the total turnover is dominated by a few units. The restriction to "simple" units under set-up (b) reduces this skewness only to a limited extent. For set-up (c), the turnover distributions are symmetric by construction.

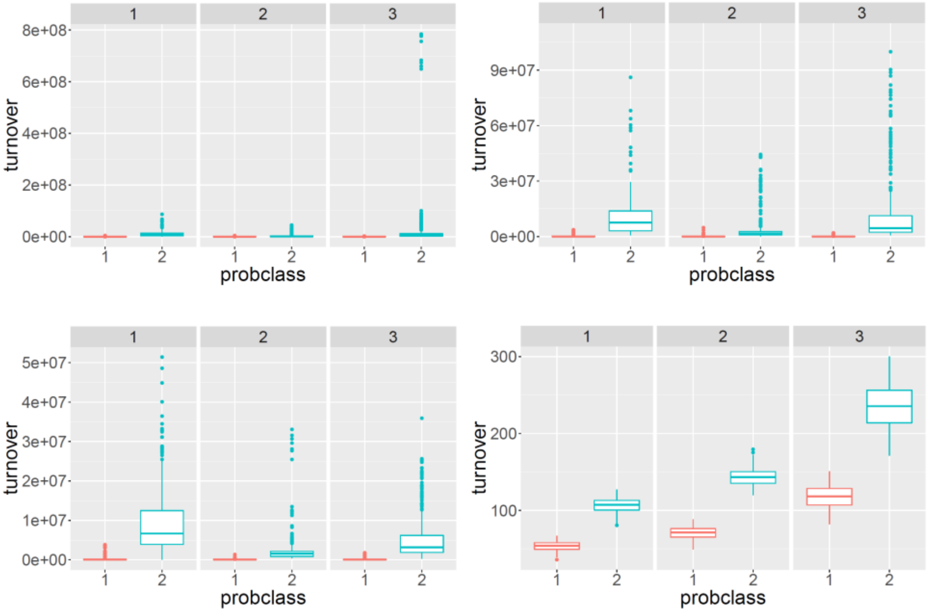


Figure 2. Boxplots of turnover distribution (all eight quarters) for each stratum (1,2,3) and probability class (1,2), for different set-ups. Top row, left panel: set-up (a). Top row, right panel: set-up (a), with turnover values of largest unit in stratum 3 excluded. Bottom row, left panel: set-up (b). Bottom row, right panel: set-up (c).

For each set-up, a single population was created by the above procedure. Next, using this population as the target population, observed stratum codes with classification errors were introduced according to the model of Section 7. For the first year, the following level matrices were used for the two probability classes:

$$\mathbf{P}_1^L = \begin{pmatrix} 0.90 & 0.10 \times 0.70 & 0.10 \times 0.30 \\ 0.20 \times 0.50 & 0.80 & 0.20 \times 0.50 \\ 0.30 \times 0.30 & 0.30 \times 0.70 & 0.70 \end{pmatrix},$$

$$\mathbf{P}_2^L = \begin{pmatrix} 0.95 & 0.05 \times 0.70 & 0.05 \times 0.30 \\ 0.05 \times 0.50 & 0.95 & 0.05 \times 0.50 \\ 0.05 \times 0.30 & 0.05 \times 0.70 & 0.95 \end{pmatrix}.$$

Note that this amounts to using the following matrix of conditional off-diagonal probabilities $\psi(g,h)$ as defined in (69):

$$\begin{pmatrix} - & 0.70 & 0.30 \\ 0.50 & - & 0.50 \\ 0.30 & 0.70 & - \end{pmatrix}.$$

For the transition to the second year, the following probabilities were assigned to the events NOTICE, RESTORE and SPURIOUS CHANGE for the two probability classes:

$$\begin{aligned} (p_{R1}, p_{N1}, p_{S1}) &= (0.10, 0.16, 0.01), \\ (p_{R2}, p_{N2}, p_{S2}) &= (0.70, 0.80, 0.001). \end{aligned}$$

For the spurious changes, the transition probabilities $\rho(j,k)$ were set equal to the corresponding conditional off-diagonal probabilities $\psi(j,k)$. Note that the above classification error probabilities are in line with the idea that the observed stratum codes of larger units (i.e., units in probability class 2) are checked more thoroughly than those of smaller units (i.e., units in probability class 1). The chosen parameter values are in line with the case study described in Section 9. The values of probability class 1 and 2 resemble those of case study probability class a and c respectively, see Table 11 in Subsection 9.4.

For each set-up, we first estimated the true bias and variance of the turnover growth rates $\hat{g}_h^{q,q-1}$ and $\hat{g}_h^{q,q-4}$ ($h \in \{1,2,3\}$) by Monte Carlo simulation. We generated $R = 10{,}000$ random sets of observed stratum codes for the population from the above classification error model and computed the observed growth rates for each set. Since the true growth rates $g_h^{q,q-1}$ and $g_h^{q,q-4}$ were known, the true bias and variance could be estimated analogously to (66) and (67), without additional bootstrapping.

We then computed the analytical approximations $AB\left(\hat{g}_h^{q,q-1}\right)$ and $AV\left(\hat{g}_h^{q,q-1}\right)$ from Subsection 4.4 and $AB\left(\hat{g}_h^{q,q-4}\right)$ and $AV\left(\hat{g}_h^{q,q-4}\right)$ from Subsection 5.3. Since the number of strata was small in this study ($M = 3$), we did not define subsets of 'non-significant' probabilities to be replaced by an average probability (i.e., we included all strata in $\mathcal{G}_h$ and all pairs of strata in $\mathcal{T}_h$).

Note that, as the true stratum codes were known in this simulation study, we did not have to estimate any of the unknown quantities in the analytical formulas. Thus, we could compare the analytical formulas to a Monte Carlo simulation of errors, both evaluated on the true population data. In practice — in particular in the case study of Section 9 — this would usually not be possible and instead we could then compare the *estimated* analytical formulas to a *bootstrap* simulation of errors, both evaluated on the *observed* population data. Since the two situations are equivalent (cf. Figure

1), we expect that they would lead to similar conclusions about the agreement between analytical approximation and direct simulation of bias and variance.

8.2 Results

For set-up (a), Figure 3 shows a comparison between the bias and standard errors of growth rates according to a direct simulation and according to the analytical formulas.

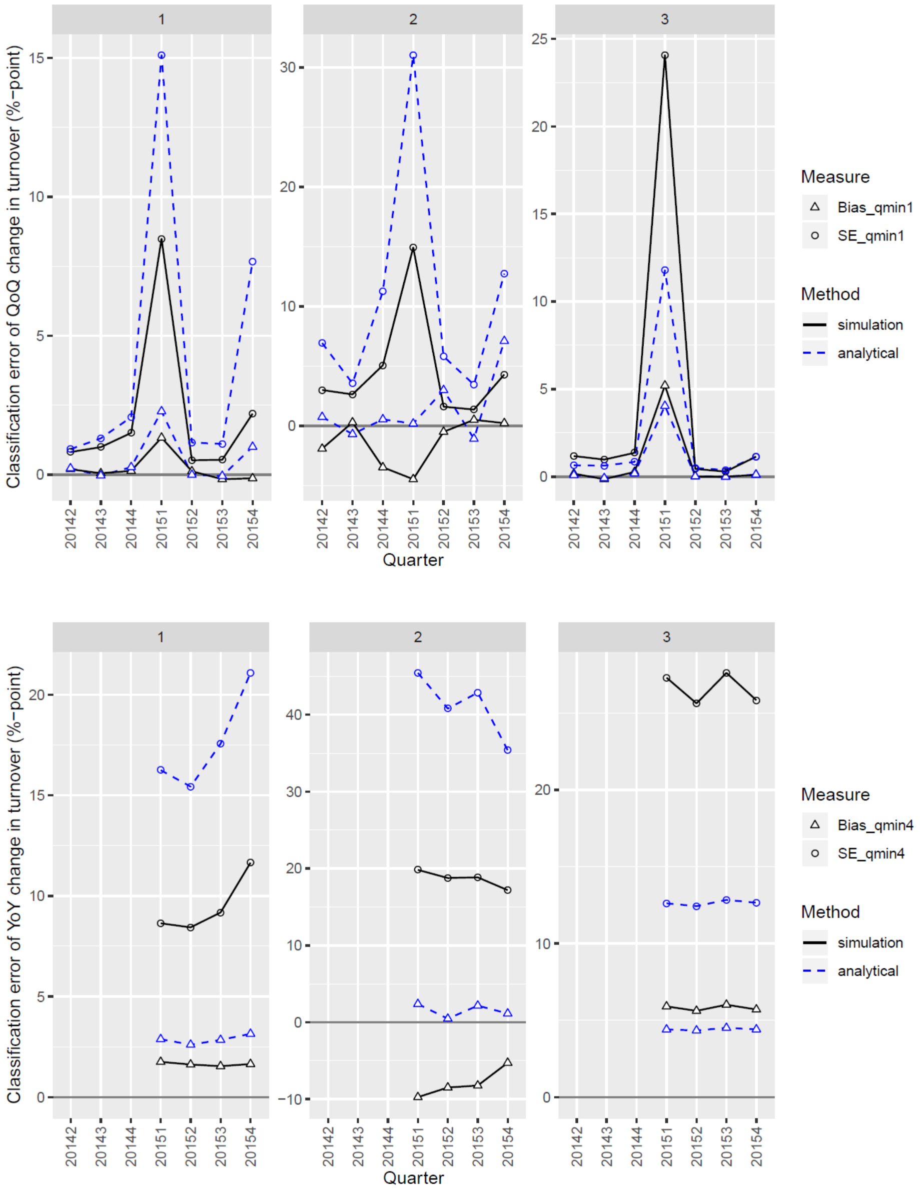


Figure 3. Comparison of bias (triangles) and standard errors (circles) in each stratum (1,2,3) based on direct simulation (solid black lines) and analytical formulas (dashed blue lines) for set-up (a). Top: quarter-on-quarter growth rates. Bottom: year-on-year growth rates.

It is seen that the analytical approximation is often substantially different from the simulated true bias and standard error, in particular for the year-on-year growth rates $\hat{g}_h^{q,q-4}$. The two approaches do agree to some extent on the shape of the development of the bias and standard errors across quarters, but they often disagree on the level. For the standard errors, the formulas yield overestimates in strata 1 and 2 and underestimates in stratum 3 (which contains the largest turnover values).

Figure 4 shows the same results for set-up (b). Recall that the only difference with respect to set-up (a) is that complex units (typically with large turnover values) are not included in the population any more.

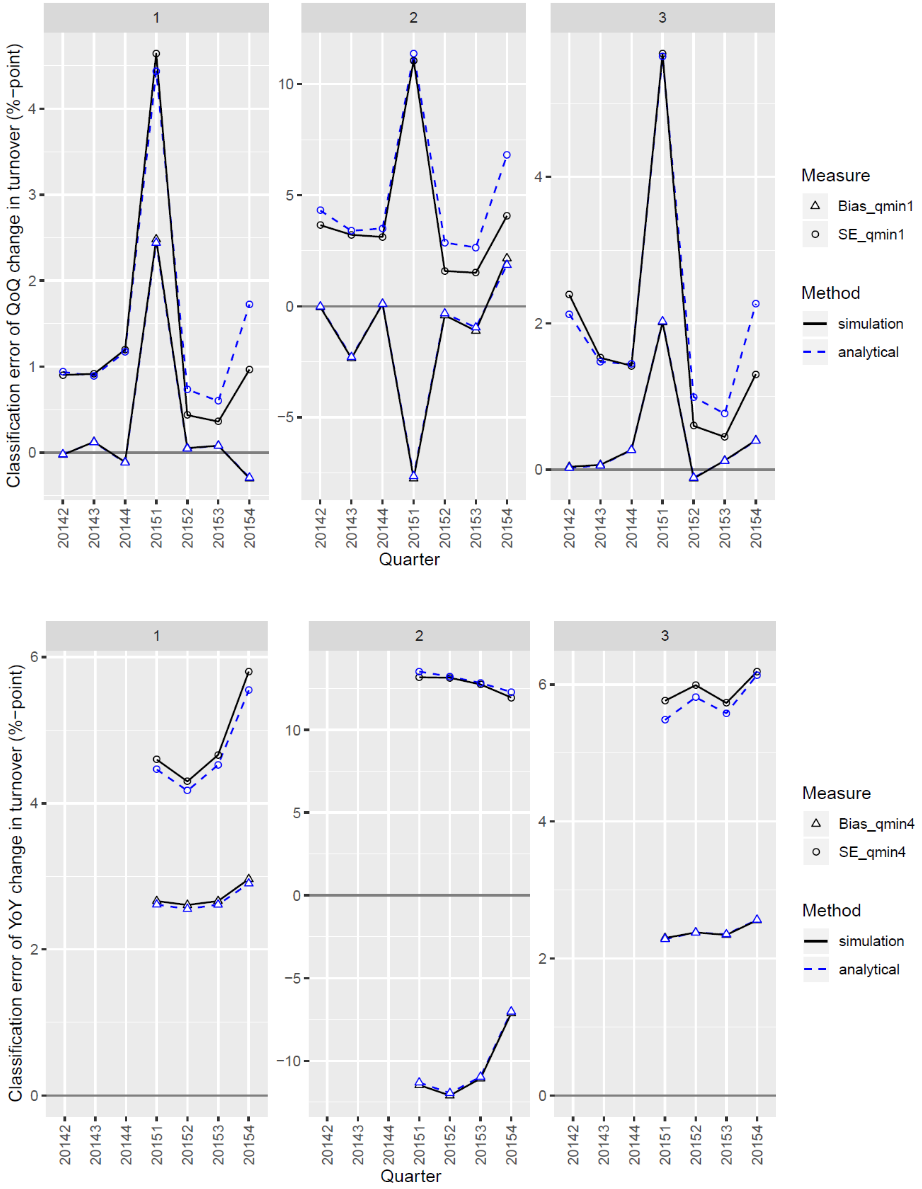


Figure 4. Comparison of bias (triangles) and standard errors (circles) in each stratum (1,2,3) based on direct simulation (solid black lines) and analytical formulas (dashed blue lines) for set-up (b). Top: quarter-on-quarter growth rates. Bottom: year-on-year growth rates.

It is seen that the analytical bias and variance approximations now agree much more closely with direct simulation. This suggests that the bad performance of the analytical formulas for set-up (a) may be related to the extreme skewness of the turnover distributions under that set-up.

This suggestion is confirmed by the results for set-up (c) shown in Figure 5: when the y_i values follow a normal distribution, the analytical bias and variance approximations are in close agreement with the outcome of direct simulation. This result holds both for quarter-on-quarter and year-on-year growth rates.

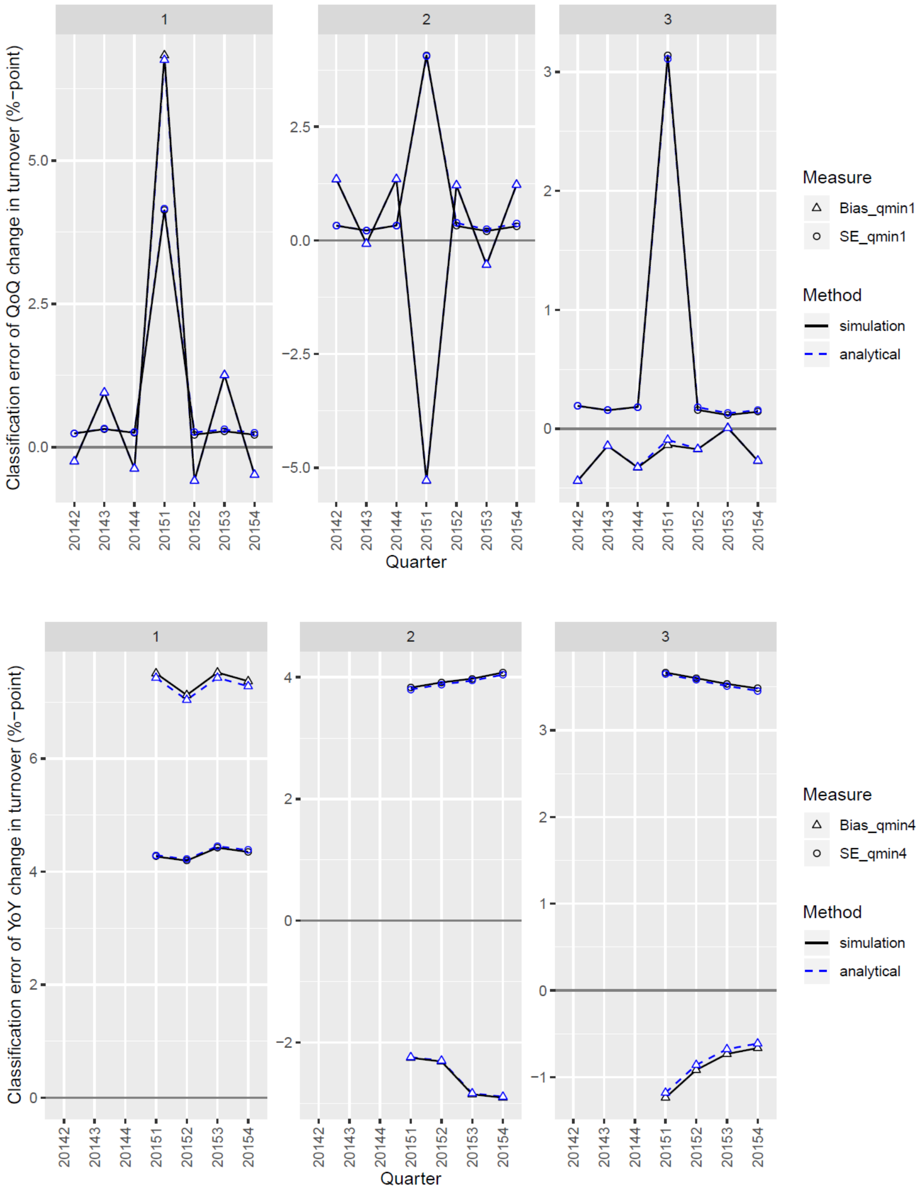


Figure 5. Comparison of bias (triangles) and standard errors (circles) in each stratum (1,2,3) based on direct simulation (solid black lines) and analytical formulas (dashed blue lines) for set-up (c). Top: quarter-on-quarter growth rates. Bottom: year-on-year growth rates.

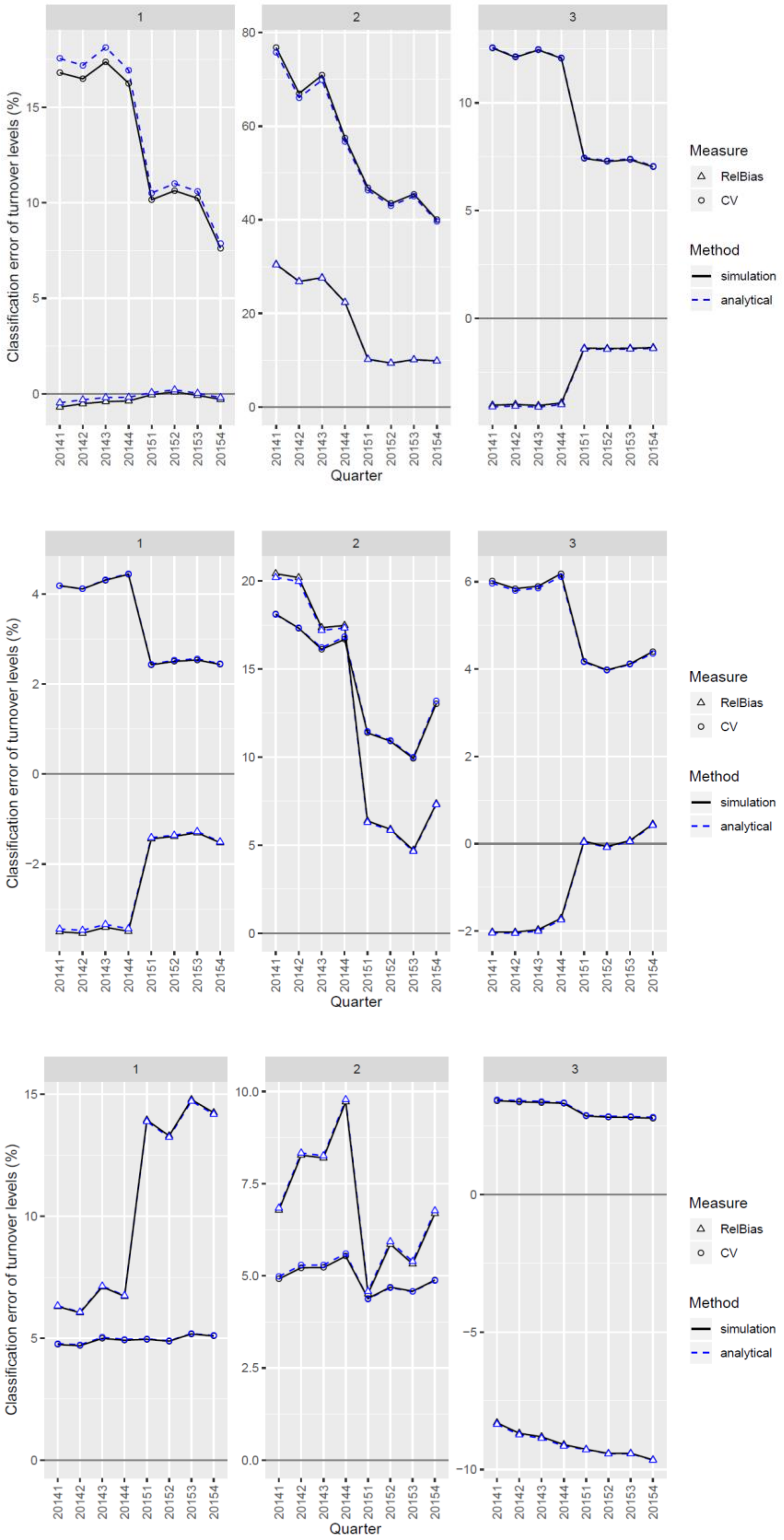


Figure 6. Comparison of relative bias (triangles) and coefficient of variation (circles) for turnover levels in each stratum (1,2,3) based on direct simulation (solid black lines) and analytical formulas (dashed blue lines). Top to bottom: set-ups (a), (b), and (c).

Finally, Figure 6 shows results for the accuracy of quarterly turnover *levels* under each of the three set-ups. We have plotted the relative bias $RB\left(\hat{Y}_h^q\right) = B\left(\hat{Y}_h^q\right)/Y_h^q$ and the coefficient of variation $CV\left(\hat{Y}_h^q\right) = \sqrt{V\left(\hat{Y}_h^q\right)}/Y_h^q$. It is seen that, in contrast to the above results for growth rates, for level estimates the skewed turnover distributions of set-ups (a) and (b) do not have an adverse effect on the accuracy of the analytical approximations.

9. Case study

9.1 Introduction

In this section we describe an application of the analytical and bootstrap approaches to estimate the bias and variance of growth rates in a case study on the short-term statistics for a small economic sector: car trade (NACE G45). Parts of this case study have been described before in van Delden et al. (2016a and 2016b). Within car trade, there are six publication cells for which short-term statistics estimates are published and there are nine underlying industries. Based on those nine industries all publications of Statistics Netherlands on car trade that use turnover (short-term statistics, structural business statistics, National Accounts) can be produced.

We begin by introducing the various data sets used in this case study (Subsection 9.2). The classification error models introduced in Section 7 were used for this study. The estimation of these models is discussed for the level matrix in Subsection 9.3 and for the change matrix in Subsection 9.4. The resulting bias and variance estimates for the growth rates within car trade are shown in Subsection 9.5.

9.2 Data

The Dutch short-term statistics on turnover are derived from two data sources on businesses. More specifically, the data concern businesses as statistical units, known as enterprises. 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. For the present application, we used turnover data for eight quarters in 2014 and 2015.

Table 5 shows the average number of units and quarterly turnover (for 2014 and 2015) observed in the nine car trade industries. The industries are listed in descending order of average quarterly turnover: the largest industry within car trade is industry code 45112 and the smallest industry is code 45194.

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 frequently inspect the largest 25 units in each industry and make corrections if needed, 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.

Table 5. Number of units and quarterly turnover (in millions of Euros) per car trade industry (average values for the period Q2 2014 – Q4 2015).

Industry	Description of economic activity	Number of units	Turnover
45112	Sale and repair of passenger cars and light motor vehicles (no import of new cars)	18 618	7 961
45111	Import of new passenger cars and light motor vehicles	157	2 910
45310	Wholesale and commission trade of motor vehicle parts and accessories	1 961	2 194
45191X	Sale and repair of trucks and trailers	1 329	1 202
45200	Specialised repair of motor vehicles	6 022	687
45401	Wholesale and commission trade of motorcycles and related parts	446	374
45402	Retail trade and repair of motorcycles and related parts	1 246	150
45320	Retail trade of motor vehicle parts and accessories	841	89
45194	Sale and repair of caravans	363	79

The classification error models of Section 7 for the level and change matrix contain a number of parameters that need to be specified and estimated. For this we used, in addition to data that were available from the GBR, two audit samples of units for which the true NACE code was determined by experts. Both audit samples were drawn from the observed car trade population according to the GBR, from the subpopulation of simple units only.

The 2014 audit sample. The first audit sample was drawn from the population in the GBR of 1 July 2014. We therefore refer to this sample as "the 2014 audit sample". From each of the nine car trade industries, 25 units were drawn at random. The observed NACE codes in the GBR were used as strata. A NACE code specialist reviewed the activities of the enterprises and determined the true (current) NACE codes, using information on the website of the enterprise and contacting the enterprise by telephone if more information was needed. A subject-matter specialist then checked the results of the NACE code specialist, and the two auditors were asked to come up with one final true NACE code for each unit in the audit sample. This sample was used previously by van Delden et al. (2016b) to estimate level matrices. Below, we will expand on their results and assume that the resulting estimated level matrices can be applied to all quarters in 2014 and 2015.

The 2015 audit sample. The second audit sample was set up to not only provide data for a level estimate of the amount of NACE classification errors, but also to estimate the relevant probabilities for the change matrix. The sample was drawn from the car trade population in the GBR of 1 July 2015. A detailed description of the design of this sample can be found in van Delden et al. (2016a). The sample was again stratified by observed NACE code, but each stratum was subdivided into four substrata which were referred to as audit strata (AS). These AS depend on the combination of observed industry codes of 1 July 2014 and 1 July 2015:

- AS 1: continuing units for which the observed industry code in the GBR has changed between 1 July 2014 and 1 July 2015;
- AS 2: continuing units for which the observed industry code in the GBR has not changed and there is a relatively large probability that the observed code contains an error in at least one of the periods;
- AS 3: continuing units for which the observed industry code in the GBR has not changed and there is a relatively small probability that the observed code contains an error in at least one of the periods;
- AS 4: new-born units between 1 July 2014 and 1 July 2015.

The demarcation of AS 1 and AS 4 followed directly from the observed GBR data. The distinction between AS 2 and AS 3 was made on the basis of a model with background variables that predict the presence of an error in the observed industry code; see van Delden et al. (2016a) for details. For each observed industry code, the sample consisted of 33 units, with 10 units allocated to each of AS 1, AS 2 and AS 3, and the remaining 3 units allocated to AS 4. Given the relative sizes of the audit strata in the population, this implies that AS 1 and AS 2 — which contain the most relevant information for estimating the change matrix — were severely oversampled.

The 2015 audit sample used the same auditors as the 2014 audit sample. For the 2015 sample, the auditors were asked to determine both the true current NACE code and the true NACE code of 12 months earlier. For the code of one year earlier they could make use of the internet archive (‘archive.org’) or contact the enterprise.

To collect the relevant information, we had to conduct the above audit samples specifically for this case study, 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 alternative ways to estimate the input parameters also at Statistics Netherlands (see Section 10).

9.3 Input parameters for quarter-on-quarter growth rates

9.3.1 Probability classes and their probabilities

To apply formula (43) for the estimated bias and formula (46) 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}^{L} (for $g \in \mathcal{G}_h$) and $\tilde{p}_{(-\mathcal{G}_h)hc}^{L}$.

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

1. simple units with fewer than 10 employees (EMP) that consist of at most two legal units (LU);
2. simple units with fewer than 10 EMP that consist of at least three LU;
3. simple units with 10–19 EMP, complex units with fewer than 20 EMP, and most complex units with fewer 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.

It should be noted 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 $\mathbf{P}_c^L$, van Delden et al. (2016b) distinguished between rows and columns within car trade and misclassifications between car trade and other economic sectors:

$$\begin{array}{cc} \text{true} \setminus \text{observed} & \begin{array}{cc} \text{car trade industries} & \text{other industries} \end{array} \\ \begin{array}{c} \text{car trade industries} \\ \text{other industries} \end{array} & \left(\begin{array}{cc} \text{within car trade} & \text{missed} \\ \text{erroneously included} & \text{outside car trade} \end{array} \right) \end{array}$$

For probability classes 1 and 2, the classification errors probabilities “within car trade” were estimated from the 2014 audit sample, using a logistic regression model for the diagonal elements and a log-linear model for the off-diagonal elements, as described in Subsection 7.2. 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 more details.

Table 6. 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

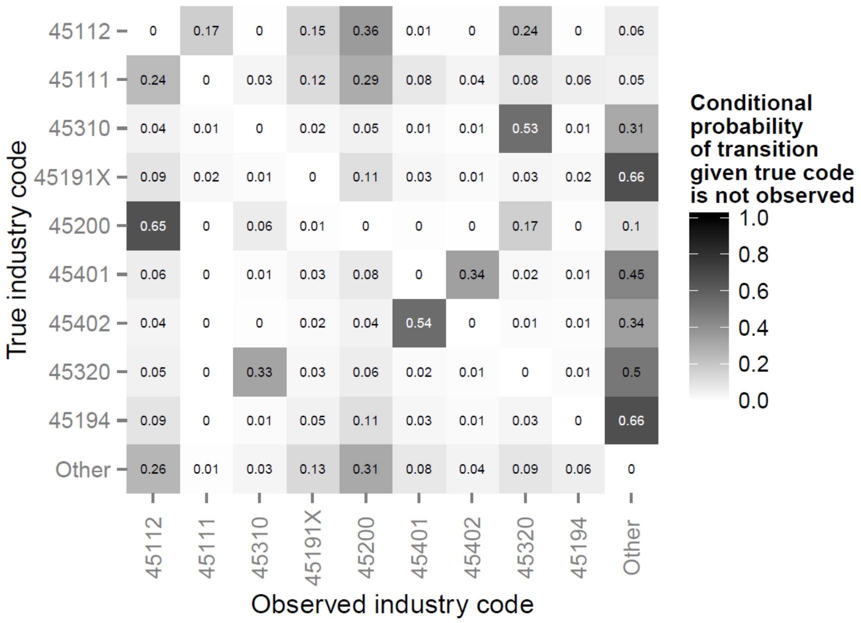


Figure 7. Estimated conditional probabilities for the off-diagonal elements of the level matrix (taken from van Delden et al., 2016b). Each row adds up to 1.

The estimated probabilities for the five probability classes as found by van Delden et al. (2016b) are summarised in Table 6 and Figure 7. 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; under the model, these probabilities are the same for all probability classes.

From the information in the table and figure, one can compute all elements of the level matrices $\mathbf{P}_c^L$ 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 with industry code 45200 is (1 – 0.88) × 0.36, so about 0.04.

Figure 7 also contains "overall" conditional probabilities for the "erroneously included" units, with the non-car trade industries collapsed into one industry "Other". In van Delden et al. (2016a, 2016b), the components "erroneously included" and "outside car trade" of the matrix $\mathbf{P}_c^L$ were not computed explicitly at a more detailed level. Here, we estimated these detailed probabilities by a slightly different approach, while still re-using some results from these previous studies.

First, for probability classes 1 to 4 we estimated the overall probability that a unit with true industry code outside car trade is misclassified into a car trade industry. (For probability class 5, this probability is assumed to be zero.) Van Delden et al. (2015, Appendix A.2) estimated this probability across all probability classes as 0.000625. Here we applied the same approach for each probability class separately, which yielded similarly-sized probabilities. Next, within each class, by multiplying the obtained probability with the conditional probabilities in the last row of Figure 7, 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.

Secondly, within probability classes 1 to 4, we computed the average number of units in each industry as observed in the GBR, across the four quarters in 2014; we denote these numbers as N_{gc}. For each probability class, by multiplying the total number of units outside car trade ($\sum_{g \in \mathcal{H}_{\text{full}} \setminus \mathcal{H}} N_{gc}$) by the above unconditional probabilities in the last row of the collapsed level matrix, we obtained approximate expected numbers of non-car trade units that are erroneously included in each car trade industry. Table 7 shows the resulting approximate expected counts for each probability class. The total number of units observed outside car trade in each probability class — averaged across four quarters in 2014 — is shown in the last row.

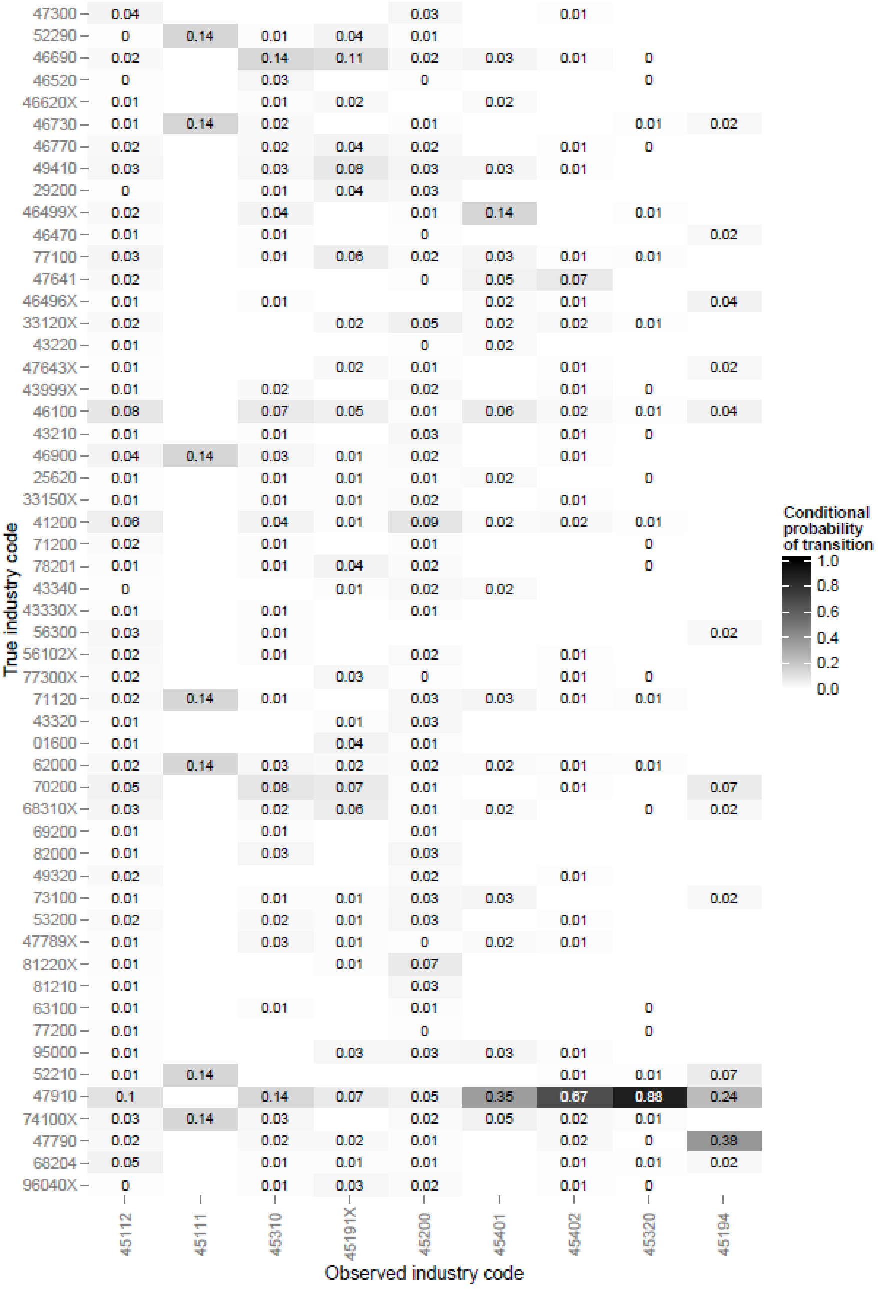
Table 7. Expected numbers of units with true industry code outside car trade that are observed in a car trade industry, for probability classes 1 to 4.

Observed industry code	PC 1	PC 2	PC 3	PC 4
45112	151.2	18.8	2.4	0.6
45111	3.5	0.4	0.1	0.0
45310	20.0	2.5	0.3	0.1
45191X	76.6	9.5	1.2	0.3
45200	182.0	22.7	2.8	0.8
45401	48.5	6.0	0.8	0.2
45402	23.2	2.9	0.4	0.1
45320	52.5	6.5	0.8	0.2
45194	34.6	4.3	0.5	0.1
Total misclassified	592.2	73.8	9.3	2.5
Total outside car trade	940 839.3	28 420.3	25 211.3	20 978.0

Finally, for *specific* true industry codes outside car trade, we computed 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 (i.e., $\mathcal{H}_{\text{full}} \setminus \mathcal{H}$). Based on observed yearly transitions in the GBR, van Delden et al. (2015) identified 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 restricted attention to this subset of non-target industries. In what follows, we denote this subset of 54 industries as $\mathcal{H}_{\text{spec}}$ and the other non-target industries as $\mathcal{H}_{\text{rest}}$, so $\mathcal{H}_{\text{full}} = \mathcal{H} \cup \mathcal{H}_{\text{spec}} \cup \mathcal{H}_{\text{rest}}$.

The relative contributions among the 54 industries to the total number of erroneously included units in car trade industries are given in Figure 8. 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 in probability class 1 that are misclassified in industry 45112. According to Table 7, we expect 151.2 units with true industry code outside car trade to be misclassified in this car trade industry. The first column of Figure 8 specifies the relative distribution of these 151.2 units across the non-car trade industries. Thus, about 0.04 × 151.2 ≈ 6 units in true industry 47300 are expected to be misclassified in industry 45112, etc.

Finally, these industry-specific expected numbers of misclassifications (say n_{ghc}, with $g \in \mathcal{H}_{\text{spec}}$ and $h \in \mathcal{H}$) were translated into classification error probabilities by dividing them by the total number of units in each non-car trade industry (averaged over 2014): $p^L_{ghc} = n_{ghc}/N_{gc}$, for $g \in \mathcal{H}_{\text{spec}}$ and $h \in \mathcal{H}$.



	45112	45111	45310	45191X	45200	45401	45402	45320	45194
47300	0.04				0.03		0.01		
52290	0	0.14	0.01	0.04	0.01				
46690	0.02		0.14	0.11	0.02	0.03	0.01	0	
46520	0		0.03		0			0	
46620X	0.01		0.01	0.02		0.02			
46730	0.01	0.14	0.02		0.01			0.01	0.02
46770	0.02		0.02	0.04	0.02		0.01	0	
49410	0.03		0.03	0.08	0.03	0.03	0.01		
29200	0		0.01	0.04	0.03				
46499X	0.02		0.04		0.01	0.14		0.01	
46470	0.01		0.01		0				0.02
77100	0.03		0.01	0.06	0.02	0.03	0.01	0.01	
47641	0.02				0	0.05	0.07		
46496X	0.01		0.01			0.02	0.01		0.04
33120X	0.02			0.02	0.05	0.02	0.02	0.01	
43220	0.01				0	0.02			
47643X	0.01			0.02	0.01		0.01		0.02
43999X	0.01		0.02		0.02		0.01	0	
46100	0.08		0.07	0.05	0.01	0.06	0.02	0.01	0.04
43210	0.01		0.01		0.03		0.01	0	
46900	0.04	0.14	0.03	0.01	0.02		0.01		
25620	0.01		0.01	0.01	0.01	0.02		0	
33150X	0.01		0.01	0.01	0.02		0.01		
41200	0.06		0.04	0.01	0.09	0.02	0.02	0.01	
71200	0.02		0.01		0.01			0	
78201	0.01		0.01	0.04	0.02			0	
43340	0			0.01	0.02	0.02			
43330X	0.01		0.01		0.01				
56300	0.03		0.01						0.02
56102X	0.02		0.01		0.02		0.01		
77300X	0.02			0.03	0		0.01	0	
71120	0.02	0.14	0.01		0.03	0.03	0.01	0.01	
43320	0.01			0.01	0.03				
01600	0.01			0.04	0.01				
62000	0.02	0.14	0.03	0.02	0.02	0.02	0.01	0.01	
70200	0.05		0.08	0.07	0.01		0.01		0.07
68310X	0.03		0.02	0.06	0.01	0.02		0	0.02
69200	0.01		0.01		0.01				
82000	0.01		0.03		0.03				
49320	0.02				0.02		0.01		
73100	0.01		0.01	0.01	0.03	0.03			0.02
53200	0.02		0.02	0.01	0.03		0.01		
47789X	0.01		0.03	0.01	0	0.02	0.01		
81220X	0.01			0.01	0.07				
81210	0.01				0.03				
63100	0.01		0.01		0.01			0	
77200	0.01				0			0	
95000	0.01			0.03	0.03	0.03	0.01		
52210	0.01	0.14					0.01	0.01	0.07
47910	0.1		0.14	0.07	0.05	0.35	0.67	0.88	0.24
74100X	0.03	0.14	0.03		0.02	0.05	0.02	0.01	
47790	0.02		0.02	0.02	0.01		0.02	0	0.38
68204	0.05		0.01	0.01	0.01		0.01	0.01	0.02
96040X	0		0.01	0.03	0.02		0.01	0	

Figure 8. 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 in $\mathcal{H}_{\text{spec}}$ are listed in descending order of average turnover.

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

After the computations in the previous subsection, we have obtained a level matrix $\mathbf{P}_c^L$ for each of the five probability classes. This matrix contains the probabilities $p_{ghc}^L = P(\hat{s}_i^q = h | s_i^q = g)$ for units i in probability class c, for all combinations of g and h. Under assumption A2'' in Subsection 4.3, for each observed industry code $h \in \mathcal{H}$ we may select a small subset $\mathcal{G}_h$ of special cases for which p_{ghc}^L within $\mathbf{P}_c^L$ is considered separately, and use an average probability for all cases $g \notin \mathcal{G}_h$. Recall that the idea behind this assumption is that it will simplify the bias and variance computations and also make the bias and variance estimates more stable. In our situation the use of a small subset $\mathcal{G}_h$ does not simplify the actual estimation of the parameters, since we compute these simplifications after we have first estimated all parameters. However, if the computations show that parameters can be combined to a great extent whithout affecting the accuracy outcomes, then it may help in future to simplify the parameter estimation also. To test this idea in practice, we applied it to this case study with different selections for $\mathcal{G}_h$. Recall that for simplicity we assumed that the same groups $\mathcal{G}_h$ apply to all probability classes. Therefore we included industries as special cases if they appeared to be important for at least one probability class.

Let $h \in \mathcal{H}$ be a car trade industry. To decide whether to include $g \in \mathcal{H}_{\text{spec}}$ in the subset $\mathcal{G}_h$, we focused on two properties:

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

We defined criteria to find the cases with the largest probabilities and the largest expected numbers of misclassified units (see below) and applied these criteria to the estimated probabilities from the previous subsection. An element of the level matrix was selected as a special case if it satisfied *at least* one of these criteria, for at least one probability class.

For the first criterion, we used $p_{gh,max}^L = \max_c p_{ghc}^L$ and selected as special cases those combinations with $p_{gh,max}^L > 0.05$:

$$\mathcal{G}_{h1} = \{g | p_{gh,max}^L > 0.05\}.$$

The second criterion requires some more explanation. To obtain an expected number of units with true industry code g that are observed in industry h, we multiplied each p_{ghc}^L by the number of units in industry g (according to the GBR). 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 computed $f_{gh,max} = \max_c f_{ghc}$ for each $h \in \mathcal{H}$. For a given $h \in \mathcal{H}$, 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 defined

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

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

Table 8 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 were fourteen industry codes outside car trade (of the 54 codes within $\mathcal{H}_{\text{spec}}$) where units have a relatively large probability to be misclassified into (at least one of) the car trade industries.

Table 8. Selection of groups based on the two criteria. Rows indicate true industry codes g; columns indicate observed industry codes h. Cells with $p^L_{gh,max} > 0.05$ are highlighted in boldface (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	45111	45112	45191X	45194	45200	45310	45320	45401	45402
45111	**0.847**	**0.005**	**0.026**	**0.034**	**0.018**	0.004	**0.018**	**0.071**	0.023
45112	0.891	**0.992**	0.216	0.019	0.147	0.002	0.362	0.039	0.013
45191X	0.044	0.004	**0.965**	0.031	**0.016**	0.004	0.017	0.064	0.020
45194	0.001	0.002	0.010	**0.968**	**0.007**	0.002	0.007	0.028	0.009
45200	0.000	**0.050**	0.006	0.008	**0.959**	0.032	0.146	0.016	0.005
45310	0.020	0.002	0.011	0.014	0.007	**0.997**	**0.272**	0.030	0.009
45320	0.001	0.002	0.010	0.016	**0.006**	**0.083**	**0.609**	0.024	0.008
45401	0.001	**0.001**	0.007	0.011	**0.005**	0.001	0.007	**0.896**	**0.225**
45402	0.000	0.000	0.001	0.002	0.001	0.000	0.002	**0.170**	**0.975**
46100	0.000	0.002	0.007	0.008	0.001	0.002	0.001	0.025	0.003
46499X	0.000	0.000	0.000	0.000	0.001	0.001	0.001	0.056	0.000
46690	0.000	0.000	0.016	0.000	0.002	0.003	0.000	0.012	0.001
47641	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.019	0.009
47790	0.000	0.001	0.003	**0.070**	0.001	0.000	0.000	0.000	0.002
47910	0.000	0.002	0.010	0.045	0.005	0.003	0.090	0.136	0.083
49410	0.000	0.001	0.011	0.000	0.003	0.001	0.000	0.012	0.001
52210	0.005	0.000	0.000	0.012	0.000	0.000	0.001	0.000	0.001
70200	0.000	0.001	0.010	0.012	0.001	0.002	0.000	0.000	0.001
71120	0.005	0.001	0.000	0.000	0.002	0.000	0.001	0.012	0.001
73100	0.000	0.001	0.001	0.004	0.002	0.000	0.000	0.012	0.000
74100X	0.005	0.001	0.000	0.000	0.002	0.001	0.001	0.019	0.002
77100	0.000	0.001	0.009	0.000	0.002	0.000	0.001	0.012	0.001
95000	0.000	0.000	0.004	0.000	0.003	0.000	0.000	0.012	0.001

The results in this table illustrate that the two criteria complement each other to some extent. For instance, for target industry 45200 the first criterion selected six industries (45111, 45191X, 45194, 45200, 45320, and 45401), including three which were not selected by the second criterion for any of the above choices of α (45194, 45320, and 45401). The latter three industries contain relatively few units, so their contributions to 45200 in terms of $f_{gh,max}$ are necessarily small. Conversely, the second criterion selected an additional industry (45112) that was not selected by the first criterion. This industry is relatively large in terms of units, so it may have a sizeable contribution to 45200 in terms of misclassified units, even when the associated misclassification probability is relatively small.

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

Having selected the groups $\mathcal{G}_h = \mathcal{G}_h(\alpha)$, we retained the classification error probabilities p^L_{ghc} for $g \in \mathcal{G}_h$ from Subsection 9.3.1 and we replaced the probabilities p^L_{ghc} for $g \notin \mathcal{G}_h$ by their mean value $\tilde{p}^L_{(-\mathcal{G}_h)hc}$ according to assumption A2'' in Subsection 4.3. In this application, we restricted the latter computation to the nine car trade industries and the 54 non-car trade industries shown in Figure 8.

Table 9. Original estimated level matrix for simple units with fewer than 10 EMP that consist of at least three LU.

Industry	45111	45112	45191X	45194	45200	45310	45320	45401	45402
45111	0.02464	0.23808	0.12070	0.05446	0.28666	0.03152	0.08275	0.07647	0.03657
45112	0.02108	0.87576	0.01872	0.00057	0.04446	0.00033	0.03044	0.00080	0.00038
45191X	0.01110	0.04343	0.53162	0.00993	0.05230	0.00575	0.01510	0.01395	0.00667
45194	0.00032	0.04450	0.02256	0.51882	0.05358	0.00589	0.01547	0.01429	0.00683
45200	0.00003	0.16968	0.00179	0.00081	0.74087	0.01653	0.04339	0.00114	0.00054
45310	0.00279	0.01090	0.00553	0.00249	0.01313	0.74832	0.13381	0.00350	0.00167
45320	0.00031	0.04289	0.02174	0.00981	0.05164	0.27259	0.16433	0.01378	0.00659
45401	0.00039	0.05343	0.02708	0.01222	0.06433	0.00707	0.01857	0.14569	0.28974
45402	0.00007	0.00954	0.00484	0.00218	0.01148	0.00126	0.00332	0.14704	0.72769
46100	0.00000	0.00300	0.00091	0.00036	0.00055	0.00035	0.00007	0.00073	0.00013
46499X	0.00000	0.00140	0.00000	0.00000	0.00075	0.00036	0.00014	0.00339	0.00000
46690	0.00000	0.00061	0.00180	0.00000	0.00082	0.00060	0.00003	0.00033	0.00006
47641	0.00000	0.00450	0.00000	0.00000	0.00132	0.00000	0.00000	0.00396	0.00294
47790	0.00000	0.01761	0.00753	0.06421	0.00755	0.00182	0.00072	0.00000	0.00195
47910	0.00000	0.00984	0.00357	0.00562	0.00562	0.00185	0.03083	0.01127	0.01035
49410	0.00000	0.00137	0.00186	0.00000	0.00186	0.00017	0.00000	0.00047	0.00004
52210	0.00146	0.00415	0.00000	0.00668	0.00000	0.00000	0.00085	0.00000	0.00038
70200	0.00000	0.00038	0.00027	0.00012	0.00012	0.00008	0.00000	0.00000	0.00001
71120	0.00007	0.00045	0.00000	0.00000	0.00066	0.00003	0.00008	0.00022	0.00002
73100	0.00000	0.00026	0.00021	0.00021	0.00128	0.00005	0.00000	0.00043	0.00000
74100X	0.00021	0.00187	0.00000	0.00000	0.00126	0.00023	0.00024	0.00095	0.00016
77100	0.00000	0.00888	0.00855	0.00000	0.00571	0.00034	0.00082	0.00286	0.00025
95000	0.00000	0.00377	0.00605	0.00000	0.01414	0.00000	0.00000	0.00406	0.00069

Table 10. Approximated level matrix for simple units with fewer than 10 EMP that consist of at least three LU ($\mathcal{G}_h$ based on $\alpha = 1\%$). Grey text indicates non-special cases. The last row refers to 40 other industries outside car trade.

Industry	45111	45112	45191X	45194	45200	45310	45320	45401	45402
45111	0.02464	0.23808	0.12070	0.05446	0.28666	*0.00110*	0.08275	0.07647	0.03657
45112	0.02108	0.87576	0.01872	0.00057	0.04446	*0.00110*	0.03044	0.00080	0.00038
45191X	0.01110	*0.00506*	0.53162	0.00993	0.05230	*0.00110*	0.01510	0.01395	0.00667
45194	*0.00006*	*0.00506*	0.02256	0.51882	0.05358	*0.00110*	*0.00077*	0.01429	*0.00046*
45200	*0.00006*	0.16968	*0.00225*	*0.00018*	0.74087	0.01653	0.04339	0.00114	*0.00046*
45310	0.00279	*0.00506*	0.00553	0.00249	*0.00412*	0.74832	0.13381	0.00350	*0.00046*
45320	*0.00006*	*0.00506*	*0.00225*	0.00981	0.05164	0.27259	0.16433	0.01378	*0.00046*
45401	*0.00006*	0.05343	*0.00225*	0.01222	0.06433	*0.00110*	*0.00077*	0.14569	0.28974
45402	*0.00006*	*0.00506*	*0.00225*	*0.00018*	*0.00412*	*0.00110*	*0.00077*	0.14704	0.72769
46100	*0.00006*	*0.00506*	*0.00225*	*0.00018*	*0.00412*	*0.00110*	*0.00077*	0.00073	*0.00046*
46499X	*0.00006*	*0.00506*	*0.00225*	*0.00018*	*0.00412*	*0.00110*	*0.00077*	0.00339	*0.00046*
46690	*0.00006*	*0.00506*	0.00180	*0.00018*	*0.00412*	*0.00110*	*0.00077*	0.00033	*0.00046*
47641	*0.00006*	*0.00506*	*0.00225*	*0.00018*	*0.00412*	*0.00110*	*0.00077*	0.00396	*0.00046*
47790	*0.00006*	*0.00506*	*0.00225*	0.06421	*0.00412*	*0.00110*	*0.00077*	*0.00012*	*0.00046*
47910	*0.00006*	*0.00506*	0.00357	0.00562	*0.00412*	*0.00110*	0.03083	0.01127	0.01035
49410	*0.00006*	*0.00506*	0.00186	*0.00018*	*0.00412*	*0.00110*	*0.00077*	0.00047	*0.00046*
52210	*0.00006*	*0.00506*	*0.00225*	0.00668	*0.00412*	*0.00110*	*0.00077*	*0.00012*	*0.00046*
70200	*0.00006*	*0.00506*	0.00027	0.00012	*0.00412*	*0.00110*	*0.00077*	*0.00012*	*0.00046*
71120	*0.00006*	*0.00506*	*0.00225*	*0.00018*	*0.00412*	*0.00110*	*0.00077*	0.00022	*0.00046*
73100	*0.00006*	*0.00506*	*0.00225*	*0.00018*	*0.00412*	*0.00110*	*0.00077*	0.00043	*0.00046*
74100X	*0.00006*	*0.00506*	*0.00225*	*0.00018*	*0.00412*	*0.00110*	*0.00077*	0.00095	*0.00046*
77100	*0.00006*	*0.00506*	*0.00225*	*0.00018*	*0.00412*	*0.00110*	*0.00077*	0.00286	*0.00046*
95000	*0.00006*	*0.00506*	*0.00225*	*0.00018*	*0.00412*	*0.00110*	*0.00077*	0.00406	*0.00046*
other	*0.00006*	*0.00506*	*0.00225*	*0.00018*	*0.00412*	*0.00110*	*0.00077*	*0.00012*	*0.00046*

As an illustration, Table 9 shows the original level matrix $\mathbf{P}_c^L$ for the second probability class (simple units with fewer than 10 EMP that consist of at least three LU) and Table 10 shows the corresponding matrix $\widetilde{\mathbf{P}}_c^L$ for the largest set $\mathcal{G}_h$ ($\alpha = 1\%$). Cells that contain approximate probabilities based on $\tilde{p}_{(-\mathcal{G}_h)hc}^L$ are printed in grey.

9.4 Input parameters for year-on-year growth rates

To apply formula (64) for the estimated bias and formula (65) for the estimated variance of a year-on-year growth rate, the following input is needed in addition to the parameters in Subsection 9.3:

– 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.

For this case study, the matrix $\mathbf{R}$ followed directly from the observed yearly transitions between NACE codes in the GBR. The probabilities p_{Rc}, p_{Nc}, p_{Sc} were estimated by van Delden et al. (2016a) based on the 2015 audit sample (excluding the audit stratum of new-born units). Four probability classes were distinguished, which are unfortunately not identical to the five probability classes listed above for the level matrix. The four probability classes are:

a. simple units with fewer than 20 EMP;
b. simple units with 20–49 EMP, complex units with fewer than 20 EMP, and most complex units with fewer than 10 EMP;
c. simple units with at least 50 EMP, complex units with 20–49 EMP, and most complex units with 10–19 EMP;
d. complex units with at least 50 EMP, most complex units with at least 20 EMP, and all units in supplemental editing.

Table 11 shows the estimated probabilities for these four probability classes as found by van Delden et al. (2016a). The values for probability class a were found by applying the EM algorithm from Subsection 7.3.2 to the data from the 2015 audit sample, with sampling weights that varied by observed industry code (see van Delden et al., 2016a, for a detailed discussion). The values for probability class d were fixed under the assumption that these units are always classified correctly. Finally, the values for probability classes b and c were obtained by a linear interpolation between the values for probability classes a and d.

Table 11. Estimated probabilities for the probabilities p_{Rc}, p_{Nc}, p_{Sc} of the change matrix within car trade (values taken from van Delden et al., 2016a).

Probability class	Parameter		
	p_{Rc}	p_{Nc}	p_{Sc}
a	0.043	0.159	0.00054
b	0.362	0.439	0.00036
c	0.681	0.720	0.00018
d	1.000	1.000	0.00000

From the estimated parameters p_{Rc}, p_{Nc}, p_{Sc} and the matrix $\mathbf{R}$, the probabilities $p^{C}_{gklhc} = P\left(\hat{s}_i^q = h \middle| s_i^{q-4} = g, s_i^q = k, \hat{s}_i^{q-4} = l\right)$ for units in each probability class can be computed using the expressions in Table 3 or Appendix C.

Once the probabilities in the level and change matrices have been estimated, the corresponding probabilities $\mathbb{p}^{LC}_{gkhhc} = p^{L}_{ghc} p^{C}_{gkhhc}$ and $p^{LC}_{gkhc} = \sum_{l=1}^{M} p^{L}_{glc} p^{C}_{gkhlc}$ can be derived. This does require that we define probability classes within which both the level and change matrix are constant. For this case study, these probability classes may be obtained by taking the intersection of the five classes defined for the level matrix and the four classes defined for the change matrix. This yields seven probability classes in total; see Table 12. In the remainder of this section, the term probability class will always be used refer to this classification into seven classes.

Table 12. Relation between final probability classes and classes defined for the level and change matrix separately.

Probability class (final)	Probability class (level)	Probability class (change)
1	1	a
2	2	a
3	3	a
4	3	b
5	4	b
6	4	c
7	5	d

In this way, we have obtained estimates of the probabilities p_{gkhc}^{LC} and $\mathbb{p}_{gkhhc}^{LC}$ for each of the seven probability classes for all combinations of $g \in \mathcal{H} \cup \mathcal{H}_{\text{spec}}$, $k \in \mathcal{H} \cup \mathcal{H}_{\text{spec}}$ and $h \in \mathcal{H}$. Under assumption A3 at the end of Subsection 5.2, these probabilities can be simplified by defining subsets $\mathcal{T}_h$ of special cases based on $\mathcal{G}_h$ (for the level matrix) and $\mathcal{K}_g$ (frequently occurring true industries in quarter q given true industry g in quarter $q-4$). We introduced several variants of $\mathcal{G}_h$ above in Subsection 9.3.2. For simplicity, in this case study we did not investigate the additional effect of choosing different variants of $\mathcal{K}_g$; instead we simply defined $\mathcal{K}_g = (\mathcal{H} \cup \mathcal{H}_{\text{spec}})\backslash\{g\}$ for all g. This implies that the probabilities p_{gkhc}^{LC} and $\mathbb{p}_{gkhhc}^{LC}$ were only simplified according to assumption A3 based on $\mathcal{G}_h$.

9.5 Results

9.5.1 Comparison of analytical approach and bootstrap approach

Using the estimated parameters from Subsection 9.3 and Subsection 9.4 as input, we estimated the bias and variance of the observed turnover levels and growth rates for the nine car trade industries between the first quarter of 2014 and the fourth quarter of 2015. We applied the analytical method, using expressions (43) and (46) for quarter-on-quarter growth rates within the same year, expressions (64) and (65) for other growth rates, and using the bias and variance expressions in van Delden et al. (2016b) for turnover levels. For the results in this subsection, we used the most detailed approximation, with all industries treated as “special cases”. We also applied the bootstrap method from Section 6, with $R = 10000$ replications. This seemed to be a sufficient number of replications to obtain stable bias and variance estimates in all industries; at 5000 replications the variance estimates for the smallest industry (45194) had not fully converged yet.

All computations were done in the R environment for statistical computing. The bootstrap was run using parallel processing on seven cores; it took several days to run all replications. Analytical formulas were computed on a single core in minutes.

Figure 9 and Figure 10 compare the results for the bootstrap and analytical method, for quarter-on-quarter growth rates and year-on-year growth rates, respectively. We have plotted the bias and the standard error (both in percentage points).

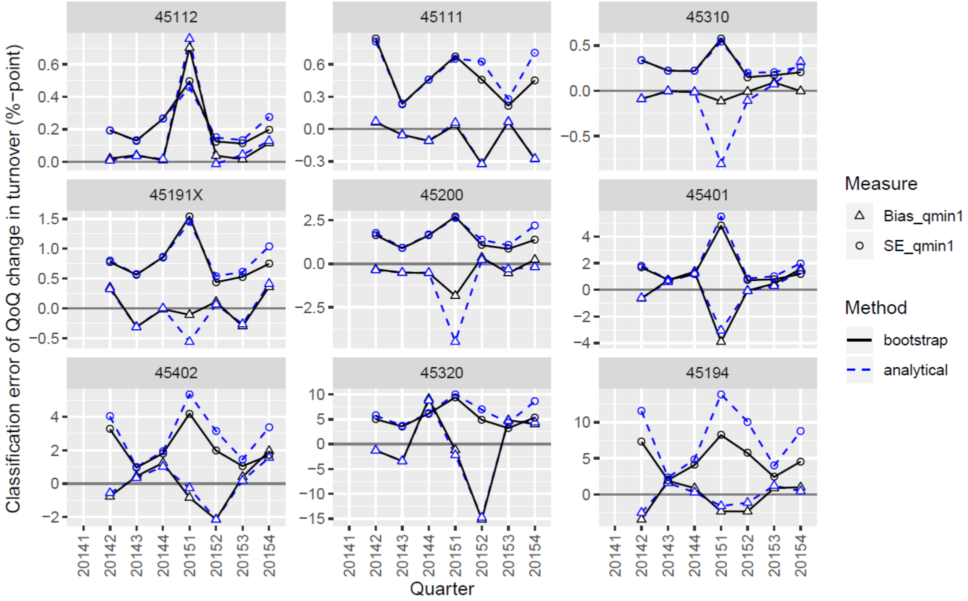


Figure 9. Comparison of bias (triangles) and standard errors (circles) of quarter-on-quarter turnover growth rates for nine industries in car trade, based on bootstrap simulation (solid black lines) and analytical formulas (dashed blue lines).

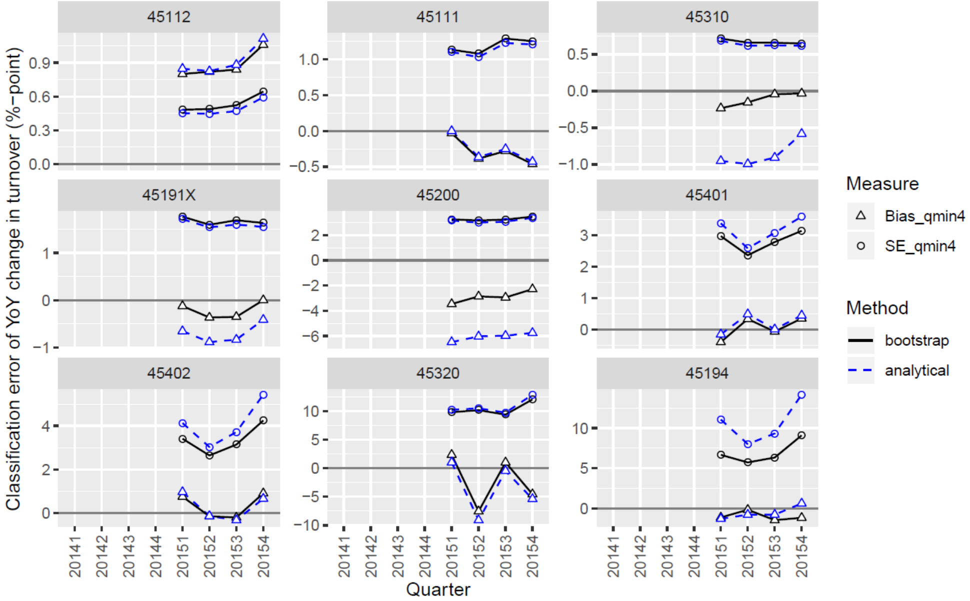


Figure 10. Comparison of bias (triangles) and standard errors (circles) of year-on-year turnover growth rates for nine industries in car trade, based on bootstrap simulation (solid black lines) and analytical formulas (dashed blue lines).

Figure 11 compares the corresponding results for the quarterly turnover levels. Here, we have plotted the relative bias $RB\left(\hat{Y}_h^q\right) = B\left(\hat{Y}_h^q\right)/Y_h^q$ and the coefficient of variation $CV\left(\hat{Y}_h^q\right) = \sqrt{V\left(\hat{Y}_h^q\right)}/Y_h^q$ (both in percentages).

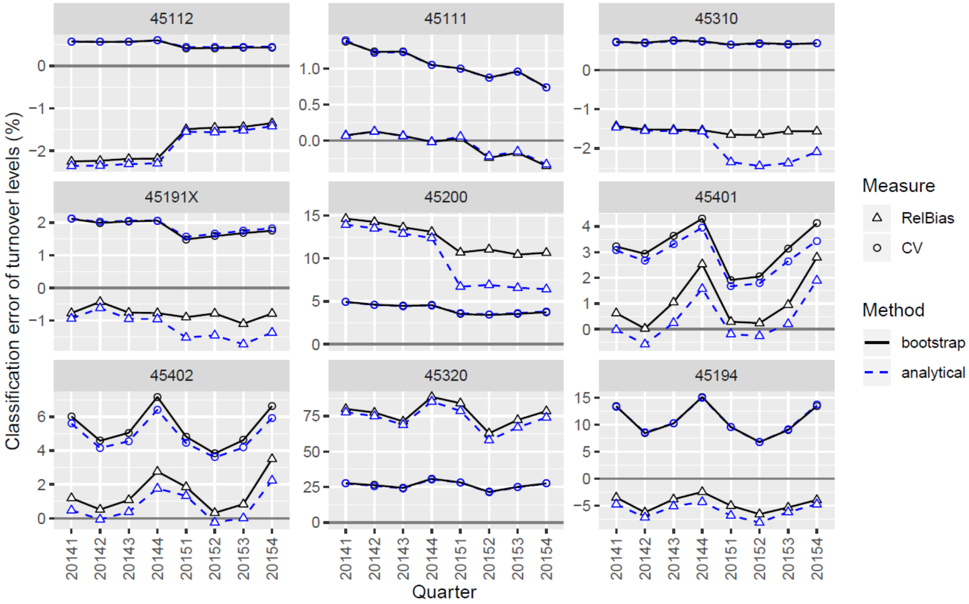


Figure 11. Comparison of relative bias (triangles) and coefficient of variation (circles) for quarterly turnover levels for nine industries in car trade, based on bootstrap simulation (solid black lines) and analytical formulas (dashed blue lines).

For growth rates, it is seen that the results of the bootstrap and the analytical approximation were usually in close agreement, in particular for quarter-on-quarter growth rates within the first year. For growth rates that involve quarters from different years, the two approaches agreed less well for some industries. From the results on simulated data in Section 8, we conclude that these differences can be explained by the skewness of the turnover distribution. In general, when the two approaches differ, we consider the bootstrap results as more reliable since they are not based on a Taylor series approximation.

For the turnover levels in Figure 11, the two approaches yielded results that are virtually identical for most industries. Again, the largest deviations occurred for the bias for quarters in the second year. The magnitude of the relative bias and coefficient of variation for each industry in Figure 11 also agrees quite well with the corresponding results of van Delden et al. (2016b), who applied an earlier version of the bootstrap with the same input parameters to turnover levels within car trade for a different period in time (the first quarter of 2012 to the second quarter of 2014). Note that larger coefficients of variation occur for industries with smaller total turnover levels.

9.5.2 Comparison of analytical approximations based on different $\mathcal{G}_h$

In this subsection, we will compare the results for different analytical approximations based on the selection of special cases in $\mathcal{G}_h$. Recall from Subsection 9.3.2 that we defined three variations of $\mathcal{G}_h(\alpha) = \mathcal{G}_{h1} \cup \mathcal{G}_{h2}(\alpha)$, where industries were included in $\mathcal{G}_{h2}(\alpha)$ only if they were expected to contribute at least $\alpha = 5\%$, $\alpha = 2\%$ or $\alpha = 1\%$ of the total number of misclassified units in industry h (for at least one probability class). The original analytical results in Subsection 9.5.1 were obtained without such a selection mechanism. In general, we expect the approximation to get closer to the results in Subsection 9.5.1 as α decreases.

Figure 12 and Figure 13 show the estimated bias and standard error, respectively, for quarter-on-quarter growth rates based on different selections of special cases. Figure 14 and Figure 15 show the same results for year-on-year growth rates.

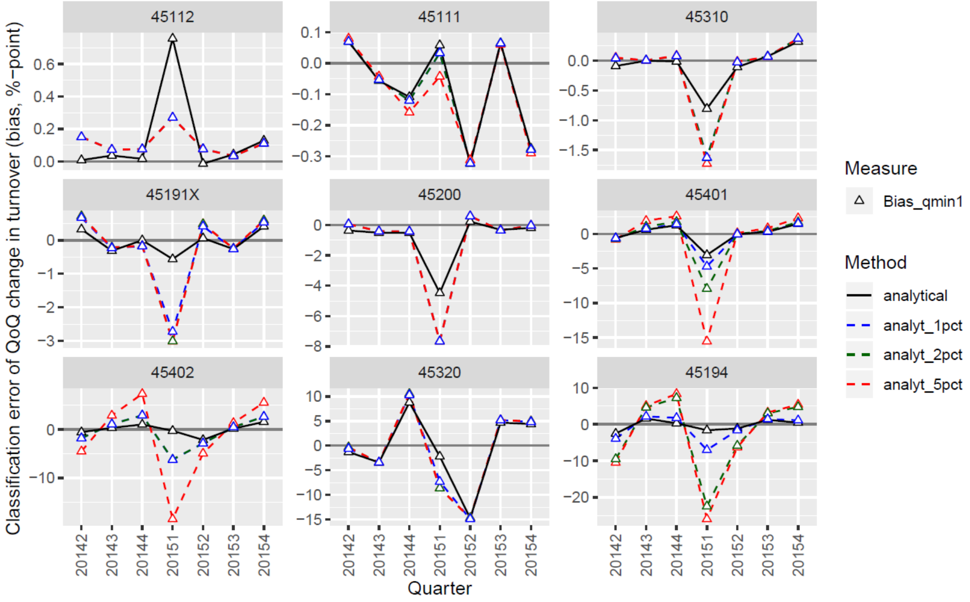


Figure 12. Comparison of analytical bias approximations of quarter-on-quarter turnover growth rates for nine industries in car trade, based on different selections of special cases: no selection (black); selection based on 1% criterion (blue); selection based on 2% criterion (green); selection based on 5% criterion (red).

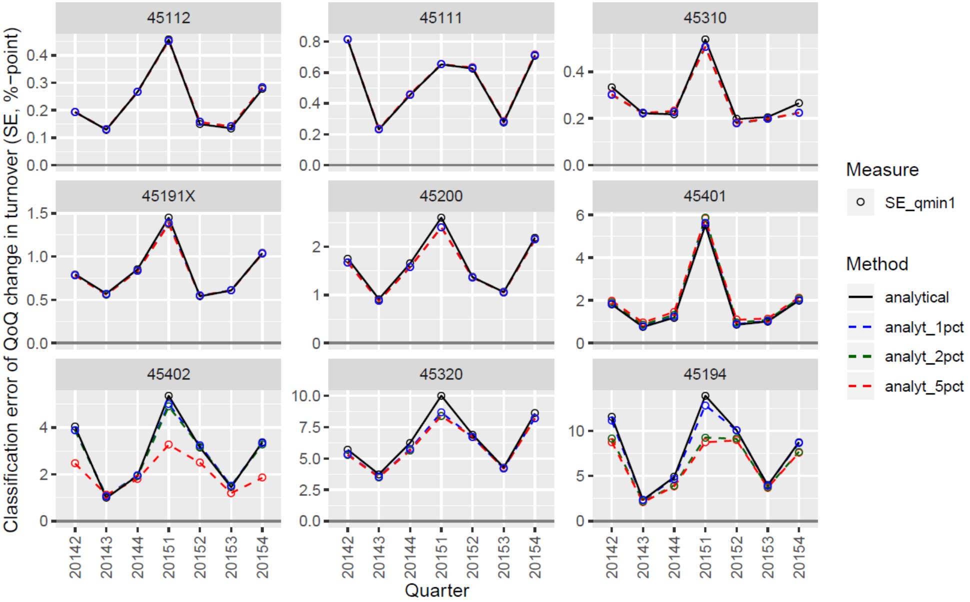


Figure 13. Comparison of analytical standard error approximations of quarter-on-quarter turnover growth rates for nine industries in car trade, based on different selections of special cases (colors as in Figure 12).

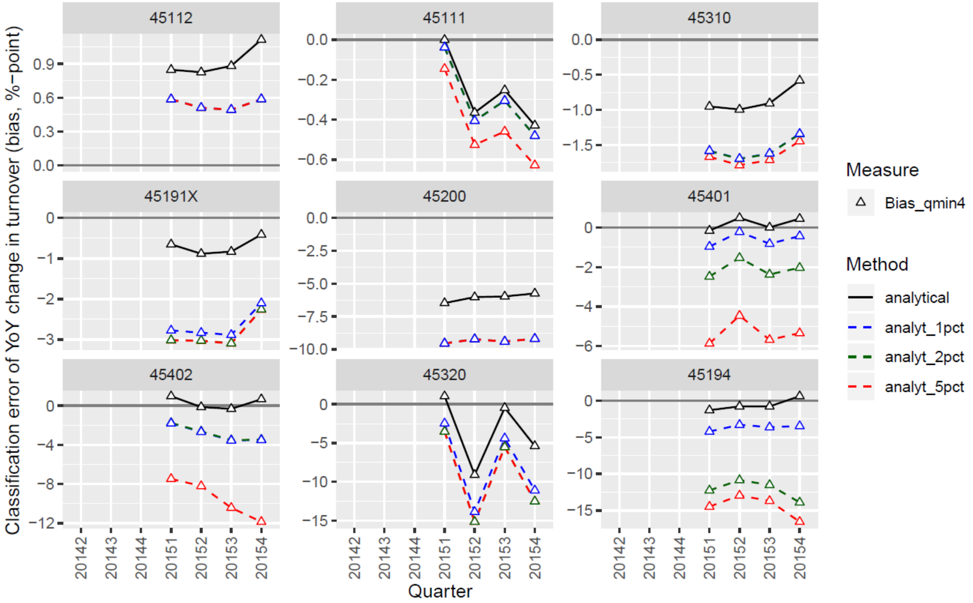


Figure 14. Comparison of analytical bias approximations of year-on-year turnover growth rates for nine industries in car trade, based on different selections of special cases (colors as in Figure 12).

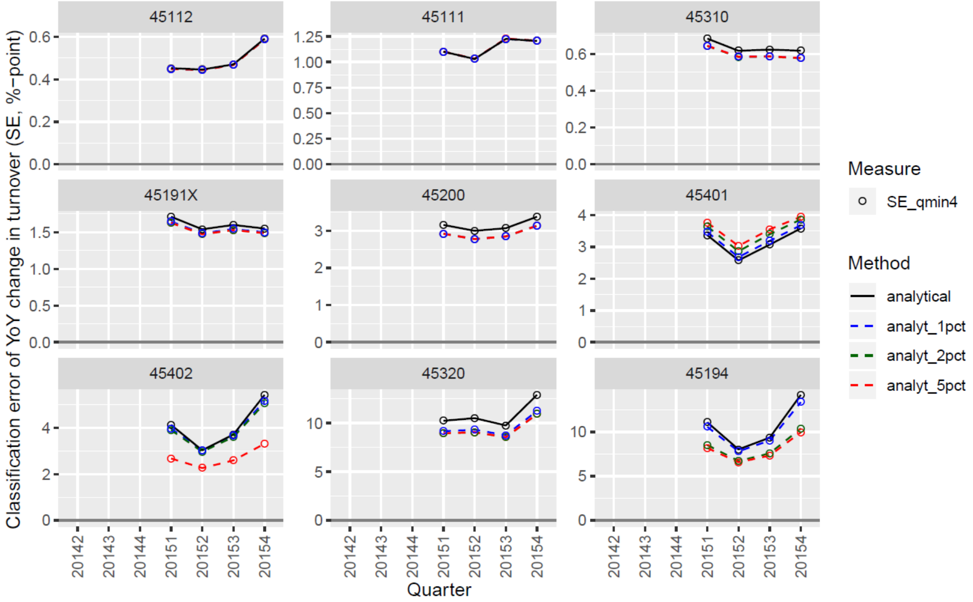


Figure 15. Comparison of analytical standard error approximations of year-on-year turnover growth rates for nine industries in car trade, based on different selections of special cases (colors as in Figure 12).

Both for quarter-on-quarter (Figure 12) and year-on-year growth rates (Figure 14), it is seen that the bias estimates are often rather sensitive to the selection of special cases. As expected, the most inclusive selection criterion based on $\alpha = 1\%$ leads to bias estimates that are closest to the original analytical results from Subsection 9.5.1, but even for this choice the differences are often substantial. By contrast, for the estimated standard errors (Figure 13 and Figure 15) it is seen that the choice $\alpha = 1\%$ yields results that are close to the original analytical results from Subsection 9.5.1.

Moreover, the estimated standard errors with $\alpha = 2\%$ are mostly close to these results as well, with a few exceptions.

Figure 16 and Figure 17 show the same comparison for the relative bias and coefficient of variation of the quarterly turnover levels. Here, it is seen again that the selection of special cases has a large influence on the analytical bias approximation, whereas the analytical variance approximation is less sensitive to this selection.

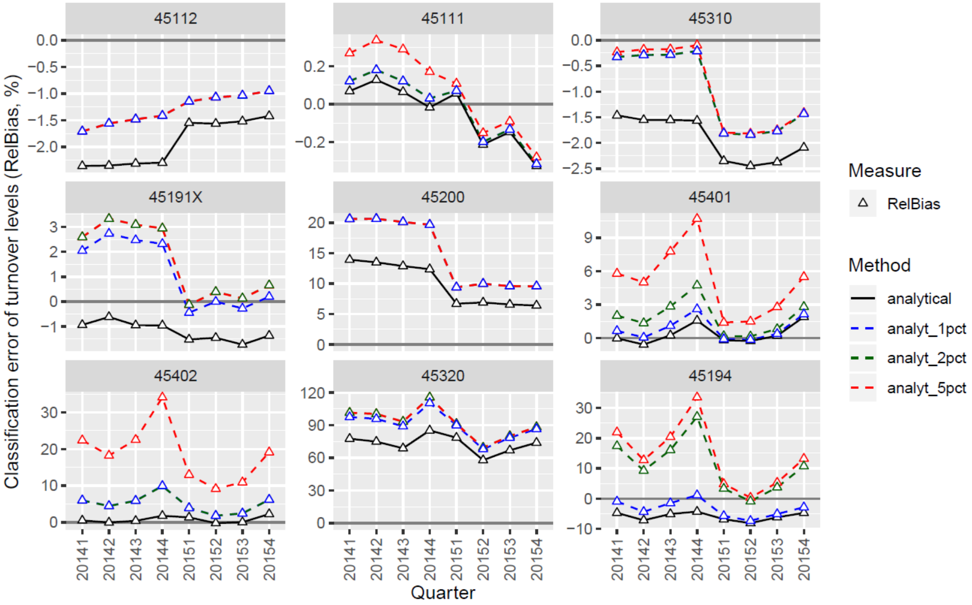


Figure 16. Comparison of analytical approximations to relative bias of quarterly turnover levels for nine industries in car trade, based on different selections of special cases (colors as in Figure 12).

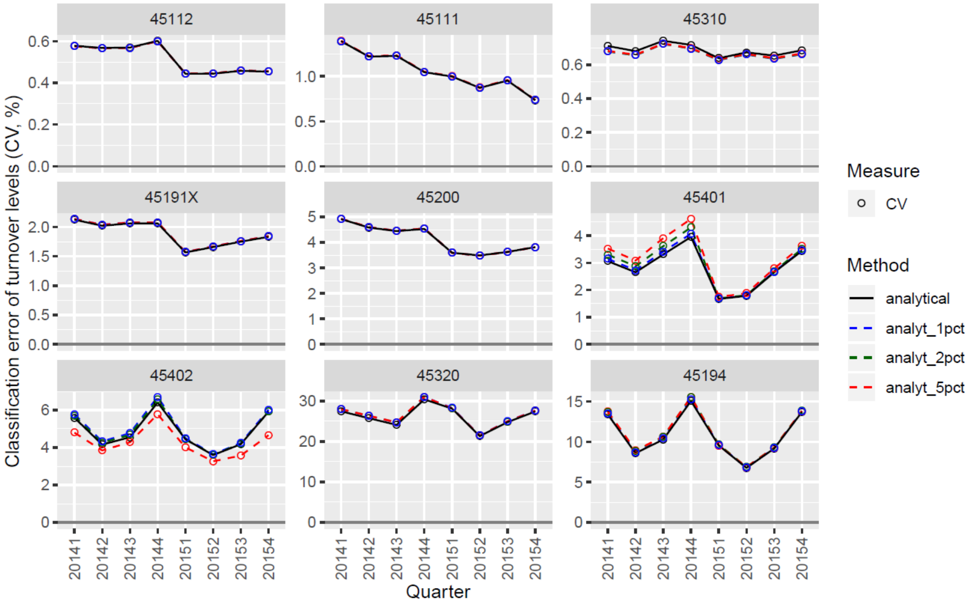


Figure 17. Comparison of analytical approximations to coefficient of variation of quarterly turnover levels for nine industries in car trade, based on different selections of special cases (colors as in Figure 12).

9.5.3 Analytical approximations based on a uniformity assumption

In practice, it is difficult to get accurate estimates of all parameters of the classification error models based on a relatively small audit sample. This holds in particular for the off-diagonal elements of the level matrix. The number of such elements increases quadratically with the size of $\mathcal{H}$. An alternative, rather crude approach could be to estimate only the diagonal probabilities of the level matrix, and assume that the misclassified units are distributed uniformly over the other codes. (Essentially, this is assumption A2' from Subsection 4.3.1.) It is interesting to see to what extent the analytical bias and variance estimates are affected by this approximation.

To test this, we have repeated the computation of the estimated level and level-change matrices outlined in Subsection 9.3 and 9.4, but this time with the conditional off-diagonal probabilities $\psi(g,h)$ in Figure 7 — as defined in (69) — all replaced by $\psi(g,h) = 1/9$. Note that this value is chosen in order to satisfy the restriction that $\sum_{h\neq g} \psi(g,h) = 1$ for all g. The remaining steps of the computation of the classification error probabilities were left unchanged. As an illustration, Table 13 shows the resulting estimated probabilities for the rows of the level matrix corresponding to car trade industries for the second probability class; this table may be compared to the original estimates in Table 9. Note that the diagonal elements are the same in both tables.

Table 13. Estimated level matrix for simple units with fewer than 10 EMP that consist of at least three LU, using uniform conditional off-diagonal probabilities.

Industry	45111	45112	45191X	45194	45200	45310	45320	45401	45402
45111	0.02464	*0.10837*	*0.10837*	*0.10837*	*0.10837*	*0.10837*	*0.10837*	*0.10837*	*0.10837*
45112	*0.01380*	0.87576	*0.01380*	*0.01380*	*0.01380*	*0.01380*	*0.01380*	*0.01380*	*0.01380*
45191X	*0.05204*	*0.05204*	0.53162	*0.05204*	*0.05204*	*0.05204*	*0.05204*	*0.05204*	*0.05204*
45194	*0.05346*	*0.05346*	*0.05346*	0.51882	*0.05346*	*0.05346*	*0.05346*	*0.05346*	*0.05346*
45200	*0.02879*	*0.02879*	*0.02879*	*0.02879*	0.74087	*0.02879*	*0.02879*	*0.02879*	*0.02879*
45310	*0.02796*	*0.02796*	*0.02796*	*0.02796*	*0.02796*	0.74832	*0.02796*	*0.02796*	*0.02796*
45320	*0.09285*	*0.09285*	*0.09285*	*0.09285*	*0.09285*	*0.09285*	0.16433	*0.09285*	*0.09285*
45401	*0.09492*	*0.09492*	*0.09492*	*0.09492*	*0.09492*	*0.09492*	*0.09492*	0.14569	*0.09492*
45402	*0.03026*	*0.03026*	*0.03026*	*0.03026*	*0.03026*	*0.03026*	*0.03026*	*0.03026*	0.72769

As the level matrix is especially relevant for level estimates, we begin by comparing the results of the two analytical approximations for the estimated quarterly turnover levels. Figure 18 shows the estimated accuracy of the turnover level estimates for both sets of estimated probabilities. It is seen that the estimated coefficients of variation (circles) using the uniformity assumption were reasonably accurate for the five industries with the largest total turnover, but less accurate for the remaining, smaller industries. The bias estimates (triangles) with the uniform assumption were less well-behaved.

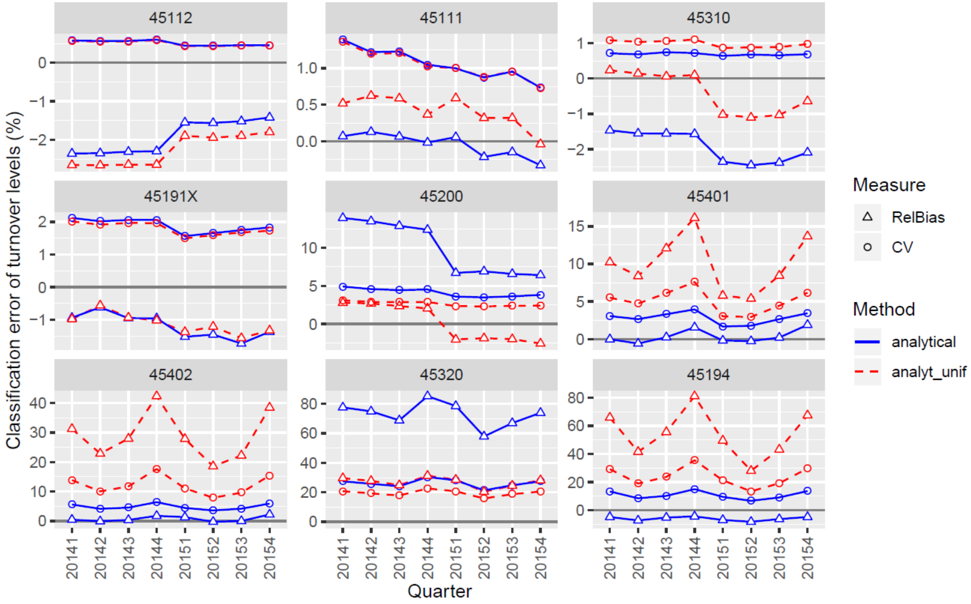


Figure 18. Comparison of relative bias (triangles) and coefficient of variation (circles) of quarterly turnover levels for nine industries in car trade, based on estimated (blue) or uniform (red) conditional misclassification probabilities.

To put these results into perspective, we computed the weighted average diagonal probability in each car trade industry, taking into account the share of turnover for each probability class; see Table 14. It is seen that the size of the deviation for the coefficients of variation in Figure 18 correlates reasonably well with the average diagonal probability: the largest deviations occurred for industries with the smallest average probability of correct classification. For the relative bias, the association is less clear.

Table 14. Average diagonal probability per car trade industry, weighted by turnover share of probability classes (averaged over Q2 2014 – Q4 2015).

Industry	Average diagonal probability
45112	0.970
45111	0.984
45310	0.977
45191X	0.948
45200	0.946
45401	0.948
45402	0.925
45320	0.715
45194	0.836

The results for turnover growth rates are shown in Figure 19 (quarter-on-quarter growth rates) and Figure 20 (year-on-year growth rates). Here, it is seen that the uniformity assumption mostly led to accurate standard errors (circles), even for the smaller industries. Moreover, in the cases where the standard errors based on the uniformity assumption deviated substantially from the original analytical standard errors, they usually provided a conservative estimate of precision. The bias estimates (triangles) were also quite accurate for the largest industries, but more erratic for the smallest industries.

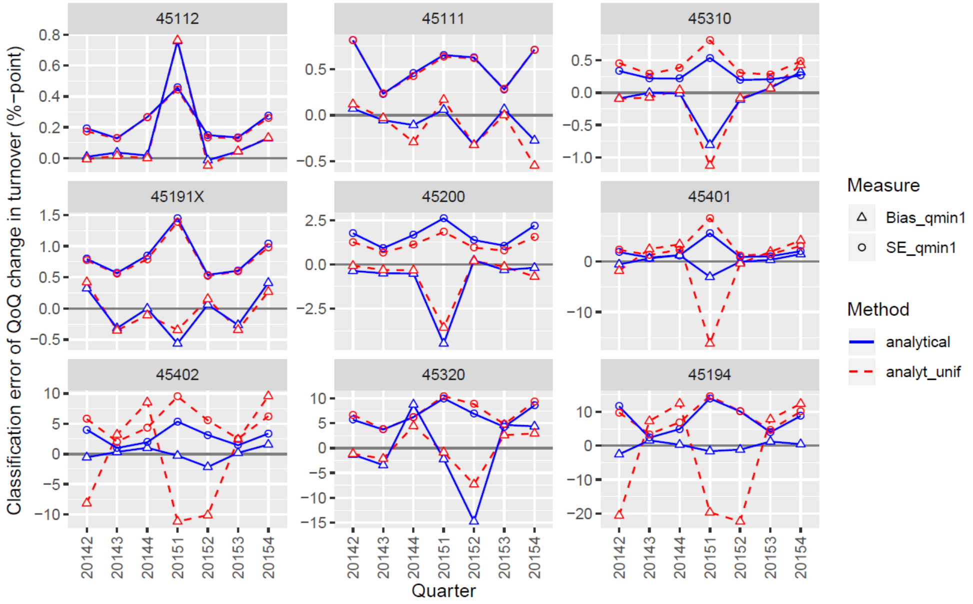


Figure 19. Comparison of bias (triangles) and standard errors (circles) of quarter-on-quarter turnover growth rates for nine industries in car trade, based on estimated (blue) or uniform (red) conditional misclassification probabilities.

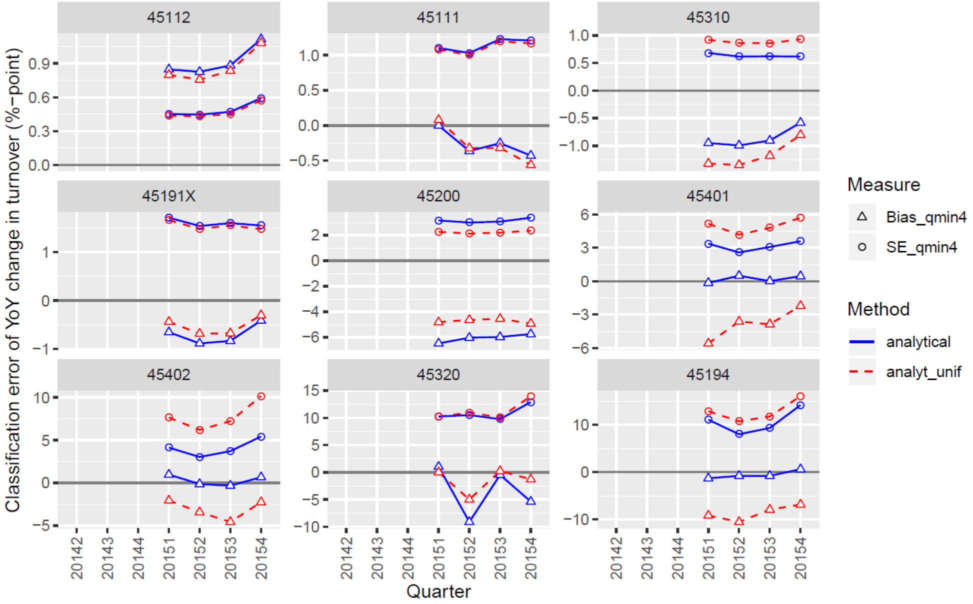


Figure 20. Comparison of bias (triangles) and standard errors (circles) of year-on-year turnover growth rates for nine industries in car trade, based on estimated (blue) or uniform (red) conditional misclassification probabilities.

10. Discussion

In this paper, we have presented analytical approximations and a bootstrap approach that can be used to estimate the bias and variance of growth rates by domain when random errors occur in the classification of units to domains. We have focussed in particular on an application to growth rates of turnover by industry (NACE code) when the classification of businesses in the GBR is affected by errors. In practice, each year the observed NACE code changes only for about 3 to 4% of all units in the GBR. Given the estimated classification error probabilities in Table 6, it follows that many errors in the observed NACE code persist over time, so that classification errors in this variable are strongly correlated over time.

Both approaches made use of a model for random classification errors. The model describes both 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. In principle, the error probabilities may also vary across units. In practice, one would usually allow the probabilities to vary only as a function of background variables (cf. the probability classes we introduced in Section 4). Although we have focussed on NACE codes in this paper, the same model could in principle also be applied to other classification variables. More restrictive classification error models can be obtained as special cases, for instance an exchangeable errors model (cf. Subsection 4.2.1) or a model that assumes that errors at different time points are independent. Several types of classification error probabilities occur as unknown parameters in the model that have to be estimated in practice.

NSIs have some flexibility in applying the approaches discussed in this paper. The formulas that have been derived in Sections 4 and 5 are actually a family of bias and variance approximations, in which the number of parameters can be varied by defining narrow or wide probability classes. 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 estimators 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 estimators may be more stable.

Results on simulated data in Section 8 showed that the analytical approximations and the bootstrap method should give similar outcomes for the bias and variance of growth rates when the target variable being aggregated has a distribution that is not highly skewed. In particular, this is true when one is interested in changes of population counts over time. When the target variable has a very skewed distribution (as in the case of turnover), the two approaches should still give similar results as long as the units with the largest values have a small to negligible probability of being misclassified. The results on real data in Section 9 confirmed this. In practice, the assumption that the largest or most influential units are classified with high accuracy in statistical production usually seems reasonable.

When the two approaches give different results, we consider the bootstrap to be more reliable because it is not based on a Taylor series approximation. Compared with the bootstrap, the main advantage of the analytical approximation is that it is much simpler to compute. The analytical approach could therefore be used to evaluate the effect of a potential improvement to a statistical process "on the fly", which, for large populations, would not be possible with the bootstrap. With the current state of computing technology, bootstrapping for large populations may require parallel processing and dedicated hardware. In addition, the analytical formulas provide an easy way to interpret the (approximate) causes of bias and/or high variance — e.g., which are the most influential misclassification patterns between domains —, and thereby indicate potential ways to improve an estimator.

In the case study of Section 9, the required parameter estimates were derived from estimated classification error probabilities in previous studies. These estimates were based on audit samples of businesses, historical data on observed transitions between industries in the GBR, and interviews with experts. 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. A question for future research is therefore whether error probabilities can be estimated without recourse to an audit sample. One option might be to collect paradata on misclassifications as part of the editing process in regular production. Here, selectivity of the paradata may be a problem, as production editing tends to focus on the most influential observations. Another idea is to try to obtain an alternative classification of enterprises to industries from a source that is independent of the GBR, for instance a text mining algorithm that is applied to information on the websites of the enterprises. From these (at least) two independent sets of observed industry codes, classification errors could be modelled by a latent class model (Biemer, 2011). The estimated error probabilities from this model could then be used as input for the analytical or bootstrap approach in this paper.

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 have 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. Subsection 7.2.2 and van Delden et al., 2016b) or in the above-mentioned latent class model. The latter approach seems to be more natural for new applications.

In practice, the bias and variance estimators in Subsections 4.5 and 5.4 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 can be shown that the bootstrap estimators for bias and variance suffer from the same theoretical bias; cf. van Delden et al. (2016b). 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. We did not take this into account here; in principle, this could be done by extending the bootstrap method to include the uncertainty in the classification error probabilities. Meertens et al. (2019a) proposed a Bayesian method to evaluate the effect of classification errors which takes the fact that the error probabilities are estimated into account in the posterior distribution of the estimator.

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 of which the first- and second-order partial derivatives are supposed to exist. 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} that is affected by classification errors. Then, in analogy with the derivation in Appendix A, it can be shown that

$$
\begin{aligned}
B(\hat{\theta}_h) &\approx \frac{1}{2}\sum_{q=1}^{Q}\sum_{r=1}^{Q}\frac{\partial^2 f}{\partial x_q \partial x_r}(\mu_{1h}, \dots, \mu_{Qh})\, C(\hat{X}_{qh}, \hat{X}_{rh}) + f(\mu_{1h}, \dots, \mu_{Qh}) \\
&\qquad - \theta_h, \qquad (74) \\
V(\hat{\theta}_h) &\approx \sum_{q=1}^{Q}\sum_{r=1}^{Q}\frac{\partial f}{\partial x_q}(\mu_{1h}, \dots, \mu_{Qh})\frac{\partial f}{\partial x_r}(\mu_{1h}, \dots, \mu_{Qh})\, C(\hat{X}_{qh}, \hat{X}_{rh}),
\end{aligned}
$$

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 such as that of Section 7. The precise form of the resulting bias and variance approximations depends on the assumptions in this model. Alternatively, the bootstrap approach of Section 6 can be readily applied to statistics $\hat{\theta}_h$ of this form, and potentially even to statistics that are not linearisable by a Taylor series, such as domain medians (Efron and Tibshirani, 1993).

As a possible example of a different application, suppose that one is interested in statistics on persons classified by education level. An interesting target parameter could be the average 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 (74) with $\theta_h = X_{1h}/X_{2h}$. In fact, modelling the classification errors for this example would be much simpler than for the growth rates we considered in this paper, since it does not involve changes over time.

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 domains. It is then natural to wonder how the accuracy could be improved, i.e., how the effect of classification errors on those domains could be reduced. Two different approaches are: computing a bias-corrected estimator using the estimated bias, or improving the accuracy by editing observed codes at the micro level.

For the first approach, Meertens et al. (2019b) examined the accuracy of several bias-corrected estimators for a case study. The accuracy of the estimator is not guaranteed to be improved by this approach: the bias estimator itself may have a large uncertainty and — as noted in Subsections 4.5 and 5.4 — it may be biased in practice. The micro-level approach 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

P.P. Biemer (2011). *Latent Class Analysis of Survey Error*. Hoboken, New Jersey: John Wiley & Sons.

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: 19 June 2019).

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.

S. van der Doef, P. Daas, and D. Windmeijer (2018). Identifying Innovative Companies from their Website. Presentation at BigSurv18 conference. Abstract available at http://www.bigsurv18.org/program2018?sess=34#205 (retrieved: 19 June 2019).

B. Efron and R.J. Tibshirani (1993). *An Introduction to the Bootstrap*. London: Chapman & Hall/CRC.

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/web/products-manuals-and-guidelines/-/KS-RA-07-015 (retrieved: 19 June 2019).

J. Kuha and C. Skinner (1997). Categorical Data Analysis and Misclassification. In: Lyberg, Biemer, Collins, de Leeuw, Dippo, Schwarz, and Trewin (eds.), *Survey Measurement and Process Quality*. New York: John Wiley & Sons, pp. 633–670.

R.J. Little and D.B. Rubin (2002). *Statistical Analysis with Missing Data* (2nd Edition). Hoboken, New Jersey: John Wiley & Sons.

P. McCullagh and J.A. Nelder (1989). *Generalized Linear Models* (2nd Edition). London: Chapman & Hall.

Q.A. Meertens, C.G.H. Diks, H.J. van den Herik, and F.W. Takes (2018). A Data-Driven Supply-Side Approach for Measuring Cross-Border Internet Purchases. Available at arXiv:1805.06930v1 [stat.AP] (retrieved: 19 June 2019).

Q.A. Meertens, C.G.H. Diks, H.J. van den Herik, and F.W. Takes (2019a). A Bayesian Approach for Accurate Classification-Based Aggregates. In: *Proceedings of the 19th SIAM International Conference on Data Mining, Calgary*, pages 306–314. Preprint available at arXiv:1902.02412v1 [stat.ML] (retrieved: 19 June 2019).

Q.A. Meertens, A. van Delden, S. Scholtus, and F.W. Takes (2019b). Bias Correction for Predicting Election Outcomes with Social Media Data. Paper presented at the 5th International Conference on Computational Social Science, Amsterdam. Available at http://www.researchgate.net/publication/333661444_Bias_Correction_for_Predicting_Election_Outcomes_with_Social_Media_Data (retrieved: 27 August 2019).

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 A. Derivation of (1)

In this appendix to Subsection 2.2, approximation (1) for the bias and variance of a generic estimated growth rate under classification errors will be derived. –

Bias

To obtain a formula for the approximate bias of the estimated growth rate $\hat{g}_h^{q,q-u} = \hat{G}_h^{q,q-u} - 1$, 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:

$$\begin{aligned} \frac{u}{v} &\approx \frac{u_0}{v_0}\left\{1 + \frac{1}{u_0}(u-u_0) - \frac{1}{v_0}(v-v_0) \right. \\ &\qquad\qquad \left. + \frac{1}{2}\left[0 - \frac{2}{v_0 u_0}(v-v_0)(u-u_0) + \frac{2}{v_0^2}(v-v_0)^2\right]\Bigg]\right\} \qquad \text{(A1)} \\ &= \frac{u_0}{v_0}\left\{1 + \frac{u-u_0}{u_0} - \frac{v-v_0}{v_0} + \frac{(v-v_0)^2}{v_0^2} - \frac{(v-v_0)(u-u_0)}{v_0 u_0}\right\}. \end{aligned}$$

Using (A1), the second-order Taylor expansion of $\hat{g}_h^{q,q-u} = \hat{Y}_h^q / \hat{Y}_h^{q-u} - 1$ in a neighbourhood of the point $\left(E(\hat{Y}_h^q), E(\hat{Y}_h^{q-u})\right)$ is given by:

$$\begin{aligned} \hat{g}_h^{q,q-u} \approx \frac{E(\hat{Y}_h^q)}{E(\hat{Y}_h^{q-u})}&\left\{1 + \frac{\hat{Y}_h^q - E(\hat{Y}_h^q)}{E(\hat{Y}_h^q)} - \frac{\hat{Y}_h^{q-u} - E(\hat{Y}_h^{q-u})}{E(\hat{Y}_h^{q-u})} \right. \\ &\quad + \frac{\left[\hat{Y}_h^{q-u} - E(\hat{Y}_h^{q-u})\right]^2}{\left[E(\hat{Y}_h^{q-u})\right]^2} \qquad \text{(A2)} \\ &\quad \left. - \frac{\left[\hat{Y}_h^{q-u} - E(\hat{Y}_h^{q-u})\right]\left[\hat{Y}_h^q - E(\hat{Y}_h^q)\right]}{E(\hat{Y}_h^{q-u})E(\hat{Y}_h^q)}\right\} - 1. \end{aligned}$$

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

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

where $V(.)$ denotes a variance and $C(.,.)$ denotes a covariance. To simplify the notation, let $\breve{G}_h^{q,q-u} = E(\hat{Y}_h^q)/E(\hat{Y}_h^{q-u})$. The bias $B(\hat{g}_h^{q,q-u})$ can now be written as:

$$\begin{aligned} B\left(\hat{g}_h^{q,q-u}\right) &= E\left(\hat{g}_h^{q,q-u}\right) - g_h^{q,q-u} \\ &\approx \breve{G}_h^{q,q-u}\left\{\frac{V\left(\hat{Y}_h^{q-u}\right)}{\left[E\left(\hat{Y}_h^{q-u}\right)\right]^2} - \frac{C\left(\hat{Y}_h^{q-u}, \hat{Y}_h^q\right)}{E\left(\hat{Y}_h^{q-u}\right)E\left(\hat{Y}_h^q\right)}\right\} + \left(\breve{G}_h^{q,q-u} - 1\right) \\ &\quad - \left(G_h^{q,q-u} - 1\right) \\ &= \frac{1}{\left[E\left(\hat{Y}_h^{q-u}\right)\right]^2}\left\{\breve{G}_h^{q,q-u} V\left(\hat{Y}_h^{q-u}\right) - C\left(\hat{Y}_h^{q-u}, \hat{Y}_h^q\right)\right\} \\ &\quad + \left(\breve{G}_h^{q,q-u} - G_h^{q,q-u}\right). \end{aligned} \tag{A4}$$

Expression (A4) can be refined further by making use of the assumption that classification errors are independent across units. The variance $V\left(\hat{Y}_h^{q-u}\right)$ can therefore be written as:

$$V\left(\hat{Y}_h^{q-u}\right) = V\left(\sum\nolimits_{i\in U^{q-u}} \hat{a}_{hi}^{q-u} y_i^{q-u}\right) = \sum\nolimits_{i\in U^{q-u}} (y_i^{q-u})^2 V\left(\hat{a}_{hi}^{q-u}\right), \tag{A5}$$

Furthermore, recall that $\hat{Y}_h^{q-u}, \hat{Y}_h^q$ partly refer to different sets of units. We can write: $\hat{Y}_h^{q-u} = \hat{Y}_{hSEP}^{q-u} + \hat{Y}_{hOLP}^{q-u}$ where $\hat{Y}_{hSEP}^{q-u} = \sum_{i\in U_D^{q-u,q}} \hat{a}_{hi}^{q-u} y_i^{q-u}$ and $\hat{Y}_{hOLP}^{q-u} = \sum_{U_O^{q-u,q}} \hat{a}_{hi}^{q-u} y_i^{q-u}$. Likewise, $\hat{Y}_h^q = \hat{Y}_{hSEP}^q + \hat{Y}_{hOLP}^q$ where $\hat{Y}_{hSEP}^q = \sum_{i\in U_B^{q-u,q}} \hat{a}_{hi}^q y_i^q$ and $\hat{Y}_{hOLP}^q = \sum_{U_O^{q-u,q}} \hat{a}_{hi}^q y_i^q$. The covariance $C\left(\hat{Y}_h^{q-u}, \hat{Y}_h^q\right)$ can therefore be written as:

$$\begin{aligned} C\left(\hat{Y}_h^{q-u}, \hat{Y}_h^q\right) &= C\left(\hat{Y}_{hSEP}^{q-u} + \hat{Y}_{hOLP}^{q-u}, \hat{Y}_{hSEP}^q + \hat{Y}_{hOLP}^q\right) \\ &= C\left(\hat{Y}_{hOLP}^{q-u}, \hat{Y}_{hOLP}^q\right) \\ &= C\left(\sum\nolimits_{i\in U_O^{q-u,q}} \hat{a}_{hi}^{q-u} y_i^{q-u}, \sum\nolimits_{i\in U_O^{q-u,q}} \hat{a}_{hi}^q y_i^q\right) \\ &= \sum\nolimits_{i\in U_O^{q-u,q}} y_i^{q-u} y_i^q C\left(\hat{a}_{hi}^{q-u}, \hat{a}_{hi}^q\right). \end{aligned} \tag{A6}$$

Combining (A4), (A5) and (A6), we obtain:

$$\begin{aligned} B\left(\hat{g}_h^{q,q-u}\right) \approx \frac{1}{\left[E\left(\hat{Y}_h^{q-u}\right)\right]^2}\Bigg\{ &\breve{G}_h^{q,q-u} \sum\nolimits_{i\in U^{q-u}} (y_i^{q-u})^2 V\left(\hat{a}_{hi}^{q-u}\right) \\ &- \sum\nolimits_{i\in U_O^{q-u,q}} y_i^{q-u} y_i^q C\left(\hat{a}_{hi}^{q-u}, \hat{a}_{hi}^q\right)\Bigg\} + \left(\breve{G}_h^{q,q-u} - G_h^{q,q-u}\right), \end{aligned}$$

where furthermore $E\left(\hat{Y}_h^{q-u}\right) = \sum_{i\in U^{q-u}} y_i^{q-u} E\left(\hat{a}_{hi}^{q-u}\right)$ and $\breve{G}_h^{q,q-u} = \sum_{i\in U^q} y_i^q E\left(\hat{a}_{hi}^q\right) / \sum_{i\in U^{q-u}} y_i^{q-u} E\left(\hat{a}_{hi}^{q-u}\right)$. This yields the expression for the approximate bias in formula (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 (cf. expression (A1)):

$$\frac{u}{v} \approx \frac{u_0}{v_0}\left\{1 + \frac{u}{u_0} - \frac{v}{v_0}\right\} = \frac{u_0}{v_0} + \frac{1}{v_0}\left(u - \frac{u_0}{v_0} v\right) \tag{A7}$$

Using (A7), $V\left(\hat{g}_h^{q,q-u}\right)$ can be approximated as follows:

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

where in the fifth and sixth line we used (A5) and (A6). This yields the expression for the approximate variance in formula (1).

Appendix B. Derivation of (9)

We start by writing $\mathbf{P}_i^C = (p_{gklhi}^C)$ as a two-dimensional matrix. Let the rows denote all possible combinations of (g, k, l) and the columns denote all possible h; this makes $\mathbf{P}_i^C$ a $M^3 \times M$ matrix. The $M \times M$ sub-matrix of $\mathbf{P}_i^C$ that applies to units with $s_i^{q-4} = g$ and $s_i^q = k$ is denoted as $\widetilde{\mathbf{P}}_{i|gk}^C$. [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 $\mathbf{P}_i^C$ consists of these blocks of rows, and we suppose that they are ordered lexicographically, so starting with $\widetilde{\mathbf{P}}_{i|11}^C$, $\widetilde{\mathbf{P}}_{i|12}^C$, etc., and ending with $\widetilde{\mathbf{P}}_{i|MM}^C$.

Since the true strata s_i^{q-4} and s_i^q are considered fixed, it will be useful to have a short-hand expression for the matrix that selects the $M \times M$ block $\widetilde{\mathbf{P}}_{i|gk}^C$ from $\mathbf{P}_i^C$ 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

$$\mathbf{\Pi}_i^C = \mathbf{\Lambda}_i \mathbf{P}_i^C, \text{with } \mathbf{\Lambda}_i = \left(\boldsymbol{a}_i^{q-4} \otimes \boldsymbol{a}_i^q\right)^T \otimes \mathbf{I}_M, \tag{B1}$$

where $\mathbf{I}_M$ denotes the $M \times M$ identity matrix and $\otimes$ denotes a Kronecker product. The operation of pre-multiplying $\mathbf{P}_i^C$ by the $M \times M^3$ matrix $\mathbf{\Lambda}_i$ selects out the M rows of $\mathbf{P}_i^C$ that correspond to (g, k, l) with $g = s_i^{q-4}$ and $k = s_i^q$, i.e., exactly those rows that apply to unit i. It follows from formula (B1) that

$$\mathbf{\Pi}_i^C = \sum_{g=1}^{M} \sum_{k=1}^{M} a_{gi}^{q-4} a_{ki}^q \widetilde{\mathbf{P}}_{i|gk}^C. \tag{B2}$$

The matrix $\mathbf{\Pi}_i^C = \mathbf{\Lambda}_i \mathbf{P}_i^C$ contains the relevant transition probabilities from $\hat{s}_i^{q-4}$ to $\hat{s}_i^q$, just like $\mathbf{P}_i^L$ contains transition probabilities from s_i^{q-4} to $\hat{s}_i^{q-4}$. By analogy to (2), this implies that

$$\begin{aligned} E\left(\widehat{\boldsymbol{a}}_i^q \middle| \hat{s}_i^{q-4}\right) &= (\mathbf{\Pi}_i^C)^T \widehat{\boldsymbol{a}}_i^{q-4}, \\ V\left(\widehat{\boldsymbol{a}}_i^q \middle| \hat{s}_i^{q-4}\right) &= \operatorname{diag}\left[(\mathbf{\Pi}_i^C)^T \widehat{\boldsymbol{a}}_i^{q-4}\right] - (\mathbf{\Pi}_i^C)^T \operatorname{diag}(\widehat{\boldsymbol{a}}_i^{q-4})\, \mathbf{\Pi}_i^C. \end{aligned} \tag{B3}$$

Using standard expansion rules for conditional expectations and (co)variances, we obtain from (2) and (B3):

$$E\left(\widehat{\boldsymbol{a}}_i^q\right) = E\left[E\left(\widehat{\boldsymbol{a}}_i^q \middle| \hat{s}_i^{q-4}\right)\right] = (\mathbf{\Pi}_i^C)^T (\mathbf{P}_i^L)^T \boldsymbol{a}_i^{q-4} \tag{B4}$$

and

$$\begin{aligned} V\left(\widehat{\boldsymbol{a}}_i^q\right) &= E\left[V\left(\widehat{\boldsymbol{a}}_i^q \middle| \hat{s}_i^{q-4}\right)\right] + V\left[E\left(\widehat{\boldsymbol{a}}_i^q \middle| \hat{s}_i^{q-4}\right)\right] \\ &= E\left\{\operatorname{diag}\left[(\mathbf{\Pi}_i^C)^T \widehat{\boldsymbol{a}}_i^{q-4}\right] - (\mathbf{\Pi}_i^C)^T \operatorname{diag}(\widehat{\boldsymbol{a}}_i^{q-4})\, \mathbf{\Pi}_i^C\right\} + V\left[(\mathbf{\Pi}_i^C)^T \widehat{\boldsymbol{a}}_i^{q-4}\right] \\ &= \operatorname{diag}\left[(\mathbf{\Pi}_i^C)^T (\mathbf{P}_i^L)^T \boldsymbol{a}_i^{q-4}\right] - (\mathbf{\Pi}_i^C)^T \operatorname{diag}\left[(\mathbf{P}_i^L)^T \boldsymbol{a}_i^{q-4}\right] \mathbf{\Pi}_i^C \\ &\qquad + (\mathbf{\Pi}_i^C)^T \left\{\operatorname{diag}\left[(\mathbf{P}_i^L)^T \boldsymbol{a}_i^{q-4}\right] - (\mathbf{P}_i^L)^T \operatorname{diag}(\boldsymbol{a}_i^{q-4})\, \mathbf{P}_i^L\right\} \mathbf{\Pi}_i^C \\ &= \operatorname{diag}\left[(\mathbf{\Pi}_i^C)^T (\mathbf{P}_i^L)^T \boldsymbol{a}_i^{q-4}\right] - (\mathbf{\Pi}_i^C)^T (\mathbf{P}_i^L)^T \operatorname{diag}(\boldsymbol{a}_i^{q-4})\, \mathbf{P}_i^L \mathbf{\Pi}_i^C \end{aligned} \tag{B5}$$

and

$$\begin{aligned}
C\left(\hat{\boldsymbol{a}}_i^{q-4}, \hat{\boldsymbol{a}}_i^{q}\right) &= E\left[C\left(\hat{\boldsymbol{a}}_i^{q-4}, \hat{\boldsymbol{a}}_i^{q} \middle| \hat{s}_i^{q-4}\right)\right] + C\left[E\left(\hat{\boldsymbol{a}}_i^{q-4} \middle| \hat{s}_i^{q-4}\right), E\left(\hat{\boldsymbol{a}}_i^{q} \middle| \hat{s}_i^{q-4}\right)\right] \\
&= 0 + C\left[\hat{\boldsymbol{a}}_i^{q-4}, (\boldsymbol{\Pi}_i^C)^T \hat{\boldsymbol{a}}_i^{q-4}\right] \\
&= V\left(\hat{\boldsymbol{a}}_i^{q-4}\right) \boldsymbol{\Pi}_i^C \\
&= \left\{\operatorname{diag}\left[(\mathbf{P}_i^L)^T \boldsymbol{a}_i^{q-4}\right] - (\mathbf{P}_i^L)^T \operatorname{diag}\left(\boldsymbol{a}_i^{q-4}\right) \mathbf{P}_i^L\right\} \boldsymbol{\Pi}_i^C.
\end{aligned} \tag{B6}$$

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 $\operatorname{diag}\left(\boldsymbol{a}_i^{q-4}\right) = \boldsymbol{a}_i^{q-4}\left(\boldsymbol{a}_i^{q-4}\right)^T$. For the derivation below, it is useful to re-write (B5) and (B6) as follows:

$$\begin{aligned}
V\left(\hat{\boldsymbol{a}}_i^{q}\right) &= \operatorname{diag}\left[(\boldsymbol{\Pi}_i^C)^T (\mathbf{P}_i^L)^T \boldsymbol{a}_i^{q-4}\right] - (\boldsymbol{\Pi}_i^C)^T (\mathbf{P}_i^L)^T \boldsymbol{a}_i^{q-4}\left(\boldsymbol{a}_i^{q-4}\right)^T \mathbf{P}_i^L \boldsymbol{\Pi}_i^C, \\
C\left(\hat{\boldsymbol{a}}_i^{q-4}, \hat{\boldsymbol{a}}_i^{q}\right) &= \operatorname{diag}\left[(\mathbf{P}_i^L)^T \boldsymbol{a}_i^{q-4}\right] \boldsymbol{\Pi}_i^C - (\mathbf{P}_i^L)^T \boldsymbol{a}_i^{q-4}\left(\boldsymbol{a}_i^{q-4}\right)^T \mathbf{P}_i^L \boldsymbol{\Pi}_i^C.
\end{aligned} \tag{B7}$$

Thus, $E\left(\hat{\boldsymbol{a}}_i^{q}\right)$ is equal to $(\boldsymbol{\Pi}_i^C)^T (\mathbf{P}_i^L)^T \boldsymbol{a}_i^{q-4}$ according to (B4) and this quantity [as well as its transpose $\left(\boldsymbol{a}_i^{q-4}\right)^T \mathbf{P}_i^L \boldsymbol{\Pi}_i^C$] also occurs in $V\left(\hat{\boldsymbol{a}}_i^{q}\right)$ and $C\left(\hat{\boldsymbol{a}}_i^{q-4}, \hat{\boldsymbol{a}}_i^{q}\right)$ according to (B7). In fact, $(\boldsymbol{\Pi}_i^C)^T (\mathbf{P}_i^L)^T \boldsymbol{a}_i^{q-4}$ is a column vector of length M of which the elements are given by [cf. (B2)]:

$$\begin{aligned}
\left[(\boldsymbol{\Pi}_i^C)^T (\mathbf{P}_i^L)^T \boldsymbol{a}_i^{q-4}\right]_h &= \sum_{g=1}^{M} a_{gi}^{q-4} (\mathbf{P}_i^L \boldsymbol{\Pi}_i^C)_{gh} \\
= \sum_{g=1}^{M} a_{gi}^{q-4} &\left[\sum_{l=1}^{M} p_{gli}^L \sum_{j=1}^{M} \sum_{k=1}^{M} a_{ji}^{q-4} a_{ki}^{q} p_{jklhi}^C\right] \\
&= \sum_{g=1}^{M} \sum_{k=1}^{M} a_{gi}^{q-4} a_{ki}^{q} \sum_{l=1}^{M} p_{gli}^L p_{gklhi}^C \\
&= \sum_{g=1}^{M} \sum_{k=1}^{M} a_{gi}^{q-4} a_{ki}^{q} p_{gkhi}^{LC}.
\end{aligned} \tag{B8}$$

In the third line, we used the fact that $a_{gi}^{q-4} a_{ji}^{q-4} = 0$ when $j \neq g$ and $\left(a_{gi}^{q-4}\right)^2 = a_{gi}^{q-4}$. In the last line, we used the notation $p_{gkhi}^{LC} = \sum_{l=1}^{M} \mathbb{p}_{gklhi}^{LC} = \sum_{l=1}^{M} p_{gli}^L p_{gklhi}^C$ defined in Subsection 2.4.

It follows directly from (B4) and (B8) that

$$E\left(\hat{a}_{hi}^{q}\right) = \left[(\boldsymbol{\Pi}_i^C)^T (\mathbf{P}_i^L)^T \boldsymbol{a}_i^{q-4}\right]_h = \sum_{g=1}^{M} \sum_{k=1}^{M} a_{gi}^{q-4} a_{ki}^{q} p_{gkhi}^{LC}.$$

Similarly, it follows from (B7) and (B8) that

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

The second term can be re-arranged as:

$$\left(\sum_{g=1}^{M}\sum_{k=1}^{M} a_{gi}^{q-4} a_{ki}^{q} p_{gkhi}^{LC}\right)^2 = \sum_{g=1}^{M}\left(a_{gi}^{q-4}\sum_{k=1}^{M} a_{ki}^{q} p_{gkhi}^{LC}\right)^2 = \sum_{g=1}^{M}\sum_{k=1}^{M} a_{gi}^{q-4} a_{ki}^{q}\left(p_{gkhi}^{LC}\right)^2,$$

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

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

Finally, for the covariance between $\hat{a}_{hi}^{q-4}$ and $\hat{a}_{hi}^{q}$, it follows from (B7), (B8) and (B2) that

$$\begin{aligned}
C\left(\hat{a}_{hi}^{q-4}, \hat{a}_{hi}^{q}\right) &= \sum_{g=1}^{M} a_{gi}^{q-4} p_{ghi}^{L} (\boldsymbol{\Pi}_i^{C})_{hh} - \left(\sum_{g=1}^{M} a_{gi}^{q-4} p_{ghi}^{L}\right)\left(\sum_{g=1}^{M}\sum_{k=1}^{M} a_{gi}^{q-4} a_{ki}^{q} p_{gkhi}^{LC}\right) \\
&= \sum_{g=1}^{M} a_{gi}^{q-4} p_{ghi}^{L} \sum_{k=1}^{M} a_{ki}^{q} p_{gkhhi}^{C} - \sum_{g=1}^{M} a_{gi}^{q-4} p_{ghi}^{L} \sum_{k=1}^{M} a_{ki}^{q} p_{gkhi}^{LC} \\
&= \sum_{g=1}^{M}\sum_{k=1}^{M} a_{gi}^{q-4} a_{ki}^{q} \mathbb{p}_{gkhhi}^{LC} - \sum_{g=1}^{M}\sum_{k=1}^{M} a_{gi}^{q-4} a_{ki}^{q} p_{ghi}^{L} p_{gkhi}^{LC}.
\end{aligned}$$

This completes the derivation of the three expressions in (9).

Appendix C. Summary of the classification error model of Table 3 in matrix notation

The model for $p_{gklhi}^{C} = P\left(\hat{s}_i^q = h \middle| s_i^{q-4} = g, s_i^q = k, \hat{s}_i^{q-4} = l\right)$ defined in Table 3 can be summarised as follows in terms of the matrix $\mathbf{\Pi}_i^C = \mathbf{\Lambda}_i \mathbf{P}_i^C$ introduced in (B1)–(B2):

$$\begin{aligned}
\mathbf{\Pi}_i^C &= \mathbf{\Pi}^C\left(\boldsymbol{a}_i^{q-4}, \boldsymbol{a}_i^q\right) \\
&= \left(\boldsymbol{a}_i^{q-4}\right)^T \boldsymbol{a}_i^q \left\{\operatorname{diag}\left(\boldsymbol{a}_i^q\right) \mathbf{A} + \left[\mathbf{I}_M - \operatorname{diag}\left(\boldsymbol{a}_i^q\right)\right] \mathbf{B}\left(\boldsymbol{a}_i^q\right)\right\} \\
&\qquad\qquad + \left[1 - \left(\boldsymbol{a}_i^{q-4}\right)^T \boldsymbol{a}_i^q\right] \left\{\operatorname{diag}\left(\boldsymbol{a}_i^q\right) \mathbf{A} + \left[\mathbf{I}_M - \operatorname{diag}\left(\boldsymbol{a}_i^q\right)\right] \mathbf{C}\left(\boldsymbol{a}_i^q\right)\right\} \\
&= \left[\mathbf{I}_M - \operatorname{diag}\left(\boldsymbol{a}_i^q\right)\right] \left\{\left(\boldsymbol{a}_i^{q-4}\right)^T \boldsymbol{a}_i^q \mathbf{B}\left(\boldsymbol{a}_i^q\right) + \left[1 - \left(\boldsymbol{a}_i^{q-4}\right)^T \boldsymbol{a}_i^q\right] \mathbf{C}\left(\boldsymbol{a}_i^q\right)\right\} \\
&\qquad\qquad + \operatorname{diag}\left(\boldsymbol{a}_i^q\right) \mathbf{A},
\end{aligned} \tag{C1}$$

with

$$\begin{aligned}
\mathbf{A} &= \mathbf{I}_M + p_S(\mathbf{R} - \mathbf{I}_M), \\
\mathbf{B}\left(\boldsymbol{a}_i^q\right) &= \frac{1}{1 - p_R p_S}\Big[(1 - p_R)(1 - p_S)\mathbf{I}_M + (1 - p_R)p_S \mathbf{R} \\
&\qquad\qquad + p_R(1 - p_S)\mathbf{1}_M \left(\boldsymbol{a}_i^q\right)^T\Big], \\
\mathbf{C}\left(\boldsymbol{a}_i^q\right) &= \frac{1}{1 - p_N p_S}\Big[(1 - p_N)(1 - p_S)\mathbf{I}_M + (1 - p_N)p_S \mathbf{R} \\
&\qquad\qquad + p_N(1 - p_S)\mathbf{1}_M \left(\boldsymbol{a}_i^q\right)^T\Big],
\end{aligned} \tag{C2}$$

where furthermore $\mathbf{R}$ is an $M \times M$ matrix that 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. Expressions (C1) and (C2) therefore follow from the description in Table 3 by observing that $\operatorname{diag}\left(\boldsymbol{a}_i^q\right) \mathbf{A} + \left[\mathbf{I}_M - \operatorname{diag}\left(\boldsymbol{a}_i^q\right)\right] \mathbf{B}\left(\boldsymbol{a}_i^q\right)$ is the form of $\mathbf{\Pi}_i^C$ when $s_i^{q-4} = s_i^q$ and $\operatorname{diag}\left(\boldsymbol{a}_i^q\right) \mathbf{A} + \left[\mathbf{I}_M - \operatorname{diag}\left(\boldsymbol{a}_i^q\right)\right] \mathbf{C}\left(\boldsymbol{a}_i^q\right)$ is the form of $\mathbf{\Pi}_i^C$ 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 Table 3, and the matrices $\mathbf{B}\left(\boldsymbol{a}_i^q\right)$ and $\mathbf{C}\left(\boldsymbol{a}_i^q\right)$ contain the probabilities that hold for unit i under Situations B and C, respectively.

The notation $\mathbf{\Pi}_i^C = \mathbf{\Pi}^C\left(\boldsymbol{a}_i^{q-4}, \boldsymbol{a}_i^q\right)$ in (C1) 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_{gklhi}^C. Since $s_i^{q-4} = g$ and $s_i^q = k$ are already included as indices in the notation of these probabilities, we can remove the index i and write $p_{gklhi}^C = p_{gklh}^C$.

Explanation of symbols

Empty cell	Figure not applicable
.	Figure is unknown, insufficiently reliable or confidential
*	Provisional figure
**	Revised provisional figure
2017–2018	2017 to 2018 inclusive
2017/2018	Average for 2017 to 2018 inclusive
2017/’18	Crop year, financial year, school year, etc., beginning in 2017 and ending in 2018
2013/’14–2017/’18	Crop year, financial year, etc., 2015/’16 to 2017/’18 inclusive

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

Colophon

Publisher
Centraal Bureau voor de Statistiek
Henri Faasdreef 312, 2492 JP Den Haag
www.cbs.nl

Prepress
Statistics Netherlands, CCN Creation and visualisation

Design
Edenspiekermann

Information
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contactform: www.cbsl.nl/information

© Statistics Netherlands, The Hague/Heerlen/Bonaire 2018.
Reproduction is permitted, provided Statistics Netherlands is quoted as the source.