Case studies B: Illustrative examples of the application of particular techniques

B.1. Illustrative record linkage example

In this section the record linkage algorithm is applied by matching several files of microaggregated micro data (a general description of this method can be found above) on the German structure of costs survey, abbreviated: SCS.

In these examples, the microaggregation method first divides the set of variables into groups. Within a group, the variables are standardised and summed up for each record, such that the records can be sorted by those called Z-scores. Afterwards, for a pregiven number k (in our case $k = 3$), the records with the greatest and smallest z-scores are classified together with their $k-1$ nearest neighbours (with respect to Euclidean distance) and their values are averaged. Hence, in the class of microaggregated data from some confidential original source, the weakest degree of anonymisation is achieved where each variable forms its own group (here, the structure of data is essentially preserved), whereas putting all variables into the same group creates the strongest degree of anonymisation because there are triples of records which agree in all numerical variables and hence can only be separated using the categorical ones. That is, from the class of microaggregated data of an original source file, the data distributing institution may extract the variant with the desired degree of anonymisation. We choose the variants of microaggregation by 1, 8, 11 and 33 groups of variables, which have been extracted by the German national project on anonymisation of business microdata. The weakest of the considered variants of microaggregation, where every numerical variable defines its proper group, is the 33-group variant MA33G. Using this variant the structure of data is widely preserved. The strongest is the multidimensional microaggregation MA1G, where all numerical variables are grouped together. The variant MA8G is obtained by forming eight groups of a size between two elements (smallest group) and twelve elements (largest group), where highly correlated variables are put together. The variant MA11G is obtained by partitioning the set of numerical variables into three-element groups. We also consider the weakest possible form of anonymisation, formal anonymisation, consisting essentially in the deletion of direct identifiers like name, address and so on (FORMAL).

In the first place we give a brief description of the German structure of costs survey. In the second place we treat the realistic scenario, where the data intruder's additional knowledge consists of an external database. For this simulation, we use as external databases both the German turnover tax statistics and the commercially available MARKUS database. In the third place the previously obtained results are contrasted with those obtained by matching records of the original German structure of costs survey with different variants of anonymisation of the survey. This may be regarded as the worst-case scenario, where the data intruder possesses the original data as the best possible external data. However, one should not presume that the data intruder possesses information about all 33 numerical variables of the survey. Realistically, the external database of the data intruder will contain only a few key variables as in subsections below. Regarding examinations as in subsections below, there are in general many more difficulties to be

expected for experiments with data of different sources and fewer key variables, not least because of the fact that the data intruder has – besides the reliable total distance of the assignment – no facility to evaluate his results. It is not least for that reason that the author feels it makes sense – as a concession to the data users – to run experiments also for the worst-case scenario A with variables most likely to be found in commercial enterprise databases in addition to the experiment including all variables.

The target data used

The German structure of costs survey, limited to the manufacturing industry, is a projectable sample and includes a maximum of 18,000 enterprises with 20 or more employees. All enterprises with 500 or more employees or those in economic sectors with a low frequency are included. That is, a potential data intruder has knowledge about the participation of large enterprises in the survey. We consider the survey of the year 1999, covering 33 numerical variables (among which are *Total turnover*, *Research and Development* and the *Number of employees*) and two categorical variables, namely the *Branch of economic activity* (abbreviated: NACE), broken down to the 2-digit level, and the *Type of administrative district* (abbreviated: BBR9), which has 9 values depending on the degree of urbanisaton of the region considered. The complete list of variables available in the German structure of costs survey is given below.

Table 0.1 below contains an excerpt of the German structure of costs survey, classified by the categorical variables mentioned above.

Table 0.1. Partitioning of the German structure of costs survey.

nace2\bbr9	1	2	3	4	5	...	8	9	Sum
10	5	5	2	4	0	...	7	0	39
14	7	19	15	4	2	...	24	8	157
⋮						...			
20	38	54	50	15	8	...	57	42	504
22	356	154	57	23	91	…	54	18	950
24	267	174	82	32	37	...	66	14	901
25	97	187	90	25	16	...	85	41	867
26	116	108	73	49	35	...	100	72	965
27	120	152	44	21	18	...	29	16	593
30	33	28	11	2	12	...	13	0	153
⋮						...			
37	13	15	6	9	9	...	11	2	94
Sum	2, 920	2, 994	1, 379	486	788	...	1, 488	677	16, 918

Totally, there are 26 economic sectors and hence 234 data blocks of a size between 0 and 670 records under consideration.

External versus microaggregated confidential data

In the following we simulate the scenarios B1 and B2 mentioned in section **Error! Reference source not found.** using a sample of around 9,300 records of the German turnover tax statistics (subsection 0) and a sample of around 9,400 records of the commercially available so-called MARKUS database (subsection 0) as the data intruder's additional knowledge.

The German turnover tax statistics

Turnover tax statistics (TTS) are based on an evaluation of monthly and quarterly advance turnover tax returns to be provided by entrepreneurs whose turnover exceeds EUR 16,617 and whose tax amounts to over EUR 511 per annum. Also excluded are enterprises with activities which are generally non-taxable or where no tax burden accrues (e.g. established medical doctors and dentists without laboratory, public authorities, insurance agents, agricultural holdings). The key variables available are:

- Branch of economic activity (NACE2, blocking variable)
- Type of administrative district (BBR9, blocking variable)
- Total turnover (matching variable).

Classifying the number of true matches into intervals relating to the number of employees, we obtain Table 0.2

Table 0.2. Re-identifications (TTS) classified by the number of employees*.

target data	total	1	2	3	4	5	6
MA1G	404 0.0435	103 0.0330	61 0.0259	55 0.0261	64 0.0755	47 0.0916	74 0.2151
MA8G	1, 177 0.1270	366 0.1173	223 0.0949	246 0.1168	137 0.1616	96 0.1871	109 0.3169
MA11G	2, 551 0.2748	824 0.2641	602 0.2561	570 0.2705	238 0.2807	180 0.3509	137 0.3983
MA33G	2, 695 0.2903	894 0.2865	639 0.2718	580 0.2753	246 0.2901	189 0.3684	147 0.4273
FORMAL	2, 677 0.2884	890 0.2853	635 0.2701	574 0.2724	247 0.2913	189 0.3684	142 0.4128

*1=20–49, 2=50–99, 3=100–249, 4=250-499, 5=500-999, 6=1000 and more.

The table contains in each cell the absolute (first row) and relative (second row) frequency of successful attempts. The second row contains the relative frequency of correctly matched pairs concerning the number of enterprises contained in the size classes regarding the external data. It can be observed that the distribution of the shares rapidly approaches the corresponding distribution of scenario B1 (last row in Table 0.2) as the degree of anonymisation goes down. The smallest ratios of correct assignment are obtained for enterprises with 50 to 249 employees. We should like to point out here that

some caution needs to be exercised in interpreting the results as the distribution may change considerably when the employee size classes are formed differently.

Although it is normal that for larger enterprises the microaggregation procedures cause more pronounced changes in the variables, the column on the right of Table 0.2 shows a notably high risk of re-identification for enterprises with at least 1,000 employees. Even in the case of the MA1G variant, about 21 per cent of the large enterprises could be re-identified.

As expected, the number of re-identifications rose considerably as we passed over from variant MA8G to variant MA11G. That is due to the fact that for the MA8G variant the numerical variable *Total turnover* was microaggregated in a group containing 12 elements (including variables 8, 9 and 32, see appendix) and thus modified strongly. In variant MA11G *Total turnover* is found in a group of three elements (together with variables 9 and 15, see appendix), in which every two variables correlate with at least 0.92.

Data incompatibilities are a major reason for incorrect matchings. While only about 1 % of the enterprises have been classified differently with regard to the regional information, nearly 25 % of the enterprises covered by the structure of costs survey have been assigned to another branch of economic activity than their respective records of turnover tax statistics. With regard to the variable *Number of employees* there also are significant differences in both surveys. *Total turnover* figures match relatively well. Only some 18.8 % of the enterprises show deviations of more than 10 % in the surveys.

The MARKUS database

The MARKUS database (in German, Marketinguntersuchungen) covers selected enterprises reported on by ”Creditreform”. It is readily available as a CD-ROM from shops and is published quarterly, with only about 4 % of all enterprises replaced per edition. Generally, the MARKUS database contains enterprises recently examined and not having blocking notes due to insolvency. Therefore, it is not a representative sample of the population. The key variables available are (one additional variable with respect to previous subsection):

- Branch of economic activity (NACE2, blocking variable)
- Type of administrative district (BBR9, blocking variable)
- Total turnover (matching variable)
- Number of employees (matching variable).

For variant MA8G, the two numerical key variables used were mico aggregated in a common group. This means that smaller differences between the variables were lost. For variant MA11G, the variable *Number of employees* was microaggregated in a 3-element group (together with variables 5 and 23) like the variable *Total turnover*, so that the values of these two variables were modified to a lesser degree.

In these experiments, the numerical key variables have been weighted with the same value. Note that a data intruder might prefer the variable *Total turnover* to *Number of*

employees if he had knowledge on the data incompatibilities discussed in the previous subsection.

In line with Table 0.2 we get

Table 0.3. Re-identifications (MARKUS) classified by the number of employees*.

target data	total	1	2	3	4	5	6
MA1G	353 0.0376	59 0.0219	35 0.0150	71 0.0309	60 0.0581	53 0.0897	75 0.1667
MA8G	1, 845 0.1964	343 0.1274	347 0.1490	503 0.2187	279 0.2703	210 0.3553	163 0.3622
MA11G	2, 273 0.2420	419 0.1556	448 0.1924	609 0.2648	355 0.3440	244 0.4129	198 0.4400
MA33G	2, 289 0.2437	420 0.1560	443 0.1902	609 0.2648	370 0.3585	246 0.4162	201 0.4467
FORMAL	2, 294 0.2442	420 0.1560	442 0.1898	610 0.2652	373 0.3614	247 0.4179	202 0.4489

*1=20–49, 2=50–99, 3=100–249, 4=250-499, 5=500-999, 6=1000 and more.

Here the difference between variant MA8G to MA11G is not as pronounced as in the preceding experiment. This also holds for the transition from the enterprise size class of 50 – 999 employees to the class containing enterprises with more than 999 employees. The weaker anonymisation variants MA11G, MA33G and FORMAL produce lower hit rates for smaller and medium-sized enterprises (20 to 249 employees) than in the previous experiment. It is somewhat surprising that the hit rate for variant MA8G increased against the previous experiment as there are more pronounced deviations here in both surveys. While the deviation amounting to about 24 % for all enterprises in the classification of economic activities is in line with the preceding experiment as are the slight deviations in the regional data of less than 2 %, there are much more marked differences regarding *Total turnover*. About 50 % of the enterprises deviate from each other by more than 10 % in the two surveys.

Original versus microaggregated data

In order to get an upper bound for the disclosure risk, the results of the foregoing section are contrasted with the results obtained assuming the worst-case scenario, in which the external database equals the original data without direct identifiers. In the following, we choose several subsets of the numerical variables as matching variables. At first, the whole of 33 numerical variables is used as matching variables (worst-case scenario). We also carry out matching experiments using one matching variable, namely *Total turnover* (in order to contrast the result with the one contained in the realistic scenario of the above subsection on the German turnover tax statistics), two matching variables, namely *Total turnover* and *Number of employees* (in order to contrast the result with the one contained

in the realistic scenario in the above subsection about Markus), and three matching variables, namely *Total turnover, Number of employees* and *Total intramural R&D expenditure*. The latter variable can be in some cases obtained e.g. via internet searches. As in the previous section, in all experiments the categorical variables BBR9 and NACE2 were used for blocking the data.

Table 0.4 shows the relative frequency of true matches. The first and second entries in each cell refer to the relative and the absolute frequency of true matches obtained using 1, 2, 3 and 33 matching variables.

Table 0.4. Re-identifications classified by the number of matching variables.

microaggreg. data	33 variables	3 variables	2 variables	1 variable
MA1G	8, 941 0.5285	2, 156 0.1274	2, 076 0.1227	1, 096 0.0648
MA8G	16, 792 0.9926	12, 820 0.7578	11, 127 0.6577	3, 621 0.2140
MA11G	16, 853 0.9962	16, 732 0.9890	16, 765 0.9910	12, 066 0.7132
MA33G	16, 918 1.0000	16, 918 1.0000	16, 912 0.9996	16, 757 0.9905
FORMAL	16, 918 1.0000	16, 918 1.0000	16, 918 1.0000	16, 918 1.0000

The protection increases notably if the data intruder has only one matching variable available (*Total turnover*) instead of two matching variables (*Total turnover* and *Number of employees*). For the transition from two matching variables to three matching variables only slight differences are observed, in the case of MA11G the hit rate actually decreases. Anyway, the weak protection effect of MA11G, as already observed in the previous section, is confirmed.

As in Table 0.2 and Table 0.3, we consider the distribution of the frequency of re-identifications among the employee size classes, starting with the one matching variable experiment:

Table 0.5. Re-identifications using one matching variable classified by the number of employees*.

target data	total	1	2	3	4	5	6
MA1G	1, 096 0.0648	243 0.0459	161 0.0391	164 0.0420	151 0.0859	145 0.1336	232 0.3069
MA8G	3, 621 0.2140	1, 043 0.1970	681 0.1653	765 0.1959	417 0.2372	354 0.3263	361 0.4775

MA11G	12, 066 0.7132	3, 841 0.7255	2, 852 0.6924	2, 706 0.6928	1, 252 0.7122	800 0.7373	615 0.8135
MA33G	16, 757 0.9905	5, 236 0.9890	4, 084 0.9915	3, 873 0.9916	1, 741 0.9903	1, 078 0.9935	745 0.9854
FORMAL	16, 918 1.0000	5, 294 1.0000	4, 119 1.0000	3, 906 1.0000	1, 758 1.0000	1, 085 1.0000	756 1.0000

*1=20–49, 2=50–99, 3=100–249, 4=250-499, 5=500-999, 6=1000 and more.

The increase in the hit rate is quite marked for the transition from MA8G to MA11G. The results of Table 0.5 may be related to the results of the realistic scenario in Table 0.2) as the same common variables were available in the additional knowledge (Turnover tax statistics) used there.

Dually to Table 0.5 we obtain in Table 0.6 the distribution of re-identifications among the employee size classes concerning the experiment with two matching variables.

Table 0.6. Re-identifications using two matching variables classified by the number of employees*.

target data	total	1	2	3	4	5	6
MA1G	2, 076 0.1227	394 0.0744	344 0.0835	420 0.1020	311 0.0796	275 0.1564	332 0.3060
MA8G	11, 127 0.6577	3, 344 0.6317	2, 610 0.6336	2, 578 0.6600	1, 206 0.6860	769 0.7088	620 0.8201
MA11G	16, 765 0.9910	5, 237 0.9892	4, 076 0.9896	3, 879 0.9931	1, 746 0.9932	1, 079 0.9945	748 0.9894
MA33G	16, 912 0.9996	5, 294 1.0000	4, 117 0.9995	3, 906 1.0000	1, 756 0.9989	1, 085 1.0000	754 0.9974
FORMAL	16, 918 1.0000	5, 294 1.0000	4, 119 1.0000	3, 906 1.0000	1, 758 1.0000	1, 085 1.0000	756 1.0000

*1=20–49, 2=50–99, 3=100–249, 4=250-499, 5=500-999, 6=1000 and more.

Regarding MA11G and MA33G, it is observed that the percentage of true matches in the class of 1,000 or more employees is actually recessive. As in the experiment with the MARKUS database (see Table 0.3), there is a pronounced increase in the hit rate between MA1G and MA8G, while the difference between MA11G and MA33G almost seems to be negligible.

Summary

On the whole, the general conjecture is confirmed that larger enterprises are easier to re-identify than smaller ones. In this context, it is fortunately observed that the method of microaggregation is more effective for very large enterprises (above a certain total number of employees). In our application, this inflecting point is lowest for the method of microaggregation by 11 groups.

From the theoretical viewpoint, experiments drawing upon additional knowledge taken from reality always have to be regarded as exemplary. The data distributing institution can never be 100 per cent sure that a potential data intruder does not have better additional knowledge at his disposal than the one used for simulation. In order to make concessions to the data users the present paper proposes an approach accounting for both, scenarios using available databases for potential data intruders and the worst-case scenario matching the original data against the anonymised data in order to determine an upper bound for the disclosure risk associated with the anonymised data.

To handle the record linkage algorithm, first of all the areas at risk have to be identified within the data material. In the present paper, size classes of employees have been examined for that purpose. Furthermore, the analyses have shown that some economic sectors (rows in Table 0.1) are more insecure than other sectors and require particularly confidential treatment. Here it seems necessary that branches of economic activity containing only a small number of values are excluded or aggregated further. In general, the following holds: The coarsening or exclusion of categorical variables contributes considerably to anonymising and, provided that the scientist can do without the information thus lost, makes it possible to modify the numerical variables to a smaller extent. It has turned out, for example, that coarsening the BBR9 code leads to a marked reduction of the disclosure risk calculated in section 8. Here it appears a trade-off, namely that on the one hand the number of mismatches within the blocks is reduced through coarsening, while on the other very large blocks are created making it much more difficult to find true matches.

Appendix: The German Structure of Costs Survey

The following variables are available in the German structure of costs survey.

1. Branch of economic activity (NACE - Classification of Economic Activities)
2. Type of administrative district (BBR[1] 9 code – so-called category 9)
3. Size class of employees
4. Working proprietors
5. Employees (salary and wage earners)
6. Part-time employees
7. Part-time employees in full-time equivalent units
8. Total of active persons
9. Turnover of the unit's own products
10. Turnover of goods for resale
11. Total turnover (does not correspond to the sum of items 9 and 10)
12. Initial stocks of work in progress and finished products manufactured by the unit measured against turnover of the unit's own products

[1]Federal Agency for Construction and Regional Planning

13. Final stocks of work in progress and finished products manufactured by the unit measured against turnover of the unit's own products
14. Change in stocks of work in progress and finished products
15. Gross output/production value
16. Initial stocks of raw materials and other intermediary products purchased and consumables, measured against turnover of the unit's own products
17. Final stocks of raw materials and other intermediary products purchased and consumables, measured against turnover of the unit's own products
18. Consumption of raw materials
19. Energy consumption
20. Initial stocks of goods for resale measured against turnover of goods for resale
21. Final stocks of goods for resale measured against turnover of goods for resale
22. Input of goods for resale
23. Wages and salaries
24. Statutory social security costs
25. Other social security costs
26. Payments for agency workers
27. Costs of contract processing
28. Repair costs
29. Renting and leasing
30. Other costs
31. Interest on borrowed capital
32. Total costs
33. Value-added at factor cost
34. Net value-added at factor cost
35. Total intramural R&D expenditure
36. Total number of R&D personnel