Matching survey data on wealth to register data on pension entitlements: what challenges need to be addressed?

Peter Lindner, Martin Schürz[1]

This paper focuses on the challenges that need to be tackled when matching different sources of data. We first present details on how to set up the matched data, before discussing variants of statistical matching of survey data and register data. For our statistical matching exercise, we use data from the Household Finance and Consumption Survey (HFCS) in Austria as recipient data. Additionally, we use the full range of data from Austria's social security register, including target information on pension entitlements, as donor data to enrich information obtained from the HFCS on households' balance sheets.

The way we define wealth influences how we measure wealth inequality. As a rule, wealth inequality is high; yet, measurements of wealth inequality often do not take into account people's pension entitlements. The conceptual challenge here is to capture the wealth component of pension entitlements which reflects people's capability to finance future consumption. There is a difference between private and public pension plans, however. As regards the latter, we would need to know about the future parameters of public pension systems and know enough about people's employment careers, which is not feasible. Nevertheless, broad and comprehensive measurements including pension entitlements may increase our understanding of wealth inequality and saving behavior.

Pension entitlements are quantitatively important. Pension wealth is relatively more important for people with low wealth. Although pension wealth rises with other wealth components, the ratio of pension wealth to household wealth decreases as household wealth increases. One motive for wealth accumulation is to finance consumption in retirement. When generous public pension schemes are in place, the incentive to save for retirement will be lower (see Fessler and Schürz, 2018).

This paper is structured as follows: Section 1.1 discusses the data obtained from the Household Finance and Consumption Survey (HFCS) in Austria. In section 1.2, we present the data on pension entitlements taken from Austria's social security register. Section 2, the core part of the paper, explains the basics of statistical data matching and shows the results based on the matched data. Finally, section 3 concludes.

1 Data

1.1 HFCS data for Austria

The Eurosystem HFCS is a comprehensive survey collecting data on balance sheets of households across euro area countries and beyond. The data capture households' real assets, financial assets, debt levels, income and expenditures, allowing for in-depth scientific analyses of households' balance sheets in line with international

[1] *Oesterreichische Nationalbank, Economic Analysis Division, peter.lindner@oenb.at, martin.schuerz@oenb.at. Opinions expressed by the authors of studies do not necessarily reflect the official viewpoint of the OeNB or the Eurosystem. The authors would like to thank Stefan Humer, Pirmin Fessler and Nicolas Albacete (all OeNB) for helpful comments and valuable suggestions.*

standards. Moreover, the data are comparable across all countries participating in the HFCS thanks to the ex ante harmonization of the survey and the survey methods applied. In Austria, the HFCS was first carried out in 2010/2011, a second time in 2014/2015 and most recently in 2016/2017. All three survey waves were conducted by the Oesterreichische Nationalbank (OeNB). The data obtained allow us to analyze households' investment and consumption decisions. Household-level finance and expenditure data are indispensable for a central bank, as they contribute significantly to improving monetary policy and financial stability analyses.

The fieldwork for the third wave[2] of the HFCS Austria was carried out between late November 2016 and July 2017. 70 interviewers, who were specifically trained for this survey (and had mostly gained experience from previous HFCS waves), conducted the interviews using computer-assisted personal interviewing (CAPI). This technique allows for dynamic interviews and data checks already during the interview. The target population included all households in Austria, irrespective of household members' nationality and citizenship, but excluding households that are institutionalized, such as households living in homes for elderly people, military compounds, monasteries, prisons and boarding schools.

A complete list of postal addresses of all households in Austria was used as the sampling frame. Moreover, a two-stage cluster sampling design was employed with enumeration districts as the primary sampling unit (PSU) and postal addresses as the secondary sampling unit (SSU). The sample was stratified by NUTS 3 regions and eight municipality size categories. The gross sample comprised a total of 614 PSUs and 6,280 SSUs in 180 strata. All households received a personalized letter from the governor of the OeNB and an information leaflet before they were contacted by the interviewers. Quality controls were key features during the field phase. Hence, it was possible to contact households again in case some information was not clear (for further details, see Albacete et al. (2018)).

The gross sample consisted of 6,280 households, of which 3,072 were successfully interviewed. Thus, the unit response rate came close to 50%. In addition to achieving an acceptable survey response rate, another issue that needed to be tackled was partial completeness of the data, i.e. the fact that the answers to some questions were missing for some households (item nonresponse rate). These missing values were imputed using multiple imputation with chained equations (broad conditioning approach). Final household-level weights were computed with nonresponse and poststratification adjustments to design weights. We used a model-based adjustment combined with weighting-class adjustment, based on an algorithm for the optimal number of classes (nonresponse adjustment) as well as a cell adjustment for poststratification. To allow for variance estimation without having the full sampling information – which cannot be disclosed – replicate weights were produced. The HFCS uses a rescaling bootstrap procedure, the idea of which is to mimic the sampling design, including 137 pseudo strata to generate 1,000 replicate weights. All adjustments made to design weights to obtain replicate weights were identical to adjustments made to obtain the final weights. Also, finite population corrections were applied to all replicate weights.

[2] *For the purposes of this paper, we use the third wave of the HFCS. For a detailed documentation on the HFCS, see Fessler et al. (2018) and Albacete et al. (2018).*

The HFCS collects information at both the household and the personal level (with some information only being collected for persons aged 16 or over). Since the main focus of the survey is the household level, all weighting information pertains to this level. Below, we therefore present unweighted as well as weighted results using household weights also for personal information. While all wealth items in the HFCS relate to the household level, information on sociodemographic characteristics as well as household members' occupation, main income sources and future pension arrangements are collected at the personal level.

Table 1

Personal information in the HFCS

Occupation	Age category		
	≤16	16–62	≥62
Child	938	0	0
Has never worked	0	373	96
Retired	0	254	1,264
Employed	0	2,915	66
Self-employed	0	271	19
Farmer	0	49	0
Civil servant	0	161	8
Total	938	4,023	1,453

Source: HFCS 2017, OeNB.

The 3,072 households that were successfully interviewed represent 6,414 household members, of which 938 are children aged under 16 and 1,453 are adults aged 62 or over (see table 1).

We implement the matching procedure at the personal level. This allows for more precise matching, as the information for each person is considered. We split the sample by age, as children have not yet accrued any pension entitlements; retired persons are treated separately. We use data for age cohorts born in 1955 and later (see section 1.2), excluding data on persons who have never worked (about 500 persons, see table 1) and have therefore not acquired any pension entitlements. After data matching, all persons are again aggregated to the household to which they belong.

1.2 Data from Austria's social security register

The donor data were obtained from Austria's social security register. More specifically, we use a complete snapshot of the data available in the social security register as of fall 2020 (in the following, this dataset will be abbreviated as SSR 2020). The data are split into two subsamples, i.e.

- *economically active persons* who have contributed to social security and were born between 1955 and 2001; and
  - The reasons for containing the years of birth include, on the one hand, legal reasons, as there is no information available on pension accounts of persons born before 1955 and, on the other hand, conceptual reasons, as most of the personal data collected by the HFCS only refer to persons aged 16 or over.
- persons receiving a pension.

For economically active persons, the following information is available:

- postal code: refers to the address to which mail is sent;
- gender;
- year of birth: grouped into five-year age brackets, with the earliest age bracket comprising six years;

- social security institution:
  - public pension insurance fund (Pensionsversicherungsanstalt – PVA);
  - social security fund for the self-employed (Sozialversicherungsanstalt der Selbstständigen (Bereich Gewerbliche Wirtschaft) – SVS-GW);
  - social security fund for farmers (Sozialversicherungsanstalt der Selbstständigen (Bereich Landwirtschaft) – SVS-LW);
  - social security fund for public sector employees, railways and mining (Versicherungsanstalt öffentlich Bediensteter, Eisenbahnen und Bergbau – BVAEB), including
    - civil servants employed by the federal government (including Österreichische Post AG, Telekom Austria AG, Österreichische Postbus AG, Austrian Federal Railways, Federal Theaters Holding Company); and
    - civil servants employed by the provincial governments (including secondary teachers);
- initial pension credits credited to individuals' notional (pension) accounts, including pension entitlement periods that could be purchased retroactively;
- pension credits for 2016 (HFCS reference period for income);
- total pension credits added to individuals' notional accounts as of January 1, 2017;
- additional information on civil servants still pending.

The data were selected in accordance with the HFCS field period and reference period for income. Table 2 shows the number of observations for economically active persons by social security institution and gender.

As can be seen in table 2, this subsample includes about 4 million economically active persons in Austria, most of which are covered by the PVA.

For persons receiving pension income, the following information is available:

- postal code;
- gender;
- exact year of birth;
- social security institution;
  - Essentially, these are identical to the institutions for economically active persons. Information on civil servants is not available, however.
- type of pension:
  - pension annuity (including supplements from the miners' pension fund (Knappschaftssold));
  - disability pension for blue- and white-collar workers as well as miners;
  - disability pension for the self-employed and civil servants;
  - widow's/widower's pension;
  - orphan's pension;
- monthly gross pension income as of December 2016;
- starting date of pension payments.

Table 2

Economically active persons by social security institution and gender

	Male	Female
BVAEB (railways and mining)	36,234	9,786
Civil servants employed by the federal government	70,600	19,599
Civil servants employed by provincial governments	5,562	23,235
PVA	1.925,982	1.725,657
SVS-GW	204,914	91,827
SVS-LW	39,486	34,833
Total	2.282,778	1.904,937

Source: SSR 2020.

Table 3

Persons receiving a pension by social security institution and gender

	Male	Female
BVAEB (railways and mining)	13,552	11,576
PVA	520,857	855,218
SVS-GW	68,122	72,050
SVS-LW	47,449	95,352
Total	649,980	1.034,196

Source: SSR 2020.

As shown in table 3, this subsample contains about 1.7 million persons. Most of them are employees and are thus covered by the PVA.

When interpreting table 3, we need to bear in mind that this subsample does not include data on special pension arrangements for civil servants or people who receive their pension benefits directly from their former employer (e.g. former central bank employees). The lack of these data is one of the weaknesses of the SSR 2020. Additionally, there are about 550 observations with obvious mistakes, such as non-existent postal codes or negative values for total pension credits. These observations amount to some 0.01% of the data and are excluded from the statistical matching exercise.

2 Statistical matching of HFCS data and SSR data

In this seven-part section, we first introduce the matching procedure in theoretical terms. This includes information on the comparison of the underlying datasets as well as on data uncertainty (section 2.1). In a next step, we introduce the matching variables available (section 2.2), and present some information on data alignment (section 2.3). Section 2.4 describes the matching variables in greater detail. Finally, we round up this section by discussing data challenges (section 2.6) and matching uncertainty (section 2.7).

2.1 Theoretical concepts

Ideally, we would observe two distinct sets of information – *Y* and *Z* – in one dataset. As this is not the case, we would like to have a common identifier *ID*, which can be used to link information *Y* from data *A* to information *Z* from data *B*, so that there is only one observation in *B* for each observation in *A* and we know how they are linked.

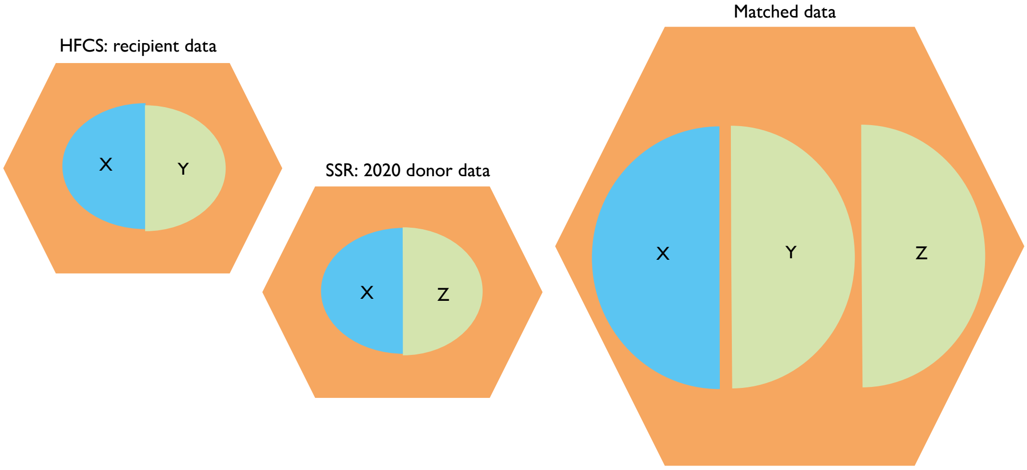
Matching procedures

Often, this is also not possible and we face the following situation: We have two distinct datasets *A* and *B*, where *A* contains information *Y* (say on wealth or demographics) and *B* contains information *Z* (say on pension credits). What we are lacking is information on how to link these two datasets, however. We can only identify a set of common characteristics *X* that can be found in both datasets and can be used to match observations.

Figure 1 shows an overview of the aims of such a statistical matching exercise. We use HFCS data, which contains information *X* and *Y* (box to the left), and SSR 2020 data, which contains information *X* and *Y* (box in the middle). By matching these two datasets, we aim to analyze both *Y* and *Z* in one dataset (box to the right).

Figure 1

Overview of matching procedure



Authors' compilation.

There are different approaches to matching data in such a scenario. Following D'Orazio et al. (2006), we can distinguish between

- parametric methods (e.g. regression type methods);
- non-parametric methods (e.g. hot deck procedures); and
- combinations of these methods.

The OECD Framework for Statistics on the Distribution of Household Income, Consumption and Wealth (OECD, 2013; chapter 7) describes these techniques in detail and considers their potential benefits, limitations and implications (e.g. for collection design). One way to think about the approaches listed above is to consider a regression as the simplest case with just one constant which is used to predict Z in the recipient data. This gives the mean imputation for each observation in the recipient data. Including explanatory variables in X in the regression allows for more flexibility, so that we can think of a conditional mean imputation/forecast. Conditional on X=x, i.e. the realization of the X matrix for a specific observation, the mean of all these observations in the donor data is imputed in the recipient data. If we allow for many variables in X and its interactions, we can mimic the non-parametric approach. For the non-parametric approach, however, we usually do not take the mean of a specific combination, but a random observation thereof in the donor data. This allows for the random hot deck approach.

Conditional independence assumption

Every matching approach depends on the conditional independence assumption (CIA), i.e. on

$$f(x,y,z) = f_{y|x}(y|x)f_{z|x}(z|x)f_X(x) \qquad \forall x \in X, y \in Y, z \in Z.$$

When using the CIA, the following considerations have to be taken into account:

- The assumption implies that the correlations between Y and Z are captured by X. This means that – conditional on X – Z and Z are completely independent.

- The assumption cannot be tested.
- A broad set of "good" conditioning variables in *X* seems to justify the use of this assumption. Thus, researchers tend to opt for a relatively broad set of variables in *X*.

Since the CIA cannot be tested but is central to the statistical matching of two distinct datasets, we opt for two different matching procedures. The stability of the results points to the major impact resulting from this choice of modeling.

The choice of variables in depends heavily on what kind of data is available. However, the data underlying the correlation between the variables also plays a major role. In the case at hand, people's occupational status and number of working years influence pension credits.[3] Additionally, we would like to have a relatively similar distribution of observations along these matching variables. In other words, the characteristics of the observations defined by the matching variables should be distributed similarly in both the recipient and the donor data. If we want to match income to a dataset only including observations for men and the donor data only contains observations for women, for instance, our endeavor is deemed to fail due to the gender wage gap. More formally, we can estimate the similarity of the distribution of observations over matching variables and combinations thereof. This is commonly done using the so-called Hellinger distance.

Hellinger distance

Let p_{Aj} be the relative frequency of observations with specific characteristics from the matching variables in data *A*, and p_{Bj} be the same for data *B* for all categories *J=1...J*. The Hellinger distance (HD) is then defined as

$$d_{H,12} = \sqrt{1 - B_{12}}$$

where B_{12} is the Bhattacharyya coefficient given by

$$B_{12} = \sum_{j=1}^{J} \sqrt{p_{Aj} * p_{Bj}}.$$

The HD ranges between 0 and 1, with 1 indicating maximum dissimilarity of the two distributions involved. Additionally, according to D'Orazio et al. (2006), we can show that the HD is connected to the dissimilarity index $\Delta_{12} = \frac{1}{2}\sum_{j=1}^{J}|p_{Aj} - p_{Bj}|$ through the following inequality:

$$d_{H,12}^{2} \leq \Delta_{12} \leq d_{H,12}\sqrt{2}.$$

Furthermore, we should analyze in depth not only the basic inputs of the matching procedure but also the (additional) uncertainty of statistical matching.

Fréchet bounds

One approach found in the literature to measure the uncertainty of the matching process is to use Fréchet bounds. This works for categorical variables. First, we estimate conditional bounds of relative frequencies for specific target information. We do this for a contingency table of net wealth levels *Y* on total pension credits

[3] *For a detailed description of the variables underlying the matching process, see section 2.2.*

added to an individual account Z. As we need categorical information about these continuous variables, we group the observations into five categories.

For each of the 25 cells (j=1...5,k=1...5), we can then estimate a lower and upper bound for the corresponding share of persons (see D'Orazio et al., 2006) and the accompanying R program StatMatch[4] by

$$p^{(low)}_{Y=j,Z=k} = \sum_{i=1}^{I} p_{X=i} \max\,(0; p_{Y=j|X=i} + p_{Z=k|X=i} - 1)$$

and

$$p^{(up)}_{Y=j,Z=k} = \sum_{i=1}^{I} p_{X=i} \min\,(p_{Y=j|X=i}; p_{Z=k|X=i}).$$

Additionally, we can look at the relative frequency under the CIA, which is calculated by

$$p_{Y=j,Z=k} = p_{Y=j|X=i} * p_{Z=k|X=i} * p_{X=i}.$$

In doing so, we can evaluate the uncertainty inherent in the statistical matching process. For this to work properly, the marginal distribution of observations in both the donor and the recipient data needs to be similar. If this is not the case to the extent necessary, we need to harmonize the distribution across our data. As proposed in the literature, we can achieve this alignment by means of a multidimensional iterative proportional fitting estimation.[5]

2.2 Matching variables

For the CIA to hold, we gather all information related to the target variables on pension wealth which can be found in both the SSR 2020 and the HFCS datasets. This includes occupational information, such as income, type of employment contract as well as gender and age. One caveat is that we were not able to obtain information on how long people were registered with the social security system and on how long they paid into the system.

However, we can use the following variables for the matching procedure:

Geographical information

While the SSR 2020 contains information on postal codes, the HFCS provides information on the community identification number, a number sequence allowing for the identification of cities, villages, municipalities as well as districts in Vienna. The information sets provided do not overlap completely, as some cities and villages have more than one postal code. However, since there is always one main postal code, i.e. the one where the municipality is located, this code is selected to achieve a one-to-one assignment between the donor and the recipient data.

[4] *The entire matching exercise as well as the estimation of the Hellinger distance and Fréchet bounds are done using the StatMatch package (for further details, see https://cran.r-project.org/web/packages/StatMatch/StatMatch.pdf (accessed on February 22, 2021)). We gratefully acknowledge the contribution to our work.*

[5] *For further information, see the mipfp R package (Barthélemy and Suesse, 2018) underlying our estimation procedure.*

Gender

Both the donor and the recipient data include information on people's gender (female/male); we can therefore use this information in the matching process.

Age

The donor data provides relatively precise age categories for economically active persons (see table 4).

Table 4

Economically active persons by year of birth

Year of birth	Observation
1955–1961	376.203
1962–1966	650.771
1967–1971	609.732
1972–1976	515.086
1977–1981	508.914
1982–1986	537.223
1987–1991	525.346
1992–1996	412.004
1997–2001	52.436

Source: SSR 2020.

Social security institution

From the SSR 2020, we can moreover use information on the social security institution people are insured with. This information is closely connected to people's working arrangements, as employees, civil servants (for the federal government and the provincial governments) and the self-employed (farmers and others) are covered by separate institutions. Information on civil servants cannot be broken down by federal and regional levels and are thus aggregated to one category (civil servants). For further details, see section 1.2 on SSR 2020 data in this paper.

From the HFCS, we use information on respondents' occupation and workplace as well as related contractual arrangements.

Income

Information on income refers to the year 2016 in both datasets. Based on the pension credits for 2016, as indicated in the SSR 2020, we can calculate a measure of income for individual *i* (inc_i), which is defined as

$$inc_i \equiv \begin{cases} 0 & if\ PC_i = 0 \\ \frac{PC_i}{0{,}0178} & if\ 0 < PC_i < 1.211{,}11. \\ 68.040 & else \end{cases}$$

This measure reverses the rules regulating contribution payments, yielding information on income levels. Below an annual income of EUR 5,820 no pension benefits are paid out, and above an annual income of EUR 68,040 no additional social security contributions have to be paid. Thus, we do not obtain any information on income levels below and above these two thresholds and need to make assumptions about people either having no income or having an income equivalent to the upper threshold level. Within the two thresholds, we obtain precise information on people's income levels.

The HFCS offers information on income in different forms. To be as close as possible to the basis for calculating pension credits (i.e. to the measure of income introduced above), we take gross annual income from

- employment;
- self-employment;
- unemployment benefits; and
- private business (household level).

Table 5

Economically active persons by income groups

	Lower threshold	Upper threshold
Low	×	5,820
Decile 1	5,820	9,800
Decile 2	9,800	16,000
Decile 3	16,000	20,800
Decile 4	20,800	24,700
Decile 5	24,700	29,000
Decile 6	29,000	33,100
Decile 7	33,100	37,600
Decile 8	37,600	43,600
Decile 9	43,600	53,000
Decile 10	53,000	68,040
High	68,040	×

Source: Authors' calculations based on deciles used in the SSR 2020.

Note: Upper thresholds are included in the deciles with the exception of the upper threshold in income group „high". In decile 10, the upper threshold is excluded as well.

Due to the fact that there is a lower and upper limit for pension credits, we construct an indicator for low and high income levels.

For income levels in-between, we calculate ten income deciles based on the information from the SSR 2020 and transfer the data to the HFCS. Decile thresholds are given by the values reported in table 5.

2.3 Data alignment

SSR data were gathered from the official register in 2020. HFCS data were obtained from the third survey wave which was conducted in 2017. As regards the two datasets, there are several distinctions that are worth discussing:

Survey vs. register

HFCS data come with household weights, since the survey covers a small fraction of the target population. Consequently, estimates for the total population need to be weighted. As for SSR 2020 data, we can rely on the full range of observations made for the entire target population in Austria. Thus, simple calculations provide population estimates. We therefore present both weighted and unweighted results for the HFCS.

Individual weights are not available. Hence, even when we present weighted estimates based on individual persons, we use household weights and refrain from any adjustments.

Person vs. household

The reference unit used in the SSR is an individual, in the HFCS a household. Since a lot of information (related to individuals' occupation in particular) is available at the personal level, we can conduct the statistical matching exercise at this level. After the matching is completed, the information is aggregated to the household level in order to take into account all survey characteristics and make results comparable.

Multiple imputation

As is common in surveys, item non-response indicates incomplete observations. To obtain unbiased estimates, the HFCS applies multiple imputations that are based on a Bayesian imputation procedure (for further details, see Albacete et al. (2018)). For the data matching, each implicate for every person in the HFCS is taken as a separate observation to account for the imputation structure and the uncertainty modeled therein. This, however, means that one person in the HFCS can be matched to several different persons in the SSR using implicates.

Timing

We aligned the information on total pension credits accumulated on individuals' notional accounts with a cutoff period ranging from the beginning of 2017 to the first half of the HFCS 2017 field phase. In line with the reference period for income, we also have information on individuals' pension credits for 2016. Initial pension credits are fixed (beginning of 2014), as is individuals' year of birth. This information will therefore not change over time. The remaining data taken from the SSR are available for 2020; caveats e.g. due to people changing their main residence or occupation in the period from 2017 to 2020 are acknowledged.

2.4 Descriptive statistics

In this section, we introduce and discuss some key descriptive statistics of the datasets used.

Geographical information

Table 6 shows that the share of economically active persons, broken down by provinces, is similar in size across the two datasets. Results for the HFCS are split into unweighted and weighted results using household weights.

As can be seen from the unweighted HFCS results, there are relatively more economically active persons in Vienna than in Tyrol and Vorarlberg as well as Carinthia. This is due to the sampling method used for the HFCS. It neatly fits that the weighted results converge toward the results obtained for the SSR 2020 dataset. Annex 1 shows similar results for persons receiving pension income.

Gender

In the HFCS, both sexes are about equally represented. In the SSR, there are more men than women. This slight difference might be due to female workers whose earnings were below the threshold for mandatory social security contributions but who stated in the HFCS that they worked. For the data matching, this difference

Table 6

Share of economically active persons by provinces

Postal code	SSR	HFCS	
		Unweighted	Weighted
	%		
Vienna [1XXX]	22.3	24.2	21.1
Lower Austria (Northern Burgenland) [2XXX]	9.7	7.3	8.8
Lower Austria [3XXX]	8.2	7.7	9.0
Upper Austria [4XXX]	15.0	17.5	16.9
Salzburg [5XXX]	8.3	9.5	8.6
Tyrol and Vorarlberg [6XXX]	13.4	10.6	11.9
Southern Burgenland [7XXX]	3.1	4.0	3.3
Styria [8XXX]	12.8	13.2	13.4
Carinthia [9XXX]	7.1	5.9	6.9

Source: HFCS 2017, OeNB; SSR 2020.

does not seem to be problematic (for further details, see also the Hellinger distance in section 2.5).

The results for persons receiving pension income are shown in annex 1. As regards this share of the population, women are represented to a greater extent in the SSR than in the HFCS.

Table 7

Share of economically active persons by gender

	SSR	HFCS	
		Unweighted	Weighted
	%		
Female	45.5	50.4	50.3
Male	54.5	49.6	49.7

Source: HFCS 2017, OeNB; SSR 2020.

Table 8

Share of economically active persons by year of birth

	SSR	HFCS	
		Unweighted	Weighted
	%		
1955–1961	9.0	12.7	12.8
1962–1966	15.5	13.2	14.0
1967–1971	14.6	15.1	14.9
1972–1976	12.3	11.9	12.4
1977–1981	12.2	12.6	12.6
1982–1986	12.8	11.7	11.6
1987–1991	12.5	10.0	9.5
1992–1996	9.8	8.7	8.2
1997–2001	1.3	4.0	3.9

Source: HFCS 2017, OeNB; SSR 2020.

Age

Information on people's age is recorded in five-year age brackets indicating people's year of birth in the SSR 2020. The HFCS, in contrast, records respondents' exact age, thus allowing us to use this information without any restrictions and aggregate it to the level used in the SSR. People's year of birth was deducted from the information on age and the year in which the interviews were conducted for the HFCS.

As shown in table 8, the share of economically active persons, broken down by year of birth, is quite similar across the two datasets. The biggest differences can be detected at the tails, i.e. for the very young born between 1997 and 2001 and the elderly born between 1955 and 1961. For these age brackets, the shares in the HFCS are higher than in the SSR.

For persons receiving a pension (see table A3 in annex 1), we could use people's actual year of birth. Distributions were again similar across the two datasets. When restricting the years of birth to those observed in the HFCS, we find that there are more elderly persons in the SSR 2020.

Social security institution and occupational information

One of the key matching variables is the institution providing social security services. Table 9 indicates the distribution of economically active persons across social security institutions (for details on persons receiving pension income, see table A4 in annex 1).

A large majority of economically active persons is covered by the public pension insurance fund (PVA). Civil servants – split into those employed by the federal government and those employed by the provincial governments – are insured with the social security fund for public sector employees, railways and mining (BVAEB) and constitute a small group. For former civil servants, there is no information available. We therefore only look at employed and self-employed persons, with the latter being covered by the social security fund for the self-employed and that for farmers (SVS-GW and SVS-LW).

Table 9

Share of economically active persons by social security institution

	SSR	HFCS	
		Unweighted	Weighted
	%		
Civil servants employed by the federal government	2.2	4.5	4.6
Civil servants employed by provincial governments	0.7		
PVA	88.3	86.1	85.3
SVS-GW	7.1	8.0	8.4
SVS-LW	1.8	1.4	1.7

Source: HFCS 2017, OeNB; SSR 2020.

Note: The category PVA includes observations from category BVAEB (railways and mining), which come to 1.1% of the SSR population.

At close to 10% (nearly twice as many are in retirement), the self-employed constitute a sizable minority, with the subgroup of farmers being rather small, however.

Distributions across the social security institutions are relatively similar for both datasets, which speaks to the quality of the HFCS sampling method.

Income

We include a measure of income in our statistical matching exercise, as pension benefits in Austria are based on people's income levels.

Table 10 shows the results obtained from indirect income measurements for the SSR 2020 as well as from income information corresponding as closely as possible to these measurements for the HFCS. More specifically, table 10 presents the share of persons whose earnings are below (above) the threshold for mandatory (capped) social security contributions as well as the distribution of income levels from the 1st percentile to the 99th percentile. The share of persons whose income exceeds EUR 68,040 is almost identical for both datasets. There is a large fraction of people in the SSR with low income (about 16%). This might be due to the fact that people who were economically active and paid social security contributions in the past are kept in the register until pension benefits are paid out, even though they might not be economically active at the moment.

Table 10

Economically active persons by income levels (basis for pension credits)

	SSR	HFCS	
		Unweighted	Weighted
	Share in %		
Low	16.3	8.7	9.2
High	4.8	4.7	5.2
	Distributional information in EUR thousand		
Mean	27.1	30.1	30.5
P1	0.0	0.0	0.0
P10	0.0	7.2	6.8
P20	5.5	14.0	13.9
P30	14.6	19.1	18.7
P40	20.8	23.0	22.9
P50	25.9	26.6	27.0
P60	31.2	30.6	30.7
P70	36.7	35.0	35.4
P80	44.2	40.8	41.8
P90	57.6	53.1	53.6
P99	68.0	113.0	121.6

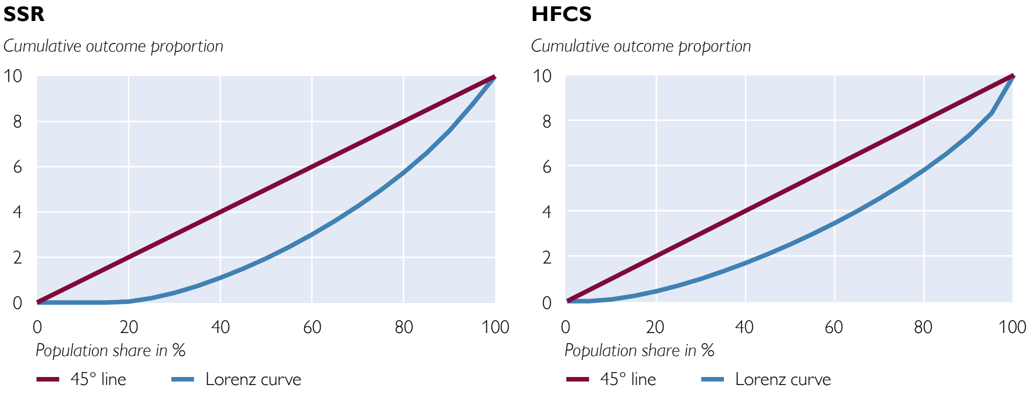
Source: HFCS 2017, OeNB; SSR 2020.

Note: In the SSR, income is estimated indirectly and pertains to the personal level. In the HFCS, not all income measures are included (e.g. pension income is not part of the basis for pension credits).

When looking at the distribution of income levels, we find similar results for both datasets, except for the values at the top, which are higher for the HFCS than for the SSR. Furthermore,

Chart 1

Lorenz curve of income measure



Source: HFCS 2017, OeNB; SSR 2020.

chart 1 shows the Lorenz curves of income distribution. Due to the large fraction of people without pension credits in the SSR 2020 dataset, the corresponding Lorenz curve deviates more from the 45 degree line than is the case in the HFCS dataset. This implies a higher degree of inequality in the former dataset.

2.5 Similarity between matching variable distributions

As indicated in section 2.1, the HD is a measure of similarity between the matching variable distributions. Table 11 shows the results for the HD and the dissimilarity index.

Both the HD and the dissimilarity index point in the same direction, namely that observations seem to be distributed in a relatively similar manner for each indicator. In the literature (see European Commission, 2013), an HD below 0.05 is considered to be desirable. Our findings are close to that level and, in the case of the gender variable, even below that threshold.

The combination of the two indicators (see last line in table 11) shows that the finer the stratification, the higher the dissimilarity of the distribution of observations across these strata. Overall, it seems that we have already achieved a relatively precise stratification. We must keep in mind that the measure of income (i.e. the basis for pension credits) is used differently in two distinct matching procedures.

Table 11

Similarity between matching variable distributions for economically active persons

	Dissimilarity index	HD
Geographical information	0.068	0.056
Gender	0.049	0.035
Age	0.074	0.082
Social security insitution	0.051	0.072
Basis for pension credits	0.077	0.083
Combined	0.435	0.478

Source: HFCS 2017, OeNB; SSR 2020.

2.6 Data challenges associated with matching

While we can rely on the full range of observations from the SSR, we need to perform aggregations in the matching categories for the HFCS.

The combinations of matching variable categories (geographical information, gender, age, social security institution, measure of income) yield a total of

$$9 \text{ X } 2 \text{ X } 9 \text{ X } 5 \text{ X } 12 = 9.720$$

strata that are divided into the observations for the recipient and the donor data. Not every stratum is occupied in the two datasets. In the HFCS, we find 2,044 occupied strata, with a median (mean) number of 5 (8.3) observations per strata.[6] In the SSR, these figures come to 8,496 occupied strata, with a median (mean) number of 50 (492.8) observations per strata. These are desired results in the sense that enough strata with many observations are occupied in the donor data. Despite these relatively large numbers, there are 421 strata with only 1 complete record and 968 strata with 2 to 5 complete records.[7] Due to the fact that the full range of data is available from the SSR, the strata with few donor observations are not problematic.

There are, however, occupied strata in the HFCS that cannot be found in the SSR. In this context, we must keep in mind the imputation structure of the HFCS, i.e. the Bayesian-based multiple imputation procedure with chained regression. In a specific implicate, there might be an observation which cannot be found in the SSR. To avoid missing observations that need to be excluded from the analysis, these observations are aggregated to another category. This is particularly the case for young civil servants as well as young farmers. In total, only about 20 persons are concerned who mostly belong to one implicate of the multiple imputations. These persons are aggregated to employed persons, i.e. they are matched to information from the PVA.

In view of the above, we choose two versions of a random hot deck procedure to match data from the SSR (donor) to the HFCS (recipient). Both versions are one-to-one matching procedures, with one observation in the donor data being matched to a specific observation in the recipient data. Every implicate for every person in the HFCS is taken as a separate observation. The main distinction between the two versions is the treatment of our income measure. For what we refer to as matching I, we take the income deciles including the lowest (no pension credits) and highest income category (no additional social security contributions), so that we only have discrete matching variables. For matching II, we consider income as continuous information and use the Manhattan metric distance function, which allows us to identify the seven closest matching observations in the SSR. Of these, one observation is taken randomly and matched to an observation in the HFCS.

2.7 Matching uncertainty – Fréchet bounds

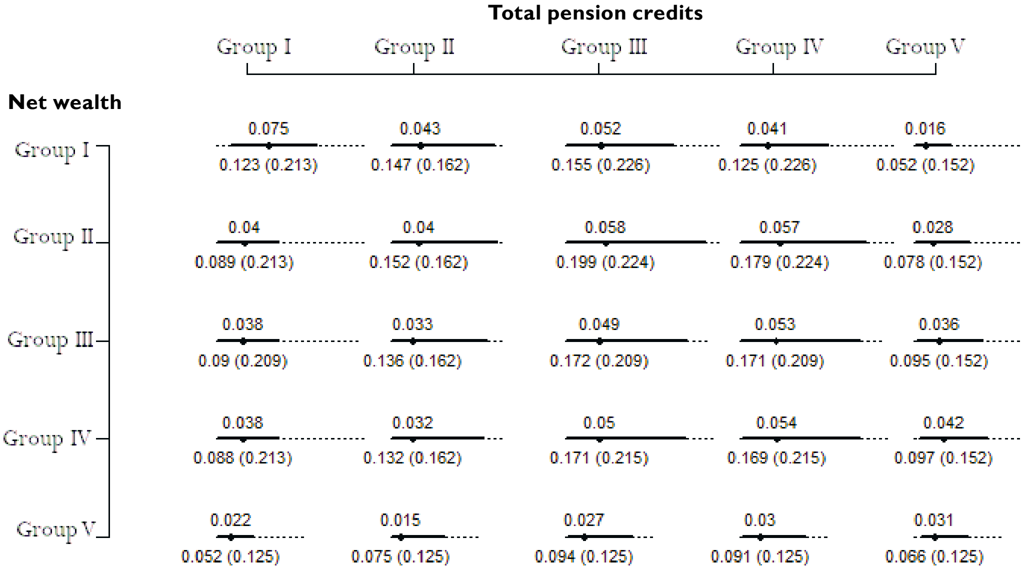
For this exercise, we split people by household wealth levels and total pension credits into five groups ranked by ascending values. The groups are defined in absolute terms rather than by a specific share of persons to allow for a varying share of persons in each group. Group thresholds are set at 12,000, 50,000, 200,000 and 500,000 persons for the HFCS, and at 1,500, 4,000, 10,000 and 20,000 persons for the SSR. The combination of these five groups for each dataset gives us a 5 by 5 matrix of shares.

[6] *For this kind of information, each implicate of the imputation procedure is taken as a distinct observation. Thus, the mean and median can be divided by five.*

[7] *All the results as well as specific information on retired persons are provided in table A6 in annex 2.*

Chart 2

Fréchet bounds



Source: HFCS 2017, OeNB; SSR 2020.

Chart 2 shows the shares for all groups and their variability. In annex 1, table A1 reports similar results for persons in retirement, cross-tabulating net wealth and income levels.

3 Conclusionary remarks

In this paper, we show that we can enhance the analytical potential of existing data sources by employing different data matching techniques. In doing so, we moreover demonstrate that the related challenges, particularly as regards conceptual questions, are abundant.

References

Albacete, N., S. T. Dippenaar, P. Lindner and K. Wagner. 2018. Eurosystem Household Finance and Consumption Survey 2017. Methodological notes for Austria. In: Monetary Policy & the Economy Q4/18. OeNB. https://www.oenb.at/dam/jcr:98e41d22-0b43-4173-aa97-bcca7ff703b7/mop_%20q4_18_addendum.pdf

D'Orazio, M., M. Di Zio and M. Scanu. 2006. Statistical Matching: Theory and Practice. West Sussex, England: John Wiley & Sons Ltd (Wiley Series in Survey Methodology).

European Commission. 2013. Statistical matching of EU-SILC and the Household Budget Survey to compare poverty estimates using income, expenditures and material deprivation. Luxembourg: Publications Office of the European Union.

Fessler, P., P. Lindner and M. Schürz. 2018. Eurosystem Household Finance and Consumption Survey 2017 for Austria. In: Monetary Policy & the Economy Q4/18. OeNB. 36–66. https://www.oenb.at/dam/jcr:8b5436a3-72a2-415a-a0b3-ea476c371a0c/4_mop_%20q4_18_screen.pdf

Fessler, P. and M. Schürz. 2018. Private Wealth Across European Countries: The Role of Income, Inheritance and the Welfare State. In: Journal of Human Development and Capabilities 19(4). 521–549.

Barthélemy, J. and T. Suesse. 2018. mipfp: An R Package for Multidimensional Array Fitting and Simulating Multivariate Bernoulli Distributions. In: Journal of Statistical Software, Code Snippets 86(2). 1–20. DOI: 10.18637/jss.v086.c02.

OECD. 2013. OECD Framework for Statistics on the Distribution of Household Income, Consumption and Wealth. https://www.oecd-ilibrary.org/docserver/9789264194830-en.pdf?expires=1627895336&id=id&accname=id5760&checksum=0EB3675F7231CA1C-DBEE379B55D3F9F1

Annexes

Table A1

Share of persons receiving a pension by provinces

	SSR	HFCS	
		Unweighted	Weighted
Postal code	%		
Vienna [1XXX]	18.0	25.0	21.7
Lower Austria (Northern Burgenland) [2XXX]	11.4	11.2	13.5
Lower Austria [3XXX]	9.2	6.6	7.5
Upper Austria [4XXX]	16.2	11.8	11.8
Salzburg [5XXX]	7.8	11.3	9.4
Tyrol and Vorarlberg [6XXX]	11.1	9.2	10.6
Southern Burgenland [7XXX]	4.1	3.5	4.0
Styria [8XXX]	14.6	15.4	14.8
Carinthia [9XXX]	7.7	6.2	6.8

Source: HFCS 2017, OeNB; SSR 2020.

Table A2

Share of persons receiving a pension by gender

	SSR	HFCS	
		Unweighted	Weighted
	%		
Female	60.3	53.0	52.9
Male	39.7	47.0	47.1

Source: HFCS 2017, OeNB; SSR 2020.

Table A3

Share of persons receiving a pension by year of birth

	SSR	HFCS	
		Unweighted	Weighted
	%		
1918	0.0	0.1	0.1
1922	0.2	0.1	0.1
1923	0.3	0.1	0.2
1924	0.3	0.2	0.3
1925	0.5	0.6	0.6
1926	0.7	0.3	0.2
1927	0.8	0.3	0.3
1928	1.0	0.8	0.9
1929	1.3	0.6	0.7
1930	1.5	0.7	0.9
1931	1.8	1.0	1.0
1932	1.9	0.8	0.9
1933	2.1	1.0	1.1
1934	2.3	1.5	1.4
1935	2.4	2.1	2.0
1936	2.7	1.9	1.9
1937	2.8	1.9	2.1
1938	3.3	2.2	2.3
1939	4.6	4.2	4.3
1940	4.8	5.0	4.8
1941	4.8	6.2	6.5
1942	4.4	4.5	4.6
1943	4.5	4.3	4.4
1944	4.6	3.8	3.7
1945	3.6	4.4	4.1
1946	4.4	4.5	3.9
1947	5.1	5.2	5.1
1948	5.1	5.7	5.5
1949	4.9	5.9	5.9
1950	4.9	4.8	4.6
1951	4.8	5.9	6.4
1952	4.6	7.8	7.6
1953	4.6	5.9	5.7
1954	4.4	5.6	6.0

Source: HFCS 2017, OeNB; SSR 2020.

Table A4

Share of persons receiving a pension by social security institution

	SSR	HFCS	
		Unweighted	Weighted
	%		
BVAEB (railways and mining)	1.6	85.7	85.5
PVA	80.6		
SVS-GW	8.9	14.3	14.5
SVS-LW	8.9		

Source: HFCS 2017, OeNB; SSR 2020.

Table A5

Share of persons receiving a pension by type of pension

	SSR
	%
Disability pension (blue- and white-collar workers as well as miners)	14.6
Disability pension (self-employed and civil servants)	1.5
Pension annuity	65.7
Widow's/widower's pension	17.9
Orphan's pension	0.2

Source: SSR 2020.

Table A6

Strata from matching variables

	Economically active persons	Persons receiving a pension
Number of strata	9.720	1224
Occupied in HFCS	2.044	530
Average number of observations	8,3	12,8
Median number of observations	5	10
Occupied in SSR	8.496	1224
Average number of observations	492,8	1140,9
Median number of observations	50	578
Only 1 donor	421	4
Only 2 to 5 donors	968	15
Missing matched	none	none

Source: HFCS 2017, OeNB; SSR 2020.

Notes: The average and median numbers of observations take each implicate as a separate observation.