Quality & Quantity (2022) 56:4747–4774
https://doi.org/10.1007/s11135-022-01333-9

Assessing dissimilarity of employment history information from survey and administrative data using sequence analysis techniques

Babette Bühler[1] · **Katja Möhring**[2] · **Andreas P. Weiland**[2]

Accepted: 24 January 2022 / Published online: 16 February 2022

Abstract

Life course data is frequently gathered either using retrospective surveys or linking records with administrative data. Yet, each strategy has specific advantages and disadvantages. We study the consistency between both types of data sources and reasons for mismatch using the linked data set SHARE-RV, which combines retrospective life history data from the Survey of Health, Ageing and Retirement in Europe (SHARE) with respondents' administrative data from German pension insurance records ($N = 1679$). Utilizing sequence analysis techniques with Hamming distance, Optimal Matching and OMspell as matching algorithms, we examine mismatches between survey and administrative data covering detailed, 30-year employment histories, and analyze how inconsistencies are associated with life-course characteristics, demographic and socio-economic factors. Our results show that life-course complexity and spells of atypical employment are associated with more mismatches. Furthermore, gender differences are pronounced and appear to be sensitive to the applied matching algorithm.

Keywords Life course · SHARE · Hamming distance · Optimal matching · OMspell · Data linkage

✉ Katja Möhring
moehring@uni-mannheim.de

Babette Bühler
babette.buehler@uni-tuebingen.de

Andreas P. Weiland
andreas.weiland@uni-mannheim.de

[1] Hector Research Institute of Education Sciences and Psychology, University of Tübingen, Europastraße 6, 72072 Tübingen, Germany

[2] School of Social Sciences, University of Mannheim, A 5, 6, 68159 Mannheim, Germany

1 Introduction

As the popularity of life-course analysis increases, so does the need for detailed and reliable life trajectory data. Two frequently used sources of such data are retrospective surveys and administrative records. The literature describes the advantages and disadvantages of both strategies in great detail. The most important drawbacks of retrospectively collected data are that recall errors might occur (Solga 2001), respondents might oversimplify their working career, and underreport certain states such as unemployment (Manzoni et al. 2010). In addition, respondents may perceive or define their employment or social situation differently than it is recorded in the administrative data. Yet, respondents must consent to having their administrative data linked to their survey responses; some may refuse, resulting in a biased sample (Korbmacher and Schröder 2013; Jenkins et al. 2006). Since in many cases only one source might be available, there is a growing interest in comparing data from both sources (Huber and Schmucker 2009; Wahrendorf et al. 2019). Inconsistencies between the two sources might point to problems in the survey of retrospective information, especially if these discrepancies are aligned with respondents' socio-economic characteristics, for example if the quality of retrospective data is significantly higher or lower for specific social groups or at particular stages of the life course. Comparing the data sources might also help reveal blind spots in administrative records.

We use the SHARE-RV data set, which combines retrospective survey data gathered in the SHARELIFE interviews from the Survey of Health, Ageing and Retirement in Europe (SHARE) program (Börsch-Supan 2019) with administrative data from the German pension insurance fund (SHARE-RV VSKT) (Forschungsdatenzentrum der Rentenversicherung, Max-Planck-Institut für Sozialrecht und Sozialpolitik 2019) to analyze differences between the data sources. The data set contains information on 1679 respondents, covering 30-year employment trajectories for West and East Germany. We study the agreement between the two sources throughout individuals' lives in order to determine whether certain stages in the life course are prone to inconsistencies. We first compare individuals' employment trajectories using sequence analysis with different distance algorithms. We then apply regression analysis using calculated distance values as dependent variables and examine how socio-economic factors and characteristics of the life histories are related to inconsistencies.

We go beyond previous research as our linked data set covers a longer time span, includes more complex sequences, and is based on a representative sample of both East and West Germany. The latter has been characterized by a free market economy, supporting a male breadwinner model, the former by a centrally planned socialist economy with a high degree of labor market engagement of men and women, until the profound system transformation following German reunification in 1990 (Trappe et al. 2015). These divergent institutional frameworks are particularly interesting in relation to life-course trajectories. The data covers the ages of 21 to 50 for all respondents, which includes school-to-work as well as retirement transitions based on a precise definition of employment status with nine categories. The literature on sequence analysis discusses a range of distance measures that quantify the degree of divergence between two sequences (Studer and Ritschard 2016). We apply Hamming distance (Hamming 1950), a measure based on common attribute quantity, which is very sensitive to timing. We also employ Optimal Matching (OM) (Abbott and Forrest 1986), a more flexible edit dissimilarity measure. Alongside those very common distance measures used in previous research (Huber and Schmucker 2009; Wahrendorf et al. 2019), we apply the more context-sensitive Optimal Matching between sequences of spells (OMspell), proposed by Studer and Ritschard (2016), which accounts for the length of spells.

The article proceeds as follows. Next, we review previous studies that compare survey and administrative data, and discuss possible sources of inaccuracy. We then describe the three sequence analysis algorithms that we use for the comparison in detail. The following section examines our data and analysis strategy. We then present and discuss the descriptive and multivariate results on the determinants of dissimilarities in both data sources.

2 Background

2.1 Previous research

There has long been a general interest in matching survey results with administrative data. These efforts have focused on increasing the accuracy of interviewee characteristics such as educational attainment (e.g. Adriaans et al. 2020) and income (e.g. Kreiner et al. 2015; Valet et al. 2019), counteracting the recall bias of isolated life-course events such as retirement (Korbmacher 2014), or decreasing the measurement error associated with entire dimensions of the life course such as union histories (Kreyenfeld and Bastin 2016) and employment biographies (e.g. Huber and Schmucker 2009; Kreuter et al. 2010; Wahrendorf et al. 2019).

For instance, Huber and Schmucker (2009) matched a year's worth of survey and administrative employment data on a monthly basis in order to compare employment biographies and found high levels of agreement between data originating from the two sources: data from only 5 percent of respondents featured deviations. The authors conducted a multivariate analysis, which demonstrated that a higher number of state transitions, for instance between employment and unemployment, increases the probability of a mismatch between the two types of sources. Age, education and income also have a significant effect. The results show that the probability of mismatches decreases until the age of 45 and increases afterwards. Moreover, having no educational degree or belonging to the middle-income category were linked to a smaller probability of diverging sequences. Other variables like sex, nationality and further training had no significant effect. However, it is notable that the regarded biographic episode of one year was rather short and not far in the past at the time of the retrospective survey.

Wahrendorf et al. (2019) examined the differences between self-reported employment histories from the Heinz Nixdorf Recall Study (executed in three German Cities in the Ruhr area) and administrative data from the German Institute for Employment Research; they note that the size and composition of the two data sources are different. They state high and constant levels of agreement over time with an average of only 4 years of diverging information per person for an observation period of 36 years, which corresponds to a median level of agreement of 89 percent. They find that self-reported employment sequences are less complex than those originating from administrative data, which confirms that respondents tend to oversimplify their biographies during surveys (Manzoni et al. 2010). Using sequence analysis, Wahrendorf et al. (2019) find larger differences for women and people working in the tertiary sector. These effects vanish in multivariate analyses that take into account transitions and the years spent in part-time work and non-employment. This leads the authors to conclude that women, who often work in the tertiary sector, have more complex sequences with frequent status changes and more part-time work or non-employment

(Widmer and Ritschard 2009) and therefore display greater disparities in the different types of data sources. However, the classification into only three employment states is rather rough; for instance, Wahrendorf et al. (2019) combine unemployment and childcare into a single category of non-employment, which prevents a detailed look at gender-based differences.

2.2 Sources of mismatch: misrepresentation and non-response

Why might survey and administrative data on the life history of the same person yield different information? Possible sources of mismatch are reporting errors in survey or administrative data, as well as diverging measurement concepts and scopes in the two types of data. For instance, survey respondents may provide inaccurate information or no information at all. Likewise, recall bias is a central source of inaccuracy in retrospective life-course interviews: respondents may not recall sections of their life history or may misremember the nature of life-course episodes, their time frame or their temporal order (e.g. Korbmacher 2014; Schröder 2011; Wagner and Philip 2019). Alternatively, they may withhold or misrepresent sensitive aspects of their past such as (un-)employment spells, earnings or partnership history due to social desirability bias (Krumpal 2013; Valet et al. 2019). The survey design, interview mode and interviewer characteristics may exacerbate or mitigate the likelihood and severity of these biases (Kreuter et al., 2008; Kühne 2018; West and Blom 2017).

Administrative data are generally considered to be more robust to misrepresentations and non-response than surveys (Kreuter et al. 2010) since they are gathered within the framework of administrative processes using a systematic approach. However, inaccuracies may occur through measurement error, as well as missing documentation and diverging measurement constructs associated with the scope and purpose of the data (Groen 2012). Depending on the type of administrative data and how it is collected and processed, the nature or exact timing of an event may be inaccurately reported, for instance in employers' social security notifications, although this is unlikely (Abowd and Stinson 2011; Kreuter et al. 2010). The convergence between survey and administrative data is contingent on the (dis-)similarity and compatibility of measurement concepts in both sources, such as how they define employment states. Further, legislative and institutional changes may produce mismatches through the (*ex post*) redefinition of measurement concepts in administrative sources (Mika 2009). Finally, depending on their scope and purpose, administrative data may lack coverage of certain individual characteristics or parts of the survey population, thus effectively producing non-responses for these items (Korbmacher and Czaplicki 2013; Sakshaug et al. 2017).

2.3 Dissimilarity measures in sequence analysis

Our analysis depicts respondents' life trajectories as sequences. A sequence is defined as an ordered list of items (Brzinsky-Fay et al. 2006)—in this case units of time (e.g. 1 year)—categorized as different states (e.g. employment, education, etc.). Several consecutive items that share the same state are called an episode or spell. Sequence analysis uses distance measures to quantify the level of inconsistency between two sequences. Calculating these distance measures is often seen as a first step. The resulting distance matrix, which contains the distances between all given sequences, is then used to cluster the sequences, for example (Brzinsky-Fay and Kohler 2010). In this analysis we only calculate distances for each person between the two sequences (originating from the survey and administrative data) because we are interested in the level of inconsistency between the two *sources*—not between the *respondents*. In a later step we use this measure of dissimilarity per person to investigate the level of dissimilarity between different groups, and to examine the determinants influencing the degree of divergence between the two sources. We employ three different distance measures for the pairwise comparison of individual sequences from administrative and survey data.

First, we calculate the Hamming distance, which depicts the difference between two sequences by simply counting the number of non-matching items (Hamming 1950; Studer and Ritschard 2016). The resulting distance value denotes the number of years in which self-reported states differ from administrative states. This measure is very sensitive to the correct timing—i.e. the exact year in which a state appears. In our case of retrospective interviewing, it can be difficult for respondents to remember the correct timing of an event, so a measure that is more forgiving in this regard may be appropriate.

The second distance measure we use is OM, which describes the difference in the two sequences as the minimal cost of transforming one into another by allowing and specifying the costs of insertion (inserting an element into a specific position) and deletion (deleting an element from a certain position) which are subsumed under the term indel, as well as substitution (changing one element into another) operations (Abbott and Forrest 1986; Abbott and Hrycak 1990; Studer and Ritschard 2016). We use the standard costs of 2 for each substitution and 1 for each indel operation. Thus, the OM distance allowing for alignment operations takes into account time shifts in the sequences and is less sensitive to correct timing than the Hamming measure.

Although OM is the most widely used distance measure for sequence analysis in the social sciences (Halpin 2010) and has previously been used to examine similar research questions (Huber and Schmucker 2009; Wahrendorf et al. 2019), it has been criticized for being sociologically meaningless (Wu et al. 2000). Halpin (2010) notes that OM was designed to assess discrete-time sequences, yet life trajectory data is continuous in time. He argues that trajectories should be considered sequences of spells. Moreover, previous research has shown very high correlations between Hamming and OM distances for life-course data (Halpin 2010; Wahrendorf et al. 2019). A range of refined measures has been proposed to address some of the critiques of OM (e.g. Elzinga 2003; Hollister 2009; Halpin 2010; Lesnard 2010; Elzinga and Wang 2012).

Third, we use one such refined, context sensitive measure, OMspell, which calculates OM for several consecutive years in the same state, and is therefore sensitive to the duration spent in distinct successive states (Studer and Ritschard 2016). It introduces a correction factor function for the spell length with a weighting factor, the so-called expansion cost. The indel and substitution costs introduced for OM are extended as follows (Studer and Ritschard 2016):

$$c_I^s\left(a_t\right) = c_I(a) + \delta \quad (t - 1)$$

$$\gamma^s\left(a_{t_1}, b_{t_2}\right) = \left\{ \delta \left| \begin{array}{ll} t_1 - t_2| & if\, a = b, \\ \gamma(a, b) + \delta\left(t_1 + t_2 - 2\right) & otherwise \end{array} \right. \right.$$

In this formula, c_I^s denotes the spell indel costs and γ^s the spell substitution cost of a given spell a_t, which depicts state a during t years. δ denotes the expansion cost. Past studies have successfully used this measure for sequence analysis (e.g. Lee et al. 2017; Squires et al. 2017). As before, we use the standard cost of 2 for substitution $\gamma(a,b)$, 1 for indel operations c_I and 0.5 for expansion cost δ.

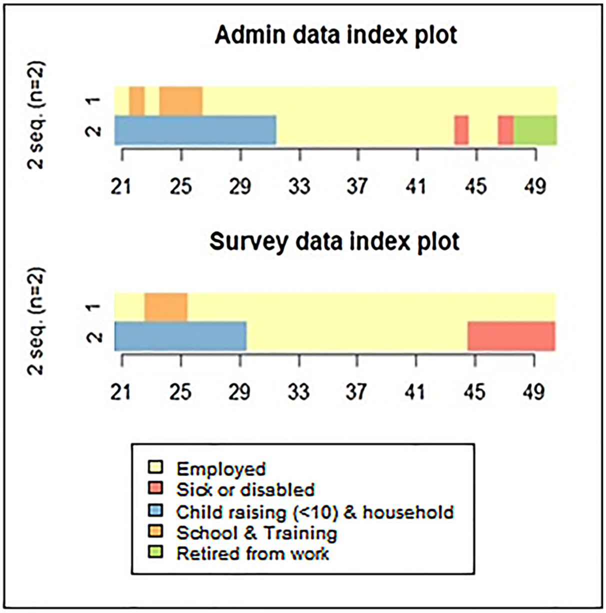
We compare different measures of dissimilarity to assess their ability to capture and emphasize different dimensions of distinction between the two sequences and to investigate which aspects are meaningful in the context of our research question. As stated above, the naive Hamming distance measure is very sensitive to the point in time when the state appears. For example, a respondent may report a state a year earlier than the administrative data due to recall errors. He or she may report the sequence A–B–C, while the administrative data states C–A–B. This sequence has the same Hamming distance value as, for instance, if the respondent had reported a third (totally different) state for that year, such as D–D–D. The information he or she provided in the first case may be inaccurate concerning the exact timing, but from a social science perspective it is more similar to the administrative sequence than the second case (Halpin 2010). OM, however, allows for slight time shifts due to the alignment through indel operations, which result in lower costs and therefore have a smaller distance value.

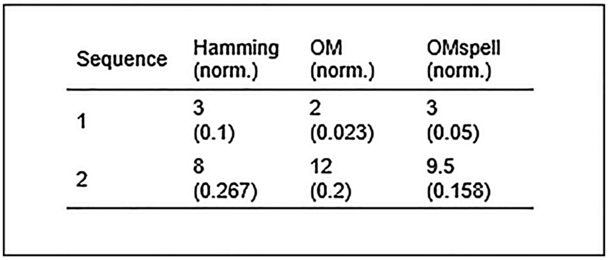
However, OM does not take the context into account. For instance, it makes no distinction between reporting a different item that is part of a two-year episode and misreporting an item that is part of a ten-year spell, which might be less consequential from a sociological point of view (Halpin 2010). OMspell also includes the length of the spells and thus the structure of the sequences. It prefers expanding or compressing existing spells before inserting new spells for alignment. By calculating OM between sequences of spells and therefore taking episodes/spells (consecutive years spent in the same state) as the unit of analysis, it accounts for the continuous character of life trajectory data. Figure 1 depicts two example sequences from the employed data, with their respective Hamming, OM and OMspell distances.

In sequence 1 at the beginning of the observation period, after a short period of employment, some years spent in school or training (ST) can be identified, followed by a very long employment (EM) spell for the rest of the period. The sequences from both sources look similar but are not identical. The differences are found in the first six years. The sequence from the administrative data for the first 6 years is: EM-ST-EM-ST-ST-ST, while the sequence from survey data is: EM-EM-ST-ST-ST-EM. The remaining 24 years are coded EM in both sources. In total, three years at age 22, 23 and 26 differ. The Hamming distance therefore takes a value of 3, normalized to 0.1. OM-based distance, which considers that these three divergences can be resolved by one insertion and one deletion operation, takes a value of 2, normalized to 0.023, which is a much smaller distance than that calculated using the Hamming method. Hamming can take a maximum value of 30, in case that every single year is different, while OM and OMspell have a (theoretical) maximum of 60. For OMspell, spells as a whole count as units for indel and substitutions operations, and substitutions of the same state can be expanded to different lengths, i.e. to include several elements. Therefore, based on OMspell, two compression operations with a cost of 0.5 and two insertion operations with a cost of 1 each are needed to convert the survey into the admin sequence. This results in an OMspell value of 3, normalized to 0.05, which is slightly higher than OM but still considerably lower than the Hamming measurement.

Sequence 2 contains a total of 9 years' difference between the administrative and survey data, resulting in a Hamming distance of 8, normalized to 2.67. In the first half of the observation period, there is a shift in the transition from childcare (CH) to employment (EM), which in the survey data occurs two years earlier. The remaining six differences are found at the end of the period. While the administrative data show several changes between Sick or Disabled (SD), return to Employment and finally to Retirement

Fig. 1 Example Sequences. Source: Authors' own depiction





Sequence	Hamming (norm.)	OM (norm.)	OMspell (norm.)
1	3 (0.1)	2 (0.023)	3 (0.05)
2	8 (0.267)	12 (0.2)	9.5 (0.158)

(RT) status, the survey data show a one-time change to SD. Using the OM approach, the survey data sequence can be transformed into the admin sequence by inserting two CH items and one SD item, and deleting two SD and one EM. These six indel operations have a cost of 6. Furthermore, three SD years have to be substituted with RT years, which costs 3*2. This results in a total OM distance of 12, normalized to 0.2. Calculating the OMspell cost consists of substituting the nine-year CH spell in the survey data with an 11-year episode of the same state. Since this is only an expansion of an existing spell, the cost results from the expansion factor 0.5 times the amount of the difference in spell length, i.e., a cost of 1. A similar calculation applies to the following compressions of the EM spell from 15 to 12 years and the SD spell from six years to one year. Subsequently, a two-year EM spell, a one-year SD spell and a three-year RT spell have to be inserted. The costs are calculated by the insertion cost (1) plus the weighting factor (0.5) times the spell length (3) minus 1, represented as $1+0.5*(3–1)$ for the three-year RT episode. This results in an OMspell distance of 9.5, normalized to 0.158. In this case OMspell is smaller than OM because it favors the expansion or compression of existing spells over the insertion of new spells. OM scores the same regardless of whether the inserted CH items show the same state as previous or following spells. OMspell rewards the fact that the considered sequences have the same spell structure in the first two-thirds—CH, EM, SD.

3 Data and methods

3.1 Data

Our analysis draws on a combination of the retrospective life history data included in the SHARELIFE survey that was part of SHARE Wave 7 (Börsch-Supan 2019; Börsch-Supan et al. 2013) and the VSKT dataset of SHARE-RV with administrative records from the German Pension Insurance fund (DRV) (Forschungsdatenzentrum der Rentenversicherung, Max-Planck-Institut für Sozialrecht und Sozialpolitik 2019). SHARE Wave 7 surveyed 3821 respondents in Germany, 2916 of whom gave their consent for their responses to be linked to their pension data contained in the VSKT.

SHARELIFE collects retrospective life-course data in face-to-face-interviews using a Life History Calendar approach, and covers information for the period 1920 to 2017 (Börsch-Supan 2019). Respondents are asked about biographical details from the time of their schooling up to the time of the interview. In an attempt to minimize recall bias, the Life History Calendar documentation follows a modular approach. The interviewer starts with questions concerning salient life-course events such as the birth years of the respondent's children. These events are recorded in the calendar in real time during the interview and serve as visible anchors for subsequent questions about the respondent's history of partnership and cohabitation. Although the interviews may slightly modify this order, life-course information regarding respondents' socio-economic episodes is by default constructed around these events. Subsequently, the study gathers information on the start and end years of different socio-economic episodes, such as school, training, employment and childcare, from which sequences of yearly states can then be deduced for each respondent (Schröder 2011).

The VSKT includes the DRV's monthly administrative records under the statutory pension scheme for individuals aged 14 to 65 years. As the DRV only records pension-relevant employment situations, work in self-employment and public service are not covered and thus result in 'no information' states in the data. In order to take this into account, the corresponding states in the survey data ('self-employment' and 'public service') were also recoded to 'no information' states. Respondents who never had public-pension-relevant employment do not appear in the data at all. In the same way, childcare and homemaking situations are only covered by pension insurance if they are relevant under pension law. In Germany, child-raising periods are only credited until the child reaches the age of ten and therefore only appear in the VSKT data up to that time. Thus, childcare states in the survey data were converted into 'no information' states after the tenth birthday of the youngest child to prevent inconsistencies arising from different recording procedures when comparing the sequences.

Since a high proportion of 'no information' states limits the meaningfulness of discrepancies between administrative and survey data, we exclude respondents with more than 25 percent 'no information' states in their pension data life history from the analysis. Therefore, our sample underrepresents self-employed persons, civil servants, and homemakers. The share of self-employed persons in Germany (West Germany before reunification) amounts to an average of 10 percent of the working age population between 1960 and 2017, while civil servants represent about 6 percent during this period (Source: Destatis 2019a; own calculations). The average share of labor market inactive persons, i.e. those neither engaging in nor seeking gainful employment (e.g. homemakers), are estimated at about 30 percent of the working age population between 1960 and 2017 (Source: Destatis 2019b; own calculations).

Furthermore, we limit the period under consideration for all respondents to the age range 21 to 50 since not all respondents had reached the age of 65 by the time of the interview and the data quality for the first years of life is significantly worse due to a high share of missing values in the survey data. This results in a dataset containing 30 years of live history data, from age 21 to 50, for 1679 respondents.

Appendix Table 4 shows sample information by selected covariates. The data set consists of almost equal shares of men and women. The vast majority (90.83 percent) of the respondents were aged 49–79 at the time of the interview. Nearly a quarter (22.87 percent) lived in the former East Germany. Education levels were classified using the ISCED scale: low (pre-primary, primary, lower secondary education), middle (upper secondary, post-secondary non-tertiary education) and high (first, second stage of tertiary education). Job sector is based on the job type of the last job stated in the interview: energy and agriculture (agriculture, hunting, forestry, fishing; mining and quarrying), manufacturing (manufacturing; electricity, gas and water supply; construction), services (wholesale and retail trade; hotels and restaurants; transport, storage and communication; financial intermediation; real estate, renting and business activity; public administration and defense; education; health and social work), and other sectors (other, don't know). The total number of pension earning points relates to pension entitlements under the statutory pension scheme and was recorded in the VSKT, while self-perceived health was queried in the SHARE survey.

3.2 Data linkage strategy

We harmonize the sequences from the two different sources in two steps in order to be able to compare them. First, the original states in the different sources have to be mapped to a common catalogue of possible social income situations (see Appendix Table 5 for detailed information on mapping). This process permits a detailed differentiation between nine different states: no information; employed; unemployed; sick or disabled; childcare; school or training; retired; military, civil service, war prisoner or equivalent; and other. The status of 'missing', which only appears in the survey data, is also regarded as a state.

Second, the monthly administrative data must be aggregated to make it comparable to the annual survey data. To do this, we chose a mode-based aggregation and used the status that was held for the longest time, i.e. the most months, in a given year as the annual status. Based on this aggregation method, none of the respondents had an average low aggregation quality (defined as 8 months or less spent in the denoted status per year). Only 12.69 percent of the respondents have a medium aggregation quality (9–11 months) and an overwhelming majority (87.31 percent) have a high-quality aggregation (12 months). This indicates that frequent changes within a year are rather rare and that mode aggregation at the annual level appears to be well justified. Appendix B explains why we favored mode- over rule-based aggregation as used by Huber and Schmucker (2009).

3.3 Analysis strategy

First, we conducted a descriptive analysis using sequence distribution plots and other graphical representations that consider the average time spent in each state, the number of transitions, the within-sequence entropy, the longest episode and the average complexity of the sequences. Gabadinho et al. (2011) defines the complexity index as:

$$\text{Complexity Index} = \sqrt[2]{\left(\frac{\text{num. transitions}}{\text{max . num. transitions}} \times \frac{\text{entropy}}{\text{max . entropy}}\right)}$$

The number of transitions indicates the number of state changes within one sequence and accounts for the complexity caused by the state ordering, while the entropy accounts for complexity stemming from the state distribution in a sequence (Gabadinho et al. 2011). The entropy is zero if a sequence only consists of the same state for all 30 years and takes the maximum value of 1 if a sequence consists of all possible states for the same number of years (Gabadinho et al. 2009).

We next consider the consistency between the two sources over time. This reveals whether certain stages in an individual's life course are more susceptible to inconsistencies than others. Each year is considered individually, and the relative number of matches between the two sources is measured for all respondents. The agreement by state, based on administrative data, is examined to determine whether certain statuses contain more differences than others. Our analysis also accounts for the different—and, over the life course, changing—institutional framework conditions between Eastern and Western Germany.

We then compare the two data sources by applying the three above-mentioned distance measures (Hamming, OM and OMspell). They are calculated for each respondent between the survey and administrative sequences using the TraMineR package in R (Gabadinho et al. 2020). Initially, the calculated distance values are presented and compared bivariately along various dimensions. Linear regression analyses are then conducted to detect potential systematic mismatches for certain groups or sequences in the sample. We include the age at the time of the interview, gender, education, residency in East or West Germany, the sum of pension earning points (as an approximation of income), the job sector of the last job, and self-perceived health as explanatory variables representing possible determinants of the dissimilarity. We also consider sequence-specific variables such as the complexity index and the number of years spent in each possible state based on the administrative data. We control for the quality of the aggregation in all models. We calculate two models for each distance measure. The first model includes all variables related to the respondents' socio-demographic characteristics, and the second adds the variables describing the sequences.

4 Results

4.1 Descriptive comparison of sequences

Figure 2 depicts the state distributions of the resulting sequences, revealing slight but remarkable discrepancies between the survey data plot (left) and the administrative data plot (right). In particular, it highlights the smaller proportion of childcare, retirement, and

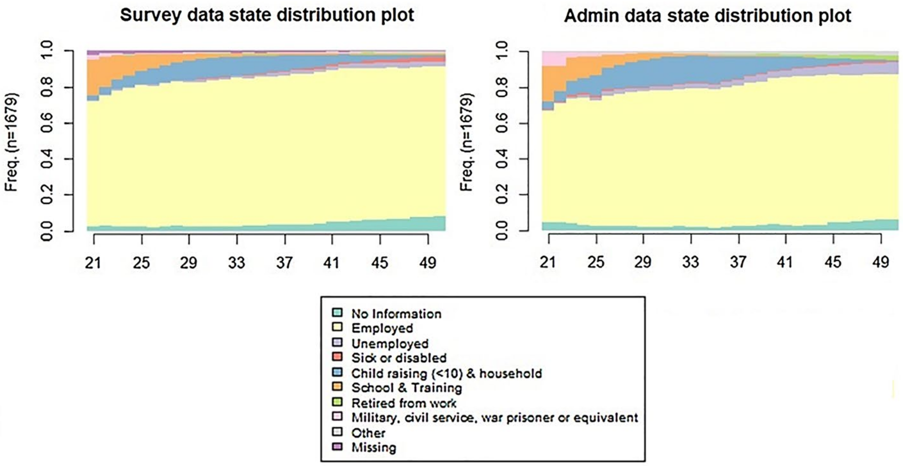


Fig. 2 State distribution plots. Source: Authors' calculations based on SHARE-RV version 7–0-0 and SHARE Wave 7 version 7–0-0

Table 1 Sequence Summary Statistics

	Survey data Mean (Sd)	Admin data Mean (Sd)
Duration (years)		
Employed	24.345 (7.562)	23.185 (6.837)
Retired	0.088 (1.144)	0.215 (1.418)
Unemployed	0.494 (2.017)	0.973 (2.314)
Child raising and homemaking	1.941 (4.707)	2.691 (4.945)
Sick or disabled	0.241 (1.684)	0.208 (0.627)
School and training	1.167 (2.856)	1.075 (2.223)
Military, civil service	0.102 (1.22)	0.243 (0.544)
No information	1.201 (3.796)	0.952 (1.832)
Other	0.124 (1.233)	0.456 (2.28)
Missing	0.297 (1.937)	0 (0)
Complexity index	0.083 (0.083)	0.155 (0.113)
Transitions	1.432 (1.562)	3.270 (2.618)
Entropy	0.149 (0.153)	0.226 (0.166)
Longest episode	23.814 (6.636)	20.154 (7.175)

Source: Authors' calculations based on SHARE-RV version 7–0-0 and SHARE Wave 7 version 7–0-0

military service states and the larger proportion of sickness or disability spells in the survey data compared to the administrative data. The survey data (but not the administrative data) contain missing states.

Table 1 shows further statistics on the created sequences. The survey data indicates that respondents spend an average of 24 out of the 30 years under consideration in employment, while in the administrative data the average amounts to about 23 years. The considerably smaller proportion of years in childcare in the survey data, approximately 1.9 years on average, compared to the administrative data, about 2.7 years, stands out. The average proportion of unemployed years is also significantly lower in the survey data than in the administrative data. Conversely, the survey sequences contain a larger share of 'no information' and years of sickness or disability as well as missing spells. Furthermore, the sequences resulting from the survey data have a considerably lower average complexity index. This indicates that the administrative data represents more complex trajectories, considering the entropy and number of transitions. This is evident when looking at the average number of transitions, namely status changes, per sequence—1.40 for the survey data and 3.27 for the administrative data. The average within-sequence entropy, highlighting the distribution of states, is smaller for survey than for administrative data, which on average contains slightly shorter longest episodes.

Overall, when comparing the survey and administrative data, 14.00 percent of respondents ($n = 235$) do not have a single inconsistent year in the entire time observed. All others had conflicting information in at least one year (5.65 years on average). The most frequent type of mismatch found in the data (16.85 percent of all deviations) is the respondent stating she was employed while the administrative data show a child raising and homemaker state. Another very common error (9.63 percent) was 'no information' in the survey data but employed status in the administrative data. The third most frequent discrepancy (8.65 percent of all deviations) is an employed state in the survey and unemployed in the administrative data.

4.2 Agreement over time

Figure 3 displays the consistency of the data over the life course—the relative frequency of agreement between the survey and administrative data, in total and for the three main states, based on administrative data by year. The solid (dotted) lines denote respondents in Western (Eastern) Germany.

The plot shows a relatively constant total agreement over time of approximately 80 percent of respondents from Western Germany. The first years of the observation period, i.e. the years furthest from the date of interview, show slightly more differences. Although respondents from the East also display less agreement in the early years, their values are

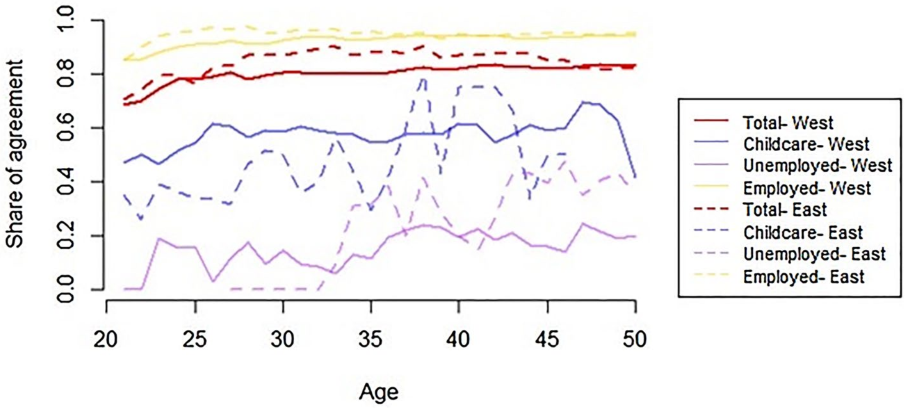


Fig. 3 Agreement over time by state (admin data) and residence. Source: Authors' calculations based on SHARE-RV version 7–0-0 and SHARE Wave 7 version 7–0-0

considerably higher throughout almost the whole observed life span (near 90 percent agreement), while towards the end of the observation period they drop significantly and are even below the level of Western Germany. The agreement over time for administrative employment states is slightly higher than the total agreement for all states for both East and West and follows a very similar pattern, which can be explained by the large overall share of employment states. However, it is notable that the relative consistency of employment spells in Eastern Germany is higher in the first half of the observation period, while in the second half it decreases to the same level as in Western Germany.

Respondents from Eastern Germany demonstrate up to over 95 percent consistency in their employment state between the ages of 25 and 35, which constitutes the highest observed agreement. For most of the observation period, the childcare state agreement for Western Germany is constantly between 50 and 70 percent. Lower relative agreement in this state is evident at the beginning and end of the observation period. The values for Eastern Germany are similar but show particularly high levels of agreement in the second half. However, the strong fluctuations result from relatively low case numbers in the corresponding years. Respondents who are unemployed according to administrative data show a very low level of agreement over time, not exceeding 30 percent for West Germans and 40 percent for East Germans, with high variation.

4.3 Determinants of sequence distances

We first assess the dissimilarity between the survey and administrative data based on the three distance measures. Table 2 shows the calculated normalized distances and their standard deviations by covariate. Hamming has the highest average distance values, followed by OM and OMspell. Regarding the covariates, women on average display greater distances than men in our data. Age at the time of the interview has no clear linear relationship to the distances, while people with high and medium levels of education show slightly smaller distances than those with low education. Respondents in East Germany have smaller average distances between the administrative and survey data. Likewise, people working in the service or other non-manufacturing sectors show greater distances than those working in the manufacturing sector. The lowest distances are found in the energy and agricultural sectors. The higher the accumulated pension points (which serve as a proxy for lifetime earnings), the lower the average distance values. We find no clear pattern for self-perceived health. Although the three distance measures have different average values, the relationships between the groups of covariates and the respective distance measures are very similar.

Table 3 reports the results of the multivariate linear regression analysis using the Hamming distance, OM and OMspell measures as dependent variables. The independent variables were checked for collinearity using the variance inflation factor test. For all three distance measures we first calculated a model containing only the socio-demographic covariates. We then added variables on the characteristics of the sequences in a second model. We controlled for aggregation quality in all models. In the first range of models, which did not control for sequence characteristics, we find that gender, sum of earnings points, high educational level, East Germany, and job sector are significantly related with the level of all three distance measures. East Germans have significantly smaller distances between the survey and administrative data; the higher the sum of the earnings points, the smaller the distance. Those with high levels of education have significantly larger distances

Table 2 Distances between survey and administrative data by covariates

Variables	Hamming Mean (SD)	OM Mean (SD)	OMspell Mean (SD)
Sex			
Male	0.154 (0.207)	0.139 (0.196)	0.119 (0.126)
Female	0.22 (0.206)	0.197 (0.192)	0.18 (0.13)
Age			
49–64 years	0.206 (0.202)	0.183 (0.188)	0.171 (0.128)
65–79 years	0.167 (0.206)	0.15 (0.194)	0.131 (0.13)
80 years or older	0.2 (0.246)	0.185 (0.235)	0.14 (0.142)
Education			
Low	0.252 (0.254)	0.225 (0.236)	0.169 (0.149)
Middle	0.178 (0.204)	0.161 (0.193)	0.149 (0.134)
High	0.191 (0.202)	0.168 (0.188)	0.148 (0.121)
Residency			
West	0.199 (0.219)	0.179 (0.206)	0.153 (0.134)
East	0.154 (0.168)	0.134 (0.154)	0.142 (0.121)
Job Sector			
Energy and Agriculture	0.113 (0.136)	0.099 (0.118)	0.11 (0.103)
Manufacturing	0.144 (0.176)	0.13 (0.164)	0.118 (0.117)
Services	0.201 (0.21)	0.179 (0.196)	0.175 (0.143)
Other sectors	0.233 (0.243)	0.21 (0.233)	0.175 (0.143)
Sum of pension earning points			
0–19	0.381 (0.257)	0.342 (0.24)	0.229 (0.141)
20–39	0.236 (0.206)	0.209 (0.195)	0.184 (0.131)
40–59	0.12 (0.163)	0.109 (0.154)	0.118 (0.117)
60 or more	0.114 (0.179)	0.105 (0.174)	0.091 (0.108)
Self-perceived health			
Excellent	0.192 (0.222)	0.164 (0.201)	0.14 (0.136)
Very good	0.178 (0.207)	0.161 (0.198)	0.138 (0.121)
Good	0.171 (0.192)	0.154 (0.18)	0.141 (0.124)
Fair	0.201 (0.212)	0.179 (0.198)	0.161 (0.134)
Poor	0.215 (0.234)	0.195 (0.218)	0.168 (0.149)
Aggregation quality			
Good	0.328 (0.199)	0.278 (0.175)	0.273 (0.112)
Perfect	0.168 (0.203)	0.153 (0.194)	0.133 (0.124)
Total	0.188 (0.209)	0.169 (196)	0.15 (0.131)

Source: Authors' calculations based on SHARE-RV version 7–0-0 and SHARE Wave 7 version 7–0-0

than those with a low level. The same applies to those working in the service and manufacturing sectors compared to the energy and agricultural sectors.

For most covariates, the results are in the same direction for all three distance measures. With respect to effect strength, distances are highest for Hamming and lowest for OMspell. This reflects the fact that OMspell is less time sensitive than Hamming and considers context more than OM. Likewise, it 'punishes' typical errors related to recall bias, such as forgetting short spells and the exact duration of spells, to a lesser extent than the other

Table 3 Linear regression models for Hamming distance, OM and OMspell as dependent variables

	(1) Hamming 1	(2) Hamming 2	(3) OM 1	(4) OM 2	(5) OMspell 1	(6) OMspell 2
Gender						
Male (Ref.)						
Female	−0.0241*	−0.0272*	-0.0200*	−0.0212*	0.0148*	0.00318
	(−2.29)	(−2.50)	(-1.97)	(−2.01)	(2.27)	(0.45)
Age in 2017	−0.00000446	0.000794	0.0000748	0.000637	−0.000736*	−0.0000971
	(−0.01)	(1.73)	(0.15)	(1.43)	(−2.28)	(−0.33)
Sum of earning points	−0.00509***	−0.000483	−0.00455***	−0.000230	−0.00183***	0.0000289
	(−14.28)	(−1.33)	(−13.31)	(−0.65)	(−8.30)	(0.12)
Education						
Low (Ref.)						
Middle	−0.0192	−0.0123	−0.0130	−0.00612	0.00635	0.00510
	(−1.16)	(-0.90)	(−0.82)	(−0.46)	(0.62)	(0.58)
High	0.0507**	0.00957	0.0453*	0.0115	0.0347**	0.0153
	(2.77)	(0.59)	(2.58)	(0.73)	(3.06)	(1.46)
Residency						
West (Ref.)						
East	−0.0540***	−0.0144	−0.0508***	-0.0147	−0.0261***	−0.0108
	(−4.94)	(−1.53)	(−4.85)	(-1.61)	(−3.86)	(−1.79)
Job Sector						
Energy and agriculture (Ref.)						
Manufacturing	0.0533*	0.00543	0.0466	0.00329	0.0355*	0.00716
	(2.14)	(0.26)	(1.95)	(0.16)	(2.31)	(0.54)
Services	0.0874***	0.0329	0.0761**	0.0275	0.0532***	0.0210
	(3.60)	(1.64)	(3.26)	(1.41)	(3.55)	(1.62)
Other sectors	0.0830**	0.0383	0.0754**	0.0350	0.0504**	0.0270*
	(3.27)	(1.83)	(3.10)	(1.72)	(3.22)	(2.00)
Subjective rated health						
Excellent (Ref.)						
Very good	−0.0127	0.00161	−0.00135	0.0121	−0.000373	0.000879
	(-0.48)	(0.07)	(−0.05)	(0.57)	(−0.02)	(0.06)
Good	−0.0164	−0.00900	−0.00547	0.000834	0.00525	0.00370
	(−0.67)	(−0.45)	(−0.23)	(0.04)	(0.35)	(0.29)
Fair	−0.0127	−0.00122	−0.00199	0.00806	0.00659	0.00573
	(−0.51)	(−0.06)	(−0.08)	(0.41)	(0.43)	(0.43)
Poor	−0.00247	0.00849	0.0113	0.0202	0.0158	0.0186
	(−0.09)	(0.37)	(0.43)	(0.92)	(0.93)	(1.27)
Aggregation quality	−0.126***	−0.0139	−0.0937***	−0.00270	−0.136***	−0.0398***
	(−8.81)	(−0.92)	(−6.85)	(−0.18)	(-15.50)	(−4.10)
Complexity index		0.564***		0.435***		0.630***
		(8.41)		(6.68)		(14.58)
Years spent in states						

Table 3 (continued)

	(1) Hamming 1	(2) Hamming 2	(3) OM 1	(4) OM 2	(5) OMspell 1	(6) OMspell 2
Employed		−0.00373		−0.00386		0.00845
		(−0.47)		(−0.50)		(1.64)
Unemployed		0.0120		0.0113		0.0101*
		(1.51)		(1.46)		(1.97)
Sick or disabled		0.000334		0.00172		0.0116
		(0.03)		(0.18)		(1.82)
Child-raising and household		0.00289		0.00321		0.00722
		(0.36)		(0.41)		(1.40)
School and training		−0.00279		−0.00330		0.000972
		(−0.35)		(−0.43)		(0.19)
Retired		0.0187*		0.0190*		0.0132*
		(2.29)		(2.40)		(2.51)
Other		0.0237**		0.0242**		0.0190***
		(2.94)		(3.09)		(3.66)
No Information		0.00914		0.00857		0.00982
		(1.15)		(1.11)		(1.91)
Constant	1.779***	0.270	1.364***	0.134	1.756***	0.230
	(11.10)	(0.92)	(8.87)	(0.47)	(17.74)	(1.22)
Observations	1676	1676	1676	1676	1676	1676

Source: Authors' calculations based on SHARE-RV version 7–0-0 and SHARE wave 7 version 7–0-0

t statistics in parentheses

* $p < 0.05$

** $p < 0.01$

*** $p < 0.001$

measures do, while Hamming and OM overly 'punish' small errors. For gender, however, effect strength and direction differ between the three distance measures. Using Hamming and OM, women have significantly lower sequence dissimilarity even after controlling for sequence characteristics in the second range of models. In contrast, women have significantly greater dissimilarity between survey and administrative data when OMspell is applied, which is in line with the descriptive results. However, with OMspell the relationship becomes insignificant after controlling for sequence characteristics.

After including sequence characteristics in the second range of models, when the Hamming and OM approaches are used, only gender differences remain significant: women have significantly smaller distances between the survey and administrative data. When using OMspell, only those working in 'other' job sectors have significantly higher dissimilarity between the survey and administrative data. All other socio-demographic covariates become non-significant in the second range of models when sequence variables are added. Among the sequence characteristics, the strongest relationship to the distance measures is found for the sequence complexity index: the higher the complexity, the more dissimilar are the survey and administrative data. In contrast to the socio-demographic characteristics, this relationship is strongest for OMspell, which considers spell length to a greater extent. Furthermore, an increasing number of years in the administrative data of the states 'retired' and 'other' has a significant positive relationship with all three distance measures. For OMspell, the number of years spent in unemployment also displays a significant positive relationship. Aggregation quality is negatively related to sequence dissimilarity in the first range of models, but becomes insignificant after controlling for sequence characteristics in the second range of models when applying the Hamming and OM measurement approaches. By contrast, for OMspell aggregation quality is highly significant in both models. Sequences with poorer aggregation quality have significantly higher OMspell values, even after considering sequence-specific variables such as complexity.

5 Discussion

The descriptive analysis demonstrates clear differences between the survey and administrative data. In particular, the lower complexity indices, transition rates and entropy values confirm that respondents oversimplify their life courses in retrospective surveys. Analysis of the agreement between the two types of data over time shows a relatively constant proportion of inconsistency across all age levels. Only the beginning of the observation period seems to be a slightly more error-prone life stage. Employment shows the highest levels of agreement, while unemployment data shows comparatively low levels of agreement. This result might be due to underreporting of unemployment states in the retrospective data, which may arise from two sources. First, people may feel too ashamed to report they have been unemployed in a survey. Second, the problem might be related to the survey design. The job spell concept applied in SHARELIFE focuses on reporting subsequent occupational sequences; it gives non-employment and unemployment a lower weight, which supports the underreporting thesis. With regards to regional differences, East Germans display very high agreement in the early phase of life. Here, during the socialist period, full-time employment was the norm with a high degree of labor market integration for both men and women, resulting in generally homogeneous trajectories and lower potential for mismatches. The system transformation following German reunification in 1990 is associated with an increased discontinuity and heterogeneity of East German employment biographies, including frequent spells of unemployment and atypical employment arrangements (Möhring and Weiland, 2021). Subsequently, in later-life-trajectories, which for most of the participants was after German reunification, there is slightly lower consistency in Eastern than in Western Germany.

Overall, the Hamming and OM distance measures are relatively similar, with the latter slightly undercutting the former as it allows for alignment. OMspell, which takes spell length into account, more frequently shows smaller distance values and almost no high values. This is not unexpected, given that the present data shows very high values for the average longest episodes. However, the calculated values of all three measures are highly correlated (see Appendix Table 6), as observed in the context of life-course data for Hamming and OM in previous research (Halpin 2010; Wahrendorf et al. 2019); it also seems to apply to OMspell. When comparing distances by covariates, we observe smaller distances for men, more educated respondents, those working in the energy and agricultural sectors, and participants who have a large number of earning points. Additionally, as suggested by the agreement over time, East Germans have lower distances overall.

The multivariate regression analysis with the distance values as the dependent variable shows that the complexity of a sequence is positively related to the distance. The significant association of socio-demographic covariates vanishes when sequence-specific variables are added to the model indicating that personal characteristics such as age, earning points, and employment sector are highly correlated with the complexity of a respondent's employment history. However, even after controlling for characteristics of respondents' employment histories, for women the differences between administratively recorded and personally reported trajectories measured by the Hamming and OM methods remain significant. In contrast to the descriptive results, women here have lower sequence dissimilarity than men—perhaps because women are overrepresented among service sector and low-income workers and typically have more complex employment histories. The gender differences we detect are reversed after controlling for these factors.

While the regression results using Hamming and OM as dependent variables reveal very similar relationships for all the independent variables, there are remarkable differences when OMspell is used. This distance measure takes into account the spell structure and the context alignment operations. For example, being female is positively related to sequence dissimilarity only when the OMspell approach is used, which may be because OMspell considers spell length to a greater extent and gives less weight to dissimilarities in longer spells. As women have more fragmented life courses with short spells and men have more continuous careers with longer spells, the latter are less 'punished' in the more context-sensitive OMspell.

Our findings are in line with those of previous studies. We find that retrospective life trajectories are more likely to diverge from administrative data the more versatile a life course is. Furthermore, even though the reason for disagreement is not rooted in their group affiliation, certain social groups exhibit more complex employment histories and therefore are especially prone to larger differences, such as women, respondents with a low level of education and people employed in the tertiary sector. For these groups, researchers must account for the fact that survey life history data underestimates life-course complexity. Moreover, with respect to gender differences, researchers applying sequence analyses to either administrative or survey data have to keep in mind that the extent and direction of gender differences vary not only depending on the data source, but also between the distance measures.

This study has two main limitations. First, self-employed, civil servants and homemakers are not represented in the analysis, which also reveals a blind spot in the employed administrative data. Second, and as reported by Wahrendorf et al. (2019), respondents must consent to have their data linked, and not all SHARE participants gave permission. Therefore, further research is needed, not only to compare life history data from different sources that also cover employment states, which we were unable to include, but also to systematically compare the results obtained from different distance measures.

Appendix A

See Tables 4, 5, 6.

Table 4 Sample statistics

Variables	Frequency	Percent
Gender		
Male	813	48.42
Female	866	51.58
Age at interview		
49–64 years	781	46.52
65–79 years	744	44.31
80 years or older	154	9.17
Education		
Low	148	8.81
Middle	1,009	60.10
High	522	31.09
Residency		
West	1,295	77.13
East	384	22.87
Sum of pension earning points		
0–19	147	8.76
20–39	670	39.90
40–59	624	37.16
60 or more	238	14.18
Job sector		
Energy and agriculture	62	3.69
Manufacturing	464	27.64
Services	811	48.30
Other sectors	342	20.37
Self-perceived health		
Excellent	60	3.57
Very good	212	12.63
Good	692	41.22
Fair	536	31.92
Poor	176	10.48
Total	1679	100

Source: Authors’ calculations based on SHARE-RV version 7–0-0 and SHARE Wave 7 version 7–0-0

Table 5 Mapping of VSKT and SHARELIFE states

Mapping	VSKT	SHARELIFE
No info	No info Unpaid care Self-employed	Looking after home and family, IF youngest child in household > 10 years old
		Civil servant
		Self-employed
Missing	-	Missing
School and Training	Vocational training School	Training Further fulltime education School
Child raising and homemaking	Child raising and homemaking	Looking after home and family, IF youngest child in household < 10 years old and first child born
Sick or disabled	Incapacity to work/ illness	Sick or disabled
	Supplementary period	
Unemployed	Unemployed: unemploymentbenefit/ALGII	Unemployed searching/ not searching for job
	unemployed: Unemployment allowance (including unemployment benefit up to 2000)	
	Unemployed: credit period	
Military, civil service, war prisoner or equivalent	Military and civilian service	Military services, war prisoner or equivalent
Employed	Marginally employed	Employed, Short term job (less than 6 months)
	Gainfully employed and obligated to pay social insurance	
Retired from work	Pension provision (own insurance)	Retired from work
Other	Other	Leisure, travelling or doing nothing
		Managing assets
		Voluntary or community work
		Forced labor or in jail
		Exiled or banished

Table 5 (continued)

Mapping	VSKT	SHARELIFE
		Labor camp
		Concentration camp

Table 6 Pairwise correlations of distance measures

Distance measures	(1)	(2)	(3)
(1) Hamming	1.000		
(2) OM	0.984*	1.000	
(3) OMspell	0.822*	0.829*	1.000

* indicates significance at the 0.000 level

Source: Authors' calculations based on SHARE-RV version 7–0-0 and SHARE Wave 7 version 7–0-0

Appendix B: Aggregation methods

We apply two different aggregation methods, as illustrated in Fig. 4. The mode aggregation uses the most frequent state of the respective year. The rule-based aggregation summarizes the year based on a state hierarchy similar to applications in previous research (Huber and Schmucker 2009). This SHARE hierarchy is deduced from the hierarchy used to deal with concurrent states during the transformation into monthly Social Income Situations for Share RV (VSKT User Information Release 7–1-0; SHARE RV 2019). Accordingly, we ranked the states in the following order: employed, military,

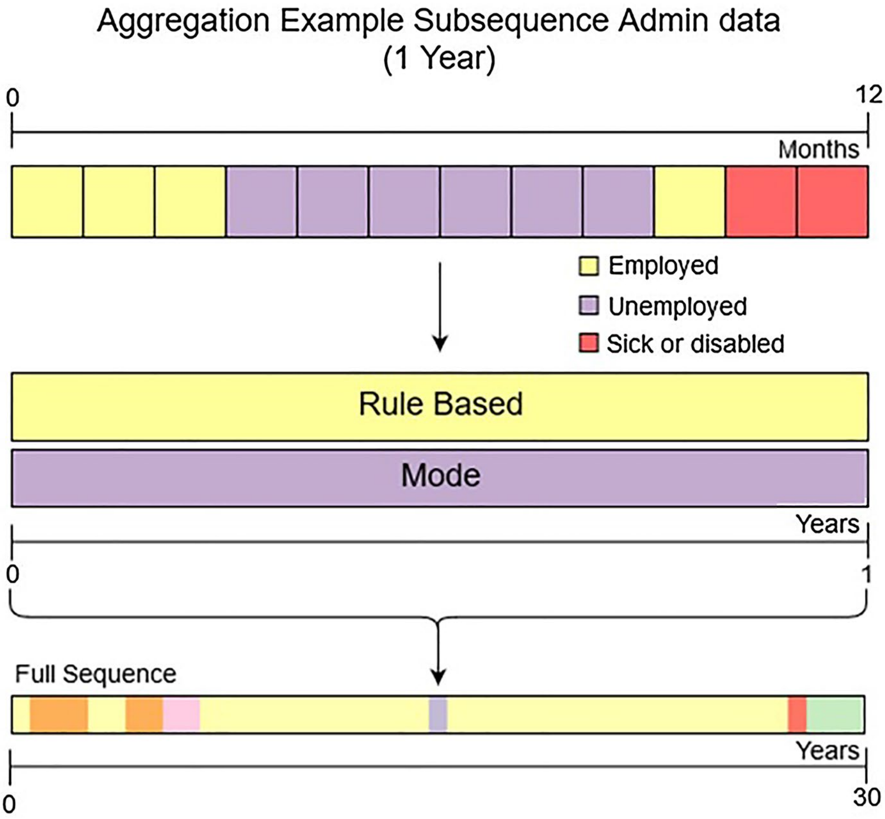


Fig. 4 Aggregation methods Source: Authors' own depiction

Table 7 Social Income Situation VSKT (according to VSKT User Information Release 7–1-0) and State Aggregation Hierarchies

Hierarchy of Social Income Situations VSKT	Hierarchy of States Data Aggregation
Compulsory contribution except of child-raising	Employed Military
Voluntary contribution	
Creditable activities	Unemployed
	Sick or disabled
	School or training
(Credited) substitute activities	
Voluntary additional insurance	
Pension provision	Retired
Child raising period	Child raising
Other activities taken into account	Other, no information

Table 8 Aggregation quality
Source Authors' calculations based on SHARE-RV version 7–0-0 and SHARE Wave 7 version 7–0-0

	Rule based	
Mean quality	Freq.	Percent
6–8 months	5	0.30
9–11 months	375	22.33
12 months	1299	77.37

Table 9 Distances between survey and administrative data by covariates

	Admin data rule based mean (sd)
Duration (years)	
Employed	24.182 (6.85)
Retired from work	0.235 (1.547)
Unemployed	1.075 (2.503)
Child raising & household	2.799 (5.244)
Sick or disabled	0.134 (0.598)
School & training	1.120 (2.417)
Military, civil service	0.256 (0.701)
No information	0.995 (2.083)
Other	0.497 (2.439)
Missing	0 (0)
Complexity index	0.117 (0.104)
Transitions	2.388 (2.259)
Entropy	0.177 (0.162)
Longest episode	22.118 (7.242)

Table 10 Pairwise correlations of distance measures for rule-based aggregation *Source* Authors' calculations based on SHARE-RV version 7–0-0 and SHARE Wave 7 version 7–0-0

	Rule based		
Distance measures	(1)	(2)	(3)
(1) Hamming	1.000		
(2) OM	0.987*	1.000	
(3) OMspell	0.854*	0.851*	1.000

* shows significance at the 0.000 level

unemployed, sick or disabled, school or training, retired, child raising, other, no information (see Appendix Table 7). This means that a respondent receives the highest-ranked state that they display in at least one month in the corresponding year.

The results of the two forms of aggregation differ only marginally. They produce different outcomes in only 5 percent of the overall years of all respondents. Only 0.3 percent of the sample shows a low average quality of aggregation, while 22.3 percent have a medium quality and 77.4 percent a high quality (compare with Table 8). Yet, we preferred the mode aggregation for two reasons. First, its aggregation quality is considerably better. Second, this method is independent from subjective decisions about a certain hierarchy of the states. The following supplemental material presents further comparisons between the two forms of aggregation. Table 9 presents the summary statistics of sequences generated by rule-based data aggregation. Comparing the two aggregation procedures that we applied demonstrates that in addition to the general similarity of the overall statistics, the rule-based aggregation results in less complex sequences, with fewer transitions and less entropy, most probably due to preferences for certain states over others. In these states they naturally also demonstrate longer average years. However, this seems to make them more similar to the survey data on the whole.

To rule out the possibility that differences in the quality of aggregation distort the analysis of sequence distances, we investigated the extent to which the average quality of the aggregation per person, namely the actual number of months spent in the assigned state in a given year, is associated with the resulting distance measures. Our ANOVA test shows significant differences in the distance values by quality. However, since most respondents show a very high average aggregation quality (see Appendix Table 8), this should not greatly distort the results. In addition, frequent status changes within a single year that are associated with low aggregation quality can also indicate generally more complex sequences, which have a higher expected distance. Comparing the distance measures (Hamming, OM and OMspell), indicates that the average distances are lower for rule-based, than for mode aggregation for all three distance measures. Hamming has the highest average distance values in both forms of aggregation, followed by OM and OMspell. As can be seen in Table 10, all distance measures are highly correlated.

Acknowledgements This paper uses data from SHARE Wave 7 (DOI: 10. 6103/ SHARE.w7.710), see Börsch-Supan et al. (2013) for methodological details. The SHARE data collection has been funded by the European Commission through FP5 (QLK6-CT-2001-00360), FP6 (SHARE-I3: RII-CT-2006-062193, COMPARE: CIT5-CT-2005-028857, SHARELIFE: CIT4-CT-2006-028812), FP7 (SHARE-PREP: GA N°211909, SHARE-LEAP: GA N°227822, SHARE M4: GA N°261982, DASISH: GA N°283646) and Horizon 2020 (SHARE-DEV3: GA N°676536, SHARE-COHESION: GA N°870628, SERISS: GA N°654221, SSHOC: GA N°823782) and by DG Employment, Social Affairs & Inclusion. Additional funding from the German Ministry of Education and Research, the Max Planck Society for the Advancement of Science, the U.S. National Institute on Aging (U01_AG09740-13S2, P01_AG005842, P01_AG08291, P30_AG12815, R21_AG025169, Y1-AG-4553-01, IAG_BSR06-11, OGHA_04-064, HHSN271201300071C) and from various national funding sources is gratefully acknowledged (see www.share-project.org). The R code of the Project is stored here: https://osf.io/45xz3/

Funding Open Access funding enabled and organized by Projekt DEAL. Project funding from the German Research Network on Pensions (FNA).

Availability of data and material Data access is restricted; see http://www.share-project.org/data-access/share-conditions-of-use.html for further information.

Code availability The R code of the Project is stored here: https://osf.io/45xz3/

Declarations

Conflicts of interest No conflicts of interest.

References

Abbott, A., Forrest, J.: Optimal matching methods for historical sequences. J. Interdiscip. Hist. **16**(3), 471–494 (1986). https://doi.org/10.2307/204500

Abbott, A., Hrycak, A.: Measuring resemblance in sequence data: an optimal matching analysis of musicians' careers. Am. J. Sociol. **96**(1), 144–185 (1990)

Abowd, J.M., Harrison Stinson, M.: Estimating measurement error in SIPP annual job earnings: a comparison of census bureau survey and SSA administrative data. SSRN Electron. J. (2011). https://doi.org/10.2139/ssrn.1894690

Adriaans, J., Valet, P., Liebig, S.: Comparing administrative and survey data: is information on education from administrative records of the German Institute for Employment Research consistent with survey self-reports? Qual. Quant. **54**(1), 3–25 (2020). https://doi.org/10.1007/s11135-019-00931-4

Börsch-Supan, A., Brandt, M., Hunkler, C., Kneip, T., Korbmacher, J., Malter, F., Schaan, B., Stuck, S., Zuber, S.: Data resource profile: the survey of health, ageing and retirement in Europe (SHARE). Int. J. Epidemiol. **42**(4), 992–1001 (2013)

Börsch-Supan, A.: Survey of health, ageing and retirement in Europe (SHARE) Wave 7. Release version: 7.1.1. SHARE-ERIC. Data set (2019). https://doi.org/10.6103/SHARE.w7.711

Brzinsky-Fay, C., Kohler, U., & Luniak, M.: Sequence analysis with Stata. Stata J. **6**(4), 435–460 (2006)

Brzinsky-Fay, C., Kohler, U.: New developments in sequence analysis. Sociol. Methods Res. **38**(3), 359–364 (2010). https://doi.org/10.1177/0049124110363371

Destatis, Genesis-Online: 12211–9000: Bevölkerung, Erwerbstätige, Erwerbslose, Erwerbspersonen, Nichterwerbspersonen [jeweils im Alter von 15 bis unter 65 Jahren]: Deutschland, Jahre (bis 2019), Geschlecht. Datenlizenz by-2–0. https://www-genesis.destatis.de/genesis//online?operation=table&code=12211-9000&bypass=true&levelindex=0&levelid=1638704581271#abreadcrumb (2021a). Accessed 15 December 2021

Destatis, Genesis-Online: 12211–9005: Erwerbstätige: Deutschland, Jahre (bis 2019), Stellung im Beruf, Geschlecht. Datenlizenz by-2–0. https://www-genesis.destatis.de/genesis//online?operation=table&code=12211-9005&bypass=true&levelindex=0&levelid=1638707272562#abreadcrumb (2021b). Accessed 15 December 2021

Elzinga, C.H.: Sequence similarity: a nonaligning technique. Sociol. Methods Res. **32**(1), 3–29 (2003)

Elzinga, C.H., Wang, H.: Kernels for acyclic digraphs. Pattern Recogn. Lett. **33**(16), 2239–2244 (2012)

Gabadinho, A., Ritschard, G., Studer, M., Müller, N. S.: Mining sequence data in R with the TraMineR package: Auser's guide. Department of Econometrics and Laboratory of Demography, University of Geneva, Geneva (2009)

Gabadinho, A., Ritschard, G., Müller, N.S., Studer, M.: Analyzing and visualizing state sequences in R with TraMineR. J. Stat. Softw. **40**(1), 1–37 (2011). https://doi.org/10.18637/jss.v040.i04

Gabadinho, A., Studer, M., Müller, N., Bürgin, R., Fonta, P.-A., Ritschard, G.: TraMineR: Trajectory miner: a toolbox for exploring and rendering sequences (2020). https://CRAN.R-project.org/package=TraMineR. Accessed 31 May 2020

Groen, J.A.: Sources of Error in Survey and Administrative Data: The Importance of Reporting Procedures. J. Off. Stat. **27**(2), 173–198 (2012)

Halpin, B.: Optimal matching analysis and life-course data: the importance of duration. Sociol. Methods Res. **38**(3), 365–388 (2010). https://doi.org/10.1177/0049124110363590

Hamming, R.W.: Error detecting and error correcting codes. Bell Syst. Tech. J. **29**(2), 147–160 (1950). https://doi.org/10.1002/j.1538-7305.1950.tb00463.x

Hollister, M.: Is optimal matching suboptimal? Sociol. Methods Res. **38**(2), 235–264 (2009)

Huber, M., Schmucker, A.: Identifying and explaining inconsistencies in linked administrative and survey data: the case of German employment biographies. Hist. Soc. Res. **34**(3), 230–241 (2009). https://doi.org/10.12759/hsr.34.2009.3.230-241

Jenkins, S.P., Cappellari, L., Lynn, P., Jäckle, A., Sala, E.: Patterns of consent: evidence from a general household survey. J. r. Stat. Soc. Stat. Soc. **169**(4), 701–722 (2006). https://doi.org/10.1111/j.1467-985X.2006.00417.x

Korbmacher, J., Czaplicki, C.: Linking SHARE survey data with administrative records: first experiences from SHARE-Germany. In: Malter, F., Börsch-Supan, A. (eds.) Share wave 4: innovations & methodology. MEA, Max Planck Institute for Social Law and Social Policy, Munich (2013)

Korbmacher, J.M., Schroeder, M.: Consent when linking survey data with administrative records: the role of the interviewer. Surv. Res. Methods **7**(2), 115–131 (2013)

Korbmacher, J.M.: Recall Error in the Year of Retirement. SHARE Working Paper Series 21–2014, 42 (2014)

Kreiner, C.T., Lassen, D.D., Leth-Petersen, S.: Measuring the Accuracy of Survey Responses using Administrative Register Data: Evidence from Denmark,. In: Carroll, C. D., Thomas F. Crossley, T. F., Sabelhaus, J. (eds.) Improving the Measurement of Consumer Expenditures, Vol. 74, 289–307. University of Chicago Press, Chicago (2015)

Kreuter, F., Presser, S., Tourangeau, R.: Social desirability bias in CATI, IVR, and web surveysthe effects of mode and question sensitivity. Public Opin. q. **72**(5), 847–865 (2008). https://doi.org/10.1093/poq/nfn063

Kreuter, F., Müller, G., Trappmann, M.: Nonresponse and measurement error in employment research: making use of administrative data. Public Opin. q. **74**(5), 880–906 (2010). https://doi.org/10.1093/poq/nfq060

Kreyenfeld, M., Bastin, S.: Reliability of union histories in social science surveys: blurred memory, deliberate misreporting, or true tales? Adv. Life Course Res. **27**, 30–42 (2016). https://doi.org/10.1016/j.alcr.2015.11.001

Krumpal, I.: Determinants of social desirability bias in sensitive surveys: a literature review. Qual. Quant. **47**(4), 2025–2047 (2013). https://doi.org/10.1007/s11135-011-9640-9

Kühne, S.: From strangers to acquaintances? Interviewer continuity and socially desirable responses in panel surveys. Surv. Res. Methods **12**(2), 121–146 (2018). https://doi.org/10.18148/srm/2018.v12i2.7299

Lee, K. O., Smith, R., Galster, G.: Neighborhood trajectories of low-income US households: An application ofsequence analysis. J. Urban Aff. **39**(3), 335–357 (2017)

Lesnard, L.: Setting cost in optimal matching to uncover contemporaneous socio-temporal patterns. Sociol. Methods Res. **38**(3), 389–419 (2010)

Manzoni, A., Vermunt, J.K., Luijkx, R., Muffels, R.: Memory bias in retrospectively collected employment careers: a model-based approach to correct for measurement error. Sociol. Methodol. **40**(1), 39–73 (2010)

Mika, T.: The effects of social and institutional change on data production. The case of welfare state reforms on the rise and decline of unemployment and care-giving in the German Pension Fund Data. Historical Social Research/Historische Sozialforschung, 115–137 (2009)

Möhring, K., Weiland, A.P.: Couples' life courses and women's income in later life: a multichannel sequence analysis of linked lives in Germany. Eur. Sociol. Rev. (2021). https://doi.org/10.1093/esr/jcab048

Forschungsdatenzentrum der Rentenversicherung, Max-Planck-Institut für Sozialrecht und Sozialpolitik: SHARE-RV. Release version: 7.0.0. SHARE-ERIC. Dataset (2019). doi: https://doi.org/10.6103/SHARE.SHARE-RV.710

Sakshaug, J., Antoni, M., Sauckel, R.: The quality and selectivity of linking federal administrative records to respondents and nonrespondents in a general population sample survey of Germany. Surv. Res. Methods **11**(1), 63–80 (2017). https://doi.org/10.18148/srm/2017.v11i1.6718

Schröder, M.: Concepts and topics. In: Schröder, M. (ed.) Retrospective Data Collection in the Survey of Health, Ageing and Retirement in Europe. SHARELIFE Methodology. MEA, Mannheim (2011)

Solga, H.: Longitudinal surveys and the study of occupational mobility: panel and retrospective design in comparison. Qual. Quant. **35**(3), 291–309 (2001). https://doi.org/10.1023/A:1010387414959

Squires, P., Kaufman, H. G., Togelius, J., & Jaramillo, C. M.: A comparative sequence analysis of career pathsamong knowledge workers in a multinational bank. 2017 IEEE International Conference on Big Data (Big Data).3604-3612 (2017). https://doi.org/10.1109/BigData.2017.8258354

Studer, M., Ritschard, G.: What matters in differences between life trajectories: a comparative review of sequence dissimilarity measures. J. R. Stat. Soc. A. Stat. Soc. **179**(2), 481–511 (2016). https://doi.org/10.1111/rssa.12125

Trappe, H., Pollmann-Schult, M., Schmitt, C.: The rise and decline of the male breadwinner model: Institutional underpinnings and future expectations. Eur. Sociol. Rev. **31**(2), 230–242 (2015)

Valet, P., Adriaans, J., Liebig, S.: Comparing survey data and administrative records on gross earnings: nonreporting, misreporting, interviewer presence and earnings inequality. Qual. Quant. **53**(1), 471–491 (2019). https://doi.org/10.1007/s11135-018-0764-z

Wagner, M., Philip, J.T.: SHARELIFE. SHARE Wave 7 Methodology: Panel innovations and life histories (2019)

Wahrendorf, M., Marr, A., Antoni, M., Pesch, B., Jöckel, K.-H., Lunau, T., Moebus, S., Arendt, M., Brüning, T., Behrens, T., Dragano, N.: Agreement of self-reported and administrative data on employment histories in a German cohort study: a sequence analysis. Eur. J. Popul. **35**(2), 329–346 (2019). https://doi.org/10.1007/s10680-018-9476-2

West, B.T., Blom, A.G.: Explaining interviewer effects: a research synthesis. J. Surv. Stat. Method. **5**(2), 175–211 (2017). https://doi.org/10.1093/jssam/smw024

Widmer, E.D., Ritschard, G.: The de-standardization of the life course: Are men and women equal? Adv. Life Course Res. **14**(1), 28–39 (2009). https://doi.org/10.1016/j.alcr.2009.04.001

Wu, L.L.: Some comments on "Sequence analysis and optimal matching methods in sociology: review and prospect." Sociol. Methods Res. **29**(1), 41–64 (2000)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.