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

03|2021 EN ”Identifying Couples in Administrative Data“ for the years 2001–2014

Ann-Christin Bächmann, Corinna Frodermann, Benjamin Lochner, Michael Oberfichtner, Simon Trenkle

Bundesagentur für Arbeit

”Identifying Couples in Administrative Data“ for the years 2001–2014

Ann-Christin Bächmann (LIfBi)
Corinna Frodermann (IAB)
Benjamin Lochner (FAU, IAB)
Michael Oberfichtner (IAB)
Simon Trenkle (IZA, IAB)

Die FDZ-Methodenreporte befassen sich mit den methodischen Aspekten der Daten des FDZ und helfen somit Nutzerinnen und Nutzern bei der Analyse der Daten. Nutzerinnen und Nutzer können hierzu in dieser Reihe zitationsfähig publizieren und stellen sich der öffentlichen Diskussion.

FDZ-Methodenreporte (FDZ method reports) deal with methodical aspects of FDZ data and help users in the analysis of these data. In addition, users can publish their results in a citable manner and present them for public discussion.

Contents

List of Figures

List of Tables

Abstract

We apply the couple identification developed by Goldschmidt *et al.* (2017, *Journal for Labour Market Research*) to German administrative data from the years 2001 through 2014. The identification builds upon married couples sharing their surname and living at the same address. The resulting dataset includes around 42 million couple-year observations from mixed-sex couples with an absolute age-difference of less than 15 years. These observations stem from around 8 million different couples. The longitudinal dimension of the couple identifier broadens the range of potential application in comparison to the original cross-sectional identifier.

Zusammenfassung

Wir wenden die von Goldschmidt *et al.* (2017, *Journal for Labour Market Research*) entwickelte Paar-Identifikation auf administrative Daten aus Deutschland für die Jahre 2001 – 2014 an. Die Identifikation baut darauf auf, dass Verheiratete den selben Nachnamen tragen und an der selben Adresse wohnen. Der finale Datensatz beinhaltet etwa 42 Millionen Paar-Jahr-Beobachtungen von gemischtgeschlechtlichen Paaren mit einem absoluten Altersunterschied von weniger als 15 Jahren. Diese Beobachtungen stammen von ungefähr 8 Millionen unterschiedlichen Paaren. Die Längsschnittdimension des Datensatzes erweitert die Anwendungsmöglichkeiten im Vergleich zum ursprünglichen Querschnittsdatensatz.

Keywords

couple identification, household information, improving administrative data

Acknowledgments

We thank our colleagues at IAB for their support throughout this project. In particular, we thank Peter Haller and Martina Oertel for their guidance on using the geo-referenced data, Dana Müller for sharing the results of earlier work conducted at IAB's Research Data Center, as well as Deborah Goldschmidt, Wolfram Klosterhuber and Johannes Schmieder who produced the original couple identification that this project is built upon for generously sharing code and data. As always, errors are ours.

1 Introduction

Administrative data are widely used in sociological and economic research, because they offer large sample sizes, long observation periods (often the population followed over decades) as well as high quality information. Administrative data are however typically provided at the person-level and lack household-level information. This shortcoming limits the scope of analyses with administrative data. For instance, labour market researcher can often not to examine long-standing questions such as household labor supply, household investment decisions in human capital, and within-household income differences using administrative data.

To address this limitation, Goldschmidt/Klosterhuber/Schmieder (2017), henceforth GKS, develop a procedure to identify married partners in administrative labor market data from Germany. Their identification procedure builds upon married couples sharing their surname and living at the same address. Applying the procedure to data from 2008, they generate a cross-sectional dataset that contains around 3.3 million mixed-sex couples with an absolute age-difference of less than 15 years. GKS and we consider such couples as married couples. Although the procedure does clearly not identify all married couples in the data, they show that the identified couples are indeed married couples. The procedure hence yields few false positives and the identified couples are a reliable base for further research.

We update their procedure and apply it to the Integrated Employment Biographies (IEB) for the years 2001 until 2014 using newly available geocoded data. We identify more than 8 million married couples – roughly 3 million couples (or 6 million persons) in each of the 14 years. Comparing our results for 2008 with those of GKS shows that we identify a similar set of couples. The increased number of observations along with the longitudinal dimension of the couple identifier broadens the range of potential application in comparison to the original cross-sectional identifier. Our resulting dataset is available via IAB's department "Data and IT-Management".

The report is structured into three main sections. The next section shortly introduces the GKS procedure before then describing the application to the years 2001-2014. It highlights where we updated the original procedure. For a full presentation and detailed discussion, see GKS. Section 3 describes the resulting dataset and how to access the data. In Section 4 we provide a descriptive analysis of the dataset and perform consistency checks that compare our resulting dataset to GKS's.

2 The Identification Procedure

2.1 The GKS Procedure

We closely follow the procedure proposed by GKS to identify couples in the administrative data. The GKS method identifies couples based on same surname, same exact residential location, and a uniqueness condition. In particular it applies the following restrictions:

- **Same surname**: Persons have to have the same surname reported in the administrative data. Double names (e.g., Mueller-Schmidt) are allowed, if they are separated by a "-". This rules out married couples who are not sharing a surname. Requiring the same surname also rules out cases in which the surname is written in different ways (e.g., due to typos), such that a standard preparation of names does not catch the difference.
- **Same exact residential location**: Married persons have to share the exact same residential location, including municipality, postal code, street name and street number. This restriction rules out couples who do not cohabit, report different addresses or report the same addresses in different ways such that they are handled differently by the geo-referencing software.
- **Uniquely identified**: Only two persons with the same surname may report the exact residential location.

GKS additionally impose restrictions on the sex composition, i.e., mixed-sex couples only, and on the age composition, i.e., an absolute age-difference of less than 15 years. Although we keep couples irrespective in our dataset, researchers will want to impose these two additional restrictions for most analyses.

When imposing the additional sex and age restrictions, the procedure yields few type II errors, i.e. markings of persons as couple who are not an actual couple. GKS show that 89% - 94% of their identified couples are indeed married couples. At the same time, many couples that are actually married are not identified. GKS conclude that the procedure identifies about one third of all couples in which both spouses are employed subject to social security or unemployed. Furthermore, the procedure cannot identify couples, in which one or both partners are not subject to social security, for instance because they are out of the labour force, self-employed or civil servants.

2.2 Implementation: Choice of Data-Sets and Sample Restriction

2.2.1 IEB

The main source of person level data are the Integrated Employment Biographies (*IAB Integrierte Erwerbsbiografien (IEB) V13.00.01-171010, Nürnberg 2017*) described in Jacobebbinghaus/Seth (2007). The IEB originate from social security records. They comprise all persons who are employed subject to social security, marginally employed, recipients of unemployment insurance, registered job searchers, or participants of employment or training programs. The IEB covers around 80 percent of the working population in Germany, only excluding civil servants and self-employed workers.

From the IEB, we create cross-sections as of June 30th of each year from 2001 to 2014 – the years for which address information are available. We restrict our sample to persons who are at least 16 years old, the legal minimum age to marry in Germany.

2.2.2 Handling of Geocodes

For each person in the IEB, we are interested in the main place of residence. To this end, we use the IEB GEO (*Geocodes von Wohn- und Arbeitsorten der Personen und Betriebe aus IEB V12.00 (IEB GEO) V01.00.00-201504. Nürnberg 2018*). Persons' addresses originate from three sources: the social security records (*DEÜV-Meldungen*) according to the social code IV, the base data of the federal employment agency, and the base data of authorised municipal authorities according to the social code II (*zugelassene kommunale Träger*). These addresses are processed by a geoinformation system (GIS) with reference date March 2015. We rely on the 'World Geodetic System 1984', which delivers distinct X- and Y-coordinates for the exact addresses of the persons in our sample.

It is important to note that the sample construction and handling of geocodes differ substantially from that in an earlier geocoded version of IAB data (Scholz et al., 2012) that GKS used. The main changes are:

- Expanding the sample to 2000 to 2014 (previously pure cross-section for 2007-2009)[1]
- Including all establishment and person addresses (previously one selected address)

[1] To avoid initial data quality problems, we use only data from 2001 onwards.

- Changing and expanding the geocoding projecting system to Gauss-Krueger, UTM32, WGS84 (previously only Lambert projection)
- Refining of data grids

Furthermore, the IEB GEO relies on a different GIS. The redesign improved the mapping between persons from the IEB and other sources of administrative data (e.g, faster updates of moves, changing postal codes, etc.). Another important improvement has been a better handling of quality indicators.

We use two quality indicators. The first counts the number of addresses which can be matched to distinct geocodes. A lower number indicates that the matching was more precise. The second indicator measures the quality of matching with respect to i) postal code, ii) name of the place of residence, iii) street name, and iv) street number. We exclude geocodes of poor matching quality, that is cases for which we cannot reliably assign a person's address to a geocode. Specifically, we exclude observations with missing information in each of i)-iv).

After assigning the geocodes to the selected persons in the IEB, we exclude observations with geocodes for which we observe only one person because these geocodes cannot host a couple.

2.2.3 Handling of Surnames

As GKS, we assume that partners in a couple have the same surname, which almost 90% of marrying couples do (Gesellschaft für deutsche Sprache, 2018). Recall that a couple needs to be 'uniquely identified'. If we find more than two persons in one location with the same surname, we cannot reliably determine whether two person form a potential couple. We hence exclude these persons.

Surnames in our data again originate from the following three sources: the social security records (*DEÜV-Meldungen* messages, social code IV), the base data of the federal employment agency, and the base data of authorised municipal authorities (*zugelassene kommunale Träger*, social code II).

We closely follow GKS in cleaning the names of errors and typos, the handling with German special notation (e.g., ä,ö,ß etc.) and handling hyphenated names. For details, see GKS.

After these preparations, we merge the person level data that includes the geocodes to their names. Table 1 shows the success of this linkage for each of our sample years. For 94-97 % of the persons in our data, we can link the surname.

Table 1: Number of Individuals in IEB at same Geo-Code Merging of Names

year	successful merge	not merged	rate
2001	29,130,590	837,371	97%
2002	28,925,808	1,207,249	96%
2003	28,751,868	1,296,042	95%
2004	29,044,682	1,375,828	95%
2005	30,065,801	1,523,942	95%
2006	30,612,971	1,602,344	95%
2007	30,533,392	1,580,659	95%
2008	30,455,591	1,556,183	95%
2009	30,532,805	1,555,286	95%
2010	30,620,774	1,564,289	95%
2011	30,884,924	1,655,118	95%
2012	30,960,911	1,768,897	94%
2013	31,241,358	1,907,715	94%
2014	31,474,449	2,036,334	94%

Notes: The table shows the success in merging the individual level data that includes the geocodes to their last names for each sample year. Source: IEB, DEÜV messages (social code IV), base data of the Federal Employment Agency, and base data of authorised municipal authorities (social code II);

3 Data Structure and Access to the Data Set

3.1 Structure of the Data Set

The procedure yields an annual relation of persons to couples. This information is stored in a long-formatted dataset that contains the following six variables:

- *prs_id*: person identifier as in IEB V13.00.01
- *couple_id*: unique couple identifer, linking two prs_id to the same couple. Derived as described in Section 2
- *jahr*: year in which the two persons are identified as a couple
- *n_coordinate*: number of persons in IEB at geocode of this couple
- *couple_type*: indicator for sex composition and age difference in couple as in IEB V13.00.01
- *birth_year*: birth year of person as in IEB V13.00.01

The resulting dataset is an unbalanced panel on the *prs_id* $\times$ *jahr* level. If two persons form a couple in several years, they have the same *couple_id* in all years. For instance, if two persons

form a couple from 2004 to 2008 and again in 2012, the *couple_id* is identical in all years. For any given year, individuals appear in the data only as couple if they where identified as belonging to a couple in that particular year.

n_coordinate is included in the dataset as a quality indictor. The likelihood of two persons coincidentally fulfilling the criteria of the identification procedure increases in the number of persons living at the same address. This is in particular relevant for wide-spread surnames, such as Schmidt and Meyer. By imposing a maximum number of persons who live at an address, researchers can reduce the risk of including falsely-identified couples in their analysis. Such a restriction is however likely to select the sample towards a more rural population. As this trade-off will depend on the specific research question, we leave this decision to researchers using the couple identifiers.

The variable *couple_typ* allows researchers to select a specific subset of couples as appropriate for their analysis. The variable takes the following values:

1. mixed sex, absolute age difference < 15
2. mixed sex, absolute age difference $>= 15$
3. both female, absolute age difference < 15
4. both female, absolute age difference $>= 15$
5. both male, absolute age difference < 15
6. both male, absolute age difference $>= 15$

For many research questions, restricting to *couple_typ* $= 1$ appears appropriate.

The variable *birth_year* enables researchers to identify merges that have gone bad. As the couple identifier will in practice be used in combination with other data sets, comparing this indicator with the information on both individuals in the other dataset is a simple plausibility check for data merges.[2]

The variables *couple_typ* and *birth_year* record information as of the IEB V13.00.01. In rare cases, the sex and/or the date of birth of a person changes in the underlying data source. Therefore, a small number of discrepancies is to be expected. We recommend to check whether the couples whom we identified in our data also fulfill the basic criteria regarding sex and age in other dataset that are used.

[2] The age difference computed from *birth_year* differs from the age difference that is used to generate the variable *couple_typ*. To construct *couple_typ*, we calculate both partners' age as of 30 June in the respective year using their exact birthdate.

3.2 Access to the Dataset

The dataset is available ”as is“ within the IAB. Researches who want to use the data set can contact the department ”Data and IT-Management“ and obtain access to the dataset under the request number (*AMS-Nummer*) 15238. The data set will be available at least until September 2031, though the department ”Data and IT-Management“ does not provide further support.

Upon request, the department ”Data and IT-Management“ will also update the person identifier to a newer version. By using the data and requesting such updates, researchers consent to other researchers also using the updated version.

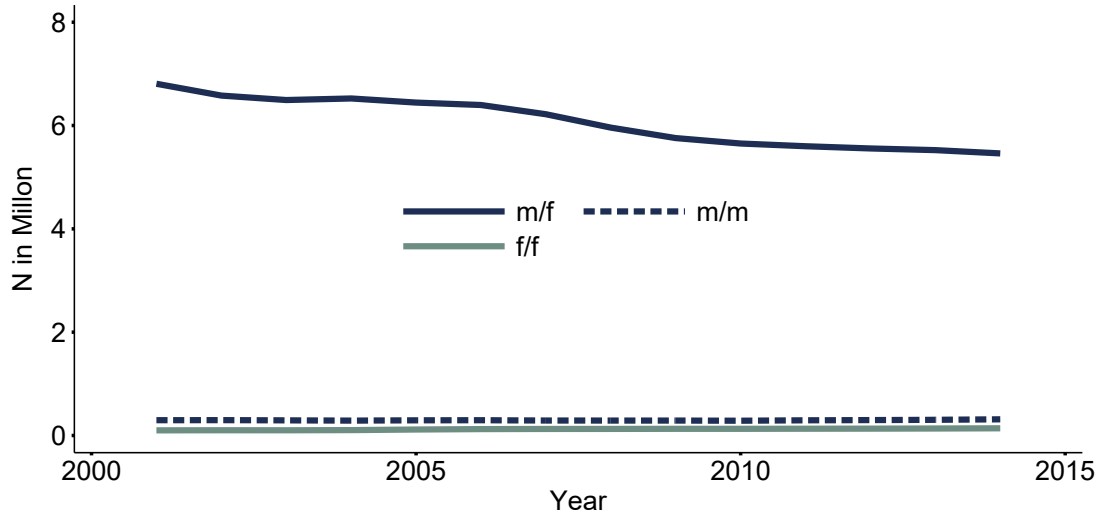
4 Descriptive Statistics

In this section, we turn to a description of the resulting data set, paying particular attention to compare our data to the to the 2008 data constructed by and reported in GKS. In particular, we repeat a number of consistency checks suggested by GKS and compare it with their findings.

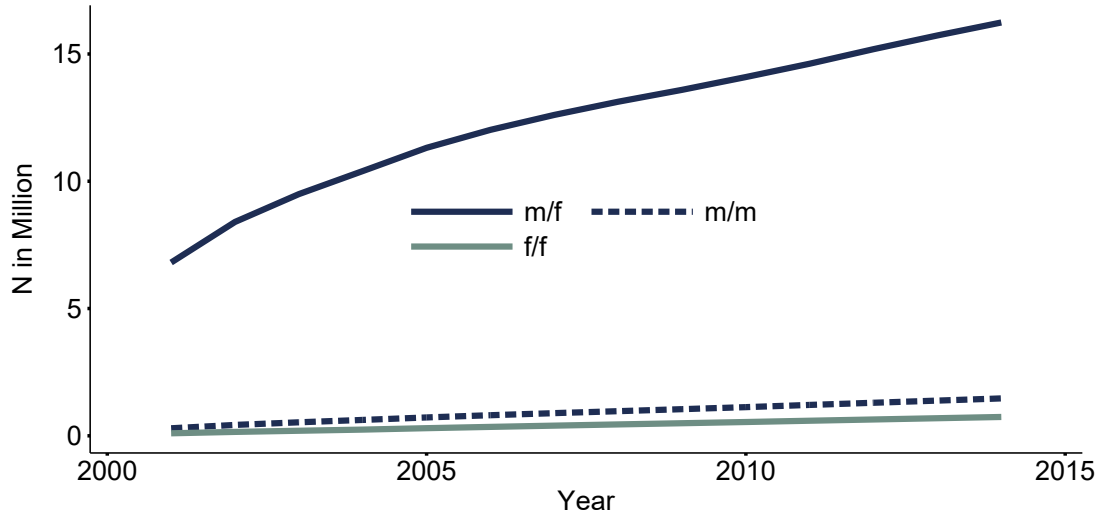
4.1 Number of Persons over Time, Individual Characteristics

In Figure 1 we plot the number of persons that are in mixed-sex couples over time in comparison to the number of observations in same sex-couples, both restricted to couples with an absolute age-difference of less than 15 years. The number of observations for male/female couples is decreasing over time, with about 6.8 million persons being in a couple in 2001 as opposed to about 5.4 million in 2014. This decline happens steadily over time without major breaks. The cumulative number of couples displayed in panel (b) increases steadily over time to about 15 million mixed-sex couples in 2014. Figure 2 plots age and age-difference in years of identified mixed-sex couples over time. It shows that the age-difference, computed as male - female, stays roughly the same within couples with males being on average 2.5 years older than their likely spouse, whereas the average age increases from roughly 42 years in 2001 to above 46 in 2014.

Figure 1: Number of Individuals in Couples by Couple-Status and over Time



(a) Number of Individuals in Couples by Year



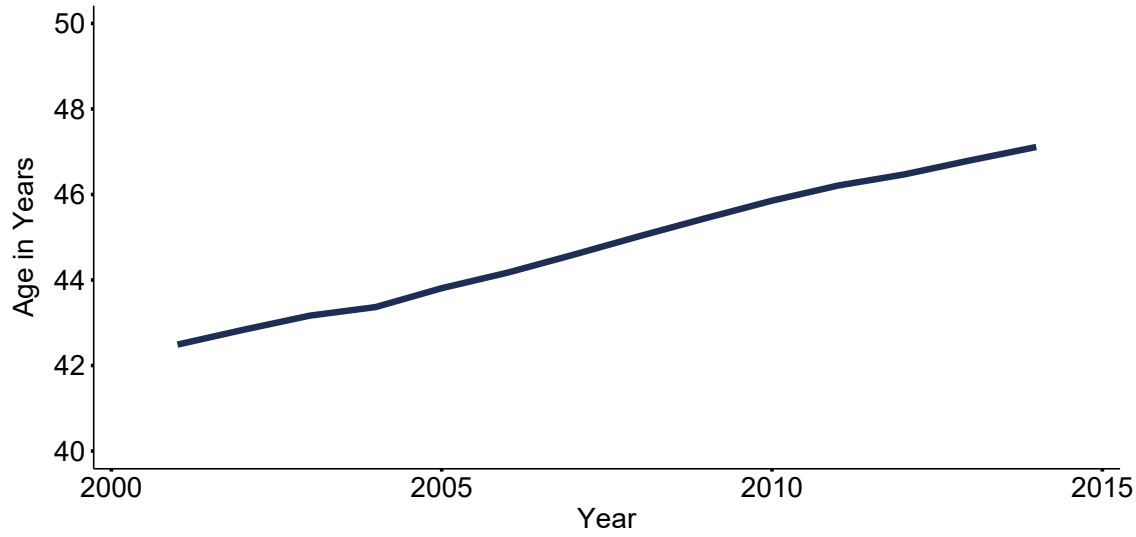
(b) Cumulative Number of Couples over Time

Notes: This figure shows the number of individuals identified as couples over time and by couple-status for couples with an age difference of below 15 years. Figure (a) shows the number of individuals that are observed in a couple in a given year for the years 2001 to 2014. In 2001 there are about 6.8 million persons in a couple and this number declines continuously to about 5.4 million in 2014. Figure (b) shows the cumulative number of individuals that are observed as a new couple in a given year. For mixed-sex couples the number in 2001 is again at about 6.8 million observations, and culminates to about 15 million in 2014.

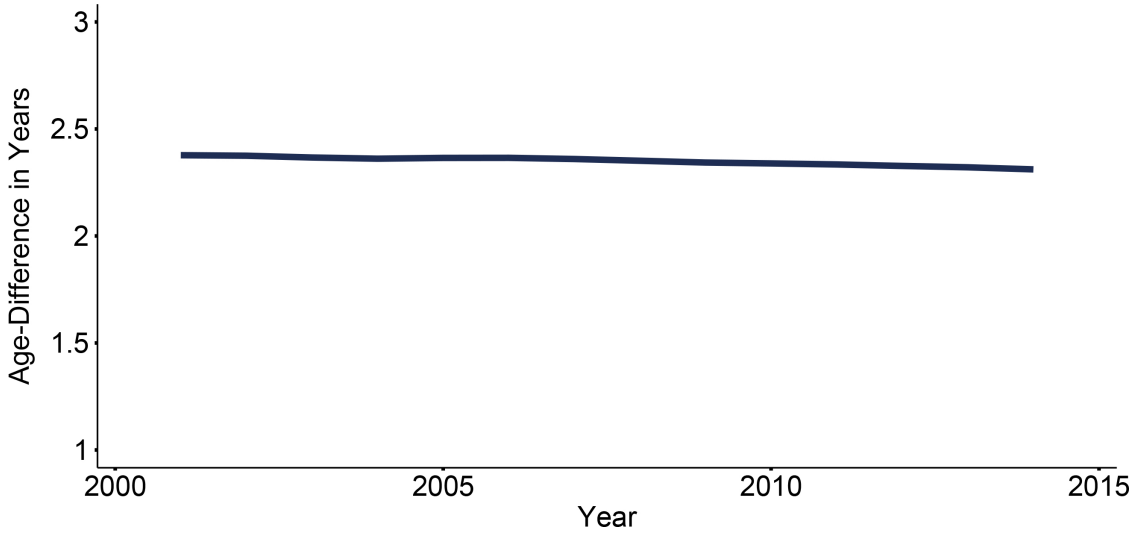
4.2 Validity of Procedure and Comparison to GKS in the Aggregate

Figure 3 plots the distribution of the age-difference in mixed-sex couples by year. The distribution follows the same pattern as already documented in GKS: Most identified pairs show

Figure 2: Age and Age-Difference within Couples over Time



(a) Age in Years over Time

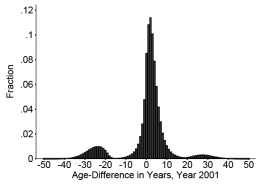


(b) Age-Difference in Years within Couples over Time

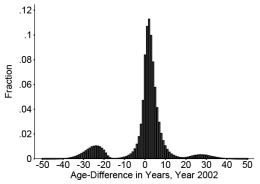
Notes: This figure shows the (a) age in years of identified individuals between 2001 and 2014. Mean age in 2001 is slightly above 42 years and increases linearly to on average about 47 in 2014. Figure (b) shows the age difference in years within a couple which lies constantly at about 2.5 years throughout the observation period.

an absolute age-difference of less then 10 years. The density is lowest at around +/- 15 years and then starts increasing again in both directions, with local maxima at around +/- 30. These local maxima indicate that they identify a parent-child link where only one parent is present (in the IEB data). Among couples with a large age difference, the share of falsely identified couples is likely larger. We therefore follow GKS and recommend excluding couples with an absolute age difference of 15 years and larger. Figure 4 shows the age-difference for the years 2001, 2008 and 2014 and compares it to the couples identified by GKS in 2008. Reassuringly,

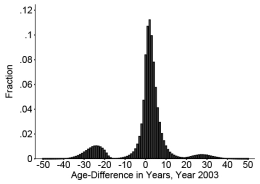
Figure 3: Age-Difference by Year



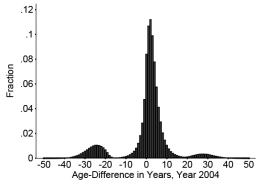
(a) Age-Difference 2001



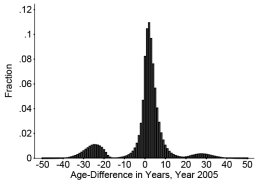
(b) Age-Difference 2002



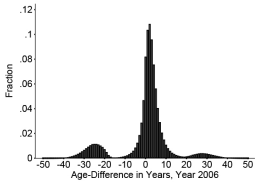
(c) Age-Difference 2003



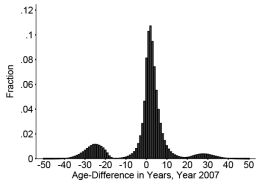
(d) Age-Difference 2004



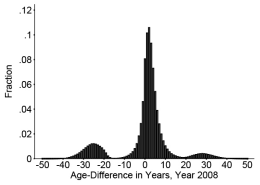
(e) Age-Difference 2005



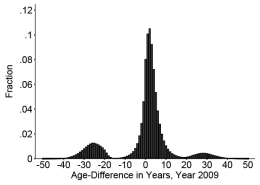
(f) Age-Difference 2006



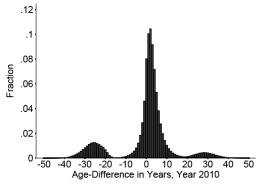
(g) Age-Difference 2007



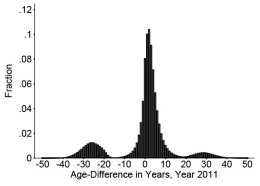
(h) Age-Difference 2008



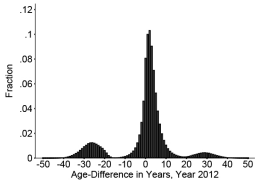
(i) Age-Difference 2009



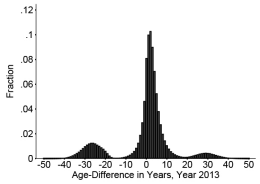
(j) Age-Difference 2010



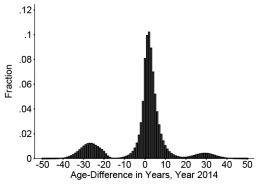
(k) Age-Difference 2011



(l) Age-Difference 2012



(m) Age-Difference 2013



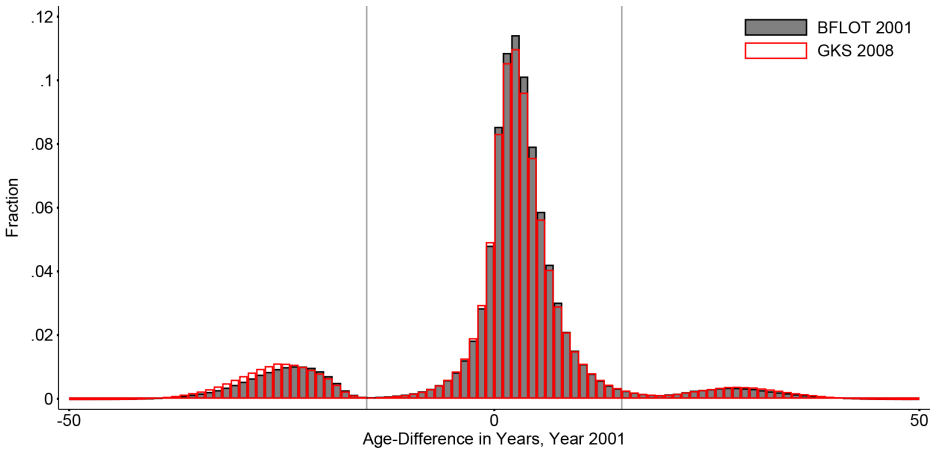
(n) Age-Difference 2014

Notes: This figure shows the distribution of age-differences in years between the male and the female partner for male/female couples by year for the years 2001 until 2014. For each year, only couples that are identified as couples in that particular year are included in the graph. Couples with an absolute age difference of more than 50 years are excluded. The distribution in most years peaks at about 2-3 years, with most observations located within an absolute age-difference of 10 years. The density is lowest at around +/- 15 years and then starts increasing again in both directions, with local maxima at around +/- 30.

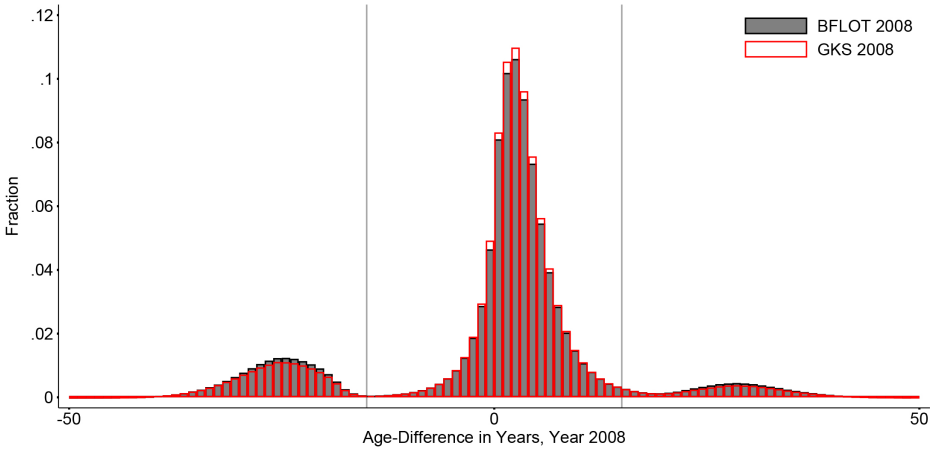
the distribution of GKS closely mimics the distribution in our data, especially in 2008. This indicates that — despite using different geocodes and some other smaller differences — the resulting data are similar in the aggregate. Moreover, the distribution in 2001 and 2014 look also quite comparable speaking to a relatively high stability of the aggregate data over time.

Table 2 presents — comparable to GKS, table 3 — the numbers of observations for the years 2001, 2008, 2014 and the GKS data for 2008, separately for persons in mixed-sex couples above and below an age difference of 15 as well as for female/female and male/male couples. The relative share of each of these groups is again similar to GKS, with a slightly higher absolute number of persons in GKS's data than in our data. We follow GKS and use the share of same-sex couples compared to mixed-sex couples among couples with an age-difference below 15 to construct an estimate of the share of correct matches. Assuming that same-sex couples are siblings and that same-sex siblings are as likely falsely identified as couples as mixed-sex siblings, we can use the ratio of same-sex couples to mixed-sex couples to the share of mixed-sex couples that are siblings. According to Figure 5 (a) about 95% of identified couples are true couples in 2001 this share decreases slightly over time to about 92%. Panel (b) shows that the matching accuracy decreases with the number of individuals at a geo-cordinate for selected years. The level and pattern for 2008 is in line with the corresponding finding of GKS.

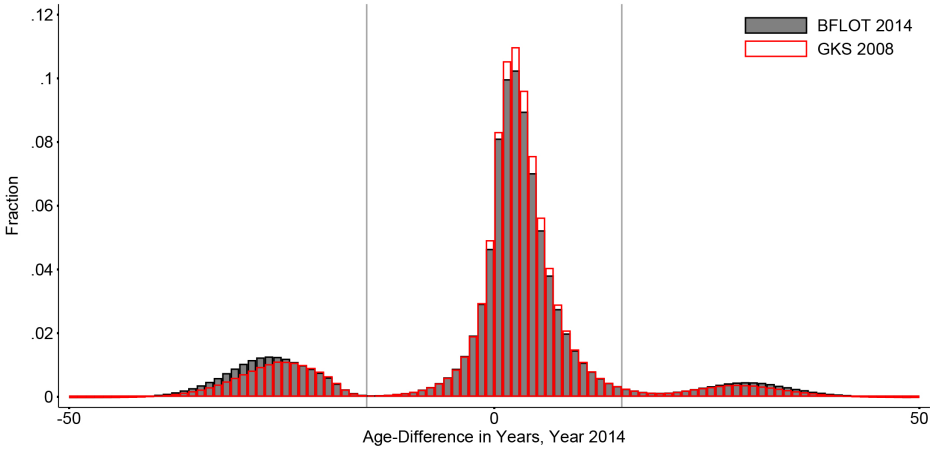
Figure 4: Distribution of Age-Difference - Selected Years Compared with GKS 2008



(a) Histogram 2001



(b) Histogram 2008



(c) Histogram 2014

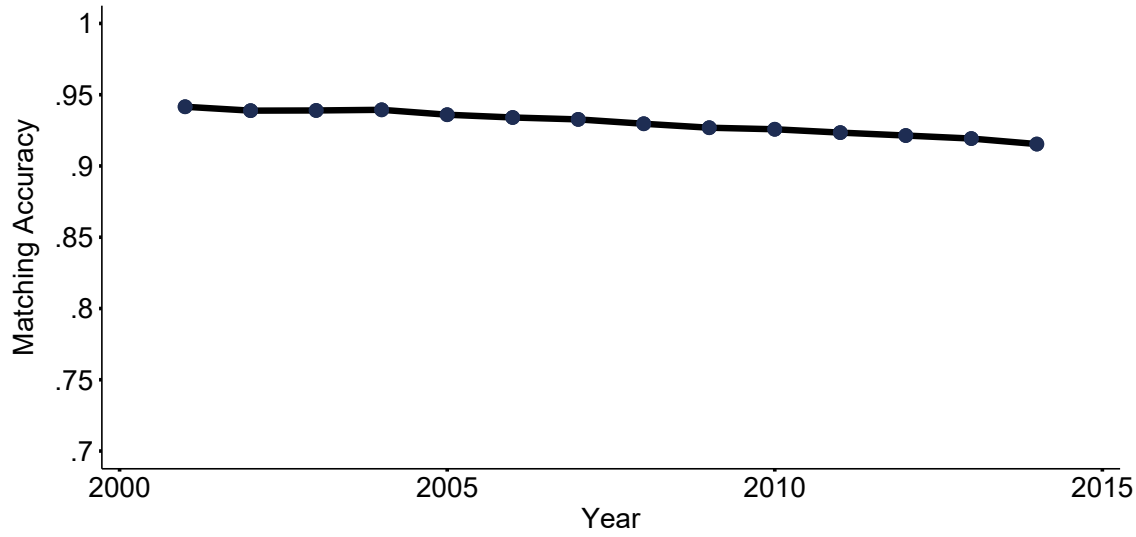
Notes: This figure shows the distribution of age-differences by year between the male and the female partner for male/female couples for selected years and compares it with the corresponding age-distribution in the GKS 2008 data. The vertical grey lines at +/- 15 years indicate the (default) range that is used in the final couple data. The age difference in our data and GKS data follow each other closely throughout to whole distribution and are also comparable for the selected years: 2001, 2008 and 2014.

Table 2: Gender Composition by Age Difference of Matched Potential Couples for Selected Years

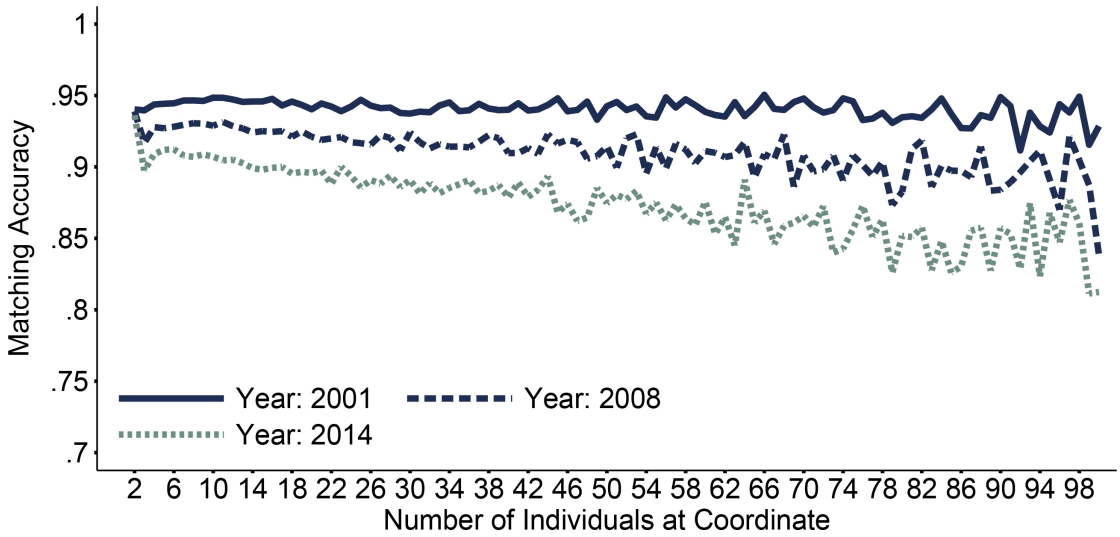
Matches	All matches		Age Difference$<$15		Age Difference$\geq$ 15	
	Absolute	Percent (%)	Absolute	Percent (%)	Absolute	Percent (%)
Year 2001						
Male/female	4,129,633	82.96	3,403,926	94.50	725,707	52.75
Male/male	528,841	10.62	148,291	4.12	380,550	27.66
Female/female	319,442	6.42	49,910	1.39	269,532	19.59
Total	4,977,916	100.00	3,602,127	100.00	1,375,789	100.00
Year 2008						
Male/female	3,812,469	79.79	2,979,794	93.46	832,675	52.38
Male/male	513,853	10.75	144,614	4.54	369,239	23.23
Female/female	451,654	9.45	63,942	2.01	387,712	24.39
Total	4,777,976	100.00	3,188,350	100.00	1,589,626	100.00
Year 2008 GKS						
Male/female	4,084,516	81.72	3,281,651	94.65	802,859	52.44
Male/male	482,891	9.66	131,550	3.79	351,341	22.95
Female/female	430,679	8.62	53,763	1.55	376,916	24.62
Total	4,998,086	100.00	3,466,970	100.00	1,531,116	100.00
Year 2014						
Male/female	3,566,198	78.18	2,730,314	92.34	835,884	52.10
Male/male	525,119	11.51	156,687	5.30	368,432	22.96
Female/female	470,074	10.31	69,946	2.37	400,128	24.94
Total	4,561,391	100.00	2,956,947	100.00	1,604,444	100.00

Notes: This table compares the number of observations for different gender-compositions on the couple level for selected years and by whether individuals above an absolute age-difference of 15 years and more are excluded or not.

Figure 5: Matching Quality



(a) Matching accuracy over time



(b) Matching accuracy by number of individuals at the same geo-code for selected years

Notes: This figure shows the estimated matching accuracy defined as $\frac{N_{fm}-N_{ff}-N_{mm}}{N_{fm}}$ over time (figure (a)) with N_{fm} being the number of individuals in mixed-sex couples, N_{ff} the number of individuals in female/female pairs and N_{mm} the number of individuals in male/male pairs. The calculated matching accuracy is at about 95% in 2001, and decreases slightly over time to about 92% in 2014. Figure (b) shows the matching accuracy by the number of individuals at the same coordinate, *n_coordinate*, for selected years. It shows that the matching accuracy slightly decreases with *n_coordinate*.

4.3 Comparison of GKS and our 2008 Data on the individual level

In a next step, we compare the couples that we identified for the year 2008 with the couples that GKS identified. Taking together the combined data-set of GKS and of our identification yields more than 7.5 million persons who are part of a mixed-sex couple with an absolute age-difference of less than 15 years, though in each of the separate datasets the number of couples amounts to roughly three millions (or 6 million persons). While the majority of these persons (64 %) were tagged as part of a couple by GKS and by us, 22 % were only identified as part of a couple by GKS and 14 % only by us (see Table 3).

Taking a closer look at the almost 5 million persons who were part of a couple in our and the GKS sample, shows that we do not only identify the same persons as part of a couple but exactly the same couples, i.e. the same combination of persons, as GKS in more than 99 % of these cases. This suggests a high accuracy in the procedure in reducing false positives.

The substantial number of couples that were only identified by GKS or by us reiterates that the identification does not identify all couples. Further checks suggest that differences in the used geo-codes cause that some couples are only identified by GKS and others only by us. Although it would in principle seem feasible to use all couples for further analysis, this would only be possible for 2008. To obtain a consistent picture over all years, our data-set thus comprises only the couples that we identified.

Table 3: Comparing individuals in GKS Couple vs. Our Data for 2008

	Freq.	Percent	Cum.
Only by GKS	1,656,120	21.75	21.75
Only by us	1,052,394	13.82	35.56
By GKS & us	4,907,194	64.44	100.00
	Total	7,615,708	100.00

Notes: This table shows the number of individuals that are in an GKS couple, in our 2008 data or in both. It restricts to mixed-sex couples with an absolute age-difference of below 15 years.

4.4 The Final Data Set

Table 4 shows the numbers of observations in the data set restricting the sample to mixed-sex couples with an absolute age-difference of less than 15 years (i.e., restricting to *couple_typ*$= 1$). We recommend applying these restrictions for analysing married couples. With these restrictions, the dataset includes more than 14 million persons from more than 8 million couples, with about 10% of persons being in more than one couple. The average duration we observe someone to be in the couple — calculated as the duration between the first and the last year we observe someone to be in a couple – is 9.3 years, 8.11 years of which we observe the couple in the data. Over all years, this results in a data set with almost 85 million person $\times$ year observations.

Table 4: Summary of final M/F Couple-Data

	All M/F Couples All Years	M/F Couples in Year 2001	M/F Couples in Year 2008	M/F Couples in Year 2014
Duration in Couple				
Year First Time in Couple	2003	2001	2003.08	2005.863
	[3.076]	[0]	[2.463]	[4.862]
Year Last Time in Couple	2011.36	2008.07	2011.93	2014
	[3.502]	[4.926]	[2.284]	[0]
Years in Couple -Total (Min until Max Year)	9.36	8.066	9.85	9.13
	[8.067]	[4.92]	[3.515]	[4.86]
Years in Couple -Only Observed Years	8.11	6.66	8.77	7.42
	[8.067]	[4.26]	[3.43]	[4.292]
Individual in more than one Couple	.12	-	-	-
	[.325]	-	-	-
Number of Observations				
N Observations	84,973,728	6,807,852	5,959,588	5,460,628
N Couples	8,118,732	3,403,926	2,979,794	2,730,314
N Individuals	14,701,609	6,807,852	5,959,588	5,460,628

Notes: This table summarizes the final couple dataset after restricting to male/female couples with an age-difference of less than 15 years. It shows the number of observations, number of individuals and number of couples and the mean of variables summarizing mean start and end year and durations in the couple-data, with corresponding SD in brackets.

5 Conclusion

This report documented a new dataset of married couples in German administrative data. Applying the imputation procedure of GKS to the years 2001-2014, we identify more than 8 million mixed-sex couples with an absolute age-difference of less than 15 years. Two factors were crucial for the feasibility of this project: First, the development of an imputation procedure by GKS including code-sharing by the authors. Second, the systematic construction, maintenance, and provision of geocodes at the IAB.

We encourage readers to further improve the data situation on households in German administrative data. In particular, we see three limitations that further research could address. First, the data focuses on married couples where both partner are observed in the IAB data. Additional analyses on differences between identified couples and other couples, building on the analysis in GKS, would help to account for differences between identified and non-identified couples. A second and related point concerns the degree of measurement error due to false positives when identifying couples. While the available evidence suggests a low share of false positives, further research could provide additional information on this measurement error by for example comparing the linked couples with administrative household information on recipients of unemployment assistance *Grundsicherung* or survey data. Third, continued development of the administrative data and geocodes offers the opportunity to update the couple identification in the future.

References

Gesellschaft für deutsche Sprache (2018): Familiennamen bei der Heirat und Vornamenprognose 2018. In: , URL https://gfds.de/familiennamen-bei-der-heirat-und-vornamenprognose-2018/, 19-12-2018.

Goldschmidt, Deborah; Klosterhuber, Wolfram; Schmieder, Johannes F (2017): Identifying couples in administrative data. In: Journal for Labour Market Research, Vol. 50, No. 1, p. 29–43.

Jacobebbinghaus, Peter; Seth, Stefan (2007): The German Integrated Employment Biographies sample IEBS. In: Schmollers Jahrbuch, Vol. 127, No. 2, p. 335–342.

Scholz, Theresa; Rauscher, Cerstin; Reiher, Jörg; Bachteler, Tobias (2012): Geocoding of German administrative data. In: The Case of the Institute for Employment Research.

Imprint

FDZ-Methodenreport No 03|2021

Publication Date
21 June 2021

Publisher
Research Data Centre (FDZ)
of the Federal Employment Agency (BA)
in the Institute for Employment Research (IAB)
Regensburger Straße 104
D-90478 Nuremberg
Germany

Download
http://doku.iab.de/fdz/reporte/2021/MR_03-21_EN.pdf

Documentation version
DOI: 10.5164/IAB.FDZM.2103.en.v1

All publications in the series "FDZ-Methodenreport" can be downloaded from
https://fdz.iab.de/en/FDZ_Publications/FDZ_Publication_Series/FDZ-Methodenreporte.aspx

Website
https://fdz.iab.de/en.aspx

Corresponding author
Corinna Frodermann
+49 911 179-6334
corinna.frodermann2@iab.de
Michael Oberfichtner
+49 911 179-4473
michael.oberfichtner@iab.de