Online Appendix: Population Aggregates from Administrative Data Samples – How good are they? (Philipp vom Berge)

A. The effect of different approximation variants in simple linear regression examples at the district level

In this section I show that all the approximation variants discussed in Section 2 of the main article – when used at the level of districts – perform reasonably well in some simple applications with univariate linear regression. To this end, I use approximated employment per district (in logs) as the right-hand-side variable in regressions with three different outcomes. The results are depicted in Table A.1.

The first outcome I consider is the district-specific unemployment rate, which is derived from official statistics and therefore based on the full population of workers and the unemployed. Consequently, approximation error is limited to the right-hand-side of the regression equation. The first column of Table A.1 shows the benchmark case where employment is derived from the full population of the BHP and thus there is no approximation error. The coefficient of 0.332 implies that districts with a larger labour force have a higher unemployment rate. Although this observation would certainly be only the starting point of an empirical investigation of local labour markets, I will not discuss it further here, but compare it directly with the approximation values. The second, third and fourth column show the results of a similar regression using the BHP-based variant, the IAB Individual File and the SIAB Basis Establishment File, respectively. All variants are very close to the benchmark, with SIAB Basis Establishment File-based variant performing particularly well. T-tests comparing coefficients to the benchmark coefficient in column 1 fail to reject the null in all three cases.

The second outcome is the district-specific average daily log wage, which I derive from the BHP. Here, I assume that the researcher can observe the true outcome, and therefore keep this variable constant through all four cases. This again means that any approximation error is only present on the right-hand side. Once more, the benchmark coefficient of 0.119 is very accurately reflected in all three alternative approximations.

The third outcome is closely related to the second. This time, however, the district-specific average daily log wage is approximated, i.e., I derive it from the same dataset used to calculate the right-hand-side variable and also use the same approximation method. Approximation error can therefore occur

on both sides of the regression equation. Nevertheless, the approximations work very well and no significant bias in the regression coefficients can be observed.[1]

Overall, I conclude that – at least at the high level of aggregation considered here – the choice of variant should not be too important for an approximation.

B. The effect of different approximation variants in simple linear regression examples at the district#3-digit-industry level

In this section, parts of the analysis from Online Appendix A are repeated, but this time at the finer level of district#3-digit-industry cells considered in Section 3 of the main paper. Again, I use approximated employment per cell (in logs) as the right-hand-side variable in the regressions. I omit the unemployment rate as an outcome because there are no official unemployment rates at this level of aggregation.[2] Results are shown in Table B.1.

The first outcome I consider is the district-specific average daily log wage derived from the full population of the BHP. Assuming that the researcher can observe the true outcome, I hold this variable constant in the results shown in panel A. This implies that any approximation error is only present on the right-hand side. I also recognize that the aggregated data now contain a large number of cells that are very sparsely populated. Their small size affects approximation error and could significantly affect the results. In such a case, many researchers resort to weighted regression, so I show both weighted and unweighted results for comparison.

The first column of Table B.1 shows the benchmark case where employment is derived from the BHP population and thus there is no approximation error. The coefficients of 0.074 in the unweighted case and 0.093 in the size-weighted case imply a positive relationship between employment and the wage level. All main three approximation variants – shown in columns 2 to 4 – lead to estimates broadly consistent with the benchmark.[3] The best performing approximation is the one based on the SIAB Basis Establishment File (column 4), which is statistically indistinguishable from the benchmark in the

[1] I omit the regression using the approximation based on the SIAB Individual File in this particular case, because I cannot replicate the wage imputation method used for the BHP in the SIAB exactly. As a result, any differences between the approximated and the benchmark measure would be a combination of approximation error and deviations due to the different imputation method, and I would not be able to disentangle the two.

[2] Also, it is debatable whether measuring an unemployment rate at the industry level makes a lot of sense.

[3] They are now statistically significantly different from the benchmark, however, due to the larger estimation sample and implied higher precision.

case of the weighted regression and very close, although statistically different, in the case of the unweighted regression.

Panel B also approximates the district-specific average daily log wage derived from the same dataset used to calculate the right-hand-side variable. Approximation error can therefore occur on both sides of the regression equation. This time, all approximations produce estimates that are statistically different from the benchmark case.[4] In the unweighted case, the BHP-sample-based variant performs best, with a coefficient of 0.079 compared to the benchmark of 0.074. In the weighted case – which is arguably more relevant – the variant based on the SIAB Basis Establishment File comes closer, with a coefficient of 0.087 compared to the benchmark of 0.093.

Table B.1 also shows the impact of the smoothing procedure discussed in Section 4 of the main paper on the regression results (column 5). Its performance appears comparatively weak for the unweighted regression case, especially in panel B. This is to be expected, as smoothing introduces bias in truly empty cells. However, in the size-weighted regression variants, the smoothing procedure produces results that are relatively close to the (unsmoothed) SIAB Basis Establishment File variant shown in column 4.

Overall, I conclude that the approximations for these finer levels of aggregation discussed in the main paper could still be useful for many applications. The SIAB-based variants still show preferential properties for employment aggregates, at least when observations are size-weighted in a regression. This is true despite the fact that the (unweighted) matrix of cell aggregates is much more sparsely populated in the SIAB than in the BHP. Data users should not discard the SIAB just because it has a lower sampling probability and appears 'smaller'.

C. The effect of censoring in simple linear regression examples at the district#3-digit-industry level

In this section, I show how some of the results presented in Table B.1 change if the underlying datasets require additional censoring after disclosure control, as outlined in Section 5 of the main article. I use average daily log wage as the left-hand side variable and approximated employment per cell (in logs) as the right-hand-side variable in the regressions (this repeats panel A of Table B.1). For ease of comparison, columns 1, 2 and 4 of Table C.1 replicates the results from columns 1, 2 and 3 of Table B.1, based on the uncensored datasets.

[4] I omit the regression for the SIAB Individual File for the same reason given in footnote 5 of the main paper.

Column 3 of Table C.1 shows that the estimated coefficients for the censored data version are significantly biased downward compared to the uncensored version (column 2). For the version based on the SIAB Individual File, the bias is slightly weaker, and upwards (comparison of column 5 to column 4).

Column 6 shows the regression results based on the approximation with random weights as suggested in Section 5 of the main paper. These results come much closer to the version with uncensored fixed weights version in column 4 than the version with censored fixed weights in column 5. Column 7 also reports the results for the smoothed approximation based on the SIAB Basis Establishment File. In summary, both proposed variants show good alignment with the target results, especially when cell size is taken into account. The examples presented here suggest that they may be superior to the results based on "classically" censored aggregated data.[5]

References for Online Appendix

Frodermann, Corinna, Alexandra Schmucker, Stefan Seth and Philipp vom Berge (2021). 'Sample of Integrated Labour Market Biographies (SIAB) 1975 - 2019.' FDZ Datenreport 01–2021, Institute for Employment Research, DOI:10.5164/IAB.FDZD.2101.en.v1.

Ganzer, Andreas, Lisa Schmidtlein, Jens Stegmaier and Stefanie Wolter (2021). 'Establishment History Panel 1975-2019.' FDZ Datenreport 16–2020 Version 2, Institute for Employment Research, DOI: 10.5164/IAB.FDZD.2016.en.v2.

Bundesinstitut für Bau-, Stadt- und Raumforschung – BBSR (2020). 'Indikatoren und Karten zur Raum- und Stadtentwicklung. INKAR. Ausgabe 2020.' © 2020 Bundesamt für Bauwesen und Raumordnung, Bonn.

[5] Like in the main paper, I want to note that I only performed primary cell suppression, skipping the secondary suppression part that would actually be necessary, but tedious. Not skipping secondary suppression might very well lead to even more bias in the results in columns 3 and 5 of Table C.1.

Online Appendix Tables

Table A.1 – Bias in regressions using approximated worker level data - district level

	(i)	(ii)	(iii)	(iv)
	BHP 100%	BHP 50%	SIAB Individual File	SIAB Basis Establishment File
Unemployment rate				
Est.	0.332	0.292	0.344	0.332
S.E.	(0.068)	(0.068)	(0.067)	(0.068)
p_diff	-	0.557	0.854	0.997
Average daily log wage (true)				
Est.	0.119	0.117	0.118	0.119
S.E.	(0.003)	(0.003)	(0.003)	(0.003)
p_diff	-	0.579	0.730	0.917
Average daily log wage (approx.)				
Est.	0.119	0.122	-	0.120
S.E.	(0.003)	(0.003)	-	(0.003)
p_diff	-	0.410	-	0.902
N	4,010	4,010	4,010	4,010

Notes: The table shows coefficients (Est.) and robust standard errors (S.E.) of linear univariate regressions for 3 different outcomes (official unemployment rate, average daily log wage calculated from the full population, average daily log wage approximated from the respective sample) at the district level. The right-hand-side variable is always (approximated) log employment. Calculations are based on the full population and 50 percent sample of the BHP, the SIAB Individual File and the SIAB Basis Establishment File, respectively. Coefficients for the constants are omitted from the table. 'p_diff' shows p-values from a t-test against the full-population coefficient.

Sources: Establishment History Panel (BHP) – Version 7519 v2 (DOI: 10.5164/IAB.BHP7519.de.en.v2); Sample of Integrated Labour Market Biographies (SIAB) – Version 7519 v1 (DOI: 10.5164/IAB.SIAB7519.de.en.v1); BBSR (2020); own calculations.

Table B.1 – Bias in regressions using approximated employment - district#3-digit-industry level

	(i)	(ii)	(iii)	(iv)	(v)
	BHP 100%	BHP 50%	SIAB Individual File	SIAB Basis Establishment File	SIAB Basis Establishment File (smoothed)
	Panel A - Average daily log wage (true)				
unweighted					
Est.	0.074	0.072	0.068	0.076	0.061
S.E.	(0.000)	(0.000)	(0.001)	(0.001)	(0.000)
p_diff	-	0.003	0.000	0.000	0.000
size weighted					
Est.	0.093	0.102	0.097	0.096	0.098
S.E.	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
p_diff	-	0.000	0.062	0.144	0.016
N	743,269	672,667	537,209	537,209	741,005
	Panel B - Average daily log wage (approx.)				
unweighted					
Est.	0.074	0.079	-	0.035	0.003
S.E.	(0.000)	(0.001)	-	(0.001)	(0.000)
p_diff	-	0.000	-	0.000	0.000
size weighted					
Est.	0.093	0.111	-	0.087	0.085
S.E.	(0.002)	(0.002)	-	(0.002)	(0.002)
p_diff	-	0.000	-	0.005	0.000
N	743,269	647,450		517,218	1,038,670

Notes: The table shows coefficients (Est.) and robust standard errors (S.E.) of linear univariate regressions for 2 different outcomes, average daily log wage calculated from the full population (panel A) and average daily log wage approximated from the respective sample (panel B) at the district#3-digit-industry level. The right-hand-side variable is always (approximated) log employment. Regression are done both unweighted and weighted by cell size (analytic weights). Calculations are based on the full population and 50 percent sample of the BHP, the SIAB Individual File, the SIAB Basis Establishment File, and the SIAB Basis Establishment File after smoothing (see section 4), respectively. Coefficients for the constants are omitted from the table. 'p_diff' shows p-values from a t-test against the full-population coefficient.

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

Table C.1 – Bias in regressions using approximated employment- district#3-digit-industry level - with censoring

	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)
	BHP 100%	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)	SIAB Basis Establishment File (smoothed)
	Average daily log wage (true)						
unweighted							
Est.	0.074	0.072	0.045	0.068	0.087	0.066	0.061
S.E.	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)
p_diff	-	0.003	0.000	0.000	0.000	0.000	0.000
size weighted							
Est.	0.093	0.102	0.081	0.097	0.101	0.096	0.098
S.E.	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
p_diff	-	0.000	0.000	0.062	0.000	0.162	0.016
N	743,269	672,667	672,667	537,209	537,209	537,209	741,005

Notes: The table shows coefficients (Est.) and robust standard errors (S.E.) of linear univariate regressions for average daily log wage as an outcome, calculated from the full population at the district#3-digit-industry level. The right-hand-side variable is always (approximated) log employment. Regression are done both unweighted and weighted by cell size (analytic weights). Calculations are based on the full population and 50 percent sample of the BHP (fixed weight: censoring below 20 establishments), the SIAB Individual File (fixed weight: censoring below 20 workers; random weight: censoring below 4 workers) and the SIAB Basis Establishment File after smoothing (see section 4), respectively. Coefficients for the constants are omitted from the table. 'p_diff' shows p-values from a t-test against the full-population coefficient.

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