Social Science History (2020), **44**, 555–570
doi:10.1017/ssh.2020.15

CAMBRIDGE
UNIVERSITY PRESS

SPECIAL ISSUE ARTICLE

Selection Bias Encountered in the Systematic Linking of Historical Census Records

Luiza Antonie[1], Kris Inwood[2,3,*], Chris Minns[4] and Fraser Summerfield[5]

[1]School of Computer Science, University of Guelph, [2]Department of Economics and Finance, University of Guelph, [3]Department of History, University of Guelph, [4]Department of Economic History, London School of Economics and Political Science and [5]Department of Economics, St Francis Xavier University
*Email: kinwood@uoguelph.ca

ABSTRACT
Linked historical records typically are unrepresentative of the population from which they are drawn even if the method of linking is restricted to time-invariant matching criteria. An example drawn from Canadian census records illustrates the nature of bias that may afflict even a carefully linked sample. The use of potentially time-varying match criteria doubles the size of a linked sample at a modest cost in terms of additional bias. This trade-off is attractive for some research purposes if care is taken in the uses to which the data are put. Reweighting to mitigate the effects of bias in visible characteristics is desirable.

Introduction

Inexpensive computing and the mass digitization of records have enabled the emergence of innovative "big data" approaches to history and the social sciences (Bloothoft et al. 2015; Fourie 2016; Gutmann et al. 2018; Maxwell-Stewart 2016). A second wave of studies reduces reliance on a single source by linking together multiple series with "machine learning" techniques adapted from computing and information science (Feigenbaum 2018; Ruggles et al. 2018). The linking of independently generated data using sophisticated algorithms constitutes an important new source for the study of social mobility, intergenerational and early life influences on adult health, and many other topics. Methodologies used for the systematic linking of records have been carefully developed over the past 50 years (Christen 2012; Feigenbaum 2016; Fellegi and Sunter 1969; Ferrie 1996; Winkler 2006).

Census data afford a useful example. Historical census data, while not perfect, offer the most comprehensive and unbiased representation of many past populations (Hacker 2013; Thorvaldsen 2017). A well-designed census enumeration characterizes many aspects of a population in a representative manner. The most common method for identifying the same person in successive censuses is to identify one *and only one* record in both years with the same name, sex, birth year,

https://doi.org/10.1017/ssh.2020.15 Published online by Cambridge University Press

birthplace, and a consistent marital status (Goeken et al. 2011). Records are connected using these time-invariant characteristics to minimize the risk of biasing the linked sample. If we were to rely on a characteristic that might change, for example occupation, the linked data would underrepresent those who changed occupation. Limiting the match criteria to time-invariant characteristics avoids this problem (Ruggles 2006).

Three potential problems are associated with any implementation of this approach. There is a risk of "false positive" matching, or incorrectly connecting records of different people with near-identical characteristics. A second potential weakness is that the number of unique and exact matches can be small. A low rate of unique matching does not handicap studies of the entire population of a large country such as the United States, but research focusing on smaller countries or on particular subgroups may be constrained by the small number of linked records. A third potential weakness is unrepresentativeness. Are the linked records a balanced representation of the population from which they are drawn?

The three problems are interconnected, insofar as a solution for one problem may aggravate another. For example, relaxing the criteria for identification of the same name, sex, and birthplace will expand the number of links but it also increases the incidence of false positive links and of records linked multiply (rather than uniquely). Alternately, we might expand the number of linked records if we match using additional characteristics that change over time. Here, the risk is that our linked sample will overrepresent people whose characteristics, while mutable, do not in fact change. In this article we focus on the second trade-off. Can the benefit of a larger sample exceed the cost of lost representativeness? Our answer is a "qualified yes."

Our Approach

We explore the selection bias generated by systematically linking "full count" census records of the entire Canadian population collected at 10-year intervals: 3.5 million records in 1871 rising to 5.4 million in 1901. First, we identify the same person in multiple enumerations using a small set of time-invariant individual characteristics (birth year and place, sex, name) to minimize bias from the linking process (Antonie et al. 2014, 2015). Our method, which uses an initial set of known or true links and the classification of all possible matches with support vector machine software, is broadly representative of a commonly used approach to historical record linkage (Christen 2012).

With this approach we can identify a unique match in 1881 for less than one-fifth of the 1871 Canadian population. More than one-half of the 1871 records are linked multiply. By this, we mean the 1871 record is matched to more than one 1881 record or it is one of several 1871 records matched to a single 1881 record. Thus, more than half of the 1871 records cannot be used for longitudinal analysis because of an inability to identify which of the multiple matches is the right one. A possible next step would be to use additional information to select the correct match from among the set of multiple or potential links.

https://doi.org/10.1017/ssh.2020.15 Published online by Cambridge University Press

Table 1. The problem of multiple potential links

Last Name	First Name	Gender	Age	Birthplace	Marital Status
1871 census records					
Barns	Mary	Female	11	Ontario	single
Barns	Mary	Female	9	Ontario	single
Barns	Mary	Female	8	Ontario	single
Barns	Mary	Female	12	Ontario	single
Barns	Mary	Female	10	Ontario	single
Barns	Mary	Female	10	Ontario	single
1881 census records					
Barns	Mary	female	20	Ontario	single
Barns	Mary	female	22	Ontario	Single

We take this step of introducing broader criteria that will recover additional unique links among records that are matched multiple times. The example of Mary Barns in table 1 illustrates our inability to determine the correct link whenever more than one person (in either census year) shares a common age, birthplace, name, and sex. The only way to discriminate among multiple potential links is to rely on additional information. We describe this process with a term borrowed from computing science: *disambiguation* (ibid.).

We follow other researchers in the choice of family coresidence as a criterion for disambiguation (Fu et al. 2011, 2014). For example, there might be many records for a "Mary Barns" with the same age and birthplace, but only one of them lived in the same household as a sister named Anastasia. While Mary and Anastasia remained in the same family over the decade, we can disambiguate among the multiple matches. A generalization of this method, described in detail elsewhere (Richards 2013), relies on the Jacquard similarity measure. It roughly doubles the size of linked sample with no change in the risk of false positive or mistaken matching (table 2).[1]

The disadvantage of disambiguation with family coresidence is that it creates a bias to families or portions of families that remain together. We can follow Mary and Anastasia from 1871 to 1881 if and only if they are the kind of sisters who continued to live together over 10 years. It will not be possible to use this method for families in which members do not remain together. Consequently, linked data that have been disambiguated in this way will overrepresent people in families that maintain stable patterns of coresidence. Our goal in this article is to assess the nature and extent of this deviation from representativeness.

The problem of representativeness plagues all longitudinal data because not everybody survives over time, and death tends to be selective. Modern longitudinal samples obtained from repeat surveys of an initially representative population

[1] A Jacquard similarity measure identifies the total number of items in the intersection of the two households (in different years) and divides it by the total number of items in the union of the two households.

https://doi.org/10.1017/ssh.2020.15 Published online by Cambridge University Press

Table 2. Number of linked records and link rate, with and without disambiguation

	No. of Linked Records	Link Rate
Linking with time-invariant individual characteristics		
1871–81	550,726	16%
1881–91	635,161	15%
1891–1901	712,318	15%
After disambiguating with coresident family members		
1871–81	1,103,713	32%
1881–91	1,209,865	28%
1891–1901	1,265,998	26%

typically lose representativeness through selective attrition originating with the death or disappearance of some subjects. The linking of records from historical or administrative sources shares this problem, and in addition has other complications that are peculiar to the historical source. As is well known, even if we restrict the historical link criteria to a small set of time-invariant characteristics, we still expect to be more successful in linking historical people with uncommon names and the kind of people who report more precisely to the census enumerator (Antonie et al. 2015; Bailey et al. 2020b; Ferrie 1996).

Representativeness and Disambiguation with Family Data

Our question in this article is the extent to which representativeness diminishes through disambiguation. During the late nineteenth century the Canadian government enumerated its population at 10-year intervals, at roughly the same time each year (April). Thus, we are able to consider changes in the population between 1871 and 1881, 1881 and 1891, and 1891 to 1901.[2] In each case we compare the population at the beginning of the decade to the subset of people linked with time-invariant individual characteristics and to the set of links expanded through disambiguation.

As expected, disambiguation increases the number of unique links; the link rate changes roughly from 15 percent of records to nearly 30 percent (table 2). Any increase in sample size is welcome for a population as small as Canada, but the potential cost in terms of lost representativeness remains to be assessed.

Both methods underrepresent women relative to men in each of the three decadal intervals (table 3). On this point there is no difference between the basic and extended methods. Both methods overrepresent married people. Unexpectedly, this bias is smaller for the disambiguated data. Both methods also underrepresent Catholics,

[2] We are grateful to the Family Search Unit of the Church of Jesus Christ of Latter-Day Saints for sharing their census indexes with the Historical Data Research Unit at the University of Guelph.

Table 3. Characteristics of linked records before and after disambiguation

	Population at Start of Decade	Original Linked Records	Links after Disambiguation
Share of records describing women			
1871–81	0.49	0.45	0.46
1881–91	0.49	0.46	0.46
1891–1901	0.48	0.46	0.46
Share of records describing married people			
1871–81	0.31	0.39	0.34
1881–91	0.32	0.40	0.36
1891–1901	0.33	0.39	0.38
Share of records describing people who report as Catholic			
1871–81	0.43	0.34	0.36
1881–91	0.42	0.35	0.37
1891–1901	0.41	0.33	0.35

and again the disambiguated data are slightly closer to the full population, in each of the three decades.

These examples suggest that disambiguation with family coresidence is less damaging for representativeness than we might have expected. Indeed, the tendency to overrepresent married people is *reduced* though disambiguation. This effect is even more clear if men and women are examined separately (not shown). At first glance, then, the consequences for representativeness of increasing sample size through disambiguation with family coresidence are remarkably benign. Admittedly, we examine only visible characteristics. Even if disambiguation with family coresidence information does not change the composition of the population in terms of visible characteristics (gender, marital status, religion, and so on), the linked sample still might be different from the population at large in other characteristics that interact with time-varying information used in the linkage. For instance, those who lived with other family members could be less mobile, more risk averse, and less healthy. We have no ability to assess the impact of either linking strategy on characteristics not recorded by the census.

To investigate more closely we turn to a subset of the same data for which additional information is available. More complete information for each person is available in a randomly selected 5 percent of the 1871 and 1891 records. Here we examine only those links that fall within the rich 5 percent samples. Subject to this limitation we are able to consider a broader range of characteristics associated with the propensity to find a unique link. We begin with average link rates by select characteristics, and then report a multivariate logistic regression that identifies the contribution of several characteristics simultaneously to the odds of establishing a link.

In figure 1 substantial variation in the link rate by age is apparent for 1871–81. The pattern for 1881–91 is similar. The most conspicuous effect with the original

https://doi.org/10.1017/ssh.2020.15 Published online by Cambridge University Press

https://doi.org/10.1017/ssh.2020.15 Published online by Cambridge University Press

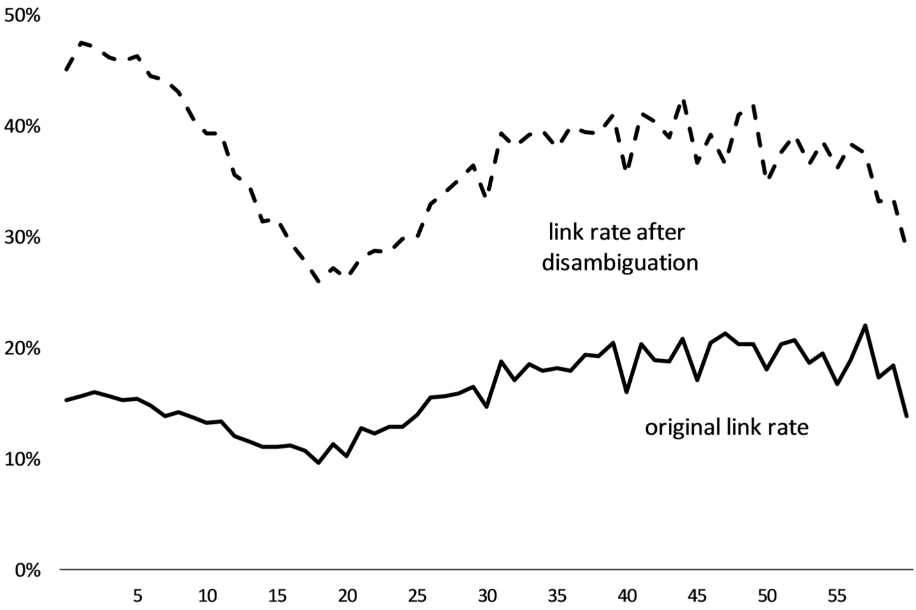

Figure 1. Rate of linking 1871–81, by age (yrs) in 1871.

linked data is an underrepresentation of children and young adults. The adolescent and young adult propensity to reinvent themselves as they leave home reduces the ability of both methods to identify them in the following census. Name changing at marriage for women is a particularly important influence although following young men, who do not change names at marriage, from one census to another is also a challenge. The lower rates of linking for people at ages 30, 40, 45, and 50 years reflect age heaping, which aggravates the problem of multiple links.

The effect of disambiguation varies considerably by age. Unsurprisingly, it is most efficacious at stages of the life course in which pairs or groups of people are more likely to remain together over the decade. Disambiguation has its biggest impact on link rates for young children, many of whom are still with their parents 10 years later. After disambiguation, the youngest children have the highest rate of linking in the population. The rate for adolescents improves but to a lesser extent, as expected. Disambiguation does not appear to reduce the effect of age heaping on the linking of people at select ages.

Additional detail in table 4 shows that with time invariant criteria we link 19 percent in New Brunswick during the 1870s against only 13 percent of the Quebec records. In the following decade, link rates range from 10 percent in western Canada to 25 percent in the small easternmost provinces. In addition to variation by gender, marital status, and religion, link rates differ by ethnicity and literacy. French-Canadians and those who did not read and write are harder to link. Again, we see that linking with time-invariant individual characteristics does not select from the population in a perfectly representative manner. Some of the biases are substantial.

Table 4. Average link rates, by characteristic

	1871–81		1891–81	
	Original	Disambiguated	Original	Disambiguated
Canada	0.15	0.38	0.18	0.35
BC			0.11	0.14
Manitoba			0.14	0.24
New Brunswick	0.19	0.37	0.26	0.39
Nova Scotia	0.18	0.42	0.24	0.42
Ontario	0.15	0.41	0.18	0.35
PEI			0.26	0.43
Quebec	0.13	0.35	0.17	0.33
Territories			0.10	0.18
Gender				
Female	0.14	0.36	0.17	0.33
Male	0.17	0.41	0.19	0.37
Marital status				
Single	0.13	0.38	0.17	0.38
married/divorced	0.19	0.39	0.20	0.32
Ethnicity				
French-Canadian	0.12	0.34	0.16	0.32
Other	0.16	0.40	0.19	0.36
Religion				
Catholic	0.12	0.34	0.16	0.32
Other	0.18	0.41	0.20	0.37
Literacy				
can read	0.18	0.38	0.19	0.36
not read	0.14	0.40	0.16	0.29
Birthplace				
born Canada	0.15	0.39	0.19	0.36
born elsewhere	0.17	0.34	0.18	0.27

Disambiguation broadly reproduces these biases for gender, ethnicity, religion, and birthplace. Bias by age increases with disambiguation especially for women. In contrast, *a priori* considerations would not predict the apparent changes in representativeness for province, marital status, and literacy. Overall, there is no obvious generalization about the impact of disambiguation on selection bias. Both strategies

https://doi.org/10.1017/ssh.2020.15 Published online by Cambridge University Press

Table 5. Odds ratios for establishing a link, 1870s, by region

	Original		Disambiguated	
	Odds Ratio	z-stat	Odds Ratio	z-stat
Ontario				
Male	1.20	9.18	1.26	16.04
Single	0.60	−15.03	0.60	−19.97
Old	0.69	−9.31	0.53	−20.17
Young	0.91	−2.85	1.22	7.45
French**	0.80	−4.29	0.35	−25.65
high status	0.84	−1.99	0.68	−5.75
born Canada**	0.96	−1.41	1.36	16.02
Quebec				
Male	1.35	12.72	1.28	14.71
Single	0.74	−7.47	0.72	−11.43
Old*	0.75	−5.90	0.54	−16.27
Young*	0.87	−3.46	1.21	6.73
French**	0.64	−15.04	1.01	0.65
high status	0.92	−0.92	0.73	−4.44
born Canada**	1.75	10.74	1.55	11.42
Atlantic				
Male	1.29	8.86	1.30	11.54
Single	0.64	−9.68	0.67	−10.53
old*	0.82	−3.49	0.59	−11.23
Young	0.99	−0.28	1.14	3.34
French	0.61	−9.70	0.51	−17.50
high status*	0.91	−0.85	0.69	−3.78
born Canada**	2.09	13.03	1.69	12.97

*Odds ratio moves .2 or more away from 1.0: increase in bias.
**Odds ratio moves .2 or more towards 1.0: decrease in bias.

for linking, in both periods, deviate from representativeness in complicated ways, and the marginal impact of disambiguation is complex.

Some of these effects may be interconnected. It is worth examining if the patterns of bias survive a multivariate analysis that considers multiple effects simultaneously. We report in table 5 and table 6 the association of different characteristics with the odds of finding a unique link. We estimate multinomial logit regressions with membership in the “individual only” and “disambiguated” samples against the baseline of the full population. We estimate separately by decade and by region (because of the

Table 6. Odds ratios for establishing a link, 1880s, by region

	Original		Disambiguated	
	Odds Ratio	z-stat	Odds Ratio	z-stat
Ontario				
male	1.10	5.23	1.20	12.39
single*	0.94	−2.58	1.38	16.53
old*	1.79	20.74	2.01	27.95
young	0.83	−7.49	0.88	−6.23
illiterate	0.66	−9.61	0.62	−13.77
French*	1.03	0.59	0.51	−15.03
high status	1.04	0.65	0.90	−2.25
born Canada*	1.03	1.39	1.53	22.96
Quebec				
male	1.34	12.31	1.32	14.93
single*	1.11	3.27	1.66	20.29
old	1.74	15.88	1.62	15.71
young	0.74	−9.28	0.85	−6.25
illiterate	0.81	−7.10	0.79	−10.20
French**	0.68	−13.1	1.01	0.21
high status	0.87	−1.92	0.76	−4.72
born Canada	1.96	13.05	2.02	15.98
Atlantic				
male	1.27	9.18	1.30	11.29
single*	0.97	−0.78	1.42	11.06
old	1.44	9.39	1.51	11.45
young	0.89	−3.33	0.87	−4.41
illiterate	0.82	−5.20	0.78	−7.15
French	0.65	−8.75	0.56	−13.81
high status	1.06	0.75	0.96	−0.49
born Canada	1.83	11.41	1.82	13.58
Manitoba				
male	1.07	1.12	1.05	1.00
single*	1.03	0.38	1.25	3.59
old	2.66	9.40	2.60	9.77
young	0.88	−1.71	0.91	−1.60

(*Continued*)

Table 6. (*Continued*)

	Original		Disambiguated	
	Odds Ratio	z-stat	Odds Ratio	z-stat
illiterate	0.79	−1.73	0.62	−3.92
French	1.01	0.11	0.85	−1.77
high status	0.89	−0.59	0.87	−0.92
born Canada*	2.04	10.78	2.83	18.75
West				
male	0.90	−1.34	0.86	−2.42
single	0.92	−1.08	0.94	−0.96
old*	2.50	7.49	3.05	10.25
young	1.27	3.21	1.30	4.12
illiterate	0.18	−12.5	0.12	−16.55
French	0.29	−3.67	0.21	−4.99
high status	1.07	0.40	1.10	0.64
born Canada	2.28	12.04	2.77	17.12

*Odds ratio moves .2 or more away from 1.0: increase in bias.
**Odds ratio moves .2 or more towards 1.0: decrease in bias.

differences identified in the preceding text). A coefficient of 1.0 indicates no effect on the odds of being linked. A coefficient smaller/larger than 1.0 indicates a characteristic that reduces/increases the probability of a record being linked.

The results provide additional detail about the pattern of demographic selections noted previously. Using time-invariant individual information being male and Canadian born tends to increase the likelihood of being linked. The young, singles, French, and illiterate are less likely to be linked. The elderly are also unlikely to be linked presumably because many do not survive into the next census enumeration. Having a high-status occupation reduces slightly the odds of being linked although the effect generally is not significant. There is some variation in these effects by province, especially during the second decade.

Again, by comparing the first and second columns, we are able to consider whether disambiguation exacerbates or reduces the pattern of selection biases evident in the original linking with time-invariant individual characteristics. No simple test statistic permits a straightforward test of the hypothesis that disambiguation increases or diminishes bias. Accordingly, in table 5 and table 6 we identify with a single asterisk the coefficients that move 0.2 or more away from 1.0—an increase in bias due to disambiguation. Identification of a 0.2 threshold difference, or a roughly 20 percent change in the odds ratio, does not derive from formal statistical reasoning; rather it is a heuristic measure of differences that seem large enough to matter (in the spirit of Ziliak and McCloskey 2004). Coefficients that converge toward 1.0 by a similar magnitude signifying that bias diminished as a result of disambiguation are reported with two asterisks.

https://doi.org/10.1017/ssh.2020.15 Published online by Cambridge University Press

By this metric, a majority of the coefficients do not change as a result of disambiguation. Only 9 of the 21 rows in table 5 and 10 of the 40 rows in table 6 see a change in the estimated coefficient larger than 0.2. Of those that do change, disambiguation makes it even more likely to link the Canadian born, that is, there is an increase in the overrepresentation of those born locally. In contrast, the underrepresentation of singles is reversed for the 1880s (not for the 1870s). Underrepresentation of those reporting French ethnicity diminishes in Quebec (where they are a majority of the population) and increases in Ontario (where the French are a minority). Overall, some biases are magnified while others are diminished as a result of disambiguation. The ability to link younger children and the middle aged benefits the most.

In summary, disambiguation does not change the extent of bias for a majority of the comparisons. Where we do see the changes, the effects are rather diverse, and the patterns differ by province. There is no obvious basis for a generalization along the lines of increasing or diminishing the problem of selection bias in linked data as a result of disambiguation.

Observations

Several observations emerge from this brief review of representativeness after linking Canadian census records in a conventional way and then disambiguating with coresident family members. Variations on the particular technique used for linking would produce slightly different results, but the patterns are unlikely to differ qualitatively.

1. Even the most parsimonious linking may inadvertently generate a selection bias that would prejudice the testing of some hypotheses. This is because people who can be followed from one census to another are somewhat atypical even if the census is a perfect representation of the population and the criteria for linking are unbiased.
2. The patterns of bias are complicated and not easily predicted. A number of factors can be seen to make it easier or harder to establish unique links from one census to another. Some of this selectivity originates with characteristics and imperfections in the census. Such influences would include, at a minimum, the size of population sharing a characteristic (e.g., birthplace), age at leaving home, and any nonrandom imprecision with which information is reported.
3. Disambiguation of multiple links increases sample size markedly. While the additional observations are useful, there is a cost in terms of added bias. Fortunately, the marginal increase in selectivity is less severe than anticipated. In some important respects, selectivity is diminished. Disambiguation also helps to reduce the rate of false positive errors (not reported here) without a marked aggravation of selection bias, especially for adults.
4. Reweighting is a useful strategy to mitigate the effect of nonrepresentative linking (Bailey et al. 2020a). Disambiguation is helpful in this regard because it expands the number of observations in each cell and thereby enables more precise parameter estimates.

An example illustrates the final point. As mentioned already, we have linked all records in the 1871 Canadian enumeration to all records in 1881, all 1881 records to 1891, and all 1891 to 1901. The use of time invariant characteristics yields more than half a million linked records in each interval, and more than a million records after disambiguation (table 2). Of course, they are not the same people in each decade. The set of people who can be found in each of the four enumerations allowing them to be followed over a full life course is much smaller.

We use these fully linked records in a separate paper to study social mobility, by comparing the occupations of fathers of boys who in 1871 had not yet entered the labor market against the sons' occupations as adults in 1901 (Antonie et al. 2020). We ignore women because their occupational reporting was inconsistent in this period. In 1901 the sons were roughly the same age, on average, as their fathers had been in 1871. Thus, we compare father and son at similar points in the life cycle. These restrictions are desirable for a study of social mobility, but the sample is reduced to 12,315 records using links established with time invariant characteristics and 22,357 records if we also rely on disambiguation.

Both sets of records are small. The descriptive review confirms that neither set of records is fully representative and that both require reweighting in any analysis. Happily, our focus on a precisely defined demographic group removes some sources of nonrepresentativeness. Given the biases identified in the preceding text, we reweight by province, French ethnicity, and whether or not the individual has left his province or country of birth by 1901.[3] The question we now ask is if the size of cells suffices for a credible reweighting.

In table 7 we report the distribution of cell sizes for boys who can be followed over 30 years, using the two methods, with cells defined by province (in 1871), French ethnicity, and mover versus stayer. Some cells have a large number of records, for example boys in each province who report the dominant ethnicity of the province and do not move. Other cells have as few as two records. The median cell size is 149 records with time-invariant linking rising to 246 records after disambiguation. The number of cells with fewer than 100 records falls from six to three through disambiguation.

The usefulness of reweighting is limited by the standard errors of the original size of cell. There is no obvious generalization about the size needed to obtain parameter estimates sufficiently precise for hypothesis testing. An appropriate threshold size will depend on the variability of the underlying data and the research question being examined. The burden of small sample size obviously weighs more heavily for research that might focus on the smaller provinces of New Brunswick and Nova Scotia. Even for Ontario, however, if we wish to compare the mobility of French-origin men with some other group, we will be driven to reweight cells with a relatively small number of observations. The larger, disambiguated sample diminishes although it does not eliminate this problem.

In our example, even though we begin the analysis using the entire Canadian population and we limit the reweighting to a small number of categories, the

[3] Religion is ignored here because it correlates strongly with ethnicity. Literacy is ignored because all the 1871 boys acquired the ability to read and write. We ignore age because almost all the boys will have left the parental home at some point during the 30 years.

https://doi.org/10.1017/ssh.2020.15 Published online by Cambridge University Press

Table 7. Number of records for each subgroup to be reweighted, by link strategy

	Original	Disambiguated
Nova Scotia, mover, French	2	2
New Brunswick, mover, French	7	10
Ontario, mover, French	18	30
New Brunswick, mover, not French	76	117
Nova Scotia, mover, not French	78	122
Nova Scotia, stayer, French	89	137
Quebec, mover, French	109	178
New Brunswick, stayer, French	151	234
Quebec, mover, not French	146	257
Ontario, stayer, French	216	297
Quebec, stayer, not French	772	1,282
Ontario, mover, not French	1,011	1,602
New Brunswick, stayer, not French	1,113	1,751
Nova Scotia, stayer, not French	1,570	2,572
Quebec, stayer, French	2,147	4,209
Ontario, stayer, not French	4,810	9,557
Average	770	1,397
Median	149	246

number of linked records is small enough on either linking strategy to reduce the precision of parameter estimates for a number of cells. Many other applications will have fewer records and/or more complicated reweighting than our example. For all of them, as for us, the problem of undersized cells is more severe if we are limited to records linked with time-invariant criteria. The obvious conclusion is that, if we are going to have to reweight anyway to mitigate the effects of selecting linking, then a larger sample achieved through disambiguation is preferable.

Concluding Comments

In this article we have examined Canadian census records. Would the same conclusions emerge from a similar treatment of US and British census records? Of interest would be any differences in enumeration practice that increased or diminished the incidence of people reporting the same name, age, and birthplace. In fact, census practice in the three countries was broadly similar (Dillon 1997, 2000). Broadly similar demographic detail was requested in each country. The enumeration principles and operating procedures of the Canadian census were heavily influenced by British and American practice.

https://doi.org/10.1017/ssh.2020.15 Published online by Cambridge University Press

To the extent that national censuses differed in ways that might affect linkage outcomes, the differences are likely to be small. Detail in the British census may have been more precise because the enumeration was *de facto* rather than *de jure*.[4] The *de jure* principle of enumeration used in North America permitted the recording of information for people absent at the time of enumeration and possibly having departed several months earlier. Indeed, information about some individuals was reported by people who were not family members and not closely familiar with them. Thus, the North American enumerations may harbor a greater incidence of imprecision and error, which in turn could lead to an increased rate of multiple records.

The granularity with which information was reported also influences the incidence of duplication using name, age, and birthplace. In Britain, despite some variation in the reporting of birthplace, there was a strong tendency to identify local communities and individual parishes, which for the most part had smaller populations than the states and provinces used for reporting in the North American censuses. The North American enumerations also include a higher proportion of foreign born, for whom birthplace typically was reported as an entire country. For all these reasons, the North American census is likely to have been less precise, and strategies for disambiguation more important, than for British data.

The small Canadian population makes disambiguation more important than it is for national-level research in a large country such as the United States. A wide range of Canadian analyses will be possible with disambiguated but not with standard linked data, simply because of the constraint of sample size. Even research about the United States, if it targets states, regions, or subpopulations defined in some other way, would find it useful to increase the size of linked samples through some kind of disambiguation.

A strategy for reweighting ameliorates the concern for bias on characteristics visible in the census, but it does not help with any bias in characteristics *not* recorded in the census. For example, disambiguation inevitably is less helpful for people at an age for which new households are being formed. Disambiguation raises sample size by a smaller proportion for this group. More importantly, those who can be disambiguated with continuous coresidence will differ from other young people in ways that are invisible to the researcher, to the extent that early marriage is selective on characteristics not reported in the census.

Disambiguation may be less useful for research questions focusing on young people than is it for analysis targeting incredibly young children or mature adults, whose sample mass can be expanded to a greater extent and with limited additional selection bias. More generally, we cannot expect to investigate family composition after disambiguating with family coresidence, just as social mobility cannot be examined if occupation is used to link or to disambiguate the data, and spatial mobility cannot be examined credibly if remaining in one place is a criterion for establishing a match. Recognition of these biases will limit some kinds of longitudinal research, although it may also suggest new research possibilities.

[4] Information in the population registration systems of several European countries is likely to have been even more precisely reported than either Britain or North America.

Clearly, there is no preferred method of constructing longitudinal data for all research questions. Rather individual investigators will benefit from a choice of link strategy that relies on criteria most appropriate for their own research projects. If care is taken to match link criteria with the hypotheses being examined, disambiguation will deliver larger samples and enable research that otherwise would not be possible. Fortunately, the advance of computational power makes custom linking by individual researchers a realistic possibility.

Acknowledgments. We gratefully acknowledge financial support of this research from the Canadian Foundation for Innovation, the Natural Sciences and Engineering Research Council of Canada and the Social Sciences and Humanities Research Council of Canada. The paper has benefited considerably from comments by participants at the 2017 European Historical Economics Society conference; specialized workshops at Northwestern University (2019), the University of Guelph (2018), and Cambridge University (2016); and the anonymous referees of this journal.

References

Antonie, Luiza, Kris Inwood, and J. Andrew Ross (2015) "Dancing with dirty data: Problems in the extraction of life-course evidence from historical censuses," in Gerrit Bloothooft, Peter Christen, Kees Mandemakers, and Marijn Schraagen (eds.) Population Reconstruction. Cham, Switzerland: Springer International Publishing AG: 217–41.

Antonie, Luiza, Kris Inwood, Dan Lizotte, and J. Andrew Ross (2014) "Tracking people over time in 19th century Canada." Machine Learning **96** (S1): 129–46.

Antonie, Luiza, Kris Inwood, Chris Minns, and Fraser Summerfield (2020) "When did the American dream move to Canada? Intergenerational mobility and the geography of opportunity, 1871–1901." Presentation to the Nuffield Historical Mobility Conference, Oxford, January 31.

Bailey, Martha, Connor Cole, and Catherine Massey (2020a) "Simple strategies for improving inference with linked data: A case study of the 1850–1930 IPUMS linked representative historical samples." Historical Methods, doi: 10.1080/01615440.2019.1630343.

Bailey, Martha, Connor Cole, Morgan Henderson, and Catherine Massey (2020b) "How well do automated methods linking perform? Evidence from the LIFE-M Project." Journal of Economic Literature: forthcoming.

Bloothooft, Gerrit, Peter Christen, Kees Mandemakers, and Marijn Schraagen, eds. (2015) Population Reconstruction. Cham, Switzerland: Springer International Publishing AG.

Christen, Peter (2012) Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Cham, Switzerland: Springer International Publishing AG.

Dillon, Lisa (1997) "Integrating nineteenth-century Canadian and American census data sets." Computers and the Humanities (30): 381–92.

—— (2000) "Integrating Canadian and U.S. historical census microdata: Canada (1871 and 1901) and the United States (1870 and 1900)." Historical Methods **33** (1): 85–194.

Feigenbaum, James (2016) "Automated census record linking: A machine learning approach." Working paper. http://scholar.harvard.edu/files/jfeigenbaum/files/feigenbaum-censuslink.pdf (accessed November 1, 2018).

—— (2018) "Multiple measures of historical intergenerational mobility: Iowa 1915 to 1940." Economic Journal (128): F446–F481.

Fellegi, Ivan P., and A. B. Sunter (1969) "A theory for record linkage." Journal of the American Statistical Association (64): 1183–1210.

Ferrie, Joseph P. (1996) "A new sample of males linked from the Public Use Micro Sample of the 1850 U.S. federal census of population to the 1860 U.S. federal census manuscript schedules." Historical Methods (29): 141–56.

Fourie, Johan (2016) "The data revolution in African history." Journal of Interdisciplinary History (47): 192–212.

https://doi.org/10.1017/ssh.2020.15 Published online by Cambridge University Press

Fu, Zhichun, Peter Christen, and Max Boot (2011) "Automatic cleaning and linking of historical census data using household information." ICDM Workshops: 413–20.

Fu, Zhichun, Mac Boot, Peter Christen, and Jun Zhou (2014) "Automatic record linkage of individuals and households in historical census data." International Journal of Humanities and Arts Computing **8** (2): 204–25.

Goeken, Ron, Lap Huynh, Thomas Lenius, and Rebecca Vick (2011) "New methods of census record linking." Historical Methods (44): 7–14.

Gutmann, Myron, Emily Klancher Merchant, and Evan Roberts (2018) "'Big data' in economic history." Journal of Economic History (78): 268–99.

Hacker, J. David (2013) "New estimates of census coverage in the United States, 1850–1930." Social Science History **37** (1): 71–101.

Maxwell-Stewart, Hamish (2016) "Big data and Australian history." Australian Historical Studies (47): 359–64.

Richards, Laura (2013) "Disambiguating multiple links." MSc thesis, University of Guelph.

Ruggles, Steven (2006) "Linking historical censuses: A new approach." History and Computing **14** (1–2): 213–24.

Ruggles, Steven, Cathy Fitch, and Evan Roberts (2018) "Historical record linkage." Annual Review of Sociology (44): 19–37.

Thorvaldsen, Gunnar (2017) Censuses and Census Takers: A Global History. London: Routledge.

Winkler, William E. (2006) "Overview of record linkage and current research directions." United States Census Bureau Research Report Series: Statistics #2006-2. Washington, DC: US Census Bureau.

Ziliak, Stephen T., and Deirdre N. McCloskey (2004) "Size matters." Journal of Socioeconomics (33): 527–46.

Cite this article: Antonie, Luiza, Kris Inwood, Chris Minns, and Fraser Summerfield (2020) "Selection Bias Encountered in the Systematic Linking of Historical Census Records," *Social Science History* **44**:555–570. doi:10.1017/ssh.2020.15