Using auxiliary data sources in the structural survey in the service sector

Thomas Zimmermann, Destatis

As a consequence of the decision taken by the Federal Administrative Court of Germany in March 2017, the currently applied sampling scheme in the structural survey in the service sector might be subject to changes. At the moment, the survey is conducted using a stratified random sampling design where the strata are constructed as a full cross-classification of the 16 German Bundesländer (states) by 83 branches of economic activity and 8 size classes. The sample size allocation is such that reliable estimates are obtained for stratum groups defined by cross-classifying the Bundesländer and sectors of economic activity. Owing to the heterogeneity in the population, highly unequal sampling fractions result with many strata that are fully enumerated.

An alternative sampling design could avoid take-all strata as much as possible to spread to response burden more evenly among the target population. This may yield larger coefficients of variation under the currently used Horvitz-Thompson estimation method, especially when regional estimates are desired. Hence, also the issue of an appropriate estimation method needs to be addressed.

In our paper, we explore potential sources of auxiliary information that can be incorporated at the estimation stage. We discuss advantages and disadvantages of different sources of auxiliary information and estimation methods in light of their efficiency and reliability.

1 Introduction

The German structural survey in the service sector is an annual survey that provides relevant information on medium-term developments and structural changes for the service sector. It covers the following sections of the classification of economic activity: transportation and storage (section H), information and communication (section J), real estate activities (section L), professional, scientific and technical activities (section M), administrative and support service activities (section N) as well as the division S 95, which comprises repair of computers and personal and household goods. To collect the information, a survey sample is conducted where the maximum sampling fraction is restricted to 15 per cent of the total number of

units in the population. A requirement of the survey is that highly accurate estimates have to be produced for a large number of domains according to regional breakdowns as well as for subgroups determined by size classes and branches of economic activity. To achieve this goal, the units in the sampling frame are stratified according to their NUTS1 region (16 states), the first four digits of their NACE-classification (NACE4 hereafter, 83 sectors) and their size class as measured by turnover or alternative the number of employees (8 classes). In a second step, sample sizes are determined such precise estimates can be obtained for total turnover in stratum groups defined as cross-classifications from the NUTS1 region with NACE4-classes, and additionally, efficient estimates for higher level of aggregations are obtained.

Formally, the sample sizes in the strata are obtained by minimizing the maximum of the weighted coefficients of variation using a Horvitz-Thompson estimator for total turnover among stratum groups g, $g = 1, \ldots, G$:

$$F = \max_{g=1,\ldots,G} W_g^q \cdot CV(\widehat{Y}_g) = \max_{g=1,\ldots,G} \frac{W_g^q}{Y_g} \sqrt{\sum\nolimits_{h \in g} N_h^2 S_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right)^2}, \quad (1)$$

subject to

$$m_h \leq n_h \leq M_h, \quad \forall h, \quad (2)$$

and

$$\sum\nolimits_h n_h = n. \quad (3)$$

In (1), W_g is a measure of importance associated with stratum group g, $CV(\widehat{Y}_g) = \sqrt{Var\left(\widehat{Y}_g\right)}/Y_g$ denotes the coefficient of variation in stratum group g and the exponent q is a constant between 0 and 1. Furthermore, n_h, N_h and S_h^2 refer to the sample size, number of units and population variance of the variable of interest in stratum h, respectively. In (2), m_h and M_h denote the box-constraints for the sample sizes in the strata. Objective functions similar to (1) have been used for a long time at the German Federal Statistical Office (Destatis), since they ensure that the coefficients of variation are inversely proportional to W_g^q [cf. Schäffer, 1961]. Moreover, minimizing (1) under the constraints (2) and (3) can be shown to be a special case of the generalized power allocation developed by Hohnhold [2010], who extends the concept of the power allocation due to Bankier [1988]. In the structural survey, we set $m_h = 3\ \forall h$, $M_h = N_h\ \forall h$ as well as $W_g = Y_g$, $q = 0.2$. Since the actual turnover is estimated using the survey, the turnover variable in the sampling frame was used to compute the sample sizes. The computations were performed using the OptAlloc macro for SAS which implements generalized power allocations. Further details and a description of the algorithm are given in Hohnhold [2010].

Since the population of businesses is very heterogeneous with respect to turnover and the heterogeneity is most pronounced for the largest enterprises, minimizing (1) gives rise to highly different sampling fractions n_h/N_h among the strata. Furthermore, many take-all

strata which are fully enumerated are obtained, especially in the group of large enterprises. Owing to a decision taken by the Federal Administrative Court of Germany in March 2017, however, the sampling design in the structural survey might undergo changes in the future. Specifically, the court required that the response burden should be spread as evenly as possible and that take-all strata are only acceptable if they are imperative to obtain estimates which are sufficiently representative [Bundesverwaltungsgericht, 2017]. Note that while these requirements could be addressed by incorporating additional constraints when minimizing (1), the precision of the resulting estimates is likely to deteriorate. This is easily explained as the sampling variance of a take-all stratum is zero and owing to the skewed distribution of turnover, the population variance within the take-all strata typically exceeds the population variance within other strata by far. Hence, the sampling variance of the stratum groups increases.

A natural approach is to consider alternative estimation methods using auxiliary information in order to alleviate the efficiency losses due to potential changes in the sampling design to some degree. In the following, we explore different sources of auxiliary information and offer first insights regarding their applicability for the structural survey in the service sector. For this purpose we use survey data for the reporting year 2014, where $n = 192{,}110$ enterprises were sampled.

2 Estimation methods

To produce and publish results in a flexible manner, estimators that can be written as a linear combination of the observations in the sample are frequently used by statistical agencies, i.e.

$$\widehat{Y} = \sum_{k \in S} w_k y_k, \tag{4}$$

where w_k and y_k are the weight and the observed value of the variable of interest for unit k, respectively, while S denotes the set of sampled units. A popular choice for w_k are the inverse inclusion probabilities, i.e. $w_k = d_k = 1/\pi_k$, where $\pi_k = Pr(k \in S)$, leading to the Horvitz-Thompson estimator, which is design-unbiased. Another class of estimators are the calibration estimators due to Deville and Särndal [1992], where the weights are found by minimizing

$$\sum_{k \in S} G(w_k, d_k) \quad \text{subject to} \quad \sum_{k \in S} w_k \mathbf{x}_k = \mathbf{X}. \tag{5}$$

In (5) $G(w_k, d_k)$ denotes a distance function, whereas $\mathbf{x}_k$ and $\mathbf{X}$ refer to the vector of auxiliary information for unit k and the known vector of population totals of the auxiliary information, respectively. Hence, under the calibration approach weights are determined which are close to the inverse inclusion probabilities but which reproduce known totals when applied to the auxiliary variables present in calibration constraints. Estimators based on calibrated weights

have many attractive properties [cf.Deville and Särndal, 1992], but for our purpose their ability to produce estimates with a smaller variance than the Horvitz-Thompson estimator is particularly important. Popular choices for the distance function include

$$G(w_k, d_k) = \frac{(w_k - d_k)^2}{2 \cdot d_k} \tag{6}$$

leading to the generalized regression (GREG) estimator and

$$G(w_k, d_k) = w_k \cdot \log(w_k/d_k) - w_k + d_k, \tag{7}$$

which is also known as the raking ratio approach. An advantage of (7) is that it guarantees strictly positive weights, unlike (6).

3 Data and Analyis

The first potential data source for auxiliary information to be used at the estimation stage is the sampling frame itself. The sampling frame is based on the Unternehmensregister (business register) for the reporting year 2013, which was the latest register data available when the sample was drawn. Since the reporting year for the survey is 2014 an alternative data source is the business register 2014 (BR). A third option would be to consider the Verwaltungsdatenspeicher (administrative data record), which are available 6 months after the reporting month and are therefore, the latest source of auxiliary information. So far, we focussed on the sampling frame and the business register with reporting year 2014 in our analyses.

Table 1 displays the number of missing values of potential calibration variables after matching auxiliary information to the survey data for the different size classes (SC). Columns beginning with SF refer to auxiliary information taken from the sampling frame, whereas columns with BR refer to variables obtained from the business register 2014. Moreover, TUR indicates the variables for turnover and EMP the variables with the number of employees. The large fraction of units with missing auxiliary information for turnover is explained by the fact that the size class 1 was designed to comprise units with missing or very small values for the reported turnover in the sampling frame. Moreover, large shares of missing information for the number of employees in the sampling frame is evident. Except for the latter variable, the share of missing values for enterprises in size classes 5 and higher is below 3 %.

Furthermore, the analysis of the survey data showed that 15,206 enterprises were misclassified, i.e. they were sampled but do not belong to the population of interest. It should be noted that this issue is also relevant for the non-sampled units, but the information whether a unit is misclassified is only known for the sampled part. Thus, to avoid calibrating to inflated population totals we fitted a misclassification model using the survey data with the purpose to delete units in the auxiliary data files. Specifically, the logistic regression model estimated by pseudo maximum likelihood included the size class, NUTS1, NACE4 and an interaction

Table 1: Percentage of missing values after matching auxiliary information to survey data

SC	SF-TUR	BR-TUR	SF-EMP	BR-EMP
1	81.4	71.6	0	8.0
2	1.3	8.6	85.9	8.3
3	1.6	5.9	68.2	4.8
4	0.7	3.3	43.4	2.8
5	0.9	2.8	24.7	2.0
6	1.0	2.4	15.3	1.6
7	0.8	2.0	8.5	1.7
8	0.3	1.2	3.5	1.7

term between size class and NUTS1 as covariates. We then used the coefficients from this model to obtain predictions for misclassification in auxiliary data files. Finally, a unit was set to be misclassified when the realization of a uniformly distributed random variable between 0 and 1 took a value less than the predicted probability of misclassification.

We calibrated against variables taken from the sampling frame and the business register. An overview for some of the specifications we tried is given in Table 2. For each specification, we first applied GREG-weighting and then, in case that some negative weights where obtained, we additionally applied the raking ratio approach, where the distance function is given by (7). The last column in Table 2 shows the R^2 from the implicit model, i.e. a regression from the reported turnover in the survey on the variables given in the column *Calibration Constraints*. As an example the model SF-1 uses auxiliary data from the sampling frame, calibrates against the number of units in the population for each NUTS1-, NACE2-, and SC-level as well as the total turnover for each NUTS1-, NACE2-, and SC-level, where the coefficient of determination from the corresponding regression model equals $R^2 = 0.94$.

A comparison between SF-1 and the more parsimonious model SF-2 shows that the former model has a better fit. However, owing to the large number of calibration constraints, a few negative GREG-weights occured, whereas SF-2 did not yield negative weights for the GREG. Note that we did not incorporate calibration constraints for the number of employees when calibrating to variables from the sampling frame, because of the large share of missing values [cf. Table 1]. Negative GREG-weights also occurred under models BR-1 and BR-2.

A comparison of the g-weights defined as $g_k = w_k/d_k$ is shown in Figure 1. It can be seen that the most parsimonious model SF-2 yields g-weights which are much less dispersed than all other choices. Furthermore, we observe the specification BR-2, which calibrates against a smaller number of constraints than BR-1, leads to a smaller range of g-weights as compared to BR-1. Additionally, larger average values result for the weights under the business register, which is due to the fact that estimated population size in terms of numbers of units is larger in business register than in the sampling frame.

Table 2: Models for predicting turnover

Model	Source	Calibration constraints	R^2
SF-1	Sampling Frame	N for NUTS1, NACE2, SC total TUR for NUTS1, NACE2, SC	0.94
SF-2	Sampling Frame	N for NUTS1, NACE2, SC total TUR for NUTS1	0.86
BR-1	Business Register	N for NUTS1, NACE2, SC total TUR for NUTS1, NACE2, SC	0.91
BR-2	Business Register	N for NUTS1, NACE2, SC total TUR for NUTS1 total EMP for NUTS1	0.90

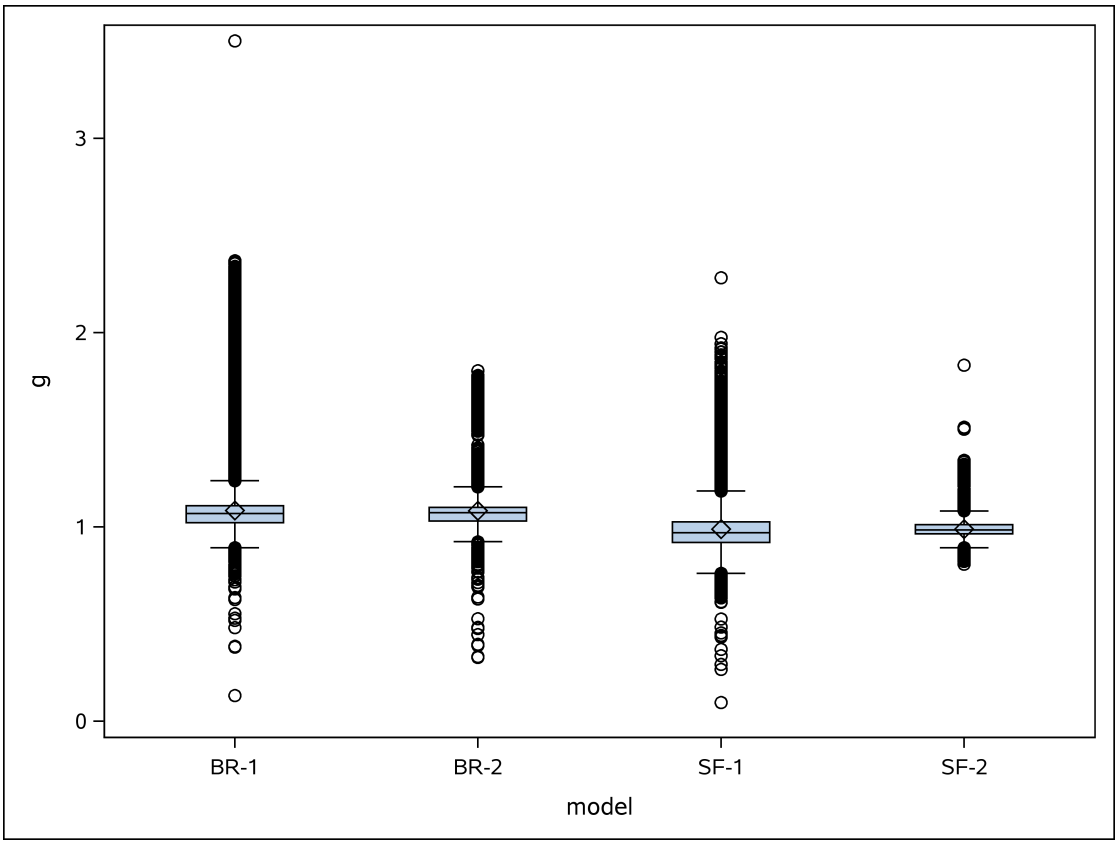

Figure 1: Distribution of the g-weights

4 Summary and Outlook

A preliminary result of our work is that by using auxiliary data sources, calibration estimators have the potential to produce estimates with smaller variances than the Horvitz-Thompson method. This issue might be relevant in the future, if the sampling design faces restrictions with respect to take-all strata. The choice of the auxiliary data source is of particular importance, as using data from the business register of the reporting year of the survey or the current administrative data record may foster coherence of the estimate with other currently produced statistics.

A prerequisite for the successful application of calibration estimators is that the calibration constraints accurately reflect the target population. Moreover, the choice of the calibration constraints has to be carefully chosen, where a trade-off between the explanatory power of the statistical model and the variation of the g-weights exists. Additionally, it should be noted that while our analyses focussed on predicting turnover so far, the structural survey is used to produce estimates for other variables such as the total number of employees as well. This issue should also be addressed when specifying the implicit model.

References

Michael D. Bankier. Power allocations: Determining sample sizes for subnational areas. *The American Statistician*, 42(3):174–177, 1988. ISSN 00031305. URL `http://www.jstor.org/stable/2684995`.

Bundesverwaltungsgericht. BVerwG 8 C 6.16 vom 15.03.2017: Heranziehung zur Dienstleistungsstatistik. 2017.

Jean Claude Deville and Carl Erik Särndal. Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418):376–382, 1992.

Henning Hohnhold. Generalized power allocations. Statistisches Bundesamt, Wiesbaden, 2010.

Karl August Schäffer. Planung geschichteter Stichproben bei Vorgabe einer Fehlerabstufung. *Allgemeines Statistisches Archiv*, 45:350–361, 1961.