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

FDZ-Methodenreport

09/2014
EN

Methodological aspects of labour market data

Identifying Couples in Administrative Data

Deborah Goldschmidt,
Wolfram Klosterhuber,
Johannes F. Schmieder

Bundesagentur für Arbeit

Identifying Couples in Administrative Data

Deborah Goldschmidt (Boston University)
Wolfram Klosterhuber (IAB)
Johannes F Schmieder (Boston University, NBER, IZA)

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

Abstract

We develop a new method for identifying married couples in administrative data. Using address and name data from the universe of employment records in Germany we find around 3.3 million pairs of individuals who are living at the same location, have a matching last name and are less than 15 years apart in age. We show supporting evidence that around 89 to 94 percent of these pairs are indeed married couples and provide careful consistency checks. Using information from the German Microcensus, we show that our method identifies about 18 percent of all married couples in Germany and about 25 percent of couples where the husband is younger than 65. Our method thus opens the door for household level analyses benefiting from the precision and very large number of observations available in administrative data.

Zusammenfassung

Wir entwickeln eine neue Methode zur Identifizierung von Ehepaaren in administrativen Daten. Anhand von exakten Adressen und Nachnamen von allen Sozialversicherungsmeldungen in Deutschland im Jahre 2008, identifizieren wir ca. 3,3 Millionen Paare von Individuen, die an derselben Adresse wohnen, einen gemeinsamen Nachnamen teilen und weniger als 15 Jahre Altersunterschied haben. Wir zeigen mittels verschiedener Konsistenzchecks, dass es sich bei ca. 88 bis 94 Prozent dieser Paare um verheiratete Ehepaare handelt. Vergleiche mit dem Mikrozensus zeigen, dass wir mit unserer Methode ca. 18 Prozent aller Ehepaare in Deutschland identifizieren und ca. 25 Prozent aller Ehepaare in denen der Ehemann jünger als 65 Jahre ist. Unser Verfahren ermöglicht zahlreiche neue Forschungsansätze zur Untersuchung von Haushalten mit hochwertigen administrativen Daten und großen Beobachtungszahlen.

Keywords: couples, geocoding, administrative data, household analysis

We would like to thank Stefan Bender, Claudia Olivetti and Daniele Paserman for many helpful comments. All errors are our own.

1. Introduction

Recent years have witnessed a dramatic rise in the use of administrative data in economic research, facilitated by increases in computing power and the availability of new administrative data sources. The main advantages of administrative data have been large sample sizes compared to survey data, often covering the entire universe; the ability to follow the units of observation over time; and the high quality of recorded information. This shift has been particularly forceful in Labor and Public Economics, where the availability of individual level employment and tax records has led to the rise in new research designs such as regression discontinuity, regression kink or bunching designs that rely on very large sample sizes. While administrative data offer many advantages, they also come with limitations and the scope of available variables is often quite limited compared to household surveys. In particular administrative employment records are typically on the individual level only and it is often not possible to link individuals to other household members. For this reason administrative data have played a much smaller role in studying traditional questions in labor economics, such as household labor supply, household investment decisions in human capital or within household income differences.

In this project we develop a new method to impute household identifiers in the administrative employment records data in Germany to increase the scope of research questions that can be addressed. Our approach is to identify pairs of individuals who are, with a high probability, married couples using information on addresses, family names and dates of birth. In Germany it is still very common that at the time of marriage one spouse (in the vast majority of cases the wife) adopts the other spouse's last name, either fully or as part of a double name. If two individuals with matching last names are living together at the same address, they are likely related, though they could also be in a sibling or parent-child relationship. To further narrow it down to married couples we take pairs of a woman and a man with matching last names with an age difference of less than 15 years, which should exclude most parent-child relationships. We present a detailed analysis of the likely extent of errors when applying this method.

Germany has a long tradition of women taking on their husbands' last name at the time of marriage. The German Civil Code from 1896 unequivocally required that the wife takes on the name of her husband.[1] A reform in 1953 allowed for the wife to keep her birth name as part of a double (or hyphenated) last name, but she was still required to take on her husband's name as the family name. The family name law was revised again in 1970 allowing

[1] See Sperling (2012) for a discussion of the legal history of the family name law in Germany.

that a couple could decide to take on the wife's name as the family name, but kept the requirement of a common family name for both spouses. Furthermore if a couple could not come to an agreement with respect to which name would become the family name the decision was up to the husband. This only changed with a decision by the German constitutional court in 1991 and a subsequent revision of the family name law in 1994, after which both spouses were allowed to keep their own birth names, while the traditional option of taking on one of the birth names or a hyphenated double name for one of the spouses continued to exist. In practice it appears that it is still the case that the vast majority of women take on their husband's names either fully or at least as part of a double name. While we are not aware of representative surveys or official registry data for Germany that would allow us to calculate the share of couples with matching last names, we found various press reports from city level wedding registries that seem to suggest that even among newly wedded couples around 85 to 90 percent still have a matching last names.[2] Among couples married for longer (and in particular before 1994), the ratio is likely significantly higher.

We implement the method of identifying likely couples using last names, addresses and age using a cross-section of the administrative data from the Institute of Employment Research in Germany spanning the universe of employment and unemployment records for 2008. This data covers all individuals who are employed in employment subject to social security contributions, receive benefits from the unemployment insurance (UI) system, or who are registered as job seekers. This data covers around 80 percent of employees, in particular excluding public servants and the self-employed. By design we are only able to identify married couples where both spouses are covered in the IAB data. While this is certainly not a representative sample and excludes a sizable part of the population of couples we are still able to identify over 3 million couples who are likely married to each other. The two main concerns with this approach are the potential for false positives and false negatives. False positives may arise because people with matching last names may live at the same address either purely by chance, or because they are related to each other but not married. Using the distribution of same-sex matching name pairs, as well as information on family status for a subset of individuals we show that likely around 88 - 94 percent of our sample of couples are indeed married to each other. Even if both spouses of a married couple are in our data, false negatives may arise, because we may not match them to each other. Either they do not have

[2] All-in (2006) report that in Kempten in 2006 around 14 percent of newly married couples keep separate names. Louis (2010) reports that a small survey among marriage registries in 5 German cities yielded that around 8 to 20 percent of couples keep separate names, with the higher number coming from more liberal cities. This also seems to refer to newly married couples, which suggests that the ratio of couples with separate names among the pool of existing couples is likely much lower.

matching names or there are more than 2 matching individuals at a location, making it impossible to tell who is married to whom. False negatives will also arise whenever one or both members of a marriage are not covered in the IAB data, which for example would include all self-employed, public servants or individuals not in the labor force, but also all individuals older than age 65. Using information from the Microcensus, we show that we can identify roughly 20 percent of the 18 million married couples in Germany. Furthermore given that many couples in Germany are older than 65, a group of individuals who are not covered in our administrative data, we capture around 30 percent of the couples where the husband is younger than 65. We also compare the observable characteristics of our matched couples with the official census data to show how our sample differs from the general population of married couples.

This paper is related to other research that uses the special features of administrative data to impute information that is not directly available. For example Jacobson, Lalonde and Sullivan (1993) use the combination of individual and firm identifiers in UI records from Pennsylvania to impute plant closings and mass-layoffs by observing when large numbers of individuals are moving away from firm identifiers and are scattered across many other employers. Hethey-Maier and Schmieder (2013) use a similar approach to identify new plant openings in administrative data, relying on worker flow information to distinguish plant openings from spurious changes in firm identifiers. Goldschmidt and Schmieder (2014) identify outsourcing of labor services in large firms employing an algorithm based on a combination of worker flows, industry and occupation codes.

The next section describes the data used in this project. Section 3 describes our method for identifying couples and presents the results based on individuals in 2008. In section 4 we show supportive evidence that our method does in fact largely identify married couples and develop bounds on the fraction of false positives. We then present characteristics of the couples that we identify with our method and compare them to the general population in the German employment data, as well as to other data sources. Section 5 concludes.

2. Data Sources

In this chapter, the sources of the data are explained in detail. Section 2.1 describes the Integrated Employment Biographies (IEB) data, while the geocoded location data and the individual name data are discussed in 2.2 and 2.3.

2.1 Integrated Employment Biographies

The Integrated Employment Biographies (IEB) of the Institute for Employment Research (Institut für Arbeitsmarkt- und Berufsforschung - IAB) stem from the notification process of the social security system of the Federal Employment Agency (BA). The IEB consolidate completed, historicized and edited process data from different data sources, which come from different operative systems. It comprises all persons registered with the Federal Employment Agency due to the following:

- Employment subject to social security or marginal part-time employment
- Receipt of unemployment insurance benefits in accordance with Social Code Book II or III
- Job search registered with local employment agencies
- Planned or actual participation in an employment or training programs

The IEB includes demographic variables such as nationality, birthdate, gender, education and family status. Information on employment, benefit receipt and job search include daily wage, daily benefit rate, occupational and employment status or economic activity. Additionally location data such as place of residence or work on different aggregated levels are provided. There are at least 40 million working individuals in Germany, about 80 percent of whom have at least one record in the IEB. The biggest groups which are not included in the biographies are self-employed workers and public servants called *Beamte*.[3]

2.2 Geocoded Data

Our method relies on finding individuals living at the same location. In principle individuals can be matched to other individuals at the same location either by directly comparing addresses, or by first geocoding addresses into latitude / longitude coordinates and then comparing coordinates. Matching addresses directly is complicated by the fact that these can often be written in a variety of ways and need to be carefully cleaned. We instead match individuals on geographic coordinates, where the address processing was done using GIS software, which allows for careful error correction methods. The geocoding was done in a project between the Research Data Centre (FDZ) and the University of Duisburg-Essen for a cross-section of all individuals in the IEB data as of June 30th, 2008. This project used data from the Federal Agency for Cartography and Geodesy, and includes 22 million addresses of

[3] see Schild et al. (2014) page 3

German buildings and their geographic coordinates and it was possible to successfully geocode 94.6 percent of the IEB records.[4]

2.3 Names

One of the criteria that we use for determining couples is whether the last names of two people match. We therefore also obtained data on last names covering the universe of individuals who have a record in the IEB as of June 30th, 2008. In order to improve the probability of success in matching, we first clean the names of errors and typos, and ensure consistency in terms of special characters and titles. With the support of the German Record Linkage Centre (GermanRLC) and their algorithm, the names of the individuals were cleaned, taking into account certain patterns and potential discrepancies.[5] Umlauts were substituted (ä → ae and so forth) as well as ß to ss. All blank spaces in the front, middle or end of the name were removed. Professional and nobility titles (such as Dr., Prof., Freiherr von) were removed as well, and special characters (e. g. ~ or %) and non-ASCII characters (e. g. © or ™) were deleted.

The only special character that was retained is the hyphen (-), which is used to indicate double names. While the family name law in the civil code book states that a spouse can add their birth name to the family name does not specifically mention a hyphen, in practice this appears to be the only option. In fact a court decision from 2013 specifically ruled that a couple was not allowed to combine the birth names of two spouses without a hyphen (Kammergericht Berlin 2013). Furthermore individuals are not allowed to create last name chains that involve more than one hyphen (for example if at the time of marriage an individual already has a double name from a previous marriage). We thus assume that double names are always separated by a hyphen and we describe below how we use hyphenated names in our name-matching algorithm. At the end of the cleaning process all letters were converted to upper case.

Although individuals have a consistent personal identifier, the Einheitliche Statistische Person (ESP), the last name may vary across different data sources. If, after the name cleaning process was completed, discrepancies persisted in the names across data sources, the individual was dropped. The exception was when an individual had a double last name in one source and an overlapping single last name in another (e.g. MUELLER-MEIER in one source and MEIER in another). In this case, the double last name was kept.

[4] see Scholz et al. (2012)
[5] See for example Schild et al. (2014) page 4 ff.

3. Identifying Couples

As mentioned previously, although the IEB data consists of a large amount of information on the majority of the German population, it – like many administrative data sets – does not include any information on the household. To circumvent this issue, we combine the IEB data with the geocoded location data and information on names to infer probable married couples. We use the following criteria to ensure that the matches we identify are most likely married couples and not simply two people with some other type of relationship (or no relationship at all):

1. Same home location

2. Uniquely matching last name

3. 1 male, 1 female, with an age difference of less than 15 years

We go into more detail on each of these requirements below.

3.1 Location

The first step in identifying potential married couples is finding people who live at the same location, since most married couples live together. We start by looking at the distribution of the number of individuals at a particular location, using each person's geocoded coordinates, for the ~33 million people in our data. The second column of Table 1 shows this distribution. Coordinates with a small number of individuals likely represent single-family homes, while coordinates where a larger number of individuals live are likely apartment buildings or other multi-unit residences. About 5 million individuals live alone at a coordinate – we eliminate these people from our set of potential couples, leaving us with about 28 million individuals. About 7.4 million individuals live at a location with exactly 1 other person in the dataset; as the number of people living at a coordinate gets larger, the absolute number of people living in this type of residence decreases.

3.2 Names

Next, we look at the cleaned names of the individuals living within any given location. We require that our identified married couples share a last name. In situations where any of the people in the location has a hyphenated name, we consider two names to be a match if at least one part of the hyphenated name is identical to another name at the location. In locations with multiple people, we additionally require that a maximum of two people have matching names. Otherwise, we have no way to determine which two individuals are likely to be a couple and which may be unrelated, or related in other ways. The following examples help to clarify the procedure.

In Example 2.1, there are 5 individuals living at a particular coordinate. Two have the last name COHLE, and there are no others names COHLE at this location, so they are kept as a potential match. Two are named HART, with no others named HART, and so they are also kept as a potential match. Finally, there is a single person named MEIER, who is dropped from our potential group of couples. In Example 2.2, we again have 5 individuals living at the same coordinate: three have the last name COHLE, one has the last name HART, and one has the last name HART-MEIER. Because there are more than 2 individuals at this location with the last name COHLE, we can not be certain which of these are part of a couple and which are not, so we drop all three. Because HART and HART-MEIER share a partial name, even though one is hyphenated, they are kept as a potential match. In Example 2.3, there are again 5 individuals at the same coordinate. Because COHLE, COHLE and COHLE-MEIER all match in terms of their names, we must eliminate all three, since we have no way of knowing which two could really be a couple. Similarly, MEIER, MEIER-MUELLER and COHLE-MEIER must all be dropped, despite their names matching. Therefore, in this example, there is no match chosen.

After running this algorithm over the 28 million individuals, we are left with about 5 million pairs (ten million individuals) who share a location and last name. The third and fourth columns of Table 1 show the number and percent of people that were matched through this algorithm, organized by the number of individuals at a location. For coordinates with only 2 individuals, almost 70 % had matching names. At coordinates with 3 or more people found at the same location, the match rate is between 20 % and 30 %.

There are several limitations to this criterion. First, while the majority of married couples in Germany share a last name (or part of a double name), not all women (or men) change their last name upon marriage, and we are certain to miss those couples. Second, in locations with multiple people where more than two share a last name, since we can not be certain which two members are married (if any) we must drop them all, eliminating more potential matches from our sample. Finally, we may be capturing two people with the same last name living in the same coordinate who are related but not married. In addition, particularly in multi-unit residences, there may be two people who are unrelated but have the same last name, and we may erroneously be including them in our sample. Our next criteria, on gender and age, will eliminate some of these falsely matched people from our sample, but not all.

3.3 Gender and Age

Finally, we take our set of potential couples – groups of two people who share a last name and a location – and impose gender and age restrictions. Since we are currently only identifying heterosexual couples, we require that each couple be composed of one male and one

female, information that is available in the IAB records. The second column of Table 3 presents the gender composition breakdown for the 5 million identified potential couples. More than 4 million of these pairs consist of one male and one female, while the remainder is made up of either two males or two females. We drop the single-sex households and move on to the age difference requirement.

We first look at the distribution of age differences among matched pairs by gender composition. Figure 1 graphs the distribution of the age difference between the two members of the couple, defining the difference as the man's age minus the woman's age. The majority of the mass lies between -15 and +15. This likely includes the majority of married couples, although it could also include brother-sister pairs (or other closely-aged family members, such as cousins). It may also include some unrelated people who simply live in the same location and have the same last name. There is a smaller mass for pairs where the female is 20-40 years older than the male, which is likely to include mothers living with their sons, and an even smaller mass for pairs where the male is 20-40 years older than the female, which likely includes father-daughter pairs. These parent child relationships may either be single parents or families where only one of the parents are working in employment covered in the IEB. The fact that there seem to be more mother-son pairs than father-daughter pairs is likely explained by the fact that there are more single mothers than single fathers.

Figures 2 and 3 show the age difference distribution for matched pairs with the same gender, where the age difference is defined as the older age minus the younger age. For both of these, the majority or pairs fall between 15 and 40, which is likely to consist mainly of mother-daughter or father-son pairs. There is also some mass for pairs with an age difference of 0-15 years; these may be siblings or other familial relationships, homosexual couples, or other pairs of people who coincidentally have the same last name in the same location. While homosexual couples can form a civil union in Germany since 2001 which allows them to adopt a common family name, these still seem to be relatively rare, with only 34,000 same sex civil unions in 2011 (Statistisches Bundesamt 2012). Thus while a small part of the same sex matches might be same sex couples most of them are not. The fact that the number of same sex matched individuals in our sample is quite small, suggests that there are relatively few cases where people live together with the same last name for other reasons than being married to each other and that in turn most matched individuals who are living with each other in this age group are in fact married to each other.

For determining our sample of couples, we require that the difference in age of the matched man and woman be less than 15 years. This should eliminate any mother-son or father-daughter pairs from the set of couples. The remaining pairs – consisting of one man and one

woman, with matching last names, who live in the same location and are less than 15 years apart in age – make up our final sample. Columns 4-5 of Table 3 show the results when we impose our age difference restriction. We retain 80 % of our male-female couples, leaving us with a final sample of about 3.3 million couples. This sample should be primarily composed of true couples, although some share will be "false positives", made up of male-female siblings or family members who are similar in age, or unrelated people with the same name living at the same coordinates.

4. Consistency Checks

Errors in our matching algorithm could occur in two ways. First, we have false positives – two people who are matched to each other by our algorithm, but who are not really a married couple. Second, there are couples that we do not pick up with our matching method, for various reasons. We discuss these two issues, and the steps we take to quantify their magnitude, below.

4.1 False Positives

One type of error that could occur is when our algorithm matches two people who are not really married to each other, also known as type 1 error. Pairs in our sample may be wrongly matched if: (1) they are brother and sister, or have some other family relationship, are close in age, and live in the same location; or (2) they are unrelated, but living in a multi-unit residence, such as an apartment building, and happen to have the same last name and are close in age.

We can try to measure the size of this type of error in our final sample of couples in a few ways. First, we can use the distribution of same-sex matches to give us a sense of what share of our sample are wrongly matched if we make the following two assumptions. The first assumption is that opposite-sex family members who are close in age (i.e. brother and sister) are as likely to live together as same-sex family members (two sisters, for example). The second is that it is as likely for two people of the opposite sex who live in the same building to share a last name as it is for two people of the same sex. Using these assumptions, we can look at the number of same-sex matched pairs that fall within our age difference restriction (ages within 15 years of each other), using the numbers provided in Table 3 – these couples are likely either pairs of family members living in the same location, or unrelated people with the same last name in the same building. We find that there are 185,313 male/male and female/female pairs that fall within our age restriction. So, it is likely that approximately 185,000 couples in our sample of matched male-female couples with age difference under 15 years are also wrongly matched. In fact, since there are some same-sex civil unions

where partners share a family name, this arguably overestimates the number of false positives by a small amount.[6] Using this methodology, our accuracy rate is around 94 % (final sample is 3,281,657; estimated wrongly matched is 185,313; correctly matched = 3,281,657-185,313 = 3,096,344; accuracy rate = correctly matched / final sample = 3,096,344 / 3,281,657 = 94 %). So, according to this method, only about 6 % of our sample is wrongly matched and our sample does indeed identify couples who with a high degree of certainty are indeed married to each other.

Next, we use the "Family Status" variable to perform an additional check on the validity of our sample. This variable is available as part of the Jobseeker-History ((X)ASU) dataset, and thus is only filled in for a small subset of people – those who are registered as job seekers as of June 30th, 2008.[7] From our sample of approximately 10 million matched individuals, about 1.5 million have the family status variable filled in. The variable takes on four possible values: living alone, cohabiting, single parent, or married. Table 4 depicts the distribution of family status values across all individuals with a matched name within their location. Although 85 % are missing the family status variable, of those in the data with a family status listed, approximately 64 % are listed as married, 22 % are listed as living alone, while the rest are either cohabiting or are single parents. We investigate further by looking at the combination of family status for matched pairs, shown separately by gender composition and age difference (Table 5). When we look at male-female pairs with an age difference under 15 years, we see that, for couples with at least one family status listed, they are listed as either both married or one married-one missing family status 89 % of the time. This is far higher than for same-sex pairs with age difference under 15 years, who are listed as both married or one married-one missing only 9 % of the time. Male-female couples with an age difference of 15 years or more are listed as both married or one married, one missing 25 % of the time. This could either indicate that there are some married couples with an age difference of larger than 15 years, but could also be because these are indeed parent-child relationships where the spouse is not covered in the data (or does not share a last name).

Using the information in Table 5, we can also estimate the share of matches in our final sample that are likely to be true couples and not wrongly matched people (i. e. our "accuracy rate") using the subsample of couples with at least one family status listed. If we think that

[6] Statistisches Bundesamt (2012) states that there are about 34 000 same sex civil unions in Germany in 2011. We do not know how common it is for same sex couples to adopt a common family name, nor that they would both be employed and covered in our data. It appears that due to the small number of same sex civil unions our method for identifying male-female marriages would not work as well for identifying same sex civil unions.

[7] These are typically either people who are unemployed (in particular unemployment insurance recipients are required to register as job seekers) or who expect to be unemployed soon.

the family status variable is accurate, then the set of "true" couples in our sample should be 578,088: the number of couples who are listed of either being both married or one married, on missing family status. Even within these there may be individuals who were mistakenly matched. For example, there may be a job-seeking man with the last name MUELLER, whose wife is out of the workforce (and hence is not included in the IEB data), living at the same coordinates as a similarly-aged jobseeker woman with the last name MUELLER whose husband is not in the IEB data either. Our matching algorithm would connect these two jobseekers, who are both listed as being married, even though they are not actually married to each other. If we think that it is as likely for two individuals of the same gender to be wrongly matched in this way as it is for two opposite-gender individuals, then we can use the information on family status for same-sex pairs for our accuracy estimate. Specifically, there are 5,173 (637 + 4,536) same-sex matched pairs with age difference less than 15 years where family status is listed as both married or married-missing.[8] Since we know that these are wrongly matched pairs, we can assume that the same number of opposite-sex pairs was wrongly matched as well. So, the estimated "true" number of couples in the subsample of couples with family status is 572,915 (578,088 matched M-F with age difference < 15 and family status married-married or married-missing minus 5,173 same-sex pairs with age difference < 15 and married-married or married-missing status). Since our full sample of matched couples (with family status) is made up of 649,643 (3,281,657 – 2,632,014) couples, our estimated accuracy rate is 88.2 % (572, 915 "true" couples / 649,643 total couples in our final sample of couples with family status filled in for at least one of the members), or 11.8 % error rate.

We may expect fewer errors of this type in our matching algorithm if we restrict our focus to coordinates with exactly two people – in this case, there are likely to be fewer mismatched pairs of the type described above. When we repeat the accuracy rate estimation, restricting our sample to matched couples living at coordinates where exactly 2 people live, we find that to be the case: our estimated error rate is likely a bit lower, around 8.6 % (see Appendix Table A1).

While using the job-seeker data is helpful for estimating the likely fraction of false positives, it should be kept in mind that neither is this subsample representative, nor necessarily is family status measured without errors. It may well be the case that we are overestimating or underestimating the number of false positives here. Overall, based on the two approaches dis-

[8] We are again being conservative here, assuming that among the same-sex matched couples, none are true couples (same-sex civil unions). As discussed before this is likely a very small group.

cussed, we estimate that the fraction of false positives lies somewhere in the range 6 % to 11.8 %.

4.2 Missing Couples

Given the data we are using and the matching algorithm we have developed, we are likely to have missed many true married couples, either among individuals who are in our dataset (a form of type 2 error) or where at least one spouse is not covered in the IEB. According to the Microcensus (Statistisches Bundesamt 2012), there were 18,008,000 married couples in 2011; of those, about 8.8 million had 2 or more working individuals in the household. About 12.5 million of the couples had a husband under 65 years old. In our final sample, we have 3.2 million couples. Therefore, we capture about 18 % of the total number of married couples, or 26 % of those with a husband under age 65.

There are several types of couples that we are likely to miss in creating our final sample. Any couple where one or both members is not in the IEB would be omitted – for example, if one member is not in the labor force, or is a public servant or other type of worker not covered by the IEB. It covers about 80 % of the German workforce, which includes approximately 60 % of the German population age 15 and above, according to the World Bank.[9]

If the couple does not share a last name (or part of a hyphenated name), then we would not capture them with our algorithm. Until 1991 it was required by German law that married couples share a last name, and even afterwards most change or hyphenate their last name upon marriage. Although we were not able to find official statistics on this topic, according to several newspaper articles the share of new couples who share a last name is around 85-90 %. Couples where one or both members are non-German are the least likely to share a last name.

Couples where the age difference between the husband and wife is more than 15 years are omitted from our sample in an effort to ensure that we do not mistakenly include parent-child pairs in our sample. Although there are certainly married couples with a 15-year or larger age difference, the number of these types of couples is quite small. For example, in the Microcensus, a representative survey of German households, the share of couples with a 16-year or more age difference was only 2 % in 2011.

We also investigated the likely impact of our age restriction using the marital status variable available in the job seeker data. For the subsample of couples where we have the marital

[9] See data.worldbank.org, Labor Force Participation Rate

status for at least one of the two individuals, in Figure 4 we plotted the share of couples where either both were reported as married or one person was married and the other person's marital status was missing. Matched couples where both are married seem to be very rare when the woman is older than 15 years than the man. This suggests that there are almost no true couples that we are missing with the 15 years age difference restriction. On the other end there is still a high share of couples where the man is around 15 to 20 years older than the woman where both are reported as married. If these are true couples, then we are excluding them from our set of likely married couples. Notice however that while the share is significant, Figure 1 shows that there are almost no couples in the 15 to 20 years age window (consistent with the information from the Microcensus), again suggesting that the 15 years age difference restriction does not exclude many true couples.[10] There are more matched pairs in Figure 1 where the man is around 25 years older than the woman, but Figure 4 shows that that is exactly where the share of married/married is falling to zero, thus suggesting that here we have mainly pairs who are not matched to each other.

Couples not living together on June 30th, 2008 are impossible for us to identify with our data; however, we believe that this situation is likely to be rare.

If the couple lives at a location with more than 2 people with the same last name at the same coordinate, we have no way of knowing which two people are part of a couple, and so all are dropped (about 5.2 million).

We drop people who have inconsistent names across data sources, thus potentially omitting more couples from our sample (about 1.8 million).

We can get a sense of how representative our final sample of couples is by comparing their characteristics to those of a truly representative sample of couples, those in the Microcensus. Table 6 compares individual characteristics of people in our final sample of couples (column 4) to couples in the Microcensus in 2011 (column 7). In terms of the age distribution, our men and women tend to be a bit younger than those in the census couples; this can be explained by the fact that our sample only includes people in the workforce, so older workers who are more likely to be retired are excluded. In addition, anyone married to a retired person will be omitted from our final sample, since their spouse will not be in our original dataset. If we exclude the individuals over age 65 and rescale the Microcensus numbers, they come closer to ours, although still skewed a bit older - 11 % of men are < 35, 25 % between 35 and 45, and 64 % between 45 and 65. Similarly for women, re-scaling the Microcensus

[10] This can also be seen from Table 5 if we look at the subsample of our matched couples with the family status variable available. Of the male-female pairs where both are listed as married, only 3 % have an age difference of 15 years or more.

numbers brings them a bit closer to ours: 15 % age < 35, 26 % between age 35 and 45, and 59 % between age 45 and 65. If we look only at couples in our final sample who live at coordinates with exactly 2 people (column 4), the age distribution becomes even closer (couples living at 2-person locations may be a more accurate sample, since there is less of a possibility of false matches; it may also favor older couples, who may be more likely to live in single-family homes rather than multi-family apartment buildings).

Looking next at the labor force status, we do not have the full range of labor force status options that are available in the Microcensus, since the IAB data only includes people in the labor force but omits self-employed and public servants. If we again rescale the census numbers for the categories that are available in the IEB, we find them to be relatively close to ours – 96 % employees in the Microcensus versus 88 % employees in our final sample. Again the sample with exactly 2 people in a location is even closer to the census, with 93 % employees.

Turning to Table 7, we can compare the characteristics of couples in the different data sets. The distribution of age difference within couples of our final sample (column 3) is almost exactly the same as that of the Microcensus. The couples in our sample are more likely to be both German and less likely to be both non-German than those of the Microcensus; as mentioned earlier, non-Germans are less likely to change their name at marriage than Germans are, and so are more likely to be omitted by our matching algorithm. Finally, household income tends to be higher in our sample than in the Microcensus, although this is likely due to the fact that the census reports net income while the income data in the IEB is gross. In addition, the Microcensus contains individuals who are retired or otherwise out of the workforce and therefore likely to have a low or no income, while our sample includes only those who are in the workforce. Overall, although we miss many couples in our data set and may mistakenly include some pairs who are not truly married, the couples that we identify seem fairly similar to the universe of couples in Germany in the below 65 age group and are likely to be at least roughly representative for couples where both spouses are participating in a labor market in either employment subject to social security contributions or by receiving UI benefits.

5. Conclusion

We present a method for identifying a very large number of pairs of individuals who are likely married to each other in the German administrative data. While room for type 1 (false positives) and type 2 (false negatives) errors exists, our analysis suggests that our final sample still contains about 89 to 94 percent true couples and that we have a fairly representative

sample of couples where both individuals would be covered in the Integrated Employment Biographies. The method appears accurate enough to open the door for future research projects analysing research questions in labor and public economics that rely on household (couple) identifiers using administrative data. The particular strength of this data will undoubtedly be the very large sample sizes possible with this approach as well as the possibility to link the cross-sectional information to longitudinal employment histories.

References

All-in. (2006). *Immer mehr behalten ihren Geburtsnamen - Zahl der Ehen mit Doppelnamen bleibt seit Jahren gleich*, retrieved September 1, 2014, from http://www.all-in.de/nachrichten/lokales/Immer-mehr-behalten-ihren-Geburtsnamen;art26090,215128

Goldschmidt, D., & Schmieder, J. F. (2014). You're In Then You're Out - The Incidence and Effects of Being Outsourced, *mimeo*, Boston University.

Hethey-Maier, T., & Schmieder, J. F. (2013). Does the Use of Worker Flows Improve the Analysis of Establishment Turnover? Evidence from German Administrative Data. *Journal of Applied Social Science Studies – Schmollers Jahrbuch 2013*, Vol. 133, No. 4: 477–510.

Jacobson, L. S., LaLonde, R. J., & Sullivan, D. G. (1993). Earnings losses of displaced workers. *The American Economic Review*, 685-709.

Kammergericht Berlin (2013). *Eheregistereintragung: Schreibweise von Ehenamen und Begleitnamen*, retrieved September 1, 2014, from http://www.gerichtsentscheidungen.berlin-brandenburg.de/jportal/?quelle=jlink&docid=KORE209412013&psml=sammlung.psml&max=true&bs=10

Louis, C. (2010). *Aus Liebe?: Ministerin Köhler ist zurückgetreten*. Emma, retrieved August 31, 2014, from http://www.emma.de/artikel/aus-liebe-ministerin-koehler-ist-zurueckgetreten-265036

Schild, C.-J., & Antoni, M. (2014). Linking survey data with administrative social security data - the project "Interactions between capabilities in work and private life". German Record-Linkage Center*. Working paper series,* 2014-02, Nürnberg, 11 p.

Scholz, T., Rauscher, C., Reiher, J., & Bachteler, T. (2012). Geocoding of German Administrative Data. *FDZ-Methodenreport*, 2012-09, Nürnberg, 9 p.

Sperling, F. (2012). Familiennamensrecht in Deutschland und Frankreich: eine Untersuchung der Rechtslage sowie namensrechtlicher Konflikte in grenzüberschreitenden Sachverhalten. Mohr Siebeck 2012, XX, 226 p.

Statistisches Bundesamt (2012). Bevölkerung und Erwerbstätigkeit - Haushalte und Familien Ergebnisse des Mikrozensus. Statistisches Bundesamt, Wiesbaden, Fachserie 1 Reihe 3.

TABLES

Table 1: Distribution of the Number of Individuals at the Same Coordinate

number of individuals on a coordinate	total number of individuals	number of individuals with matched names	percent matched
1	4,956,761		
2	7,443,038	5,082,600	68.29%
3	4,911,162	1,024,758	20.87%
4	3,061,944	651,742	21.29%
5	1,998,695	473,896	23.71%
6	1,589,814	396,944	24.97%
7	1,345,134	347,244	25.81%
8	1,154,712	305,390	26.45%
9	971,325	259,734	26.74%
10	807,360	219,600	27.20%
11	673,090	182,466	27.11%
12	548,928	147,280	26.83%
13	451,828	120,658	26.70%
14	366,646	96,724	26.38%
15	304,245	79,844	26.24%
16	254,032	66,272	26.09%
17	209,984	53,700	25.57%
18	177,840	45,022	25.32%
19	151,734	37,638	24.81%
20	131,940	32,064	24.30%
>20	1,540,207	372,596	24.19%
Total	33,050,419	9,996,172	30.25%

Notes: Second column includes all geocoded data as of June 30th 2008. Third column includes all individuals with geocoded location for whom we were able to match according to our name-matching algorithm, described in the text.

Table 2: Examples of the name-matching procedure

Example 2.1:

number of individuals on a coordinate	last name	potential couple
5	COHLE	match
5	HART	match
5	COHLE	match
5	MEIER	no match
5	HART	match

➔ matches HART and COHLE are chosen

Example 2.2:

number of individuals on a coordinate	last name	potential couple
5	COHLE	no match
5	HART	match
5	COHLE	no match
5	COHLE	no match
5	HART-MEIER	match

➔ match (HART-)MEIER is chosen

Example 2.3:

number of individuals on a coordinate	last name	potential couple
5	COHLE-MEIER	no match
5	MEIER	no match
5	COHLE	no match
5	COHLE	no match
5	MEIER-MUELLER	no match

➔ no match is chosen

Note: These are provided as examples only, and are not taken from the actual data.

Table 3: Gender Composition of Matched Potential Couples

matches	All matches		Age Difference < 15		Age Difference >= 15	
	absolute	percent	absolute	percent	absolute	percent
male/female	4,084,516	81.72%	3,281,657	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%

Notes: Includes all individuals with geocoded location for whom we were able to match according to our name-matching algorithm, described in the text.

Table 4: Family Status, by individual

family status	absolute	percent	accumulated
living alone	340,722	3.41%	3.41%
cohabiting	113,153	1.13%	4.54%
single parent	109,783	1.10%	5.64%
married	986,480	9.87%	15.51%
missing	8,446,034	84.49%	100.00%
Total	9,996,172	100.00%	

Notes: Includes all individuals who we were able to match by location and name (according to our name-matching algorithm). Only individuals who are registered as job-seekers have the family status variable filled in.

Table 5: Family Status Composition, for matched couples

combinations	opposite sex				same sex			
	age diff < 15		age diff >= 15		age diff < 15		age diff >= 15	
	absolute	percent	absolute	percent	absolute	percent	absolute	percent
alone-alone	5,762	0.89%	9,073	3.98%	9,854	17.65%	6,987	3.51%
alone-missing	26,692	4.11%	69,514	30.50%	28,148	50.43%	61,258	30.76%
alone-cohabit	3,124	0.48%	6,066	2.66%	2,538	4.55%	5,197	2.61%
alone-single parent	1,795	0.28%	16,050	7.04%	594	1.06%	14,573	7.32%
alone-married	9,207	1.42%	15,670	6.88%	1,391	2.49%	15,553	7.81%
cohabit-cohabit	3,248	0.50%	2,401	1.05%	1,337	2.40%	2,197	1.10%
cohabit-missing	7,001	1.08%	13,607	5.97%	4,331	7.76%	12,815	6.44%
cohabit-single parent	757	0.12%	9500	4.17%	196	0.35%	9348	4.69%
cohabit-married	5,870	0.90%	6,764	2.97%	303	0.54%	7,370	3.70%
single parent-single parent	85	0.01%	58	0.03%	219	0.39%	399	0.20%
single parent-missing	5,331	0.82%	22,240	9.76%	1,595	2.86%	21,261	10.68%
single parent-married	2,683	0.41%	1,055	0.46%	136	0.24%	1,147	0.58%
married-married	229,279	35.29%	8,078	3.54%	637	1.14%	1,111	0.56%
married-missing	348,809	53.69%	47,851	20.99%	4,536	8.13%	39,925	20.05%
Both Missing	2,632,014		574,932		129,498		529,116	
Total	3,281,657		802,859		185,313		728,257	

Notes: Includes all couples who we were able to match by location and name (according to our name-matching algorithm). Only individuals who are registered as job-seekers have the family status variable filled in.

Table 6: Stats – Individual level

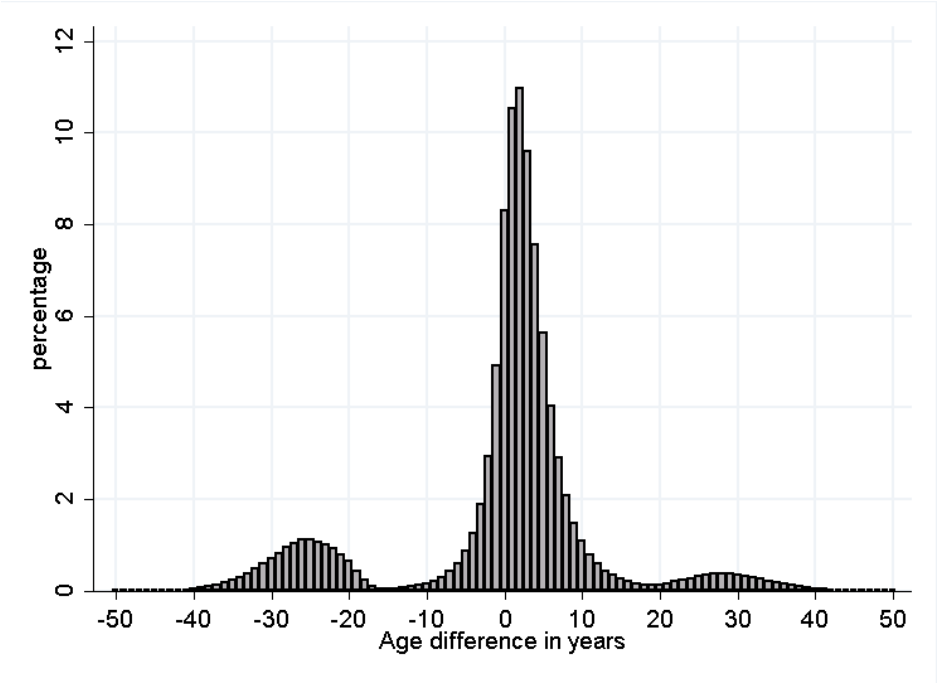
Sample	**All**		**Final Matched Sample**			**Microcensus 2011**
Restriction		**2 People at Coordinate**		**2 People at Coordinate**	**> 2 People at Coordinate**	
Number of individuals on coordinate	6.44	2.00	5.65	2.00	9.53	-
Age husband						
< 35	0.33	0.24	0.12	0.07	0.17	0.08
>= 35 and < 45	0.26	0.28	0.31	0.33	0.29	0.18
>= 45 and < 65	0.39	0.45	0.54	0.57	0.51	0.44
>= 65	0.03	0.03	0.03	0.03	0.03	0.31
Age wife						
< 35	0.31	0.23	0.17	0.11	0.23	0.12
>= 35 and < 45	0.25	0.3	0.34	0.39	0.29	0.20
>= 45 and < 65	0.42	0.44	0.47	0.48	0.47	0.45
>= 65	0.02	0.02	0.01	0.01	0.02	0.24
Labor Force Status						
in Labor Force	-	-	-	-	-	0.64
self employed	-	-	-	-	-	0.10
public servant	-	-	-	-	-	0.04
family workers	-	-	-	-	-	0.00
employee	0.84	0.90	0.88	0.93	0.83	0.47
unemployed	0.13	0.08	0.10	0.06	0.13	0.02
not in Labor Force	-	-	-	-	-	0.36
Education						
Secondary / intermediate school leaving certificate	0.78	0.80	0.82	0.81	0.83	0.71
Upper secondary school leaving certificate	0.21	0.21	0.18	0.20	0.17	0.25
Living in East Germany	0.15	0.15	0.17	0.16	0.17	-
Number of observations	33,050,419	7,443,038	6,563,314	3,384,124	3,179,190	36,016,000

Table 7: Stats – Couple level

Sample	All male/female Matches	Final Matched Sample	Final Matched Sample	Final Matched Sample	Microcensus 2011
Restriction			**2 People at Coordinate**	**> 2 People at Coordinate**	
Age difference					
no age difference	0.08	0.10	0.11	0.10	0.10
>= 1 and < 4	0.41	0.51	0.52	0.49	0.48
>= 4 and < 7	0.20	0.25	0.25	0.25	0.25
>= 7 and < 11	0.09	0.11	0.10	0.12	0.12
>= 11 and < 16	0.03	0.03	0.03	0.04	0.04
>= 16	0.19	0.00	0.00	0.00	0.02
Nationality					
both German	0.90	0.90	0.96	0.83	0.86
one German	0.06	0.07	0.03	0.10	0.07
both non-German	0.04	0.04	0.01	0.06	0.07
Monthly household income					
< 1300	0.09	0.08	0.05	0.10	0.07
>= 1300 and < 3200	0.19	0.17	0.15	0.20	0.53
>= 3200	0.72	0.75	0.80	0.70	0.34
Number of observations	4,084,516	3,281,657	1,692,062	1,589,595	18,008,000

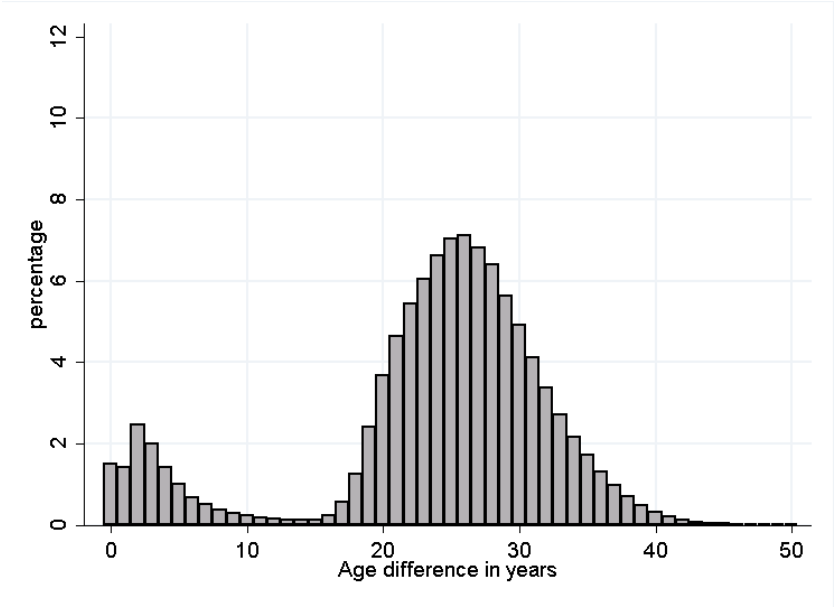
FIGURES

Figure 1: Distribution of age differences of matches, male/female



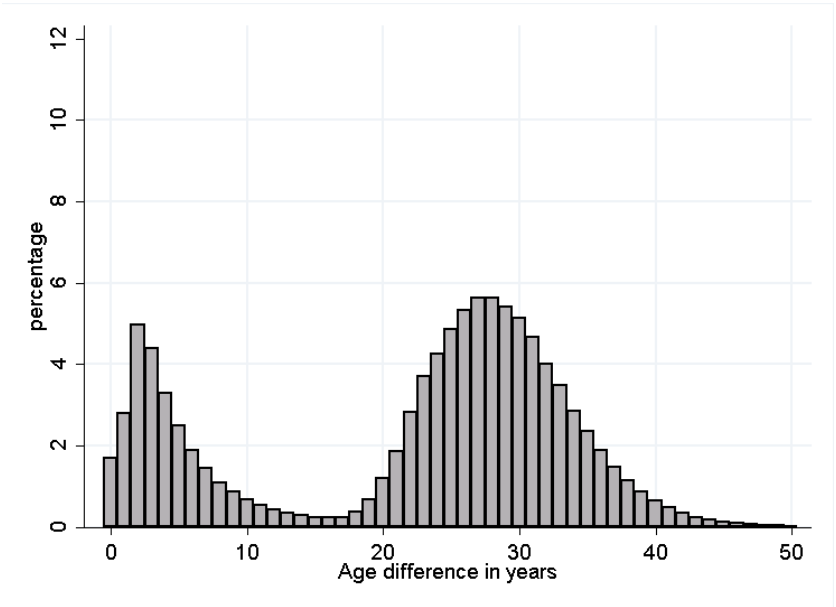
Note: Includes all male-female pairs of individuals who we were able to match by location and name (according to our name-matching algorithm). Age difference is calculated as man's age – woman's age.

Figure 2: Distribution of age differences of matches, female/female



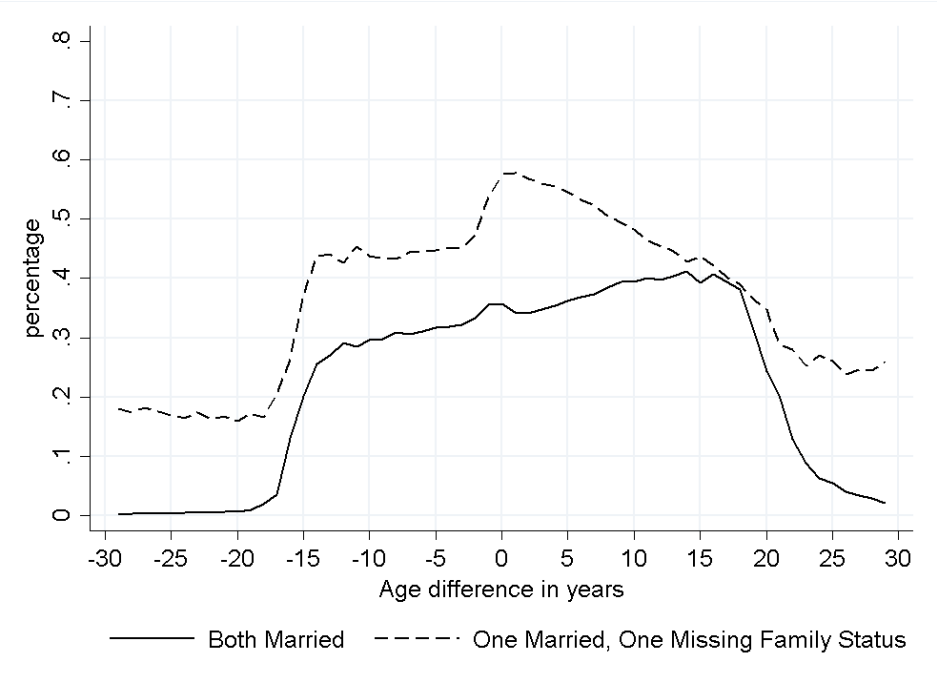
Note: Includes all female-female pairs of individuals who we were able to match by location and name (according to our name-matching algorithm). Age difference is calculated as older age – younger age.

Figure 3: Distribution of age differences of matches, male/male



Note: Includes all male-male pairs of individuals who we were able to match by location and name (according to our name-matching algorithm). Age difference is calculated as older age – younger age.

Figure 4: Share of matched pairs listed as married-married or married-missing



Note: Includes all male-female pairs of individuals who we were able to match by location and name (according to our name-matching algorithm), and where at least one member has the family status variable filled in. Age difference is calculated as man's age – wife's age.

Appendix

Table A1: Family Status Composition, for matched couples living at coordinates with exactly 2 people total

combinations	different sex				same sex			
	age diff < 15		age diff >= 15		age diff < 15		age diff >= 15	
	absolute	percent	absolute	percent	absolute	percent	absolute	percent
alone-alone	1,228	0.54%	1,504	2.08%	2,385	14.38%	1,278	2.01%
alone-missing	9,956	4.42%	30,437	41.99%	10,765	64.92%	27,634	43.44%
alone-cohabit	412	0.18%	624	0.86%	338	2.04%	542	0.85%
alone-single parent	293	0.13%	1922	2.65%	95	0.57%	1864	2.93%
alone-married	1,170	0.52%	4,106	5.67%	280	1.69%	3,840	6.04%
cohabit-cohabit	431	0.19%	226	0.31%	132	0.80%	213	0.33%
cohabit-missing	1,742	0.77%	3,622	5.00%	903	5.45%	3,537	5.56%
cohabit-single parent	98	0.04%	1103	1.52%	13	0.08%	1136	1.79%
cohabit-married	915	0.41%	1118	1.54%	39	0.24%	1122	1.76%
single parent-single parent	22	0.01%	8	0.01%	15	0.09%	40	0.06%
single parent-missing	1,595	0.71%	4,420	6.10%	339	2.04%	4,154	6.53%
single parent-married	357	0.16%	212	0.29%	11	0.07%	223	0.35%
married-married	47,922	21.25%	1,404	1.94%	77	0.46%	211	0.33%
married-missing	159,344	70.67%	21,774	30.04%	1,190	7.18%	17,816	28.01%
both missing	1,466,577		326,445		67,477		302,644	
Total	1,692,062		398,925		84,059		366,254	

Notes: Includes all couples who we were able to match by location and name (according to our name-matching algorithm), restricted to couples living at coordinates where no other

people are listed. Only individuals who are registered as job-seekers have the family status variable filled in.

Figure A1: Share of matched pairs listed as married-married or married-missing; 2 people at a coordinate

Note: Includes all male-female pairs of individuals who we were able to match by location and name (according to our name-matching algorithm), and where at least one member has the family status variable filled in. Restricted to couples living at coordinates where exactly 2 people are located. Age difference is calculated as man's age – wife's age.

Imprint

FDZ-Methodenreport 09/2014 Englisch

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, Heiner Frank

Technical production
Heiner Frank

Download
http://doku.iab.de/fdz/reporte/2014/MR_09-14_EN.pdf

Internet
http://fdz.iab.de/

Corresponding author:

Wolfram Klosterhuber

The Research Data Centre (FDZ)
of the Federal Employment Agency
at the Institute for Employment Research
Regensburger Str. 100
90478 Nuremberg

Phone: +49(0)911/179-7007
mailto: Wolfram.Klosterhuber@iab.de