### Effect of classification errors on domain level estimates in business statistics

#### General setting

<table>
<thead>
<tr>
<th>Configuration</th>
<th>Configuration 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type of sources</td>
<td>Combination of administrative with survey data</td>
</tr>
<tr>
<td>Statistic of interest</td>
<td>Domain (strata) estimates in business statistics</td>
</tr>
<tr>
<td>Type of errors</td>
<td>Classification errors</td>
</tr>
<tr>
<td>Quality measure</td>
<td>Estimation of bias and variance of the statistic of interest</td>
</tr>
</tbody>
</table>

#### Method (use one block per method)

| Description | The method uses three basic steps. First classification errors are modelled, second the error sizes are estimated by collecting independent data and third the accuracy is estimated using a bootstrap approach. We consider the relatively simple case where classification errors are the only errors that affect the publication figures. This implies also that there are no sampling nor non-response errors, so we have values for all units in the population. |

*First step.* Consider a population of units \((i = 1, ..., N)\) that is divided into categories of a classification, e.g. enterprises classified by economic activity (industries) as derived in a business register. We now further explain the method by using industries as an example, but it can also be applied to other classifications.

Denote the total set of industries by \(H_{\text{full}}\). Each unit (enterprise) \(i\) has an unknown fixed true industry code \(s_i = g\) and a stochastic observed industry code \(\hat{s}_i = h\), where \(g, h \in H_{\text{full}}\). We suppose that for each unit random classification errors occur, independently across units, according to a known (or previously estimated) transition matrix \(P_i = (p_{ghi})\), with \(p_{ghi} = P(\hat{s}_i = h | s_i = g)\).

An example of \(P_i\) (Fig. 1) shows the labels 45111, ..., 45402 that are target industries for which estimates are produced, and the label “other” represents non-target industries. Rows represent true values and columns observed values with errors.
Second step. To apply the method, we need to estimate $P_t$. An estimate for $P_t$ can be based on an audit sample of units from the observed industries for which experts determine the correct values. In van Delden et al. (2016) the diagonal elements of the transition matrix are estimated by a logistic model, where the exact values depend on the background variables, so that is how the $P_t$ matrix varies with unit $i$. Then, the probabilities given that a unit is misclassified (the off-diagonal elements) are estimated in van Delden et al. (2016) by using an adapted log-linear model, using the audit data and some simplifying assumptions. One assumption is that the relative frequency of industry transitions as found in the business register are related to misclassification errors (natural transitions that occur more often will also more often erroneously applied).

Third step. Let $\theta = f(y_1, ..., y_N, s_1, ..., s_N)$ denote a target parameter. Based on the observed data, this parameter is estimated by $\hat{\theta} = f(y_1, ..., y_N, \hat{s}_1, ..., \hat{s}_N)$. Here, the assumption is used that classification errors are the only errors that occur. We now use the estimated values for $P_t$ to estimate the accuracy of $\hat{\theta}$. First assume that for all units $i$ in the population we know the true industry code $s_i$ (see Fig. 2) and we have observations for the target variable $y_i$. For each unit we draw a value for the observed industry code $\hat{s}_i$ according to $P_t$. Based on the results for this draw, denoted by $r$, we compute $\hat{\theta}_r$. We repeat this procedure $R$ times, thus $r = 1, ..., R$, and use the set of outcomes $\hat{\theta}_r$ to compute the bias and variance of $\hat{\theta}$:

$$B_R(\theta) = m_R(\theta) - \theta,$$  

$$V_R(\theta) = \frac{1}{R - 1} \sum_{r=1}^{R} (\hat{\theta}_r - m_R(\theta))^2.$$  

with $m_R(\theta) = \frac{1}{R} \sum_{r=1}^{R} \hat{\theta}_r$.

Obviously, in practice, we do not know the true industry codes for all units in the population. Instead, a bootstrap approach is used. We start with the set of data that contains the observed industry code $\hat{s}_i$ for all units $i$ in the population (and the observations for target variable $y_i$). We now draw a new industry value $\hat{s}_i^*$ according to $P_t$ and use this to compute $\hat{\theta}_r^*$. Next this procedure is repeated $R$ times, where bias and variance are estimated as (Efron and Tibshirani, 1993):

$$B^*_R(\theta) = m_R(\theta^*) - \theta,$$  

$$V^*_R(\theta) = \frac{1}{R - 1} \sum_{r=1}^{R} (\hat{\theta}_r^* - m_R(\theta)^*)^2.$$
\[ \hat{V}_r(\hat{\theta}) = \frac{1}{R-1} \sum_{r=1}^{R} \left( \hat{\theta}_r^* - m_R(\hat{\theta}^*) \right)^2. \] 

with \( m_R(\hat{\theta}^*) = \frac{1}{R} \sum_{r=1}^{R} \hat{\theta}_r^*. \)

<table>
<thead>
<tr>
<th>Reality</th>
<th>Bootstrap</th>
</tr>
</thead>
<tbody>
<tr>
<td>( s_i )</td>
<td>( \hat{s}_i )</td>
</tr>
<tr>
<td>( p_i )</td>
<td>( \hat{p}_i )</td>
</tr>
</tbody>
</table>

Fig. 2. Applying the transition matrix \( P \) to estimate accuracy.

In van Delden et al. (2016), this bootstrap method is applied to the total turnover per industry: \( \theta = Y_h = \sum_{i=1}^{N} a_{hi} y_{li} \), with

\[ a_{hi} = I(s_i = h) = \begin{cases} 1 & \text{if } s_i = h, \\ 0 & \text{if } s_i \neq h. \end{cases} \]

In practice, \( Y_h \) is estimated by \( \hat{\theta} = \hat{Y}_h = \sum_{i=1}^{N} \hat{a}_{hi} y_{li} \), with \( \hat{a}_{hi} = I(\hat{s}_i = h) \). The bootstrap replicates of this estimate are given by \( \hat{\theta}_r^* = \hat{Y}_{hr}^* = \sum_{i=1}^{N} \hat{a}_{hi}^* y_{li} \), with \( \hat{a}_{hi}^* = I(\hat{s}_i^* = h) \).

**Assumptions**

The method is based on the following assumptions:
- independent classification errors across units;
- the probabilities in the transition matrix \( P \) for unit \( i \) can be modelled as a function of background variables of the units;
- the obtained scores on the categories for the audit data are error-free.

In addition, van Delden et al. (2016) used the following simplifying assumptions in the second step:
- the probability that a unit \( i \) is observed in the correct category can be modelled by a logistic regression;
- the probability distribution that a unit \( i \) is observed in category \( h \) whereas its true code is \( g \), given it is misclassified \( g \neq h \) is independent of unit \( i \) (i.e. does not depend on background variables);
- the probabilities that a unit \( i \) is observed in the correct category is close to 1 for the larger “business” units;
- the number of units that are erroneously observed in the target population equals the number of missed units in the target population (see Fig. 1).

**Advantages**

- The approach is generic: we expect that it can be extended to other error types.
- Mathematically it is a straightforward approach.

**Disadvantages**

- The bootstrap estimates of the accuracy are biased, a correction for this is needed. van Delden et al. (2016) used an analytical solution for this, but that might not be feasible for more complex situations.
- The estimation of the error probabilities requires collection of independent data (e.g. audit data).
- Another disadvantage is that the approach is computationally intensive.

**Case study (per method if needed)**
<table>
<thead>
<tr>
<th>Agency – country</th>
<th>Statistics Netherlands</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Topic</strong></td>
<td>Effect of NACE classification errors in the general business register on accuracy of short-term statistics on turnover level estimates</td>
</tr>
<tr>
<td><strong>Data sets used</strong></td>
<td>Quarterly values of 2012 first quarter – 2014 second quarter for units within the car trade industry. The data concerns value added tax micro data for the smaller and simple enterprises and census survey data for the largest and most complex units. Audit sample of 225 units, for which true NACE codes were obtained.</td>
</tr>
<tr>
<td><strong>Results (e.g. different methods)</strong></td>
<td>The method resulted in plausible estimates of accuracy (bias and variance).</td>
</tr>
</tbody>
</table>

**Final remarks**

- The method so far is only suitable for level estimates. Extensions to estimates of change, such as growth rates, need to be developed.
- It would be advantageous to use currently collected /available data to estimate error size, thereby avoiding a separate audit sample.
- Development of a derived, simplified, indicator that can more easily be understood by employees working in daily production.

**Other remarks**
The method can be extended to other types of errors and also to other configurations.

**References**