Record Linkage of Birthbooks

Martin Karlsson

CINCH, University of Duisburg-Essen
University of Gothenburg

September 2, 2021

1 Introduction

This note describes how the 1910 cohort of the Swedish birth books – a data source originally compiled by Statistics Sweden, which is now publicly available on the internet – can be linked to a census constructed by combining the 1950 census and the deathbook. Our approach is based on probabilistic record linkage. Probabilistic record linkage normally requires the estimation of probabilities that a linked pair with certain characteristics is a true match; however, due to some special features of the Swedish data, these probabilities can be estimated as sample equivalents, based on a subsample of perfect matches.

2 Ground Truth Based on Married Couples

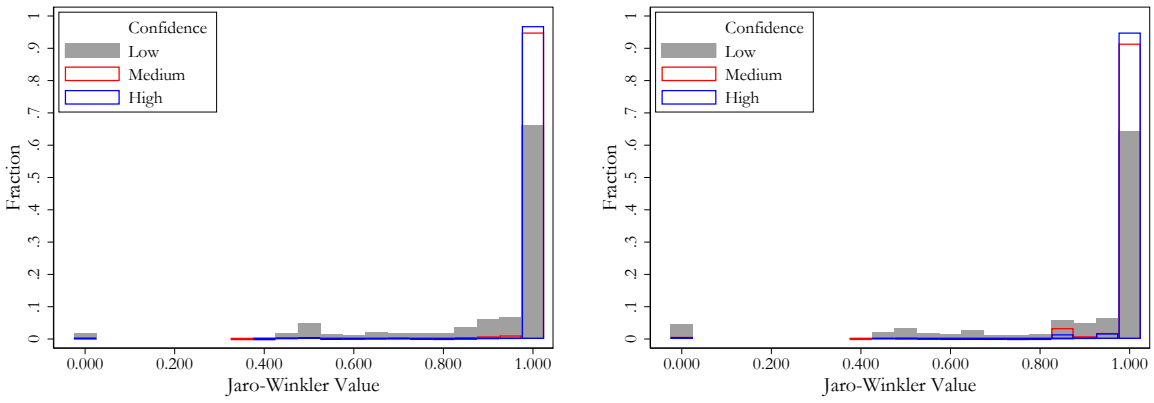
In order to assess the distribution of name match quality for true and false matches, a sample of married couples from the 1950 census was used. Compared to the parents of the 1910 births, this is a highly selective sample, considering that both spouses need to have survived 40 years after becoming parents. It is also selective with regard to marital status, since only parents who are married in 1910 and in 1950 will be included. However, there is no reason to suspect that these sample selection rules correlate systematically with the difficulties in transcribing names.

The sample of 1950 couples was constructed based the criterion that the joint birth dates of both spouses were unique. This applies to N% of all couples.

2.1 Transcription Confidence and Match Quality

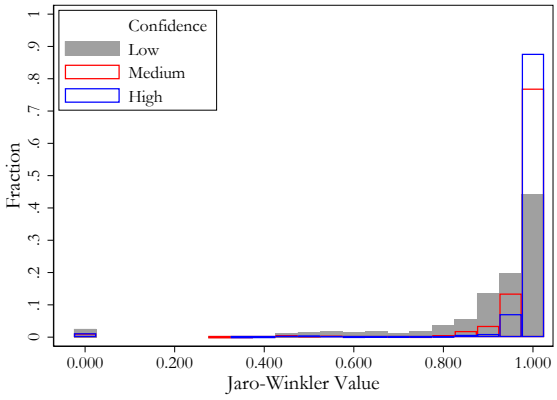
As a first check, we investigated whether the transcription confidence variables carry information relevant for match quality. In the sample of matched couples and for each name variable, we split the sample into terciles based on the confidence in the transcription. Figure 1 shows the result. Whereas the figures make very clear that the confidence variable carries some information – in the sense that the higher tercile have greater JW similarity, even the first tercile has very high similarity in general.

Since even records with poor confidence scores apparently carry information, we decided not to consider confidence in the sample selection.

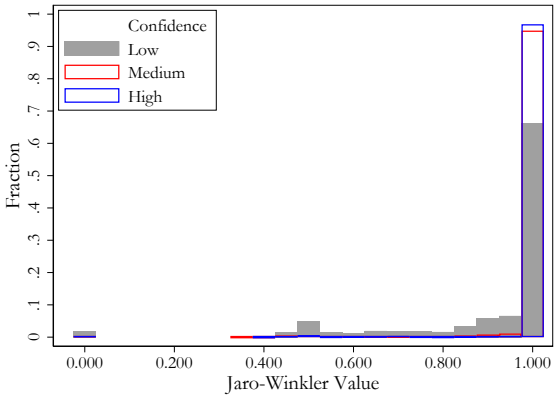


(a) Female First Names

(b) Male First Names



(c) Female Last Names



(d) Male Last Names

Figure 1: Jaro-Winkler Similarity Scores by Confidence.

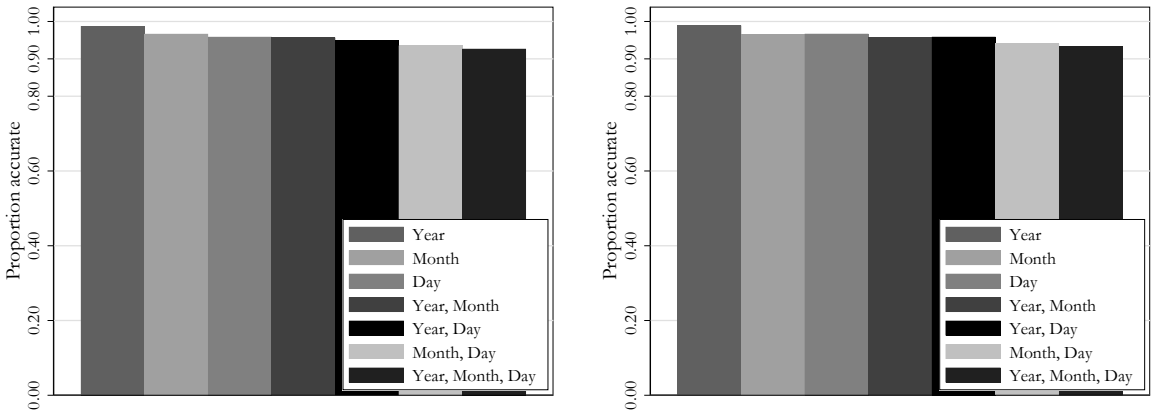
2.2 Accuracy of Transcribed Birth Dates

The dataset consisting of married couples was also used to assess the accuracy of transcribed birth dates. We used couples which were unique with regard to both spouse's first and last (maiden) names, and the birth date of one of the spouses. Thus, one of the spousal birth dates was left out in each round. When the mothers' birth dates were left out, a total of 10,591 parent couples from the birthbook could be matched. Figure 2 shows the proportion of matches that had different elements of the maternal birth dates correct.

According to these estimates, 92.7 per cent of maternal birth dates were completely accurate. 98.7 per cent had a correct birth year. The variable that was most often inaccurate was the day of the month, which was inaccurate in 4.1 per cent of cases. Thus, there is quite high accuracy in general, but there is clearly scope for improvement by allowing for birth date variables to be inaccurate. Among the 7.3 per cent mothers with an inaccurate birth date, 88 per cent had only one inaccurate variable. Therefore, it appears that most of the inaccurate birth dates gets captured if one of the birth date variables is allowed to be wrong.

3 Record Linkage Procedure

For children, the record linkage was carried out in five distinct steps. In principle, the parish of birth, which is available in both datasets, is a one-to-one mapping; thus in principle, it would



(a) Mothers (b) Fathers

Figure 2: Accuracy of Transcribed Birth Dates.

be possible to require an exact match on birth parish. However, there are minor discrepancies between the two in some cases, for instance due to institutional deliveries, which are also included in the birth books. If records with differing parish of birth information are not discarded, it will be possible to create a variable representing institutional delivery, and to correctly assign the parents' actual parish of residence to each child.

Therefore, we did not require a perfect match on birth parish, but instead considered a distance cutoff of 35-60 kilometres.[1]

The five rounds of record linkage are defined as follows.

1. Exact match on **birth date**, **region** (län); fuzzy match on up to **two forenames**; distance radius **35 km**.

2. Exact match on **birth date**, **initials of forenames**; fuzzy match on **parental surname**; distance radius **35 km**.

3. Exact match on **birth parish**; fuzzy match on **month of birth**, **day of birth**, and up to **two forenames**.

4. Exact match on **parental surname** and **month of birth**; fuzzy match on up to **two forenames** and **day of birth**; distance radius **60 km**.

5. Exact match on **parental surname** and **day of birth**; fuzzy match on up to **two forenames** and **month of birth**; distance radius **60 km**.

All steps were carried out for the entire sample, with the exception of step 3 which, for computational reasons, was only carried out for previously unmatched birthbook records. Whenever parental surnames were used, priority was given to the paternal surname; the maternal surname was used only in case the paternal surname was missing. This procedure thus emulates the conventions used in practice: children would derive their surname from their fathers, unless the mother was single. In the latter case, the father would not be entered in the birthbook. In step 2, the transcribed paternal surnames were used, whereas in steps 4-5 the linked surnames of parents (see below) were used. Since in those steps, the surnames were taken from the same source, an exact match could be required.

[1]This criterion was not applied to the 4 per cent of individuals in the census dataset who do not have a birth parish recorded; for them the criterion was instead that they were born in the same region.

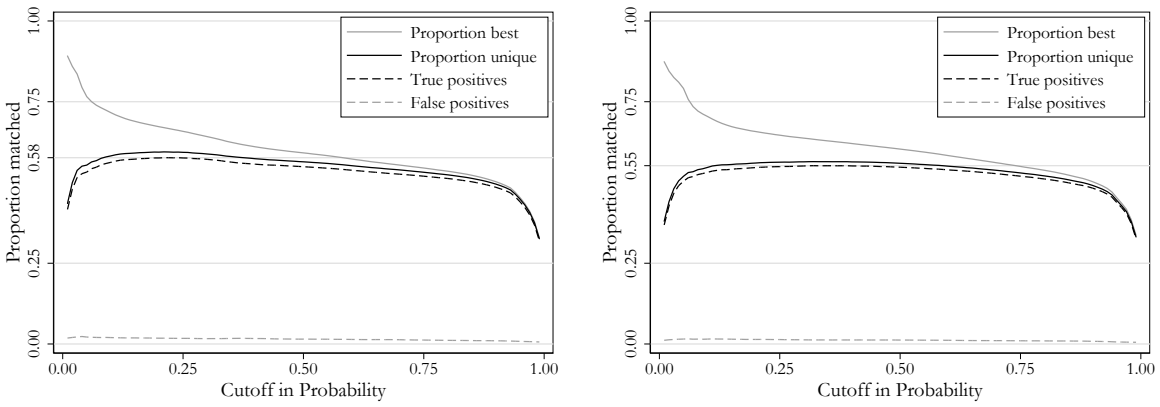
For **mothers** and **fathers**, the record linkage was carried out in four distinct steps:

1. Exact matching on **sex**, **birth year** and **birth month**, fuzzy matching on **birth day**, **first name**, **last name and place of residence**.
2. Exact matching on **sex**, **birth day** and **birth month**, fuzzy matching on **birth year**, **first name**, **last name and place of residence**.
3. Exact matching on **sex**, **birth year** and **birth day**, fuzzy matching on **birth month**, **first name**, **last name and place of residence**.
4. Importing the identity of the spouse from census 1910.
5. Importing the identity of the spouse from census 1950.

4 Record Linkage Performance

4.1 Parents

For mothers, the first iteration led to 59 per cent of the 153,000 complete records being matched. This result was achieved by setting a cutoff in the estimated match probability at 0.20; accepting as a true link any record for which the best match has probability at least 0.20, whereas the second best match has probability less than 0.2. The result for any level of the cutoff is provided in Figure 3. A probability of 0.2 might sound very low, but an analysis of the false positive rate[2] suggests it is not a concern: the false positive rates are low throughout, among the 57 per cent of mothers that got matched, at least 96 per cent were estimated to be correct true positives (cf. Figure 3b). For fathers, the link rates are slightly worse, which is to be expected, given that the fathers were not entered in case the mother was single. The first round thus led to 55 per cent of fathers being linked, and also for them, the estimated false positive rate was very low.



(a) Mothers (b) Fathers

Figure 3: Record Linkage Performance for Parents: Round 1.

Table 1 summarises the outcome of the linkage in all steps: the two leftmost columns provides statistics for each individual round, and the three rightmost columns provide cumulative statistics. The overall false positive rate is estimated at between 2.5-3.1 per cent. In 4-6 per cent

[2]False positive rates were estimated using the sample of married couples described in Section 2 above.

of cases, there is at least one conflict among the matched cases. The two statistics are not in disagreement with each other since the false positive rate is based on the entire sample, whereas the conflict rate is based on individuals who got matched in each round.

Table 1: Record Linkage Results: Parents

	PERFORMANCE				
	Round		**Cumulative**		
Round	Matched	FP	Matched	Conflict	FP
MOTHERS					
1	0.594	0.030	0.594	0.000	0.030
2	0.596	0.030	0.618	0.001	0.033
3	0.588	0.028	0.634	0.022	0.035
4	0.431	0.009	0.708	0.050	0.033
5	0.171	0.001	0.726	0.057	0.031
FATHERS					
1	0.564	0.022	0.564	0.000	0.022
2	0.567	0.022	0.586	0.001	0.025
3	0.546	0.021	0.602	0.026	0.026
4	0.173	0.011	0.638	0.058	0.026
5	0.130	0.001	0.645	0.038	0.025

4.2 Children

For children, there is no information to go by to estimate false positive rates; however, since the information available for children is richer – the parish of birth is available in both datasets – and likely to be more accurate as well – birth dates of children allows perfect matching on birth year – it is likely that the false positive rates are lower than those estimated for parents.

The outcome of the first round of record linkage procedure for children is presented in Figure 4. Overall, 80 per cent of birthbook records get linked in the first round.

Table 2 summarises the outcome of the linkage of children. In total 88.5 per cent of children get linked, and the conflict rate between the rounds is very low.

5 Linked Dataset – Descriptives

A desirarble feature of the linked dataset is that it is representative of the target population. Therefore, we now present descriptive statistics for how the linked sample compares to the overall population. More specifically, we compare three distinct samples:

1. The universe of individuals who were born in 1910 who are present in either the deathbook or in the 1950 census.

2. Individuals from sample (1) with at least one link in the birthbook dataset.

3. Individuals from sample (2) with at least one parental link in the birtbook dataset.

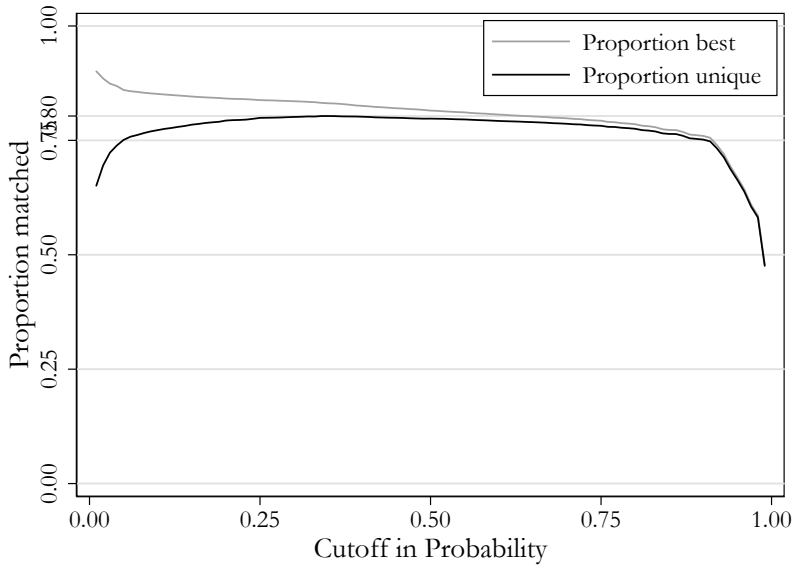


Figure 4: Record Linkage Performance: Children, first round.

Table 2: Record Linkage Results: Children

	PERFORMANCE		
	Round	**Cumulative**	
Round	Matched	Matched	Conflict
1	0.803	0.803	0.000
2	0.455	0.821	0.005
3	0.055	0.875	.
4	0.341	0.884	0.006
5	0.336	0.885	0.004

Table 3 compare the three samples for some characteristics; we provide the number of individuals with non-missing information, the mean of each variable, the difference between the linked samples and the target population, and the standardised difference. Overall, the representativeness appears to be very good. The largest discrepancy is noted for household head marital status in the child and parent sample: 92.8 per cent of household heads are married in this sample, compared to 91.1 per cent in the population. This discrepancy is to some extent expected, given that it is easier to find at least one parental link if the parents are married.

Table 3: Representativeness

	Population		Child Linked				Child and Parent Linked			
Variable	N	Mean	N	Mean	Δ_{12}	Std. Dif	N	Mean	Δ_{13}	Std. Dif
Individual Characteristics										
Male	134,010	0.516	119,552	0.516	-0.000	-0.001	101,027	0.518	-0.002	-0.004
Age at Death	131,717	64.720	119,021	65.106	-0.386	-0.013	100,605	65.617	-0.897	-0.031
Urban	134,035	0.300	119,559	0.303	-0.003	-0.007	101,030	0.282	0.018	0.040
Household Characteristics										
Family Size	109,976	5.440	102,066	5.435	0.006	0.002	87,544	5.488	-0.048	-0.020
N Kids	109,976	3.292	102,066	3.287	0.005	0.002	87,544	3.339	-0.047	-0.021
Household Head Characteristics										
Age	109,976	36.650	102,066	36.588	0.062	0.006	87,544	36.414	0.236	0.022
Married	109,976	0.911	102,066	0.912	-0.000	-0.001	87,544	0.928	-0.016	-0.060
Widow(er)	109,976	0.040	102,066	0.039	0.001	0.003	87,544	0.035	0.005	0.026
Farmer	109,976	0.415	102,066	0.411	0.005	0.009	87,544	0.426	-0.011	-0.022
Industrial Worker	109,976	0.445	102,066	0.449	-0.004	-0.008	87,544	0.443	0.002	0.005
Retired	109,976	0.001	102,066	0.001	0.000	0.003	87,544	0.001	0.000	0.005
Labour Force	109,976	0.949	102,066	0.950	-0.000	-0.001	87,544	0.957	-0.007	-0.035
Disabled	109,976	0.002	102,066	0.002	0.000	0.001	87,544	0.002	0.000	0.006