Estimation methods for computing a branch's total value added from incomplete annual accounting data

Working Paper Research

by Stijn Vansteelandt, François Coppens, Dries Reynders, Mario Vackier and Laurent Van Belle

April 2019 No 371

NationalBank
OF BELGIUM
Eurosystem

Editor

Pierre Wunsch, Governor of the National Bank of Belgium

Statement of purpose:

The purpose of these working papers is to promote the circulation of research results (Research Series) and analytical studies (Documents Series) made within the National Bank of Belgium or presented by external economists in seminars, conferences and conventions organised by the Bank. The aim is therefore to provide a platform for discussion. The opinions expressed are strictly those of the authors and do not necessarily reflect the views of the National Bank of Belgium.

The Working Papers are available on the website of the Bank: http://www.nbb.be

ISSN: 1375-680X (print)
ISSN: 1784-2476 (online)

Abstract

Timely monitoring of the economic performance of a particular sector is generally hindered by the fact that not all companies have deposited their annual accounts by the time that an evaluation is made. In view of this, we develop several imputation strategies that each enable predicting a company's value added based on available information from past and current years for those companies where the value added was not timely reported.

For each proposed strategy we discuss the assumptions which must be fulfilled for unbiased estimation and calculate the estimation uncertainty. In particular, the proposed imputation procedures all rely on an assumption of missing at random, namely that the values added in companies that did not yet deposit their annual accounts are similar (in some way) to those in companies with the same characteristics (e.g. the same historical data) that did deposit their accounts by the evaluation date. We show how to retrospectively assess the validity of this assumption, and how to adjust the imputation procedure in case the assumption fails.

The importance of the availability of the uncertainty margins should not be underestimated because they will result in faster and higher quality publications.

Finally we retrospectively apply each strategy to data from the Belgian Port sector and compare their performance at several evaluation dates. All the proposed methods show good results on these data. The method using (ordinary least squares) regression is preferred because it is very flexible in the use of auxiliary variables, requires weaker assumptions, has smaller estimation uncertainty and is easily automatable.

Authors:
Stijn Vansteelandt, Ghent university, Belgium and London School of Hygiene and Tropical Medicine, U.K. – e-mail: stijn.vansteelandt@ugent.be
Corresponding author: François Coppens, National Bank of Belgium – e-mail: francois.coppens@nbb.be
Dries Reynders, Ghent University, Belgium – e-mail: dries.reynders@ugent.be
Mario Vackier, National Bank of Belgium – e-mail: mario.vackier@nbb.be
Laurent Van Belle, National Bank of Belgium – e-mail: laurent.vanbelle@nbb.be

The views expressed in this paper are those of the authors and do not necessarily reflect the views of the National Bank of Belgium or any other institution to which the authors are affiliated.

Contents

1 Introduction

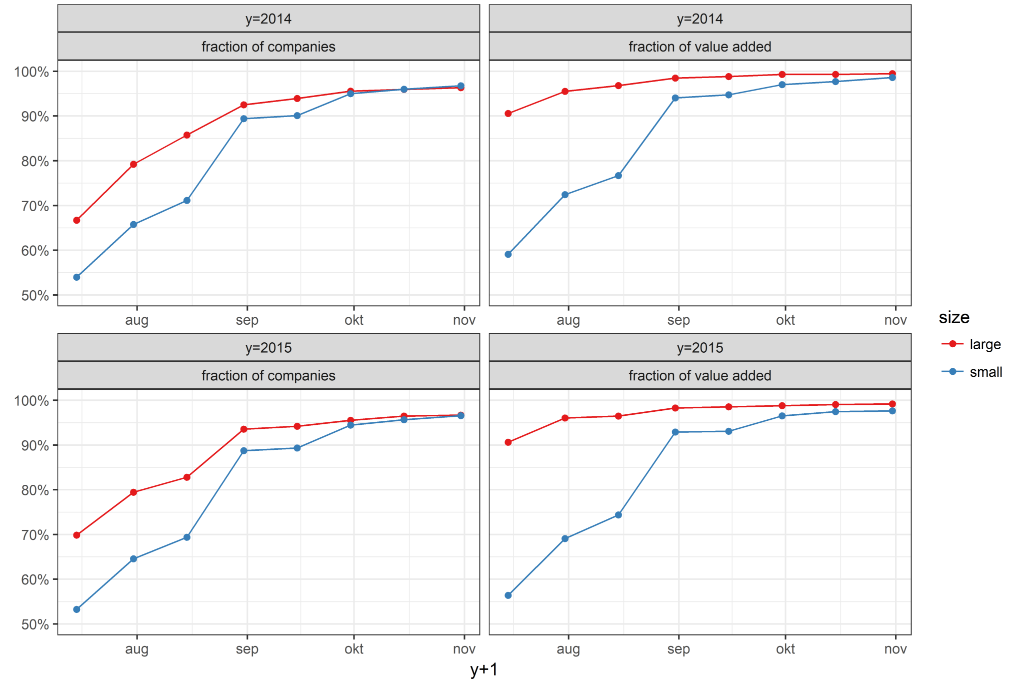
The Central Balance Sheet Office (CBSO) of the National Bank of Belgium (NBB) is responsible for collecting the annual accounts of ±300,000 companies[1]. Around 5% of these annually collected accounts are so-called 'full schemes' and the other are abbreviated and (from the accounting year 2016 on) also micro schemes. The micro schemes contain a subset of the data in the abbreviated schemes, the latter contain a subset of the data that is reported in the full schemes. Companies with the same scheme are similar with respect to information availability. Companies depositing a full scheme are called 'large', the others are categorized as 'small'.

The annual accounts must be filed with the CBSO within thirty days after they have been approved and no later than seven months after the end of the financial year. In practice and for several reasons, the CBSO receives the major part of the accounts for a certain financial year y between July $y + 1$ (notation: $y + 7m$) and August $y + 1$ (notation: $y + 8m$), but it can take until February-March of $y + 2$ (notation: $y + 14m$ - $y + 15m$) before all accounts for year y are (considered to be) complete. As an example, the accounts for the accounting year 2015 start to arrive massively from July-August 2016, but only in February 2017 they are (almost) complete. The arrival rate of the annual accounts for the companies in the Ports[2] is illustrated in figure 1 for the years $y \in \{2014, 2015\}$. The horizontal axis shows the dates u in $y + 1$ at which the number of accounts is counted, the vertical axis is the fraction of the accounts registered at the CBSO at date u for the year y. The fraction is computed in terms of the number of companies and in terms of value added and separately for large companies (i.e. the companies that report a full scheme) and small companies (that report an abbreviated or a micro scheme). At the beginning of August around 65% of the small companies reported their accounts (representing

[1] The legal background for the compilation and submission of annual accounts, consolidated accounts and the social balance sheet with the Central Balance Sheet Office is mainly based on European and Belgian laws and implementing decrees, see https://www.nbb.be/en/central-balance-sheet-office/filing-annual-accounts/legal-background for the details.

[2] see e.g. Coppens et al., 2018 for the definition of the population of the Belgian Ports

Figure 1: Fraction of accounts for year y deposited at the CBSO in $y + 1$ (for the large and small in the Ports study)



around 70% of their value added), while for the large companies this fraction is around 80% (accounting for more than 95% of their value added). In this context, it is important to note that the large companies account for around 97% of the total value added of the Ports. Each year there are around 4 300 companies in the Port studies, of which 1 600 are large.

The annual accounts are an important source for several studies and statistics published by the NBB. These studies analyse a number of economic variables (like value added or employment) for a particular sector (defined as a list of companies with a similar activity). In practice, a sector is a list of company identification numbers. Because of new entries and exits, the list can change from year to year. The economic variables can be computed from information in the (full, abbreviated, micro) annual accounts.

For this reason, most sectoral studies are published with a large time lag, e.g. the study on the Ports sector over the year 2015 (Mathys C., 2017) was published in June 2017. As Port authorities have asked for shorter publication lags, the Bank developed a "flash estimate" of the ports major aggregated variables in October $y + 1$ (NBB Press release, 2016). This "flash estimate" estimates ('imputes') the missing values for the not-yet-received accounts using ad hoc methods. The variable of interest is the sum of the (imputed and observed) value added over all the companies in a year[3]. The currently used, ad hoc, methods are labour-intensive and time consuming, and the ad hoc nature implies that their accuracy and precision remain unclear.

The aim of this paper is to develop accurate, flexible and easily automatable methods for estimating major aggregated variables of a branch (like ports) in the presence of incomplete reporting of the annual accounts, along with an assessment of the uncertainty in the obtained estimate. This uncertainty assessment will serve to decide when and at what level of detail a so called flash estimate or a study can be published. In view of this, we develop several imputation strategies that each enable predicting a company's value added based on available information from past and current years for those companies where the value added was not timely reported.

As such the proposed methods should increase the quality and shorten the publication delay of these sectoral studies.

2 Methodology

2.1 Aim

As explained in the introduction, we aim to infer the total value added for year y ($T(y)$) across a pre-defined population of companies, $i = 1, ..., n$. We will denote this as

$$T(y) \equiv \sum_{i=1}^{n} Y_i(y),$$

where $Y_i(y)$ denotes the value added of company i in the given year y. While straightforward to calculate when each of the considered companies has submitted data on the value added for year

[3] The methods presented can (under similar conditions) be applied to variables other than value added.

y by the time of the assessment, a timely monitoring of the general economy requires evaluation prior to the value added of all companies being available. In other words, at the assessment date u not all the companies have submitted their data for the year y. Let therefore $R_i(y;u)$ indicate 1 if the value added for year y of company i is available at the assessment date u, and 0 otherwise. Then we will estimate the total value added by substituting $Y_i(y)$ for companies whose value added is missing at the time of the assessment, by a prediction $\hat{Y}_i(y;u)$ based on available, actual and historical data X_i (e.g. the value added of previous years, fiscal data, the number of employees in the company, the change in number of employees since the previous year, the production activity of the company, ...) available at the assessment date u.

Note that whenever $R_i(y;u) = 1$ then the value added for company i for year y is known at date u and for such a company it holds that $R_i(y;u)Y_i(y) = Y_i(y)$, while for a company j with $R_j(y;u) = 0$, the value added is not observed and will be estimated by a value $\hat{Y}_j(y;u)$. In this case it also holds that $\hat{Y}_j(y;u) = (1 - R_j(y;u))\hat{Y}_j(y;u)$. Therefore we will estimate the total value added at the assessment date u as $\hat{T}(y;u) \equiv \sum_{i=1}^{n} \left[R_i(y;u)Y_i(y) + (1 - R_i(y;u))\hat{Y}_i(y;u) \right]$. To simplify the notation we will drop the assessment date u and the year y for which we compute the value added and write:

$$\hat{T} \equiv \sum_{i=1}^{n} \left[R_i Y_i + (1 - R_i)\hat{Y}_i \right]$$

In practice u is at least equal to July of year $y + 1$, where y is the year for which we compute the total value added (notation: $u \geq y + 7m$).

In Section 3, we will discuss various strategies $s = 1, 2, \ldots, k$ for calculating $\hat{Y}_i$ (k is the number of strategies). Depending on the chosen strategy (s) we will obtain different values for $\hat{Y}_i^{(s)}$ and therefore different estimators for the total value added $\hat{T}^{(s)}$. For each of these strategies we will moreover assess the underlying assumptions just as well as the accuracy of each of the resulting estimates $\hat{T}^{(s)}$ under the assumption that the data for the different companies

$(Y_i, X_i, R_i), i = 1, ..., n$, are mutually independent. In particular, we will assess under what conditions the resulting estimates are unbiased (i.e., not systematically under- or overestimated) and moreover assess their precision (i.e., how different they can be expected from the target T).

Using the (un)biasedness and precision and the "ease of use" of each of these estimators $\hat{T}^{(s)}$, we will propose our 'preferred' solution.

2.2 Accuracy

To reason about the accuracy of each of these estimators we will focus on the difference between the target value T (when the value added for all the companies is known) and the estimated value $\hat{T}$ based on partially known information at the assessment date u. It should be clear that this difference is only determined by the data from companies that did not yet submit their financial data (i.e. companies i for which $R_i = 0$). Such companies i have $(1 - R_i) = 1$ and contribute a difference of $(\hat{Y}_i - Y_i)$ to $\hat{T} - T$. As such, the total difference is given by:

$$\hat{T} - T = \sum_{i=1}^{n} (1 - R_i)(\hat{Y}_i - Y_i). \qquad (1)$$

The estimated value $\hat{Y}_i$ for such a company will be based on other data (X_i) for that company that is known at the time of the assessment (like e.g. the value added in the previous year, the number of employees at the company, ...). For a company i we also know whether the value added is available or not (i.e. we know R_i). We will therefore study the behaviour of this difference $\hat{T} - T$ conditional on knowing $\{X_i, R_i; \forall i\}$[4]. We condition on this information to express that inaccuracy arises from what we don't know, which concerns the value added Y_i, rather than what we know. In particular, inaccuracy may arise as a result of an individual company's value added Y_i differing from what is predicted, $\hat{Y}_i$. Conditioning on $\{X_i, R_i; \forall i\}$ is especially indicated since we are making an evaluation of a full population of companies, and thus there is no variability related to the sampling of companies. The only sources of variability

[4] ∀ reads "for all".

that must be considered when evaluating the inaccuracy of $\hat{T}$ are due to the fact that the quality of the prediction $\hat{Y}_i$ may be poor when based on limited observations, and the fact that even if the prediction $\hat{Y}_i$ were based on data for all companies, that prediction for company i would still differ from the true value added for that company, due to random noise (i.e. the disturbance term ϵ_i appearing in all imputation models infra).

We will say that $\hat{T}$ is an unbiased estimate of T when the average difference between the estimated ($\hat{T}$) and real (T) total value added, given the values of X_i and R_i (expressed as $...|\{X_i, R_i; \forall i\}$) is zero:

$$E\left(\hat{T} - T|\{X_i, R_i; \forall i\}\right) = 0.$$

In that case, $\hat{T}$ does not systematically under- or overestimate T or the predictions $\hat{T}$ are "on average" equal to the "true" value T.

From equation (1) it follows that this is satisfied when $\hat{Y}_i$ unbiasedly estimates the average value added for companies with missing value added and the same available data X_i, as company i. Indeed by equation (1) the bias equals $E\left(\hat{T} - T|\{X_i, R_i; \forall i\}\right) = \sum_{i=1}^{n}(1 - R_i)E\left(\hat{Y}_i - Y_i|\{X_j, R_j; \forall j\}\right)$, which is zero when

$$E\left(\hat{Y}_i|\{X_j, R_j, \forall j\}, R_i = 0\right) = E\left(Y_i|\{X_j, R_j, \forall j\}, R_i = 0\right) = E\left(Y_i|X_i, R_i = 0\right).$$

Unfortunately, we cannot guarantee unbiased predictions without making both testable and untestable assumptions. The fact that Y_i is missing for all companies with $R_i = 0$ means that none of the available data is directly informative about the conditional mean $E\left(Y_i|X_i, R_i = 0\right)$. We will therefore proceed under the so-called missing at random (MAR) (Rubin, 1976) assumption that $Y_i \perp\!\!\!\perp R_i|X_i$ (meaning that the missingness (R_i) for given X_i does not depend on the value that is missing $(Y_i))^5$, and thus in particular that

$$E\left(Y_i|X_i, R_i = 0\right) = E\left(Y_i|X_i, R_i = 1\right).$$

This assumption states that companies whose value added is not available at the time of the assessment are comparable (in terms of value added) to companies with the same characteristics X_i whose data are available at the time of the assessment (and hence the predictions for companies i with $R_i = 0$ can be based on the analysis of companies j with $R_j = 1$). It can be made more plausible by incorporating more characteristics X_i into the analysis. The missing at random assumption is untestable, i.e. unverifiable from the available data at the time of the assessment. Interestingly, however, in our case it can be assessed retrospectively, once the data on the value added Y_i have come available for all companies (i.e. at $u \geq y + 14m$). We will discuss this in Section 5, as well as how to deal with violations of this assumption[6].

The missing at random assumption is sufficient for inferring T because we can infer the mean $E\left(Y_i|X_i, R_i = 1\right)$ from the available (thus for companies i with $R_i = 1$) data and, by the missing at random assumption, use it as a substitute for the unknown mean $E\left(Y_i|X_i, R_i = 0\right)$. Estimating the mean $E\left(Y_i|X_i, R_i = 1\right)$ will typically necessitate the use of additional (testable) modelling assumptions, however; see Section 3.

The unbiasedness of $\hat{T}$ expresses only one aspect of its (in)accuracy. In addition, we will quantify the imprecision of $\hat{T}$, which expresses how far we can expect $\hat{T}$ to be located from T. It can be assessed as

$$\begin{aligned} E\left\{\left(\hat{T} - T\right)^2 |\{X_i, R_i; \forall i\}\right\} \quad &= \quad \sum_{i=1}^{n}(1 - R_i)E\left\{\left(\hat{Y}_i - Y_i\right)^2 |\{X_i, R_i; \forall i\}\right\} \qquad (2) \\ &+\sum_{i \neq j}(1 - R_i)(1 - R_j)E\left\{\left(\hat{Y}_i - Y_i\right)\left(\hat{Y}_j - Y_j\right) |\{X_i, R_i; \forall i\}\right\}, \end{aligned}$$

[5] Note that for MAR to hold, the missingness may depend on X_i and, as Y_i depends on X_i it could then also depend (but indirectly, through X_i) on Y_i. As long as the dependence of R_i on the value Y_i that is missing is only indirect, via the X_i, MAR is fulfilled.

[6] When the MAR assumption is violated, we say that the missingness is not at random (MNAR).

which, by definition, is expected to shrink as the assessment time comes later (for then $R_i = 1$ for more companies).

The magnitude of the imprecision is driven by 3 sources of error: bias (e.g. due to failure of missing at random or misspecification of the prediction model), the lack of sufficient data to model $E(Y_i|X_i, R_i = 1)$ at the assessment time, as a result of which $\hat{Y}_i$ may differ from $E(Y_i|X_i, R_i = 0)$ (i.e., sampling variability), and the fact that Y_i will generally differ from what is expected, $E(Y_i|X_i, R_i = 0)$.

3 Imputation strategies

In the following sections, we will propose various strategies for calculating imputations $\hat{Y}_i$. We will moreover evaluate under what conditions these imputations are unbiased, and show how to calculate the imprecision of an estimate $\hat{T}$ based on these imputations.

3.1 Proportional imputation

3.1.1 Definition and assumptions of the estimator $\hat{T}^{prop'}$

A first imputation strategy chooses the auxiliary variable X_i, used to predict $\hat{Y}_i$, to be merely the value added of company i from the past year, and proceeds under the assumption that

$$Y_i(y) = \beta X_i + \epsilon_i \quad (\equiv \beta Y_i(y-1) + \epsilon_i) \text{ with } E(\epsilon_i|X_i) = 0. \tag{3}$$

Note that this model implicitly assumes that $Y_i(y-1)$ is known at the assessment date u, which is *not* the case for companies that enter the population between $y-1$ and y. Therefore we will later propose an alternative that takes these 'new' companies into account.

In model (3), β can be estimated based on the available data as

$$\hat{\beta} = \frac{\sum_{i=1}^{n} R_i Y_i}{\sum_{i=1}^{n} R_i X_i}.$$

This is the weighted least squares estimator[7], based on weights equal to $1/X_i$ (Cochran, 1977). It follows that $\hat{\beta}$ is an efficient (i.e., most precise) estimator of β in the model (3) when the variance of the value added Y_i, conditional on X_i, is proportional to X_i (Greene, 2008). This estimated $\hat{\beta}$ can be used to predict missing values as $\hat{Y}_i^{prop'}(y) = \hat{\beta} X_i$, so that

$$\begin{aligned}
\hat{T}^{prop'} &= \sum_{i=1}^{n} \left(R_i Y_i + (1 - R_i) \hat{Y}_i^{prop'} \right) \\
&= \sum_{i=1}^{n} R_i Y_i + \sum_{i=1}^{n} (1 - R_i) \hat{Y}_i^{prop'} \\
&= \hat{\beta} \sum_{i=1}^{n} R_i X_i + \sum_{i=1}^{n} (1 - R_i) \hat{\beta} X_i \\
&= \hat{\beta} \sum_{i=1}^{n} \left(R_i X_i + (1 - R_i) X_i \right).
\end{aligned}$$

It follows that

$$\hat{T}^{prop'} = \hat{\beta} \sum_{i=1}^{n} X_i, \tag{4}$$

where $\hat{\beta}$ is as supra, which is also known as the 'ratio estimator'; the ratio $\hat{\beta}$ is estimated on the available companies for which $Y_i(y)$ and $X_i \equiv Y_i(y-1)$ are known and then applied to the total value added in the previous year for the population ($\sum_{i=1}^{n} X_i$).

3.1.2 (Un)biasedness of the estimator $\hat{T}^{prop'}$

Under assumption (3), that for each company i (including the exiting and entering companies) its value added in year y is, up to a random error, proportional to the one in the previous year, along with the missing at random assumption, the ratio estimator is unbiased because

$$\begin{aligned}
E\left(\hat{Y}_i | \{X_j, R_j, \forall j \neq i\}, X_i, R_i = 0 \right) &= E\left(\hat{\beta} X_i | \{X_j, R_j; \forall j\}, R_i = 0 \right) \\
&= E\left(\frac{\sum_{j=1}^{n} R_j \beta X_j}{\sum_{j=1}^{n} R_j X_j} X_i | \{X_j, R_j; \forall j\}, R_i = 0 \right) \\
&= \beta X_i = E\left(Y_i | X_i, R_i = 0 \right).
\end{aligned}$$

[7] The value for β that minimizes the weighted sum of squares $\sum_i R_i w_i (Y_i - \beta X_i)^2$ is $\hat{\beta} = \frac{\sum_i R_i w_i X_i Y_i}{\sum_i R_i w_i X_i^2}$, where w_i is the weight for observation i.

This unbiasedness is however questionable in practice, because the linear regression through the origin, model (3), is likely misspecified, and moreover, adjustment for merely the value added from the past year renders the missing at random assumption rather implausible.

3.1.3 (Im)precision of the estimator $\hat{T}^{prop'}$

In what follows, we will assess the imprecision of $\hat{T}$ under the assumption that $\hat{T}^{prop'}$ is unbiased. To assess imprecision due to $\hat{Y}_i$ differing from $E\left(Y_i|X_i, R_i = 0\right)$, remember that we are studying a complete population of companies. There is thus no variability related to the sampling of companies, although there is variability related to the fact that $\hat{\beta}$ is only based on data from the subset of companies with $R_i = 1$.

The precision of the estimator $\hat{T}^{prop'}$ can be assessed by substituting $\hat{Y} = \hat{\beta} X_i$ and $Y_i = \beta X_i + \epsilon_i$ in equation (2). It is approximately given by

$$E\left\{\left(\hat{T} - T\right)^2 |\{X_i, R_i; \forall i\}\right\} \approx \sigma^2 \sum_{i=1}^{n} (1 - R_i),$$

where $\sigma^2 = \text{Var}\left(\epsilon_i|X_i\right)$ can be estimated as

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} R_i \left(Y_i - \hat{\beta} X_i\right)^2}{\left(\sum_{i=1}^{n} R_i\right) - 1}. \tag{5}$$

This expression is only approximate as it ignores the imprecision in $\hat{\beta}$ (which is generally small when large numbers of companies are considered). A more accurate result is given by (see annex A.1 for details):

$$\begin{aligned} &E\left\{\left(\hat{T} - T\right)^2 |\{X_i, R_i; \forall i\}\right\} \\ &= \sigma^2 \left[\sum_{i=1}^{n} (1 - R_i) \left\{1 + \left(1 - \frac{X_{(O)}}{X_{(P)}}\right)\right\} + \left(\frac{X_{(P)}}{X_{(O)}} - 1\right)^2 \sum_{i=1}^{n} \left(R_i - \frac{X_{(O)}}{X_{(P)}}\right)^2\right]. \end{aligned} \tag{6}$$

where $\sum_{i=1}^{n} R_i X_i \equiv X_{(O)}$ is the total of X for the companies for which the value added is observed at date u (because in that case $R_i = 1$), while $\sum_{i=1}^{n} X_i \equiv X_{(P)}$ is the total of X for the whole population that will eventually be observed. This expression incorporates finite population corrections to acknowledge that $\hat{\beta}$ has no imprecision when data from all companies are observed.

The above expression relies on the homoscedasticity assumption that $\sigma^2 \equiv \text{Var}\left(\epsilon_i|X_i\right)$ does not depend on X_i, which is quite strong. Its violation is likely, but, as argued in the annex, generally does not impose major concerns, unless when R is strongly dependent on X.

3.2 Proportional imputation, accounting for new companies

3.2.1 Definition and assumptions of the estimator $\hat{T}^{prop}$

In practice, the proportional imputation procedure needs adjustment for the fact that new companies may arise, and that others may have left the population. For the companies $i = 1, 2, \ldots, n$ belonging to the population in year y, let $S_i = 1$ if that company already existed in the previous year, and 0 otherwise. Let $R_i = 1$ be defined as before. The proportional imputation procedure may then be revised by redefining model (3) as[8]

$$Y_i = \beta X_i + \epsilon_i, \tag{7}$$

for all the companies that exist in year y, i.e. $S_i = 1$, and moreover assuming that for the new companies (for which $S_i = 0$)

$$Y_i = Z_i + \nu_i, \tag{8}$$

with Z_i a proxy for value added in the current year (Y_i) that can be derived from the fiscal data of company i[9] and ν_i independent of Z_i (conditional on $S_i = 0$). It then follows that:

$$Y_i = S_i(\beta X_i + \epsilon_i) + (1 - S_i)(Z_i + \nu_i) \tag{9}$$

In the proportional imputation we choose $X_i = Y_i(y-1)$ so equation (7) expresses that for companies that should deposit in y and $y-1$, proportionality between $Y_i(y)$ and $Y_i(y-1)$

[8] There are other alternatives to adjust the ratio estimator to account for new companies, but we describe the method that is currently used and thus serves as our benchmark.

[9] Z_i derived from fiscal data is an (imprecise) proxy for the value added Y_i. This holds for all companies. As it is imprecise, it is only used where no accounting data is available, i.e. for the new companies that did not deposit their accounts at the assessment date.

holds. Equation (8) states that for new companies the value added can just be copied from fiscal data. The latter implicitly assumes that the fiscal value added is an unbiased estimator for the expected value added (computed from balance sheet data), although our later estimates of the imprecision in $\hat{T}$ will acknowledge imprecision that may result from bias. Equation (9) expresses that, for new companies we use a proxy derived from fiscal data, while for companies that already existed last year, we use the value added from accounting data.

With this choice, the imputation estimator $\hat{T}^{prop}$ equals

$$\hat{T}^{prop} = \sum_{i=1}^{n} R_i Y_i + (1 - R_i) \left(S_i X_i \frac{\sum_{i=1}^{n} R_i Y_i}{\sum_{i=1}^{n} R_i X_i} + (1 - S_i) Z_i \right),$$

which is also known as the '*corrected* ratio estimator'.

Note that R_i equals one when the value Y_i is observed and zero otherwise, so $\sum_{i=1}^{n} R_i Y_i$ is the sum of all observed Y_i[10]. Further $\sum_{i=1}^{n}(1 - R_i) S_i X_i \frac{\sum_{i=1}^{n} R_i Y_i}{\sum_{i=1}^{n} R_i X_i}$ is the ratio estimator applied to companies that exist in y and $y - 1$ ($S_i = 1$ if i exists in $y - 1$) but for which $Y_i(y)$ is not observed ($R_i = 0$) and $\sum_{i=1}^{n}(1 - R_i)(1 - S_i) Z_i$ is the sum of fiscal value added for those companies that are new ($S_i = 0$) and for which Y_i is not observed ($R_i = 0$).

3.2.2 (Un)biasedness of the estimator of the estimator $\hat{T}^{prop}$

It is trivially seen that this estimator is unbiased under assumptions (7) and (8) when MAR holds and ϵ_i and μ_i have mean zero.

3.2.3 (Im)precision of the estimator of the estimator $\hat{T}^{prop}$

To assess the imprecision of $\hat{T}^{prop}$ let us write $Y_i = \beta X_i + \epsilon_i$ for companies with $S_i = 1$, as before, and $Y_i = Z_i + \nu_i$ for companies with $S_i = 0$. Then the imprecision equals (we used the

[10] This includes the value added of new companies (i.e. with $S_i = 0$) for which $R_i = 1$, i.e. which already reported their data.

result in equation (6))

$$
\begin{aligned}
& E\left(\left[\sum_{i=1}^{n}(1-R_i)\left\{S_iX_i(\beta-\hat{\beta})+S_i\epsilon_i+(1-S_i)\nu_i\right\}\right]^2 |\{S_iX_i, Z_i, R_i, S_i; \forall i\}\right) \\
= \; & \sigma^2\left[\sum_{i=1}^{n}(1-R_i)S_i\left\{1+\left(1-\frac{X_{(O)}}{X_{(P)}}\right)\right\}+\left(\frac{X_{(P)}}{X_{(O)}}-1\right)^2\sum_{i=1}^{n}S_i\left(R_i-\frac{X_{(O)}}{X_{(P)}}\right)^2\right] \\
& +\sum_{i=1}^{n}(1-R_i)(1-S_i)E\left(\nu_i^2|Z_i\right), \qquad (10)
\end{aligned}
$$

where $X_{(O)}$ and $X_{(P)}$ are defined as before, but restricted to companies with $S_i = 1$. Here, σ^2 can be estimated as before, but now restricted to companies with $S_i = 1$, i.e.

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} R_iS_i\left(Y_i-\hat{\beta}X_i\right)^2}{\left(\sum_{i=1}^{n} R_iS_i\right)-1}. \qquad (11)$$

Further, $E\left(\nu_i^2|Z_i\right)$ can be unbiasedly estimated as

$$\hat{\sigma}_\nu^2 = \frac{\sum_{i=1}^{n} R_i(Y_i-Z_i)^2}{\sum_{i=1}^{n} R_i} \qquad (12)$$

when $E\left(\nu_i^2|Z_i\right)$ does not depend on Z_i and, moreover, $Y_i - Z_i$ is equally distributed for companies with $R_i = 1$ and companies with $R_i = S_i = 0$. Note that this estimate of $E\left(\nu_i^2|Z_i\right)$ expresses both imprecision due to the fiscal data being a biased assessment of the value added, as well as due to random error. Note also that, while violation of assumption (8) may introduce a bias in the estimates, this bias is taken into account in the above formula for the imprecision.

As an alternative we also assessed the imprecision using a semiparametric bootstrap procedure to compute 95% percentile prediction intervals as follows (see Efron and Tibshirani, 1994):

1. Draw, with replacement, a bootstrap sample $X_{i(k)}, Y_{i(k)}$ with size $\sum_i R_iS_i$ from the companies i that have $R_iS_i = 1$;

2. Draw $\sum_i(1-R_i)S_i$ outcomes $\epsilon_{m(k)}$ randomly from a normal distribution with zero mean and (estimated) variance given by equation (11);

3. Draw $\sum_i (1-R_i)(1-S_i)$ outcomes $\nu_{l(k)}$ randomly from a normal distribution with zero mean and (estimated) variance given by equation (12);
4. Calculate $\hat{T}_{(k)} = \sum_{i=1}^n \left[R_i Y_i + (1-R_i)\left(S_i X_i \frac{\sum_{i=1}^n Y_{i(k)}}{\sum_{i=1}^n X_{i(k)}} + (1-S_i) Z_i \right) \right] + \sum_m \epsilon_{m(k)} + \sum_l \nu_{l(k)}$;
5. Repeat this for $k = 1, 2, \ldots, K = 2500$ to find bootstrap replicates $\hat{T}_{(1)}, \hat{T}_{(2)}, \ldots, \hat{T}_{(2500)}$ and obtain a 95% prediction interval based on the 2.5% and 97.5% percentiles of these replicates. If one wishes to avoid reliance on normal distributions then $\epsilon_{m(k)}$ and $\nu_{l(k)}$ may alternatively be drawn from the observed residuals from models (7) and (8).

   Note finally that that the above bootstrap procedure, as well as those that will follow later, ignore finite population corrections.

3.3 Ordinary least squares imputation

3.3.1 Definition and assumptions of the estimator $\hat{T}^{ols}$

A second, preferable, imputation strategy chooses $\boldsymbol{X}_i$ to be a rich collection of company characteristics[11], including 1 to allow for an intercept. It proceeds under the assumption that

$$Y_i = \boldsymbol{\beta}' \boldsymbol{X}_i + \epsilon_i, \tag{13}$$

where $\boldsymbol{\beta}$ is estimated using ordinary least squares estimation (and both $\boldsymbol{\beta}$ and $\boldsymbol{X}_i$ are $p \times 1$ vectors[12] [13], for a given dimension p, and R_i and Y_i are scalars), so

$$\hat{\boldsymbol{\beta}} = \overbrace{\left(\sum_{i=1}^n R_i \boldsymbol{X}_i \boldsymbol{X}_i' \right)^{-1}}^{p \times p} \overbrace{\sum_{i=1}^n R_i \boldsymbol{X}_i Y_i}^{p \times 1}$$

Ex post, when all companies have deposited their accounts, we know the whole population and thus also the population value for $\boldsymbol{\beta}$. In practice however, we are in a situation where some

[11] Proportional imputation is a special case where there is only one characteristic X_i being the value added in the previous year.

[12] We use boldface notation to distinguish vectors from scalars.

[13] $\boldsymbol{\beta}'$ denotes the transpose of $\boldsymbol{\beta}$.

of the companies did not yet deposit their accounts. In those cases the population parameter $\boldsymbol{\beta}$ is unknown and has to be estimated from available data.

3.3.2 (Un)biasedness of the estimator $\hat{T}^{ols}$

It follows by the properties of OLS estimators that, under MAR, this estimation algorithm guarantees unbiasedness under model (13), resulting in an imputation estimator $\hat{T}$ equal to

$$\hat{T}^{ols} = \sum_{i=1}^{n} R_i Y_i + (1 - R_i)\hat{\boldsymbol{\beta}}' \boldsymbol{X}_i.$$

3.3.3 (Im)precision of the estimator $\hat{T}^{ols}$

The imprecision of the estimator $\hat{T}^{ols}$ is approximately given by

$$\sigma^2 \sum_{i=1}^{n} (1 - R_i), \tag{14}$$

where $\sigma^2 = \mathrm{Var}\,(\epsilon_i | X_i)$ can be estimated as the residual variance

$$\frac{\sum_{i=1}^{n} R_i (Y_i - \hat{\boldsymbol{\beta}}' \boldsymbol{X}_i)^2}{(\sum_{i=1}^{n} R_i) - p}, \tag{15}$$

with p the dimension of $\boldsymbol{\beta}$. This expression is only approximate as it ignores the imprecision in $\hat{\beta}$ (which is generally small when large numbers of companies are considered). A more accurate result, which involves finite population corrections, is given by (see annex A.2 for details):

$$\begin{aligned} &\sigma^2 \left[\sum_{i=1}^{n} (1 - R_i) \left\{ 1 + \boldsymbol{X}_i' \boldsymbol{A}^{-1} \boldsymbol{r} \left\{ \sum_{i=1}^{n} (1 - R_i) \boldsymbol{X}_i \right\} \right\} \right] \\ &+ \sigma^2 \left\{ \sum_{i=1}^{n} (1 - R_i) \boldsymbol{X}_i \right\}' \boldsymbol{A}^{-1} \left(\sum_{j=1}^{n} \boldsymbol{B}_j \boldsymbol{X}_j \boldsymbol{X}_j' \boldsymbol{B}_j' \right) \boldsymbol{A}^{-1} \left\{ \sum_{i=1}^{n} (1 - R_i) \boldsymbol{X}_i \right\}. \end{aligned} \tag{16}$$

where $\boldsymbol{r} = (\sum_{i=1}^{n} R_i \boldsymbol{X}_i \boldsymbol{X}_i')\,(\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i')^{-1}$, $\boldsymbol{A} = \sum_{i=1}^{n} R_i \boldsymbol{X}_i \boldsymbol{X}_i'$ and $\boldsymbol{B}_i = (R_i \boldsymbol{I}_{p \times p} - \boldsymbol{r})$.

The homoscedasticity assumption is quite strong. Its violation is likely, but, as before, generally does not impose major concerns since interest lies in the 'total' residual variance across companies with incomplete data, rather than the variance of individual records.

We also assessed the imprecision using a semiparametric bootstrap procedure to compute 95% percentile prediction intervals as follows (see Efron and Tibshirani, 1994) in a similar way as for the estimator $\hat{T}^{prop}$ at the end of section 3.2.3:

1. Draw, with replacement, a bootstrap sample $\boldsymbol{X}_{i(k)}, Y_{i(k)}$ with size $\sum_i R_i$ from the companies i that have $R_i = 1$ and estimate $\hat{\boldsymbol{\beta}}_{(k)}$ from that sample;

2. Draw $\sum_i (1 - R_i)$ outcomes $\epsilon_{l(k)}$ randomly from a normal distribution with zero mean and (estimated) variance given by equation (15);

3. Calculate $\hat{T}_{(k)} = \sum_i \left[R_i Y_i + \sum_i (1 - R_i) \hat{\boldsymbol{\beta}}'_{(k)} \boldsymbol{X}_i \right] + \sum_l \epsilon_{l(k)}$;

4. Repeat this for $k = 1, 2, \ldots, K = 2500$ to find bootstrap replicates $\hat{T}_{(1)}, \hat{T}_{(2)}, \ldots, \hat{T}_{(2500)}$ and use these to compute 95% percentile prediction intervals. As before $\epsilon_{l(k)}$ may be drawn from the empirical distribution of the residuals from model (13) if one wishes to avoid normality assumptions.

3.4 Advantages of ordinary least squares imputation over proportional imputation

The ordinary least squares imputation strategy has a number of major advantages over proportional imputation.

First, it can easily accommodate imputation of new companies; this can be done by letting $\boldsymbol{X}_i$ include S_i. Remember that $\boldsymbol{X}_i$ is a vector of p auxiliary variables $\boldsymbol{X}_i$ for company i. We could for instance let X_{1i} be the value added for company i of the previous year, X_{2i} be the fiscal data for the current year, X_{3i} be the indicator S_i that is one for a company that existed in the previous year and let X_{4i} be 1 for the intercept in the model. Then one may choose imputation models with e.g. $\boldsymbol{X}_i = [1 \; X_{1i} S_i \; X_{2i} \; S_i]'$, i.e.

$$Y_i = \beta_0 + \beta_1 X_{1i} S_i + \beta_2 X_{2i} + \beta_3 S_i + \epsilon_i.$$

This model assumes that the mean value added for new companies is $\beta_0 + \beta_2 X_{2i}$, while for existing companies it is $\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3$.

Second, in the same way one can easily include additional auxiliary variables that strongly correlate with Y_i like e.g. employment or salaries.

Third, the missing at random assumption in model (13) is much less strong than in (3) as a result of adjusting for a larger collection of variables $\boldsymbol{X}_i$. Adjusting for more variables $\boldsymbol{X}_i$ in the ordinary least squares regression approach helps to 'explain' missingness (see also footnote [5]), thereby rendering missing at random more plausible.

Fourth, model (3), unlike model (13) (e.g. with $\boldsymbol{X}_i = [1 \ X_{1i} S_i \ X_{2i} \ S_i]'$ as supra), excludes the possibility of a systematic growth, for instance, by assuming that companies with small (negative) value added in the past year have small (negative) value added (on average) in the current year.

Fifth, adjusting for more company characteristics results in more accurate predictions $\hat{Y}_i$. This is mainly expressed in the estimate of the residual variance (15) typically being much smaller than the corresponding estimate (5). This is quite important if one considers that the key component of the imprecision of $\hat{T}$ equals $\sum_{i=1}^{n}(1 - R_i)\text{Var}\,(\epsilon_i|\boldsymbol{X}_i)$.

By the same token, outlying outcome measurements are less likely in the ordinary least squares imputation approach as they are more likely explained by additional predictors.

3.5 Linear mixed model imputation

3.5.1 Definition and assumptions of the estimator $\hat{T}^{mix}$

The imputation strategy of the previous section does not immediately lend itself to the modelling of many branches, in view of their high-dimensionality. One may remedy this by using linear mixed models. In particular, let $\boldsymbol{X}_i = (\boldsymbol{W}_i', U_i, Z_i)'$ be a vector with company characteristics $\boldsymbol{W}_i$ (e.g. fiscal data, the number of employees in the company, the change in number of employees since the previous year, the production activity of the company, ...), U_i is the value added of previous year (and is also included in the fixed effects vector $\boldsymbol{W}_i$) and Z_i the company's branch code. Further, let $I(Z_i = j), j = 1, 2, \ldots, c$ be an indicator function which takes the value 1 when the ith company's branch code equals j, and 0 otherwise. Then one may proceed under the assumption that[14]

$$Y_i = \boldsymbol{\beta}'\boldsymbol{W}_i + \sum_j b_{0j} I(Z_i = j) + \sum_j b_{1j} U_i \times I(Z_i = j), \tag{17}$$

where b_{ij} is assumed to be a mean zero, normal variate with variance $\sigma^2_{b_i}$ for $i = 0, 1$; here, the summation runs over all possible branch codes. The assumption that all coefficients b_{ij} originate from a mean zero normal distribution with variance $\sigma^2_{b_i}$ ensures regularisation of the corresponding coefficient estimates. In particular, it prevents that the lack of information gives rise to highly variable estimates, and therefore instability. It moreover ensures that the coefficients of the $c-1$ dummies are essentially replaced by only one parameter $\sigma^2_{b_i}$ to estimate; doing so we accommodate the loss of degrees of freedom. This then results in an imputation estimator $\hat{T}^{mix}$ equal to

$$\sum_{i=1}^{n} R_i Y_i + (1 - R_i) \left\{ \hat{\boldsymbol{\beta}}'\boldsymbol{W}_i + \sum_j \hat{b}_{0j} I(Z_i = j) + \sum_j \hat{b}_{1j} U_i \times I(Z_i = j) \right\},$$

where $\hat{\boldsymbol{\beta}}$ denotes the so-called restricted maximum likelihood (REML) estimator of $\boldsymbol{\beta}$ under model (17), and $\hat{b}_{ij}$ represents the so-called empirical best linear unbiased predictor (empirical BLUP) corresponding to b_{ij}.

3.5.2 (Un)biasedness of the estimator $\hat{T}^{mix}$

Unlike the estimators of the previous sections, this estimator of T will typically have some bias as a result of shrinkage in the empirical BLUPs that results in some attenuation (i.e. in $\hat{b}_{ij}$ being closer to zero, on average, than b_{ij}), which diminishes as the number of companies per branch code grows.

[14] This model implies that the intercept and the coefficient of U_i (value added in the previous year) may depend on the branch of activity of the company. The coefficients of the other explanatory variables included in $\boldsymbol{W}_i$ are assumed not to depend on the branch code.

3.5.3 (Im)precision of the estimator $\hat{T}^{mix}$

The aforementioned bias, as well as the fact that empirical BLUPs tend to follow a complex, non-standard distribution, makes analytical expressions of the imprecision of $\hat{T}$ difficult to obtain; while analytical approximations are possible, these approximations are known to be poor in practice. In view of this, one may make use of the parametric bootstrap to assess the imprecision of $\hat{T}$. In particular, writing

$$Y_i = \boldsymbol{\beta}'\boldsymbol{W}_i + \sum_j b_{0j} I(Z_i = j) + \sum_j b_{1j} U_i \times I(Z_i = j) + \epsilon_i,$$

note that the imprecision is given by the conditional expectation of

$$\left[\sum_{i=1}^{n}(1 - R_i)\left\{(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\boldsymbol{W}_i + \sum_j (b_{0j} - \hat{b}_{0j}) I(Z_i = j) + \sum_j (b_{1j} - \hat{b}_{1j}) U_i \times I(Z_i = j) + \epsilon_i\right\}\right]^2,$$

given $\{\boldsymbol{X}_i, R_i; \forall i\}$. In the expression for

$$(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\boldsymbol{W}_i + \sum_j (b_{0j} - \hat{b}_{0j}) I(Z_i = j) + \sum_j (b_{1j} - \hat{b}_{1j}) U_i \times I(Z_i = j),$$

we will use $\hat{\boldsymbol{\beta}}$ and $\hat{b}_{ij}$ as substitutes for $\boldsymbol{\beta}$ and b_{ij}, respectively. Next, we will repeatedly (for $k = 1, ..., K$ for some pre-specified number, e.g. $K = 2500$) simulate new observations Y_i^* for the value added of each company i with $R_i = 1$ by drawing normal variates with mean $\hat{\boldsymbol{\beta}}'\boldsymbol{W}_i + \sum_j b_{0j}^* I(Z_i = j) + \sum_j b_{1j}^* U_i \times I(Z_i = j)$ and variance $\hat{\sigma}^2$, where b_{ij}^* is a random draw from a mean zero normal distribution with variance $\hat{\sigma}_{b_i}^2$, the restricted maximum likelihood estimate of $\sigma_{b_i}^2$. We next analyse each such simulated dataset $k = 1, ..., K$, in the same way as the observed data to arrive at estimates $\hat{\boldsymbol{\beta}}_{(k)}^*$ and $\hat{b}_{ij(k)}^*$. For each company i, we further take a draw ϵ_i^* from the distribution of ϵ_i. This could be a random draw from a mean zero normal distribution with variance σ^2, which can be consistently estimated as $\hat{\sigma}^2$ using restricted maximum likelihood procedures under the assumption of constant residual variance. This could alternatively be the estimated residual $Y_j - \hat{\boldsymbol{\beta}}'\boldsymbol{W}_i + \sum_j \hat{b}_{0j} I(Z_i = j) + \sum_j \hat{b}_{1j} U_i \times I(Z_i = j)$ for a randomly selected company. This allows us to estimate the imprecision as

$$\frac{1}{K}\sum_{k=1}^{K}\left[\sum_{i=1}^{n}(1-R_i)\left\{(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_k^*)'\boldsymbol{W}_i+\sum_j(\hat{b}_{0j}-\hat{b}_{0j(k)}^*)I(Z_i=j)\right.\right.$$
$$\left.\left.+\sum_j(\hat{b}_{1j}-\hat{b}_{1j(k)}^*)U_i\times I(Z_i=j)+\epsilon_i^*\right\}\right]^2.$$

Drawing ϵ_i^* from the estimated residuals has the advantage of not assuming normality, but the drawback that it relies on biased estimates $\hat{b}_{ij}$ of the random effects and that estimated residuals tend to be attenuated towards zero.

The above procedure is likely to lead to slight overestimation of the imprecision of $\hat{T}^{mix}$ by not involving finite-population corrections on the distribution of $\hat{\boldsymbol{\beta}}$ and $\hat{b}_{ij}$. This is not a major concern however, since for the methods where an analytical formula for the precision could be derived, the "finite population correction" shows to be relatively small (as can be seen in annex B.)

Further, note that we deliberately make use of the parametric bootstrap rather than the more common nonparametric bootstrap. The reason is that the nonparametric bootstrap does not enable us to condition on the observed data on $\boldsymbol{X}_i$; this is especially problematic as certain branch codes might otherwise not appear in certain resamples, thereby yielding no estimates of the corresponding coefficient b_{ij}.

4 Robustness against model extrapolation

A drawback of the imputation strategies of the previous section is that they all imply a risk of extrapolation, which may occur when the companies who did versus did not submit their data are rather different in terms of the observed variables $\boldsymbol{X}_i$. Although the imputation models correct for such differences, correct model specification can be quite crucial when the differences are large, for then even mild misspecifications over the observed data range can induce large biases. In view of this, we will make an 'in principle' preferable (though more complicated) proposal below, which explicates the uncertainty due to extrapolation and has the added advantage of generalising rather straightforwardly to missing not at random data.

4.1 Definition and assumptions of the weighted imputation estimator $\hat{T}^{wls}$

To lessen the risk of model extrapolation, we will avoid sole reliance on imputation models. We will do this by considering instead a model which describes how likely a company with characteristics $\boldsymbol{X}_i$ has submitted its financial data at the time of the assessment. For instance, we may postulate that

$$\text{logit}\left(P(R_i = 1|\boldsymbol{X}_i)\right) = \boldsymbol{\gamma}'\boldsymbol{X}_i, \tag{18}$$

which can be fitted using logistic regression; $\hat{P}(R_i = 1|\boldsymbol{X}_i)$ is then obtained as the fitted value from this logistic model, i.e. $\hat{P}(R_i = 1|\boldsymbol{X}_i) = \frac{e^{\hat{\gamma}'\boldsymbol{X}_i}}{1+e^{\hat{\gamma}'\boldsymbol{X}_i}}$, where $\hat{\boldsymbol{\gamma}}$ is the maximum likelihood estimator of γ.

This model merely quantifies what percentage of companies with data $\boldsymbol{X}_i$ has submitted its financial data; it thereby avoids the possible extrapolation in imputation models (which model the outcome in companies with $R_i = 1$ and extrapolate to companies with $R_i = 0$). With this logistic model, we then propose fitting the imputation model (13) to companies with observed outcome data using ordinary least squares regression, weighting the ith $(i = 1, ..., n)$ company's data by weight

$$w_i = \frac{\hat{P}(R_i = 0|\boldsymbol{X}_i)}{\hat{P}(R_i = 1|\boldsymbol{X}_i)} = e^{-\hat{\gamma}'\boldsymbol{X}_i},$$

resulting in estimates $\hat{\boldsymbol{\beta}}^{(w)}$ for β and in an imputation estimator

$$\hat{T}^{wls} = \sum_{i=1}^{n} R_i Y_i + (1 - R_i)\hat{\boldsymbol{\beta}}^{(w)'}\boldsymbol{X}_i.$$

4.2 (Un)biasedness of the estimator $\hat{T}^{wls}$

The use of such weights does not harm the unbiasedness of the imputations when the imputation model is correctly specified, even when the logistic model for the weights is misspecified.

However, it provides additional assurance in the following sense. At an intuitive level, these weights are large at covariate levels $\boldsymbol{X}_i$ where there is a lot of 'missing data' (i.e. $P(R_i = 1|\boldsymbol{X}_i)$ is closer to 0). They thus ensure that the imputation model fits well at covariate values $\boldsymbol{X}_i$ of companies who did not yet submit their financial data; they thus target a good fit in the region of the covariate space where predictions will be made (Vansteelandt, Carpenter and Kenward, 2010). More formally, when the imputation model is misspecified so that biased imputations are obtained, the imputation estimator $\hat{T}^{wls}$ remains unbiased (in large samples) provided that the logistic model for the weights is correctly specified (Seaman and Vansteelandt, 2018). Indeed, in that case, the imputation estimator can be rewritten as

$$\begin{aligned}
\hat{T}^{wls} &= \sum_{i=1}^{n} R_i Y_i + (1 - R_i)\hat{Y}_i \\
&= \sum_{i=1}^{n} \left(R_i Y_i + (1 - R_i)\hat{Y}_i + R_i \frac{\hat{P}(R_i = 0|\boldsymbol{X}_i)}{\hat{P}(R_i = 1|\boldsymbol{X}_i)} \left(Y_i - \hat{Y}_i \right) \right) \\
&= \sum_{i=1}^{n} \left(R_i Y_i + (1 - R_i)\hat{Y}_i + R_i \left\{ \frac{1 - \hat{P}(R_i = 1|\boldsymbol{X}_i)}{\hat{P}(R_i = 1|\boldsymbol{X}_i)} \right\} \left(Y_i - \hat{Y}_i \right) \right) \\
&= \sum_{i=1}^{n} \left(\frac{R_i}{\hat{P}(R_i = 1|\boldsymbol{X}_i)} Y_i + \left\{ 1 - \frac{R_i}{\hat{P}(R_i = 1|\boldsymbol{X}_i)} \right\} \hat{Y}_i \right). \qquad (19)
\end{aligned}$$

Here, the second equality holds because the weighted least squares predictions $\hat{Y}_i$ satisfy:

$$\sum_{i=1}^{n} R_i \frac{\hat{P}(R_i = 0|\boldsymbol{X}_i)}{\hat{P}(R_i = 1|\boldsymbol{X}_i)} \left(Y_i - \hat{Y}_i \right) = 0$$

The unbiasedness now follows because, under MAR, R_i has mean equal to $P(R_i = 1|\boldsymbol{X}_i)$, conditional on $\boldsymbol{X}_i$ and Y_i, so that the first term in (19) averages to $E(Y_i)$ and the second term to 0, when the model for $P(R_i = 1|\boldsymbol{X}_i)$ is correctly specified. Since the resulting method thus gives (approximately) unbiased results when either the imputation model (13) or the missingness model (18) is correctly specified, it is called double robust.

4.3 Stability of the estimator $\hat{T}^{wls}$

The parameter $\boldsymbol{\gamma}$ indexing model (18) can be estimated using maximum likelihood for logistic regression yielding weights $\hat{w}_i = e^{-\hat{\boldsymbol{\gamma}}'\boldsymbol{X}_i}$. However, this is likely leading to instability as a result of $P(R_i = 1|\boldsymbol{X}_i)$ converging to 1 as the assessment time comes later because 'in the limit' all values Y_i are observed and therefore $R_i = 1$. In that case, especially the intercept in model (18) can be expected to diverge. The intercept is nonetheless irrelevant for the procedure of the previous section. Indeed, writing $\boldsymbol{\gamma}'\boldsymbol{X}_i = \gamma_0 + \boldsymbol{\gamma}_1'\boldsymbol{X}_{-1,i}$ (where $\boldsymbol{X}_{-1,i}$ is the vector $\boldsymbol{X}_i$ excluding the component for the intercept), note that the weights equal $e^{-\hat{\boldsymbol{\gamma}}'\boldsymbol{X}_i} = e^{-\hat{\gamma}_0}e^{-\hat{\boldsymbol{\gamma}}_1'\boldsymbol{X}_{-1,i}}$, where the factor $e^{-\hat{\gamma}_0}$ is constant and hence can be ignored in the maximisation of the weighted sum of squared residuals. We will therefore design a novel estimation procedure which does not require estimation of the intercept γ_0.

In particular, we propose estimating $\boldsymbol{\gamma}_1$ as the solution to the system of unbiased equations

$$\frac{\sum_{i=1}^n (1-R_i)\boldsymbol{X}_{-1,i}}{\sum_{i=1}^n (1-R_i)} = \frac{\sum_{i=1}^n R_i e^{-\boldsymbol{\gamma}_1'\boldsymbol{X}_{-1,i}}\boldsymbol{X}_{-1,i}}{\sum_{i=1}^n R_i e^{-\boldsymbol{\gamma}_1'\boldsymbol{X}_{-1,i}}}.$$

Under MAR, the left hand side (LHS) is the mean of the covariates in the companies with missing data and the right hand side (RHS) is the weighted mean of the covariates in the companies with observed data, which is equal when MAR can be assumed[15] . Both sides of the equation are thus asymptotically unbiased estimators of $E(\boldsymbol{X}_{-1,i}|R_i = 0)$, as a result of which the solution to this equation is a consistent estimator of $\boldsymbol{\gamma}_1$ (Newey and McFadden, 1994). In cases where the equation is difficult to solve, we recommend minimising the least squares distance between both sides of the equation. In particular, we recommend minimising

$$\sum_{j=2}^{p}\left(\frac{\sum_{i=1}^n (1-R_i)\boldsymbol{X}_{ji}}{\sum_{i=1}^n 1-R_i} - \frac{\sum_{i=1}^n R_i e^{-\boldsymbol{\gamma}_1'\boldsymbol{X}_{-1,i}}\boldsymbol{X}_{ji}}{\sum_{i=1}^n R_i e^{-\boldsymbol{\gamma}_1'\boldsymbol{X}_{-1,i}}}\right)^2,$$

where $\boldsymbol{X}_{ji}$ is the jth element of $\boldsymbol{X}_i$. The solution to this equation guarantees more stable weights $e^{-\boldsymbol{\gamma}_1'\boldsymbol{X}_{ji}}$. Indeed, applying these weights to the companies that submitted their data

[15] This follows from the fact that MAR implies mean independence ($E(Y_i|\boldsymbol{X}_i, R_i = 1) = E(Y_i|\boldsymbol{X}_i, R_i = 0)$ and from our model assumption ($Y_i = \boldsymbol{\beta}'\boldsymbol{X}_i + \epsilon_i$)

ensures that their (weighted) covariate means match those of companies that did not yet submit their financial data.

Although the above procedure is designed to return stable weights, one may sometimes still observe the weights for certain companies to be large. We have observed this often to be the result of outlying predictor values $\boldsymbol{X}_i$ in companies with $R_i = 1$. For instance, when companies with $R_i = 0$ had added values below a certain threshold x in the previous year, and some companies with $R_i = 1$ had added values above x, then much greater stability can be achieved by defining $P(R_i = 1|\boldsymbol{X}_i) = 1$ for companies with added value in the past year above x, and assuming that $\text{logit}\,(P(R_i = 1|\boldsymbol{X}_i)) = \boldsymbol{\gamma}'\boldsymbol{X}_i$ in the remaining companies. This enables restricting the above minimisation procedures to all companies whose added value in the past year was below x. This is valid, since the weights are then zero for companies whose added value in the past year was above x so that they can effectively be eliminated from the imputation procedure. If the remaining weights continue to be unstable, one may consider truncating them at the 99% percentile (see Cole and Hernan, 2008).

4.4 (Im)precision of the estimator $\hat{T}^{wls}$

When the imputation model (13) is correctly specified, then, in large samples from an infinite population, the uncertainty due to the estimation of $\boldsymbol{\gamma}$ can be ignored when evaluating the imprecision of $\hat{T}$, because inconsistent estimation of $\boldsymbol{\gamma}$ then does not affect the consistency of $\hat{T}^{wls}$ (Newey and McFadden, 1994; Theorem 6.1). The resulting approximation can be expected to be even smaller in finite populations, and will therefore be ignored. In that case, the imprecision of $\hat{T}^{wls}$ can be assessed along similar lines as before. In particular,

$$\begin{aligned} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} &= \left(\sum_{i=1}^{n} R_i e^{-\gamma_1' \boldsymbol{X}_{-1,i}} \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1} \sum_{i=1}^{n} R_i e^{-\gamma_1' \boldsymbol{X}_{-1,i}} \boldsymbol{X}_i Y_i - \left(\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1} \sum_{i=1}^{n} \boldsymbol{X}_i Y_i \\ &= \left(\sum_{i=1}^{n} R_i e^{-\gamma_1' \boldsymbol{X}_{-1,i}} \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1} \sum_{i=1}^{n} \left(R_i e^{-\gamma_1' \boldsymbol{X}_{-1,i}} \boldsymbol{I}_{p\times p} - \boldsymbol{r}\right) \boldsymbol{X}_i \epsilon_i + o_p(n^{-1/2}), \end{aligned}$$

where

$$\boldsymbol{r} = \left(\sum_{i=1}^{n} R_i e^{-\boldsymbol{\gamma}_1' \boldsymbol{X}_{-1,i}} \boldsymbol{X}_i \boldsymbol{X}_i'\right) \left(\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1}.$$

In practice $\boldsymbol{\gamma}_1$ can be substituted by $\hat{\boldsymbol{\gamma}}_1$.

If we introduce the notations $\boldsymbol{A} = \sum_{i=1}^{n} R_i e^{-\boldsymbol{\gamma}_1' \boldsymbol{X}_{-1,i}} \boldsymbol{X}_i \boldsymbol{X}_i'$ and $\boldsymbol{B}_i = \left(R_i e^{-\boldsymbol{\gamma}_1' \boldsymbol{X}_{-1,i}} \boldsymbol{I}_{p\times p} - \boldsymbol{r}\right)$ then this becomes:

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \;=\; \boldsymbol{A}^{-1} \sum_{i=1}^{n} \boldsymbol{B}_i \boldsymbol{X}_i \epsilon_i,$$

The imprecision of $\hat{T}^{wls}$ (as well as the residual variance σ^2) may then be calculated as for $\hat{T}^{ols}$ (see section 3.3.3), but using the matrices $\boldsymbol{A}, \boldsymbol{B}_i$ and $\boldsymbol{r}$ defined above. The homoscedasticity assumption is quite strong. Its violation is likely, but, as argued in the annex, generally does not impose major concerns.

5 Missing not at random estimator $\hat{T}^{mnar}$

In the previous section we argued how a violation of the model assumption can be mitigated by using a doubly robust estimator. Remember that, besides the model assumption, we also assumed that the data is missing at random. In this section we show how the results in the previous section can be extended (1) to construct a retrospective hypothesis test for the MAR assumption and (2) if the data are missing not at random, how we can accomodate that.

When the data are missing not at random in the sense that $E(Y_i|R_i = 0, \boldsymbol{X}_i) \neq E(Y_i|R_i = 1, \boldsymbol{X}_i)$ for some $\boldsymbol{X}_i$, then we may postulate a model for $P(R_i = 1|\boldsymbol{X}_i, Y_i)$, which additionally allows for a dependence on the current outcome. For instance, we may postulate that

$$\text{logit}P(R_i = 1|\boldsymbol{X}_i, Y_i) = \gamma_0 + \boldsymbol{\gamma}_1' \boldsymbol{X}_{-1,i} + \theta Y_i. \tag{20}$$

Here, θ describes the extent to which companies with the same observed data (i.e. fiscal data, employment data, historical value addeds, ...), but different value added in the current year, have different probabilities of having submitted their data at the assessment time. In particular, $\theta = 0$ corresponds with Missing at Random.

One may thus retrospectively test the null hypothesis of Missing at Random once all financial data have been submitted, using a standard Wald test that $\theta = 0$ in model (20)[16].

If the hypothesis test concludes that $\theta \neq 0$ then the missingness is not at random. In that case we can, by the same reasoning as in section 4, give a higher weight to observations where the probability of being missing is higher. In other words, we then define weights equal to the reciprocal of $P(R_i = 1|\boldsymbol{X}_i, Y_i)$.

However, at the time of the assessment, the observed data carry no information about θ, because Y_i is missing when $R_i = 0$. We will therefore first explain how to proceed for a prespecified choice of θ. A reasonable guess for θ can be obtained by fitting model (20) to the data from the previous year using maximum likelihood (since the financial data are all available in that case), although there are no guarantees that the same value of θ would apply to the current year.

For given θ, we then recommend estimating $\boldsymbol{\gamma}_1$ by minimising

$$\sum_{j=2}^{p} \left(\frac{\sum_{i=1}^{n}(1-R_i)\boldsymbol{X}_{ji}}{\sum_{i=1}^{n} 1 - R_i} - \frac{\sum_{i=1}^{n} R_i e^{-\boldsymbol{\gamma}_1' \boldsymbol{X}_{-1,i} - \theta Y_i} \boldsymbol{X}_{ji}}{\sum_{i=1}^{n} R_i e^{-\boldsymbol{\gamma}_1' \boldsymbol{X}_{-1,i} - \theta Y_i}} \right)^2,$$

as before. Note that these equations ensure that Y_i is only needed for companies with $R_i = 1$, so that the equations can be calculated. This is important since standard maximum likelihood would give infeasible estimators.

Given an estimate $\hat{\boldsymbol{\gamma}}_1$ of $\boldsymbol{\gamma}_1$, an estimate $\hat{T}^{mnar}$ of T is now obtained as in Section 4 upon substituting $\hat{P}(R_i = 1|\boldsymbol{X}_i)$ by $\hat{P}(R_i = 1|\boldsymbol{X}_i, Y_i)$. The weights thus become $e^{-\boldsymbol{\gamma}_1' \boldsymbol{X}_{-1,i} - \theta Y_i}$, the term θY_i in the exponent explicitly accounts for the missingness being not at random.

[16] For the hypothesis test the coefficients of the logistic model (20) can be estimated by maximising the likelihood function. In this paper we used Firth regression (Firth , 1993, Heinze and Schemper , 2002) to eliminate small sample bias.

6 Application to the port study

6.1 Retrospective testing procedure

In the previous sections we have introduced several estimators ($\hat{T}^{prop}, \hat{T}^{ols}, \hat{T}^{mix}$) for the total value added of a population of companies in the presence of missing data. We analysed the underlying imputation model's assumptions and assessed their (un)biasedness and (im)precision when the data are missing at random. We also argued how the imputation models could be made more robust against model extrapolation by using estimated probabilities of missingness as weights in a weighted linear regression (i.e. the estimator $\hat{T}^{wls}$). As the missing at random assumption is crucial, we also presented a hypothesis test to verify its credibility. Finally we presented a method that can be used when the data are not missing at random (i.e. the estimator $\hat{T}^{mnar}$).

In this section we will compare the performance of each of these estimators on a real data set, namely the data that was used for the population of the Belgian ports study (e.g. Coppens et al. (2018), Mathys (2017)). The reason for using this study as a benchmark is that there has been an explicit demand for faster publication. We used the population for the years 2014 and 2015, because at the time we started working on this paper (October 2017) the total value added for these years was known, just as well as the value added for each company[17].

To make things clear, let's take e.g. $y = 2015$. At the time of writing the paper (October 2017) we knew the target value $T(y = 2015)$ (because October 2017 is more than 14 months later than end of 2015). At each assessment date u, $2016/07/31 \leq u \leq 2017/02/28$ we can extract the data for the companies that had already deposited their financial accounts at u and exclude those that reported after u[18]. As we fix the deposit date at u, some of the companies in

[17] Pro memorie: the CBSO receives the major part of the accounts for a certain financial year y between July $y+1$ and August $y+1$, but it can take until February-March of $y+2$ before all accounts for year y are (considered to be) complete. Therefore, in October 2017 we knew (approximately) what the "true" value of $T(y)$ for the years 2015 and 2016 would be.

[18] The CBSO registers, for each account, the date at which the account was received. Therefore we can

the population for 2015 have not yet reported and their value added has to be estimated with one of the described imputation models ($\hat{T}^{prop}, \hat{T}^{ols}, \hat{T}^{mix}$). Moreover, we know the target value for $T(y = 2015)$, so we can compare each of these estimated values ($\hat{T}^{prop}(y = 2015), \hat{T}^{ols}(y = 2015), \hat{T}^{mix}(y = 2015)$) to $T(y = 2015)$. Note that this comparison can only be made ex post, when $T(y)$ is known, which is not the case at the assessment date u. Therefore we also computed the precision of each estimator as a prediction interval at the 95% confidence level using only information available at u.

This paper assesses the imprecision using the bootstrap (see Efron and Tibshirani, 1994) based on which we constructed a 95% prediction interval[19]; for $\hat{T}^{mix}$ we used the parametric bootstrap (see section 3.5.3); for the other estimators the non-parametric bootstrap was used (see sections 3.2.3 and 3.3.3). For reasons of computational efficiency we also derived analytical formulas for the (im)precision of all estimators except for $\hat{T}^{mix}$ (see equations (10) and (16) for the imprecision of the proportional and the OLS imputation). This section will compare the outcome of the analytical formulas to the bootstrap-based intervals.

Remember that the missing at random assumption is crucial for the unbiasedness of the predictions. In section 5 a retrospective test was presented to verify the validity of that assumption. In the current section this test will be applied at each assessment date u. Note again that this test can only be performed ex post, when the outcome Y_i is known for all companies i.

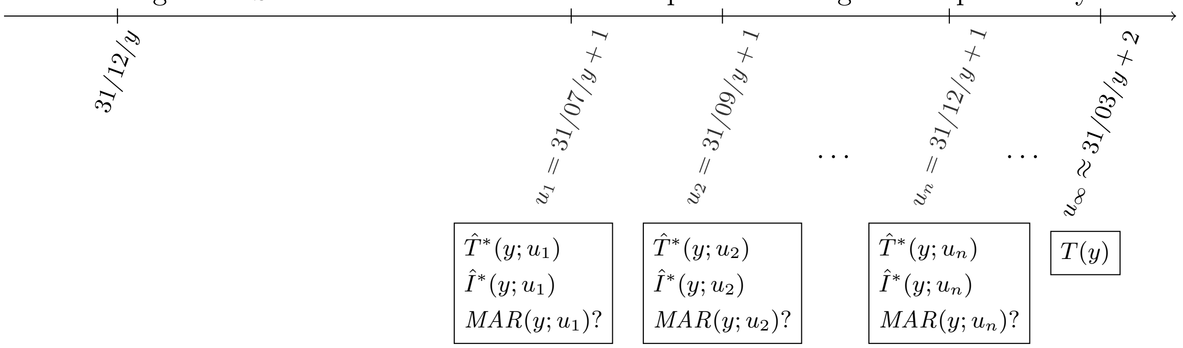
The retrospective testing process is schematised in figure 2 for year y: at e.g. $u_1 = 31/07/y + 1$ some companies deposited their accounts for y, while others did not. Using the above estimators, we can, at assessment date u_1, estimate the total value added $\hat{T}^{(s)}(y; u_1)$ for $y \in \{2014, 2015\}$, where $(s) \in \{prop, ols, mix\}$. For each estimator we can also estimate a 95% prediction interval $\hat{I}(y; u)$ (either using an analytical formula or using the bootstrap). Moreover, at each assessment date u we can test the MAR assumption. Fourteen months after the end of

retrospectively test the state of that database at each such date u.

[19] We used the 95% 'percentile' type intervals, see e.g. (Efron and Tibshirani, 1994, sect. 13.3)

year y we know the target value $T(y)$, so at the time we started working on the paper (October 2017) we knew the target values $T(2014), T(2015)$[20].

Figure 2: Schematic overview of the retrospective testing for the port study



6.2 Detailed specification of the OLS- and Mix- estimators

The estimator $\hat{T}^{prop}$ given by equation (9) does not require further specifications (it was already mentioned that X_i is the value added in the previous year and that Z_i is a proxy for the value added derived from fiscal data). This is the method that is currently applied and therefore it serves as our benchmark. In its current application, the ratio $\hat{\beta} = \sum_{i=1}^{n} R_i Y_i / \sum_{i=1}^{n} R_i X_i$ is computed by branch[21] of activity and by size class.

The regression models will be estimated by size class. The branch of activity is included as an explanatory variable in the OLS procedure and as a random effect in the mixed effect model.

The estimators $\hat{T}^{ols}$ and $\hat{T}^{mix}$ require us to specify the components of $\boldsymbol{X}_i$. For the current retrospective testing evaluation we defined $\boldsymbol{X}_i$ to be

$$\boldsymbol{X}_i = [1 \; X_{1i} \; X_{2i} \; X_{3i} \; X_{4i} \; X_{5i} \; X_{6i} \; X_{7i}]$$

where

[20] The attentive reader will notice that e.g. for $y = 2015$ the total $T(y)$ is not identical to the figures in Mathys, 2015. The reason is that in Mathys, 2015 there are additional corrections after the imputation step e.g. companies for that are only partially included in the study.

[21] A branch of activity is defined as a group of NACE-codes. In this paper we grouped companies by the first position of the NACE code.

– $X_{1i} = Y_i^*(y-1)$ is value added in the previous year (with a zero when it is not available, hence the '*' superscript);

– $X_{2i} = Z_i^*(y)$ is value added derived from fiscal data in year y (with a zero when it is not available);

– $X_{3i} = e_i(y)$ is the number of persons employed by company i;

– $X_{4i} = D_i^{(-1)}(y-1)$ is an indicator for the availability of the value added in the previous year;

– $X_{5i} = D_i^{(f)}(y)$ is an indicator for the availability of fiscal data in the current year;

– $X_{6i} = N1_i(y)$ is the branch of activity (the first position of the Nace code);

– X_{7i} is the interaction effect between the activity branch and X_{1i}.

For the estimator $\hat{T}^{mix}$ we decompose $\boldsymbol{X}_i$ into $(\boldsymbol{W}_i, U_i, Z_i)'$ where

$$\boldsymbol{W}_i \ = [1 \ X_{1i} \ X_{2i} \ X_{3i} \ X_{4i} \ X_{5i}] \,.$$

These were the explanatory variables used for the purpose of the retrospective testing in this paper. For NBB-internal use, additional (confidential) variables may be added (like salaries paid) that might further reduce the uncertainty.

The "corrected ratio estimator" copies a proxy value from fiscal data for new companies that did not yet deposit their accounts. By the choice of the variables supra, it can be seen that the OLS- and mixed estimators correct the proxy value via regression adjustment.

6.3 Definition of the population for year y

The imputation methods assume that we know all the companies that belong to the population in year y at u, where $u \geq y + 7m$. Moreover, for each company i the X_i values must be observed. After $19m$ we can reasonably assume that the population for $y-1$ ($\mathcal{P}(y-1)$) is known (Note that $y + 7m = (y-1) + 19m$, it is assumed that all companies have deposited their account for $y - 1$ 19 months after the account is closed).

The population for y can be defined as follows:

$$\mathcal{P}(y) = (\mathcal{P}(y-1) \setminus \mathcal{E}(y)) \cup \mathcal{N}(y),$$

where $\mathcal{E}(y)$ is the set of companies exiting the population during year y and $\mathcal{N}(y)$ is the set of companies entering during year y. $\mathcal{E}(y)$ can be derived from other data sources with relatively high precision; $\mathcal{N}(y)$ can only be defined approximately, but it should be noted that new companies are usually the smaller ones (cfr infra for the definition of small companies and their impact on the estimation results).

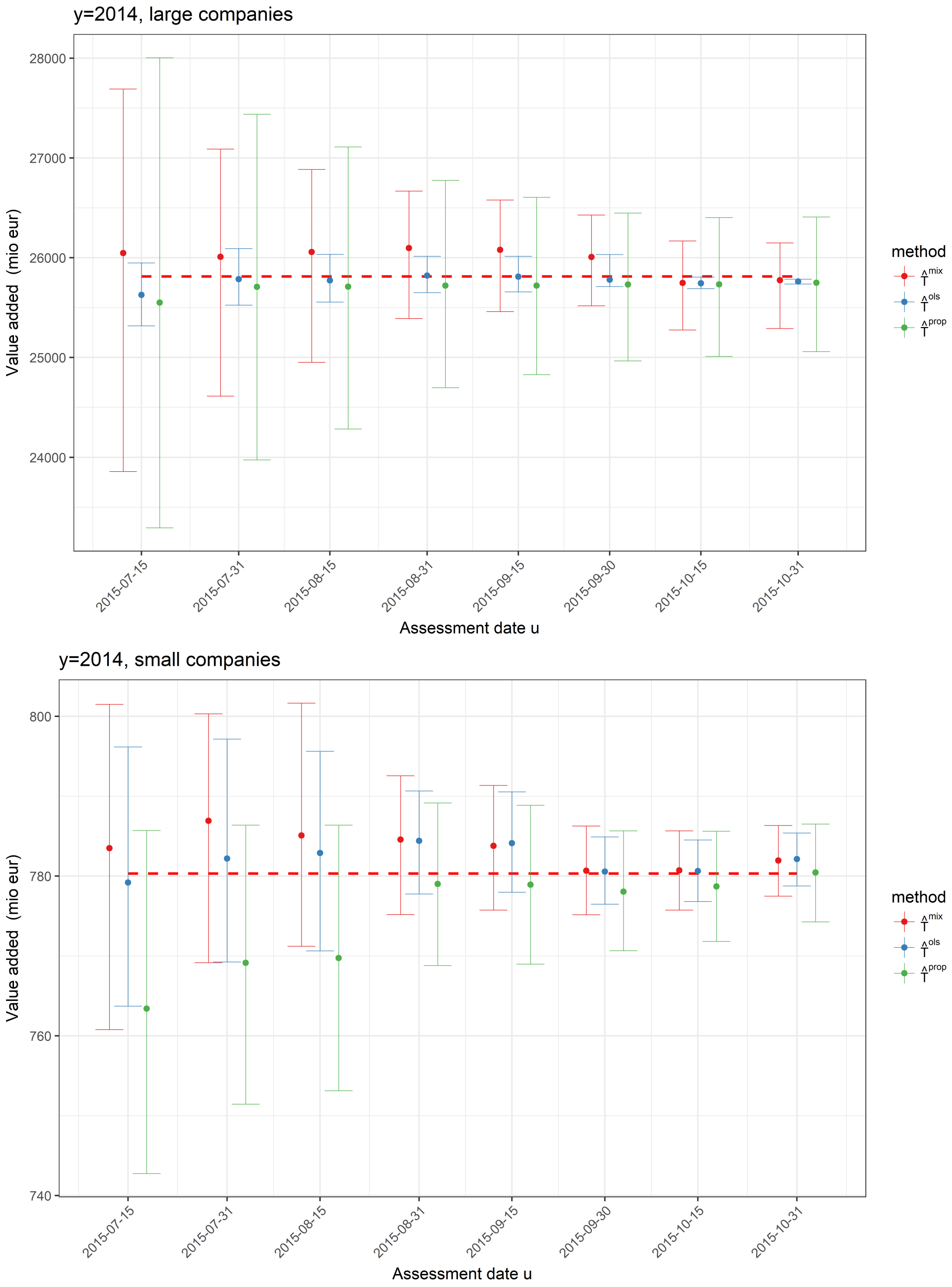
Note that $\boldsymbol{X}_i$ is observed for all the companies i in the population (for year y) $\mathcal{P}(y)$.

6.4 Simulation of (un)biasedness and (im)precision

Figure 3 illustrates the estimated values $\hat{T}^{prop}(y), \hat{T}^{ols}(y), \hat{T}^{mix}(y)$ along with 95% bootstrap prediction intervals (the error bars) for the year $y = 2014$. The top panel shows the results for the companies that report full schemes or the 'large' ones; the bottom panel for those that report abbreviated or micro schemes, called the 'small' companies. The dashed horizontal red line represents the true (ex post known) value $T(y)$. On the horizontal axis one finds the different values for the (simulated) assessment date u while on the vertical axis one finds the estimated values and their uncertainty (i.e. the 95% bootstrap prediction interval) and the target value $T(y)$. Note that the scales on the vertical axis are very different for both subpanels. Figure 4 gives similar results for year $y = 2015$.

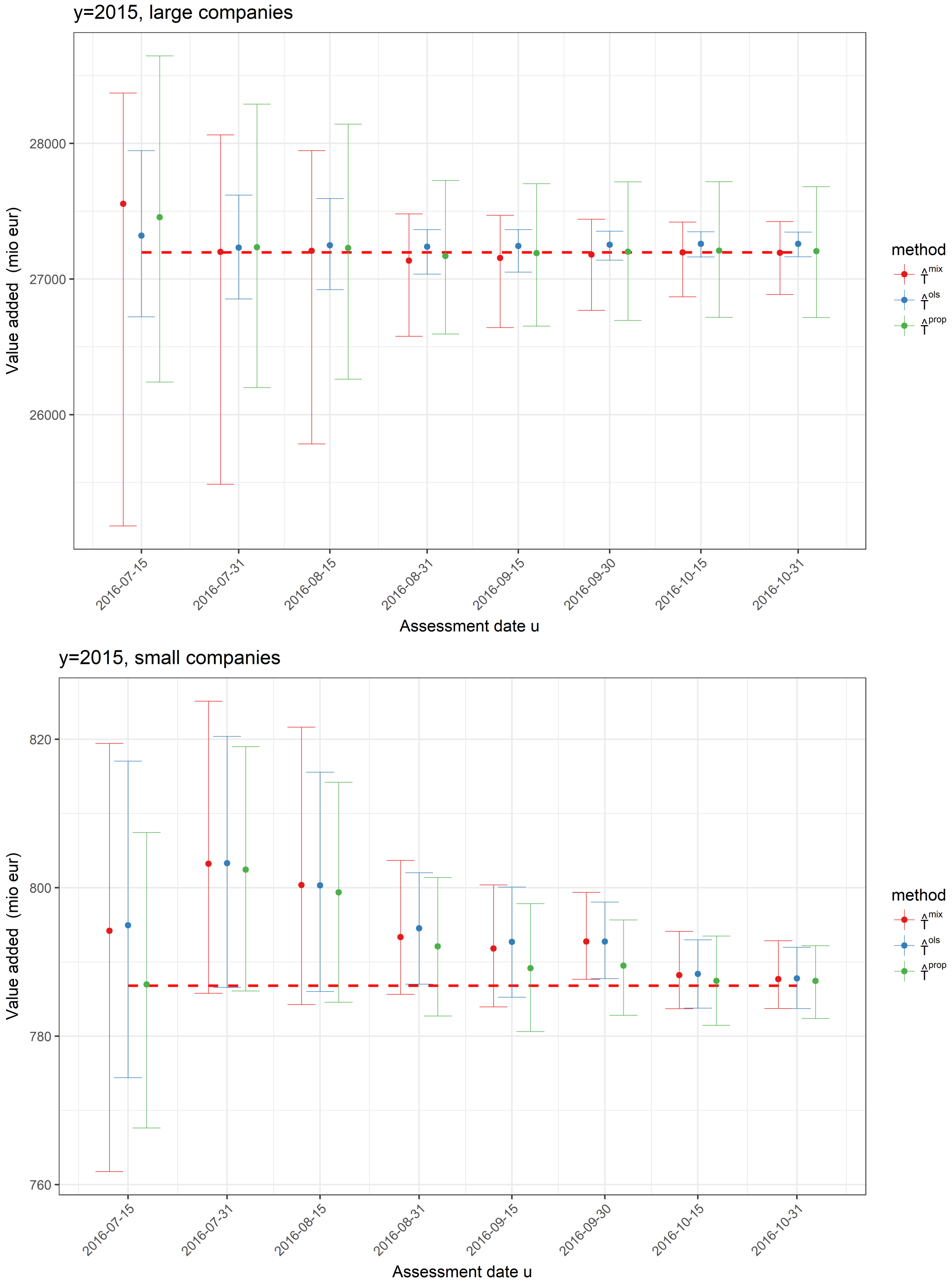
As expected, the estimated value (for all the methods) converges to the ex-post value when the total value is estimated at later dates u, because the number of companies for which the value added is observed increases (and thus the number of companies for which value added must be estimated decreases). The width of the estimated 95% prediction intervals also become

Figure 3: Estimation results and bootstrap intervals for $y = 2014$


y=2014, large companies

y=2014, small companies

Figure 4: Estimation results and bootstrap intervals for $y = 2015$



smaller with increasing u.

The ex-post observed value added (and thus the "true" value added) for the large companies for year $y = 2015$, i.e. $T^{large}(2015)$, is 27 196.86 million euro. As an example, if we would estimate the total value added for that year for the large companies at date u =2016-08-31 using the OLS-estimator, we find $\hat{T}^{ols,large}(2015; u)$ =27 238.33 million euro, or an estimation error of 0.15%. This error can only be observed ex-post, when the "true" value $T(2015)$ has been observed (in practice this is after February 2017). At the assessment date u (August 2016) $T(2015)$ is unknown and we can only calculate an estimate of the imprecision of the estimate $\hat{T}^{ols,large}(2015; u)$ using data available at u. Figure 4 shows the uncertainty as 95% prediction intervals for $y = 2015$ (for $y = 2014$, see figure 3), computed using the bootstrap. The bootstrap intervals are computed by repeatedly (2500 times) predicting the value added using other data (see sections 3.2.3 and 3.3.3 for more details). This is computationaly intensive and therefore we also derived analytical formulas for computing this uncertainty (except for the imputation based on mixed models).

The 95% bootstrap (percentile type) intervals for different estimators are shown as error bars in figures 3 and 4. As an example, the bootstrap estimated 95% prediction interval for the large companies in year 2015 and for the OLS-estimator, $\widehat{Ib}^{ols,large}(2015; u) = [\widehat{Ib}_{low}^{ols,large}(2015; u); \widehat{Ib}_{high}^{ols,large}(2015; u)] = [27\ 036.93\ ;\ 27\ 364.58]$ million euro at u =2016-08-31. Note that (see also figure 4) the estimated prediction intervals contain the ex-post observed value and therefore there are no signs of bias[22]. The percentile type bootstrap intervals are (in general) asymmetric around the estimated value $\hat{T}^{ols,large}(2015; u)$ =27 238.33 million euro. Considering the fact that large companies account for a much larger share in total value added (97.2% in 2015), we conclude that the OLS-estimator outperforms the other estimators. This confirms the theoretical arguments given in section 3.4.

[22] The prediction intervals are computed at a 95% level, so, in (infinitely) repeated samples, 95% of the computed intervals must contain the "true" (ex post) value.

We define the relative uncertainty[23] of the estimate as the width of the 95% bootstrap interval relative to the estimated value or

$$\mu_b^{ols,large}(2015;u) = \frac{\widehat{Ib}_{high}^{ols,large}(2015;u) - \widehat{Ib}_{low}^{ols,large}(2015;u)}{\hat{T}^{ols,large}(2015;u)}$$

Using equation (16), the imprecision of the OLS-estimator can be computed analytically[24]. The analytically computed 95% prediction interval is defined as

$$\widehat{Ia}^{ols,large}(2015;u) = \hat{T}^{ols,large}(2015;u) \pm 2\sqrt{E\left((\hat{T}^{ols,large}(2015;u) - T^{large}(2015))^2\right)}$$

The analytical uncertainty $\mu_a^{ols,large}(2015;u)$ is defined in a similar way as the bootstrap uncertainty. The estimated value added and their relative uncertainty for the large companies at two assessments dates[25] is shown in table 1 for the estimators $\hat{T}^{ols}, \hat{T}^{mix}, \hat{T}^{prop}$. Table 2 is similar but for the small companies.

Table 1: Estimated values and their analytical uncertainty (μ_a) and bootstrap uncertainty (μ_b) for the large companies for y=2015 (mio eur)

method	u	$\hat{T}^{*,large}(2015;u)$	$\mu_b/2$ (%)	$\mu_a/2$ (%)
ols	2016-08-15	27 249.60	±1.2(%)	±1.3(%)
mix		27 208.09	±4.0(%)	-
prop		27 229.77	±3.5(%)	±3.8(%)
wls		27 281.41	-	±1.5(%)
ols	2016-08-31	27 238.33	±0.6(%)	±0.4(%)
mix		27 136.24	±1.7(%)	-
prop		27 170.31	±2.1(%)	±2.1(%)
wls		27 247.66	-	±0.8(%)
ols	2016-10-31	27 259.97	±0.3(%)	±0.3(%)
mix		27 193.27	±1.0(%)	-
prop		27 205.93	±1.8(%)	±1.9(%)
wls		27 248.19	-	±0.4(%)

[23] Note that, if the intervals would have been symmetric, this would imply that the interval could be written as $\hat{T}^{ols,large}(2015;u) \cdot (1 \pm \frac{1}{2}\mu_b^{ols,large}(2015;u))$

[24] The imprecision can be approximated using the simpler formula in equation (14), i.e. $\hat{\sigma}^2 \sum_i (1 - R_i)$, instead of the exact formula (16). $\sum_i (1 - R_i)$ is the number of companies for which the value added is unknown at u, $\hat{\sigma}^2$ is the estimated residual variance given by equation (15). The latter value is part of the standard output of a linear regression. The impact of this approximation is shown in annex B

[25] We choose Mid August and end August because these are the assessment dates we aim at. The conclusions for other assessment dates u are similar however.

The tables show that, as expected, the analytically computed uncertainty is very similar to the bootstrapped uncertainty, but the former has the advantage of being more efficient to compute and can also be computed at the individual company level (when homoscedasticity is assumed)[26]. We did not derive an analytical formula for the uncertainty of the mixed model-estimator. Remember that the ex-post observed value for the large companies is 27 196.86 million euro and 786.78 million euro for the small companies. For u =2016-08-15 the ex post observed estimation error varies by estimator: 0.19% (ols), 0.04% (mix), 0.12% (prop), 0.31% (wls) for the large companies and 1.72% (ols), 1.73% (mix), 1.6% (prop), 1.86% (wls) for the small companies. It is noted once more that these errors are unknown at date u because the total value added $T(2015)$ cannot be observed at early dates u. The estimated errors μ_b and μ_a on the other hand can be computed using data available at the assessment date u.

Note that the largest part of the uncertainty μ_a of the estimator $\hat{T}^{prop}$ for the large companies comes from the term $\sum_{i=1}^{n}(1-R_i)(1-S_i)E\left(\nu_i^2|Z_i\right)$ in equation (10). This term reflects the uncertainty due to using fiscal data as a proxy for the value added in the annual accounts.

6.5 Violations of the assumptions; double robustness, MAR-test and the MNAR-estimator.

Figures 3 and 4 did not show any signs of (strong) bias. From the theoretical considerations in the first sections this is expected (except for the mixed model imputation) when the model assumptions and the MAR assumption are fulfilled.

In section 4 the double robust estimator $\hat{T}^{wls}$ was introduced and it was shown that, even if the (OLS-) imputation model assumptions are violated, the estimator is still unbiased provided that the model for missingness is correctly specified. Therefore, even if the OLS- imputations would give biased results (when the underlying assumptions of the linear model are violated) the WLS- imputation will be unbiased if the missingness model (18) is correctly specified. At

[26] This can be seen by analyzing equation (16): it is the sum of the estimation errors for the companies where $R_i = 0$.

an intuitive level, the WLS estimator gives more weight to observations with covariate levels $\boldsymbol{X}_i$ where there is a lot of 'missing data' (i.e. $P(R_i = 1|\boldsymbol{X}_i)$ is closer to 0). They thus ensure that the regression model fits well at covariate values $\boldsymbol{X}_i$ of companies who did not yet submit their financial data; they thus target a good fit in the region of the covariate space where predictions will be made (Vansteelandt, Carpenter and Kenward, 2010). In figure 5 the results for the OLS- and the WLS- imputation are compared. The two estimators show comparable performance. For the small companies (where there is more missingness) the prediction intervals are narrower, as expected (cfr. supra). It must be said however that the WLS-estimator is computationaly much more challenging and probably also more difficult to understand by non (expert) statisticians.

A bias could also arise from a violation of the MAR-assumption. This holds for all the imputation models $\hat{T}^{ols}, (\hat{T}^{wls}), \hat{T}^{prop}, \hat{T}^{prop}$. However, as argued in section 2.2, violation of the MAR assumption is less probable when including more $\boldsymbol{X}_i$ variables. This is a disadvantage of the proportional imputation strategy because it is based on a single $\boldsymbol{X}_i$ variable namely $X_i = Y_i(y-1)$.

The MAR-assumption can also be tested retrospectively. Firth regression (see Firth (1993), Heinze and Schemper (2002)) was used for fitting the logistic model

$$\text{logit}P(R_i = 1|\boldsymbol{X}_i, Y_i) = \gamma_0 + \boldsymbol{\gamma}_1'\boldsymbol{X}_{-1,i} + \theta Y_i$$

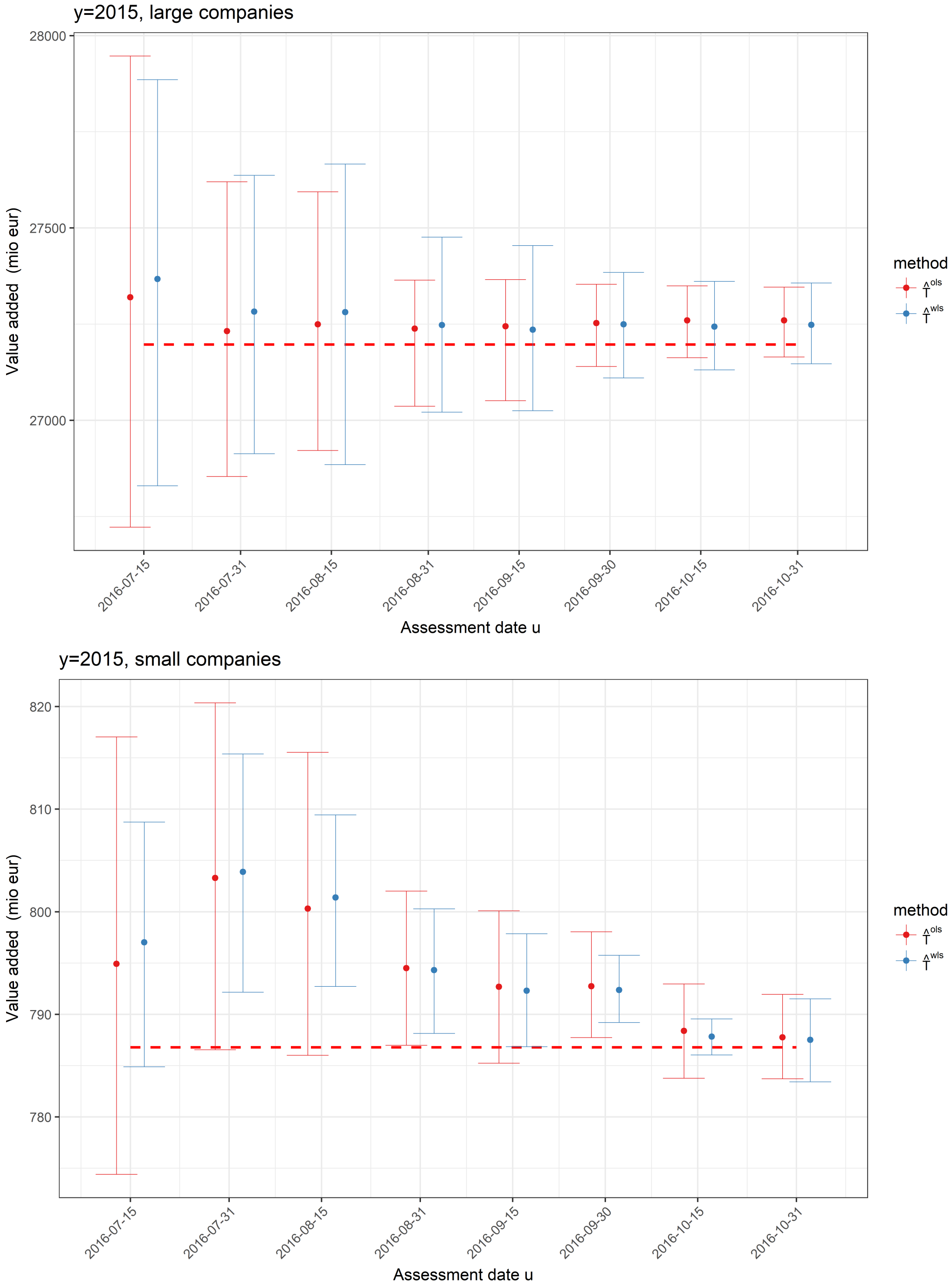
and testing $H_0 : \theta = 0$ (i.e. missingness is at random) versus $H_1 : \theta \neq 0$ (i.e. missingness is not at random). Alternatively the hypothesis $H_0 : \theta_{log} = 0$ (i.e. missingness is at random) versus $H_1 : \theta_{log} \neq 0$ (i.e. missingness is not at random) was tested for the model:

$$\text{logit}P(R_i = 1|\boldsymbol{X}_i, Y_i) = \gamma_0 + \boldsymbol{\gamma}_1'\boldsymbol{X}_{-1,i} + \theta_{log} log(Y_i)$$

To avoid the multiple testing problem, we also performed a likelihood ratio (LR) test[27] for

[27] The fact that Firth regression uses a penalised likelihood was taken into account.

Figure 5: Comparison of OLS and WLS estimations, result for $y = 2015$



comparing the nested models $\text{logit}P(R_i = 1|\boldsymbol{X}_i, Y_i) = \gamma_0 + \boldsymbol{\gamma}_1'\boldsymbol{X}_{-1,i} + \alpha Y_i + \alpha_{log}log(Y_i)$ and $\text{logit}P(R_i = 1|\boldsymbol{X}_i, Y_i) = \gamma_0 + \boldsymbol{\gamma}_1'\boldsymbol{X}_{-1,i}$.

Tables 3 and 4 show the results for $y = 2014$ and $y = 2015$. $\hat{\theta}$ and $\hat{\theta}_{log}$ are the estimates of θ and θ_{log}, repectively, in the models supra, p_Y and $p_{log(Y)}$ are the p-values of the corresponding tests. p_{LR} is the p-value of the likelihood ratio test for the comparison of the nested models. The likelihood ratio test corrects for multiple testing. With one exception ($y =$2014,$u =$2015-08-15), the null hypothesis (of the LR test) of missingness at random cannot be rejected at the 5% significance level. Therefore, while the date at which a company deposits its annual account may be related to its size (larger companies tend to deposit faster), as well as various historical company characteristics, it seems unlikely that there would be a further dependence on the value added in the current year. The missing at random assumption is therefore considered to be reasonably plausible.

If the MAR-assumption would have been rejected then the estimator $\hat{T}^{mnar}$ is recommended, but in that case it must be assumed that the coefficient θ (θ_{log}) is relatively stable in consecutive years. Note that tables 3 and 4 can only be compiled ex post, when all $Y_i(y)$ are observed (so for $y = 2014$ this is at earliest at u=2016-02-29 and for $y = 2015$ at earliest at u=2017-02-28).

7 Discussion

In this paper, we have developed several estimation methods for computing a branch's total value added from incomplete annual accounting data under a missing at random assumption, along with careful uncertainty margins that incorporate finite population corrections. The importance of the availability of these uncertainty margins should not be underestimated because they will result in faster and higher quality publications.

For each proposed method we analyse the underlying assumptions, the estimation bias and the estimation uncertainty. The proposed imputation procedures all rely on an assumption

Table 2: Estimated values and their uncertainty (μ_a, μ_b) for the small companies for y=2015 (mio eur)

method	u	$\hat{T}^{*,small}(2015; u)$	$\mu_b/2$ (%)	$\mu_a/2$ (%)
ols	2016-08-15	800.31	±1.8(%)	±1.9(%)
mix		800.36	±2.3(%)	-
prop		799.36	±1.9(%)	±2.0(%)
wls		801.38	-	±1.1(%)
ols	2016-08-31	794.51	±0.9(%)	±1.0(%)
mix		793.35	±1.1(%)	-
prop		792.10	±1.2(%)	±1.2(%)
wls		794.32	-	±0.8(%)
ols	2016-10-31	787.77	±0.5(%)	±0.5(%)
mix		787.68	±0.6(%)	-
prop		787.45	±0.6(%)	±0.7(%)
wls		787.51	-	±0.5(%)

Table 3: MAR test for large companies

y	u	$\hat{\theta}$	$\hat{\theta}_{log}$	p_Y	$p_{log(Y)}$	p_{LR}
2014	2015-08-15	0.000	0.101	0.570	0.518	0.743
2014	2015-08-31	0.000	0.104	0.029	0.588	0.094
2014	2015-09-15	0.000	0.119	0.244	0.550	0.457
2015	2016-08-15	-0.000	0.343	0.899	0.053	0.161
2015	2016-08-31	-0.000	0.225	0.663	0.397	0.620
2015	2016-09-15	-0.000	0.217	0.757	0.578	0.809

Table 4: MAR test for small companies

y	u	$\hat{\theta}$	$\hat{\theta}_{log}$	p_Y	$p_{log(Y)}$	p_{LR}
2014	2015-08-15	0.000	0.805	0.855	0.004	0.002
2014	2015-08-31	0.001	0.269	0.101	0.180	0.191
2014	2015-09-15	0.001	0.262	0.131	0.206	0.244
2015	2016-08-15	0.000	0.298	0.051	0.333	0.184
2015	2016-08-31	0.001	0.226	0.046	0.282	0.071
2015	2016-09-15	0.001	0.240	0.134	0.275	0.202

of missing at random, namely that the values added in companies that did not yet deposit their annual accounts are similar to those in companies with the same characteristics (e.g. the same historical data) that did deposit their accounts by the evaluation date. The missing at random assumption was made for mathematical convenience, though it should only be used when considered plausible. In our opinion, the date at which a company deposits its annual account may be related to its size (larger companies tend to deposit faster), as well as various measured historical company characteristics, but it is unlikely that there would be a further dependence on the value added in the current year. The missing at random assumption is therefore considered to be reasonably plausible, but can be retrospectively assessed. We have moreover shown how the proposed estimation methods can be relatively easily accommodated in case the missing at random assumption fails.

Finally we retrospectively apply each strategy to data from the Belgian Port sector and compare their performance at several evaluation dates. All the proposed methods show good results on these data. The method using (ordinary least squares) regression is preferred because it is very flexible in the use of auxilairy variables, requires weaker assumptions than currently employed methods, has smaller estimation errors and is easily automatable. The use of more flexible and automatable methods, compared to the currently used ad hoc and sometimes manual methods, will result in faster publication. The automatization and the availability of estimation errors will result in higher quality of the sector studies. Estimation errors at the individual company level also fasten publication by making it possibe to focus on the cases with high imprecision.

A practical problem with the proposed strategies may arise when some companies are much larger than others, making them possibly very influential when fitting the imputation models. Removing or down-weighting outlying companies based on their current added value (or the magnitude of accompanying regression residuals) is not desirable as it may induce bias in the fitted prediction models. Removing or down-weighting outlying companies based on their company characteristics (i.e., the predictors in the imputation model) is less problematic, so long as none of the companies that did not yet submit their accounts is so extreme in terms of these characteristics. One may therefore consider basing the fitting of the imputation models only on those companies whose leverage is smaller than the largest leverage of companies with missing data, as well as than the typical cut-off $2p/\sum_{i=1}^{n} R_i$, with p the number of unknown coefficients in the regression model.

Further work remains to be done to evaluate how to best use the proposed methods in practice, in such a way that they guarantee reliable results. In particular, further study is warranted how to best model the value added in function of company characteristics with the aim of avoiding model misspecification. In future work, we will extend these methods to other sectoral and/or regional studies, just as well as to other variables than value added. We will additionally evaluate the use of flexible statistical learning methods (smoothing splines, gradient boosting, ...) to enable flexible modelling.

Also the empirical performance of estimators obtained via the proposed missing not at random strategy remains to be evaluated.

References

Brewer, K.R.W (1963). *Ratio estimation and finite populations: some results deducible from the assumption of an underlying stochastic process.* In: Australian journal of statistics 5.3, pp. 93-105.

Chamberlain, G. (1987). *Asymptotic efficiency in estimation with conditional moment restrictions*, In: Journal of Econometrics, Elsevier, vol. 34(3), pages 305-334, March 2008.

Cochran, W.G. (1977). *Sampling techniques.* Wiley and sons: New York.

Cole S. and Hernan M. *Constructing inverse probability weights for marginal structural models.*, In: American Journal of Epidemiology, vol. 168(6), pages 656-664, 2008

Coppens, F. and Mathys, C. and Merckx, J.-P. and Ringoot, P. and Van Kerckhoven, M. (2018), *The economic importance of the Belgian ports : Flemish maritime ports, Lige port complex and the port of Brussels, Report 2016*, NBB Working paper Series, No 342, April 2018.

Efron, B. and Hastie, T. (2016), *Computer age statistical inference.* Stanford University.

Efron, B. and Tibshirani, R.J. (1994), *An introduction to the bootstrap.* Chapman & Hall.

Firth, D. (1993), *Bias reduction of maximum likelihood estimates.* Biometrica, Vol. 80, 27-38.

Fitzmaurice, G.M. and Laird, N.M. and Ware, J.H. (2011), *Applied Longitudinal Analysis.* John Wiley & Sons.

Greene, W.H. (2008), *Econometric Analysis.* Prentice Hall.

Gujarati, D.N. (2003), *Basic econometrics.* Mc Graw Hill.

Heinze G. and Schemper, M. (2002), *A solution to the problem of separation in logistic regression.* Statistics in medecine, 21, 2409-2419.

Hoaglin, D.C. and Welsch R.E. (1978), *The hat matrix in regression and anova.* American statistical association, Vol. 32, No. 1, 17-22.

Hosmer, D.W. and Lemeshow S. (2000), *Applied logistic regression.* John Wiley & Sons.

Mathys C. (2017), Economic importance of Belgian Ports: Flemish maritime ports, Lige port complex and the port of Brussels, Report 2015, NBB Working paper Series, No 321, June 2017.

NBB Central Balance Sheet Office (2013-2016), *Annual accounts submitted to the Central Balance Sheet Office.* Ed. by NBB Central Balance Sheet Office.

Newey, W.K. and McFadden, D. (1994). *Large Sample Estimation and Hypothesis Testing*, in Handbook of Econometrics, Volume 4, ed. R.F. Engle and D. McFadden. Amsterdam: North Holland, 2111-2245.

Pregibon D. (1981). *Logistic regression diagnostics*, The annals of statistics, Volume 9, No. 4, 705-724.

Royall, R.M. (1970). *On finite population sampling theory under certain regression models.* In: Biometrica 57.2, pp. 377-387.

Rubin, D.B. (1976). Inference and Missing Data. *Biometrika*, **63**, 581-592.

Schafer, J.L. and J.W. Graham (2002). *Missing data: our view of the state of the art.* In: Psychological methods 7.2, pp. 147-177.

Seaman, S.R. and Vansteelandt, S. (2018). Introduction to Double Robust Methods for Incomplete Data. *Statistical Science*, **33**, 184-197.

Vansteelandt, S., Carpenter, J. and Kenward, M.G. (2010). Analysis of incomplete data using inverse probability weighting and doubly robust estimators. *Methodology*, **6**, 37-48.

A Derivations of the formula's for the variance of the estimators

A.1 Variance of $\hat{T}^{prop}$

To analyse the difference between $\hat{\beta} = \frac{\sum_{i=1}^n R_i Y_i}{\sum_{i=1}^n R_i X_i}$ and $\beta = \frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n X_i}$, let us simplify the notation and note that $\sum_{i=1}^n R_i X_i \equiv X_{(O)}$ is the total of X for the companies for which the value added is observed at date u (because in that case $R_i = 1$), while $\sum_{i=1}^n X_i \equiv X_{(P)}$ is the total of X for the whole population that will eventually be observed.

$$\begin{aligned}
\hat{\beta} - \beta &= \frac{\sum_{i=1}^n R_i Y_i}{X_{(O)}} - \frac{\sum_{i=1}^n Y_i}{X_{(P)}} \\
&= \frac{1}{X_{(O)}} \left(\sum_{i=1}^n R_i Y_i - \sum_{i=1}^n Y_i \frac{X_{(O)}}{X_{(P)}} \right) \\
&= \frac{1}{X_{(O)}} \sum_{i=1}^n \left(R_i Y_i - Y_i \frac{X_{(O)}}{X_{(P)}} \right) \\
&= \frac{1}{X_{(O)}} \sum_{i=1}^n Y_i \left(R_i - \frac{X_{(O)}}{X_{(P)}} \right) \\
&= \frac{1}{X_{(O)}} \sum_{i=1}^n (\beta X_i + \epsilon_i) \left(R_i - \frac{X_{(O)}}{X_{(P)}} \right) \\
&= \frac{1}{X_{(O)}} \left(\beta \sum_{i=1}^n X_i R_i - \beta \frac{X_{(O)}}{X_{(P)}} \sum_{i=1}^n X_i + \sum_{i=1}^n \epsilon_i R_i - \frac{X_{(O)}}{X_{(P)}} \sum_{i=1}^n \epsilon_i \right) \\
&= \frac{1}{X_{(O)}} \left(\beta X_{(O)} - \beta \frac{X_{(O)}}{X_{(P)}} X_{(P)} + \sum_{i=1}^n \epsilon_i R_i - \frac{X_{(O)}}{X_{(P)}} \sum_{i=1}^n \epsilon_i \right) \\
&= \frac{1}{X_{(O)}} \sum_{i=1}^n \epsilon_i \left(R_i - \frac{X_{(O)}}{X_{(P)}} \right).
\end{aligned}$$

Here, the terms involving $X_{(O)}/X_{(P)}$ represent finite population corrections; they would reduce to zero if an infinite population were considered.

Let $\sum_{i=1}^n (1 - R_i) X_i / \sum_{i=1}^n (1 - R_i) \equiv \bar{X}_{(0)}$, from which $\sum_{i=1}^n (1 - R_i) \bar{X}_{(0)} = X_{(P)} - X_{(O)}$. The variance can be assessed as:

$$
\begin{aligned}
& E\left((\hat{T}-T)^2|\{X_i, R_i; \forall i\}\right) \\
& = E\left[\left\{\sum_{i=1}^{n}(1-R_i)\left(X_i\hat{\beta}-Y_i\right)\right\}^2 |\{X_i, R_i; \forall i\}\right] \\
& = E\left[\left\{\sum_{i=1}^{n}(1-R_i)\left(\bar{X}_{(0)}\hat{\beta}+(X_i-\bar{X}_{(0)})\hat{\beta}-Y_i\right)\right\}^2 |\{X_i, R_i; \forall i\}\right] \\
& = E\left[\left\{\sum_{i=1}^{n}(1-R_i)\left(\bar{X}_{(0)}(\hat{\beta}-\beta)+(X_i-\bar{X}_{(0)})(\hat{\beta}-\beta)-\epsilon_i\right)\right\}^2 |\{X_i, R_i; \forall i\}\right] \\
& = \left\{\sum_{i=1}^{n}(1-R_i)\right\}^2 \left(\bar{X}_{(0)}\right)^2 E\left\{(\hat{\beta}-\beta)^2|\{X_i, R_i; \forall i\}\right\} \\
& + E\left[\left\{\sum_{i=1}^{n}(1-R_i)\left((X_i-\bar{X}_{(0)})(\hat{\beta}-\beta)-\epsilon_i\right)\right\}^2 |\{X_i, R_i; \forall i\}\right] \\
& + E\left[\sum_{i=1}^{n}\sum_{j=1}^{n}(1-R_i)(1-R_j)\bar{X}_{(0)}(\hat{\beta}-\beta)\left((X_j-\bar{X}_{(0)})(\hat{\beta}-\beta)-\epsilon_j\right)|\{X_i, R_i; \forall i\}\right] \\
& = \left\{\sum_{i=1}^{n}(1-R_i)\right\}^2 \left(\bar{X}_{(0)}\right)^2 E\left\{(\hat{\beta}-\beta)^2|\{X_i, R_i; \forall i\}\right\} + E\left[\sum_{i=1}^{n}(1-R_i)\epsilon_i^2|\{X_i, R_i; \forall i\}\right] \\
& - E\left[\sum_{i=1}^{n}\sum_{j=1}^{n}(1-R_i)(1-R_j)\bar{X}_{(0)}(\hat{\beta}-\beta)\epsilon_j|\{X_i, R_i; \forall i\}\right],
\end{aligned}
$$

where we use that $\sum_{i=1}^{n}(1-R_i)(X_i-\bar{X}_{(0)})=0$ and that ϵ_i are independent and have mean zero conditional on $\{X_i, R_i; \forall i\}$. Here, the first term equals

$$\left(\frac{X_{(P)}}{X_{(O)}}-1\right)^2\sum_{i=1}^{n}\mathrm{Var}(\epsilon_i|X_i)\left(R_i-\frac{X_{(O)}}{X_{(P)}}\right)^2;$$

it reflects the imprecision due to the estimation of β and the fact that we consider a finite population and is generally small when the number of observed companies is large. The second term equals

$$\sum_{i=1}^{n}(1-R_i)\mathrm{Var}(\epsilon_i|X_i); \tag{21}$$

it reflects the imprecision due to Y_i varying around $X_i\beta$, which does not reduce as the number of observed companies is larger. The third term equals

$$\begin{aligned}
&-\frac{1}{X_{(O)}}(X_{(P)} - X_{(O)})\sum_{j=1}^{n}(1-R_j)\mathrm{Var}(\epsilon_j|X_j)\left(R_j - \frac{X_{(O)}}{X_{(P)}}\right) \\
&= \left(1 - \frac{X_{(O)}}{X_{(P)}}\right)\sum_{j=1}^{n}(1-R_j)Var(\epsilon_j|X_j);
\end{aligned}$$

it reflects additional uncertainty coming from the fact that we consider a finite population of companies, which induces a (generally) small correlation between the estimation errors $\hat{\beta} - \beta$ and the prediction errors ϵ_i. This term would be zero if an infinite population were considered.

Under a homoscedasticity assumption that the variance in ϵ_i does not depend on X_i, the variance of $\hat{T}$ then reduces to

$$\sigma^2\left[\sum_{i=1}^{n}(1-R_i)\left\{1 + \left(1 - \frac{X_{(O)}}{X_{(P)}}\right)\right\} + \left(\frac{X_{(P)}}{X_{(O)}} - 1\right)^2\sum_{i=1}^{n}\left(R_i - \frac{X_{(O)}}{X_{(P)}}\right)^2\right],$$

where $\sigma^2 = \mathrm{Var}\left(\epsilon_i|X_i\right)$ can be estimated as

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} R_i\left(Y_i - \hat{\beta}X_i\right)^2}{\left(\sum_{i=1}^{n} R_i\right) - 1}. \tag{22}$$

The homoscedasticity assumption is quite strong. Its violation is likely, but generally does not impose major concerns. Indeed, suppose for instance that $\mathrm{Var}\left(\epsilon_i|X_i\right) = \sigma^2 X_i$ for some X_i. Then the key component of the imprecision of $\hat{T}$ equals

$$\sum_{i=1}^{n}(1-R_i)\mathrm{Var}\left(\epsilon_i|X_i\right).$$

When the homoscedasticity assumption is made in error, then the estimator (5) of $\mathrm{Var}\left(\epsilon_i|X_i\right)$ is an approximately unbiased estimator of $\sigma^2\sum_{i=1}^{n} R_iX_i/\left(\sum_{i=1}^{n} R_i\right)$. It then follows that the key component of the imprecision of $\hat{T}$ is expected to equal

$$\sigma^2\sum_{i=1}^{n}(1-R_i)\frac{\sum_{i=1}^{n} R_iX_i}{\sum_{i=1}^{n} R_i}.$$

This approximates $\sigma^2 \sum_{i=1}^n (1-R_i)X_i$ when R is independent of X. When this approximation is poor, one may alternatively make progress under a specific model for the residual variance $\mathrm{Var}\,(\epsilon_i|X_i)$. For instance, under the assumption that $\mathrm{Var}\,(\epsilon_i|X_i) = \sigma^2 X_i$, one may estimate σ^2 as

$$\frac{\sum_{i=1}^n R_i \left(Y_i - \hat{\beta} X_i\right)^2 / X_i}{\left(\sum_{i=1}^n R_i\right) - 1}.$$

This is the estimator for the residual variance upon using weighted least squares, with weights $1/X_i$ ($X_i > 0$).

A.2 Variance of $\hat{T}^{ols}$

To assess the variance of $\hat{T}^{ols}$, we proceed in a similar way as in section 3.1.3. We first evaluate the imprecision of $\hat{\boldsymbol{\beta}}$ in terms of how much it differs from

$$\boldsymbol{\beta} = \left(\sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1} \sum_{i=1}^n \boldsymbol{X}_i Y_i,$$

the value that $\hat{\boldsymbol{\beta}}$ takes when the data of all companies have come available:

$$\begin{aligned}
\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} &= \left(\sum_{i=1}^n R_i \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1} \sum_{i=1}^n R_i \boldsymbol{X}_i Y_i - \left(\sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1} \sum_{i=1}^n \boldsymbol{X}_i Y_i \\
&= \left(\sum_{i=1}^n R_i \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1} \left[\sum_{i=1}^n R_i \boldsymbol{X}_i Y_i - \underbrace{\left(\sum_{i=1}^n R_i \boldsymbol{X}_i \boldsymbol{X}_i'\right)\left(\sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1}}_{\equiv \boldsymbol{r}} \sum_{i=1}^n \boldsymbol{X}_i Y_i\right] \\
&= \left(\sum_{i=1}^n R_i \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1} \left[\sum_{i=1}^n R_i \boldsymbol{X}_i Y_i - \boldsymbol{r} \sum_{i=1}^n \boldsymbol{X}_i Y_i\right] \\
&= \left(\sum_{i=1}^n R_i \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1} \left[\sum_{i=1}^n (R_i \boldsymbol{I}_{p\times p} - \boldsymbol{r}) \boldsymbol{X}_i (\boldsymbol{X}_i' \boldsymbol{\beta} + \epsilon_i)\right] \\
&= \left(\underbrace{\sum_{i=1}^n R_i \boldsymbol{X}_i \boldsymbol{X}_i'}_{\equiv A}\right)^{-1} \sum_{i=1}^n \underbrace{(R_i \boldsymbol{I}_{p\times p} - \boldsymbol{r})}_{\equiv B_i} \boldsymbol{X}_i \epsilon_i, \\
&= \boldsymbol{A}^{-1} \sum_{i=1}^n \boldsymbol{B}_i \boldsymbol{X}_i \epsilon_i,
\end{aligned}$$

where $\boldsymbol{I}_{p\times p}$ is the $p\times p$ identity matrix, $\boldsymbol{r} = \left(\sum_{i=1}^n R_i \boldsymbol{X}_i \boldsymbol{X}_i'\right)\left(\sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1}$, $\boldsymbol{A} = \sum_{i=1}^n R_i \boldsymbol{X}_i \boldsymbol{X}_i'$ and $\boldsymbol{B}_i = (R_i \boldsymbol{I}_{p\times p} - \boldsymbol{r})$. Here, the term involving $\boldsymbol{r}$ represents a finite population correction; it would reduce to zero if an infinite population were considered. In the last step we used the fact that $\left(\sum_{i=1}^n R_i \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1}\left[\sum_{i=1}^n (R_i \boldsymbol{I}_{p\times p} - \boldsymbol{r})\boldsymbol{X}_i(\boldsymbol{X}_i'\boldsymbol{\beta})\right] = \left(\sum_{i=1}^n R_i \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1}\left(\sum_{i=1}^n R_i \boldsymbol{X}_i \boldsymbol{X}_i'\right)\boldsymbol{\beta} - \left(\sum_{i=1}^n R_i \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1}\boldsymbol{r}\left(\sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{X}_i'\right)\boldsymbol{\beta}$ and after substitution of $\boldsymbol{r}$ this becomes zero.

Let $\sum_{i=1}^n (1-R_i)\boldsymbol{X}_i / \sum_{i=1}^n (1-R_i) \equiv \bar{\boldsymbol{X}}_{(0)}$. The variance of $\hat{T}$ (assuming it is unbiased) then equals

$$E\left(\left[\sum_{i=1}^n (1-R_i)\left\{(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\bar{\boldsymbol{X}}_{(0)} + (\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{(0)}) - \epsilon_i\right\}\right]^2 |\{\boldsymbol{X}_i, R_i; \forall i\}\right).$$

This equals

$$\begin{aligned}
&E\left(\left[\sum_{i=1}^n (1-R_i)(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\bar{\boldsymbol{X}}_{(0)}\right]^2 |\{\boldsymbol{X}_i, R_i; \forall i\}\right) \\
&+E\left(\left[\sum_{i=1}^n (1-R_i)\left\{(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{(0)}) - \epsilon_i\right\}\right]^2 |\{\boldsymbol{X}_i, R_i; \forall i\}\right) \\
&+E\left[\left\{\sum_{i=1}^n (1-R_i)(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\bar{\boldsymbol{X}}_{(0)} \sum_{j=1}^n (1-R_j)\left((\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'(\boldsymbol{X}_j - \bar{\boldsymbol{X}}_{(0)}) - \epsilon_j\right)\right\} |\{\boldsymbol{X}_i, R_i; \forall i\}\right] \\
&= E\left(\left[\sum_{i=1}^n (1-R_i)(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\bar{\boldsymbol{X}}_{(0)}\right]^2 |\{\boldsymbol{X}_i, R_i; \forall i\}\right) + \sum_{i=1}^n (1-R_i) E\left(\epsilon_i^2 |\{\boldsymbol{X}_i, R_i; \forall i\}\right) \\
&-E\left[\left\{\sum_{i=1}^n (1-R_i)(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\bar{\boldsymbol{X}}_{(0)} \sum_{j=1}^n (1-R_j)\epsilon_j\right\} |\{\boldsymbol{X}_i, R_i; \forall i\}\right]
\end{aligned}$$

From the formula supra, it follows that the first term equals

$$\left\{\sum_{i=1}^n (1-R_i)\boldsymbol{X}_i\right\}' \boldsymbol{A}^{-1}\left(\sum_{j=1}^n \mathrm{Var}\left(\epsilon_j|\boldsymbol{X}_j\right)\boldsymbol{B}_j \boldsymbol{X}_j \boldsymbol{X}_j' \boldsymbol{B}_j'\right)\boldsymbol{A}^{-1}\left\{\sum_{i=1}^n (1-R_i)\boldsymbol{X}_i\right\}.$$

The second term equals

$$\sum_{i=1}^n (1-R_i)\mathrm{Var}(\epsilon_i|X_i). \tag{23}$$

The third term equals

$$-\left\{\sum_{i=1}^{n}(1-R_i)\boldsymbol{X}_i\right\}'\boldsymbol{A}^{-1}\sum_{j=1}^{n}\mathrm{Var}\left(\epsilon_j|\boldsymbol{X}_j\right)(1-R_j)\boldsymbol{B}_j\boldsymbol{X}_j \tag{24}$$

$$=\left\{\sum_{i=1}^{n}(1-R_i)\boldsymbol{X}_i\right\}'\boldsymbol{A}^{-1}\sum_{j=1}^{n}\mathrm{Var}\left(\epsilon_j|\boldsymbol{X}_j\right)(1-R_j)\boldsymbol{r}\boldsymbol{X}_j \tag{25}$$

Assuming homoscedasticity, the sum of these 3 terms reduces to

$$\sigma^2\left[\sum_{i=1}^{n}(1-R_i)\left\{1+\boldsymbol{X}_i'\boldsymbol{A}^{-1}\boldsymbol{r}\left\{\sum_{i=1}^{n}(1-R_i)\boldsymbol{X}_i\right\}\right\}\right]$$
$$+\sigma^2\left\{\sum_{i=1}^{n}(1-R_i)\boldsymbol{X}_i\right\}'\boldsymbol{A}^{-1}\left(\sum_{j=1}^{n}\boldsymbol{B}_j\boldsymbol{X}_j\boldsymbol{X}_j'\boldsymbol{B}_j'\right)\boldsymbol{A}^{-1}\left\{\sum_{i=1}^{n}(1-R_i)\boldsymbol{X}_i\right\}.$$

Here, $\mathrm{Var}(\epsilon_i|\boldsymbol{X}_i)$ can be estimated as the residual variance

$$\frac{\sum_{i=1}^{n}R_i(Y_i-\hat{\boldsymbol{\beta}}'\boldsymbol{X}_i)^2}{(\sum_{i=1}^{n}R_i)-p}, \tag{26}$$

with p the dimension of $\boldsymbol{\beta}$.

B Approximation of the uncertainty of the OLS estimator.

Remember that the analytical relative uncertainty of the OLS-estimator was defined as:

$$\mu_a^{(ols)}(y;u) = \frac{4\sqrt{E\left((\hat{T}^{(ols)}(y;u) - T^{ols}(y))^2\right)}}{\hat{T}^{ols}(y)}$$

where $E\left((\hat{T}^{(ols)}(y;u) - T^{ols}(y))^2\right)$ is the imprecision of the estimator as given by equation (16). The latter formula differs from the much simpler, but approximate, equation (14) by a "finite population correction". Analyzing equation (14) it can be seen that it is equal to the number of companies for which the value added is unknown at u times the residual variance. An estimate for the latter is part of the standard output of a linear regression. Table 5 illustrates the impact of using the much simpler but approximative formula (14) instead of the more precise but more complex formula (16) on the relative uncertainty; the column $\mu_a/2$ uses the above definition for μ_a, using equation (16) for the imprecision of the estimator (i.e. the expression under de root), column $\widetilde{\mu}_a/2$ is similar but it uses the (simpler but) approximate equation (14).

Table 5: Approximated uncertainty of the OLS estimator for year 2015 of the Ports study

u	size	$\hat{T}^{ols}$	$\mu_a/2(\%)$	$\widetilde{\mu}_a/2(\%)$
2016-08-15	large	27 249.60	±1.27%	±1.13%
2016-08-31		27 238.33	±0.44%	±0.42%
2016-10-31		27 259.97	±0.31%	±0.3%
2016-08-15	small	800.31	±1.89%	±1.55%
2016-08-31		794.51	±0.95%	±0.89%
2016-10-31		787.77	±0.52%	±0.51%

The results in this table are the same as in tables 1 and 2 for the OLS-estimator, except for the last column $\widetilde{\mu}_a/2$, where the uncertainty was approximated using the equation (14). The differences between the exact formula and the approximation are small and, as expected, become smaller when more observations become available.

NATIONAL BANK OF BELGIUM - WORKING PAPERS SERIES

The Working Papers are available on the website of the Bank: http://www.nbb.be.

317. “An estimated two-country EA-US model with limited exchange rate pass-through”, by G. de Walque, Ph. Jeanfils, T. Lejeune, Y. Rychalovska and R. Wouters, *Research series,* March 2017.
318. Using bank loans as collateral in Europe: The role of liquidity and funding purposes”, by F. Koulischer and P. Van Roy, *Research series*, April 2017.
319. “The impact of service and goods offshoring on employment: Firm-level evidence”, by C. Ornaghi, I. Van Beveren and S. Vanormelingen, *Research series*, May 2017.
320. “On the estimation of panel fiscal reaction functions: Heterogeneity or fiscal fatigue?”, by G. Everaert and S. Jansen, *Research series*, June 2017.
321. “Economic importance of the Belgian ports: Flemish maritime ports, Liège port complex and the port of Brussels - Report 2015”, by C. Mathys, *Document series*, June 2017.
322. “Foreign banks as shock absorbers in the financial crisis?”, by G. Barboni, *Research series*, June 2017.
323. “The IMF and precautionary lending: An empirical evaluation of the selectivity and effectiveness of the flexible credit line”, by D. Essers and S. Ide, *Research series*, June 2017.
324. “Economic importance of air transport and airport activities in Belgium – Report 2015”, by S. Vennix, *Document series*, July 2017.
325. “Economic importance of the logistics sector in Belgium”, by H. De Doncker, *Document series*, July 2017.
326. “Identifying the provisioning policies of Belgian banks”, by E. Arbak, *Research series*, July 2017.
327. “The impact of the mortgage interest and capital deduction scheme on the Belgian mortgage market”, by A. Hoebeeck and K. Inghelbrecht, *Research series,* September 2017.
328. “Firm heterogeneity and aggregate business services exports: Micro evidence from Belgium, France, Germany and Spain”, by A. Ariu, E. Biewen, S. Blank, G. Gaulier, M.J. González, Ph. Meinen, D. Mirza, C. Martín and P. Tello, *Research series*, September 2017.
329. “The interconnections between services and goods trade at the firm-level”, by A. Ariu, H. Breinlichz, G. Corcosx, G. Mion, *Research series,* October 2017.
330. “Why do manufacturing firms produce services? Evidence for the servitization paradox in Belgium”, by P. Blanchard, C. Fuss and C. Mathieu, *Research series,* November 2017.
331. “Nowcasting real economic activity in the euro area: Assessing the impact of qualitative surveys”, by R. Basselier, D. de Antonio Liedo and G. Langenus, *Research series,* December 2017.
332. “Pockets of risk in the Belgian mortgage market: Evidence from the Household Finance and Consumption Survey (HFCS)”, by Ph. Du Caju, *Research series,* December 2017.
333. “The employment consequences of SMEs’ credit constraints in the wake of the great recession” by D. Cornille, F. Rycx and I. Tojerow, *Research series*, December 2017.
334. “Exchange rate movements, firm-level exports and heterogeneity”, by A. Berthou and E. Dhyne, *Research series*, January 2018.

335 “Nonparametric identification of unobserved technological heterogeneity in production”, by L. Cherchye, T. Demuynck, B. De Rock and M. Verschelde, *Research series*, February 2018.

336 “Compositional changes in aggregate productivity in an era of globalisation and financial crisis”, by C. Fuss and A. Theodorakopoulos, *Research series*, February 2018.

337. “Decomposing firm-product appeal: How important is consumer taste?”, by B. Y. Aw, Y. Lee and H. Vandenbussche, *Research series*, March 2018.

338 “Sensitivity of credit risk stress test results: Modelling issues with an application to Belgium”, by P. Van Roy, S. Ferrari and C. Vespro, *Research series*, March 2018.

339. “Paul van Zeeland and the first decade of the US Federal Reserve System: The analysis from a European central banker who was a student of Kemmerer”, by I. Maes and R. Gomez Betancourt, *Research series*, March 2018.
340. “One way to the top: How services boost the demand for goods”, by A. Ariu, F. Mayneris and M. Parenti, *Research series*, March 2018.

341 “Alexandre Lamfalussy and the monetary policy debates among central bankers during the Great Inflation”, by I. Maes and P. Clement, *Research series*, April 2018.

342. “The economic importance of the Belgian ports: Flemish maritime ports, Liège port complex and the port of Brussels – Report 2016”, by F. Coppens, C. Mathys, J.-P. Merckx, P. Ringoot and M. Van Kerckhoven, *Document series*, April 2018.
343. “The unemployment impact of product and labour market regulation: Evidence from European countries”, by C. Piton, *Research series,* June 2018.
344. “Trade and domestic production networks”, by F. Tintelnot, A. Ken Kikkawa, M. Mogstad, E. Dhyne, *Research series*, September 2018.

345. “Review essay: Central banking through the centuries”, by I. Maes, *Research series,* October 2018.

346. “IT and productivity: A firm level analysis”, by E. Dhyne, J. Konings, J. Van den Bosch, S. Vanormelingen, *Research series*, October 2018.

347. “Identifying credit supply shocks with bank-firm data: methods and applications”, by H. Degryse, O. De Jonghe, S. Jakovljević, Klaas Mulier, Glenn Schepens, *Research series*, October 2018.

348. “Can inflation expectations in business or consumer surveys improve inflation forecasts?”, by R. Basselier, D. de Antonio Liedo, J. Jonckheere and G. Langenus, *Research series*, October 2018.

349. “Quantile-based inflation risk models”, by E. Ghysels, L. Iania and J. Striaukas, *Research series*, October 2018.

350. “International food commodity prices and missing (dis)inflation in the euro area”, by G. Peersman, *Research series*, October 2018.

351. “Pipeline pressures and sectoral inflation dynamics”, by F. Smets, J. Tielens and J. Van Hove, *Research series*, October 2018.

352. “Price updating in production networks”, by C. Duprez and G. Magerman, *Research series*, October 2018.

353. “Dominant currencies. How firms choose currency invoicing and why it matters”, by M. Amiti, O. Itskhoki and J. Konings, *Research series*, October 2018.

354. “Endogenous forward guidance”, by B. Chafwehé, R. Oikonomou, R. Priftis and L. Vogel, *Research series*, October 2018.

355. “Is euro area lowflation here to stay? Insights from a time-varying parameter model with survey data”, by A. Stevens and J. Wauters, *Research series*, October 2018.

356. “A price index with variable mark-ups and changing variety”, by T. Demuynck and M. Parenti, *Research series*, October 2018.

357. “Markup and price dynamics: Linking micro to macro”, by J. De Loecker, C. Fuss and J. Van Biesebroeck, *Research series*, October 2018.

358. “Productivity, wages and profits: Does firms’ position in the value chain matter?”, by B. Mahy, F. Rycx, G. Vermeylen and M. Volral, *Research series*, October 2018.

359. “Upstreamness, social upgrading and gender: Equal benefits for all?”, by N. Gagliardi, B. Mahy and F. Rycx, *Research series*, December 2018.

360. “A macro-financial analysis of the corporate bond market”, by H. Dewachter, L. Iania, W. Lemke and M. Lyrio, *Research series*, December 2018.

361. “Some borrowers are more equal than others: Bank funding shocks and credit reallocation”, by O. De Jonghe, H. Dewachter, K. Mulier, S. Ongena and G. Schepens, *Research series*, December 2018.

362. “The origins of firm heterogeneity: A production network approach”, by A. B. Bernard, E. Dhyne, G. Magerman, K. Manova and A. Moxnes, *Research series*, January 2019.

363. “Imperfect competition in firm-to-firm trade”, by A. Ken Kikkawa, G. Magerman and E. Dhyne, *Research series*, January 2019.

364. “Forward guidance with preferences over safe assets”, by A. Rannenberg, *Research series*, January 2019.

365. “The distinct effects of information technologies and communication technologies on the age-skill composition of labour demand”, by S. Blanas, *Research series*, January 2019.

366. “A survey of the long-term impact of Brexit on the UK and the EU27 economies”, by P. Bisciari, *Document series*, January 2019.

367. “A macroeconnomic model with heterogeneous and financially-constrained intermediaries”, by Th. Lejeune and R. Wouters, *Research series*, February 2019.

368. “The economic importance of the Belgian ports: Flemish maritime ports, Liège port complex and the port of Brussels – Report 2017”, by E. Gueli, P. Ringoot and M. Van Kerckhoven, *Document series*, March 2019.

369. “Does banks’ systemic importance affect their capital structure and balance sheet adjustment processes?”, by Y. Bakkar, O. De Jonghe and A. Tarazi, *Research series*, March 2019.

370 “A model for international spillovers to emerging markets”, R. Houssa, J. Mohimont and C. Otrok, *Research series*, April 2019.

371. “Estimation methods for computing a branch’s total value added from incomplete annual accounting data”, S. Vansteelandt, F. Coppens, D. Reynders, M. Vackier and L. Van Belle, *Research series*, April 2019.

National Bank of Belgium
Limited liability company
RLP Brussels – Company's number: 0203.201.340
Registered office: boulevard de Berlaimont 14 – BE-1000 Brussels
www.nbb.be

Editor

Pierre Wunsch
Governor of the National Bank of Belgium

Layout: Analysis and Research Group
Cover: NBB CM – Prepress & Image

Published in April 2019