vom Berge *J Labour Market Res* (2023) 57:5
https://doi.org/10.1186/s12651-023-00334-x

Journal for Labour Market Research

ORIGINAL ARTICLE Open Access

Population aggregates from administrative data samples–how good are they?

Philipp vom Berge*

Abstract

Researchers regularly use administrative micro-data samples to approximate subgroup aggregates from the full population. In this paper, I argue that the most commonly used method to do this is often not optimal. I outline some alternatives and compare their relative performance in selected cases. I also discuss the effect of statistical disclosure control on the aggregated data and how researchers can reduce bias resulting from censoring.

Keywords SIAB, BHP, Microdata, Aggregation, Statistical Disclosure Control

JEL Classification C43, C80

1 Introduction

Many empirical studies in the social sciences require aggregated subgroup data from a population. Examples for such data are the number of workers per district, average wages per sector or the unemployment rate for people of different nationalities. However, suitable data based on the total population are not always available. When working with German labour market data, many researchers therefore resort to approximations based on administrative microdata samples. In doing so, they often apply the "bigger-is-better" principle: Choose the data product with the highest total number of observations, aggregate without further adjustments and hope for the best.

In this paper, I review this frequently used strategy and provide some critical comments and suggestions. I do this, first, by reminding everyone of the obvious: Aggregating a sample inevitably leads to sampling error, and this error can quickly become significant, in the worst case calling into question any follow-up analysis. Second, I show that in many cases it is not advisable to rely on the intuition "bigger-is-better" and that the right choice of data source can significantly improve the performance of the approach. Third, I propose a simple data smoothing procedure that reduces the asymmetric approximation error in cases with many sparsely populated cells. Forth and finally, I discuss the impact that statistical disclosure control might have on the final aggregated data and suggest some ways to reduce bias resulting from censoring.

The easiest way to get around problems induced by aggregating from samples would, of course, be for empirical researchers to always use the entire population under study. In many cases this will not be a problem. In Germany, the Federal Statistical Office provides a variety of official statistics divided into different subgroups, many of which are based on the entire population. For labour market research, the Statistics Department of the Federal Employment Agency provides many official statistics free of charge, and individualized data extracts are also available for a fee upon request.

However, there are a number of reasons why researchers may not be able to use these data sources. A typical example is that the research question requires a long panel of aggregated subgroup data, but official statistics only provide a much shorter period. Another example is that researchers need to perform more complex initial data preparation steps before the final aggregation, but this service is not offered by the respective statistics

*Correspondence:
Philipp vom Berge
philipp.vom-berge@iab.de
Institute for Employment Research (IAB), Regensburger Str. 104,
90478 Nuremberg, Germany

department. In such cases, recourse on administrative micro-data samples might not be a second-best alternative only, but the only option available.

When it comes to accessing administrative microdata on the German labour market, the most obvious first port of call is the Research Data Centre of the Federal Employment Agency at the Institute for Employment Research (IAB-FDZ) with its broad portfolio of administrative microdata, surveys and linked data products. Access is limited to a scientific research context and three ways of accessing data, namely on-site access, remote execution and the provision of Scientific Use Files, depending on the data product. In this paper, I will focus on experience from working with data users at the IAB-FDZ who use its data products.[1] I hope that some of my remarks can be transferred to other contexts.

After several years as a staff member at the IAB-FDZ, I can confirm with some certainty that the demand for subgroup aggregates from administrative microdata samples is quite high. While many of these aggregations take place as intermediate steps of data preparation in a secure data processing environment, there are also frequent requests for the release and export of the resulting data, which then requires manual data disclosure control, a task that ties resources in a Research Data Centre (RDC). Sometimes, data users are not aware that there are ways to access similar or even better data from another source, in which case IAB-FDZ staff can help. In most cases, however, there are some good reasons why researchers decide to rely on microdata.

The Establishment History Panel (BHP) is the most frequently requested data product for studies using aggregated microdata from IAB-FDZ. It is a 50 per cent random sample of establishments in Germany based on social security notifications, currently covering the years 1975 to 2020. It contains information on up to 1.5 million establishments annually and is thus a natural candidate for aggregations of detailed subgroups. Another IAB-FDZ data product that could be considered as a competitor, the Sample of Integrated Employment Biographies (SIAB) which focuses on individuals, is “only” a 2 per cent random sample and therefore seems much less attractive because "Bigger-is-better". In my experience—which may not be representative of the typical data user—many researchers then prefer to ignore the fact that their aggregated data is constructed from a sample, perhaps due to the fact that the BHP is such a large dataset and the intuition that this will lead to small approximation errors.

In the following, I would like to provide data users of the IAB-FDZ—and users of data samples in general—with some descriptive analyses to help them get a feel for the data quality costs of aggregating administrative data samples at increasingly detailed levels. For example, using the BHP to aggregate employment at the district level results in a ratio of aggregate error to total employment of about 5 per cent. When aggregating at the district level in conjunction with 3-digit industries, this ratio increases to 32 per cent.

Not only are these errors significant, but I also provide evidence that the BHP is likely to be inferior to the SIAB in approximating employment aggregates despite its bigger sample size. Switching to the SIAB can reduce error ratios by more than half in the examples I tested. Furthermore, I show that for many applications it is best to use the BHP add-on to the SIAB (the SIAB Basis Establishment File) rather than the BHP itself.

Selecting a fine level of detail for subgroup aggregation not only increases the overall error ratio, but also leads to an asymmetric distribution of approximation errors for some aggregation methods when true cell sizes become really small and often contain only one observation. For example, when the sampling probability is small, not selecting this one observation leads to an error of -1, while selecting it leads to an error equal to the weighting factor minus one. I show that this asymmetry can be reduced by a simple data smoothing step and that this can also improve the overall performance of the approximation.

Finally, I show how statistical disclosure control, which may require censoring of data before an aggregated dataset is released, can significantly worsen the approximation error. This poses a problem for data users who plan to export their aggregated data from the secure computing environment of an RDC for use in a later phase of their project or publication. I argue that the smoothing technique also proposed in the paper, or alternatively aggregation methods with random weighting factors, can mitigate this problem and reduce the need for censoring.

The remainder of the paper is structured as follows. In Sect. 2, I present the basic approximation variants and show that bigger is not necessarily better. In Sect. 3, I move to a finer level of aggregation and document and discuss the increase of approximation error. Sect. 4 presents the data smoothing procedure. In Sect. 5, I discuss the implications of statistical disclosure control. Sect. 6 provides a conclusion.

2 Approximation and the choice of data source

Suppose a researcher wants to create a dataset of aggregate employment figures for Germany at the district level over several years, and for some reason official statistics

[1] This paper focuses on aggregates from administrative samples, omitting aggregation based on surveys or linked data products.

Table 1 Quality indicators for cell deviations from target (full population version of BHP)—district level

	(i)	(ii)	(iii)	(iv)	(v)
	Workers			Establishments	
	BHP 50%	SIAB Individual File	SIAB Basis Establishment File	BHP 50%	SIAB Basis Establishment File
Absolute deviation					
Mean	– 488	– 196	– 207	1	– 74
p-value	0.001	0	0	0.288	0
rmse	8844	2103	1539	83	515
mae	4727	1574	1137	62	351
Ratio	0.054	0.018	0.013	0.008	0.047
Percentage deviation					
Mean	0	– 0.003	– 0.002	0	– 0.006
p-value	0.849	0	0	0.039	0
Mape	0.06	0.024	0.018	0.011	0.057
Percentage deviation (size weighted)					
Mean	– 0.006	– 0.002	– 0.002	0	– 0.01
p-value	0	0	0	0.288	0
mape	0.054	0.018	0.013	0.008	0.047
N	4010	4010	4010	4010	4010

The table shows various quality indicators for cell deviations from the target dataset (the full population version of the BHP) at the district level, including absolute deviations, percentage deviations and percentage deviations weighted by cell size. Approximations are calculated for the number of workers and establishments. Calculations are based on the 50 percent sample of the BHP, the SIAB Individual File and the SIAB Basis Establishment File, respectively. Indicators are the mean error (mean), root mean squared error (rmse), mean absolute error (mae), mean absolute percentage error (mape) and ratio of the total sum of errors to the total sum of cell counts (ratio). p-values for a t-test of (mean) against zero are shown in (p-value)

Sources: Establishment History Panel (BHP)—Version 7519 v2 (https://doi.org/10.5164/IAB.BHP7519.de.en.v2); Sample of Integrated Labour Market Biographies (SIAB)—Version 7519 v1 (https://doi.org/10.5164/IAB.SIAB7519.de.en.v1), own calculations

are not readily available (in fact they are). The researcher decides to approximate the population aggregates from administrative microdata samples and is now faced with a choice, as there are several different data products available that might be suitable. In this section, I will use this example to argue that researchers should focus primarily on what they want to measure and what the relevant unit of observations is, rather than on identifying the nominally largest dataset containing the largest sample.

For this purpose, I first use the largest sample available via the IAB-FDZ, the BHP.[2] This data product contains a 50 percent random sample of establishments in Germany based on social security notifications (see Ganzer et al. 2021, for further details). I aggregate establishment-specific employment counts at the district level, forming a panel spanning the years 2008 to 2017. I then multiply the resulting employment in each district by two as a simple way to account for the sampling. To test the quality of this approximation, I need to compare it to aggregates from the 'true' target population. Here, I choose the full population version of the BHP, an IAB-internal dataset not available to external researchers via the IAB-FDZ, and aggregate it in a similar way (but without weighting). This gives me a total of 4,010 data cells that can be compared to each other.

The first column of Table 1 contains several quality indicators for this comparison. On average, the approximated district-level employment figures slightly, but statistically significantly, underestimate the actual values by 488 workers. For later comparisons, I also report the root mean squared error (rmse) of 8844, the mean absolute error (mae) of 4,727, and the ratio of the total sum of errors to the total sum of cell counts (ratio) of 0.054. In addition, I give indicators calculated for the percentage deviation from the true values at the district level, both unweighted and weighted by cell size, the latter taking into account the fact that sampling error is relatively larger in small cells. Here, the average percentage deviation is basically zero or slightly negative, and the mean absolute percentage deviation (mape) is 5 to 6 per cent, depending on weighting.

[2] For the purpose of this exercise, I use Version 7519 v2 of the BHP (https://doi.org/10.5164/IAB.BHP7519.de.en.v2).

Next, I would like to check how good this approximation is by presenting two alternative variants. The second variant is still quite simple, but is used less frequently because it is based on the "smaller" SIAB.[3] This data product contains a 2 per cent random sample of employment biographies in Germany and is based on social security notifications and other process data of the Federal Employment Agency (see Frodermann et al. 2021, for further details). After some initial data selection steps to make the worker data in the SIAB Individual File as comparable as possible to the BHP selection,[4] I aggregate employment figures at the district level and again form a panel covering the years 2008 to 2017. This time I multiply the resulting employment in each district by 50.

The third variant I would like to propose is, to my knowledge, not widely used among researchers working with IAB-FDZ data. It takes advantage of the fact that establishments in the SIAB Basis Establishment File, a BHP-style add-on to the main Individual File (which contains the individual biographies), have a sampling probability that, by construction, is proportional to the size of the establishment. Therefore, the aggregation of the establishment-specific employment figures in the SIAB Basis Establishment File—after weighting by the inverse of the probability that an establishment is in the sample—can also serve as a useful proxy. With establishment size s, the weighting factor in this case can be written as:

$$\omega = 1/\left(1 - 0.98^{s}\right).$$

To get a first overview of the relative performance of the three approximation methods described above, Fig. 1 shows the distribution of the respective deviations from the target population size. It shows that both the approximation using the SIAB Individual File (SIAB 2%) and the one using the SIAB Basis Establishment File (SIAB-estab) perform visibly better than the BHP-based aggregation (BHP 50%), which seems to be the default choice for many researchers. The graphical inspection is confirmed by the set of quality indicators in Table 1, where the second column shows the indicators for the variant using the SIAB Individual File and the third column shows the variant with the SIAB Basis Establishment File. All in all, the third variant seems to perform best.

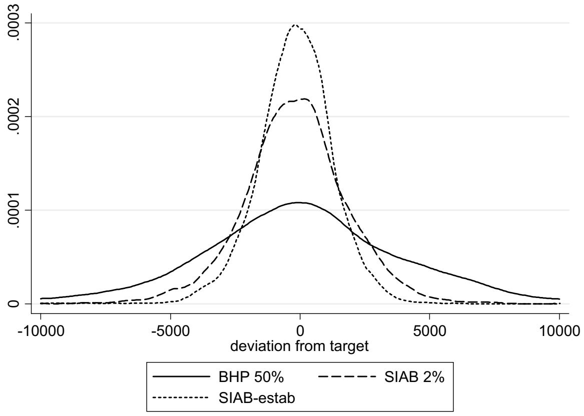

Fig. 1 Distribution of deviations from target (full population version of BHP)—district level. The figure shows the distribution of absolute cell deviations from the target dataset (the full population version of the BHP) at the district level (range ± 10,000). Approximations are calculated for the number of workers. Calculations are based on the 50 per cent sample of the BHP (BHP 50%), the SIAB Individual File (SIAB 2%) and the SIAB Basis Establishment File (SIAB-estab), respectively. Sources: Establishment History Panel (BHP)—Version 7519 v2 (https://doi.org/10.5164/IAB.BHP7519.de.en.v2); Sample of Integrated Labour Market Biographies (SIAB)—Version 7519 v1 (https://doi.org/10.5164/IAB.SIAB7519.de.en.v1), own calculations

Does this result imply that the third variant is clearly superior and should always be used? Not necessarily. There are a couple of arguments against such a far-reaching conclusion. First, one must bear in mind that the BHP—and thus also the SIAB Basis Establishment File—only allows for a limited range of aggregations because of its structure. Aggregations by detailed nationalities, occupations or age groups are only possible using the second variant (using the SIAB Individual File). Second, it still depends on what exactly one is trying to measure. To illustrate this point, the fourth and fifth columns of Table 1 repeat the previous exercise for the BHP and SIAB Basis Establishment File, but this time approximating the number of establishments per district instead of the number of workers. Here the variant using the BHP 50 per cent sample performs much better, because the relevant unit of observation is now the establishment, fitting the BHP sampling scheme.[5] Third, the extent to which the differences in approximation error matter could depend crucially on how large the approximation error is overall and what the researcher intends to do next with the aggregated dataset. If the data are used for some simple descriptions, any of the variants might work well enough. However, once the approximated variable enters regression analysis, one must consider the potential for reduced precision (in case the variable enters as a

[3] I use Version 7519 v1 of the SIAB (https://doi.org/10.5164/IAB.SIAB7519.de.en.v1).

[4] These steps include aligning the reference date, selecting person groups and keeping relevant spells. Details can be found in the code provided in the supplementary data file.

[5] Trying to approximate the number of establishments from the SIAB Individual File is hopeless, and I therefore did not include this variant in the table.

dependent variable) or attenuation bias (in case it enters as an explanatory variable).

To conclude this chapter, I would like to suggest that—at least at the high level of aggregation considered so far—the choice of variant for an approximation may not matter too much and that all variants will perform quite well in many applications. In Sect. A of the Online Appendix (Additional file 1), I provide evidence for this conjecture using three simple examples. However, this may well change when we consider finer levels of aggregation. Unfortunately, finer levels of aggregation are exactly what researchers usually want. This will be the topic of the next section.

3 Finer levels of aggregation

For some empirical applications, researchers conclude that they need very fine levels of aggregations, including very detailed categories or the combination of several categorial variables. The idea behind this is that important variation for the identification process can only be modelled properly at this level. However, this decision comes at a cost. At such a level of detail, many cells in the population are already sparsely populated, and many more in available samples. This leads to a sharp increase in the approximation error, at least when assessed in terms of percentage deviations. In addition, there will be an increasing number of cells that are not observed in the sample, although they are present in the population.

For illustration purposes, I repeat the exercise of the previous section, but this time I use the interaction of districts and 3-digit-industries as the level of aggregation instead of districts alone.[6] This leads to a total of 789,616 cells for the analysis, which is summarized in Table 2. Some results stand out in particular. First, the approximation variant using the SIAB Basis Establishment File (shown in column 3) again seems to perform best for workers, at least most of the time. Only in the unweighted percentage deviations do the SIAB-based variants perform worse than the BHP-based variant, with a mean absolute percentage deviation of more than 70 per cent. This result highlights the fact that there are now many more cells containing both few and small establishments, a combination where the BHP-based variant has less approximation error. Second, the BHP-variant (column 4) still performs best in approximating of the number of establishments. Third, regardless of the choice of variant, the quality of the approximation is significantly lower than in the case presented in the previous section. While not all quality indicators can be directly compared with the results in Table 1, some can. For workers, the ratio between the total sum of errors and the total sum of cell counts is now 0.132 (in column 3), compared to only 0.013 in Table 1. The mean absolute percentage errors are also much higher.

For applied research, this presents a dilemma. A supposed gain in precision might be countered, unnoticed, by a reduction of precision and an additional bias that could more than outweigh any potential gain. In Sect. B of the Online Appendix (Additional file 1), I provide simple examples where the choice of approximation method affects the outcome, albeit to a relatively small extent. Nonetheless, the results presented here should be a warning to data users not to underestimate the potential complications of choosing fine aggregation levels for their data. This is all the more true as some requests for aggregated data we receive at the IAB-FDZ require even finer aggregations, such as a combination of districts, 5-digit-industry codes and worker qualification. Against the background of the results discussed above, such aggregations run a considerable risk of producing biased results.

4 Smoothing out asymmetric approximation errors

A disadvantage of using fixed weighting factors to approximate full population aggregates based on random samples is that the fixed weights lead to unevenness in the distribution of approximation errors in cases with many small cells. The simplest case—and also the most frequently observed—is where there is actually only one establishment with a single worker in a cell of the population, and this establishment is randomly selected for the sample. In the SIAB-based approximations, the observation is weighted by a factor of 50, resulting in a deviation from target of 49. If the observation is not selected, the deviation from target is -1. The effect of this difference is depicted in Fig. 2, where the solid line (SIAB-estab) shows the distribution of deviations from the target value of the population for the approximation variant using the SIAB Basis Establishment File (column 3 in Table 2). Another disadvantage of using the fixed weighting factors is that if no observation is included in the sample, no adjustment is made and the cell remains empty although it is filled in the population.

One way to deal with these issues would be to work with random weighting factors (see Sect. 5), but this does not necessarily lead to better outcomes. Here I propose a smoothing procedure that worked quite well in the cases I tested. The procedure works in three simple steps. The first step starts from the fixed-weight approximation using the SIAB Basis Establishment File and calculates the difference between the weighted and unweighted cell

[6] I also performed the exercise at the level of 5-digit-industries. Results are not depicted here but can be found in the supplementary data file accompanying the article.

Table 2 Quality indicators for cell deviations from target (full population version of BHP)—district#3-digit-industry level

	(i) Workers	(ii)	(iii)	(iv) Establishments	(v)
	BHP 50%	SIAB Individual File	SIAB Basis Establishment File	BHP 50%	SIAB Basis Establishment File
Absolute deviation					
Mean	− 2.5	− 1.0	− 1.1	0.0	− 0.4
p-value	0	0	0	0.26	0
rmse	596.3	148.8	101.3	6.1	31.7
mae	144.1	82.9	59.1	3.4	14.1
Ratio	0.323	0.186	0.132	0.092	0.376
Percentage deviation					
Mean	0	0	0	0	− 0.001
p-value	0.679	0.893	0.907	0.438	0.793
mape	0.519	0.787	0.739	0.372	0.953
Percentage deviation (size weighted)					
Mean	− 0.006	− 0.002	− 0.002	0	− 0.01
p-value	0	0	0	0.26	0
mape	0.323	0.186	0.132	0.092	0.376
N	789,616	789,616	789,616	789,616	789,616

The table shows various quality indicators for cell deviations from the target dataset (the full population version of the BHP) at the district#3-digit-industry level, including absolute deviations, percentage deviations and percentage deviations weighted by cell size. Approximations are calculated for the number of workers and establishments. Calculations are based on the 50 percent sample of the BHP, the SIAB Individual File and the SIAB Basis Establishment File, respectively. Indicators are the mean error (mean), root mean squared error (rmse), mean absolute error (mae), mean absolute percentage error (mape) and ratio of the total sum of errors to the total sum of cell counts (ratio). p-values for a t-test of (mean) against zero are shown in (p-value)

Sources: Establishment History Panel (BHP)—Version 7519 v2 (https://doi.org/10.5164/IAB.BHP7519.de.en.v2); Sample of Integrated Labour Market Biographies (SIAB)—Version 7519 v1 (https://doi.org/10.5164/IAB.SIAB7519.de.en.v1), own calculations

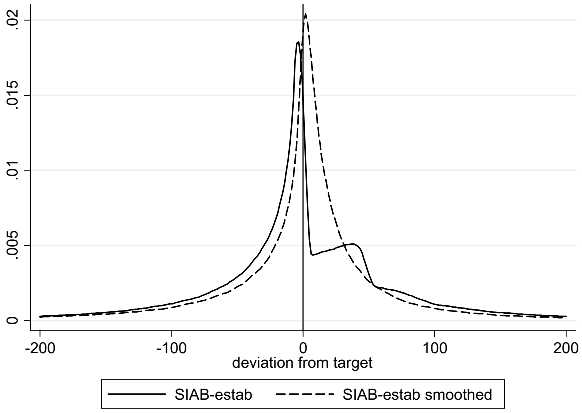

Fig. 2 Distribution of deviations from target (full population version of BHP)—district#3-digit-industry level. The figure shows the distribution of absolute cell deviations from the target dataset (the full population version of the BHP) at the district#3-digit-industry level (range ± 200). Approximations are calculated for the number of workers. Calculations are based on the SIAB Basis Establishment File (SIAB-estab), with and without a smoothing adjustment. Sources: Establishment History Panel (BHP)—Version 7519 v2 (https://doi.org/10.5164/IAB.BHP7519.de.en.v2); Sample of Integrated Labour Market Biographies (SIAB)—Version 7519 v1 (https://doi.org/10.5164/IAB.SIAB7519.de.en.v1), own calculations

counts (the uncertain part of the approximation). In the second step, this difference is fed into a Poisson model containing the cell categories as fixed effects. A Poisson model is chosen because of the count data nature of the variable, which will include many zeros. The categories should be interacted in a flexible manner, but the model should not be fully saturated, otherwise no smoothing will take place.[7] In the third step, the predictions from this model are added to the unweighted cell counts, resulting in the final smoothed measure. The performance of this measure can be seen in Fig. 2, where it is depicted as the dashed line (SIAB-estab smoothed). The approximation error is now distributed more evenly.[8]

[7] For the model used here, I added fixed effects for the interactions of (i) districts with 1-digit-industries as well as (ii) states with 3-digit-industries to the model.

[8] This approach will work much worse for the approximations using the BHP 50 per cent sample or the SIAB Individual File. This is because in these cases, the ratio of weighted to unweighted cell counts is much higher, which leads to a larger fraction of 'uncertain' observations that need to be distributed via the smoothing procedure.

Table 3 Quality indicators for cell deviations from target (full population version of BHP)—district#3-digit-industry level (with and without smoothing)

	(i)	**(ii)**	**(iii)**	**(iv)**
	Workers		**Establishments**	
	SIAB Basis Establishment File (no smoothing)	**SIAB Basis Establishment File (smoothing)**	**SIAB Basis Establishment File (no smoothing)**	**SIAB Basis Establishment File (smoothing)**
Absolute deviation				
Mean	− 0.8	− 0.8	− 0.3	− 0.3
p-value	0	0	0	0
rmse	86.2	96.9	27.0	23.6
mae	42.8	38.5	10.2	6.4
Ratio	0.132	0.119	0.376	0.234
Percentage deviation				
Mean	− 0.547	0.545	− 0.693	0.445
p-value	0	0	0	0
mape	0.849	0.808	1,035	0.83
Percentage deviation (size weighted)				
Mean	− 0.002	− 0.002	− 0.01	− 0.01
p-value	0	0	0	0
mape	0.132	0.119	0.378	0.236
N	1,090,720	1,090,720	1,090,720	1,090,720

The table shows various quality indicators for cell deviations from the target dataset (the full population version of the BHP) at the district#3-digit-industry level, including absolute deviations, percentage deviations and percentage deviations weighted by cell size. Approximations are calculated for the number of workers and establishments. Calculations are based on the SIAB Basis Establishment File, with and without a smoothing adjustment. Indicators are the mean error (mean), root mean squared error (rmse), mean absolute error (mae), mean absolute percentage error (mape) and ratio of the total sum of errors to the total sum of cell counts (ratio). p-values for a test of (mean) against zero are shown in (p-value)

Sources: Establishment History Panel (BHP)—Version 7519 v2 (https://doi.org/10.5164/IAB.BHP7519.de.en.v2); Sample of Integrated Labour Market Biographies (SIAB)—Version 7519 v1 (https://doi.org/10.5164/IAB.SIAB7519.de.en.v1), own calculations

However, there is a problem with the smoothing procedure not shown in Fig. 2. Since the researcher performing the approximation from a sample dataset does not know whether an empty cell it is a true or false negative when she observes it (since the full population is not available), she can only use a fully rectangular matrix of cell elements for the Poisson model. This means that the procedure provides positive estimates not only for the empty cells in the sample dataset that are actually not empty in the full population, but also for true negatives. In cases where the number of true empty cells is very high compared to the number of cells with positive counts in the full population but zero observations in the sample, the smoothing procedure might not work very well.

In Table 3, the smoothed approximation is contrasted with its non-smoothed counterpart, also taking into account truly empty cells, thus allowing for an overall comparison. Since regular percentage deviations from target do not work when the target is zero, I replace them by the measure

$$2 * \frac{\widehat{N}-N}{\widehat{N}+N}$$

where N denotes true employment (or establishment) counts from the full population and $\widehat{N}$ denotes the smoothed approximation. I also use $\left(\widehat{N}+N\right)/2$ instead of N when a weighting factor is required. Due to these measurement differences, the unsmoothed results in columns 1 and 3 of Table 3 are not identical to those of columns 3 and 5 of Table 2. Overall, Table 3 confirms that the smoothing procedure can improve the approximation in a non-trivial way. The only quality indicator where smoothing actually has a detrimental effect is the root mean squared error for employment, which increases from 86 to 97, probably due to some outliers. This improvement comes despite the fact that of the 1,090,720 cells that make up the fully rectangular matrix of districts, industries and years, 301,104 are themselves empty in the population and 249,629 are empty only in the sample, resulting in a ratio of about 1.2:1 of true to false empty cells.

In Sect. B of the Online Appendix (Additional file 1), I show that the smoothed approximation works about as well as the unsmoothed variant in a simple linear regression where the approximated employment number enters as the right-hand-side variable, but only in cell-size-weighted regressions. For unweighted regressions, the large number of empty cells to which some information is added can lead to problems. However, approximation smoothing brings some additional advantages when aggregated data need to be subjected to additional disclosure control, a point I will discuss in more detail in the next section.

5 Disclosure control and cell suppression

In the previous sections, I have implicitly assumed that researchers can work with the approximated aggregated datasets without further constraints. However, this will not always be the case. When data from sensitive data products are processed in the secure data processing environment of an RDC, additional disclosure control is required before the data can be exported. Data users can of course circumvent this restriction by keeping the data in the secure data processing environment throughout the research project and performing all analyses there, but sometimes they still want to export the data.[9] The reasons for this range from the desire to share data with other researchers or to facilitate the reproducibility of published results, to software restrictions within the secure computing environment, to sheer convenience.

While data protection rules for aggregated tables vary from RDC to RDC, there are some common themes. As a rule, cells with observation counts below a certain threshold are censored (primary suppression). In addition, other cells might need to be censored to prevent recalculation of the cells censored in the first step (secondary suppression). There may also be rules to avoid revealing groups information or to filter out dominant observations. For a more detailed discussion of rules and principles of output control, I refer interested readers to Brandt et al. (2010).

At the IAB-FDZ, primary suppression is performed at a threshold level of 20 observations, both for individuals and establishments. This means that researchers would not be able to export the data underlying Table 2 without adjustments, and the censoring at the fine aggregation level of districts interacted with 3-digit-industries would be massive. For example, using the SIAB Individual File to approximate the population aggregates, almost 32 per cent of available cells would remain empty due to sampling, more than 56 per cent would have to be censored, and only 12 per cent could be used for analysis. Using the BHP, 11 per cent of cells would remain empty, 69 per cent would have to be censored, and 20 per cent could be exported unadjusted.[10] This poses a high risk to the reliability of any follow-up analysis, and I therefore warn against exporting such subgroup aggregates.

In Table 4, I examine how these censoring rules affect the data quality of the aggregated data set for employment on the district#3-digit-industry level that was introduced in Sect. 3. Columns 1 and 3 show the uncensored versions of the approximation variants for the BHP and SIAB Individual File, for reference, and reproduce columns 1 and 2 from Table 2. In columns 2 and 4 of Table 4, I display the same indicators for the censored versions of the aggregated data.[11] It is obvious that the quality of the approximation deteriorates significantly. The ratio of the total sum of errors to the total sum of cell counts increases by about 75 per cent in both cases. Consequently, a researcher requesting such a data export must expect not only reluctant and highly critical RDC personnel, but also that the quality will deteriorate further for subsequent data analysis. In Sect. C of the Online Appendix (Additional file 1), I show this with a simple example.

I would like to propose two solutions for the censoring dilemma, assuming that data export cannot be avoided for good reason. The first goes back to the smoothing adjustment described in Sect. 4. Since smoothing involves a model that is not fully saturated, the cell counts or any other indicators for a given cell are a complicated combination of different values that obscure the original value and ensure privacy. In the example in Sect. 4, each cell also contains components that come from other districts in the same state and other industries within broader industry classes. If used correctly, this method can be sufficient to completely avoid cell suppression, allowing the dataset to be released without further adjustments.

In the second solution, applicable for the SIAB Individual File or comparable datasets, the fixed weighting factor (50 in case of the SIAB) is replaced by a random weighting factor tailored to the underlying sampling scheme. For the SIAB Individual File, I draw random numbers from a negative binomial distribution with success probability 0.02. The resulting integer weights account for the fact that each person in the sample represents an unknown number of additional individuals

[9] At IAB-FDZ, transferring aggregated datasets between projects is treated as an export, too.

[10] I only consider primary cell suppression, both for these shares and the following analysis, mostly because ensuring adequate secondary cell suppression would be painful for the data analyzed here. This means that the results presented in this section represent a lower bound for the actual problems arising from cell suppression.

[11] Instead of setting the censored cell to zero or missing, I replace it with the year-specific average size of all censored cells, to minimize information loss. Just ignoring censored cells—as is often done by practitioners—makes things even worse.

Table 4 Quality indicators for cell deviations from target (full population version of BHP)—district#3-digit-industry level—with censoring

	Workers				
	BHP 50% fixed weight (uncensored)	**BHP 50% fixed weight (censored)**	**SIAB Individual File—fixed weight (uncensored)**	**SIAB Individual File—fixed weight (censored)**	**SIAB Individual File—rand. weight (censored)**
Absolute deviation					
Mean	− 2.5	− 2.5	− 1.0	− 12.5	− 0.8
p-value	0	0.002	0	0	0.001
rmse	596.3	699.2	148.8	228.7	209.1
mae	144.1	250.5	82.9	149.4	111.4
Ratio	0.323	0.561	0.186	0.334	0.249
Percentage deviation					
Mean	0	12	0	1.452	0.179
p-value	0.679	0	0.893	0	0.004
mape	0.519	12.354	0.787	2.292	1.008
Percentage deviation (size weighted)					
Mean	− 0.006	− 0.006	− 0.002	− 0.028	− 0.002
p-value	0	0.002	0	0	0.001
mape	0.323	0.561	0.186	0.334	0.249
N	789,616	789,616	789,616	789,616	789,616

The table shows various quality indicators for cell deviations from the target dataset (the full population version of the BHP) at the district#3-digit-industry level, including absolute deviations, percentage deviations and percentage deviations weighted by cell size. Approximations are calculated for the number of workers, and censored cells are replaced by the year-specific mean cell size. Calculations are based on the 50 percent sample of the BHP (fixed weight: censoring below 20 establishments) and the SIAB Individual File (fixed weight: censoring below 20 workers; random weight: censoring below 4 workers), respectively. Indicators are the mean error (mean), root mean squared error (rmse), mean absolute error (mae), mean absolute percentage error (mape) and ratio of the total sum of errors to the total sum of cell counts (ratio). p-values for a test of (mean) against zero are shown in (p-value)

Sources: Establishment History Panel (BHP)—Version 7519 v2 (https://doi.org/10.5164/IAB.BHP7519.de.en.v2); Sample of Integrated Labour Market Biographies (SIAB)—Version 7519 v1 (https://doi.org/10.5164/IAB.SIAB7519.de.en.v1), own calculations

that have not been selected. Since the sampling probability in the SIAB is low, this will not only conceal the true cell size of the full population, but also of the sample. This eliminates the requirement for secondary cell suppression, as recalculation is not possible when all cells are obfuscated. Note, however, that primary suppression will still be necessary, albeit at a lower threshold.[12] Column 5 of Table 4 shows that a random weighting approximation variant for the SIAB Individual File, choosing a censoring threshold of 4 individuals in the sample dataset, performs significantly better than the fixed weights version in column 4 with a threshold of 20.[13]

In Sect. C of the Online Appendix (Additional file 1), I show that both suggestions presented here perform favourably when compared to the censored datasets in a simple regression example.

6 Conclusions

Empirical researchers regularly use samples of administrative microdata to approximate the population aggregates they ultimately care about for their data analysis. This comes at a cost, as the use of samples leads to sampling error that affect all subsequent analysis. Based on the results presented in this article, I want to draw some conclusions.

First, researchers should avoid samples altogether whenever possible for their aggregated datasets. Statistical offices can offer appropriate aggregated data based on the total population either free of charge[14] or for a fee.[15] Even if data extracts are chargeable and researchers

[12] The easiest way to see this is to think of an observer with the outside knowledge that the true cell count in the full population is one, but with no further knowledge. This observer could deduce, irrespective of the weighting factor and assuming the cell is not empty in the sample, that any additional information for that cell applies to that particular individual.

[13] Random weighting is not as effective in concealing true observations counts in the sample if the BHP is used. There will always be cases where the unweighted counts can be directly inferred from the weighted counts, no matter where we choose the censoring threshold.

[14] For Germany, it is always worthwhile visiting the Database of the Federal Statistical Office or, specifically for labor market data, the Website of the Federal Employment Agency's statistics department.

[15] For custom data extracts based on the full population of data available at the Federal Employment Agency, researchers can contact the Central Statistical Service at Zentraler-Statistik-Service@arbeitsagentur.de.

already have access to micro-data samples as part of their overall research project, they may consider the additional cost. However, it is important to remember that statistical disclosure control also applies to statistical offices, and some researchers may be surprised how quickly a country like Germany runs out of sufficient observations when detailed aggregates are requested.

Second, when aggregation of administrative microdata samples is necessary, researchers should always start by thinking about what they actually try to measure and what the appropriate unit of observation is. The "bigger-is-better" approach taken by many data users is not optimal in many cases and can lead to questionable decisions regarding data parsimony, administrative work and quality of the final research output.

Third and finally, researchers should not underestimate the effect statistical disclosure control might have on their intermediate data output, especially when working with finer levels of aggregation. Some readers may assess that the results for bias presented in this paper and the Online Appendix (Additional file 1) are still moderate and acceptable. However, I deliberately kept my examples narrow and simple, and many real applications will involve more—and more complex—measures, like ratios or flow data. Once researchers use several of those approximated measures in intricate models applying clever identification strategies to look for potentially small underlying effects, the effect of censoring becomes very difficult to assess.

As a result, data users should try to avoid getting intermediate data out of secure computing environments of RDCs, and rather perform their follow-up analyses there. They will not only gain everlasting gratitude by RDC staff, but also likely end up with more robust results. Limitations due to software, access to other data or reproducibility concerns should be addressed early on, because there might be alternatives that do not require data export.

In cases where data export cannot be avoided, I suggest two approaches—tailored around IAB-FDZ data products—that might help reduce statistical disclosure requirements significantly while still enabling meaningful data analysis. In the supplementary material to this article, I provide templates for how such approximations can be implemented (Additional File 3). Using these templates will not automatically lead to a release of aggregated export data by the IAB-FDZ, because the specifics of each dataset still need to be considered, and there might be reasons that require further censoring. Nonetheless, these methods and templates hopefully can serve as a starting point for a discussion about adequate disclosure control in those cases where data export is essential and the standard solutions prove to be tedious, error-prone and unsatisfying.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12651-023-00334-x.

Additional file 1: Online Appendix

Additional file 2: Reproduction package

Additional file 3: Templates

Acknowledgements

I thank Dana Müller, Alexandra Schmucker, Heiko Stüber and Stefanie Wolter for helpful comments.

Author contributions

PB is responsible for the concept, all analyses and writing the manuscript. He is the corresponding author. All authors read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

For my analyses, I use administrative data of the Institute for Employment Research (IAB). The data are social security data with administrative origin which are processed and kept by IAB, Regensburger Str. 104, D-90478 Nürnberg, iab@iab.de, phone: +49 911 1790, according to the German Social Code III. There are certain legal restrictions due to the protection of data privacy. The data contain sensitive information and therefore are subject to the confidentiality regulations of the German Social Code (Book I, Sect. 35, Paragraph 1). The raw data, computer programs, and results have been archived by IAB in accordance with good scientific practice. Computer programs and results can be found in the Reproduction package (Additional file 2). If you wish to access the full data for replication purposes, please contact Philipp vom Berge (philipp.vom-berge@iab.de). Please visit https://www.iab.de/en/daten/replikationen.aspx for further information.

Declarations

Competing interests

Not applicable.

Received: 31 October 2022 Accepted: 16 January 2023
Published online: 13 February 2023

References

Brandt M, Franconi L, Guerke C, Hundepool A, Lucarelli M, Mol J, Ritchie F, Seri G, Welpton R.: Guidelines for the checking of output based on microdata research (2010). https://www.academia.edu/38191675/ESSNet_SDC_-_Guidelines_for_the_checking.pdf

Frodermann C, Schmucker A, Seth S, Vom Berge P.: Sample of Integrated Labour Market Biographies (SIAB) 1975–2019.' FDZ Datenreport 01–2021, Institute for Employment Research (2021). DOI:https://doi.org/10.5164/IAB.FDZD.2101.en.v1.

Ganzer A, Schmidtlein L, Stegmaier J, Wolter S.: Establishment History Panel 1975–2019.' FDZ Datenreport 16–2020 Version 2, Institute for Employment Research (2021). DOI: https://doi.org/10.5164/IAB.FDZD.2016.en.v2.

Publisher's Note