View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by SSOAR - Social Science Open Access Repository

SSOAR

Open Access Repository
www.ssoar.info

Statistical matching of the German Aging Survey and the Sample of Active Pension Accounts as a source for analyzing life courses and old age incomes

Simonson, Julia; Romeu Gordo, Laura; Kelle, Nadiya

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
GESIS - Leibniz-Institut für Sozialwissenschaften

Empfohlene Zitierung / Suggested Citation:
Simonson, J., Romeu Gordo, L., & Kelle, N. (2012). Statistical matching of the German Aging Survey and the Sample of Active Pension Accounts as a source for analyzing life courses and old age incomes. *Historical Social Research*, *37*(2), 185-210. https://doi.org/10.12759/hsr.37.2012.2.185-210

Nutzungsbedingungen:
Dieser Text wird unter einer CC BY Lizenz (Namensnennung) zur Verfügung gestellt. Nähere Auskünfte zu den CC-Lizenzen finden Sie hier:
https://creativecommons.org/licenses/by/4.0/deed.de

Terms of use:
This document is made available under a CC BY Licence (Attribution). For more Information see:
https://creativecommons.org/licenses/by/4.0

gesis
Leibniz-Institut
für Sozialwissenschaften

Mitglied der
Leibniz-Gemeinschaft

Diese Version ist zitierbar unter / This version is citable under:
https://nbn-resolving.org/urn:nbn:de:0168-ssoar-372915

Statistical Matching of the German Aging Survey and the Sample of Active Pension Accounts as a Source for Analyzing Life Courses and Old Age Incomes

Julia Simonson, Laura Romeu Gordo & Nadiya Kelle *

Abstract: *»Statistisches Matching von Deutschem Alterssurvey und Versicherungskontenstichprobe als Quelle zur Analyse von Lebensläufen und Alterseinkommen«*. The paper examines the combination of the German Aging Survey (*Deutscher Alterssurvey* – DEAS) with the Sample of Active Pension Accounts (*Versicherungskontenstichprobe* – VSKT), as an example of how survey results may be linked together with administrative data using statistical matching. Statistical matching is a technique increasingly being applied in order to combine information from different data sources where no linkage can be made between records based on any unique identifier. This might be due to confidentiality restrictions or attempts to avoid the high attrition rates connected with informed consent requirements. The aim of this matching is to provide a combined dataset that contains more information than the data sources would on their own. In our paper, we detail some preparatory steps for making this match, such as the definition and adjustment of matching variables. The steps we describe may well be worth challenging on the basis of the divergent characteristics of the two sets of data. We also outline the procedure we used to combine the data sources, based on the Mahalanobis distance vector. Finally, we assess the quality of the matching by comparing the individual pension amounts that we can extract from each of the two matching sources as our external criterion.

Keywords: statistical matching, old age income, life course research, German Aging Survey, Sample of Active Pension Accounts.

* Address all communications to: Julia Simonson, German Centre of Gerontology (DZA), Manfred-von-Richthofenstr. 2, 12101 Berlin, Germany; e-mail: Julia.Simonson@dza.de.
Laura Romeu Gordo, German Centre of Gerontology (DZA), Manfred-von-Richthofenstr. 2, 12101 Berlin, Germany; e-mail: Laura.Romeu-Gordo@dza.de.
Nadiya Kelle, German Centre of Gerontology (DZA), Manfred-von-Richthofenstr. 2, 12101 Berlin, Germany; e-mail: Nadiya.Kelle@dza.de.
Previous versions of this paper were presented to the 4th Conference of the European Survey Research Association (ESRA) on July 18-22, 2011 in Lausanne, as well as to the workshop on 'Statistical Matching and Sequence Analysis' of the research project 'Life Course, Aging and Well-being' (LAW) on September 2-3, 2010 in Berlin. We thank all discussants for their helpful remarks and gratefully acknowledge Hans Kiesl for his substantial comments. We also thank the members of the LAW-project for their support for our work.

Historical Social Research, Vol. 37 — 2012 — No. 2, 185-210

1. Introduction

The goal of this contribution is to describe a case of statistical matching based on the German Aging Survey (*Deutscher Alterssurvey* – DEAS) and the Sample of Active Pension Accounts (*Versicherungskontenstichprobe* – VSKT), which contains pension insurance data collected by public administration. To be more specific, what we aim to do is describe the data preparation process, the selection of matching variables, the matching procedure we employ and the method we use to validate the quality or our matching. This paper, apart from the discussion it contains on the challenges facing the matching procedure, describes the potential gains to be had from linking survey and administrative data together.

Before entering into methodological issues, the first question that arises is why in the first place we want to match these two datasets; or in other words, what benefits there are in working with a VSKT-DEAS matched dataset that cannot be achieved by working with the two datasets separately.

The data matching we describe was completed within the framework of 'Life Course, Aging and Well-Being' (LAW), a project being financed by the Volkswagen Foundation (2009-2012). This project is being carried out by three cooperating institutions – the German Centre of Gerontology (*Deutsches Zentrum für Altersfragen* – DZA), the German Socio-Economic Panel (*Sozio-oekonomisches Panel* – SOEP), and the German Federal Pension Scheme (*Deutsche Rentenversicherung Bund* – DRV). In this project we analyze how the various life courses taken by German baby boomers (i.e. those born between 1956 and 1965) have changed in comparison with older cohorts, and how these changes in life courses are likely to 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 pluralization and inhomogeneity (Leisering et al. 2001). To be more specific, their employment biographies are less continuous and episodes of unemployment, part-time employment and self-employment are more common. Furthermore, their family biographies deviate more often from traditional family norms, for example, showing more non-marital partnerships and more periods of single life. All these changes in life course patterns may potentially affect the lives and material situations of baby boomers in the future – impacting especially on the protection provided by public and occupational pensions, as well as private provision made to ensure old age security.

VSKT and DEAS both are helpful in the analysis of such pluralization and inhomogenization trends in life courses and their consequences for old age provision. However, neither of them provides sufficient information. The VSKT, which records pension insurance data, covers a large number of cases and is very informative in that it records the series of life episodes of contributors, but mainly concentrates on information relevant to the calculation of the statutory pensions; e.g., periods of employment and their associated month-to-month earnings. However, statutory pensions are only one source of income in old age. There are other resources in the mix, such as private old age provision, occupational pension schemes, wealth, savings, and inheritances. Thus, while this pension data is helpful for analyses on specific working biographies, work related income, and pension entitlements from the statutory pension system, it does not contain all the relevant information needed to give us a complete picture of potential financial situations in old age.

Some of this information missing from the VSKT can be recovered using the German Aging Survey (DEAS).[1] The DEAS is a nationwide representative cross-sectional and longitudinal survey of the German population aged 40 or older that includes comprehensive information on living situations in old age and contains detailed subjective evaluations on the life quality of respondents, thus enabling us to place the analysis of income situations in old age within a broader context (Simonson et al. 2010). It does not give as complete and exhaustive a picture of employment careers as the VSKT, but it does inform us comprehensively on transitions to retirement, the various different sources of old age incomes, private and occupational pension schemes, inheritances, as well as monetary and non-monetary transfers within the family. Furthermore, the DEAS offers the information needed to analyze the prospective situation of the baby boomers, such as data on retirement plans and attitudes towards old age provision.

Summarizing, while the two data sources do not on their own offer a complete picture of the lives of baby boomers and control cohorts, matching the VSKT and the DEAS would allow us to complement information on public pension entitlements with data on other relevant determinants of future life arrangements. Using such a combined dataset it may even be possible to analyze whether the employment and family biographies of the baby boomers and their general financial situation and attitudes have changed as compared to the control cohorts and the future consequences of such changes for their lives.

[1] 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 (SOEP) also contains comprehensive information on life courses, different sources of incomes and living situations. This fact indicates that a statistical matching made between the SOEP and the VSKT would be useful in analyzing life courses and old age income situations too. For this reason, SOEP and VSKT are also combined as part of the project 'Life Course, Aging and Well-Being' (LAW).

2. Data

2.1. The Sample of Active Pension Accounts (VSKT)

The Sample of Active Pension Accounts (VSKT) is a one percent random sample of the insurance accounts kept by the German statutory pension agencies. In Germany, statutory pension insurance is mandatory for all employed persons in the private and public sector. The pension data of the German Federal Pension Scheme therefore covers more than 90 percent of the German population (Himmelreicher and Stegmann 2008). An individual account is kept for each compulsory member of the statutory pension insurance scheme, recording all periods of contribution and relevant non-contributory periods (such as child-raising phases) until retirement (Himmelreicher and Stegmann 2008).

The VSKT is a stratified random sample taken from these accounts. It contains information on all relevant periods in which contributions credits are recorded and on the pension entitlements of insured persons participating in the German statutory pension insurance system aged between 15 and 67 years. Insurance accounts data are recorded using a random sample and accounts are followed over time offering a panel structure (Himmelreicher and Stegmann 2008).

It should be noted that although the data represents all insured persons, it is not representative of the whole German population. Only persons eligible for old age or disability pension benefits are contained in the sample. People who have been employed as civil servants ('*Beamte*') or as self-employed for the whole of their employment lives are not included. In addition, some other occupation groups such as lawyers or doctors, are not insured through the statutory pension insurance scheme and therefore do not form part of the VSKT.[2] It should also be remembered that the data quality of the accounts is quite variable and will depend heavily on the validation status of each account ('*Kontenklärung*'). To fill gaps in individual accounts, the pension insurance scheme contacts the insured persons with a request to supply the information needed to close such gaps. This is done at regular intervals, but people answer these requests with a varying degree of effort, leading to a varied degree of completeness in the accounts.

For the purposes of the statistical matching procedure described here, a special VSKT dataset – the 'VSKT-LAW-Sample' – has been prepared, which is described in the following section.

[2] 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 'Life Course, Aging and Well-Being' research project (LAW), the Research Data Centre of the German Federal Pension Scheme (FDZ-RV) provided a special VSKT dataset. This VSKT-LAW sample contains information from VSKT datasets of three different years: 2007, 2005, and 2002. This is due to the fact that as part of that project we planned to analyze the life courses of three separate birth cohorts: 1936-45 (cohort 1), 1946-55 (cohort 2), and 1956-65 (baby boomers; cohort 3). Since the VSKT covers information on individual life courses for people between the age of 15 and 67, it is not possible to get full information on the life courses of the three cohorts using a single VSKT. For instance, the 2007 data does not include the birth years 1936-1939. The 2005 data, however, 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, meaning that we would not know what has happened since then if we were to use this data source alone. So, to get the most comprehensive information possible, the 2007 data was enhanced using the 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 validated with the help of insured persons, as well as accounts that could be validated without their help by virtue of the fact that they contain no gaps since the last validation. The VSKT-LAW sample covers a total of 205,828 accounts from Germans and non-German citizens. However, we use only the accounts for German citizens for the matching procedure as described later.

2.2. The German Aging Survey (DEAS)

The German Aging Survey (DEAS) is a national 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 (*Bundesministerium für Familie, Senioren, Frauen und Jugend* – BMFSFJ). It provides a comprehensive look at people in mid-life to older adulthood, including micro data for use both in social and behavioral scientific research and in reporting on social developments (Motel-Klingebiel et al. 2010). The data thus provides a source of information for decision-makers, the general public and for scientific research.

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

The third wave differentiated between three sub-samples: (1) persons who took part in the survey in both 1996 and 2002, (2) persons previously assessed only in 2002, and (3) the new group of 6,205 participants included in the study for the first time. This approach facilitates 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 identified at the points in time 1996, 2002 and 2008, and an investigation of individual personal developments that occurred over either six or twelve years (2002-2008, or 1996-2002-2008). Another perspective results from the comparison of individual developments over a six year period in the two time-frames 1996 to 2002 or 2002 to 2008, i.e. a comparison of the developments affecting the two birth cohorts within specific age segments.[3]

The first-time respondents (n=6,205) to the third wave of the DEAS are chosen for our statistical matching procedure. The DEAS 2008 sample is a stratified random sample of the population aged 40 years or older living in Germany. That means that it includes both Germans and non-Germans. However, as we explain later, only the sub-sample consisting of German citizens born between 1936 and 1965 is included in our statistical matching procedure.

3. Statistical Matching

In empirical research, statistical matching is a technique increasingly applied in order to combine information from different data sources where no one-to-one correspondence is possible via a unique identifier (D'Orazio et al. 2001; Kum and Masterson 2008). Such a difficulty may result from confidentiality restrictions or to attempts to avoid the attrition rate that is associated with having to obtain informed consent for direct matching ('record linkage') based on a unique identifier.[4]

[3] From 2011 on information from the panel will be collected every third year. In addition, a yearly panel study for participants aged 70 years and older will also be carried out. New base samples are conducted every sixth year. Further information on the German Aging Survey as well as the data collection instruments used in the previous waves can be downloaded from the internet <www.deutscher-alterssurvey.de and www.fdz-deas.de>.

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

Originally employed for the most part to avoid selection bias problems in medical studies and in analytical research evaluating the effects of treatments (Heckman 1990; Zhao 2004), statistical matching is now increasingly commonly used in the social and economic fields too. Here, however, the aim of statistical matching is often not to obtain comparable groups of treated and untreated persons to evaluate treatment effects, but to combine the information coming from different data sources. For example, Rasner et al. (2007) used statistical matching to combine the Scientific Use File Completed Insurance Biographies (SUF VVL) with the German Socio-Economic Panel Study (SOEP). Rasner et al. (2011) and Frick and Grabka (2010) used statistically matched data from the German Socio-Economic Panel Study (SOEP), the Sample of Active Pension Accounts (VSKT) and the *Statistics for Pension Rights Splitting in Case of Divorce (Versorgungsausgleichsstatistik)* 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 (SOEP) with the Sample of Active Pension Accounts (VSKT) in order to obtain a sufficient basis to analyze old age incomes.

Unlike cases where record linkage is used, statistical matching does not aim to identify the same individual in two datasets, but links the information from individuals with similar features, at least in relation to certain characteristics ('matching variables') observed in the two datasets (Frick and Grabka 2010). This means that statistical matching procedure uses variables common to both datasets to identify similar records, which are then linked together in order to obtain a combined dataset capable of facilitating analyses that would be have been impossible using a single input data source alone.

The most common method of statistical matching is propensity score matching (Rosenbaum and Rubin 1983). A propensity score is a function of several variables that one expects to affect the probability of belonging to a particular treatment group. In this way the propensity score obtains the probability that a particular subject is undergoing a particular treatment. Where matching is being done in order to analyze treatment effects, the use of propensity scores would seem a quite appropriate approach. But for matching with the purpose of combining two sets of data it is less than optimal due to the absence of any treatment variable upon which to base it. For this reason we do not use the very commonly employed propensity matching method, but instead employ a Mahalanobis distance vector based matching procedure (Gu and Rosenbaum 1993; Rubin 1980), which directly calculates the distances between possible cases in relation to variables selected as relevant without making any detour via any propensity to be receiving treatment.

3.1. Preparatory Steps in Combining the Data Sets

Statistical matching first requires some preparatory steps to be taken, such as the definition of the sample population, the selection of matching variables and the comparison of distributions; all of which are described in the sections to follow.

3.1.1. Defining the Sample Population

To do statistical matching successfully, it is useful to have data sources that represent the same population at least to some degree, especially in relation to some key dimensions such as age or gender. If the populations of the underlying data sources differ strongly, the reliability of the results could be appreciably reduced. It is important therefore that the sample population be specified correctly (Rasner et al. 2007).

The population of interest for our matching procedure consists of men and women who belong to the three birth cohorts 1936-1945, 1946-1955, and 1956-1965. Because of the relatively small number of foreign respondents in the German Aging Survey and the lower level of completeness of the foreigners' accounts in the VSKT due to the temporary residence status of many foreigners, only accounts from individuals with a German nationality are included in our statistical matching. As we will see below, there are also other restrictions relating to the population emerge from the characteristics of our two data sources. The VSKT-LAW sample that we use as one of the starting bases for our matching covers a total of 205,828 accounts for both Germans (n=151,844) and foreigners (n=53,984) insured either by the statutory schemes in Germany through their status as employees or in the German Federal Pension Scheme for some other reason. However, for our purposes, only accounts for individuals of 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 quality of the data recorded in their accounts.[5] Finally, we also drop some outliers in

[5] Until 1967 West German women could make use of a marriage reimbursement, i.e. after marriage they could receive a payout of the contributions made up until then. Their pension accounts were eliminated, but if they wanted, e.g. in case of later becoming employed anew, the women could pay their reimbursement back to the pension insurance (*retroactive payment*) and open a new pension account. If they did so, the earning points for the repaid contributions were credited to the new opening account without information where they came from (e.g. periods of education or employment). All this means that the VSKT does not give valid information on the beginning of the employment biographies of such women. Since this regulation was terminated in 1967, it involved women born up until 1949. Women who made use of the retroactive payment for marriage reimbursement can be identified in the data. Women who made use of this reimbursement and did not pay back the reimbursement later may also remain part of the VSKT, for example, if they later reentered employment. A new account with an opening balance of zero was opened for such women,

relation to retirement age from the data (n=361; 0.2 percent).[6] After all this has been done, the resulting dataset contains 148,110 insurance accounts from individuals born between 1936 and 1965, whereby most of the accounts (87 percent) come from VSKT 2007 – a plausible statistic, considering that all accounts for individuals born between 1940 and 1965 come from VSKT 2007, and only the accounts of insured persons born between 1936 and 1939 stem from VSKT 2002 (7 percent) or 2005 (6 percent).

Our second initial basis for the matching task is the third DEAS wave, in which 6,205 respondents were interviewed for the first time. The DEAS 2008 sample includes all Germans and non-Germans aged 40 or older. However, only the three birth cohorts 1936-1945, 1946-1955, and 1956-1965 (n=4,555) are included in our data for matching, and of those, only the sub-sample with German citizenship (n=4,414) is used.

To render the populations of the German Aging Survey comparable with the Sample of Active Pension Accounts, one should bear in mind that only people who are registered with an insurance account in the German Statutory Pension Scheme are included in the pension data (VSKT). People who have been employed for their entire employment history as civil servants ('Beamte') or who have always been self-employed are not included. In addition, there are some other occupational groups (independent professions, e.g. lawyers or doctors) who are not insured in the statutory pension insurance and who therefore do not form part of the VSKT.

The DEAS covers the whole population aged 40 or older independently of their current or previous occupations. For this reason it will be necessary before matching the data to exclude from the DEAS sample all persons belonging to groups with a high probability of not having a counterpart in the VSKT. The difficulty lies in how to decide which persons in the DEAS need to be excluded, since in the DEAS we do not have information on occupational positions during the subjects' entire employment careers, but only for the beginning and end of those careers (for retirees or unemployed persons) or for the beginning of careers and for the current point in time (for employed persons). For this reason we cannot define with absolute certainty who has been a civil servant or self-employed for most of his or her working life. Since their status at the end of their career is very likely to be their predominant status if it is identical to their status on starting up work, all those occupied as civil servants, as self-employed or in an independent profession both initially and currently (for employed persons) or both initially and in their last job (for retirees) are ex-

with the result that the duration of their employment (and the resulting income) is underestimated in the VSKT. Unfortunately, we have no information on which and how many women are involved.

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

cluded from the matching. This is the case for 306 persons from the sub-sample used of German citizens born between 1936 and 1965 (n=4,414), leaving a data base of 4,108 cases. Persons with outlying values for their retirement age (n=14; 0.4%) are also dropped from the DEAS. As the matching procedure excludes cases with missing values in one or more of the matching variables used in the calculation of the Mahalanobis distance, thus also excluding all such cases from the matching, the DEAS sample size we use for matching is reduced by another 97 cases. All these exclusions result in a DEAS matching source containing 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 refer to basically the same group (i.e. Germans born in the period 1936-1965 insured in the public pension scheme), the distributions of relevant socio-demographic variables in the two data sources remain divergent, 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 more strongly overrepresented in the DEAS data, which is a result of the disproportional sampling used for the DEAS. Persons belonging to at least the cohort 1936-1945 are overrepresented in the DEAS, while the pension data contains a distribution of birth cohorts that reflects the real-life predominance of baby boomers (1956-1965). According to the differences that can be seen in cohort and age distributions, we also appear to have a difference in the proportion of retirees in each data source. While in the DEAS the share of retirees is 35 percent, in our VSKT dataset it is only about 27 percent. Overall, though there are certainly some substantial differences in the distributions of socio-demographic characteristics, they do not, however, seem grave enough to seriously endanger the success of the matching.

3.1.2. Matching Variables

For the statistical matching procedure to be successful, the datasets need to share some set of variables that can be measured in comparable ways (Himmelreicher and Schröder 2010; Rasner et al. 2007). Since we want to create a data basis for use in analyzing life courses, employment histories and old age incomes, the goal of our matching procedure is to combine the data of people who exhibit similar characteristics in relation to these fields. It would therefore seem reasonable to focus on matching variables which offer us substantial information on working biographies.

Since the purposes and collecting methods involved in the two data sources are very different, our matching has to be based on a limited set of core variables (for an overview of these see Table 2). Although it is generally feasible to include a categorical variable via a set of indicator (dummy) variables (Kantor 2006) in computation of the Mahalanobis distance, it is recommended to use only continuous variables. The inclusion of categorical variables has the drawback that the computed distance function is more strongly affected by them than by continuous variables, leading to a match result that is strongly driven by such categorical variables. However, and in the knowledge that this course of action may have a negative effect on our matching results, we are forced to include some categorical variables in our matching due to a shortage of appropriate continuous matching variables.[7]

Both datasets contain some information on the employment biography and on such socio-demographic characteristics as age, gender, and region (a distinction is made between East and West Germany). In the main we also have information on education levels, but in our VSKT data source this variable unfortunately contains a substantial number of missing values (56 percent), as such information is not required for calculating pension entitlements. We have therefore decided not to use educational level as matching variable. Income too would have been a helpful variable, but, while the VSKT contains longitudinal information on contribution points accumulated, which we could use to calculate incomes over full life cycles at least roughly,[8] the DEAS income data is only available for the point in time at which the data was collected. We thus decided not to include income data in the matching.

[7] Due to the problems associated with categorical matching variables we made some effort to convert the categorical variables into continuous ones by estimating propensities using logit and probit models. Unfortunately, this attempt did not amend the matching results, so that we do not present the results here.

[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 that year (FDZ-RV 2008). Additional contribution points may be given during certain contribution periods.

Employment Duration

To provide information on periods of employment, we include the total duration of employment by subtracting the year a person was employed for the first time (start of employment) from the year that person stopped working (end of employment). In the DEAS data the start of employment is measured by a direct question asking respondents to indicate the year they first started regular employment. End of employment data is collected via a similar direct question as to when the respondent's employment ended. For individuals still in employment we took the year of the survey (2008) as the end point. In the VSKT data, the beginning and end of employment can be taken from the longitudinal information contained in the accounts. If the effective status at the time of data collection is employment, we take the collection year of the VSKT (2002, 2005 or 2007) as the end point, just as we do for the DEAS data.

Table 2: Overview of Matching Variables

variable	*scale*	*function*
existence of employment gaps	yes/no	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: illness/rehabilitation	yes/no	matching variable
duration of employment gaps	number of years	matching variable
duration of employment	number of years	matching variable
retirement age	age in years	matching variable
unemployment duration (currently unemployed only)	number of years	matching variable
invalidity pension duration (invalidity pensioners only)	number of years	matching variable
children (women only)	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

Employment Gaps

Both data sources offer information on gaps in employment or times where employment is interrupted. In DEAS respondents are asked if their employment has ever been interrupted for more than 6 months. If so, they are asked to give the information for how long their employment has been disrupted and for what reasons. Given the available information on periods of non-employment in the VSKT accounts, we incorporate this information if individuals have gaps in employment due any of five different reasons: (1) parental leave, (2) military or civilian service, (3) studies/further education, (4) unemployment, and (5) illness/rehabilitation. Whereas the pension data also includes information on the duration of the various gaps, in DEAS we have no such details, so for the matching we can only use the information on whether such an interruption took place or not, but not how long it lasted. What we do have from both data sources, however, is information on the total duration of gaps in employment. We therefore consider this information for the matching procedure too.

Retirement Age and Status of Retirement

For old age retirees, we also take age at retirement into account. We can get this information from both DEAS and VSKT. For people not yet retired, this 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 retirement status into account. We use this information (on whether subjects are already in retirement or not) as a slicing variable. Using slicing variables has the effect that matches are made within strata or 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 will also have to be non-retired.

Unemployment Duration (for Unemployed) and Invalidity Pension Duration (for Invalidity Pensioners)

In addition, for individuals who are currently unemployed or in receipt of an invalidity pension, the duration of their status to date is calculated and taken into account. For people currently having neither of these statuses, this 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 not available for men, because childcare periods are normally credited in the women's accounts only. So for all men in DEAS and VSKT the value for children is set to zero. While the maximum number of registered children is limited to ten in the pension data, 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 (which occurs only very seldom) and for better comparability all numbers of children above ten are lowered to ten.

Year of Birth, Cohort, Gender, and Region

We also take socio-demographic information on the year of birth, cohort, gender and region into account for the matching. Year of birth, which is provided in

[9] However, although we use these approximations for those not already retired for the matching, in the tables and figures relating to retirement age, we only include those who are already retired.

comparable form in both datasets, is incorporated in the matching procedure as a calendar year. However, the incorporation of year of birth as a matching variable does not necessarily mean that only people with the same birth year can be chosen as matching partners, but that people with similar though not identical years of birth can be so matched as well. One of our aims in the matching is to provide a dataset for cohort analyses. To avoid people from one cohort getting a matching partner from another cohort, which would obviously weaken the validity of our cohort analyses, we also take the birth cohorts into account and use them as a slicing variable, a decision which has the consequence that matching will be done only within the three cohorts, so 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. We did the same for region (East or West Germany), on grounds of the separate histories of the former German States of the FRG and the GDR and the still persisting differences in living conditions, life courses, pension calculations and benefits in East and West Germany. This regional information is available in both datasets. However, the logic of how this information is structured differs: In the German Aging Survey (DEAS) we use the statement of where a participant principally 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 their current region of residence (East or West). In the pension data, however, East-West affiliation is gleaned from the percentage of contribution points acquired from East or West Germany. If more than half of such points come from East Germany, a person is assigned to the East, and otherwise to the West. This means that while in the DEAS we apply a residential criterion to the region variable; in the VSKT an earnings-based concept is used. However, though these concepts clearly differ, because of the relatively low levels of labor mobility and job-related commuting that occurs in the data, the overlap between the two concepts should be very substantial.

Since the metrics used for the variables differs to some extent between the two sets of data, we need to verify whether variables in each dataset in reality measure the same thing. Thus one can compare the distributions of the selected matching variables to get a feel for the suitability of the matching variables. In Table 3 we can see that the means and percentages of some of the variables fit quite well, but that others do not.

We can observe a moderate difference in 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 divergent (about 5 years in the DEAS and 8 years in the VSKT).

In addition, one should also consider that the proportion of people for whom we measured an employment gap differs strongly between the two datasets. While 42 percent of the DEAS respondents have had an interruption in employment of more than six months, in the VSKT the proportion is about 75 percent. This difference may be at least partly an effect of the method of data collection. While the VSKT is data created through an administrative process, respondents to the DEAS were asked retrospectively if they ever had interrupted their employment for a period of more than six months. It is known from the literature (e.g. Janson 1990; Middendorff 2000) that retrospective statements are prone to recollection errors. Recollection tends to disimprove with the time span involved, and the pace of this disimprovement differs strongly depending on the matter being recollected (Janson 1990, 104). Thus, it could be that gaps in employment are sometimes misremembered, especially if they happened a long time ago. In addition, interruptions in employment might be perceived as having a stigma attached to them and therefore not reported correctly in all cases. This line of thinking is supported by the fact that most discrepancies are in episodes of unemployment, which many people see as being socially undesirable and which may lead to them being underreported. We see discrepancies for the other types of gap too; and for all types of gap, the percentages are always higher in the VSKT data. For individuals currently unemployed or receiving invalidity pensions, the duration of their current statuses tends to be longer in the DEAS data.

Table 3: Description of Matching Variables in the Data Sources

variable	*DEAS*			*VSKT*		
	%	sd	n	%	sd	n
employment gaps	0.42	0.49	3,997	0.75	0.43	148,110
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: illness/rehabilitation	0.02	0.15	3,997	0.06	0.24	148.110
	mean	sd	n	mean	sd	n
duration of employment gaps	5.00	5.46	1,667	7.86	7.59	110,851
duration of employment	32.70	10.50	3,997	28,69	12.46	148,110
retirement age (retirees only)	61.39	3.17	1,411	61.71	3.73	39.350
unemployment duration (currently unemployed only)	5.14	5.14	220	1.98	2.62	9,382
invalidity pension duration (invalidity pensioners only)	7.61	7.39	135	5.81	5.77	4,802
children (women only)	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.

There is also a difference relating to year of birth, stemming from the fact that the older cohorts are slightly overrepresented in the DEAS, as we have already pointed out. Mean retirement age shows a good correspondence, being between 61 and 62 years of age (for those who are already retired) in both data sources. The mean number of children per woman in both is quite similar in the two datasets, though it is slightly higher in the DEAS. This may 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 others for which this is not the case. This may become problematic for the quality of the matching results. If, for instance, we underestimate the proportion of people who experienced an episode of unemployment in one data source, this may lead us in the matching procedure to combine persons who incorrectly report no unemployment episode with those who have in reality never experienced such an episode. However, we do not know for certain the reasons for the lower levels of unemployment gaps and other interruptions of employment in the DEAS. Therefore it is not possible to distinguish whether employment interruptions are being underestimated or if there is some other reason for the differences. We therefore take all of the described variables into account in our matching, and then consider the differences in distributions later, while interpreting 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 vector (Gu and Rosenbaum 1993; Rubin 1980; Zhao 2004).[10] Mahalanobis distance (Mahalanobis 1936) is based on the correlations between variables and differs from Euclidean distance in that it is scale-invariant – i.e. its results are independent of the scale of the measurements made – and takes the quality of correlations between variables into account, so that strongly correlated matching variables do not enter the computation of the distance function with the same weight as weakly correlated ones.

Assuming that the (weighted) DEAS sample is representative of the German population in the relevant age groups, we match the VSKT information to DEAS data, so that 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 the statistical matching partner or 'donor'.

For the purposes of the matching, we slice the data by region, gender, cohort, and retirement status, taking into account regional, cohort and gender

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

differences in life courses and pension entitlements, as well as the differing situations of retirees and non-retirees. This means matching can only take place within these defined groups: Thus 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, we arrive at 24 slices arithmetically (3 cohorts * 2 regions * 2 genders * 2 retirement states). However, since there are no retirees within the third cohort, the matching procedure is completed in 20 slices. We used the matching variables we described in 3.1.2. The matching calculations are executed for all 3,997 persons from DEAS, and for each observation in the DEAS an observation from the VSKT is chosen as matching partner or donor.

4. Results of the Matching

4.1. Description of the Combined Data

The result of the matching procedure is a dataset that includes information from both sources covering 3,997 cases: i.e. each record in our DEAS data obtained a record from the VSKT as matching partner. However, the data includes some VSKT cases that were not matched uniquely. Actually only 58.1 percent of the chosen donors were matched uniquely; 19.5 percent were matched twice, 9.5 percent three times, and nearly 13 percent more often than three times. This relatively large amount of multiply matched cases may be problematic as it has the effect of reducing the variance in the variables coming from the donor data. However, attempting to avoid multiple matching would probably result in a deterioration in the matching fit, because if one did so for many matches an observation would be chosen other than the one with the smallest distance.

Due to the use of the slicing variables – gender, region, cohort, and retirement status – the distribution of these characteristics in the resulting data is identical, regardless of whether we consider the information from DEAS or VSKT.[11] In relation to the matching variables there is a sound correspondence between the information from DEAS and VSKT on an individual level. For all metric variables there are outstandingly high correlations (r=0.92-0.98). If we look at the means and percentages shown in Table 4, we can see that, due to the adjustment made in the VSKT to match the DEAS sample, the new distributions differ substantially from the original VSKT distributions. As we can see, distributions are now much more similar than they were in the unmatched data. The difference in the duration of employment has dropped from 4 percentage

[11] The distributions of the slicing variables are now identical to the figures in the DEAS column of Table 1.

points to 2.5, and the percentages of people who have experienced gaps in employment now agree: we now get the information from both sources that 42 percent have experienced an interruption of employment at some time. The proportions of the different types of gaps from the VSKT are now aligned with the DEAS information, as are numbers of children and year of birth information. However, for retirement age, invalidity pension duration, duration of employment gaps, and unemployment duration we can still see some differences.

Table 4: Description of the Matching Variables in the Combined Data Set

variable	*information from DEAS*			*information from VSKT*		
	%	sd	n	%	sd	n
existence of employment gaps	0.42	0.49	3,997	0.42	0.49	3,997
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: illness/ rehabilitation	0.02	0.15	3,997	0.02	0.15	3,997
	mean	sd	n	mean	sd	n
duration of employment gaps	5.00	5.46	1,667	9.84	7.67	1,665
duration of employment	32.70	10.50	3,997	30,24	11.35	3,997
retirement age (retirees only)	61.39	3.17	1,411	61.89	2.96	1,411
unemployment duration (currently unemployed only)	5.14	5.14	220	2.19	2.61	248
invalidity pension duration (invalidity pensioners only)	7.61	7.39	135	5.01	5.28	148
children (women only)	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.

What we can observe is that we get a better concordance for the dichotomous variables in general than we do for the metric variables. This is plausible, considering how distances are calculated during the matching procedure. With a dichotomous variable there are only two possible states in terms of similarity: either the two cases are maximally similar (if both have the same value for the relevant variable) or maximally dissimilar (if they do not have the same value for the relevant variable). Thus the probability of choosing a matching partner with a different value in a dichotomous variable is relatively low. In contrast, there are far more variants in terms of similarity for a metric variable. Thus, it is much more likely for a metric variable that the partner chosen for it is similar, but not absolutely identical, to it.[12]

4.2. Validation of the Matching Quality

To assess matching quality, we compare the retirement pensions from the statutory pension scheme that we can take from both data sources for the matched cases as an external criterion for matching quality. We look at pensions amounts from both data sources for those already in receipt of a retirement pension. We did not include pensions amounts as a matching variable because it is only available for a sub-group: in the DEAS the amount of retirement pension from the statutory pension assurance is part of the drop-off questionnaire, with the result that this information is only available for retired participants who did not only take part in the face-to-face interview, but additionally filled in the questionnaire. For this group, we are able to use the retirement pension as an external criterion for matching quality. Whereas in the DEAS the respondents were asked to declare their income from old age provision, in the pension data the amount of retirement pension received is not given directly. However, one can approximate the amount from the sums of individual contribution points (‘*persönliche Entgeltpunkte*’ – PSEGPT) from East and West Germany and the pension values of the relevant year.[13] For the present approximation of retirement 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 € 23.34 and for West Germany it is € 26.56 (DRV, 2008, 11). We therefore use the following equation to calculate the retirement pension amounts for those already in old age retirement only:

$$\boxed{\begin{aligned} &\textit{retirement pension amount} \\ &\quad = (\textit{PSEGTP}_{\textit{East}} * \textit{pension value}_{\textit{East2008}}) \\ &\quad + (\textit{PSEGPT}_{\textit{West}} * \textit{pension value}_{\textit{West2008}}) \end{aligned}}$$

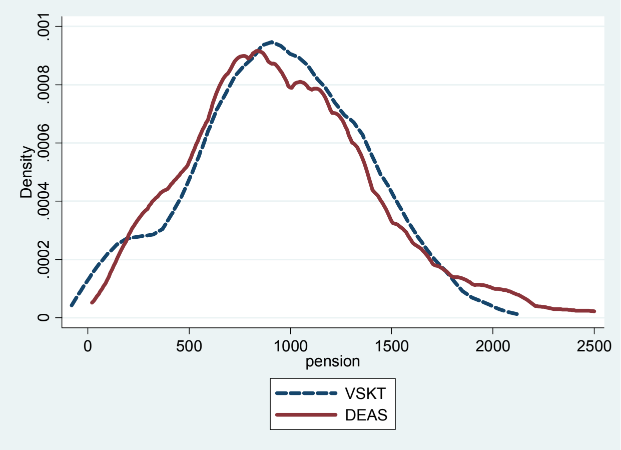
In the following we compare the DEAS and VSKT based retirement pension amounts of the retirees for whom this information is available. Given the in-

[12] As pointed out before, this may have the consequence that our matching is more strongly driven by the binary variables than by the metric ones.

[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 the case of old age retirement it is valued at 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 will reduce the factor, while late retirement will increase it (Rasner et al. 2007, 24).

formation from DEAS, the mean pension amount for those 845 retirees for whom we have information on pension amounts both from DEAS and VSKT is € 1073.8 (sd=619.2). In contrast, the average pension amount calculated from the VSKT, at € 950.8 (sd=414.8), is somewhat lower. This difference of € 123.0 is at least partly the result of a few relatively high pensions in the DEAS. If we exclude the cases of pensions in DEAS above 2500 € (n=20), the average values for the remaining 825 retirees become € 1014.1 (sd=450.7) in DEAS and € 945.5 (sd=413.8) in VSKT, giving a mean difference of 68.6 €. However, this minor difference is somewhat misleading, as positive and negative values cancel each other out. It is therefore more informative to look at the absolute difference values, which are € 404.7 (sd=474.4) for all retirees and € 357.1 € (sd=313.2) for those whose pensions do not rise above € 2500.[14]

Figure 1: Kernel Density of Retirement Pensions



Source: DEAS 2008 and VSKT-LAW 2002-2005-2007 (combined data), own calculations, n=825, pensions less than or equal to € 2,500.

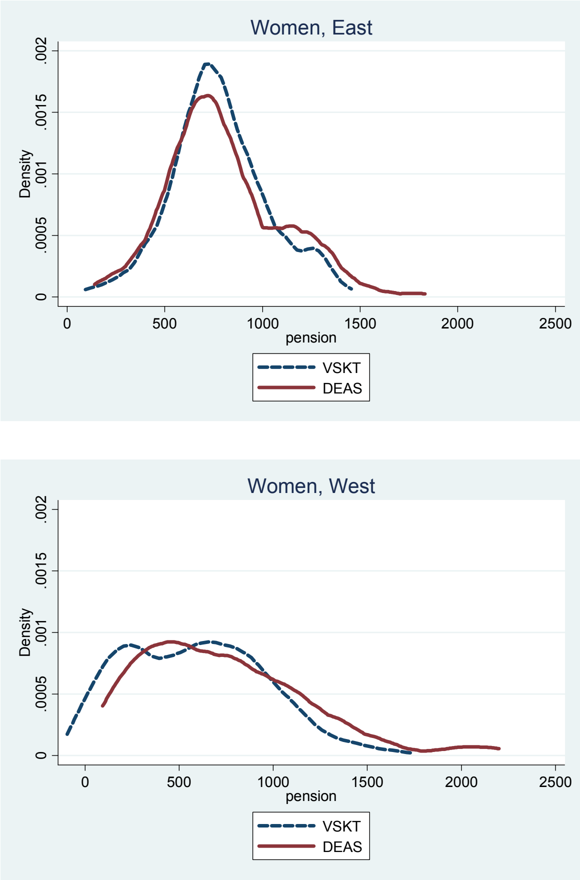
Figure 1 gives us an impression of the distribution of the retirement pensions as determined from DEAS and VSKT data via the kernel density function. All in all the distributions are quite similar. As we learn from Figure 2, which shows the densities of the pension amounts for different groups,[15] the distributions of

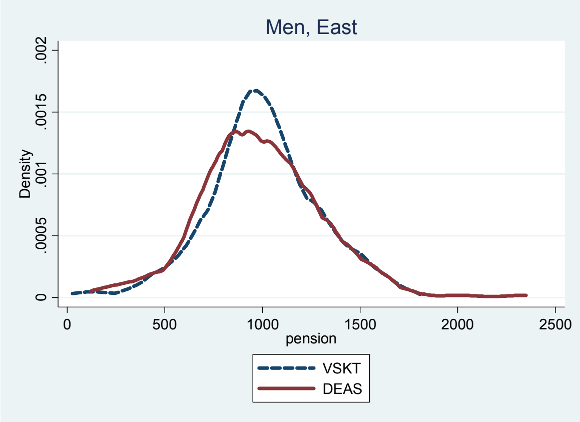
[14] For the further inspection we drop pensions above € 2,500 € because it is very unusual to achieve such high pension amounts from payments of the statutory pension only. It is hence safe to assume that at least some of the high pension amounts from DEAS do not solely reflect income from the statutory pension insurance, but a combination from different income sources.

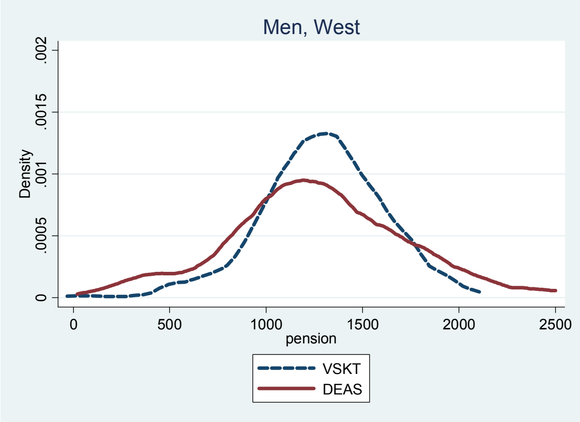
[15] We do not distinguish between cohorts on pension amounts because most of the retirees in our sample belong to the eldest cohort.

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. Pensions for West German women (whether from the DEAS or from the VSKT data) are much lower than for East German women, which is a result of different employment histories and is consistent to current findings (e.g. Himmelreicher and Frommert 2006).

Figure 2: Kernel Density of Retirement Pensions for Different Groups



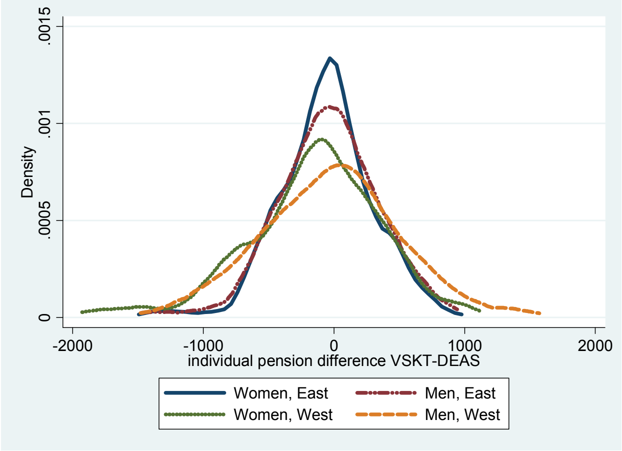




Source: DEAS 2008 and VSKT-LAW 2002-2005-2007 (combined data), own calculations, n=825, pensions less than or equal to € 2,500.

Figure 3 shows the distributions of individual differences between the amounts of pension measured by DEAS and VSKT for men and women from East and West Germany. Again, we can see that the best fit is for women from East Germany. Here the differences in comparison to the other groups are most clustered around zero with a mean absolute difference of € 268.9 (sd=238.5). For East German men as well the accordance is somewhat high, with an absolute difference of € 298.1 (sd=249.3). The worst fit we have is for West German men with a mean absolute difference of € 417.1 (sd=333.8), followed by women from West Germany with an absolute difference of € 402.6 (sd=361.5).

Figure 3: Kernel Density of Individual Differences in Retirement Pensions for the Different Groups



Source: DEAS 2008 and VSKT-LAW 2002-2005-2007 (combined data), own calculations, n=825, pensions less than or equal to € 2,500.

The relatively large differences for West Germany might be an effect of a wider range of incomes over citizens' life course, which we did not take into account in the matching process. Lifetime income is the primary determinant for the amount of pension received. Any lack of the income information can weaken our matching result especially for men, for whom variance is low due to more continuous working careers, with few employment interruptions none at all. Furthermore, we should remember that in our original data sources we saw relatively large differences in numbers of gaps in employment (especially due to unemployment) which, if they result from underestimation in occurring one data source, may weaken the individual fit of pension amounts.

5. Discussion

In this paper we have described the statistical matching of the German Aging Survey (DEAS) with administrative pension insurance data, the Sample of Active Pension Accounts (VSKT). The matching procedure consists of a number of steps, including the definition of the sample population, the selection of suitable matching variables, the matching process itself, and the assessment of matching quality. This last step is very important in verifying the usefulness of the whole process and to assess whether the resulting dataset can be used for the analyses it is intended for.

The results of our inspection with regard to the quality of matching are somewhat mixed. We compared the amounts of pensions taken from both data sources as an external criterion for 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. However, for retirees from West Germany, especially men, there were some larger differences, probably due to the absence of the income information in the matching procedure. So, in order to generate a matched data source for use as an adequate basis for analyzing provision for old age, it would seem valuable to include income data. However, as we have mentioned before, information on income is only available in the DEAS for a particular point in time, so it is likely to be difficult to use this information to improve our matching.

Although there may be limitations on the use of this combined data to analyze individual old age incomes due to some differences between the individual pension amounts coming from the two data sources, the combined dataset may yet prove useful for analyses of old age incomes on an aggregate level, as it offers us the general ability to widen 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 income from the statutory pension topped up with other such income forms as private savings or occupational pensions within a variety of subgroups. In addition to this, the combined data allows us to carry out further analyses that would not be possible with only one of the data sources alone: for instance, analyses regarding the influence of employment history (from the Sample of Active Pension Accounts) on the retirement plans and expectations (from the German Aging Survey) of baby boomers.

Besides the practical benefits of the matching for our further analyses on life courses and old age incomes, one of the things we have learned from the procedure is that for statistical matching of survey and administrative data it is very important at the outset to have a good basis of valid matching variables measured in a comparable way. Furthermore, one has to check if the existing matching variables are suited to 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 – then the absence of a longitudinal income variable can affect the quality of the results. However, as we have seen, even if the number of matching variables is limited, this does not necessarily imply low quality results. Taking these considerations into account, statistical matching of survey and administrative data would seem an appropriate and valuable way to extend the options for analysis of the relevant data sources, and hence to close a number of research gaps.

References

Bothfeld, S., U. Klammer, C. Klenner, S. Leiber, A. Thiel, and A. Ziegler, eds. 2005. *WSI FrauenDatenReport*. Berlin: edition sigma.

Dingeldey, I., and S. Reuter. 2003. Beschäftigungseffekte der neuen Verflechtung zwischen Familien- und Arbeitsmarktpolitik. *WSI Mitteilungen* 11: 659-64.

D'Orazio, M., M. Di Zio, and M. Scanu. 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., and M. M. Grabka. 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", June, 29-30, Luxembourg.

Gu, X. S., P. R. Rosenbaum. 1993. Comparison of Multivariate Matching Methods: Structures, Distances, and Algorithms. *Journal of Computational and Graphical Statistics* 2 (4): 405-20.

Heckman, J. J. 1990. Varieties of selection bias. *American Economic Review* 80: 313-18.

Himmelreicher, R. K., and D. Frommert. 2006. Gibt es Hinweise auf zunehmende Ungleichheit der Alterseinkünfte und zunehmende Altersarmut? *DIW Vierteljahreshefte zur Wirtschaftsforschung* 75 (1): 108-30.

Himmelreicher, R. K., and C. Schröder. 2010. Vorüberlegungen zur statistischen Verknüpfung von Querschnitts-Surveydaten mit prozessproduzierten Längsschnittdaten: EVS und VSKT. *Deutsche Rentenversicherung* 2: 208-16.

Himmelreicher, R. K., and M. Stegmann. 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: 647-60.

Janson, C. G. 1990. Retrospective data, undesirable behavior, and the longitudinal perspective. In *Data quality in longitudinal research*, ed. Magnusson D. and Bergman L. R., 100-21.

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., and T. Masterson. 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., R. Müller, and K. F. Schumann, eds. 2001. *Institutionen und Lebensläufe im Wandel. Institutionelle Regulierungen von Lebensläufen*. Weinheim: Beltz Joventa.

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., S. Wurm, and C. Tesch-Römer, eds. 2010. *Altern im Wandel. Befunde des Deutschen Alterssurveys* (DEAS). Stuttgart: Kohlhammer.

Rasner, A., J. R. Frick, and M. M. Grabka. 2011. Extending the Empirical Basis for Wealth Inequality Research Using Statistical Matching of Administrative and Survey Data. *SOEPpapers* 359. Berlin: DIW.

Rasner, A., R. K. Himmelreicher, M. M. Grabka, and J. R. Frick. 2007. Best of both worlds: preparatory steps in matching survey data with administrative pension records: the case of the German Socio-Economic Panel and the Scientific Use File Completed Insurance Biographies 2004. *SOEPpapers* 70. Berlin: DIW.

Rosenbaum, P. R., and D. B. Rubin. 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: 293-8.

Simonson, J., A. Motel-Klingebiel, and K. Kowalska. 2010. Alterssicherung und Alterseinkünfte im Deutschen Alterssurvey (DEAS). *Deutsche Rentenversicherung* 2: 301-13.

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

Steiner, V., and K. Wrohlich. 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.