Session Number: Parallel Session 2B
Time: Monday, August 23, PM

Paper Prepared for the 31st General Conference of
The International Association for Research in Income and Wealth

St. Gallen, Switzerland, August 22-28, 2010

Combining the German Aging Survey with the Sample of the Insured Population Pension Records via Statistical Matching as a Source for the Analysis of Life Courses and Old Age Incomes

Julia Simonson, Laura Romeu Gordo, and Nadiya Titova

For additional information please contact:

Name: Julia Simonson
Affiliation: German Centre of Gerontology, DZA

Email Address: Julia.Simonson@dza.de

This paper is posted on the following website: http://www.iariw.org

Combining the German Aging Survey with the Sample of the Insured Population Pension Records via Statistical Matching as a Source for the Analysis of Life Courses and Old Age Incomes

Julia Simonson, Laura Romeu Gordo, and Nadiya Titova
German Centre of Gerontology, DZA

Paper prepared for the 31st IARIW General Conference,
St. Gallen, Switzerland, August 22-28, 2010

Abstract

In social and economic research, statistical matching is a technique increasingly applied for combining information from different data sources when a one-to-one correspondence by an identifier is not possible (see D'Orazio et al., 2001).

The basic idea of such a procedure is to find similar cases in different data bases and to link the information referring to these cases. The similarity of these cases is measured by several characteristics like age, gender or career history. Statistical matching avoids the attrition rate linked to the informed consent requirements for direct matching based on a unique identifier. It thus provides substantial opportunities for research and allows for analyses that would be impossible from one input data source alone.

The paper examines the combination of the German Aging Survey (DEAS) with administrative pension insurance data, the Sample of the Insured Population Pension Records ('Versicherungskontenstichprobe', VSKT), via statistical matching. The aims of this data matching are to provide a combined dataset that contains more information than both the stand alone data sources. The resulting data set allows us to analyze questions on changing life courses and old age provision within the framework of the project 'Life Course, Aging and Well-Being. The aims of this project are twofold: first, we want to analyze to what extent the life courses of those birth cohorts currently in the middle of their lives show increased pluralism and inhomogeneity compared to retirees today; and second we want to investigate how these changes will affect the lives of the elderly in the future - especially the protection by public and occupational pensions as well as private forms of old age security. Furthermore, we ask how the old age security systems should be structured and re-structured in order to accommodate non-traditional types of life courses.

The paper outlines the matching procedure used to combine the data as well as the required preparatory steps. Furthermore, challenges to the matching resulting from the characteristics of the data are examined as well as outcomes related to the goodness of matching and to the information we get about life courses and old age incomes from the joint data. Finally, some practical remarks on the statistical matching of survey and register data deduced from the experience with the German Aging Survey and the administrative pension insurance data are given.

Keywords: Statistical matching, old age incomes, life course research, DEAS, VSKT

1. Introduction

The goal of the present contribution is to describe statistical matching on the basis of the German Aging Survey (DEAS) and the Sample of the Insured Population Pension Records (VSKT)[1] which is administrative pension insurance data. Specifically we aim to describe the process of data preparation, the selection of matching variables, the matching procedure we use, and the corresponding matching quality tests. This paper, apart from discussing the challenges of the matching procedure, shows the potential gains of linking survey and register data.

Before entering into the methodological questions, the first question which arises is why do we want to match these two data sets, or in other words, what are the benefits of working with a VSKT-DEAS matched data set instead of working with both data sets separately.

The data matching we describe takes place within the framework of the project 'Life Course, Aging and Well-Being' (LAW), which is financed by the Volkswagen Foundation (2009-2012). This project is carried out by three cooperating institutions – the German Centre of Gerontology (DZA), German Socio-Economic Panel (SOEP), and German Federal Pension Insurance (DRV). In this project we analyze how life courses of the German baby boomers (born between 1956 and 1965) have changed in comparison to older cohorts and how these changes in life courses affect their financial situation in old age. The material situation of retirees depends to a large extent on their employment and family biographies. According to the literature, the life courses of the baby boomers are marked by increased pluralism and inhomogeneity (Leisering et al., 2001). Specifically, their employment biographies are less continuous, and unemployment, part-time employment and self-employment episodes are more common. Furthermore, their family biographies deviate more often from traditional family norms, for example in form of non marital partnerships or more single periods. All these changes in life course patterns might affect the lives and material situations of the baby boomers in the future – especially the protection by public and occupational pensions as well as private forms of old age security.

Both VSKT and DEAS are helpful but not quite sufficient for the analysis of pluralization and inhomogenization trends of life courses and their consequences for old age provision. The VSKT, which records pension insurance data, covers a large number of cases and is very informative in the recording of several episodes, but mainly includes information that is relevant for the calculation of the statutory pensions e.g. periods of employment and related earnings on a monthly basis. However, the statutory pension is only one source of income in old age. There are other resources like private old age provision, occupational pension schemes, wealth, savings, and inheritances. So, the pension data is helpful for analyses on specific working biographies, work related income, and pension entitlements from the statutory pension system, but it does not

[1] An equivalently used name for 'Sample of the Insured Population Pension Records' is 'Sample of Active Pension Accounts'.

contain all the relevant information which helps us to have a complete picture of the potential old age financial situation.

This missing information in the VSKT can be complemented by the German Aging Survey (DEAS).[2] The DEAS is a nationwide representative cross-sectional and longitudinal survey of the German population aged 40 or older, funded by the Federal Ministry for Family Affairs, Senior Citizens, Women and Youth (BMFSFJ). The data includes comprehensive information on the living situation in old age as well as on their subjective evaluation and enables us to place the analysis of the income situation in old age within a broader context (Simonson et al., 2010). It does not give such an exhaustive and complete picture of the employment career as the VSKT, but it does inform comprehensively about transitions to retirement, different sources of old age incomes, private and occupational pension schemes, inheritances, and monetary and nonmonetary transfers within the family. Furthermore, the DEAS offers information required to analyze the future situation of the baby boomers like retirement plans and attitudes towards old age provision.

Summarizing, while both data sets on their own do not offer a complete picture of the lives of the baby boomers and the reference cohorts, the matching of VSKT and DEAS allows complementing the information on public pension entitlements with other relevant determinants of future life arrangements. With the resulting data set it is possible to analyze whether employment and family biographies of the baby boomers and their general financial situation and attitudes have changed in relation to the reference cohorts and the consequences of these changes for their future lives.

2. Data

2.1. The Sample of the Insured Population Pension Records (VSKT)

The Sample of the Insured Population Pension Records (VSKT) is a one percent random sample of the insurance accounts of the statutory pension agencies. In Germany, the Statutory Pension Insurance is mandatory for all employed persons in the private and public sector. Hence, the pension data of the German Federal Pension Insurance covers more than 90 percent of the German population (Himmelreicher and Stegmann, 2008). For each compulsory member of the

[2] Needless to say that the German Aging Survey (DEAS) is not the only data source that can be usefully combined with the pension data for the analysis of life courses and old age incomes. The German Socio-Economic Panel Study (GSOEP) also contains comprehensive information on life courses, different sources of incomes and living situations. This also makes a statistical matching of the GSOEP and VSKT beneficial for analyzing life courses and old age income situations. Therefore, within the project 'Life Course, Aging and Well-Being' (LAW) GSOEP and VSKT are also statistically matched. The parallel matching of two surveys with the pension data is due to the fact that GSOEP and DEAS both have their advantages and drawbacks. So depending on the specific research question, using either the matched DEAS-VSKT or GSOEP-VSKT data might be more appropriate.

statutory pension insurance an account is kept where all contribution periods and relevant non-contributory periods (like child-raising phases) until retirement are registered (Himmelreicher and Stegmann, 2008).

The VSKT is a stratified random sample from these accounts. It contains information on all relevant registered creditable periods and pension entitlement of insured persons in the German Statutory Pension Insurance aged between 15 and 67 years. Information is given on relevant registered creditable periods. Reporting on insurance accounts is done in a kind of random sample and is followed in time offering a panel structure. According to Himmelreicher and Stegmann (2008, 651), insured persons in the sense of the German statutory pension insurance are all persons who are registered in a German pension insurance account which

- is not closed at the time of the evaluation;
- contains contribution periods up to the census deadline 31.12. of the reference year;
- has no notification of the decease of the insured person up to the census deadline (31.12. of the reference year);
- relates to a person of a minimum age of 15 and a maximum age of 67.

It has to be taken into account that although the data represents the entirety of insured persons, it is not representative of the German population. Only persons eligible for old age or disability pension benefits can be in the sample. This means people who have been employed their whole employment history as civil servants ('Beamte') or as self-employed are not included. In addition, some other occupation groups like lawyers or doctors are not insured in the statutory pension insurance and therefore not part of the VSKT.[3]

Moreover, it should be remembered that the data quality of the accounts is quite different and depends on the status of the clarification of an account ('Kontenklärung'). The individual accounts often have gaps. To fill these gaps the pension insurance contacts the insured persons with the request to supply information required for closing the gaps. This is done at regular intervals, but people answer these requests with varying effort, which leads to a different degree of completeness of the accounts.

The VSKT contains both individual-related and life course-related information. The life course related information is given on a monthly basis, which means that for every person the information for 624 biography months is provided.

For the present statistical matching procedure a special VSKT data set, the 'VSKT-LAW-Sample', has been prepared, which will be described in the following.

---

[3] For more information on subgroups that are excluded from the pension data see Himmelreicher and Stegmann, 2008. Further information on the VSKT Scientific Use Files (SUFs) is available on the internet (http://forschung.deutsche-rentenversicherung.de/FdzPortalWeb/).

The VSKT-LAW-Sample

For the purpose of the research project 'Life Courses, Aging and Well-Being' (LAW), the Research Data Centre of the German Federal Pension Insurance (FDZ-RV) provided a special VSKT data set. This VSKT-LAW-Sample contains information from VSKT data sets from three different years: 2007, 2005, and 2002. This is because in the project we want to analyze the life courses of three birth cohorts: 1936-45 (cohort 1), 1946-55 (cohort 2), and 1956-65 (baby boomers; cohort 3). Due to the fact that the VSKT covers information on the individual life course between the age of 15 and 67, it is not possible to get full information on the life courses of the three cohorts from one VSKT: the 2007 data does not include the birth years 1936-1939. The 2005 data does include the birth years 1938-39, but not the years 1936-37. The 2002 data does include all relevant birth years but the information on the accounts ends in 2002, which means we would not know what has happened afterwards if we used only this data source. So, to get the most comprehensive information possible, the 2007 data was enhanced with information on the life courses of those born 1936-37 from the 2002 data and of those born 1938-39 from the 2005 data.

The VSKT-LAW-Sample only covers accounts cleared with the aid of the insured persons as well as accounts cleared without the aid of the insured persons and without gaps since the last clarification. In total the VSKT-LAW-Sample covers 205,828 accounts from Germans and non Germans. However, for the matching we only use the accounts of German citizens as described later.

2.2. The German Aging Survey (DEAS)

The German Aging Survey (DEAS) is a nationwide representative cross-sectional and longitudinal survey of the German population aged 40 or older, which is funded by the Federal Ministry for Family Affairs, Senior Citizens, Women and Youth (BMFSFJ). The comprehensive examination of people in mid- and older adulthood provides micro data for use both in social and behavioral scientific research and in reporting on social developments (Motel-Klingebiel et al., 2009). The data thus provides a source of information for decision-makers, the general public and for scientific research.

The first DEAS survey wave took place in 1996, the second wave followed in 2002. The third wave of DEAS was conducted in 2008. Participants are questioned in detail on their living situation. Particular issues addressed in the survey include an assessment of occupational status or living conditions after retirement, social participation and leisure activities, information on their economic and housing situation, family ties and other social contacts, as well as issues regarding health, well-being and life-goals. Data are conducted via a face-to-face interview and a self-administered questionnaire, respondents are asked to fill in after the face-to-face interview (drop-off questionnaire). There are both individuals who are asked for the first time (base samples) and individuals who were asked in past waves and take part again in the survey (panel samples).

The third wave differentiates between three subsamples: (1) Persons who took part in the survey in 1996 and 2002, (2) persons assessed in 2002, and (3) a new group of 6,205 participants included in the study for the first time. This approach enables a comprehensive description of life situations and life contexts of the German population aged over 40 in the year 2008 (current cross-sectional analysis), an analysis of social changes over the points in time 1996, 2002 and 2008, and an investigation of intra-individual development over either six or twelve years (2002-2008, or 1996-2002-2008). Another perspective results from the comparison of individual development over a six year period in the two time-frames 1996 to 2002 or 2002 to 2008, i.e. a comparison between the development of two birth cohorts in a specific age segment.[4]

For the present statistical matching, the first time respondents (n=6,205) of the third wave of the DEAS are taken into account. The DEAS 2008 sample is a stratified random sample of the population aged 40 years or older living in Germany. That means it includes both Germans and non Germans. However, in the present matching only a subsample of German citizens born between 1936 and 1965 is included as pointed out later.

3. Statistical matching

In empirical research, statistical matching is a technique increasingly applied for combining information from different data sources when one-to-one correspondence by an identifier is not possible (D'Orazio et al., 2001; Kum and Masterson, 2008). This can be due to confidentiality restrictions or attempts to avoid the attrition rate linked to the informed consent requirements for direct matching ('record linkage') based on a unique identifier.[5]

While originally predominantly used in medical and evaluation studies to analyze treatment effects by avoiding selection bias problems (Heckman, 1990; Zhao, 2004), statistical matching is now increasingly common in social and economic fields too. Here, the aim of statistical matching is often not to get similar groups of treated and not-treated persons for an evaluation of treatment effects, but to combine the information from different data sources. For example, Rasner et al. (2007) considered statistical matching to combine the Scientific Use File Completed Insurance Biographies (SUF VVL) with the German Socio-Economic Panel Study (GSOEP). Frick and Grabka (2010) used statistically matched data from the German Socio-Economic Panel Study (GSOEP), the Sample of the Insured Population Pension Records (VSKT) and the statistics on pension benefits splitting after divorce ('Versorgungsausgleichstatistik') for analyzing wealth inequality and the importance of public pension entitlements. Steiner and Geyer (2009) also statistically matched the data from the German Socio-Economic Panel Study (GSOEP) with

[4] Further information on DEAS as well as the data collection instruments of the three waves can be downloaded from the internet (www.deutscher-alterssurvey.de – www.fdz-deas.de).

[5] In Germany the record linkage is not permitted without the explicit informed consent of the respondents.

the Sample of the Insured Population Pension Records (VSKT) in order to get a profound basis for the analysis of old age incomes.

Unlike record linkage, statistical matching does not aim to find the same person in two data sets, but links the information from observations that are similar, at least in certain characteristics ('matching variables') observed in both data sets (Frick and Grabka, 2010). This means, statistical matching uses variables common to both data sets to identify similar records that can be linked in order to get a combined data set, which allows for analyses that would be impossible from one input data source alone.

The most common method of statistical matching is the propensity score matching (Rosenbaum and Rubin, 1983). The propensity score is a function of several variables which one expects to affect the probability of belonging to a treatment group. This means the propensity score provides the probability of participating in a treatment. For matching in order to analyze treatment effects, the use of propensity scores seems to be quite appropriate. But for matching in order to combine different data, it is suboptimal, because of the absence of any treatment variable. Thus we do not use the widespread propensity matching, but a Mahalanobis distance based matching procedure (Gu and Rosenbaum, 1993; Rubin, 1980), which directly calculates the distances between several cases in regard to relevant variables without making the detour via a propensity of treatment.

3.1. Preparatory steps in matching the data sets

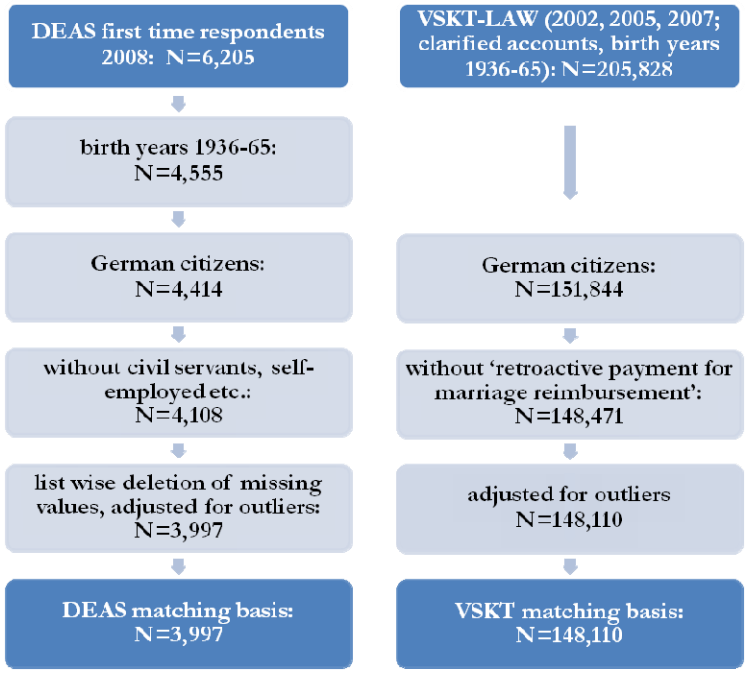
A statistical matching requires some preparatory steps like defining the sample population, finding matching variables and comparing distributions, which are described in the next sections.

Defining the Sample Population

For a successful statistical matching, it is useful to have data sources that represent the same population in some degree, especially in some key dimensions like age or gender. If the populations of the underlying data sources differ strongly, this could lessen the reliability of the results. Therefore, the sample population has to be specified correctly (Rasner et al., 2007).

An overview on our data selection processes and the resulting numbers of cases within our two data bases is given in Figure 1. The population of interest for the present matching procedure consists of men and women who belong to the three birth cohorts 1936-1945, 1946-1955, and 1956-1965, regardless of their present age. Because of the relatively low number of foreign respondents in the German Aging Survey and the lower rate of completeness of the foreigners' accounts in the VSKT due to the temporary residence of many foreigners, for the present matching only accounts from individuals with a German nationality are included. Although a separate matching of foreigners is intended later, it is not part of the described matching. Further restrictions regarding the population come from the characteristics of our two data sources as we will see in the following.

Figure 1: Overview of data selection processes



Source: own illustration

In total the VSKT-LAW-Sample that we use as one starting basis for our matching covers 205,828 accounts of both Germans (N=151,844) and foreigners (N=53,984) who are employed with compulsory insurance in Germany or insured in the German Pension Insurance for another reason. However, for the present matching, only accounts from individuals with a German nationality are included (N=151,844). We also exclude the data of those women (N=3,373) who made use of a 'retroactive payment for marriage reimbursement' ('Nachzahlung bei Heiratserstattung') due to the lower validity of their accounts.[6]

[6] Until 1967 West German women could make use of a so-called 'marriage reimbursement', i.e. after marriage they could receive a payout of the contributions made until then. Their pension accounts were eliminated, but if they wanted, e.g. in case of later being employed anew, the women could pay their reimbursement back to the pension insurance ('retroactive payment') and open a new pension account. If so, the earning points for the repaid contributions were put in the beginning of the new account without information on their sources (e.g. education or employment times). This implies that the VSKT does not give valid information about the beginning of the employment biographies of these women. Since the termination of this regulation in 1967, women up to the birth year 1949 were involved. Women who made use of a 'retroactive payment for marriage reimbursement' can be identified in the data. Women, who made use of the 'marriage reimbursement' and did not pay back the reimbursement later, can also be part of the VSKT, e.g. if they were employed again. For these women a new account with blank beginning was opened, so that for them the duration of employment (and the resulting income) in the VSKT is underestimated. However, we have no information on which and how many women are involved.

Finally, we drop some outliers relating to retirement age from the data (N=361; 0.2%).[7] After this, the resulting data set contains 148,110 insurance accounts from individuals born 1936-65, whereby most of the accounts (87 percent) come from the VSKT 2007, which is plausible, considering that all accounts of individuals born between 1940 and 1965 come from the VSKT 2007, and only the accounts of insured persons born between 1936 and 1939 stem from the VSKT 2002 (7 percent) or 2005 (6 percent).

Our second starting basis for the matching is the third DEAS wave, in which 6,205 respondents were interviewed for the first time. The DEAS 2008 sample includes both German and non Germans aged 40 or older. However, in the present matching only the three birth cohorts 1936-1945, 1946-1955, and 1956-1965 (N=4,555) and from those only the subsample of German citizens (n=4,414) are included.

To make the populations of DEAS and VSKT comparable, it should be borne in mind that in the pension data (VSKT) only people who are registered with an insurance account in the German Statutory Pension Insurance are included. People who have been employed their whole employment history as civil servants ('Beamte') or has been self-employed the whole time are not included. In addition, some other occupation groups (independent professions, e.g. lawyers or doctors) are not insured in the statutory pension insurance and therefore not part of the VSKT.

In the DEAS the population aged 40 or older is covered, independent of their current or previous occupation. Therefore persons in the DEAS who belong to groups with a high probability of not having a counterpart in the VSKT have to be excluded from the DEAS sample before matching the data. The difficulty is how to decide which persons in the DEAS have to be excluded, since in the DEAS we do not have information on the occupational positions during the whole employment career, but only for the beginning and the end of the career (for retirees or unemployed persons) or for the beginning and the current time point (for employed persons). So, we cannot define with absolute certainty who has been a civil servant or self-employed for most of his or her working life. Since the probability is high that this was their predominant status if it was their status both at the beginning and the end of their career, all those occupied as civil servants, as self-employed or with an independent profession both initally and currently or in their last job respectively are excluded from the matching. This applies to 306 persons from the used subsample of German citizens born between 1936 and 1965 (n=4,414), which left a data base of 4,108 cases. We exclude persons with outlying values in the retirement age (N=14; 0.4%) from the DEAS as well as for the VSKT. As the matching procedure we use excludes cases with missing values in one or more of the matching variables from the calculation of the Mahalanobis distance and therefore also from the matching, the sample size of DEAS we use for the matching

[7] For plausibility reasons we decided not to include data for persons with a transition age to old age retirement of below 40 years.

is additionally reduced for 97 cases. That means the resulting DEAS matching source contains a total of 3,997 cases.

Table 1: Description of the matching bases

Variable	DEAS		VSKT	
	N	%	N	%
gender				
male	1,969	49.26	64,940	43.85
female	2,028	50.74	83,170	56.15
region				
East	1,502	37.58	27,385	18.49
West	2,495	62.42	120,725	81.51
birth cohort				
1936-1945	1,469	36.75	47,498	32.07
1946-1955	1,247	31.20	46,778	31.58
1956-1965	1,281	32.05	53,834	36.35
retirement status				
retired	1,411	35.30	39,350	26.57
not retired	2,586	64.70	108,760	73.43
total	3,997	100.00	148,110	100.00

Source: DEAS 2008 and VSKT-LAW 2002-2005-2007, own calculations

Although both matching sources now basically refer to the same group (Germans of the birth years 1936-1965 who are insured in the public pension scheme), the distributions of relevant socio-demographic variables in the two data sets still differ, as we can see in Table 1. While in the DEAS the proportions of men and women are almost balanced, in the VSKT we have a quite clear predominance of women. Furthermore, persons allocated to East Germany are much more overrepresented in the DEAS data, which is a result of the disproportional sampling of the DEAS. Persons belonging at least to the cohort 1936-1945 are overrepresented in the DEAS, while in the pension data the distribution of birth cohorts reflects the factual predominance of the baby boomers (1956-1965). According to the differences in the cohort and age distributions, we also have a difference in the proportion of retirees between both data sources. While in the DEAS the share of retirees is 35 percent, in our VSKT data base it is only about 27 percent. Overall, there are some differences in the distributions of socio-demographic characteristics, but they do not seem grave enough to seriously endanger the success of matching.

Matching Variables

For the statistical matching to be successful, the datasets require to share a set of variables measured in comparable ways (Himmelreicher & Schröder 2010; Rasner et al., 2007). Since we want to create a data basis for the analysis of life courses, employment histories and old age

incomes, the goal of our matching procedure is to combine the data of people who are similar in regard to these fields. So it is reasonable to focus on matching variables which give us substantial information about working biographies.

Since the purposes and collecting modes of both data sources are quite different, our matching has to base on a limited set of core variables (for an overview see Table 2). In both datasets we have some information on the employment biography and on socio-demographic characteristics like age, gender, and region (West or East Germany). For women we can also use the information on the number of children. However, this information is not available for men in the VSKT, because childcare periods normally are credited in the women's accounts only.

We generally also have information on education levels, but unfortunately in our VSKT data source this variable has a substantial proportion of missing values (56 percent), because it is not necessary for calculating pension entitlements. Therefore, we do not use educational level as matching variable. Income too would be a helpful variable, but while in the VSKT we have the longitudinal information on earning points, with which we could at least roughly recalculate the income over the life cycle,[8] in the DEAS income is only available for the time point of data collection. So we decided not to include income for the matching.

Employment duration

To provide for employment times, we include the total duration of employment, subtracting the year a person was employed for the first time (begin of employment) from the year a person stopped working (end of employment). In the DEAS, begin of employment is measured by a direct question regarding the year of first regular employment. In the same way, end of employment is collected via a direct question in regard to the time employment ended. For individuals still in employment this question obviously is not applicable and therefore not asked. For this group we considered the year of the survey (2008) as the ending point. In the VSKT, the beginning and end of employment can be taken from the longitudinal information on the accounts. If the effective status at the time point of data collection is employment, equivalently to DEAS we take the collection year of the VSKT (2002, 2005 or 2007) as the ending point. Although there is a high probability that people in reality will work longer, it would seem appropriate to take the last observation time as the ending point, because calculating the employment duration gives their effective employment time up till then.

[8] Earning points mirror the income situation during the employment history of an insured person. One earning point per year is given if an individual earns neither more nor less than the average income of all insured persons in this year (FDZ-RV 2008). Additional earnings points are given for several contribution periods.

Table 2: Overview of matching variables

Variable	Format	Function
duration of employment	number of years	matching variable
existence of employment gaps	yes/no	matching variable
duration of employment gaps	number of years	matching variable
employment gaps: parental leave	yes/no	matching variable
employment gaps: military/ civilian service	yes/no	matching variable
employment gaps: studies/ further education	yes/no	matching variable
employment gaps: unemployment	yes/no	matching variable
employment gaps: sickness/ rehabilitation	yes/no	matching variable
retirement age	age in years	matching variable
unemployment duration (for unemployed)	number of years	matching variable
invalidity pension duration (for invalidity pensioners)	number of years	matching variable
children (for women)	number of children	matching variable
year of birth	calendar year	matching variable
birth cohort	1936-45/ 1946-55/ 1956-65	slicing variable
gender	male/female	slicing variable
region	East/West	slicing variable
retirement status	retired/not retired	slicing variable

Source: own illustration

Employment gaps

Both data sources offer information on employment gaps or times where employment is interrupted. In DEAS respondents are asked if they ever have disrupted their employment for more than 6 months. If so, they are asked to give the information for how long they have interrupted employment at all and for what reasons. Given the available information on times of non-employment in the VSKT accounts we incorporate the information if individuals have employment gaps due to five different reasons: (1) parental leaves, (2) military/ civilian service, (3) studies/ further education, (4) unemployment, and (5) sickness/ rehabilitation. Whereas in the pension data we also have information on the duration of the various gaps, in DEAS we have not, so for the matching we can only use the information whether such an interruption took place or not, but not how long it lasted. What we have from both data sources is the information on the duration of the employment gaps in total. We also consider this information for the matching.

Retirement age and status of retirement

For old age retirees, we additionally take the age at the beginning of retirement into account. This information we can get from both DEAS and VSKT. For people not yet retired, the information is substituted by their regular pension age (65, 66 or 67, depending on their year of birth),

regardless of their effective retirement age in the future.[9] We also take the retirement status into account. This information (already in old age retirement vs. not in old age retirement) we use as a slicing variable. Using slicing variables means the matching will be done within slices, defined by the chosen slicing variable. Therefore, a pensioner can only have another pensioner as matching partner, while the matching partner for a person not yet retired also has to be not yet retired.

Unemployment duration (for unemployed) and invalidity pension duration (for invalidity pensioners)

Additionally, for individuals who are currently unemployed or receive an invalidity pension, the duration of the respective status to date is calculated and taken into account. For people currently not in any of these statuses, the duration is set to zero.

Number of children (for women)

We also take the number of children into account. Unfortunately in the VSKT this information is only available for women. So for all men in DEAS and VSKT the value for children is set to zero. While in the pension data the maximum number of registered children is limited to ten, in DEAS the number of children on whom information can be given is unlimited. The empirical maximum of children in the DEAS data is twelve (very seldom) and for a better comparability all numbers of children above ten are lowered to ten.

Year of birth, cohort, gender, and region

We take at least socio-demographic information on the year of birth, cohort, gender and region into account for the matching. Year of birth is incorporated in the matching procedure as calendar year, comparably provided by both data sets. However, the incorporation of birth year as a matching variable does not necessarily mean that only people with the same birth year can be chosen as matching partners, but also people with similar though different birth years. One of our aims in the matching is to provide a data set for cohort analyses. To avoid people from one cohort getting a matching partner from another cohort, that would obviously weaken the validity of our cohort analyses, we additionally take the birth cohorts into account and use them as a slicing variable, which implies that the matching will be done within the three cohorts, meaning that a member of cohort 1 can only have a matching partner also belonging to cohort 1, and so on.

Because life courses still differ considerably between men and women, we also use gender as a slicing variable, which means data from a woman can only be matched to data from another woman (same for men). Since the different histories of the former German States FRG and GDR and the persisting differences in living conditions, life courses, pension calculations and benefits in West and East Germany, we additionally use the regional information (East or West

[9] However, although for the matching we use these approximations for those not already retired, in the tables and figures regarding retirement age, we only include those who are already retired and for whom we have an actual value.

Germany) as a slicing variable. In both data sets we have this regional information. However, the logics of this information differ: In the German Aging Survey (DEAS) we use the statement were a participant mainly lived during the division of Germany (1949-1990): in East or West Germany. For participants who mainly lived outside Germany during this period we use the information on the current region of residence (East or West). In the pension data East-West affiliation is composed via the percentage of earning points from East or West Germany. If more than half of the earning points come from East Germany, a person is assigned to the East, otherwise to the West. That means that while in the DEAS we apply a residence concept of region, in the VSKT it is an earning based concept. However, although these concepts clearly differ, because of the relatively low levels of labor mobility and job related commuting, the overlap between both concepts should be considerably high.

Since the measurement of the variables differs to some extent between the data, it is necessary to verify if variables in both data sets really measure the same. One can thus compare the distributions of the selected matching variables to get an impression of the suitability of the matching variables and the samples to match. In Table 3 we see that means and proportions of some of the variables accord quite well, but others do not.

Table 3: Distributions of the matching variables in the data bases

Variable	**DEAS**			**VSKT**		
	mean / %	sd	N	mean / %	sd	N
duration of employment	32.70	10.50	3,997	28,69	12.46	148,110
existence of employment gaps	0.42	0.49	3,997	0.75	0.43	148,110
duration of employment gaps (for persons with employment gaps)	5.00	5.46	1,667	7.86	7.59	110,851
employment gaps: parental leave	0.23	0.42	3,997	0.29	0.45	148.110
employment gaps: military/ civilian service	0.08	0.28	3,997	0.14	0.35	148.110
employment gaps: studies/ further education	0.05	0.23	3,997	0.13	0.33	148.110
employment gaps: unemployment	0.07	0.26	3,997	0.24	0.42	148.110
employment gaps: sickness/ rehabilitation	0.02	0.15	3,997	0.06	0.24	148.110
retirement age (for retirees)	61.39	3.17	1,411	61.71	3.73	39.350
unemployment duration (for unemployed persons)	5.14	5.14	220	1.98	2.62	9,382
invalidity pension duration (for invalidity pensioners)	7.61	7.39	135	5.81	5.77	4,802
children (for women)	1.88	1.18	2,028	1.71	1.31	83,170
year of birth	1949.78	8.87	3,997	1951.02	8.82	148.110

Source: DEAS 2008 and VSKT-LAW 2002-2005-2007, own calculations

We can observe a moderate difference for the mean employment durations. However, this is not completely unexpected, considering the larger proportions of women and younger birth cohorts in the VSKT. The durations of employment gaps in total are relatively different (about 5 years in the DEAS and 8 years in the VSKT). Moreover, one has to consider that the proportion of people for whom we measured an employment gap differs strongly between both data sets. While 42 percent of the DEAS respondents have an employment interruption of more than half a year, in the VSKT the proportion is about 75 percent. This difference could be at least partly an effect of the mode of data collection. While the VSKT is process produced data, in the DEAS respondents are asked retrospectively if they ever have interrupted their employment for a period of more than six months. As is known from the literature (e.g. Janson, 1990; Middendorff, 2000), retrospective statements are prone to recollecting errors. Recollecting tends to decrease with the time span involved, whereby the pace of this decrease differs strongly among types of circumstances to be recollected (Janson, 1990, 104). Thus, it might be that employment gaps are sometimes misremembered, especially if they happened a long time ago. Moreover, interruptions in employment might be perceived as socially undesirable and therefore not reported correctly in all cases. This thought is supported by the fact that most of the discrepancy comes from unemployment episodes, which many people do not perceive as socially desirable, and which might be underreported for that reason. We have differences for the other types of gaps too, whereby for all types the proportions are higher in the pension data. For individuals currently unemployed or invalidity pensioners, the present duration of these statuses is higher in the DEAS data.

There is also a difference between the birth years, which stems from the fact that older cohorts are slightly overrepresented in the DEAS, as pointed out before. We have a very good correspondence for the mean retirement age, which in both data sources is between 61 and 62 years (for those who are already retired). Also the mean number of children per woman in both data sets is quite similar, though in the DEAS it is slightly higher. This could be an effect of the higher proportion of older birth cohorts with higher fertility rates in the DEAS (cf. Table 1).

Summarizing, we have some variables whose distributions are quite similar, but there are other variables whose distributions are not. This might be problematic for the goodness of the matching results. If we, for instance, underestimate the proportion of people who had an unemployment episode in one data source, this could imply that in the matching procedure we will combine persons who wrongly report no unemployment episode with persons who really have no employment episode. However, we do not know for certain what the reasons for the lower proportions of unemployment and other employment interruptions in the DEAS are. Therefore it is not possible to decide if employment interruptions really are underestimated or if there is another reason for the differences. Thus, we will take all of the described variables into account for our matching, considering the differences of distributions in the interpretation of the results.

3.2. Matching on basis of the Mahalanobis distance

For the statistical matching we use a matching procedure based on the Mahalanobis distance (Gu and Rosenbaum, 1993; Rubin, 1980), which is relatively robust under different settings (Zhao, 2004).[10] The Mahalanobis distance is a distance measure introduced by Mahalanobis (1936) and frequently used in cluster analysis. It is based on the correlations between variables and differs from the Euclidean distance in that it is scale-invariant, i.e. independent of the scale of measurements, and takes correlations between variables into account, which implies that highly correlated matching variables do not enter the computation of the distance function with the same weight as low correlated variables.

Given the representativeness of the (weighted) DEAS sample for the German population in the relevant age, we match the VSKT information to DEAS data, i.e. the DEAS data provides the recipient file. For each observation x_i in the DEAS, the statistical software measures the Mahalanobis distance d_{ij} to each observation x_j in the VSKT on the basis of the chosen matching variables p. The VSKT observation with the smallest distance is chosen as statistical matching partner or donor.

For the matching we slice the data by region, gender, cohort, and retirement status, taking into account regional, cohort and gender differences in life courses and pension entitlements as well as the different situation of retirees and non-retirees. This means matching can only take place within the thus defined groups: A retired East German woman of cohort 1 from the DEAS can only be matched to a retired East German woman of cohort 1 from the VSKT, a non-retired West German man of cohort 2 only to a non-retired West German man of cohort 2, etc. Taking our four chosen slicing variables into account, arithmetically we get 24 slices (3 cohorts * 2 regions * 2 genders * 2 retirement states). However, we do not have any retirees within the third cohort. Therefore the matching is done within 20 slices. As matching variables we use the variables described in section 3.1. All metric variables are standardized before entering the equation of the distance function. Categorical information is recoded to dummy variables. Calculations for the matching are done for all 3,997 persons from the DEAS, and for each of these 3,997 observations in the DEAS an observation from the VSKT is chosen as statistical matching partner or donor.

4. Results of the matching

4.1. Description of the combined data

The result of the matching is a data set that includes information from both sources: VSKT and DEAS. The new data set covers 3,997 cases, which are all cases of our DEAS data base. Due to the use of the slicing variables gender, region, cohort, and retirement status the distribution of these characteristics in the resulting data is the same, regardless of whether we consider the

[10] For the Mahalanobis matching we use the MAHAPICK procedure in Stata (Kantor, 2006).

information from DEAS or VSKT, as we can see in Table 4. Due to the adjustment of the VSKT at the DEAS sample, the new distribution differs substantially from the original VSKT distribution.

Table 4: Description of the resulting data set

Variable	information from DEAS		information from VSKT	
	N	%	N	%
gender				
male	1,969	49.26	1,969	49.26
female	2,028	50.74	2,028	50.74
region				
East	1,502	37.58	1,502	37.58
West	2,495	62.42	2,495	62.42
birth cohort				
1936-1945	1,469	36.75	1,469	36.75
1946-1955	1,247	31.20	1,247	31.20
1956-1965	1,281	32.05	1,281	32.05
retirement status				
retired	1,411	35.30	1,411	35.30
not retired	2,586	64.70	2,586	64.70
total	3,997	100.00	3,997	100.00

Source: DEAS 2008 and VSKT-LAW 2002-2005-2007 (combined data), own calculations

Table 5 shows the means and proportion of the matching variables as they are in the resulting data set. As we can see the distributions are now much more similar than in the unmatched data. The difference in the duration of employment has dropped from 4 percentage points to 2.5, and the proportions of people with employment gaps is now in accordance: from both sources we get the information that 42 percent had an interruption of employment. However, the difference in the duration of the employment gaps is now greater than before. The proportions of the different types of gaps from the VSKT are now aligned with the DEAS information as well as the number of children, and birth year. For retirement age, invalidity pension duration, and unemployment duration we still have some differences.

What we can observe is that for the dichotomous variables in general we get a better concordance than for the metric variables. This is plausible, considering the calculation of distances during the matching procedure. With regard to a dichotomous variable there are only two possibilities of similarity: two cases are maximally similar (if both have the same value in the concerning variable) or maximally dissimilar (if they have not the same value in the concerning variable). Thus the probability of choosing a matching partner with a different value in a dichotomous variable is relatively low. In contrast, there are much more variants of similarity for

a metric variable. Thus in the case of a metric variable it is much more likely that the chosen partner is similar in respect to this variable but not equal.

Table 5: Distributions of the matching variables in the resulting data set

Variable	information from DEAS			information from VSKT		
	mean / %	sd	N	mean / %	sd	N
duration of employment	32.70	10.50	3,997	30,24	110.35	3,997
existence of employment gaps	0.42	0.49	3,997	0.42	0.49	3,997
duration of employment gaps (for persons with employment gaps)	5.00	5.46	1,667	9.84	7.67	1,665
employment gaps: parental leave	0.23	0.42	3,997	0.23	0.42	3,997
employment gaps: military/ civilian service	0.08	0.28	3,997	0.08	0.27	3,997
employment gaps: studies, further education	0.05	0.23	3,997	0.05	0.23	3,997
employment gaps: unemployment	0.07	0.26	3,997	0.07	0.26	3,997
employment gaps: sickness/ rehabilitation	0.02	0.15	3,997	0.02	0.15	3,997
retirement age (for retirees)	61.39	3.17	1,411	61.89	2.96	1,411
unemployment duration (for unemployed persons)	5.14	5.14	220	2.19	2.61	248
invalidity pension duration (for invalidity pensioners)	7.61	7.39	135	5.01	5.28	148
children (for women)	1.88	1.18	2,028	1.86	1.14	2,028
year of birth	1949.78	8.87	3,997	1950.42	8.64	3,997

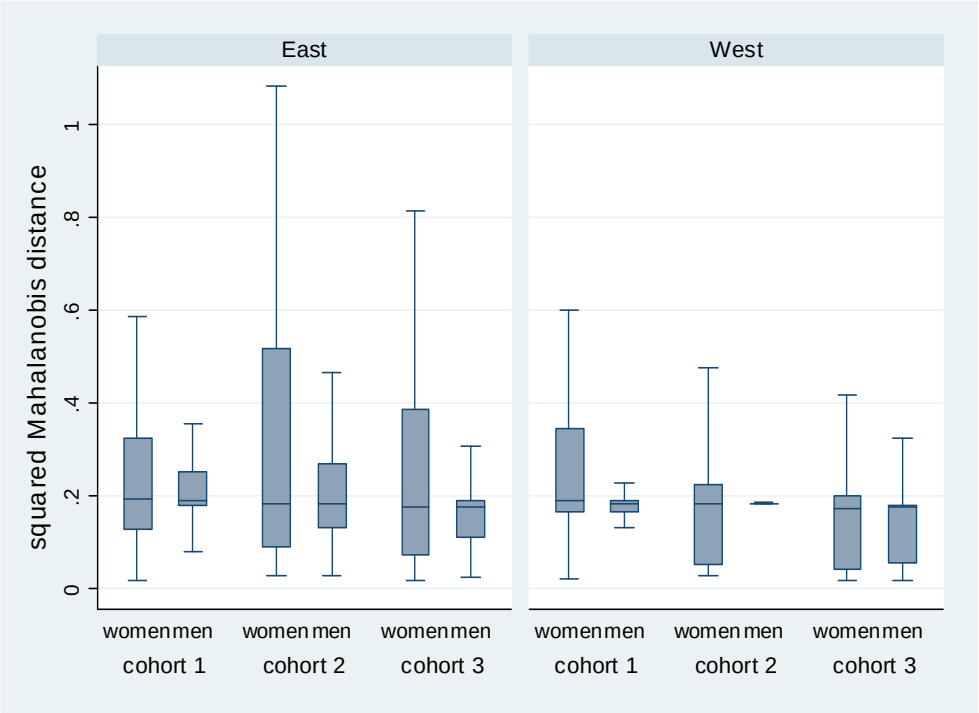
Source: DEAS 2008 and VSKT-LAW 2002-2005-2007 (combined data), own calculations

Although the relatively similar distributions give prima facie evidence for a good match of the data on the aggregate level, it does not necessarily mean that the individually matched pairs also have a high congruence. In the next section, therefore, we will go a bit deeper into question of the goodness of matching.

4.2. Assessing the goodness of matching

For assessing the quality of matching we consult different information. At first we have a look at the matching score, which in our case is the squared Mahalanobis distance. This distance gives us a first impression of the dissimilarity of the matched cases in regard to the variables we used for the matching. The second way to investigate the results of the matching is to take a look at the individual fit between the matching partners in relating to the matching variables. We compare at least the amounts of pensions we can take from both data sources for the matched cases as an external criterion for the goodness of matching.

Figure 2: Squared Mahalanobis distances



Source: DEAS 2008 and VSKT-LAW 2002-2005-2007 (combined data), own calculations

The box plots in Figure 2 show the median and interquartile range of the squared Mahalanobis distance for different groups. What we can see is that the matching results differ between the cohorts as well as between men and women and individuals from East and West Germany. For the younger cohorts the median distances are slightly lower than for older cohorts, and for the West the interquartile ranges are narrower and lower than for the East. Moreover, the interquartile ranges of the distances for women are wider than for men, which on the one hand could be an outcome of the additionally included matching variable for women (number of children). In general, the Mahalanobis distance becomes higher the more variables are included in the equation and the higher the variance of these variables is within a certain subgroup. Because information on children is not available for men in the VSKT, the value of this variable for all men is zero.[11] This means that the variance for men in this variable is bound to be zero, while for women it is much higher, which leads to greater distances of the matched pairs. On the other

[11] The value of zero was given because the allocation of a missing value would have excluded all men from the matching. Due to the use of slices in the matching process, the fact that all men have the same value in one variable does not affect the matching outcome.

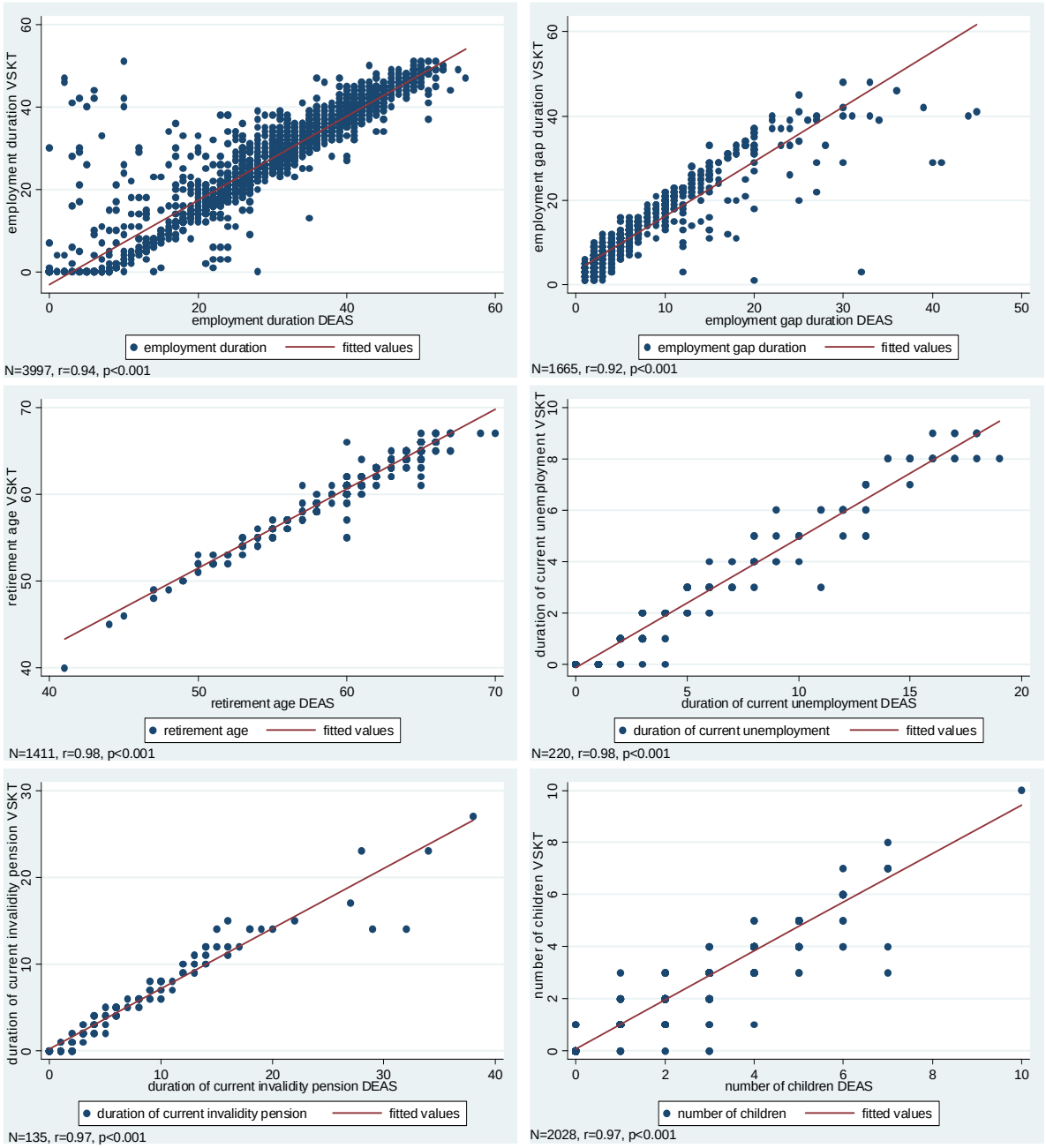
hand, the larger spread of distances among women could be a result of a higher degree of plurality in women's employment biographies than in men's.[12]

Overall, the values for the median are quite low, which means in 50 percent of the cases donor and recipient are relatively similar in regard to the matching variables. However, this does not automatically mean that we have a substantially good matching result. If the variances in the matching variables are very low, this leads to low Mahalanobis scores, but cases need not be very similar in regard to relevant outcome variables. However, even though this is not the case for our data, it might be helpful not only to look at the distances, but also on other information.

An additional way to investigate the results of the matching is to take a look at the individual relation between several metric matching variables, as given in Figure 3. What we see is a sound correspondence between the information from DEAS and VSKT on the individual level, although for all variables there are some outliers with a worse fit. However, for all variables there are outstandingly high correlations (r=0.92-0.98).

[12] Literature shows that women more often interrupt their careers to take care of children or other family members needing care (Dingeldey and Reuter 2003), and more frequently change to part-time or marginal employment (Steiner and Wrohlich 2005, Bothfeld et al. 2005).

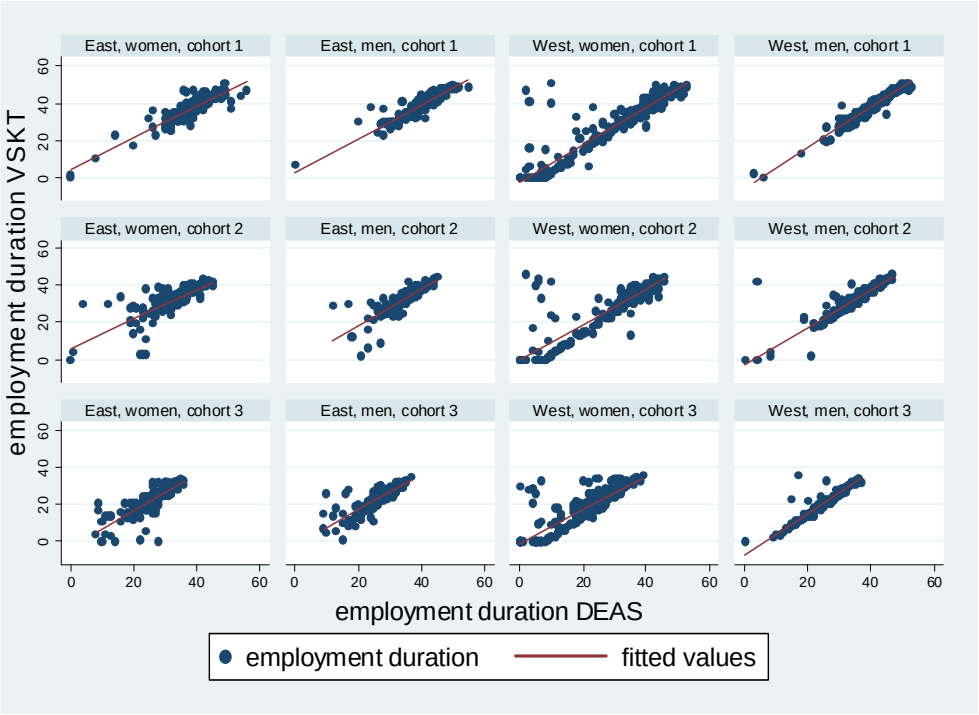
Figure 3: Associations between matching variables measured by DEAS and VSKT



Source: DEAS 2008 and VSKT-LAW 2002-2005-2007 (combined data), own calculations

Since employment is a central variable for further analyzes of life courses and old age incomes, it is useful to investigate if the fit of employment from both data sources differs between several groups. Figure 4 shows the correspondence of the employment duration from DEAS and VSKT on the individual level for 12 different groups by gender, region and cohort. What we can observe is that for all groups we have a very high positive correlation between both measurements. All correlations are strong and highly significant ($p<0.001$). They range between 0.80 (East German women of cohort 2) and 0.98 (West German men of cohort 1).

Figure 4: Associations between employment duration measured by DEAS and VSKT for different groups



Source: DEAS 2008 and VSKT-LAW 2002-2005-2007 (combined data), own calculations

The last way we want to check the results of the matching is to look on the amounts of pensions from both data sources for those who already receive an old age pension. We did not include the amounts of pensions as a matching variable because it is only available for a subgroup: in the DEAS the amount of pension is part of the drop-off questionnaire, which means that this information is only available for those already retired participants who did not only take part in the face-to-face interview, but also filled in the questionnaire. For this group, we are able to use the pension amount as an external criterion for the goodness of the matching. Whereas in the DEAS the respondents are asked to declare their income from old age provision, in the pension data the amount of pension is not given directly. However, one can approximate the pension amount from the sums of individual earning points ('PSEGPT') from East and West Germany and the pension values of the correspondent year.[13] For the present approximation of pension amounts we use the pension values of the year 2008, to keep the pension amounts comparable to those reported in the DEAS in 2008. In 2008, the pension amount for East Germany equals

[13] The sum of individual earning points ('PSEGPT') includes all full contribution periods, reduced contribution periods and non-contributory periods. In addition it takes into account the pension type factor and actuarial adjustment in case of early or late retirement. The pension type factor varies with the type of pension a person receives. In case of old age retirement it is one. The actuarial adjustment factor depends on the individual retirement age. If a person retires at the statutory retirement age, the factor equals one. Early retirement reduces; late retirement increases the factor (Rasner et al. 2007, 24).

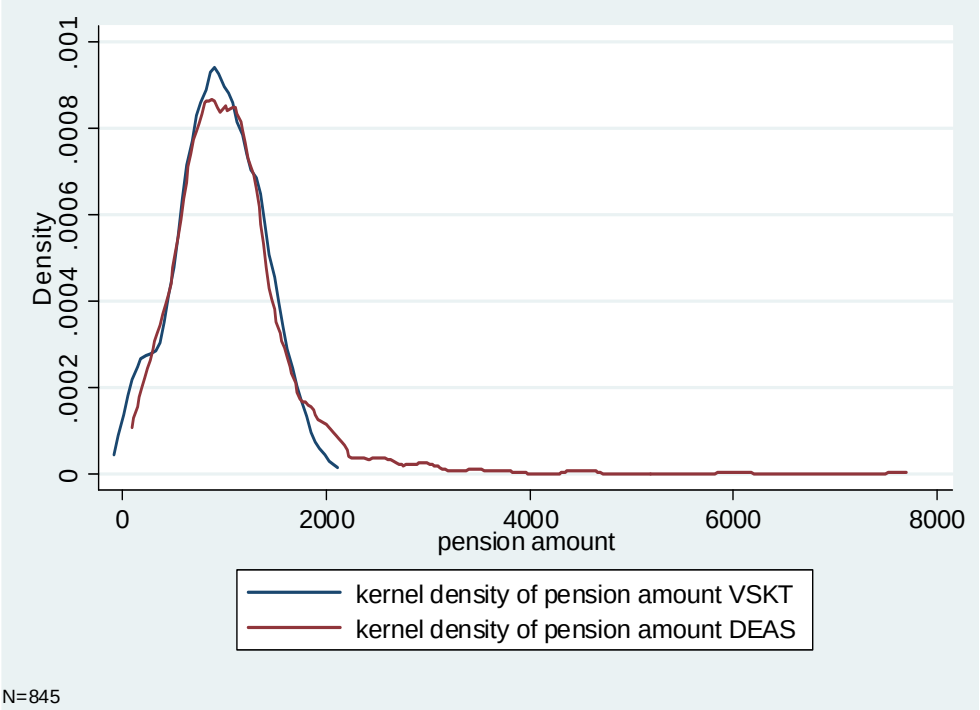
23.34 €, for West Germany 26.56 € (DRV, 2008, 11). Therefore, the following equation is used, whereby we calculate the pension amounts for those already in old age retirement only:

$$pension\ amount = \left(PSEGPT_{East} * pension\ value_{East_{2008}}\right) + \left(PSEGPT_{West} * pension\ value_{West_{2008}}\right)$$

In the following we compare the DEAS and VSKT based pension amounts of the 845 old age retirees for whom this information is available.[14] Given the information from DEAS, the mean pension amount is 1073.82 € (sd=619.21). In contrast, the average calculated pension amount from the VSKT at 950.79 € (sd=414.85), is somewhat lower. This difference is at least partly a result of a few relatively high pensions in the DEAS. This we can see in Figure 5, which gives us an impression on the distribution of the pension amounts coming from DEAS and VSKT via the kernel density function. It shows us that the distributions in the lower income segments are quite similar. While pension amounts derived from the VSKT very seldom exceed an amount of 2000 € (in our matched data we have three cases), we have a considerable number of DEAS pensions in the higher segments above 2000 € (N=39, which is about 5 percent of the 845 old age retirees for whom we have the information on pension amounts from DEAS).

[14] We dropped one person from the analysis of pension amounts who in the DEAS declared a monthly pension of 24,000 €, which is far above the usual pension amounts and probably either a mistake (yearly than monthly amount of pensions) or the sum of incomes from different sources.

Figure 5: Densities of the pension amounts in the joint data, derived from DEAS and VSKT

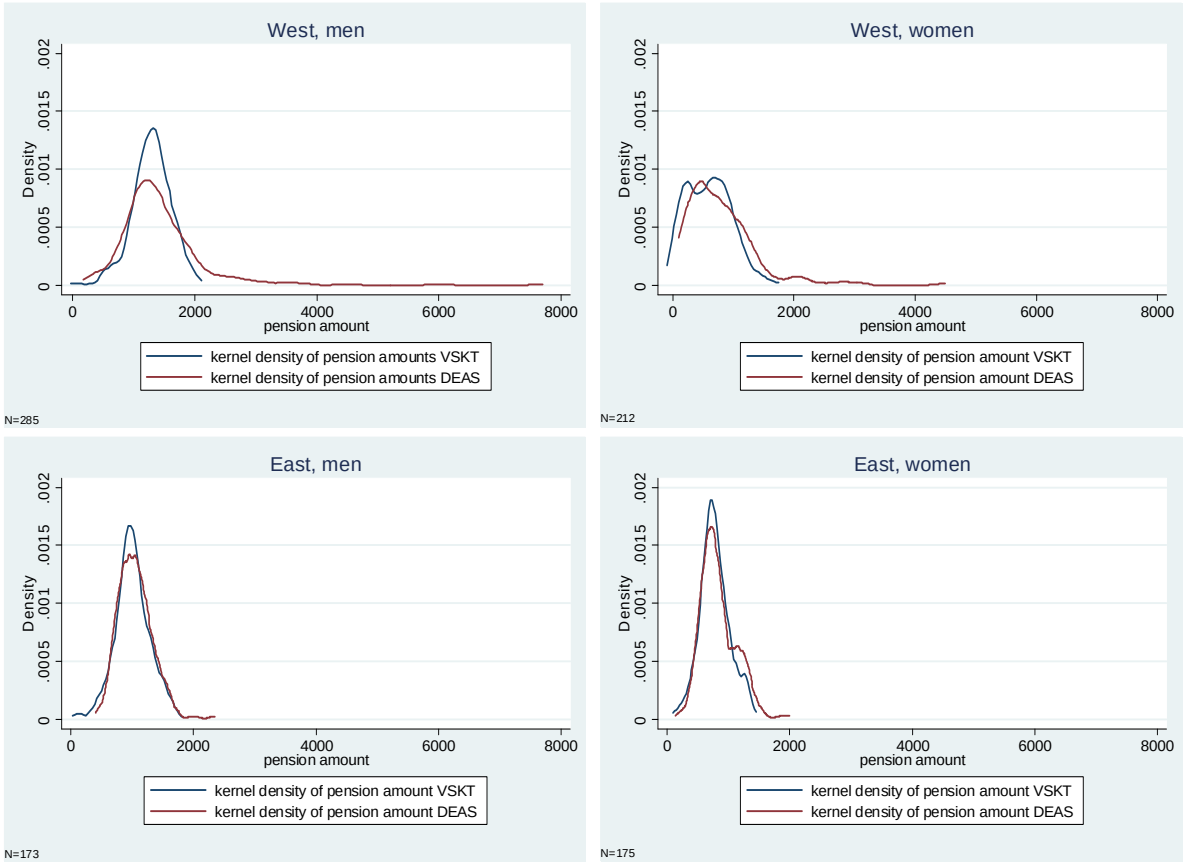


Source: DEAS 2008 and VSKT-LAW 2002-2005-2007 (combined data), own calculations

As we can learn from Figure 6, which shows the densities of the pension amounts for different groups,[15] the higher pension amounts are for retirees from West Germany (especially men). Additionally, both for East German men and women the distributions (for DEAS pensions as well as for VSKT pensions) are narrower and more concentrated in the lower segments than for their West German counterparts. Moreover, the distributions of DEAS and VSKT derived pension amounts are more similar in East than in West Germany, and the largest difference we can see is for West German men. Nevertheless, within all groups the distributions of DEAS and VSKT derived pension amounts are quite similar.

[15] We do not distinguish between cohorts on pension amounts because most of the old age pensioners in our sample belong to the oldest cohort.

Figure 6: Densities of the pension amounts in the joint data, derived from DEAS and VSKT for different groups

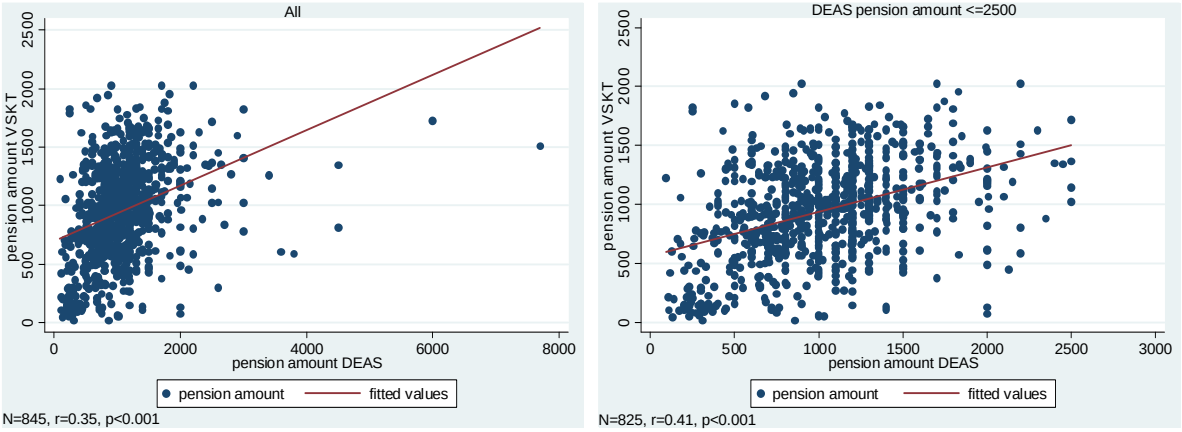


Source: DEAS 2008 and VSKT-LAW 2002-2005-2007 (combined data), own calculations

Figure 7 shows the individual concordance between the amounts of pension measured by DEAS and VSKT for those who are already in old age retirement. Although we can observe a positive association, we have to consider that the individual concordance between the pension amounts is only moderate, since we have an overall correlation of r=0.35 (Pearson correlation coefficient). If we only take into account those retirees who in DEAS reported a pension not higher than 2,500 €, the correlation increases moderately (r=0.41).[16]

[16] Pensions over 2,500 € from the statutory pension insurance are not impossible to reach, but nevertheless extremely rare. It is hence safe to assume that at least some of the high pension amounts from DEAS do not solely reflect income from the statutory pension insurance, but a combination from different income sources.

Figure 7: Associations between pension amounts derived from DEAS and VSKT



Source: DEAS 2008 and VSKT-LAW 2002-2005-2007 (combined data), own calculations

However, we have to remark that if we consider the correlations within subgroups by region and gender, they are somewhat lower. Thereby, the highest correlation (r=0.20) can be reported for West German women. So it is likely that the overall moderate correlation is partly an effect of similar distributed pension amounts within the groups and similar differences between the groups. The finding that the individual associations within the groups are much lower could be an effect of varying incomes over the life course, which we did not take into account in the matching process. Income over the life course is the primary determinant for the amount of pension, especially for men, for whom the variance due to more continuous working careers and no or only a few employment interruptions is lower. Furthermore, we should remember, that in our original data sources we had relatively high differences between the rates of employment gaps (especially unemployment), which could in case of underestimation in one data source have weaken the individual fit of pension amounts.

5. Discussion

In the present paper we have described the statistical matching of the German Aging Survey (DEAS) with administrative pension insurance data, the Sample of the Insured Population Pension Records (VSKT). The matching procedure consists of different steps including the definition of the sample population, the selection of the matching variables, the matching itself, and the assessment of the matching quality. This last step is very important to verify the goodness of the whole process and to assess whether the resulting data set can be used for the desired analyses.

The results of our quality tests with regard to the goodness of matching are somewhat mixed. At first we observed the distributions of the matching variables before and after the matching process. Whereas after matching for the metric variables there were still some differences in the distributions, for the dichotomous variables we found a remarkable improvement in resemblance.

The examination of the matching score (squared Mahalanobis distance) gave us an impression of the dissimilarity between the matched cases in regard to the variables we used for the matching. Here we found relatively low distances in general. For the younger cohorts the distances were slightly lower than for older cohorts and for the West lower than for the East. Moreover, the spread of the distances was larger for women than for men, which could point to a higher degree of plurality among women. We also investigated the results by looking at the individual fit between the matching partners with regard to the matching variables. Here we found a sound correspondence between the information from DEAS and VSKT with outstandingly high correlations for all variables.

We compared at least the amounts of pensions taken from both data sources as an external criterion for the goodness of matching. With the exception of some very high pension amounts reported in the DEAS, the distributions of pension amounts, derived from both data sources, were relatively similar. Despite this high accordance on the aggregate level, the individual correlations were only moderate. Besides statistically controlling for gender and region, the correlations between the pension amounts were even lower. These only middle-rate correlations between pension amounts could be a result of the absence of income as an important factor for pension amounts in the matching equation. So, for generating a matched data source that provides a profound basis for analyzing old age provision, the inclusion of income would be valuable. However, as we alluded before, in the DEAS information on income is only available for one point in time, so it might be difficult to use this for improving our matching.

Summarizing, the quality tests of our matching tell us that the resulting data set seems appropriate to carry out analyses on changes in employment careers, since the concordance of the matched biographies with regard to employment related variables is high and the matching hence quite successful in this respect. However, the results imply that the resulting data is more appropriate for life course analyses than for analyzing old age incomes, since there are some differences between the pension amounts coming from both data sources.

However, the combined data set might also be useful for analyses of old age incomes on an aggregate level because it generally gives us the possibility of enlarging the focus from the statutory pension to a combination of different old age income sources. Even if the individual match between pension amounts is not optimal, the data allows us to compare the distributions of the income from the statutory pension topped up with other income forms like private savings or occupational pensions, within different subgroups. Moreover, the combined data allows us to carry out further analyses that would not be possible with one of the data sources alone, for instance analyses regarding the influence of the employment history (from VSKT) on retirement plans and expectancies (from DEAS) of the baby boomers.

Besides the practical benefits of the matching for our further analyses on life courses and old age incomes, what we have learned from the procedure is that for statistical matching of survey and register data it is very important to have a good basis of valid and comparably measured matching

variables. Furthermore, one has to check if the existing matching variables suit the aim of matching. If one aim of the statistical matching, for instance, is to provide a data basis for the analysis of old age incomes, as it was in the present case, the absence of a longitudinal income variable can lower the quality of the results. However, as we have seen, even if the number of matching variables is limited, this does not necessarily imply a low-level quality of the results. Taking these considerations into account, statistical matching of survey and register data seems to be an appropriate and valuable way to raise the possibilities of analysis within the data sources, and hence to close some research gaps.

References

Bothfeld, S.. Klammer, U., Klenner, C., Leiber, S., Thiel, A., & Ziegler, A. (Eds.) (2005) WSI FrauenDatenReport. Berlin: edition sigma.

Dingeldey, I., & Reuter, S. (2003). Beschäftigungseffekte der neuen Verflechtung zwischen Familien- und Arbeitsmarktpolitik. WSI Mitteilungen 11, 659-64.

D'Orazio, M., Di Zio, M., & Scanu, M. (2001). Statistical Matching: a tool for integrating data in National Statistical Institutes (Rome, Italian National Statistical Institute. http://epp.eurostat.ec.europa.eu/portal/page/portal/research_methodology/documents/43.pdf.

DRV (Deutsche Rentenversicherung) (2008). Rentenversicherung in Zahlen 2008. Berlin: Deutsche Rentenversicherung Bund.

FDZ-RV (Forschungsdatenzentrum Rentenversicherung) (2008). FDZ-Biografiedatensätze – VSKT/VVL. Benutzerhinweise zu den Verlaufsmerkmalen und Merkmalen der Rentenberechnung (April 14, 2008).

Frick, J.R., & Grabka, M.M. (2010). Wealth Inequality and the Importance of Public Pension Entitlements. Paper prepared for the LIS conference "Inequality and the Status of the Middle Class: Lessons from the Luxembourg Income study", Luxembourg, 29-30 June 2010.

Gu, X. S., Rosenbaum, P.R. (1993). Comparison of Multivariate Matching Methods: Structures, Distances, and Algorithms. Journal of Computational and Graphical Statistics, 2(4), 405- 420.

Heckman, J.J. (1990). Varieties of selection bias. American Economic Review, 80, 313-318.

Himmelreicher, R.K., & Schröder, C. (2010). Vorüberlegungen zur statistischen Verknüpfung von Querschnitts-Surveydaten mit prozessproduzierten Längsschnittdaten: EVS und VSKT. Deutsche Rentenversicherung 2/2010, 208-216.

Himmelreicher, R.K., & Stegmann, M. (2008). New Possibilities for Socio-Economic Research through Longitudinal Data from the Research Data Centre of the German Federal Pension Insurance (FDZ-RV), Schmollers Jahrbuch 128, 647 – 660.

Janson, C.G. (1990). Retrospective data, undesirable behavior, and the longitudinal perspective. In Magnusson, D. & Bergman, L.R. (Eds.). Data quality in longitudinal research, 100-121.

Kantor, D. (2006). MAHAPICK: Stata module to select matching observations based on a Mahalanobis distance measure. Statistical Software Components, Boston College Department of Economics. http://econpapers.repec.org/software/bocbocode/s456703.htm.

Kum, H., & Masterson, T. (2008). Statistical Matching Using Propensity Scores: Theory and Application to the Levy Institute Measure of Economic Well-Being. Working Paper No. 535. The Levy Economics Institute of Bard College.

Leisering, L., Müller, R., & Schumann, K.F. (Eds.) (2001). Institutionen und Lebensläufe im Wandel. Institutionelle Regulierungen von Lebensläufen. Weinheim.

Mahalanobis, P.C. (1936). On the generalised distance in statistics. Proceedings of the National Institute of Sciences of India 2 (1): 49–55.

Middendorff, E. (2002). Panta rhei oder der mentale Fluss von Tatsachen: Zur Reliabilität retrospektiv erhobener biographischer Ereignisse. ZA-Information 46, 58-71.

Motel-Klingebiel, A., Wurm, S., Engstler, H., Huxhold, O. Jürgens, O., Mahne, K., Schöllgen, I., Wiest, M., & Tesch-Römer, C. (2009). Deutscher Alterssurvey: Die zweite Lebenshälfte. Erhebungsdesign und Instrumente der dritten Befragungswelle. Berlin: DZA (DZA Diskussionspapiere, No. 48).

Rasner, A., Himmelreicher, R.K., Grabka, M.G., & Frick, J.R. (2007). Best of both worlds: preparatory steps in matching survey data with administrative pension records: the case of the German Socio-Economic Panel and Completed Insurance Biographies 2004. Berlin: DIW.

Rosenbaum, P.R., & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika 70 (1), 41-55

Rubin, D.B. (1980). Bias Reduction Using Mahalanobis Metric Matching. Biometrics, 36, 293-298.

Simonson, J., Motel-Klingebiel, A., & Kowalska, K. (2010). Alterssicherung und Alterseinkünfte im Deutschen Alterssurvey (DEAS). Deutsche Rentenversicherung 2/2010, 301-313.

Steiner, V., & Geyer, J. (2009). Erwerbsbiografien und Alterseinkommen im demografischen Wandel – eine Mikrosimulationsstudie für Deutschland. Berlin: Forschungsnetzwerk Alterssicherung der Deutschen Rentenversicherung.

Steiner, V., & Wrohlich, K. (2005). Work Incentives and Labour Supply. Effects of the Mini-Jobs Reform in Germany, Empirica, 32: 91–116.

Zhao, Z. (2004). Using Matching to Estimate Treatment Effects: Data Requirements, Matching Metrics, and Monte Carlo Evidence. The Review of Economics and Statistics. 86(1), 91-107.