Schmollers Jahrbuch 134 (2014), 507 – 516
Duncker & Humblot, Berlin

CAMPUS-File AFiD-Panel Industrial Enterprises

By Matthias Klumpe and Michael Rößner

1. Motivation

The Research Data Centres (RDC) of the Statistical Offices of the Länder and the Federal Statistical Office offer different ways of access to a wide range of official german microdata.

On the one hand there is the possibility of analysing the microdata at a Safe Centre or via remote execution (On-site-use). In addition to that, the Off-site-use, with the access forms of Scientific Use Files (SUF) and CAMPUS-Files, allows analysing the microdata outside the safe premises of the statistical offices (Zühlke et al., 2007).

The access forms differ with respect to the level of anonymity and information potential of the used data. The closer the microdata will get to the user the stronger the anonymisation, which was applied during the preparation of the datasets, and the lower the remaining information content of the datasets will be.

As specified by the Federal Statistics Law (BStatG), there are also differences concerning the groups of people that may be given access to the microdata sets. All forms off access, with the exception of the CAMPUS-Files, may only be used by researchers at research institutions with the purpose of independent scientific research (Federal Statistics Law §16,6). Even though CAMPUS-Files are absolutely anonymised datasets, they are a good instrument for students and young professionals to learn to work with official microdata sets. In addition to that CAMPUS-Files are used to promote the different statistics of the RDC and to motivate the research community to analyse the de facto or formal anonymised microdata sets at the Safe Centres or via remote execution[1]. Most of the CAMPUS-Files provided by the RDC are personal or household data. For that reason we developed the CAMPUS-File AFiD[2]-Panel Industrial Enterprises, to extend the offer of enterprise and establishment level CAMPUS-Files in the RDC.[3]

---

1 For an explanation of de facto and formal anonymised microdata: http://www.forschungsdatenzentrum.de/en/data_access.asp.

2 The acronym stands for "*A*mtliche *Fi*rmendaten für *D*eutschland", meaning "Official Firm data for Germany".

The first chapter of this paper gives general information about the CAMPUS-Files provided in the RDC followed by a description of the data basis of the new CAMPUS-File AFiD-Panel Industrial Enterprises. The third chapter will cover the applied anonymisation methods. After that a short first comparative analysis of the CAMPUS-File and the non-anonymised dataset is carried out. This paper will be completed by a future prospect.

2. CAMPUS-Files in the RDC

CAMPUS-Files were developed especially for teaching purposes and contain absolutely anonymised microdata. They offer the possibility for students to acquire methodological knowledge while analyzing social and economic questions. CAMPUS-Files can be downloaded free of charge on the website of the Research Data Centres (www.forschungsdatenzentrum.de/CAMPUS-file.asp). Currently, the following CAMPUS-Files are available:

Table 1

Available CAMPUS-Files of the RDC

social statistics • microcensus • Continuing Vocational Training Survey (CVTS) • student statistics • exam statistics • statistics of public assistance	economic statistics • AFiD-Panel Industrial Enterprises • structure of earnings survey • cost structure survey for small and medium-sized enterprises
finance and tax statistics • wage and income tax statistics	agricultural and environmental statistics • AFiD-Panel Agriculture • census of agriculture

All these CAMPUS-Files are absolutely anonymised microdata sets. This implies that it is not possible to reidentify individuals, establishments or enterprises at any time. Absolute anonymity is the result of a large information reduction achieved by using different anonymisation methods. In consequence of this information reduction, the analytical findings computed with CAMPUS-Files compared to the results computed in the same way with the formal or de facto anonymised microdata sets are biased. Therefore the CAMPUS-Files are unsuitable for bachelor and master theses or dissertations. For further detailed research SuF, Safe Centres or remote execution should be used.

[3] To get an overview of the project "Official Firm Data for Germany" see Malchin/Voshage (2009).

Nevertheless the existing CAMPUS-Files have been and will be used regularly in lectures or in practical courses at various universities to impart methodological knowledge while at least providing a smaller part of real microdata.

3. Database of the CAMPUS-File AFiD-Panel Industrial Enterprises

The CAMPUS-File AFiD-Panel Industrial Enterprises is based on the AFiD-Panel Industrial Enterprises, which is a linked dataset of the annual report for enterprises, the survey of investment and the cost structure survey in the sections of manufacturing, mining and quarrying. All statistics were linked on enterprise level. The panel currently contains the survey years 2001 to 2013 and can be analyzed as cross-sectional and longitudinal data. Prospectively, the panel will be enhanced by current survey years.

The statistical units of the used statistics are the enterprises as well as their establishments. The regarded panel only contains datasets of enterprises. In the statistical sense, an enterprise is the smallest legal independent unit which, according to fiscal and commercial law, keeps accounts and balances. The information of the statistical units always refer to the whole enterprise, including all producing and non-producing establishments. Foreign establishments are not considered.

The group of respondents of the AFiD-Panel Industrial Enterprises is limited to a maximum of 68.000 enterprises, with generally at least 20 or more employees, by law. The enterprises are obliged to provide information for all surveys on an annual basis.

Only enterprises having their main activity (see Federal Statistical Office, 2008, 23 ff.) in the sections of manufacturing, mining and quarrying were recorded. Until the survey year 2008 the sections C and D of the German Classification of Economic Activities, Edition 2003 (corresponds to the economic sectors 10.10 “Mining and agglomeration of hard coal” to 37.20 “Recycling of non-metal waste and scrap”) were covered and from the survey year 2009 the sections B and C of the German Classification of Economic Activities, Edition 2008, (corresponds to the economic sectors 05.10 “Mining of hard coal” to 33.20 “Installation of industrial machinery and equipment”).

For the preparation of the CAMPUS-File the survey years 2003 – 2007 of the AFiD-Panel Industrial Enterprises were used. To add information about the domestic and non-domestic turnover the CAMPUS-File was enhanced by on enterprise-level aggregated information from the establishment surveys of the previously mentioned economic sectors. Therefore the CAMPUS-File contains information from the following statistics:

Table 2

Database of the CAMPUS-File

Data	EVAS[4]	survey years
annual report for enterprises in the sections of manufacturing, mining and quarrying	42221	2003 – 2007
survey of investment in the sections of manufacturing, mining and quarrying	42231	2003 – 2007
cost structure survey in the sections of manufacturing, mining and quarrying	42251	2003 – 2007
summarized annual results of the **monthly report for establishments** in the sections of manufacturing, mining and quarrying (aggregated on enterprise level)	42111	2003 – 2007
annual report for establishments[5] in the sections of manufacturing, mining and quarrying (aggregated on enterprise level)	42271	2007

The linkage of the surveys was done in the cross as well as in the longitudinal section by using the unique enterprise identification numbers.

The information about the economic sector was taken from the annual report for enterprises. In the case of missing values it was taken from the survey of investment or the cost structure survey. In the same way the regional information was treated (see Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder, 2014, 3 f.). You can find a complete list of variables of the CAMPUS-File in the metadata report which is available for download on the website of the RDC mentioned above (see Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder, 2014, 11 ff.).

4. Applied Anonymisation Methods

To prevent the (re-)identification of the reporting units, various anonymisation methods like variable suppressing, recoding, (sub)sampling, stochastic noise or microaggregation were applied during the preparation of the CAMPUS-File. In detail, the following anonymisation methods were applied according in mentioned order.

First of all the regional information for each enterprise was summarized to the regions “West” and “East”. Now the region “West” includes the federal

4 The acronym stands for the german “*E*inheitliches *V*erzeichnis *a*ller *S*tatistiken der Statistischen Ämter des Bundes und der Länder”, meaning “Integrated List of all Statistics Compiled by the Federal Statistical Office and the Statistical Offices of the Länder”.

5 To retain the group of respondents the annual report for establishments 2007 complements the results of the monthly report for establishments 2007.

states Schleswig-Holstein, Hamburg, Lower Saxony, Bremen, North Rhine-Westphalia, Hesse, Rhinland-Palatinate, Baden-Wuerttemberg, Bavaria and Saarland while the region "East" represents the federal states Berlin, Brandenburg, Mecklenburg-Western Pomeriana, Saxony, Saxony-Anhalt and Thuringia.

Information about the economic sector of each enterprise was reduced from the five-digit to the two-digit-level and replaced by random two-digit numbers in a range of 10 to 37. These random numbers are identical for all enterprises and survey years for the same economic sector.

Furthermore the enterprises were divided into employee size classes "more than 1000 employees", "500 up to below 1000 employees" and "less than 500 employees". The size class each enterprise was associated to depends on the highest number of employees in any of the survey years.

After allocating each unit to a size class all enterprises with more than 1000 employees in one survey year were deleted from the dataset. The enterprises of the size class 500 up to below 1000 employees, which were not part of the original panel in every year, were deleted as well. After that a sample of 50% was drawn from the remaining enterprises of this size class. From all enterprises of the size class "less than 500 employees" a sample of 75% was drawn.

All of the remaining enterprises with at least more than 500 employees in one year but always less than 1000 employees were microaggregated by (anonymised) economic sector, region and year (= stratum). The group members remained the same in every year. This means that every value of all metric variables was replaced by the mean of the variable of the respective group. Relevant for the construction of the different aggregation groups was the average annual number of employees over all survey years. To be more precisely, the following steps were conducted:

First of all the average annual number of employees for each enterprise of the mentioned employee size classes over all years was computed. After that the enterprises were ordered by economic sector, region in 2007 and average number of employees. If there were less than three enterprises in one stratum, all enterprises in this stratum were deleted. In a third step the microaggregation was done in groups of three in descending order of the number of employees. If the number of enterprises in a stratum did not correspond to a multiple of three, the group of enterprises with the lowest number of employees was enhanced to four or five units.

Besides this microaggregation every metric variable of the remaining enterprises was multiplied with a random factor (stochastic noise). Each random factor allocated to an enterprise was the same for all metric variables and for all survey years. Half of the enterprises got a random factor between 0.6 and 0.8 and the other half got one between 1.2 und 1.4.

Because of the stochastic noise the value of the employees of an enterprise has changed. Therefore the employee size class was adapted to the new value.

Finally, all enterprise identification numbers were replaced by artificial numbers. This happened by sorting the enterprises by a random number, so that there is no possibility to identify the original number. The artificial number is identical for the same enterprise for all years (see Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder 2014, 5 f.).

5. A Short First Comparative Analysis of the CAMPUS-File and the Original Dataset

To get a first impression about the strength of the adapted anonymisation methods the following chapter compares a few results computed with the CAMPUS-File and the non-anonymised dataset.

The percentage distribution of the enterprises by region and type of enterprise shows very plainly, that there is no major deviation between the CAMPUS-File and the original dataset (cf. Figure 1).

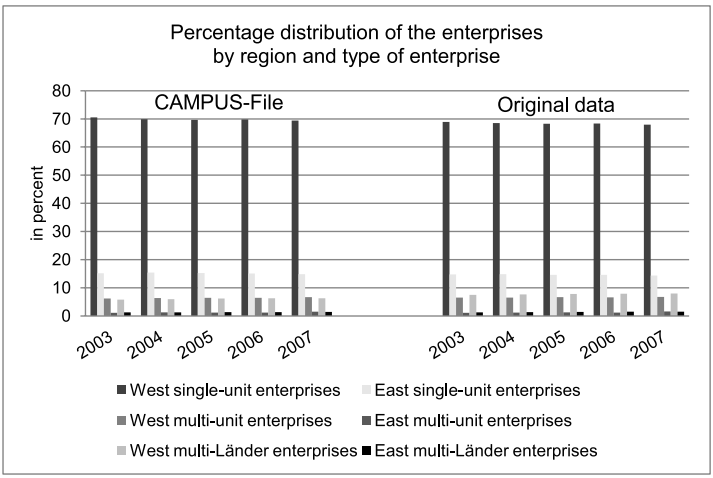

Figure 1: Percentage distribution of the enterprises by region and type of enterprise

By comparing the means of several variables, as expected because of the applied anonymisation methods, clear level differences are recognizable (cf. table 3). For example, there is a mean of about 6.5 million Euro of wages and salaries for West-Germany in 2003 on the basis of the original dataset,

Table 3

Different means (in Euro) by year and region computed with CAMPUS-File and original data

year	region	total wages and salaries		investment in properties with buildings		non-domestic turnover	
		CAMPUS-File	Original data	CAMPUS-File	Original data	CAMPUS-File	Original data
2003	West	3.044.713	6.438.485	57.264	132.021	4.346.297	14.560.651
	East	1.721.751	2.449.630	79.743	88.516	1.782.853	3.848.595
2004	West	3.075.495	6.539.748	58.509	118.959	4.795.515	16.208.872
	East	1.726.020	2.475.358	92.626	181.757	1.957.816	4.195.594
2005	West	3.106.928	6.676.215	62.963	123.039	5.227.375	17.572.994
	East	1.801.596	2.549.347	97.685	122.466	2.384.954	4.729.974
2006	West	3.219.025	6.950.545	70.039	136.109	5.996.730	19.710.132
	East	1.882.219	2.658.614	111.740	139.211	3.020.837	5.746.100
2007	West	3.432.365	7.118.970	93.710	178.970	6.852.070	21.523.784
	East	2.032.409	2.873.971	128.951	163.494	3.520.262	6.648.619

while there is only a mean of about 3 million Euro when taken CAMPUS-File as database. Nevertheless the development of the means between 2003 and 2007 is very similar in both datasets.

Another example showing the differences of both datasets more clearly is the ratio of the means between the enterprises located in East and West Germany by year. The values in table 4 represent the share of the mean of the enterprises located in East Germany in relation to the means of the enterprises located in West Germany. As shown in table 4 the results of the CAMPUS-File overestimate the original values in favor of the East German enterprises in any case. This is a result of deleting the big enterprises, which are mainly located in western Germany.

Table 4

Different share of means by year computed with CAMPUS-File and original data

year	total wages and salaries		investment in properties with buildings		non-domestic turnover	
	CAMPUS-File	Original data	CAMPUS-File	Original data	CAMPUS-File	Original data
2003	0,57	0,38	1,39	0,67	0,41	0,26
2004	0,56	0,38	1,58	1,53	0,41	0,26
2005	0,58	0,38	1,55	1,00	0,46	0,27
2006	0,58	0,38	1,60	1,02	0,50	0,29
2007	0,59	0,40	1,38	0,91	0,51	0,31

The results of a fixed effect regression, with the dependent variable being "total wages and salaries", show differences as well (cf. table 5).

While the coefficient e.g. for the average turnover per person, computed with the original dataset, is about –1.573 the same coefficient computed with the CAMPUS-File is –0.422. These differences are a consequence of the applied anonymisation methods like i.a. the dropping of enterprises with at least more than 1000 employees or the migroaggregation of every value of all metric variables. Furthermore it has to be noted, that the algebraic signs of the coefficients are the same, meaning the effect of the average turnover per person goes in the same negative direction.

The comparative analysis shows, that there are similarities between the results computed with both datasets. However, there are also significant (level) differences. While the CAMPUS-File is certainly suitable for methodological knowledge transfer, the preparation of e.g., a bachelor or master thesis as well as a dissertation should better be done with the original data and not with the absolutely anonymised CAMPUS-File.

Table 5

Fixed effect regression computed with CAMPUS-File and original data

dependent variable total wages and salaries	CAMPUS-File	Original data
independet variables		
average turnover per person	–0.422***	–1.573***
	(0.0274)	(0.140)
non-domestic turnover	0.0309***	0.0606***
	(0.000402)	(0.000385)
return on sales for tangible fixed assets	0.0346***	0.328***
	(0.00695)	(0.0164)
Observations	132,314	182,838
R-squared	0.058	0.171
Number of unr	33,397	46,116
Standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1		

6. Future Prospects

The anonymisation of microdata, especially on enterprise level, is quite difficult. Particularly big enterprises do have a high potential of reidentification. Despite the difficulties the RDC have developed a CAMPUS-File which ensures that there is no possibility to reidentify a single statistical unit and enables students to acquire methodological skills. Although the results of the CAMPUS-File do rarely correspond with the results computed with the original data, the development of further CAMPUS-Files will be an important task for the RDC in the future. The CAMPUS-Files are used to promote the different statistics of the RDC, e.g. in practical courses at german universities, and to motivate the research community to analyze the original data with a lower grade of anonymisation and a higher information content in the Safe Centres or via remote execution.

References

Bundesstatistikgesetz vom 22. Januar 1987 (BGBl. I S. 462, 565), das zuletzt durch Artikel 13 des Gesetzes vom 25. Juli 2013 (BGBl. I S. 2749) geändert worden ist.

Federal Statistical Office (2008): German Classification of Economic Activities 2008 (WZ 2008).

Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder (2014): Metadatenreport CAMPUS-File AFiD-Panel Industrieunternehmen 2003 – 2007 – Version 1.1.

Malchin, A. / *Voshage*, R. (2009): Official Firm Data for Germany, Schmollers Jahrbuch / Journal of Applied Social Science Studies 129, 3, 501 – 513.

Zühlke S. / *Christians*, H. / *Cramer*, K. (2007): Das Forschungsdatenzentrum der Statistischen Landesämter – eine Serviceeinrichtung für die Wissenschaft, AStA Wirtschafts- und Sozialstatistisches Archiv, 3 – 4, 169 – 178.