Research Data Centre (FDZ)
of the German Federal
Employment Agency (BA)
at the Institute for
Employment Research (IAB)

FDZ·Methodenreport

Methodological aspects of labour market data

05/2011
EN

Generation of time-consistent industry codes in the face of classification changes

Simple heuristic based on the Establishment History Panel (BHP)

This version: September 2014

Johanna Eberle
Peter Jacobebbinghaus
Johannes Ludsteck
Julia Witter

Bundesagentur für Arbeit

Contents

Abstract 3

Zusammenfassung 3

1 Introduction 4

2 Correspondence Tables 5

2.1 Computation of Correspondence Tables Based on the Mode 6

2.2 Quality Check of Correspondence Tables 7

3 Generation of Completed Industry Variables 8

3.1 Completing the Classification of Economic Activities 73 9

3.2 Completing the Classification of Economic Activities 93 10

4 Summary 12

Literature 14

Appendix 15

A1 Industry Codes Used at the FDZ BA 15

A2 Links to Correspondences to International Industry Classifications 17

A3 Matching Quality Outside the Period of Overlap 18

Abstract

The analysis of FDZ BA data is hampered by changing industry classifications. This report describes two methods to deal with this and provides ready-to-use working tools. The first method is the usage of directional correspondence tables. We create several correspondence tables for the classifications used in our data. The mapping is based on the mode. The quality for each mapping is indicated by the share of correct matches for each value of the base classification. The correspondence tables are supplements to this report. Due to the limitations of correspondence tables we also generate two completed variables for each establishment in our data: w73 (3-digit) and w93 (3-digit). We first extrapolate valid values whenever possible. Then we use correspondence tables to replace further missing values based on available classifications. A large share of the missing values can be replaced. The generated variables are available for our users upon request.

Zusammenfassung

Die Analyse der FDZ BA Daten wird durch wechselnde Klassifikationen der Wirtschaftszweige erschwert. Dieser Methodenreport beschreibt zwei Wege mit den wechselnden Klassifikationen umzugehen und stellt gebrauchsfertige Arbeitshilfen bereit. Wir erstellen eine Reihe von Umschlüsselungstabellen für die Klassifikationen, die in unseren Daten verwendet werden. Die Zuordnung der Werte von einer Basisklassifikation zu einer Zielklassifikation basiert auf Modalwerten. Die Güte der Zuordnung wird für jeden Wert der Basisklassifikation anhand des Anteils richtiger Zuordnungen angezeigt. Die Umschlüsselungstabellen stehen als Beilage zu diesem FDZ-Methodenreport zum Download bereit. Aufgrund ihrer begrenzten Anwendbarkeit generieren wir zusätzlich zwei vervollständigte Variablen für alle Betriebe in unseren Daten: w73 (3-Steller) und w93 (3-Steller). Dazu schreiben wir zunächst gültige Werte in Zukunft und Vergangenheit fort. Anschließen ersetzen wir weitere fehlende Werte anhand der Umschlüsselungstabellen. Dadurch kann ein Großteil der fehlenden Werte ersetzt werden. Die generierten Variablen werden unseren Nutzern auf Anfrage zur Verfügung gestellt.

Keywords: Classification of economic activities, correspondence tables, concordances, conversion of industry codes.

1 Introduction

The Research Data Centre of the Federal Employment Agency (FDZ BA) provides different datasets with information on employees that stem from the notifications of employers to the German social security system and range backwards until 1975.[1] These data include information on the branch of economic activity. Unfortunately, the coding scheme has been modified several times, in a way that there is no classification available which is consistent over time. Since it is not possible to simply map one classification onto the other, longitudinal analyses are hampered by the fact that industry classifications are not comparable over the whole period covered by the data.

Four different classifications have been used since 1975: the Classification of Economic Activities 73, 93, 03 and 08, which we will call w73, w93, w03 and w08 in the following.[2] Table 1 shows that they cover different periods which overlap in some years only. There are several means to fill in missing values in the years not covered by a certain classification scheme. One is to assume that the economic activities of the firms do not change over time and to extrapolate observed values to the years with missing information. Another method is to use correspondence tables to map values of observed classifications onto unobserved ones. A third method is to replace missing values by regression-based multiple imputation. This method is very costly and not very promising if, as in our case, the industry codes are to be imputed at a very disaggregate level and economic activity cannot be explained well by the variables at hand.

The first aim of this work is to provide new correspondence tables between the different classifications used in our data which are based on a simple and transparent mode heuristic. The second aim is to compile a dataset with completed industry variables w73 and w93 for all firms and all years in our data. Missing values are replaced by means of extrapolation and correspondence tables and the resulting industry variables can be merged with the datasets provided by the FDZ BA.

[1] Examples are the Establishment History Panel (BHP), the Sample of Integrated Labour Market Biographies (SIAB), and the Linked Employer-Employee Data from the IAB (LIAB).

[2] Comparability to international standard classifications is given for w93, w03 and w08. For these three schemes, the first 2 digits are the same as ISIC, the first 4 digits are the same as NACE (in their respective versions). To our knowledge, w73 does not correspond to an international classification. Appendix A1 gives basic information on the industry classifications.

Table 1: Availability of Industry Codes

Year	w73	w93	w03	w08
1975	XXX			
…	XXX			
1998	XXX			
1999	XXX	XXXXX		
2000	XXX	XXXXX		
2001	XXX	XXXXX		
2002	XXX	XXXXX		
2003		XXXXX	XXXXX	
2004			XXXXX	
2005			XXXXX	
2006			XXXXX	
2007			XXXXX	
2008			XXXXX	XXXXX

Notes: The Xs reflect the number of digits: w73 has 3 digits, whereas w93, w03, and w08 come with 5-digit accuracy. The classifications w93, w03 and w08 are very similar, as only minor modifications were made. Grey-shaded areas display the periods of overlap.

2 Correspondence Tables

A correspondence table “Is a tool for the linking of classifications. A correspondence table systematically explains where, and to what extent, the categories in one classification may be found in other classifications, or in earlier versions of the same classification. Methodologically, correspondence tables (…) describe the way in which the value sets of classifications are related, by describing how the units classified to the groups defined for a classification would be classified in other classifications.”[3]

The German Federal Statistical Office provides correspondence tables between w93 and w03 as well as between w03 and w08.[4] These tables list both classifications and mark which classes can be transferred directly and which cannot. But for those values that cannot be uniquely allocated, there is no guidance as to which value to take. Frank and Grimm (2010) provide a transition matrix of all establishments between w03 and w08 on the level of divisions (2 digits). This transition matrix reveals not only which categories of the old classification relate to which categories of the new classification, it also shows the relative importance

[3] See United nations Statistics Division, UN Glossary of Classification Terms; http://unstats.un.org/unsd/class/family/glossary_short.asp#C

[4] See Appendix and http://www.destatis.de/jetspeed/portal/cms/Sites/destatis/Internet/DE/Navigation/Klassifikationen/Klassifikationen.psml. These tables can be used to get consistent information since 1999.

of each link. The elements of the transition matrix can be used as weights to compute consistent time series of aggregates. If a one-to-one mapping from w03 to w08 or the other way round is needed, it helps to find the most frequent link among ambiguous choices.

We do not provide transition matrices but several directional correspondence tables that provide one-to-one mappings from base classifications to target classifications. The mapping is based on the most frequently observed link. Our correspondence tables cannot be used to compute consistent time series of aggregates. But they might be a sufficient approximation of the true economic activity in some cases.[5]

2.1 Computation of Correspondence Tables Based on the Mode

For our computations, we employ an establishment-level dataset of the FDZ BA, the Establishment History Panel (BHP)[6]. In order to compute a correspondence table between two classifications we select all observations that contain information on both of these classifications (the grey-shaded areas in Table 1). Based on these data points, we calculate which value of the target classification is the most frequent value (the mode) for each value of the base classification. Establishments with multiple observations are counted only once. If two codes equal in relative frequencies of establishments, we choose the value referring to the larger total number of employees, assuming that notifications of larger firms are more reliable. If there is still a tie, we take the value connected to the higher median wage. The latter choice is arbitrary to some extent but it makes sure that there is a distinct allocation.

The correspondence tables we provide as supplements to this report are listed in Table 2. Each of the correspondence tables contains the values of the base classification and the respective modes of the target classification. By merging the table based on the base classification, you can add the corresponding values of the target classification to your data easily. The files further include the number of firms per value of the base coding scheme and the share of firms that map onto the modal value. These variables are helpful to assess the quality of the mode computations.

In general, the finer the base classification and the broader the target classification, the better is the mapping. We found that it does not make much difference whether to compute the correspondence between base code and the coarsened target code (e.g. 3-digit) directly or to compute it for the detailed target code (e.g. 5-digit) first and then coarsen the variable (e.g. by cutting off some digits at the back). This is why we chose to provide correspondences to the detailed target codes.

[5] The correspondence tables can be downloaded from the FDZ website:
http://doku.iab.de/fdz/reporte/2011/MR_05-11_corrtab.zip

[6] The Establishment History Panel (BHP) covers all establishments with at least one employee liable to social security (between 1.3 and 2.5 million per year). It is based on the notifications submitted by employers to the social security agencies for employees covered by social security (and since 1999 for employees in marginal part-time employment). These data on individuals are aggregated to establishment level (Hethey-Maier and Seth, 2010).

Table 2: List of Correspondence Tables Provided by FDZ BA

Base variable	Target variable	Years of overlap	Observations	File name
w73 (3)	w93 (3)	1999-2002	10,044,467	corrtab_w93_3_w73_3.dta
w93 (5)	w73 (3)	1999-2002	10,044,467	corrtab_w73_3_w93_5.dta
w93 (5)	w03 (5)	2003	2,504,872	corrtab_w03_5_w93_5.dta
w03 (5)	w93 (5)	2003	2,504,872	corrtab_w93_5_w03_5.dta
w03 (5)	w08 (5)	2008	2,770,250	corrtab_w08_5_w03_5.dta
w08 (5)	w03 (5)	2008	2,770,250	corrtab_w03_5_w08_5.dta

Notes: The number of digits is given in brackets. The file corrtab_w73_3_w93_5 maps the modes of w73 (3-digit) to all values of the base variable w93 (5-digit). These modes of w73 are stored in the variable w73_3_w93_5_mod. The modes can be merged to other datasets based on the variable w93_5. The datasets include two further variables: w73_3_w93_5_firms gives the number of firms the mapping is based on and w73_3_w93_5_rel gives the share of firms for which the mode equals the value of w73. As indicated by the file extension dta the files are in Stata format.

2.2 Quality Check of Correspondence Tables

A measure of fit for the mapping of modal values is the share of observations for which the mapped mode equals the true value in the period of overlap. Table 3 shows that the mapping quality varies substantially among the correspondence tables. As expected the mapping between w93 and w03 and between w03 and w08 is good, i.e. even mappings from 5 digits to 5 digits render correct matches in about 90 per cent of all cases. The correspondence between w73 and w93 is considerably weaker. Only in 74.1 (w73 (3) to w93 (3)) and 87.7 (w93 (5) to w73 (3)) per cent of observations the mode is equal to the original value.

Table 3: Quality of Mode Computations

Base variable	Target variable	Original value = mode	Original value ≠ mode	Original value missing	Base value missing	Omitted base values	Omitted target values
w73 (3)	w93 (3)	74.13	25.63	0.21	0.03	2/303	64/222
w93 (5)	w73 (3)	87.73	12.04	0.08	0.16	1/1062	45/303
w93 (5)	w03 (5)	94.67	1.52	0.17	3.64	0/1062	31/1041
w03 (5)	w93 (5)	93.45	2.74	3.80	0.01	0/1041	50/1062
w03 (5)	w08 (5)	94.06	5.50	0.16	0.28	0/1041	99/835
w08 (5)	w03 (5)	86.36	13.20	0.44	0.00	1/835	297/1041

Notes: The number of digits is given in brackets. Columns 3 and 4 display the share of matches between actual values and the computed modes during the periods of overlap. Columns 5 and 6 give information on the amount of missing values of the base classification and the cases where no mode exists for a given non-missing value of the base classification. Column 7 and 8 contain the number of values of the base and the target classification that are not included in the correspondence table.

One disadvantage of directed correspondence tables based on the mode is that for a given value of the base classification, only one mapping is selected and the less frequent values that a value of the base classification maps onto are discarded. As a consequence, values of the target classification that do not represent the mode to any value of the base classification will not be assigned. Column 8 of Table 3 shows how many target values are not assigned. The mapping of w93 onto w03 and vice versa produces few omitted target values relative to the total number of categories (the share is below 5%). For the correspondences between w03 and w08 and between w73 and w93, the share of values not included in the correspondence tables is more significant, amounting to about 15-30%.

The indicators in Table 3 are averages for all observations in the overlapping period. Since the mapping quality varies strongly among the values of the base classification, we added the share of firms that map onto the mode for each base category to the correspondence tables. This allows the user of the correspondence table to exclude certain categories that are based on small shares of correct mappings.

The quality of correspondence outside the period of overlap, which is of primary interest, cannot be directly observed, since we do not observe values on both classification schemes. One way to assess the accuracy of fit is to compare the modes with values that were extrapolated. The results are displayed in Table 6 in Appendix A3. The results are quite similar to the ones obtained for the period of overlap.

3 Generation of Completed Industry Variables

Due to the inaccuracy of correspondence tables, especially if the classifications are quite different, we suggest to replace missing values by extrapolation first. By extrapolation we mean the replacement of missing values by observed values of the target classification in earlier and later years. Unfortunately, this is not possible for some of the FDZ BA datasets.

The SIAB, for example, is a 2 per cent sample of individuals. Employee A of establishment B may leave B before w93 is introduced. If no other employees of establishment B are in the SIAB sample, no value of w93 can be extrapolated to the observations of A in B. This is why we generate a dataset that contains completed industry variables for all of the BHP establishments. This dataset can be merged to all FDZ BA datasets by establishment number and year.[7]

The completion of an industry variable is done in two steps. First, missing values are replaced by extrapolation of valid values to all past and future observations of the establishment. We have to assume that the economic activity of an establishment does not change during the unobserved years. If extrapolation is not possible (e.g. if the firm closed before the classification to be completed was introduced), we use the correspondence tables described in the previous section and replace still missing values based on the available industry code.

We choose to complete the w73 since it is available for a long time period already and values after 2002 can be inferred from the detailed 5-digit classifications if extrapolation is not possible. The completed w73 should be used for analyses covering all years from 1975 to 2008. We also completed the w93 since it better distinguishes service sector activities and it relates to international classifications (see Appendix A1). The completed w93 should be used with caution for the years before 1999. The earlier the more values are not extrapolated but based on the correspondence table w73 to w93 which is of limited accuracy.

Besides the completed w73 and w93 variable the data file contains additional variables that allow for the distinction of original values, extrapolated values and values based on correspondence tables. The next sections describe the details of the completion for both variables.

3.1 Completing the Classification of Economic Activities 73

The completed w73 (w73_3_gen) is generated in 3 steps:

1. We replace as much missing values as possible by extrapolation back and forth, not only for w73, but also for w03 (5-digits) and w93 (5-digits).

2. We use the correspondence table corrtab_w93_5_w03_5.dta to replace missing w93 values based on the (extrapolated) w03. This affects mainly 2004 to 2008 but also earlier years.

3. We use the correspondence table corrtab_w73_3_w93_5.dta to replace missing w73 values based on the (in steps 1 and 2 completed) w93. This affects mainly the years 2003 to 2008 but also earlier years.

Table 4 shows the frequencies of original values of w73, extrapolated values, values based on correspondence tables, and missing values that could not be replaced. For the years

[7] We copy the file to our users' project folders upon request. The file cannot be merged with scientific use files.

2003-2008, a large but decreasing number of missings on w73 can be filled via extrapolation. A smaller but increasing number of missing values is replaced based on correspondence tables. Of about 2.6 million missing values on w73 per year during 2003-2008, only about 1,000 to 10,000 per year remain. Table 4 also shows that some missing values of w73 can also be replaced for the period 1975-2002.

Table 4: Composition of the Completed Variable w73_3_gen

Year	Original	Extrapollat-ed	Based on correspond-ences	Remaining missing	Number of firms
1975	1,237,596	11,320	1,930	40,486	1,291,332
1980	1,349,059	12,927	2,319	41,625	1,405,930
1985	1,394,949	15,997	2,667	37,964	1,451,577
1990	1,507,880	17,452	2,042	21,584	1,548,958
1995	1,959,655	25,175	1,757	11,772	1,998,359
1996	1,975,270	26,304	1,468	8,912	2,011,954
1997	1,974,796	27,341	1,348	6,741	2,010,226
1998	2,005,810	29,130	988	4,431	2,040,359
1999	2,489,099	9	68	104	2,489,280
2000	2,532,861	10	323	300	2,533,494
2001	2,527,719	4	1,009	829	2,529,561
2002	2,486,866	2	4,304	960	2,492,132
2003	0	2,239,910	264,080	882	2,504,872
2004	0	2,108,569	526,597	1,406	2,636,572
2005	0	1,974,183	704,556	1,609	2,680,348
2006	0	1,867,489	864,577	2,046	2,734,112
2007	0	1,770,129	997,561	2,758	2,770,448
2008	0	1,669,911	1,091,020	9,319	2,770,250
Total	47,289,735	12,061,957	4,504,857	718,561	64,575,110

Notes: Computations are based on the BHP 1975-2008. The classification scheme w73 was used until 2002. Grey-shaded rows denote years without original observations on w73.

3.2 Completing the Classification of Economic Activities 93

For the better comparability to international industry classifications, researchers may prefer to have w93 completed for the period 1975 to 2008. Therefore, we also provide a completed w93 variable with 3-digit accuracy.[8] The ISIC-equivalent code can be obtained by cutting off the last digit of this variable.[9]

The completed w93 (w93_3_gen) is generated in 3 steps:

1. We replace as much missing values as possible by extrapolation back and forth, not only for w93 (3-digits), but also for w73 (3-digits) and w03 (5-digits).
2. We use the correspondence table corrtab_w93_3_w03_5.dta to replace missing w93 values based on the (extrapolated) w03. This affects mainly 2004 to 2008 but also earlier years.
3. We use the correspondence table corrtab_w93_3_w73_3.dta to replace still missing w93 values based on the (extrapolated) w73. This affects mainly the years 1975 to 1998 but also later years.

Table 5 shows the frequencies of original values of w93, extrapolated values, values based on correspondence tables, and missing values that could not be replaced. The amount of extrapolated values decreases with growing temporal distance to the observed period 1999-2003 since establishments enter and leave the panel. The further away from the observation period, the larger is the share of values based on correspondence tables. The latter have to be regarded as less reliable, especially those based on w73.

[8] It is not reasonable to complete w93 based on w73 with 5-digit accuracy, since w73 is only available as 3-digit code and the quality of the match would thus be very low.

[9] Our tests show that an indirect computation of w93_2 (computing w93 first and then cutting off the last digit) yields similar matching quality as a direct computation. Therefore we do not provide the completed w93 as 2-digit variable.

Table 5: Composition of Completed Variable w93_3_gen

Year	Original	Extrapolated	Based on correspond-ences	Remaining missing	Number of firms
1975	0	439,596	811,245	40,491	1,291,332
1980	0	575,434	788,867	41,629	1,405,930
1985	0	722,994	690,616	37,967	1,451,577
1990	0	945,527	581,847	21,584	1,548,958
1995	0	1,510,141	476,446	11,772	1,998,359
1996	0	1,713,666	289,819	6,741	2,010,226
1997	0	1,713,666	289,819	6,741	2,010,226
1998	0	1,857,679	178,249	4,431	2,040,359
1999	2,485,419	455	3,302	104	2,489,280
2000	2,528,899	537	3,758	300	2,533,494
2001	2,524,244	0	4,488	829	2,529,561
2002	2,484,965	0	6,207	960	2,492,132
2003	2,409,675	0	94,316	881	2,504,872
2004	0	2,229,736	405,431	1,405	2,636,572
2005	0	2,072,848	605,892	1,608	2,680,348
2006	0	1,952,157	779,910	2,045	2,734,112
2007	0	1,844,291	923,400	2,757	2,770,448
2008	0	1,735,652	1,025,280	9,318	2,770,250
Total	12,433,202	32,187,534	19,235,749	718,625	64,575,110

Notes: Computations are based on the BHP 1975-2008. The classification scheme w93 was used 1999-2003. Grey-shaded rows denote years without original observations on w93.

4 Summary

The analysis of FDZ BA data is hampered by changing industry classifications. This report describes two methods to deal with this and provides ready-to-use working tools. The first method is the usage of directional correspondence tables. We create several correspondence tables for the classifications used in our data. The mapping is based on the mode. The quality for each mapping is indicated by the share of correct matches for each value of the base classification. The correspondence tables are supplements to this report.

Due to the limitations of correspondence tables we also generate two completed variables for each establishment of our data: w73 (3-digit) and w93 (3-digit). We first extrapolate valid values whenever possible. Then we use correspondence tables to replace further missing values based on available classifications. A large share of the missing values can be replaced. The generated variables are available for our users upon request.

The methods presented have some important limitations: underlying changes in economic structure are not fully reproduced in the completed variables since observed values may be extrapolated for many years. Correspondence tables based on the mode may lead to biased results since rare values of the target classification are assigned too seldom or even not at all. We provide quality indicators as well for the correspondence tables as for the completed variables that allow the researcher to decide if these working tools are adequate for his purposes.

Literature

Hethey-Maier, Tanja; Seth, Stefan (2010): Das Betriebs-Historik-Panel (BHP) 1975-2008 * Handbuch Version 1.0.2. (FDZ Datenreport, 04/2010 (de)), Nuremberg (in German).

Thomas Frank und Christopher Grimm (2010): Beschäftigungsstatistik: Umstellung der Klassifikation der Wirtschaftszweige von WZ 2003 auf WZ 2008, Methodenbericht der Statistik der BA, Nuremberg (in German).

Appendix

A1 Industry Codes Used at the FDZ BA

Classification of Economic Activities 73 (WS 73)

German title: Klassifikation der Wirtschaftszweige 1973

Classification level	Short name	Name			WZ 73	Coding
1	1-digit code (generated)	Divisions	-	-	10	0 – 9
2	2-digit code	Groups	-	-	96	00 – 99
3	3-digit code	Classes	-	-	303	000 – 998

Classification of Economic Activities 93 (WZ 93)

German title: Klassifikation der Wirtschaftszweige 1993

Classification level	Short name	Name	ISIC Rev.3	NACE Rev.1	WZ 93	Coding
1	Letter	Sections	17	17	17	A – Q
	2 Letters	Sub-sections		31	31	AA – QA
2	2-digit code	Divisions	60	60	60	01 – 99
3	3-digit code	Groups	159	222	222	01.1 – 99.0
4	4-digit code	Classes	292	503	503	01.11 – 99.00
5	5-digit code	Sub-classes			1062	01.11.1 – 99.00.3

Classification of Economic Activities 03 (WZ 2003)

German title: Klassifikation der Wirtschaftszweige 2003

Classification level	Short name	Name	ISIC Rev. 3.1	NACE Rev. 1.1	WZ 03	Coding
1	Letter	Sections	17	17	17	A – Q
	2 Letters	Sub-sections		31	31	AA – QA
2	2-digit code	Divisions	62	62	60	01 – 99
3	3-digit code	Groups	161	224	222	01.1 – 99.0
4	4-digit code	Classes	298	515	513	01.11 – 99.00
5	5-digit code	Sub-classes			1041	01.11.1 – 99.00.3

Classification of Economic Activities 08 (WZ 2008)

German title: Klassifikation der Wirtschaftszweige 2008

Classification level	Short name	Name	ISIC Rev. 4	NACE Rev. 2	WZ 08	Coding
1	Letter	Sections	21	21	21	A – U
2	2-digit code	Divisions	88	88	88	01 – 99
3	3-digit code	Groups	238	272	272	01.1 – 99.0
4	4-digit code	Classes	419	615	615	01.11 – 99.00
5	5-digit code	Sub-classes			839	01.11.0 – 99.00.0

A2 Links to Correspondences to International Industry Classifications

Eurostat

RAMON is Eurostat's Metadata Server and provides

- Classifications ISIC, NACE, U.S. NAICS, JSIC
- Correspondence tables
    - between different versions of ISIC
    - between different versions of NACE
    - ISIC and NACE
    - ISIC and US SIC
    - ISIC and NAICS
    - NACE and NAICS
    - NACE and US SIC

at
http://ec.europa.eu/eurostat/ramon/relations/index.cfm?TargetUrl=LST_REL&StrLanguageCode=EN&IntCurrentPage=1

Statistics Canada

- Classifications NAICS, SIC, ISIC, NAICS US, SIAC Mexico, NACE
- Correspondence tables
    - NAICS to ISIC
    - NAICS to NACE

at:
http://www.statcan.gc.ca/concepts/industry-industrie-eng.htm

A3 Matching Quality Outside the Period of Overlap

While the computation of modes is based on the period of overlap, the results of these computations are used for the years before and after this period. Therefore, we also try to assess how adequate the correspondence tables based on the mode work outside the period of overlap. For the completion of w73, Column 2 of Table 6 shows the relative frequencies of observations for which original or extrapolated values of w73 and w93 are available. For these observations values of w73 can be compared with values inferred from w93 using correspondence table corrtab_w73_3_w93_5.dta. The two highlighted columns display the amount of correct and incorrect matches. The further away from the overlapping period 1999 to 2002 the lower is the share of correct matches. One possible reason is that the extrapolation of w93 becomes incorrect if the economic activity of the firm has changed over time.

Table 6: Comparison of w73 and the Computed Mode Based on w93

Year	Comparison of w73 and mode			Target w73 missing	Base w93 missing	Number of firms
		Correct match	Incorrect match			
	%	%	%	%	%	
1975	34.34	78.83	21.17	3.28	62.38	1,291,332
1980	41.29	80.81	19.19	3.13	55.58	1,405,930
1985	50.25	81.90	18.10	2.80	46.95	1,451,577
1990	61.58	82.97	17.03	1.53	36.89	1,548,958
1995	76.06	83.94	16.06	0.68	23.27	1,998,359
1996	80.51	84.26	15.74	0.52	18.97	2,011,954
1997	85.62	86.37	13.63	0.40	13.98	2,010,226
1998	91.32	86.67	13.33	0.27	8.41	2,040,359
1999	99.87	87.86	12.14	0.01	0.12	2,489,280
2000	99.86	87.96	12.04	0.02	0.12	2,533,494
2001	99.80	87.96	12.04	0.07	0.13	2,529,561
2002	99.66	87.96	12.04	0.21	0.13	2,492,132
2003	89.31	87.76	12.24	10.58	0.11	2,504,872
2004	79.85	87.73	12.27	20.03	0.12	2,636,572
2005	73.54	87.69	12.31	26.35	0.11	2,680,348
2006	68.20	87.61	12.39	31.70	0.10	2,734,112
2007	63.80	87.51	12.49	36.11	0.09	2,770,448
2008	60.19	87.42	12.58	39.72	0.09	2,770,250
Total	68.38	85.33	14.67	8.09	23.53	64,575,110

Notes*:* Column 2 gives the percentage of establishments for which original or extrapolated information of w73 and w93 is available. The highlighted columns 3 and 4 show the relative frequency of correct and incorrect matches as a percentage of all those observations if the correspondence table corrtab_w73_3_w93_5.dta is used.

Imprint

FDZ-Methodenreport 05/2011

Publisher
The Research Data Centre (FDZ)
of the Federal Employment Agency
in the Institute for Employment Research
Regensburger Str. 104
D-90478 Nuremberg

Editorial staff
Stefan Bender, Dagmar Theune

Technical production
Dagmar Theune

Download
http://doku.iab.de/fdz/reporte/2011/MR_05-11_EN.pdf

Internet
http://fdz.iab.de/

Corresponding author:

Dr. Peter Jacobebbinghaus
The Research Data Centre (FDZ)
Regensburger Str. 104
D - 90478 Nuremberg
Phone: +49 (0)911 / 179-4667
E-Mail: peter.jacobebbinghaus@iab.de