German Record Linkage Center
WORKING PAPER SERIES
NO. WP-GRLC-2014-02 | JANUARY 18, 2014

Linking Survey Data with Administrative Social Security Data - the Project "Interactions Between Capabilities in Work and Private Life"

Christopher-J. Schild | Manfred Antoni

Contents

Abstract

We perform a record linkage of survey data from the project “Interactions Between Capabilities in Work and Private Life: A Study of Employees in Different Work Organizations”, a sub-project of the research group “From Heterogeneities to Inequalities” at Bielefeld University, with administrative social security data from the Institute for Employment Research (IAB) at the German Federal Employment Agency. We describe both data sets with regard to the identifiers available for the record linkage (such as names and addresses), as well as the methods and procedures that were used for this record linkage. As a final result, about 73% of the survey records could be linked, which is slightly below our average matching success rate for individual data.

Keywords: administrative data, Germany, record linkage, survey data

1 Introduction

Record linkage of individual data aims at classifying record pairs from different data files into pairs that represent the same person ("links") and pairs that do not represent the same person ("non-links") (Herzog et al., 2007). Linking survey data and administrative data offers many advantages, such as increased data reliability and smaller survey burdens to respondents. Moreover, the technical challenges have become less restrictive in recent years (Schnell, 2013).

In the project "Interactions Between Capabilities in Work and Private Life: A Study of Employees in Different Work Organizations" ("Wechselwirkungen zwischen Verwirklichungschancen im Berufs- und Privatleben")[1] 115 establishments were surveyed in the summer of 2012. From August until September of 2012, 6454 employees of these establishments were surveyed via telephone about their careers and private life goals. 5368 of these employees had a partner, of which 1888 agreed that their records be linked to administrative data of the Federal Employment Agency. Since the employees had originally been sampled from the administrative data, a unique personal linkage key was already available. Linking these respondents was therefore trivial. Regarding their partners, given the lack of a unique identifier key, it was necessary to link the individuals by using non-unique and error-prone identifiers, such as names and addresses.

2 Data sources

2.1 Survey data

The raw data from the survey to be matched (the "A-file") contained 1888 individuals (the partners of the main survey respondents). As in some cases there were several addresses provided, there were 2016 entries overall.

The raw data contained the following variables: recordno (id variable), PZ1A (last name), PZ1B (first name), PZ1C (birth name), PZ1D (day of birth), PZ1E (month of birth), PZ1F (year of birth), ORT01 (place name), PLZ01 (postal code), STR01 (street and house number).

The raw data of the A-file was of a quality typical for survey data, i.e. it contained variations (abbreviations, spelling mistakes, nicknames etc.) for names, place and street names typical for the German cultural-linguistic and geographic context. Additionally, the A-file showed an unusually large number of obvious first name - last name switches as well as obvious first name - birth name switches.

After preprocessing of the raw data (see section 3), missing value patterns were analyzed. This analysis revealed 1998 observations where both name variables, "first name" and "last name", were not missing. Of these, 33 cases had missing values of either day, month

[1] The project is a sub-project of the Collaborative Research Center (SFB) 882 research group at the University of Bielefeld, led by Prof. Martin Diewald.

or year of birth, which left 1965 observations. Of these, another 32 observations had missing values in either street name, city or postal code, leaving 1933 cases. 5 of these had missing house numbers, however this turned out to be irrelevant since B-file quality prohibited making "house number" a necessary identificator variable (see section 2.2.2).

2.2 Administrative data

2.2.1 General data base description

The administrative research data of the Institute for Employment Research (IAB) contain detailed information on the employment history of all employees liable to social security contributions and all social assistance recipients on a daily basis. There are currently more than 40 million working-age individuals in Germany, more than 80% of which have at least one record in these data. The main groups that are missing from these data are public *beamte*[2] and self-employed. These data originate from notifications of employment to social security bodies (since 1975), from data on receipt of unemployment benefit (since 1975) or unemployment benefit II (since 2005) as well as from registered job search activities and participation in active labor market policy measures of the Federal Employment Agency (BA) (both since 2000). The IAB, the research institute of the BA with limited access to these administrative data, turns these data into anonymous research data.[3]

There are usually a number of records for each individual in the administrative data, since a new record is generated whenever the individual changes employment or unemployment benefit status. Spells from different sources may also be parallel, for instance when a person is searching for a new job although he is still employed. Therefore, time-varying individual information such as addresses and last names may change from one record to the next. Individuals are identified however, by the use of a system-independent identifier that is linkable to the social security number.

The record linkage can not be conducted using the anonymous research data. Instead, we used the following identifiers from the Data Warehouse (DWH) of the BA: "first name", "last name", "sex", "birth date", "postal code" and "street name and house number".

The variation due to abbreviations, spelling mistakes, nicknames, omitted name components etc. for individuals' names, cities and streets in the B-file is considerable, in spite of presumably high standards of data entering and processing of administrative data.

[2] The German government provides two different employment schemes for civil servants: "regular employment" liable to social security and the so-called "beamte" who are not liable to social security and, therefore, are not included in the administrative data of the Federal Employment Agency. Judges, policemen, teachers, and public administration officials are usually employed as *beamte*, who are on average better educated and receive higher wages than regularly employed civil servants. In 2011, 1.7m of 4.6m civil servants were employed as *beamte*. See: http://www.bpb.de/nachschlagen/zahlen-und-fakten/soziale-situation-in-deutschland/61714/oeffentlicher-dienst

[3] A sample of these data is accessible by the general scientific community as the Sample of Integrated Labour Market Biographies (SIAB) (see vom Berge et al., 2013)

2.2.2 Preparing B-file subset suitable for record linkage

To limit the number of necessary pairwise comparisons in the record linkage, the administrative data to be linked was restricted to a subset, based on value ranges of some identifier variables in the A-file. The restriction of the B-file was based on the time of the survey as well as the geographic location (residence) and years of birth of the participants. Regarding the time period, in order to find individuals who have moved shortly before the survey, or who appear in the administrative data only in earlier years, a time window from 2004 to the year of the interview (2012) was chosen for the administrative data. Regarding geographic location, records with a postal code not present in the survey data were excluded. Regarding year of birth, records with a year of birth not present in the survey data were excluded, plus / minus a few years to allow for inexactness in birth years.

Trivially, the administrative subset was restricted to variables suitable as identifiers that were present in the A-file. Since birth names are not explicitly indicated in the administrative data, and given the available potential identifiers contained in the A-file, the B-file request to the IAB department "IT Services and Information Management"[4] was restricted to the variables “first name”, “last name”, “birth date”, “postal code”, “street name”, “house number”. Regarding the adress-related variables, it was requested that streetnames and house numbers be delivered seperately, i.e. pre-parsed by the IT department.

The raw B-file contained a total of 74.6 million entries, which included a considerable number of exact duplicate entries as well as quasi duplicate entries with respect to the identifiers “first name”, “last name”, “birth date”, “postal code”, “street name”, “house number”, “system-independent personal identifier”. The reason for these duplicates is that these administrative data are generated from several sources (see section 2.2.1). After dropping exact duplicates with respect to above mentioned identifiers, 70.3 million entries remained. Of these, the most common missing values pattern of the pre-processed B-file was “no missing values in any variable” (55 million), the second most common pattern was “only house number missing” (17 million), the third most common was “only full address missing“ (0.9 million). However, note that “non-missing” refers to whether the corresponding variable in the raw data is filled with any combination of characters and / or numbers, regardless of information content for the respective identifier.

3 Preprocessing procedures

Preprocessing of raw data is an essential step before employing similarity algorithms in record linkage (Herzog et al., 2007; Schnell et al., 2003). Preprocessing means removing spelling mistakes and known variations in correct notations (i.e. abbreviations, nicknames etc.), thus equalizing differing entries which are known to refer to the same object (“standardization”), extracting variables (such as street names and house numbers) from common text fields such as address fields (“parsing”), and eliminating implausible values (“plausibility-based elimination”, e.g. year of birth before 1900) (Herzog et al., 2007).

[4] We thank Cerstin Rauscher and Robert Jentzsch for the delivery of the raw data in the B-file.

Standardization involved steps common to all string variables (capitalizing all characters, replacing German Umlauts, removing leading and trailing blanks; see Schnell et al., 2003). For name variables, standardization also included removing name supplements such as “Junior” and “Senior” as well as academic and royal titles such as “Prof.”, “Graf” etc. For place names, common spelling mistakes, abbreviations and inconsistently used geographic name complements (such as “Frankfurt am Main” or “Berlin-Kreuzberg”) were collected and corrected. With regard to street names, the common name component “STRASSE” was standardized to its common abbreviation “STR” (“stemming”). Typical spelling mistakes of streets named after famous persons were collected and corrected. Any numbers contained in street names that could be identified as certainly being a component of the street name were spelled out (e.g., “STR DES 17. JUNI” became “STR DES SIEBZEHNTEN JUNI”), in order to increase chances of correctly parsing street names and house numbers.

Parsing mostly related to extracting street names and house numbers from address fields (A-file) and incorrectly pre-parsed street name fields and house number fields (B-file). All address related fields included a number of address information patterns not directly related to the street name or house number, such as information regarding the exact location of apartments (“5. Etage”, “Block 3”, “Apartment 10” etc.), or indications of subtenancy (“c/o Mueller”). To parse street names and house numbers from these patterns, a large set of regular expressions was employed.

The final component consisted of eliminating entries that were out of scope of a plausible value range. It was checked for implausible values for day (1-31), month (1-12) and year of birth. We also checked whether the distributions of these variables hinted at a switch of day and month, which was not the case.

4 Linkage procedures

Depending on data quality, the size of the data sets to be linked and the available identifiers, different matching methods may be optimal. The choice of the method is still largely subject to experience and thus to best practice rules. Below we briefly describe each linkage method applied. Record linkage of individual data is done in several steps, each resembling a specific matching algorithm, starting with the most strict algorithm. Cases in the A-file for which no match is found with a specific algorithm are carried over to a subsequent less strict linkage step, reducing the number of unmatched cases with each further step. An overview on the success of the subsequent matching steps is provided in table 1.

4.1 General description of employed matching techniques

With **deterministic matching**, or “exact matching”, both records have to share the exact same values for the complete set of available identifiers (Herzog et al., 2007).

Distance-based matching can be used when a record’s identifier values contain noise, such as spelling mistakes, since in these cases deterministic linkage will generate false

negatives. For comparing noisy string identifiers, string comparator algorithms are employed (for an overview see Herzog et al., 2007). Among these algorithms, **Jaro Metrics** are particularly suited to capture typical human typesetting mistakes, since they emphasize transposition of characters, i.e. switching of character positions. The **Jaro-Winkler** variant of this metric gives more weight to initial characters of strings, which can be useful if the likelihood of transpositions is lower for the first characters of a string. That is typically the case with names (Herzog et al., 2007).

Probabilistic matching is useful when similarity of some identifiers, such as the birth date, may contain more information about whether we can expect a link as the similarity of other identifiers, such as sex. Probabilistic matching incorporates information on how likely similar values are for specific identifiers within the group of matches and the group of non-matches, respectively. For linkage projects at the IAB, these likelihoods (m- and u-parameters) are chosen based on a previous frequency analysis performed with a large sample of the IAB data.

Array matching is a suitable strategy when there are several identifier variables in at least one data set, that may equal the value contained in the variable of the other data set. This is the case when, for instance, the first data set comprises the variables “last name” and “maiden name”, and the second data set comprises only “last name”, and there is no information on the marriage day or current marital status of that individual. An array match means comparing all representations of an identifier in the A-file with all representations of that identifier in the B-file, and to keep the record pair with the highest similarity value of all these comparisons.

Blocking, which is done before each distance-based or probabilistic matching step, is a very effective way of reducing calculation duration. Traditional blocking involves restricting comparisons to record pairs with exact similarities on one or more identifiers, such as postal code, which can drastically reduce the number of comparisons. Since exact blocking excludes the possibility of finding matches of individuals with erroneous or missing values in the blocking variables in one data set, this can lead to false negatives. Therefore, it is advisable to use different blocking variables in subsequent steps.

After matching quality scores were calculated, the matched data were sorted in a descending order by the matching score and then visually inspected within the presumably critical matching score value range, i.e. in the area where both matches and non-matches can be found in about equal shares. In order to calibrate an upper matching score threshold value, above which all pairs are considered as matches, the following rule-of-thumb was applied: starting from any visually found match that was adjacent to a visually found non-match (i.e. a match within the critical matching score range), further pairs were visually inspected moving upwards, i.e. higher on the matching score. Once 30 pairs were visually classified as matches consecutively, the corresponding matching score of this 30th consecutively found match was defined as the upper threshold value. The lower threshold value was determined analogously. All pairs with matching score values in between both thresholds were classified manually, i.e. by visual inspection.

Table 1: Result of Record Linkage - Overview

Type of Match	No. of Records	*cum.*	share	*cum.*
1 Deterministic / exact w. house number	651	651	34.48%	34.48%
2 Deterministic / exact w/o house number	316	967	16.74%	51.22%
3 Probabilistic w. blocking on postal code	407	1374	21.55%	72.77%
4 Probabilistic w. blocking on birth year	9	1383	0.005%	72.78%
5 No Match	505	1888	26.75%	100.0%
Total	1888			

Source: Project Survey Data, BA-Adress Data

When more than one link was found for an A-file record in the B-file, the record pair with the highest matching score was classified as a match, and the rest of the respective links was dropped.

Deterministic matching was done with Stata, for all further steps the software "Merge Toolbox (MTB)" was used.[5]

4.2 Specification of matching steps performed

In this record linkage project, we performed four matching steps.

In the first step (Step 1 in Table 1), a deterministic matching was performed. Here the set of identifiers was defined as "first name", "last name", "date of birth", "city / place", "postal code", "street", and "house number".

In the second step (Step 2 in Table 1), another deterministic matching was attempted, using the same identifiers as in the previous step, but without "house number".

The third step (Step 3 in Table 1) involved a probabilistic distance based linkage of all entries of the A-file for which no match could be found in the preceding steps. The identifier variables were "first name", "last name", "date of birth", "city / place", "postal code", "street", and "house number". Since the A-file also contained birth names, and since the A-file was characterized by switched first and last names as well as switched first and birth names, Step 3 involved an array match over all individual name-related variables. To limit computation time, Step 3 also involved blocking, where the blocking variable was the postal code (all 5 digits). Probabilistic matching was performed as described above and, regarding distance-metrics, relied on the Jaro-Winkler variant.

The fourth step (Step 4 in Table 1) was similar to Step 3, except for the blocking variable, which in this step was the year of birth.

5 Linkage result

Table 1 provides an overview on the overall success of this record linkage project. With the first deterministic steps 1 and 2, we found matches in the preprocessed B-file for 967 or about 51.22% of all 1888 records of the final preprocessed A-file. By the probabilistic steps 3 and 4, another 416 matches were found, leading to an overall matching success rate of 72.78%. This is a few percentage points below the average linkage success rate achieved in previous comparable projects with administrative, person-level data from the IAB.

We suspect that this slightly below-average success rate can be explained by specific characteristics of individuals in the A-file sample. Generally, since public *beamte* as well as self-employed are not present in the IAB data, a record linkage project with an oversampling of these groups would be expected to show lower success rates. From the design of the sampling procedure, there may be one possible theoretical indication of oversampling of individuals that are not present in the IAB data: The A-file sample consists of partners of employees drawn from the IAB data, and since (male) single-earner households are still common in Germany, one could suspect that partners of employees are more likely to be non-working (housewives) than individuals randomly drawn from the working-age population.

The survey data contains information on age, sex, education levels, paid employment, self-employment, and dummies for East German and foreign place of birth. Regrettably, the occupational information in the data set is sparse; most notably, the data do not contain information on whether individuals work as public *beamte*.

Table 2 shows frequencies for different individual characteristics available in the survey data as well as percentages of unmatched and matched survey participants for the corresponding groups. We first note that among the groups analyzed, only women are somewhat over-represented in the sample as compared to the general working-age population.[6] Therefore, we cannot conclude that there are empirical indications of over-representation of individuals typically not found in the administrative data being the reason for the relatively low matching success rate.

To further improve our understanding of the linkage process, we are also examine whether matching success can be explained by known individual characteristics of the survey participants, in a way that one should expect given that certain groups such as housewives / husbands, *beamte* and self-employed are under-represented in the IAB-file. To do so, we first compare group-specific matching success rates (see table 2), knowing that such simple group mean comparisons can only provide a first hint and do not allow for any causal interpretation.

We first note that there does not seem to be a significant correlation between gender groups and matching success. We also note that individuals with an intermediate level of educa-

[5] See Schnell (2004) for details on the MTB, which is provided freely for academic purposes at `http://record-linkage.de`.

[6] With regard to other characteristics, and in particular with regard to paid employment, the sample seems fairly representative.

Table 2: Matching Success by Individual Characteristics

	Matching Success				
	not matched		**matched**		**Total**
	Row %	95% CI	Row %	95% CI	Row %
Sex					
Men (n=726)	26.7	[23.6,30.1]	73.3	[69.9,76.4]	100.0
Women (n=1,142)	25.5	[23.0,28.1]	74.5	[71.9,77.0]	100.0
Total (n=1,868)	26.0	[24.0,28.0]	74.0	[72.0,76.0]	100.0
Pearson: Uncorrected chi2(1) =	0.3551				
Design-based F(1.00, 1867.00) =	0.3549	Pr =	0.551		
Education					
Low (n=70)	27.1	[18.0,38.7]	72.9	[61.3,82.0]	100.0
Medium (n=1,096)	23.8	[21.4,26.4]	76.2	[73.6,78.6]	100.0
High (n=702)	29.2	[26.0,32.7]	70.8	[67.3,74.0]	100.0
Total (n=1,868)	26.0	[24.0,28.0]	74.0	[72.0,76.0]	100.0
Pearson: Uncorrected chi2(2) =	6.5161				
Design-based F(2.00, 3734.00) =	3.2563	Pr =	0.039		
Employment Status					
No Empl. (n=293)	24.2	[19.7,29.5]	75.8	[70.5,80.3]	100.0
Paid Empl. (n=1,456)	25.1	[23.0,27.4]	74.9	[72.6,77.0]	100.0
Self-Employm. (n=119)	40.3	[31.9,49.4]	59.7	[50.6,68.1]	100.0
Total (n=1,868)	26.0	[24.0,28.0]	74.0	[72.0,76.0]	100.0
Pearson: Uncorrected chi2(2) =	13.7621				
Design-based F(2.00, 3734.00) =	6.8774	Pr =	0.001		
Place of Birth					
West Germany (n=1,405)	28.8	[26.5,31.3]	71.2	[68.7,73.5]	100.0
East Germany (n=311)	14.1	[10.7,18.5]	85.9	[81.5,89.3]	100.0
Foreign (n=152)	23.7	[17.6,31.1]	76.3	[68.9,82.4]	100.0
Total (n=1,868)	26.0	[24.0,28.0]	74.0	[72.0,76.0]	100.0
Pearson: Uncorrected chi2(2) =	28.9854				
Design-based F(2.00, 3734.00) =	14.4849	Pr =	0.000		
Decade of Birth					
40s (n=16)	56.3	[32.4,77.6]	43.8	[22.4,67.6]	100.0
50s (n=175)	36.6	[29.8,44.0]	63.4	[56.0,70.2]	100.0
60s (n=844)	33.3	[30.2,36.5]	66.7	[63.5,69.8]	100.0
70s (n=542)	17.9	[14.9,21.4]	82.1	[78.6,85.1]	100.0
80s (n=276)	9.8	[6.8,13.9]	90.2	[86.1,93.2]	100.0
90s (n=15)	46.7	[24.1,70.7]	53.3	[29.3,75.9]	100.0
Total (n=1,868)	26.0	[24.0,28.0]	74.0	[72.0,76.0]	100.0
Pearson: Uncorrected chi2(5) =	100.7579				
Design-based F(5.00, 9335.00) =	20.1408	Pr =	0.000		

Source: Project survey data

tion seem to be slightly over-represented in the matched group as compared to the other education categories; overall, the corresponding chi-square test discards the hypotheses of statistical independence between the education categories and matching success. The other tested group differences also show statistically significant relations between the corresponding group categorisation and the matching success rate.

Table 3: Logit Regression of Individual Characteristics on Matching Success

	Model	
Explaining Variables:		
Female	-0.237*	(-1.82)
Medium Level Educ.	0.159	(0.51)
Higher Educ.	-0.209	(-0.66)
Paid Employment	-0.101	(-0.61)
Self-Employed	-0.617**	(-2.49)
Born in 50s	0.834	(1.53)
Born in 60s	1.071**	(2.01)
Born in 70s	1.890***	(3.51)
Born in 80s	2.507***	(4.42)
Born in 90s	0.464	(0.61)
Place of Birth outside Germany	0.1000	(0.47)
Place of Birth in East Germany	0.450*	(1.88)
Female X East Germany	0.587	(1.62)
Constant	-0.200	(-0.34)
Observations	1868	

t statistics in parentheses

Reference categories: for decade of birth - 'born in 1940s', for employment status - 'no employment'.

20 cases were dropped because of missing birth year information.

Source: Project survey data

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

To analyze partial correlations between individual characteristics and the likelihood to find an individual in the administrative data, we regress linkage success on age, sex, education levels, paid employment, self-employment, as well as East German and foreign place of birth (see table 3).

This regression shows that neither paid employment nor education seem to have an effect on the likelihood of being matched. In contrast, the regression provides support for our expectation that self-employment has a negative effect on the matching likelihood.[7] Coefficients of the birth cohorts also show the expected signs. Young and old respondents are less likely to be found in the administrative data as they are less likely to be in the labor force than the intermediate age groups.

Since former East German labor market institutions may still cause women from East Germany to behave differently for example in employment participation or social assistance

[7] Note that the likelihood to find self-employed individuals is not zero. This can be explained by a) the fact that only an imperfect proxy for "self employed" was available and b) the fact that the survey data refer to the time of the survey, while the record linkage attempted to find individuals in several years of administrative data (in the case of this record linkage project 2004-2012).

take-up decisions, we include an interaction between “east” and “female”. While this interaction term is insignificant,[8] the effect of “east” is weakly positively significant, and the effect of “female” is weakly negatively significant.[9]

We conclude that, as far as the available variables allow such an analysis, the individual determinants of a successful linkage are as expected, given the specific characteristics of the administrative data.

6 Summary

We have described the record linkage for the project “Interactions Between Capabilities in Work and Private Life: A Study of Employees in Different Work Organizations”, which consisted of linking survey data from the project with administrative data from the Institute for Employment Research (IAB). We have discussed general procedures of record linkage applicable to the project, as well as challenges specific to the project.

The overall linkage success rate of roughly 73% is a few percentage points below the average of record linkages of survey data with administrative data from the IAB. Some possible explanations for the fairly low matching success were discussed and evaluated using information from the survey. A logistic regression was performed which confirmed that individual characteristics of survey participants, such as self-employment, explain the likelihood of linking these individuals in the administrative data as one should expect.

References

Herzog, Thomas N., Fritz J. Scheuren, and William E. Winkler (2007). *Data Quality and Record Linkage Techniques*. New York: Springer.

Schnell, R. (2013). *Linking Surveys and Administrative Data*. Working Paper wp-grlc-2013-03. German RLC.

Schnell, R., T. Bachteler, and S. Bender (2003). “Record Linkage Using Error-prone Strings”. In: *Proceedings of the Section on Survey Research Methods*, pp. 3713–3717.

Schnell Rainer; Bachteler, Tobias; Bender Stefan (2004). “A Toolbox for Record Linkage”. In: *Austrian Journal of Statistics* 33.1/2, pp. 125–133.

Vom Berge, Philipp, Marion König, and Stefan Seth (2013). *Sample of Integrated Labour Market Biographies (SIAB) 1975-2010*. FDZ data report 01/2013 (en). IAB, Nürnberg.

[8] However, note that we do control for “paid employment” and “self employment”.

[9] This could be due to omitted, unobserved variables. For example, we assume that, since women are more likely to be teachers than men (about 70% of teachers in Germany are women, in spite of lower female labor market participation), this result may partly capture the fact that public *beamte* cannot be found in the IAB-file. But, as mentioned above, we do not have this information for our sample.

IMPRINT

Publisher

German Record-Linkage Center
Regensburger Str. 104
D-90478 Nuremberg

Editors

Stefan Bender, Rainer Schnell

Template layout

Christine Weidmann

www.record-linkage.de