Grether, Jean-Marie; Tissot-Daguette, Benjamin

Working Paper

Multiple imputation techniques: An application to Swiss value-added data

IRENE Working Paper, No. 21-09

Provided in Cooperation with:
Institute of Economic Research (IRENE), University of Neuchâtel

Suggested Citation: Grether, Jean-Marie; Tissot-Daguette, Benjamin (2021) : Multiple imputation techniques: An application to Swiss value-added data, IRENE Working Paper, No. 21-09, University of Neuchâtel, Institute of Economic Research (IRENE), Neuchâtel

This Version is available at:
https://hdl.handle.net/10419/265176

UNIVERSITÉ DE NEUCHÂTEL

Institut de recherches économiques

IRENE Working Paper 21-09

Multiple Imputation Techniques: An Application to Swiss Value-Added Data

Jean-Marie Grether, Benjamin Tissot-Daguette

Multiple Imputation Techniques : An Application to Swiss Value-Added Data

Jean-Marie Grether, University of Neuchâtel
Benjamin Tissot-Daguette, Federal Department of Finance
September 13, 2021

Abstract

We use imputation techniques and combine official data sources to address the various shortcomings affecting the analysis of value-added data at the level of production units in Switzerland. The new *ad hoc* databases that emerge include consistent information on value added and employment at the level of geographically localized pseudo-firms over the 2011-2015 period. Our preferred sample is obtained through multiple imputation techniques, includes 18'000 pseudo-firms per year, covers two-third of Swiss municipalities and is suitable to address productivity issues at the microeconomic level.

JEL Classification: R11, R32
Keywords: Swiss firms, value-added, multiple imputation, spatial distribution

1 Introduction

A detailed analysis of productivity requires disaggregated data, preferably at the level of the establishment. However, complete and reliable data may be difficult to access because of three problems. First, confidentiality issues may limit the diffusion of firm level information. Second, available surveys may only include a selection of small firms, which are not kept for long in the sample. Third, some variables (crucially value added) may only be available at the firm level, not at the establishment level. These problems are particularly acute in Switzerland, which lags behind most OECD countries in terms of data accessibility at the firm level. Building on micro data from the Federal Statistical Office (FSO), and relying on several imputation techniques, we present here a novel way to address these problems. Our original treatment leads to a set of *ad hoc* value-added databases over the 2011-2015 period at an unprecedented level of granularity for Switzerland. [1]

The official data source for value-added which we use is the *Wertschöpftungsstatistik* (WS) of the FSO. [2] It is a yearly survey of around 22'000 firms which presents incomplete coverage for two major sets of reasons. On the one hand, the sampling frame excludes

1. We thank Sam Banatte, Nicole Mathys, Tobias Müller and Claudio Sfreddo for their very helpful recommendations, Markus Daeppen and Stephen Sonntag from the FSO for their data support, and participants at the Swiss Society for Economics and Statistics and the PhD seminar of the University of Neuchâtel in June 2019 for their comments. The usual disclaimers apply.

2. See https://www.bfs.admin.ch/bfs/en/home/statistics/industry-services/surveys/ws.assetdetail.926303.html

firms with less than 3 employees as well as several sectors (primary, banks and insurance companies, public administration and human health activities). The answer rate is slightly less than 66%, leading to a net sample of around 14'000 firms. Except from keeping them in mind, there is not much we can do regarding this first set of shortcomings.

On the other hand, some firms drop from the sample either because of occasional non-response or because they are not sampled anymore, as each year 20% of small firms (less than 50 employees) are replaced. Moreover, the survey is conducted at the firm level, not at the plant level. Thus, all production of multi-plant firms is reported at a single location (headquarters), hindering a proper geographical analysis. We propose a novel method to control for this second set of limitations. Relying on additional data sources and applying multiple imputation techniques, we estimate missing values of dropout firms and redistribute the value-added among plants of multi-plant firms. This leads to a set of *enlarged* databases for value-added at the plant level.

In addition to missing data issues, we also have to respect confidentiality rules. To do so, we have to aggregate these data at the level of the legal form, the municipality and the NOGA-4 industrial sector. [3] This is what we call a "pseudo-firm" in the present paper. It corresponds to the finest disaggregation level at which value-added data is made available in the final novel databases.

Section 2 presents the imputation techniques which are used to address firms' dropout and multi-plant firms. They are similar in design, relying mostly on hypotheses regarding productivity growth and on employment figures provided by the *Statistique Structurelle des Entreprises* (STATENT) database, a FSO business statistics mainly based on data from the Old-age and survivor's insurance (OASI) registers and available either at the firm or at the plant level. [4] Section 3 presents the final re-aggregation process at the level of the pseudo-firm. Section 4 presents an overview of the constructed datasets and a comparison with National Accounts figures. Section 5 concludes.

2 Imputation strategy

After a brief introduction to multiple imputation techniques, we characterize the missing data pattern and then provide a detailed presentation of the imputation procedures followed to complete the value-added data.

3. As of 2014, there are 2352 municipalities in Switzerland. NOGA 4-digits is a 615 levels industry classification and the FSO divides firms into 23 different legal forms.

4. https://www.bfs.admin.ch/bfs/en/home/services/geostat/swiss-federal-statistics-geodata/business-employment/structural-business-statistics-statent-from-2011-onwards.html

2.1 Dealing with missing values through (multiple) imputation

Missing data are a prevalent source of concern for the empirical researcher. Broadly speaking, there are three ways to address this concern, each one of them being considered in the present work (see Schafer and Graham (2002) for a technical discussion). The first obvious way of dealing with missing data is to keep only non-missing cases in a *restricted sample*. Such an option is easily implemented but will bias the results of the analysis if there are structural differences between the observed and the missing data. The second option is to rely on additional data sources (employment in our case) to impute values on the missing variables by using simple rules e.g. proportional or mean-preserving attributions. This type of *naive* imputation methods enlarges the sample size but fails to take properly into account the potential above-mentioned structural differences. The third option also relies on additional data but exploits them more systematically using statistical inference techniques. This third method is the only one that can capture at least part of the structural differences between the missing and the observed data.

A particularly flexible case of the third category is the *multiple imputation* procedure, which repeats the imputation routine m times, leading to a "distribution" for the missing value rather than a point estimate. This allows to take into account the uncertainty around its formation (see Rubin (1987)). Standard procedures can then be applied on the resulting m complete databases. There are several ways of implementing multiple imputation, which mostly depend on the missing data pattern. In our case, as described below, we will follow a simple monotone regression framework. [5]

2.2 *Restricted* vs. *Enlarged* samples

As mentioned above, our newly created value-added databases result from the combination of WS sample results and STATENT data at the firm level for five years (2011-2015). The WS sample survey covers around 22'000 firms. It is drawn from a sample frame of 170'000 firms with at least three employees in the secondary and tertiary sectors (except bank and insurance companies, public administration and human health activities). Large and medium-sized firms (50 employees or above) are all present in the survey. For small firms (between 3 and 49 employees), the sample is stratified according to 2-digit NOGA sectors and size categories based on the number of employees. Small firms are only kept five years in the sample, which means that every year, 20% of them are renewed. The response rate is around 90% for large firms, 70% for medium-sized ones and 55% for small ones.

5. Yuan (1994) provides a detailed presentation. He also identifies three steps in any multiple imputation procedure. First, m estimates are formed for each missing value. Then any required analysis can be applied to each of the m datasets. Finally, the results are combined in a valid statistical way (Rubin (1987)).

We first match the two databases at the firm level and compute value-added as the difference between gross output and intermediate consumption. Then we eliminate from the sample all firms which are not present in the WS survey, or never respond, or exhibit a zero or negative value-added at any given year (this to avoid unrealistic estimates in the multiple imputation procedure). This leads to a temporary sample of approximately 14'000 observations per year, among which around 55% (i.e. 7'700) are small firms.

This intermediate sample is still unsatisfactory for analysis because of missing data due to the rollover of small firms or non-responses, and because the value-added of multi-plant firms is concentrated at the headquarters' location. A number of steps are necessary to obtain more suitable databases. These steps are stylized in figure 1.

<table>
<tr><td colspan="5">Across the 2011-2015 period, is the firm...</td></tr>
<tr><td>...suitable for inclusion?[1)]</td><td>...always sampled when it is active?[2)]</td><td>...always responding when it is sampled?</td><td>...single plant?[3)]</td><td>...multi-plant?</td></tr>
<tr><td>YES</td><td>YES</td><td>YES</td><td colspan="2">1. Restricted sample</td></tr>
<tr><td></td><td>YES</td><td>NO</td><td colspan="2" rowspan="2">2a. Non-response enlargement
2b. Rollover enlargement</td></tr>
<tr><td></td><td>NO</td><td>YES/NO</td></tr>
<tr><td>NO</td><td></td><td></td><td colspan="2">Not considered</td></tr>
</table>

[1)] for at least one year, the firm is sampled and responding; it is not reporting negative value-added.
[2)] a firm is considered active when it reports positive employment in the STATENT database.
[3)] also includes multi-plant firms which always remain active within the same pseudo-firm.

Figure 1 – Sample selection and enlargement.

A first step is to limit the analysis to those firms that are not replaced, are always responding, and remain single plants (or multi-plant but always active within the same pseudo-firm i.e. the same combination of legal form, municipality and four digit sector). This corresponds to the *restricted sample* represented by the top left cell of the shaded area of figure 1. This sample is biased towards medium to large firms (unaffected by the rollover problem and responding more than small firms) and excludes multi-plant firms by definition. The number of firms drops to less than 4'000 per year.

Starting from this minimum benchmark, two *enlarged samples* are proposed, both of them relying on imputation techniques. The second step allows to enlarging the database using employment (and other) data to infer missing values due to non-response and firms' rollover. This corresponds to the intermediate left cell of figure 1. At that stage, small

firms are better represented, but multi-plant firms are still absent. The third and final step corresponds to re-distributing value-added across the various units of multi-plant firms, represented by the shaded right cell of figure 1. The consecutive increase in observations depends on both the type of enlargement and the imputation method, as sequentially discussed below.

2.3 Enlargement types

2.3.1 Non-response and rollover enlargement

Among eligible firms, and for certain years, some of them do not answer to the WS questionnaire while others disappear from the sample due to the yearly rollover of a quintile of small firms. Imputing value-added to these missing cases enlarges the sample.

Figure 2 illustrates this enlargement effect in the pure rollover case. Each year, one quintile disappears - the "old" quintile - and another one appears - the "new" quintile -. Forward imputation of the old quintiles and backward imputation of the new quintiles increases the number of observations. In the final sample 4/9th (around 45%) of observations have been imputed.

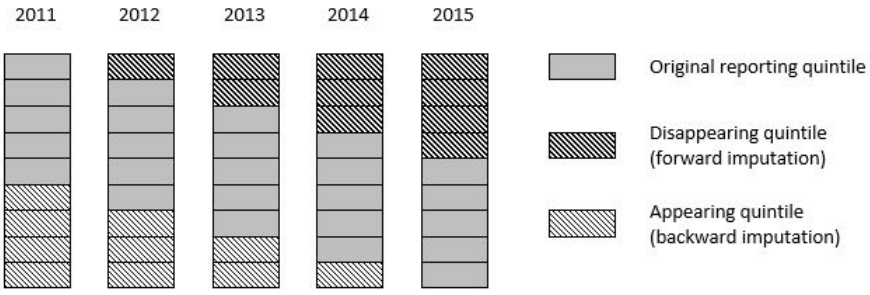


Figure 2 – Stylized firm-rollover imputation.

This is a lower bound given that our stylized reasoning so far abstracts from the non-response problem. In reality, accounting for both non-response and rollover, the share of imputed observations turns out to be 54% for small plants (less than 50 employees), 16% for medium firms (between 50 and 499 employees), and 5% for large firms (more than 499 employees).

2.3.2 Multi-plant enlargement

So far, the enlargement process has followed the WS survey definition of the reporting unit, which is the firm, not the plant. For multi-plant firms that are active

across several municipalities and/or 4-digit sectors, this is a problem (recall from the introduction that plant level data will have to be re-aggregated anyway at the level of the pseudo-firm for confidentiality reasons). More precisely, when using firm-level data, the presence of multi-plant firms leads to mixing together different 4-digits activities into a single 4-digit one and to mixing together plants in different municipalities into a single one (see Figure 3). These biases are too severe to be acceptable in any study analyzing the interconnections between performance and localization of productive units.

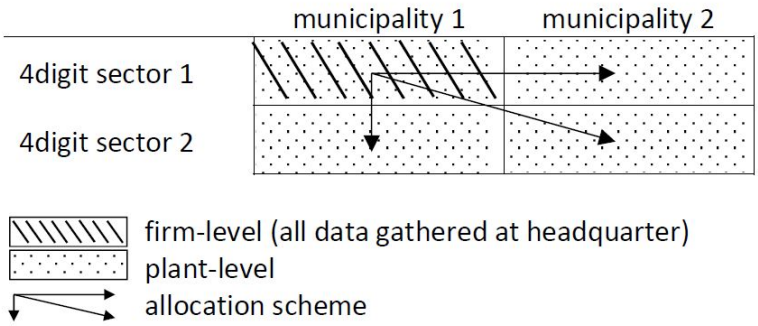


Figure 3 – Re-distribution of value-added within multi-plants.

Thus, the second enlargement consists of re-distributing the value-added among the different units of multi-plant firms. To do this, we have to estimate value-added shares or *allocation schemes* through an imputation process which is again mostly based on employment data.

2.4 Imputation methods

To enlarge the sample according to the two above-mentioned procedures, we rely on two imputation methods.

2.4.1 Naive imputation

The first method, or naive imputation, consists of using employment (full time equivalent) shares available from the STATENT data. In the multi-plant enlargement case, the "allocation scheme" is simply the share of each plant in total employment of the multi-plant firm. For the other enlargement cases, we combine STATENT shares with aggregate value added data from the National Accounts. More specifically, for every firm that is affected by rollover or non-response in a given year, we proceed as follows:

(i) A first set of value-added estimates at the firm level is obtained by assuming that value-added remains proportional to full time employment equivalents from the given to the closest year. We keep 2011 figures unchanged but adjust 2012-2015 figures in the rest of the procedure.

(ii) For 2012, we sum up value-added estimates across all imputed firms of a given sector (k). This gives us a first total for value-added of the missing value firms, denoted by TV_k^1.

(iii) From the National Accounts, we obtain the 2011-2012 growth rate of labor productivity at the sector level, denoted by γ_k^n, and we posit it also applies at the level of the completed sample. This, combined with aggregated STATENT and WS data, allows to compute a second figure for the total value-added of the missing value firms, denoted by TV_k^2.[6]

(iv) To make 2012 firm-level data consistent with national account figures, we multiply all value-added estimates obtained at step i at the firm level by the TV_k^2/TV_k^1 ratio.

(v) Steps ii-iv are repeated for the remaining three consecutive years, *mutatis mutandis*.

This adjusted proportionality rule ensures that the growth rate of labor productivity in the constructed sample is consistent with the reported growth rate from the National Accounts.

2.4.2 Multiple imputation

The second method consists of applying a multiple imputation procedure based on Rubin (1987). For the implementation, we use the multiple imputation *PROC MI* routine proposed by SAS (SAS Institute Inc. (2015)), using a monotone regression frame-

6. The demonstration is as follows. Let us denote value added by V, labor (full time equivalent) by L, the sector by k, the national level by n, the sample level by s, and any growth rate by a hat i.e ($\hat{}$). Assuming identical labor productivity growth rates at the sample and national level leads to $[(1+\hat{V}_k^n)/(1+\hat{L}_k^n)]-1 = [(1+\hat{V}_k^s)/(1+\hat{L}_k^s)]-1$. In the previous expression, the sample value added growth rate ($\hat{V}_k^s$) can be replaced by a weighted average of the missing firms value added growth rate ($V_{k,mv}^s$) and the incumbent firms value added growth rate ($V_{k,ic}^s$), i.e. $\hat{V}_k^s = \theta_k^s \cdot \hat{V}_{k,ic}^s + (1-\theta_k^s) \cdot \hat{V}_{k,mv}^s$, where $\theta_k^s = V_{k,ic}^s/V_k^s$. After simplification we obtain:

$$\hat{V}_{k,mv}^s = \frac{(1+\gamma_k^n)(1+\hat{L}_k^s) - (1+\theta_k^s \hat{V}_{k,ic}^s)}{1-\theta_k^s}$$

Applying the above growth rate (bounded between -50% and +50% as a feasibility constraint) to the firms with missing values and summing up leads to the second estimated total for the sectoral value added, TV_k^2.

work.[7] Whatever the enlargement type, we estimate the value added of the corresponding unit (the firm for the non-response and rollover cases, the plant for the multi-plant case) performing the following steps:

(i) Regression of the natural logarithm of value added (V) on the natural logarithm of employment (full time equivalent, L) and a set of categorical variables that includes year, 3-digits sector, district and legal form.

$$ln(V) = \beta_0 + \beta_1 \cdot ln(L) + \alpha_j + \gamma_i + y_t + \omega_f$$

Where α_i, γ_i, y_t and ω_f are fixed effects capturing, respectively, the sectoral effect of belonging to 3-digits industry j, district i, year t and legal form f.[8]

We control for stand-alone cases where there is no reported observation because there has been a change of the sector, region, or legal form during the sample period. We also control for dummy outliers. To do so, we run a trial imputation (only two runs) without variables log-transformation and we define as outliers those specific industries, districts or legal forms with an estimated coefficient that deviates by more than two standard deviations from their classes' means. All firms that correspond to the identified stand-alone or outlier cases are dropped from subsequent analysis.

(ii) New parameters $\tilde{\beta} = (\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\alpha}_j, \tilde{\gamma}_i, \tilde{y}_t, \tilde{\omega}_f)$ and variance $\tilde{\sigma}^2$ are simulated from the estimated parameters of the above regression, $(\hat{\beta}_0, \hat{\beta}_1, \hat{\alpha}_j, \hat{\gamma}_i, \hat{y}_t, \hat{\omega}_f)$, and estimated variance $\hat{\sigma}^2$.

$$\tilde{\sigma}^2 = \hat{\sigma}^2 \, \frac{n-k-1}{g}$$

where n is the number of non-missing observations, k the number of explanatory variables and $g \sim \chi^2_{n-k-1}$.

$$\tilde{\beta} = \hat{\beta} + \tilde{\sigma} W_h' Z$$

where $W = X'X$ and W_h is obtained by the Cholesky decomposition of W. Z is a vector of $k+1$ normalized random variables.

7. Note that unfortunately, the alternative SAS procedure that imputes the closest observed values (predictive mean matching) is too time-consuming to be implementable in our case.

8. The original specification was run for each monetary variable separately (i.e. gross output and intermediate consumption) and included more than 2800 dummies, in particular all municipalities and 4-digit sectors. Many dummies turned out non significant but were nevertheless used in the imputation procedure. As many municipalities had too few observations, this led to unrealistically low or high figures for imputed value-added. Therefore, a unique specification for value-added was selected, with substantially less dummies (slightly more than 400) by replacing municipalities with districts and 4-digit by 3-digit sectors.

(iii) Then predicted values ($\tilde{V}$) are formed using these new coefficients. For each missing observation, belonging to industry j, located in municipality i and year t:

$$\tilde{V} = exp[\tilde{\beta}_0 + \tilde{\beta}_1 \cdot ln(L) + \tilde{\alpha}_j + \tilde{\gamma}_i + \tilde{y}_t + \tilde{\omega}_f + z\tilde{\sigma}]$$

where z is a normal standard deviation.

(iv) The procedure is repeated m times with m being the number of imputations. The efficiency of the estimators depends on the number of imputations. For a relative high fraction of missing information, Graham et al. (2007) recommend a high number of imputations, up to 40 if half of the observations are missing . To balance estimator efficiency and computing time, we have selected 20 imputations, a reasonable number according to Graham et al. (2007) when 30% of observations are missing.

2.5 Combining imputation procedures

We proceed by implementing the various imputations techniques for the non-response and rollover enlargements. This leads to $m + 2$ firm-level databases (including the restricted sample and the one obtained by the naive imputation method), corresponding to the first column of figure 4. These databases are then converted into $(n + 1)(m + 2)$ plant-level databases by the multi-plant enlargement (remaining columns of figure 4), applying the allocation schemes obtained through the naive or multiple imputation techniques. [9]

The rounded average number of firms per year is indicated between parentheses in Figure 4. As expected, the *restricted* sample, which is limited to single-plant or non-problematic multi-plant firms reporting positive value-added every year they are active, is small (3'600 firms per year) and biased towards large firms. Combining *naive imputation* techniques for firms' non-response, rollover and multi-plant firms maximizes the number of firms in the sample (24'000 per year). Using *multiple imputation* techniques instead still increases the number of firms vis-à-vis the restricted sample but to a lower extent (18'000 firms per year), due to the elimination of outliers and stand-alone cases.

9. See figure A1 in the Appendix for a schematic representation of the sequence of imputations for the m=n=2 case.

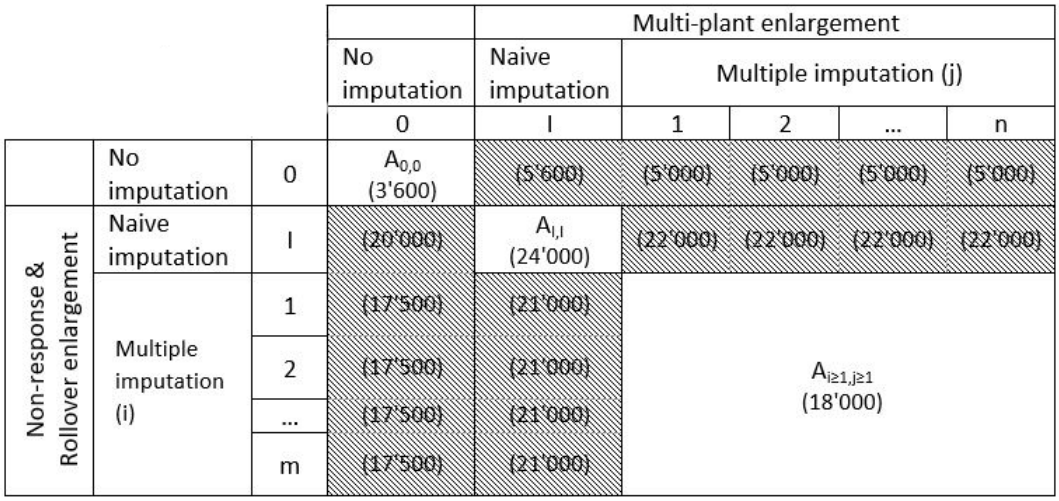


Figure 4 – Stylized imputation strategy.

Notes: Number of firms per year between parentheses. No imputation means keeping only the *restricted* sample (see figure 1); *Naive imputation* is based on employment share only; *Multiple imputation* is based on a multivariate regression and is repeated m(n) times. Dashed zones correspond to alternative datasets for robustness. A_{00}: dataset including only firms of the restricted sample (no imputation). A_{II}: dataset including all firms from the WS survey. $A_{i\geq 1,j\geq 1}$: datasets including all firms from the WS survey minus outliers and stand-alone cases.

3 Aggregation to *pseudo-firms*

In a final step, to maintain confidentiality, we re-aggregate all the plant-level databases obtained from the previous stages at the level of unique combinations of 4-digit sector, municipality and legal form. Each combination is called a "pseudo-firm".

To document this final aggregation step we calculate (in addition to the value-added and employment data) the following indicators for each pseudo-firm: number of plants, number of firms, number of active 6-digits sectors and coefficient of variation of full time employment across plants. We also construct two employment-related variables which are necessary to locate and weight the pseudo-firms in the final sample:

1. The employment-weighted geographic coordinates of the economic center of gravity of the pseudo-firms, which allow its spatial localisation (see Figure A2 in the Appendix for an example).
2. The weight attributed to each pseudo-firm, which is obtained according to one of the three procedures described below.

Whatever the procedure followed to construct pseudo-firm weights, it has to respect two principles. In the initial WS sample, weights are attributed so that, if all firms were to respond, the sum of weight-inflated employment figures would be equal to total employment in the WS sample frame. The first principle is to adjust weights in order to maintain this desirable property in each (pseudo-firm level) final sample. As a result, whatever the sample, the employment-weighted total is always the same, and the difference with respect to total employment is due to non-response. Another source of concern is that some firms disappear and other re-introduced through the selection and imputation processes described above. It is therefore not guaranteed that the distribution of pseudo-firms in the final sample is representative of the observed distribution of firms in the sample frame. The second principle is to adjust weights in order to minimize the difference between the probability density function of pseudo-firms in the final sample and the converse density function for the sample frame obtained from STATENT data.

3.1 Aggregation of original sampling weights

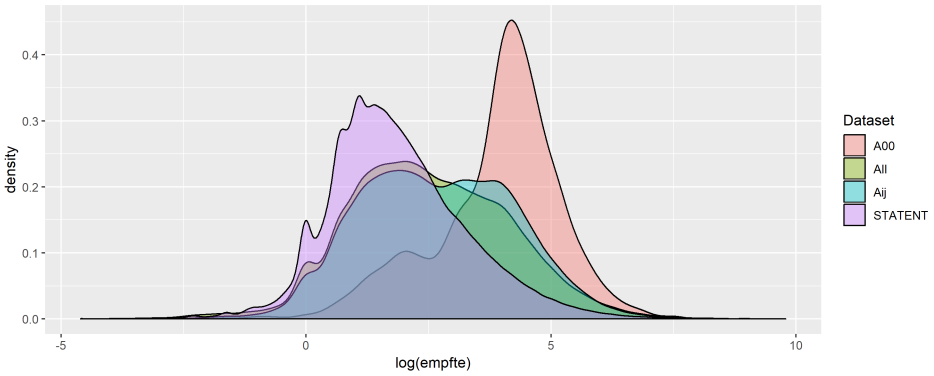
A first way to comply with the above-mentioned principles, is simply to aggregate reported weights in the original WS sample. We proceed as follows:

1. We calculate the sum of weight-inflated employment figures (full time equivalents) in the initial WS sample, denoted by L_I.
2. If, for a given year, the firm is present in the final sample but missing in the original sample, we impute for that year the average observed weight when the firm is not missing. Then we calculate the sum of weight-inflated employment figures (full time equivalents) in the final sample, denoted by L_F.

3. We calculate firm-level adjusted weights as the product between the original weight and the L_I/L_F ratio.
4. We assume identical adjusted weights across all plants of a given multi-plant firm.
5. When aggregating from the plant to the pseudo-firm level, we calculate the adjusted-weight of the pseudo-firm as the employment-weighted average of the adjusted weights of all plants belonging to that same pseudo-firm.

This first set of weights improves the matching between the distribution of employment in the final samples and the distribution of employment in the whole population. This is illustrated in Figures 5 and 6 comparing kernels of employment (full-time equivalent) probability density functions (pdf), for year 2015, for each type of final database defined in Figure 4 ($A_{0,0}$, $A_{I,I}$ and $A_{i,j}$, see Figure 4) and also for the reference population i.e. the database obtained when aggregating STATENT data at the level of pseudo-firms for the WS sample frame. [10]

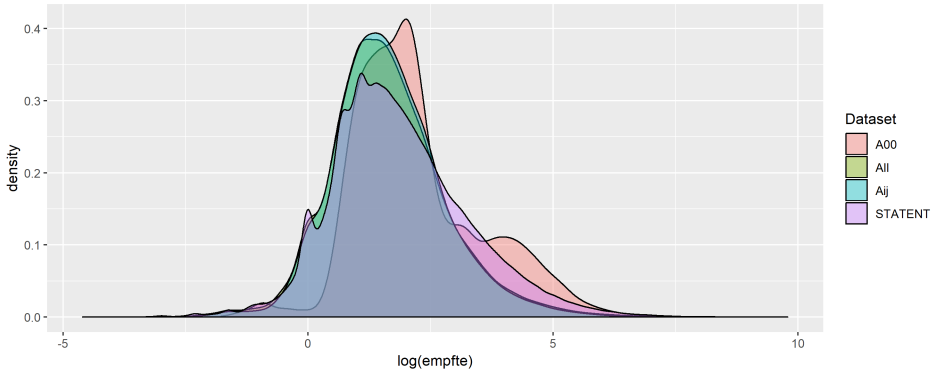
Regarding non-weighted data (Figure 5), as could be expected, the contrast is striking between the population (STATENT purple curve) and the restricted sample ($A_{0,0}$ pink curve), which is biased towards large firms. The two imputed samples (either $A_{I,I}$ or $A_{i,j}$, green and blue curves) lie as intermediate cases between these two extremes. As it should be, applying weights to the restricted and imputed samples drastically reduces these differences, as illustrated by figure 6, where all four pdfs now overlap more closely. Results for all years are similar, they are reported in figures A3 and A4 in the Appendix.



Notes: See figure 4 for a description of datasets A_{00}, A_{II} and A_{ij}

Figure 5 – Non-weighted density functions for full-time employment equivalents, 2015

10. As the WS survey is not supposed to include firms with less than three employees, those are dropped from STATENT data, except if they report more than two employees during at least one year. Sectors A, K, O and Q (partially) are not covered by the WS, so there are also dropped from all databases.



Notes: See figure 4 for a description of datasets A_{00}, A_{II} and A_{ij}

Figure 6 – Weighted density functions for full-time employment equivalents, 2015

At closer look however, there remains differences in shape, and although these differences are stronger for $A_{0,0}$ (which is due to the fact that this database is biased towards large firms) they are also present for $A_{I,I}$ and $A_{i,j}$. More precisely, with respect to the reference population, the pdf of the imputed samples are larger for small and medium range pseudo-firms, and smaller for large pseudo-firms. In other words, it suggests that these preliminary weights are too large at the bottom of the distribution. This is due to the enlargement of the database towards small firms, and requires further adjustments of the weights, as performed by the two following procedures.

3.2 Adjusted weights I: Proportional downsizing

The first procedure to adjust weights consists of finding the most appropriate mix between no weighting at all (which understates the frequency of small firms) and full weighting (which overstates it). More specifically, for each type of imputed dataset, we calculate an adjusted weight for pseudo-firm p year t given by:

$$\tilde{\omega}_{p,t} = \lambda_t \omega_{p,t} + (1 - \lambda_t)\frac{1}{n_t}$$

where $\omega_{p,t}$ is the original (normalised [11]) weight, $\tilde{\omega}_{p,t}$ the adjusted weight, n_t the number of pseudo-firms and λ_t the optimisation parameter, $\lambda_t \in [0, 1]$. These adjusted weights are then re-scaled so that the sum of weighted employment figures remains unchanged, as for the weights of subsection 3.1.

We determine numerically the optimal value of λ by 0.01 increments. Figure 7 reports the results for 2015. It turns out that the area is minimised at $\lambda = 1$ for $A_{0,0}$, $\lambda = 0.69$

11. To ease calculation of area differences between the density functions by the R package, weights are normalised so that their sum is equal to 1.

for $A_{I,I}$ and $\lambda = 0.73$ for $A_{i,j}$. Figure 8 illustrates the impact of the optimisation process on the overlap between the two pdf for the multiple imputation sample ($A_{i,j}$). Results for the other samples and years are very similar and available upon request.

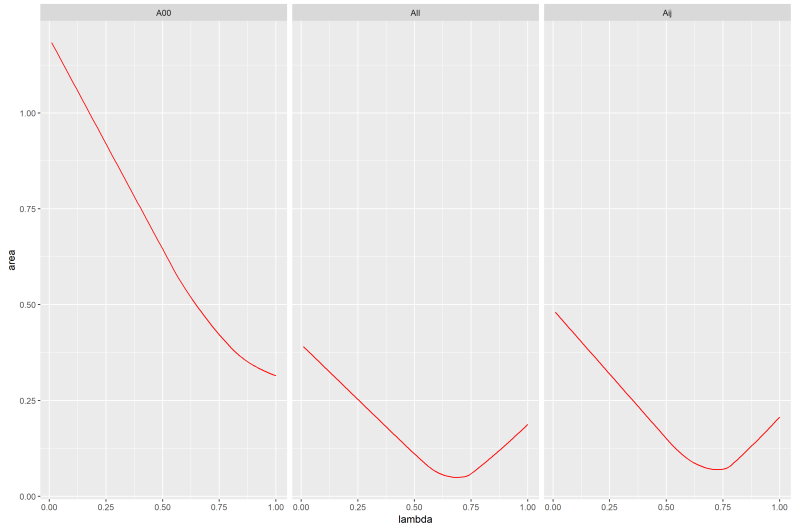


Figure 7 – Optimal values of the weighted average parameter (λ), 2015

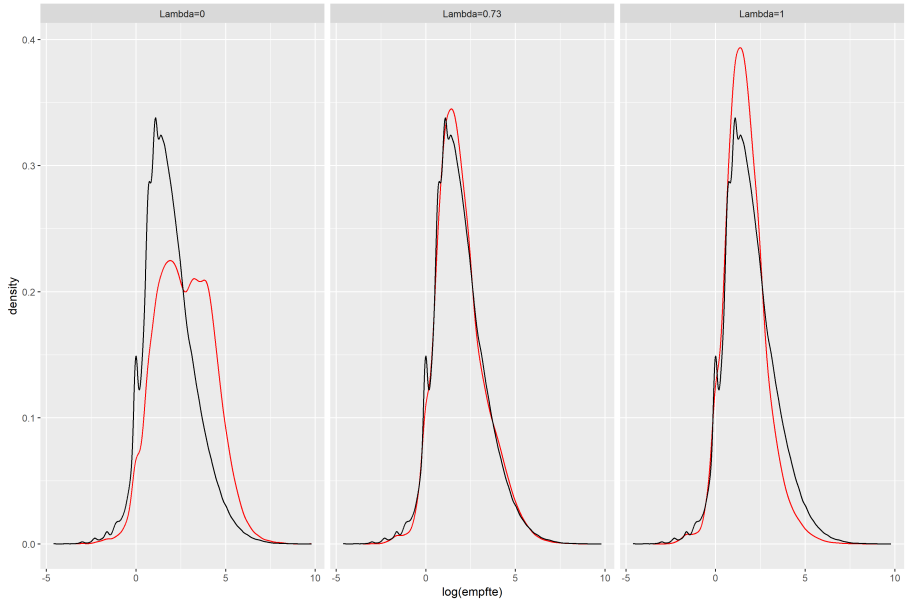


Figure 8 – Impact of optimal weights on the matching between the multiple imputation sample and the reference population, 2015

3.3 Adjusted weights II: Bottom up reconstruction

Adjustment method I (proportional downsizing) is intuitive but lacks theoretical justification and does not exploit fully the available information. The alternative proposed here is to reconstruct weights from the bottom up, making use of our knowledge of the employment distribution in the sample and over the reference population of the WS sample frame (see footnote 9 above).

General procedure

The basic idea is to decompose the population into different strata according to geography (Switzerland, large regions, cantons, districts), industry (NOGA2, NOGA3, NOGA4), legal form (yes, no) and size classes (5, 10 , 15 or 20 size classes, see figure 9 for a definition). This leads to 96 possible strata definitions (4x3x2x4). For a given definition, the weight of each particular combination is defined as the inverse of the sample full-time employment (FTE) share in the refrence population. Finally, the best definition of strata is selected by minimizing again the differential area between the population and the sample employment distributions.

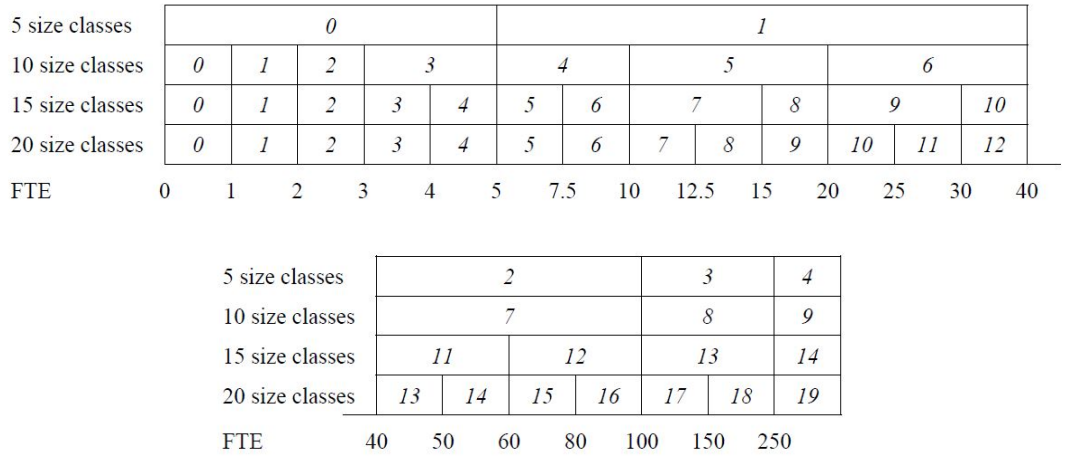


Figure 9 – Full-time employment (FTE): Size classes definitions
Notes : Class size numbers in italic / upper bound not included in the interval

Dealing with the "zero-weight" issue

Although straightforward, this procedure cannot be applied directly at the pseudo-firm level. To understand why, imagine that for a given combination of canton, 4digit industry and legal form there are three plants in the reference population, with FTE figures of 4, 8 and 9 respectively. Assume further that we work with 15 size classes (third line of figure 9) and that the plant with 4 FTE is not present in the sample (a frequent

case as small firms are under-represented in the WS sample). This leads to a sample pseudo-firm employment of 17, versus 23 for the reference population. As the threshold is 20 between size classes 8 and 9 (see figure 9), this leads to a weight of zero for class 8 and no reported employment for class 9, i.e. a complete loss of all the available information. Working at the firm rather than the pseudo-firm level doesn't necessarily solve this *zero-weight* problem (for example if the smallest and largest plant of the previous example belong to the same firm misclassification also happens).

At the end of the day, the only way to eradicate the zero-weight problem is to define weights at the plant level. However, at this stage of the procedure, we have no access to plant-level data anymore as we are already dealing with pseudo-firm (i.e. aggregated) data. Thus, we need to recover plant-level data exploiting the available information we have in STATENT and in the variables created during the pseudo-firm aggregation. We proceed in three steps (see Appendix A.6 for a more detailed description of step 3):

1. **Direct calculation**. When the pseudo-firm has less than 3 plants, knowledge of the mean ($\bar{X}$) and the variation coefficient (VC) is sufficient to calculate plant level FTE ($\bar{X}$ if there is a single plant, $\bar{X} \pm \frac{VC}{\sqrt{2}}$ if there are two plants). This corresponds to 94.5% of pseudo-firms (71% of total FTE).
2. **Combinatory analysis**. For pseudo-firms with a number of plants which is larger than two but sufficiently close to the total number of plants in STATENT, the exact combination of plants within the pseudo-firm can be retrieved computationally in a reasonable time. Let us denote by n the number of plants of the pseudo-firm, and N the number of plants for the same combination of municipality, NOGA4 and legal form values in STATENT. Then, among all the possible C_n^N combinations of n out of N plants, and provided $C_n^N < 75000$ (threshold determined by trial and error) one identifies the unique combination that leads to the same $\bar{X}$ as the one reported for the pseudo-firm (calculating the VC constitutes a proof). This corresponds to 4.5% of pseudo-firms (17% of total FTE).
3. **Gamma distribution**. For all remaining cases (1% of pseudo-firms, 12% of total FTE, i.e. essentially large pseudo-firms with many plants), we assume that plants' employment follows a Gamma distribution and distribute total FTE across plants in accordance with the reported values for n, VC and $\bar{X}$. This generates some zero-weight cases. We eliminate these cases by reshuffling employment across plants in non-zero weight categories, working with the maximum number of size classes (20) and along a systematic procedure described in Appendix A.6.

Steps 2 and 3 above are rather time consuming so their application is limited to the multiple imputation case ($A_{i,j}$ in figure 4). Moreover, they imply small adjustments of the reference population. [12]

12. When it is not possible to find a combination in step 2 that perfectly matches the reported $\bar{X}$ for the pseudo-firm, the selected sample is enriched by all available observations from STATENT data (i.e. including plants with less than three employees during the entire time period) for that particular set of

Selecting the optimal strata definition

Once plant-level employment figures have been recovered, plant-level weight are calculated (inverse of employment share) and re-aggregated back as employment-weighted averages at the pseudo-firm level. Some categories combinations are present in the population but not in our sample. To recover this missing employment, we inflate all weights by a common factor. These calculations are performed for each one of the 96 above-mentioned strata definitions. We select the most appropriate definition following the same criterion as in the previous section i.e. by minimizing the difference, in terms of area, between the population and sample employment densities. Figure 10 shows that the best strata definition turns out to be canton, NOGA3, including legal form and 15 size classes.

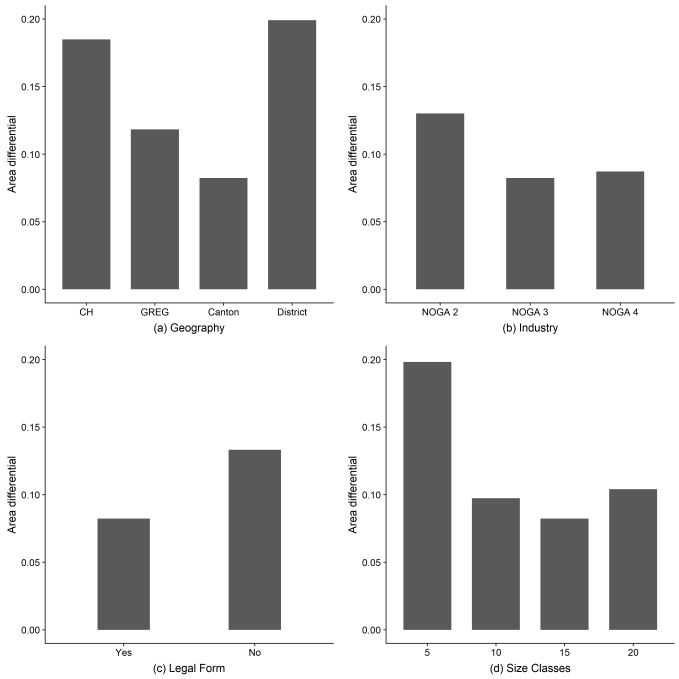


Figure 10 – Selection of the most appropriate strata definition
Notes : average differential area between the $A_{i,j}$ sample and the population density / in each panel the values of the other dimensions are set at their optimal level i.e. canton, NOGA3, legal form and 15 size classes.

As illustrated by Figure 11, the bottom-up adjustment method (weights II, in green) provides a better fit of the population distribution (in blue) than the proportional down-

year, municipality, NOGA4 and legal form values. This affects 98 pseudo-firms (i.e. 0.11% of the total). A similar enlargement of the reference population is necessary for 62 pseudo-firms in step 3(i.e. 0.07% of the total).

sizing methods (weights I, in pink). On average over the 2011-2015 period, the differential area between the sample and the population employment densities is 0.091 for weights I (proportional downsizing method) and 0.082 for weights II (bottom-up method).

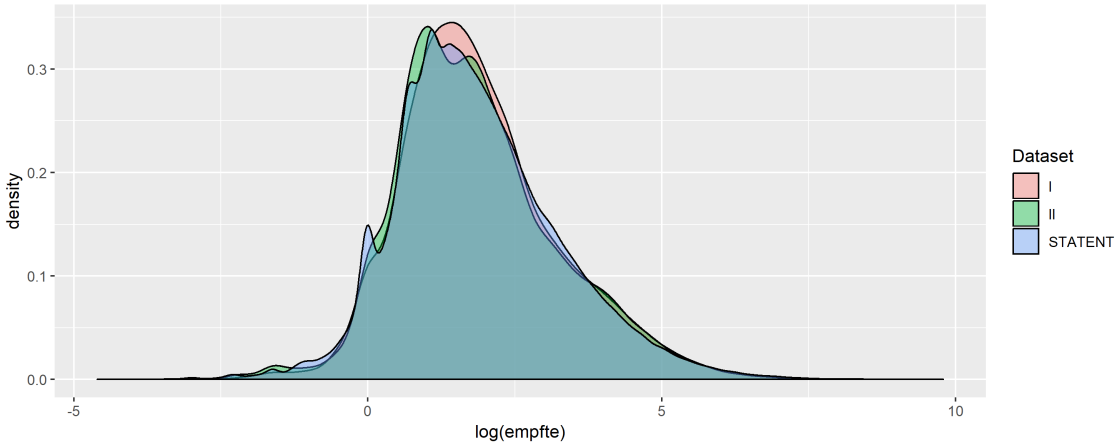


Figure 11 – Comparison of the two adjusted weights methods, $A_{i,j}$ (2015)

4 Data Overview

4.1 Data summary

Table 1 summarizes the characteristics of the main databases generated by the imputation procedures. Beware that observations here correspond to pseudo-firms, while in figure 4 the numbers between parenthesis correspond to firms. Whatever the sample, the number of pseudo-firms is almost constant over time, as entry/exit rates are roughly similar. The *restricted sample*, $A_{0,0}$, is the smallest database, with less than 3000 yearly observations. The number of plants is not much larger. This is normal, as the restricted sample is mostly composed of relatively large single-plant firms, with only some non-problematic multi-plant firms with all plants in the same pseudo-firm. The *naive imputation* sample ($A_{I,I}$) considerably increases the sample size, in terms of pseudo firms, due to the non-response and rollover enlargements, and even more in terms of plants, due to the multi-plant enlargement. On average, the *multiple imputations sample* ($A_{i,j}$) is around twice smaller than the naive imputation sample, due to the elimination of outliers and stand-alone cases. However, it remains considerably larger than the restricted sample (6 times larger in terms of pseudo-firms, 8 times in terms of plants).

Regarding the composition of each database, it appears that the restricted sample pseudo-firm is on average twice larger than in the other two databases (around 100 full

Dataset	Year	Number*	N new$^{\triangle}$	Number	Full-time equivalent				Value added**				Share of municipalities	Total employment share†	Total value-added share††		
				of plants	Min	Max	Mean	CV	Min	Max	Mean	CV	covered	non-weighted	non-weighted	weighted (I)	weighted (II)
$A_{0,0}$	2011	2936		3174	0.2	8170.98	93.8	2.12	6.0	3540885.80	22463.4	5.53	39.5%	8.37%	12.38%	90.93%	/
	2012	2812	67	3052	0.3	8085.11	97.4	2.07	8.0	3920905.00	24582.7	5.96	38.6%	8.26%	12.84%	95.61%	/
	2013	2750	62	2994	0.5	8537.80	100.3	2.12	33.0	3556642.00	24735.3	5.71	38.0%	8.16%	12.43%	97.58%	/
	2014	2753	107	2994	0.2	8549.70	101.8	2.12	12.6	3477452.00	25696.0	5.85	38.3%	8.20%	12.60%	98.69%	/
	2015	2803	147	3059	0.2	8524.84	99.8	2.14	7.7	4183055.00	24838.4	6.06	38.9%	8.15%	12.29%	94.23%	/
$A_{I,I}$	2011	33008		52687	0.0	15868.97	47.1	3.72	0.3	7996248.68	8419.5	8.14	81.2%	47.22%	52.18%	77.06%	/
	2012	33599	1743	53898	0.0	16484.06	46.8	3.81	0.3	8472771.52	8493.6	8.69	81.7%	47.42%	53.02%	77.59%	/
	2013	35006	2861	57630	0.0	16824.98	46.4	3.88	0.7	8354989.67	8410.8	8.44	83.0%	48.03%	53.79%	80.05%	/
	2014	34918	1675	57547	0.0	17268.41	46.9	3.86	1.1	7938707.26	8636.6	8.27	83.1%	47.90%	53.74%	81.04%	/
	2015	34394	1530	56645	0.0	17708.27	47.3	3.87	0.7	8495329.78	8710.2	8.54	82.8%	47.40%	52.89%	79.64%	/
$A_{i,j}$ $^{\triangledown}$	2011	17625	/	23878	0.0	15797.35	48.7	3.67	1.0 (0.54)	7896097.33 (608705.3)	8535.0 (18.6)	9.22 (0.33)	65.3%	26.07%	28.35% (0.06)	100.03% (0.27)	96.96% (0.37)
	2012	17924	834	24365	0.0	16416.76	48.5	3.78	0.9 (0.49)	8403539.36 (592480.1)	8311.0 (21.3)	9.82 (0.33)	65.9%	26.20%	27.78% (0.07)	97.43% (0.25)	94.49% (0.35)
	2013	17925	794	24398	0.0	16824.98	49.0	3.83	1.0 (0.46)	8259435.51 (696976.5)	8392.5 (23.0)	9.61 (0.38)	65.6%	26.00%	27.58% (0.08)	97.21% (0.35)	93.93% (0.40)
	2014	17860	794	24342	0.0	17268.41	49.5	3.84	1.1 (0.29)	7741245.10 (675973.8)	8608.2 (15.5)	9.11 (0.34)	65.9%	25.88%	27.49% (0.05)	96.97% (0.22)	94.43% (0.32)
	2015	17586	808	23934	0.0	17708.27	49.8	3.90	0.5 (0.16)	8362957.38 (682613.6)	8736.0 (18.4)	9.55 (0.36)	65.6%	25.53%	27.23% (0.06)	97.09% (0.27)	93.50% (0.29)

Notes: * Number of *pseudo-firms* (unique noga4-municipality-legal form combinations).

$^{\triangle}$ Number of emerging *pseudo-firms* (i.e. present in the current year but not previous one).

** In thousands CHF.

$^{\triangledown}$ With $(i, j) \in \{1; 20\}$. In parenthesis, the standard deviation of the given statistics (i.e. statistics are computed on each of the 400 datasets and the mean and standard deviation of each statistics are reported (implicitly in hundreds when needed)).

† We exclude sectors A, K and U from the reference population, as those sectors are not part of the WS survey. Weighted employment results are not reported as they are identical for all samples due to calibration.

†† We exclude sectors A, K, T and U from the reference population, as those sectors are not part of the WS survey. Weights I and II are defined in, respectively, sections 3.2 and 3.3.

Table 1 – Summary statistics of the datasets.

time-equivalent vs. 50). It is also more productive (around CHF 230000 vs. 170000 per full time-equivalent per year). This should not come as a surprise as the imputed samples were basically designed to recover small firms which are rolled over and to split value-added across the various units of multi-plant firms. Both manipulations reduce the average size of the production units. This result is consistent with the structural differences between the various sample types discussed in Section 3.

4.2 Global coverage of employment and value added

The last three columns of table 1 provide the share of employment and value added covered by the different samples. Consider employment first, the reference population being obtained from STATENT data on the basis of the WS sample frame. As weights are always calibrated in order to match the reference population, there is no point to report the weighted share for employment. Regarding the non-weighted shares, the imputation procedures do increase the coverage of the sample, which is rather low for the restricted sample ($A_{0,0}$), less than 10%), up to more than 25% for the multiple imputation sample ($A_{i,j}$) and slightly less than 50% for the naive imputation sample ($A_{I,I}$).

Regarding value-added, we use as a comparison basis the figures estimated by the National Accounts department of the FSO. They constitute the official reference in Switzerland, and the FSO has taken care over recent years to refine its procedure in order to provide robust estimates of value added aggregated at the regional and industry level (see (Federal Statistical Office, 2016)). The methodology followed by the FSO also relies partly on the WS database, but it is distinct from the procedure applied in the present paper on several counts (apart from the basic difference in objectives, i.e. we seek to provide firm-level rather than aggregated level estimates). We focus here on the major distinctions. First, it performs only a multiplant enlargement, which means that the non-response and turnover enlargement is not considered. Second, it works at a more aggregated level than we do for sectors (21 instead of 272), plant size categories (10 instead of 15), geographical units (7 instead of 26) and legal forms (none vs. 22). Third, it calibrates its results at the aggregate level in order to make them consistent with other national account calculation approaches for GDP.

Given the above-mentioned differences, imperfect coverage may be expected for our three samples. The total non-weighted value added share is indeed rather low, although slightly larger than for employment. However, and quite surprisingly, there is a rather good match for the weighted figures, particularly for the two imputed samples. The value-added share is around 125% for the restricted sample, again a reflection of the bias of that sample towards large and more productive single plant firms. The imputed samples are more representative of the distribution of firms in the population, and although their weights were obtained from a completely different perspective from the FSO methodology, they achieve a share in total value added which is quite close to 100% (104% for $A_{I,I}$, 95-98% for $A_{i,j}$).

4.3 Detailed coverage of the imputed samples ($A_{I,I}$, $A_{i,j}$)

Pursuing the analysis at a more disaggregated level, tables 2 and 3 report the coverage rates for large geographic regions and sectors for the two imputed samples (see Appendix A.7 for the restricted sample). As the naive imputation sample covers a larger share of total employment (around 50% vs. 25% for the multiple imputation sample), it leads naturally to a better employment coverage for large regions, with an approximate range of 90-110% (vs. 50-120% for $A_{i,j}$). This remains valid for value added, and to a lesser extent also for large industry groups, even if the DEPQ share (public utilities, education and health) falls to 40% for employment and 55% for value added. As employment and value added shares are positively correlated, the contrast between the two samples is less stark regarding the productivity ratio. The better coverage of the naive imputation sample must be put in balance with its major drawback namely that its two enlargement procedures are based on the explicit assumption of a constant productivity whether within firms or across periods. This makes it less appropriate than the multiple imputation sample to analyse productivity change at the micro level.

Table 2 – A_{II} coverage

	Weigthed full-time equivalent share	Weighted value added share[†]	Productivity share[▽]
Major regions			
Espace Mittlelland	99.1%	95.0%	95.9%
Région lémanique	95.5%	120.6%	126.3%
Zürich	101.1%	102.4%	101.3%
Nordwestschweiz	108.9%	110.0%	101.1%
Ostschweiz	99.6%	90.8%	91.2%
Zentralschweiz	100.9%	116.1%	115.1%
Ticino	90.7%	78.4%	86.5%
Industries			
GHIJ	124.5%	144.0%	115.7%
BCF	129.8%	135.4%	104.4%
DEPQ	41.4%	56.6%	136.6%
LMNRS	104.8%	93.3%	89.0%
Total	100.0%	104.0%	104.0%

Notes:
[†] Share in value added of the national accounts, excluding industries NOGA 1 A,K,T,O and Q(partially), which are not covered by WS.
[▽]Ratio between average productivity in the sample and "national" productivity.
"National" productivity is based on own calculations using national accounts total value added data and STATENT.
[*] Weights defined in section 3.2.

Table 3 also provides the opportunity to compare the two alternative sets of weights for the multiple imputation sample. The global coverage for value added is 3% higher for weight I. However, weights II lead to a smaller coverage range than weights I. This is valid in general, both for employment or value added, and for large regions or large sectors. Moreover, across the 400 databases, the standard deviation is generally smaller

Table 3 – A_{ij} coverage

	Weigthed full-time equivalent share		Weighted value added share†		Productivity ratio▽	
	weights I*	weights II**	weights I	weights II	weights I	weights II
Major regions						
Espace Mittelland	101.3%	99.4%	88.0%	82.8%	88.5%	83.3%
			(0.36)	(0.28)	(0.36)	(0.29)
Région lémanique	89.7%	97.2%	113.0%	116.4%	116.2%	119.7%
			(0.50)	(0.32)	(0.52)	(0.33)
Zürich	101.7%	118.7%	104.2%	114.0%	87.8%	96.0%
			(0.53)	(0.46)	(0.44)	(0.39)
Nordwestschweiz	123.8%	111.1%	130.2%	113.5%	117.2%	102.2%
			(0.42)	(0.39)	(0.38)	(0.35)
Ostschweiz	114.1%	99.7%	88.6%	76.7%	88.9%	76.9%
			(0.37)	(0.43)	(0.37)	(0.43)
Zentralschweiz	86.7%	71.2%	68.8%	57.6%	96.6%	80.8%
			(0.36)	(0.39)	(0.50)	(0.55)
Ticino	44.4%	63.2%	36.6%	50.1%	58.0%	79.2%
			(0.68)	(0.38)	(1.07)	(0.60)
Industries						
GHIJ	107.0%	120.2%	126.0%	131.4%	104.9%	109.3%
			(0.44)	(0.33)	(0.36)	(0.28)
BCF	161.7%	144.5%	160.1%	141.4%	110.8%	97.9%
			(0.32)	(0.36)	(0.22)	(0.25)
DEPQ	38.4%	35.0%	37.4%	33.7%	107.1%	96.3%
			(0.21)	(0.14)	(0.60)	(0.40)
LMNRS	88.9%	96.9%	65.6%	70.4%	67.7%	72.6%
			(0.37)	(0.30)	(0.38)	(0.31)
Total	100.0%	100.0%	97.7%	94.7%	97.7%	94.7%
			(0.17)	(0.17)	(0.17)	(0.17)

Notes:
† Share in value added of the national accounts, excluding industries NOGA 1 A,K,T,O and Q(partially), which are not covered by WS. Standard deviation in parenthesis.
▽Ratio between average productivity in the sample and "national" productivity.
"National" productivity is based on own calculations using national accounts total value added data and STATENT.
* Weights defined in section 3.2.
** Weights defined in section 3.3.

for weights II than for weights I. Thus, on balance, it appears that weights II offer a more stable representation of value added and productivity differences across Swiss firms.

5 Conclusion

Data on value-added at the level of the production unit are difficult to obtain for Switzerland. Official sources are only reported at the firm level, not the plant level, only available for a subsample of relatively large firms, and only reported with other monetary variables, not employment figures. This makes it particularly unsuitable to undertake a proper analysis of productivity at the microeconomic level.

Taking the best out of available data sources, we used several techniques to address these caveats in a novel way. The new set of three databases that results includes consistent information on value added and employment over the 2011-2015 period and at a high degree of economic and geographic granularity. To protect confidentiality, data had to be reported for "pseudo-firms", a constructed production unit at the level of

the municipality, legal form and 4-digit sector. The information loss due to that slight re-aggregation is kept minimal, as 94% of pseudo-firms have less than three plants. Moreover, broadly speaking, our results are in line with the value added estimates produced by the FSO along an entirely different methodology. However, the composition of each sample is different, which must be kept in mind for interpretations.

The *restricted* database is only representative of the upper part of the distribution of firms, as it includes around 2'800 large pseudo-firms which were present every year in the original WS survey. The two types of *imputed* databases re-integrate small production units, leading to larger samples. The *naive imputation* database is the largest one, with more than 34'000 pseudo-firms. However, it is based on the assumption of constant labor productivity through time or between plants of the same firm, and thus improper for a detailed micro-based analysis of productivity change. Our preferred option is thus the *multiple imputation* database, which relies on additional information to allocate value added across plants which were not systematically surveyed or which locate in another municipality than the headquarters. It reports an interval of 400 different estimates for the imputed value-added of the 18'000 pseudo-firms that constitute the final database.

To illustrate the potential use of these new information sources, let us suggest an application to the low-productivity performance of Switzerland in recent years. Using National Accounts data reveals that Swiss productivity growth (in terms of value added per employee and for the subset of sectors considered in the WS survey) has been roughly equal to 0.6% per year across the 2011-2015 period. According to our own estimates based on the multiple imputation sample, when re-aggregated at the national level, we get an even more disappointing figure, at 0.01% per year. The difference is probably due to our improved coverage of small firms, which are less productive than larger ones. That apart, the overall productivity performance remains fairly poor. However, it masks important sectoral and geographical differences. Yearly productivity changes vary a lot across both municipalities (between -14% and +22%) and 4-digit sectors (between -11% and +22%). These structural patterns deserve further examination and are analyzed in a companion paper (see Tissot-Daguette and Grether (2021)).

6 References

Federal Statistical Office. Produit intérieur brut par grande région et par canton. *Rapport méthodologique*, 2016.

John W. Graham, Allison E. Olchowski, and Tamika Gilreath. How many imputations are really needed? some practical clarifications of multiple imputation theory. *Prevention Science*, 8(3):206–13, 2007.

Donald B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. John Wiley and Sons, Inc., 1987.

SAS Institute Inc. *SAS/STAT 14.1 User's Guide*, chapter 75 : The MI Procedure. Cary, NC: SAS Institute Inc., 2015.

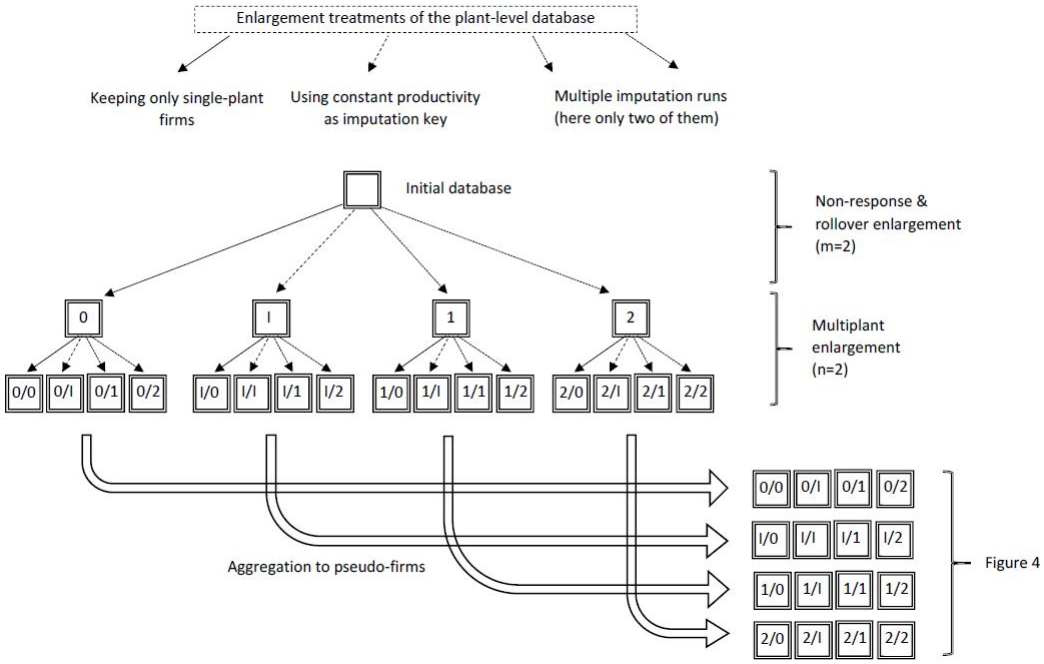
Joseph L. Schafer and John W. Graham. Missing data: Our view of the state of the art. *Psychological Methods*, 7(2):147–177, 2002.

Benjamin Tissot-Daguette and Jean-Marie Grether. Zooming in, zooming out: A shift-share analysis of productivity in switzerland based on micro data. *IRENE Working Paper, 21-10, University of Neuchâtel*, 2021.

Yang Yuan. Multiple imputation using sas software. *Journal of Statistical Software*, 45 (6), 1994.

A Appendix

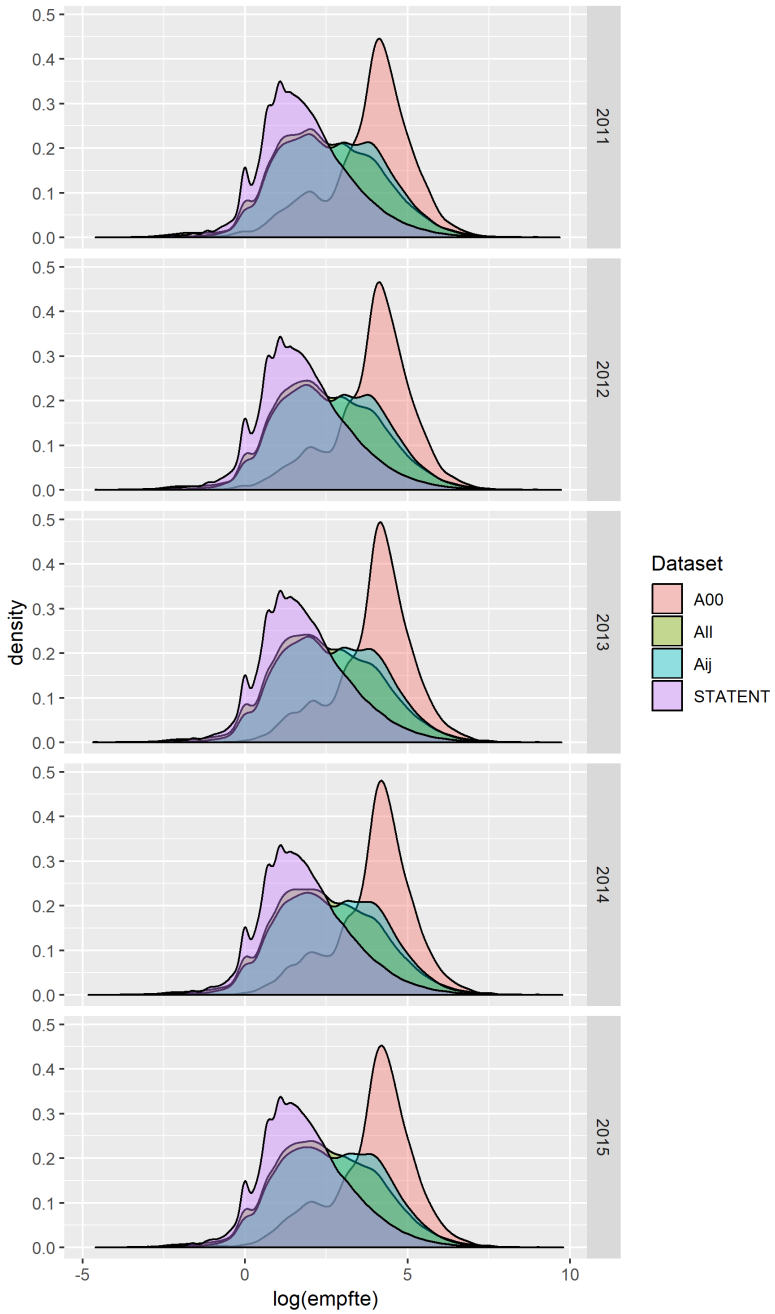
A.1 Detailed imputation strategy in the m=2, n=2 case



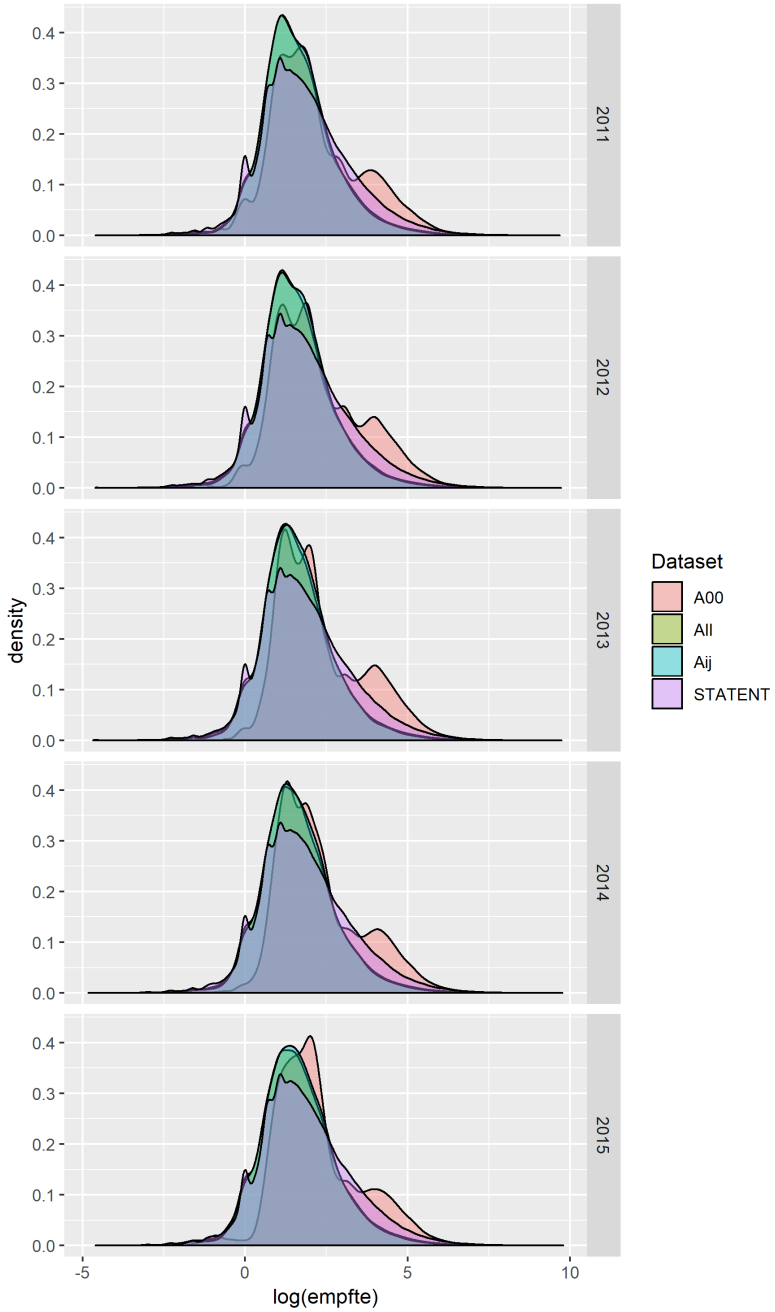
A.2 Geographic coordinates of a pseudo-firm: example

Red (blue) balloons correspond to the localization of the original firms (constructed pseudo-firm).

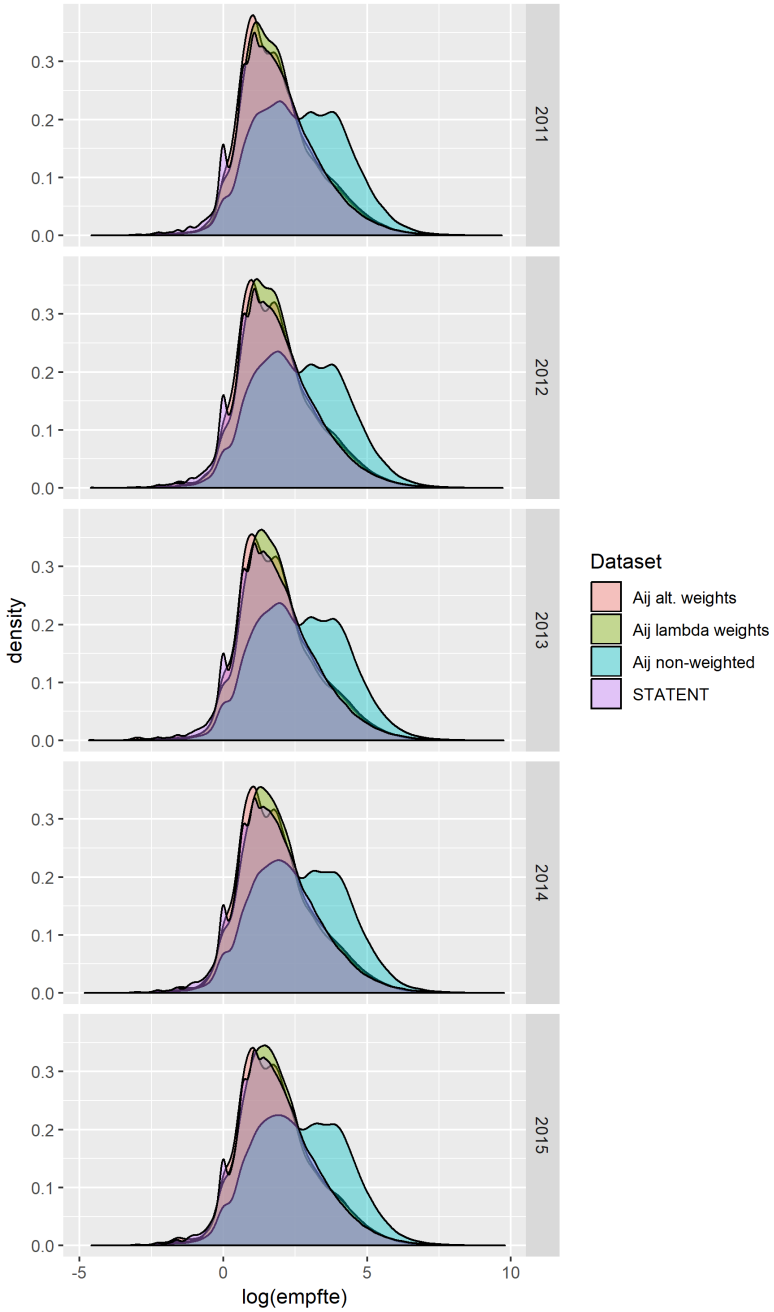
A.3 Non-weighted density functions for full-time employment equivalents, all years



A.4 Weighted density functions for full-time employment equivalents, all years



A.5 Weighted and non-weighted density functions for Aij database, all years



A.6 Recovering plant-level data employment figures: step 3

This appendix provides more details on step 3 of the general strategy to recover plant-level employment described in sub-section 3.3. For all pseudo-firms not covered by steps 1 and 2, the general principle is to apply a Gamma distribution first, then perform two further adjustments. The exact procedure is described below.

Applying the Gamma distribution assumption within pseudo-firms

We proceed as follows:

1. We infer the variance (V) of the pseudo-firm employment (X), $V = CV^2 \cdot \bar{X}^2$. Then we compute the shape (α) and rate (β) parameters of the Gamma distribution, $\alpha = \frac{\bar{X}^2}{V}$ and $\beta = \frac{\bar{X}}{V}$.
2. We divide the area below the Gamma pdf in n equi-probable intervals, with n being the number of plants in the pseudo-firm.
3. We consider the midpoint of each one of the first $n-1$ intervals as the estimated employment of the corresponding plant. We infer the n^{th} plant's employment by substracting from the pseudo-firm total employment the sum of the midpoint estimates found in point 3.

Adjusting the estimates obtained through the Gamma distribution

We define as a category the combination of year, municipality, NOGA4 and legal form that corresponds to each pseudo-firm. Within each category, we consider 20 size classes (see figure 9). This allows refining the Gamma approximation avoiding two types of inconsistencies. More precisely, we reallocate FTE among plants of a given pseudo-firm in order to eliminate all cases where

- there is a plant in the sample for that category and size class, but no plant in STATENT. This would lead to attributing a weight of zero to that plant, i.e. a *zero-weight* size class.
- there are plants in both the sample and STATENT for that category and size class, but too many plants in the sample regarding total reported FTE in STATENT. That is, even if FTE per plant was kept at a minimum in the sample (i.e. reduced to the lower bound of the size class for each plant), total FTE would remain larger in the sample. We call this an *overcrowded* size class.

Both problems signal that the Gamma-distribution-based attribution of pseudo-firm FTA across its plants is not correct. They also distort computed weights for the category and size-class and therefore deserve correction. The relative importance of both cases is presented in table 3. A detailed presentation of each adjustment type follows.

Table 4 – Adjustments of the Gamma distribution estimates

	Number of pseudo-firms	FTE	Average FTE per pseudo-firm	Share of pseudo-firms	Share of FTE
Elimination of zero-weight cases					
No zero weight	453	205404	91	0.5%	4.7%
Zero weight	710	326303	92	0.8%	7.5%
Elimination of over-crowded cases					
Not over-crowded	857	269089	63	1.0%	6.2%
Simple cases	251	192599	153	0.3%	4.4%
Non-simple cases	55	70016	255	0.1%	1.6%
Total	1163	531707	91	1.3%	12.2%

Adjustment procedure to eliminate the *zero-weight* cases

1. Identify MINS, the minimum size class with positive STATENT FTE figures and MAXS, the maximum size class with positive STATENT FTE figures.
2. Identify the number of plants with zero-weights. Classify them in three groups: UPGR = those with a size class smaller than MINS (must be lifted up), DWGR = those with a size class larger than MAXS (must be scaled down) and INGR = those with a size class in between MINS and MAXS.
3. If the UPGR is not empty, for each zero-weight case, identify the upward FTE gap, UPFG i.e. the extra FTE needed to reach the nearest larger size class with employment in STATENT. Starting from the largest plant, attribute UPFG of extra FTE to the zero-weight plant while sharing the corresponding decrease in FTE on all other plants in proportion of their maximum capacity of provision under the constraint that they do not change size class. Repeat the procedure of the last sentence until the UPGR is empty.
4. If the DWGR is not empty, for each zero-weight case, identify the downward FTE gap, DWFG i.e. the decrease in FTE needed to reach the nearest smaller size class with employment in STATENT. Starting from the smallest plant, take away DWFG of FTE from the zero-weight plant while sharing it across all other plants in proportion of their maximum capacity of absorption under the constraint that they do not change size class. Repeat the procedure of the last sentence until the DWGR is empty.
5. If the INGR group is not empty, for each zero-weight case, compute the minimum of the upward and downward FTE gap as described in the previous two steps, i.e. MNFG=min(UPFG;DWFG). Rank these cases by increasing MNFG. Starting from the smallest MNFG, adjust the FTE of the zero-weight plant up (by UPFG) or down (by DWFG) depending on which adjustment is smaller and compensate that change across all other plants in proportion of their maximum capacity of absorption or provision under the constraint that they do not change size class.

Repeat the procedure of the last sentence until the INGR is empty.

Adjustment procedure to eliminate the *over-crowded* cases

1. Identify all pseudo-firms that present one or more cases of over-crowded size classes.
2. Identify simple cases i.e. those where there is a single plant to relocate, either out of two or out of three plants in the corresponding category. For each simple case, identify the smallest amount of FTE that must be given to (or taken out of) the plant in order to shift it to the closest available size class.
3. For non-simple cases, attribute the required changes in FTE "by hand" i.e. printing the situation and finding the set of minimum changes in order to eliminate the over-crowded problem.
4. For both simple and non-simple cases, redistribute the net required FTE change (in order to maintain total FTE of the pseudo-firm unchanged) across all other plants in proportion of their maximum capacity of absorption or provision under the constraint that they do not change size class.

A.7 A_{00} coverage.

Table 5 – A_{00} coverage

	Weigthed full-time equivalent share	Weighted value added share[†]	Productivity ratio[▽]
Major regions			
Espace Mittlelland	100.1%	114.4%	114.3%
Région lémanique	74.0%	125.1%	169.2%
Zürich	69.7%	84.5%	122.1%
Nordwestschweiz	141.8%	177.0%	124.8%
Ostschweiz	130.1%	122.4%	94.0%
Zentralschweiz	107.3%	178.6%	166.5%
Ticino	106.2%	79.5%	74.9%
Industries			
GHIJ	80.2%	166.5%	207.9%
BCF	190.1%	206.9%	108.9%
DEPQ	60.9%	82.7%	135.8%
LMNRS	65.5%	58.1%	89.0%
Total	100.0%	125.4%	125.4%

Notes:
[†] Share in value added of the national accounts, excluding industries NOGA 1 A,K,T,O and Q(partially), which are not covered by WS.
[▽]Ratio between average productivity in the sample and "national" productivity.
"National" productivity is based on own calculations using national accounts total value added data and STATENT.
[*] Weights defined in section 3.2.