Socio-Economic Status and Names: Relationships in 1880 Male Census Data

Rebecca Vick, University of Minnesota

Record linkage is the process of connecting records for the same individual from two or more data sources. Linked files are uniquely rich in information about individual life change such as migration, occupational mobility and household composition. Historical linked datasets could potentially contain information that will solidify, enlighten or expand our knowledge of social science and demographic history.

To produce good quality linked datasets to be used for research one must consider how the predictor variables will affect the linking. In his 2006 paper Ruggles laid out the reasoning for limiting predictor variables for historical linking in order to avoid bias in which records are linked[1]. For academically-geared linked datasets, the linkage rate is important but representativeness is of the utmost concern so that the dataset will yield reliable research results. For example, using county of residence to help link would be very helpful. If you find someone with the same name and adjusted age living in the same county in time one and time two, you can be more confident that the link is correct compared to matching on name and age alone. However, using county of residence would lead to bias towards non-movers. In order to avoid biases, the variables used to link are often limited to variables that generally do not change over time, the most powerfully predictive being name (the main exception being women whose marital status changes from single to married over time), others being age, birthplace, sex and race.

Identifying individuals in 19th century records would be impossible without names. 19th century data rarely contains any form of identification number like we rely upon today. Not all names are created equal when it comes to predictability power. One would intuitively be more confident they found the correct match if they found a rare name like Rufus Pinkerton in two different datasets, but much less confident in matching a very common name like John Smith. Depending on the other supporting variables, common names will usually link to more than one record, which leads to ambiguous results. A powerful way to avoid harmful false positives is to simply not make a link when there is ambiguity as to which link is the correct[2]. Although avoiding false links is of primary concern, and throwing away ambiguous links the best way of minimizing false links, the final dataset will tend not to contain individuals who have common names[2]. Names are a personal identification method, something so basic that perhaps their relationship with other demographic information has been assumed to be random or benign. But is that assumption true? With record linkage there is an important reason to look deeper at this question. If there are relationships, record linkage methods that tend to exclude individuals with more common names to avoid false positives could lead to added bias in linked samples. It behooves record linkers who use name data to know whether or not this is possible. This research asks if there is a relationship between name commonness and socio-economic status. I use 1880 U.S. Census data and

the Duncan socio-economic index measure to examine this question. Here, I present some preliminary results of the study.

Data

For this inquiry I am using the IPUMS 10% sample of 1880 U.S. Census data. The IPUMS or Integrated Public Use Microdata Series is a harmonized set of census and other demographic datasets for social and economic research[3]. Using name data, I am able to compute commonness measures of first and last name combinations. The commonness measures are then attached to each individual's record. The IPUMS provides a variable called SEI, which contains a Duncan Socio-economic Index score based on occupation. The IPUMS occupation variable, OCC1950, contains the occupation codes used to assign a score to the SEI variable.

The OCC1950 coding scheme is well-established and accepted method to apply to data going back to the 19th century[4]. Although an occupational category scheme specific to 1880 would be ideal for this research, OCC1950 is suitable. Much important 19th century social science research has relied on these categorizations[4]. IPUMS coded 1880 occupational strings directly into OCC1950 coding scheme, therefore second-hand distortion that can occur from recoding from one scheme to another is not an issue. The Duncan Socio-economic Index score or SEI is an occupational standing measure. It is a composite measure that is based upon three measurable dimensions of status: income, education and prestige.[5] For more on how the SEI is constructed refer to Duncan's 1961 paper "A Socioeconomic Index for All Occupations".[6] Using a numeric measure of occupational prestige will make our evaluation simpler. There is a great amount of evidence that the socioeconomic status of occupations has been largely stable over the past two centuries, therefore, SEI based on 1950 occupational prestige, income and education, is a reliable score for 19th century data.[7]

The SEI has a maximum score of 96. It is calculated for all those with an occupational response, or OCC1950 code from 000-970. I am focusing on men only because women in the 19th century typically were not recorded as having an occupation outside the home, therefore using SEI would not be appropriate. I also focus on a subset of males who are of prime working age so as to avoid age affects. The age group I chose to look at men aged 30-50. This will avoid including young people who tend to have lower socioeconomic status, and those that are no longer in the work force because they are retired.

Methods

Males aged 30-50 who had an occupation (i.e. SEI>0) were selected from the 1880 10% sample dataset. The name data was then cleaned of non-alphabetic characters, titles and other non-pertinent characters, then parsed into first, middle and last name fields. A dictionary of standardized names was then applied to the first name data to correct for abbreviations and nicknames. For example the abbreviation “wm” was changed to William. Applying standardizations is common record linkage practice. It gives a better probability of making a name record for the same individual appear the same over time in different data sources. Any records missing a first *or* last name string were then removed. Finally, the cleaned and standardized first and last names were concatenated into full names. Names

containing initials were included. This was the final study group. The total number or records in the group is 658,541.

Analysis and Results

The clean and standardized names were tallied for how often each occurred in the data. The most common names are listed in Table 1 below.

Table 1. Twenty-five Most Common First-Last Name Combinations, Males Age 30-50, 1880 10% U.S. Census Sample

rank	first and last names	frequency	percent	cumulative frequency	Cumulative percent
1	john smith	838	.0012369	838	.0012369
2	william smith	747	.0011026	1585	.0023396
3	john brown	497	.0007336	2082	.0030732
4	william johnson	483	.0007129	2565	.0037861
5	james smith	482	.0007115	3047	.0044976
6	john williams	477	.0007041	3524	.0052016
7	john johnson	476	.0007026	4000	.0059043
8	john miller	449	.0006628	4449	.006567
9	george smith	441	.0006509	4890	.0072179
10	william jones	407	.0006008	5297	.0078187
11	john jones	396	.0005845	5693	.0084032
12	william brown	374	.000552	6067	.0089553
13	henry smith	354	.0005225	6421	.0094778
14	john davis	351	.0005181	6772	.0099959
15	charles smith	322	.0004753	7094	.0104712
16	james brown	307	.0004532	7401	.0109243
17	william davis	288	.0004251	7689	.0113494
18	john wilson	287	.0004236	7976	.0117731
19	james johnson	265	.0003912	8241	.0121642
20	george brown	263	.0003882	8504	.0125524
21	william miller	263	.0003882	8767	.0129406
22	william williams	263	.0003882	9030	.0133288
23	thomas smith	255	.0003764	9285	.0137052
24	james jones	251	.0003705	9536	.0140757
25	george washington	243	.0003587	9779	.0144344

To evaluate whether or not name commonness is associated with socio-economic status, the data was split into categories that ranged from most to least common. Because the frequency distribution of names is heavily skewed to the left with nearly 70% of all names occurring 4 or fewer times (51%

occurring only once) I first broke the frequencies into categories by looking for natural breaks then created breaks in smaller and smaller occurrence increments until reaching names that only occurred once.

Table 2. Name Commonness in 11 Categories: 1880 Males with Occupational Responses Aged 20-50

Category	Name Occurrences	Mean SEI	Frequency	Percent	Cumulative percent
1-Most Common	>=500	19.46887	1,542	0.23	0.23
2	440-499	18.85279	3,213	0.49	0.72
3	285-439	19.98373	3,012	0.46	1.18
4	200-284	18.59591	4,796	0.73	1.91
5	100-199	19.60625	16,894	2.57	4.47
6	50-99	20.44846	26,040	3.95	8.43
7	20-49	21.22089	45,724	6.94	15.37
8	10-19	21.29182	46,594	7.08	22.45
9	5-9	21.87144	57,248	8.69	31.14
10	2-4	22.03117	117,153	17.79	48.93
11-Least Common	1	21.74719	336,325	51.07	100.0
			658,541	100.0	

After considering the results of the 11 categories, I then collapsed them into four groups for easier analysis and interpretation. Groups one and two contain the most common names, and group three and four the least common. Each category's mean SEI is presented in Table 3 below.

Table 3. Name Commonness in four Categories: 1880 Males with Occupational Responses Aged 20-50

Category	Name Occurrences	Mean SEI	Frequency	Percent	Cumulative percent
1 – Most common	>=285	19.41367	7,767	1.18	1.18
2	50-284	19.96422	47,730	7.25	8.43
3	5-49	21.49199	149,566	22.71	31.14
4 – Least common	1-4	21.82056	453,478	68.86	100.0
			658,541	100.0	

Mean SEI grows from category one to category four indicating lower socio-economic status for those with common names. We can test for statistical significance in these SEI mean differences by creating a regression model. Table 4 contains results of a regression model that predicts SEI using the four name commonness categories. Categories one, two and three were represented as dummy variables and category four, which represents the least common names, was the reference category.

Table 4. Regression predicting SEI using name commonness categories

SEI	Coef.	Std. Err.

Comcat_1	-2.406885*	.224898
Comcat_2	-1.856343*	.0945714
Comcat_3	-.3285678*	.0586009
constant	21.82056*	.0291841

*statistically significant at the p=.05 level

All four name commonness categories are statistically significantly different SEI scores from that of category four, i.e. least common names. The coefficients show that those males with the most common names have SEI scores 2.4 points lower than males with the least common names.

The preliminary results of this study show that there is a statistical significantly difference in socio-economic status between those with common and uncommon first and last name combinations. Those with the common names tend to have lower status than those with less common names.

Although statistically significant, it is difficult to interpret the affects of this level of socio-economic differences between those with common and uncommon names might have on any particular linked dataset. The proportion of those with very common names is very small. The names deemed most common in this paper one comprise only a little over 1% of the overall study population. And the majority of people (69%) have uncommon names, names that are much less likely to share similarity with multiple records. However, the results do point to a relationship between socio-economic standing and name commonness, which could introduce unwanted bias into linked datasets.

Future Work

The SEI is a score that is applicable to datasets across time and does not change for different datasets. One score means the same thing in the 1850 IPUMS census data sample as it does in the 1950 IPUMS census data sample. Using SEI I plan to replicate the analysis done for 1880 10% data to the IPUMS 1850 and 1910 census records. Although I would like to do the same analysis for women's names, it is problematic in that women in the 19th century often did not have occupations outside the home, therefore using indexes that rely upon occupational data may not provide meaningful results.

I also plan to look at other IPUMS economic and socio-economic scores where available. They include the Siegel Prestige Score (PRESGL), the Nam-Powers_boyd Occupational Status Score (NPBOSS50), the Occupational Education Score (EDSCOR50), the Occupational Earnings Score (ERSCORE50) and the Occupational Income Score (OCCSCORE).

Finally, I plan to do further inquiry into the scale of the issue and if its potential effects on linked datasets.

References

1. Ruggles, S. 2006. Linking historical censuses: A new approach. History and Computing 14:213–24.

2. Goeken, Ron, Huynh, Lap , Lynch, T.A. & Vick, Rebecca. 2011. New Methods of Census Record Linking, Historical Methods: A Journal of Quantitative and Interdisciplinary History, 44:1, 7-14.

3. Ruggles, S., Trent Alexander, Katie Genadek, Ronald Goeken, Matthew B. Schroeder, and Matthew Sobek. 2010. Integrated Public Use Microdata Series: Version 5.0 [Machine-readable database]. Minneapolis: University of Minnesota.

4. IPUMS-USA Website. Chapter 4: Integrated Occupation and Industry Codes and Occupational Standing Variables in the IPUMS. Accessed on 9/27/2013. https://usa.ipums.org/usa/chapter4/chapter4.shtml .

5. IPUMS-Website. SEI variable description page. Accessed on 9/27/2013. https://usa.ipums.org/usa-action/variables/SEI#description_section

6. Duncan, O.D. 1961. "A Socioeconomic Index for All Occupations," in A. Reiss et al., Occupations and Social Status. Free Press.

7. Hauser, Robert M. and John Robert Warren. 1997. "Socioeconomic Indexes for Occupations: A Review, Update, and Critique." Sociological Methodology 27: 177-298.