Methodological Advances for Linking Historical Censuses
With an Application to Occupational Following

Alan Gelder

Department of Economics, The University of Iowa, Iowa City, IA 52242, USA

Abstract

Using comprehensive census data for Cornwall County, England, I create a panel dataset that spans six censuses (1841–1891)—possibly the largest panel dataset for Victorian England at present. I present the methodology for linking individuals and families across these censuses. This methodology incorporates recent advances in census linking (including the use of machine learning) and introduces new methods for tracking migration and changes in household composition. I achieve a forward matching rate of 43%. The additional inclusion of marriage and death records could allow for well over 60% of the population to be accounted for from one census to the next. Using this new panel, I investigate the frequency with which sons pursue the same occupations that they observed their fathers doing while growing up. For sons that did not follow in their father's footsteps, I identify some correlates that may have contributed to the change.

Keywords: Census Matching, Support Vector Machine, Data Linkage, Cornwall, Occupational Following
JEL: C81, N00, N33, N93

1. Introduction

Beginning in 1841 in the UK and 1850 in the US, census enumerators were charged with individually listing—not just tallying—every man, woman, and child in the country. In addition to recording names, the typical decennial census began to include a person's age, gender, occupation, birthplace, marital status, address, and relationship within his or her household. While census data present a wealth of demographic information for social scientists and genealogists alike, the amount of

*Tel: +1 319 621 7086; Fax: +1 319 335 1956
Email address: alan-gelder@uiowa.edu (Alan Gelder)

information conveyed by any census in isolation is limited since it only provides a cross-sectional picture. If, however, censuses were systematically linked so that each individual could be traced throughout their life at ten year increments, the result would be a dynamic and intricately detailed picture of a nation that spans generations.

Given this rich potential, it is surprising how little has been done to systematically link individual and family records across multiple censuses. Three major obstacles have limited large scale census matching projects: (1) the availability of census transcriptions; (2) imperfections in the original census data—a problem which is compounded by imperfections in census transcriptions; and (3) methodological challenges for linking the data given its underlying shortcomings.

Census transcriptions that are freely available to social scientists are primarily limited to 1% and 5% samples—the biggest exceptions being the complete transcriptions of the 1880 US and 1881 UK censuses.[1] In fact, the majority of census linking projects to date have entailed matching one of the 1% or 5% census samples with the corresponding 1880 or 1881 census.[2] The linking process is necessarily difficult since names, ages, and birthplaces frequently have irregular variations from one census to another.[3] Transcribing historic records also has its own set of problems as the quality of original documents may have been compromised by illegible handwriting, ink blots, or improper storage conditions. Transcribers may also be inexperienced, fatiqued, or unfamiliar with names and places.[4] Matching algorithms must be flexible enough to account for substantial variations in the census entries. However, as flexibility increases, so does the problem of duplicate matches and false positives (especially for people with common names). Life cycle changes are yet another con-

[1] These transcriptions were coordinated by The Church of Jesus Christ of Latter-day Saints. Complete, albeit proprietary transcriptions for the remaining UK and US censuses in the public domain (UK: 1841–1911; US: 1850–1940) are available through a variety of genealogical websites.

[2] A leader in census matching has been the North Atlantic Population Project of the Minnesota Population Center. It has linked samples from seven US censuses into the 1880 census and has also performed census linking with Norwegian censuses. Joseph Ferrie and Jason Long have also linked small samples into the 1880 US and 1881 UK censuses (see, for instance, Long and Ferrie, 2013, Appendix 2).

[3] Names may be abbreviated or nicknames may be used instead of proper names. Even though ages should systematically increase by ten years from one census to the next, they are often misreported—either accidentally or deliberately. Birthplaces may sometimes be reported as one parish in one census and as an adjoining parish in a subsequent census.

[4] For example, I have repeatedly seen the female biblical name Tamar mistakenly transcribed as James. High quality transcription projects often entail multiple levels of transcribing and checking as an attempt to minimize such errors.

Table 1: Census Observations by Year

1841	1851	1861	1871	1881	1891	Total
340,901	354,742	372,163	356,364	324,835	318,634	2,067,639

sideration: people die, people move, wives adopt their husband's surname. All of these issues should ideally be accounted for.

Here, I present my methodology and the results for linking six consecutive censuses (1841–1891) for Cornwall County, England. Comprehensive transcriptions for these censuses were generously provided by the Cornwall Online Population Project.[5] In total, these transcriptions contain more than 2 million census observations, with roughly one-third of a million observations per census year (see Table 1). Methodologically, I implement and build upon the census linking tools introduced in Fu, Christen, and Boot (2011b). A key element of Fu et al. is the utilization of household information in the matching algorithm—information which can greatly aid in selecting true matches from the many false positives that are frequently generated by pairwise comparisons alone.[6] While the algorithm in Fu et al. focuses on tracing the core members of a household through time, I adapt their algorithm to account for household members who move away. The magnitude of observations I am working with is more than a dozen times larger than in Fu et al. The Cornwall data also covers a considerably larger geographic area. Given this expanded coverage, tracking migration within Cornwall presents a challenge since comparison sets must be kept small enough to be computationally feasible. I investigate a nested matching algorithm which meets this challenge. For most censuses, I am able to successfully identify about 43% of the population in the subsequent census.

I demonstrate one potential use of the new linked dataset by investigating occupa-

[5]This group is affiliated with FreeCen, a non-profit organization whose stated mission is to provide high quality 19th Century census transcriptions free of charge. The transcription methodology and standards for FreeCen can be found at http://www.freecen.org.uk/.

[6]The Minnesota Population Center has deliberately omitted household information from its census linking projects since it is a "source[ ] of potential bias" (Goeken et al., 2011, p. 8). Here the exercise is different. I am not matching a 1% sample into a complete census with the goal of maintaining a representative sample of matches. My ultimate goal is to match (or at least account for) *all* individuals. In doing so, I take more of a jigsaw puzzle approach by putting the easy pieces together first (this paper) and then figuring out where the more difficult pieces fit in (future work). Additionally, by beginning the matching algorithm with pairwise comparisons at the person level, I avoid the pitfalls Goeken et al. allude to for matching on household traits when the composition of a household changes between censuses (p. 8–9).

tional patterns between fathers and sons. For this exercise, I look across censuses and identify a father's occupation when his son is a boy and then the son's occupation later in life as an adult. Since many of the occupations cannot be clearly ranked monetarily, and since several require unique skill sets, I analyze intergenerational patterns on a sector by sector basis. For most sectors, between one-quarter and one-half of sons pursue jobs in their father's line of work. Using multinomial logistic regression, I find that primogeniture inheritance is clearly apparent in farming but seemingly absent in other sectors.

2. Data Formatting

With the underlying goal of maintaining the maximal amount of identifying information, I follow some established procedures and implement others that are new in preparing the data for matching. There are eight variables from the census that I primarily utilize: first name, surname, birth year, birthplace, occupation, gender, address, and household. The gender variable requires little cleaning, and the birth year variable can easily be constructed from the ages reported in the census.[7] I largely follow Fu et al. (2011a) in my standardization of the name variables.[8] I also follow Fu et al. in using the occupational codes that were developed for the 1911 UK Census to numerically code occupational values.[9] The original coding system assigned three-digit codes to each occupation. However, it also grouped occupations into 23 industries, each with between one and eleven sub-industries. To account for these industry divisions and to increase the numeric difference between industries, I extended the three-digit codes into a new seven-digit code where the first two digits are the industry number, the next two digits are the sub-industry number, and the last three digits are comprised of the original three-digit code.

One of the greatest assets of the UK censuses in terms of personal identifying information (and oddly omitted from Fu et al.) is the specificity of birthplaces. In the

[7]The primary caution is to be mindful of the units in which ages are reported since children's ages are often reported in months, weeks, or even days. Census dates are also important in calculating a child's birth year: June 6 in 1841, and between March 30 and April 7 for 1851–1891.

[8]Names must first be divided into surnames and forenames (this distinction was already made in some transcriptions). I then further divided forenames into first names and middle names (due to the sparsity of middle names and the irregularity of their use from one census to another, I chose to ignore them in the matching process). After removing special characters and capitalizations, I standardized common abbreviations and name variants by assigning uniform values. For example, Chas, Chs, Charley, and Carles were all coded as Charles.

[9]Information on the occupational code system can be found on the official 1911 Census website: http://www.1911census.co.uk/content/default.aspx?127

US, "New York" or "Virginia" is a typical birthplace entry—broad enough to convey no more than trifling information for matching people within those states. The UK censuses, however, commonly specify a person's birthplace to within a radius of two or three kilometers (the one exception being the 1841 census which only identified a person as having been born either in or out of the county). That said, it is also not uncommon for one parish to be reported as a person's birthplace in one census and a neighboring parish to be reported in the next. Although geographically close, this information would be lost in the matching process without a reasonable metric for comparing place names. To account for such discrepancies, I used the UK Ordnance Survey's coordinate system to geocode birthplaces with a northing and an easting.[10] For birthplaces outside the UK, Ireland, and the Channel Islands, I used latitude and longitude degrees.

I use four nested variables to describe a person's residential location on the night of the census: civil parish, registration district, and two nested groupings of registration districts.[11] This nesting process will be useful in tracking migration. Civil parishes are fairly small, with well over 200 in Cornwall. Registrations districts are substantially larger with portions of 16 within the county, each comprised of as many as 25 civil parishes apiece.[12] The regional groupings of registration districts first divide Cornwall into six regions, and then into three.[13]

For referencing records within and across censuses, I developed a simple system which nests person, household, regional, and census level information. Individuals are numbered within a household, households are numbered within a census piece (a numbered geographic division of the national census enumeration), and census pieces are numbered within a census. For instance, reading from right to left,

[10]I used the "1:50 000 Scale Gazetteer", downloaded from http://www.ordnancesurvey.co.uk/ on 13 June 2013. This data, measured in meters and rounded to half a kilometer, provides northing and easting coordinates for thousands of places within the UK. I used the same coordinate system to identify major locations in Ireland and the Channel Islands.

[11]By census regulations, the residence of a visitor or a traveler was recorded as the place where they physically resided on the night of the census. Hence, the residence in the census may not be the person's permanent domicile.

[12]Small portions of Launceston and St. Germans extended beyond Cornwall into Devon County, while bits of Holsworthy and Tavistock crossed into Cornwall County. The remaining 12 registration districts were entirely within Cornwall. Transcription coverage along the Cornwall-Devon border varies from census to census.

[13]The small groupings consist of the following registration districts: 1. Stratton, Launceston, Camelford, and Holsworthy; 2. Liskeard, St. Germans, and Tavistock; 3. Bodmin and St. Austell; 4. St. Columb and Truro; 5. Falmouth and Redruth; and 6. Penzance, Helston, and the Scilly Islands. The larger groupings combine 1–2, 3–4, and 5–6.

record number 9-1856-1128-008 refers to the eighth person in the 1128th household of census piece 1856 in the 1891 census.[14] Harnessing such a degree of information within the identification numbers themselves has proved to be quite useful.

3. Methodology

Before covering details, I will give a brief synopsis of the census matching algorithm. First, pairwise comparisons are done to access the similarity of records across censuses. Second, a machine learning algorithm classifies each pairwise comparison as either a match or a non-match. Third, to try to separate false positive matches from true matches, household information is taken into account via a group linking process. So far this is according to the algorithm in Fu et al. (2011b). My contributions are in two additional steps. Fourth, I extend the group linking process to look for individuals who have moved away from their former household members. Fifth, I balance the goal of tracking migration within Cornwall with the computational cost of conducting vast numbers of pairwise comparisons by first identifying the people who did *not* move. These records are then removed, and steps one through four are iterated on the remaining records, looking for people who moved increasingly larger distances.

As I alluded to, it is computationally expensive to systematically compare each record in one census with each record in another. Two censuses with one-third of a million observations apiece would result in over 111 billion pairwise comparisons. In the data linking literature, the concept of blocking has been established to focus pairwise comparisons on sets of data where matches are most likely to occur. For instance, there is little sense in attempting to match a male in one census with a female in another (unless, perhaps, there was an error in the recording process).

I use three variables to define comparison blocks in the census data: gender, surname, and residential location. Thus, pairwise comparisons between censuses will be limited to observations that share the same values for each of these variables. To account for spelling variations and phonetic similarities, the surname variable is coded as a double metaphone. Residential location is based on the four nested

[14] Household distinctions are less clear in the 1841 census than they are in subsequent censuses. Household relationship data is also not included in the 1841 census, so I assigned a new household each time the surname changed (this method does, however, separate live-in servants, boarders, and relatives with different surnames into different households). In all censuses, I omit vacant buildings from the numbering process. The transcriptions used by Fu et al. apparently did not include household divisions since they developed a method for automatically detecting breaks between households based on relationship entries (2011a).

variables described earlier, with the county of Cornwall as a fifth.

Since the matching algorithm in Fu et al. (2011b) serves as the basis of the algorithm here, it is worth outlining. Censuses are compared two at a time. After being separated into blocks based on gender, surname, and residential location, observations within each block are then systematically compared across censuses.[15] Pairwise comparisons assess the similarity of records based on five attributes: first name, birth year, the northing and easting of the birthplace geocode, and the occupation code.[16] Each of these five attributes is assigned a score in the unit interval where the score is monotonically increasing in the degree of similarity—1 representing identical values and 0 reflecting no similarity.[17] At this point, a question arises as to how the five similarity scores should be interpreted. How should each score be weighted, and what is the threshold for classifying the record pair as a match? A viable solution is the use of a *support vector machine* (SVM), a technique whereby a computer is trained to identify the characteristics of a true match and to place weights accordingly.

An SVM identifies the separating hyperplane which places the most distance between data from two groups: true matches and false matches. To function, an SVM must first have a training sample of pairwise comparisons which have already been identified as either true or false matches. My training sample is comprised of 478 true matches which I identified by hand and 31,708 false matches. Pairwise comparisons had been done for these false matches because they were in the same block, yet they could not be true matches since I had already identified the unique true match.[18] Based on the characteristics of the elements in the training sample, the SVM is able to identify the desired hyperplane. This hyperplane can then be used to determine the matching status of additional pairwise comparisons.

[15] Fu et al. (2011a) only uses surnames to define blocks; Fu et al. (2011b) does not specify which variables, if any, are used at the blocking stage.

[16] Fu et al. (2011b) compares surname, first name, gender, age, occupation code, and address.

[17] The specific comparison methods are as follows: first names are compared by the Winkler string comparator (checking for the initial characters, long sets of characters that are similar, and overall similarity); birth years are compared by absolute numeric difference (a score of zero is assigned for differences over 10 years); northings, eastings, and occupational codes are compared by their percentage difference (differences above 15% are given scores of zero for northings and eastings, while 80% is used for occupational codes).

[18] For the true matches, I randomly sampled portions of the Penzance and Falmouth registration districts from both the 1861 and 1871 censuses and then looked for the matching record in the corresponding census (either 1861 or 1871). The false matches were derived from pairwise comparisons for these registration districts.

Formally, let $(\mathbf{x}_k, y_k)$ be a pairwise comparison where $\mathbf{x}_k$ is the vector of similarity scores and $y_k \in \{1, -1\}$ indicates the true or false match status. With the notation for a hyperplane of $\mathbf{w}^T\mathbf{x} + b = 0$, the SVM solves:

$$\min_{\mathbf{w},b,\xi} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_k \xi_k \qquad \text{s.t.} \;\; y_k[\mathbf{w}^T\phi(\mathbf{x}_k) + b] \geq 1 - \xi_k \;\; \text{and} \;\; \xi_k \geq 0$$

Nonlinear separations between groups are accounted for through the kernel $\phi(\cdot)$, and outliers are factored in through a penalty parameter $C > 0$ and a slack variable ξ (Cortes and Vapnik, 1995).[19] Since birthplace information is limited in the 1841 census, I train two separate SVM models: one that incorporates the birthplace geocodes and one that omits them.

After the SVM determines the matching status of each pairwise comparison, household information is used to address the problem of false positives. The underlying premise being that while an individual (e.g. Mary Smith) may have many potential matches, the probability of detecting the true match increases dramatically if her family members also have matches within the same household. The group linking process assigns a similarity score to each potential household pair. Formally, let $M_{i,j}$ be the set of pairwise comparisons between households i and j that the SVM classified as matches (with its cardinality denoted by $|M_{i,j}|$). Then for each $k \in M_{i,j}$, let S_k be the sum of the similarity scores $\mathbf{x}_k$.[20] Assuming household i is in the earlier of the two censuses and household j is in the latter, denote the number of people in household i as m_i and the number of people in household j who are at least 10 years-old as m_j.[21] The similarity score $\mathbb{S}_{i,j}$ between households i and j is computed as follows:

$$\mathbb{S}_{i,j} = \frac{\sum_{k \in M_{i,j}} S_k}{m_i + m_j - |M_{i,j}|}$$

This formula is increasing in both the number of SVM matches between households and the raw strength of their underlying similarity scores. However, it also takes the number of potential matches between households into account and places downward weight accordingly. The algorithm in Fu et al. (2011b) concludes by selecting the household pairs that have the greatest similarity.[22]

[19] I use the Gaussian Radial Basis Kernel with parameter $\gamma = 0.001$. I also set $C = 200$. My goal in selecting these parameter values was to reduce the degree of both false positives and false negatives.

[20] Fu et al. (2011b) find that this formulation outperforms the alternative where $S_k = 1$.

[21] The purpose of m_i and m_j is to provide weights for the number of potential matches. Therefore, unlike Fu et al. (2011b), I adjust m_j to only include individuals who had been born previous to the last census.

[22] Their notion of similarity, however, is one-sided. For each household in the first census, Fu et

A shortcoming of only selecting the household pair with the maximum similarity score is that it fails to account for evolving household compositions. Family members may move in or out. A household may also be comprised of multiple family units, such as live-in servants, lodgers, or extended family members. Even though core members of a household may be tracked by selecting household pairs with the maximum similarity score, peripheral members are liable to be missed. I introduce a simple multi-level group linking technique to account for such peripheral changes. The first level begins by identifying the household pairs with the highest similarity scores. All SVM matches that are between such household pairs are marked as having cleared this stage of analysis and are separated from the remaining SVM matches. In each successive level, this pattern is repeated: household pairs with the greatest similarity among the remaining SVM matches are flagged, and the SVM matches within those household pairs are separated so that the process can be repeated. This multi-level technique may be iterated any number of times. I find, however, that the number of additional matches decreases sharply with each new level; so I stop after three levels.

The final element of my matching algorithm is to search for matches over an expanding geographic region. Even after separating the data into comparison blocks by gender and surname, a full pairwise analysis for the entire county of Cornwall is—computationally—a fairly tall order. Furthermore, if such a task is difficult at the county level, consider the difficulty in doing a full pairwise comparison for the entire UK or US. Restricting comparison blocks to smaller geographic regions has clear computational benefits, but it also comes with the potential that erroneous matches may arise at the local level which prohibit true matches from being classified at the global level. The question then is whether or not the same records are consistently classified as matches under different geographic divisions. To investigate this issue, I examine three different matching types which vary in terms of their initial and intermediate geographic divisions (see Table 2).

Matching Type A, for instance, begins at the civil parish level. Therefore, all of the matching steps up to this point are conducted among individuals living in the same civil parish. Before repeating the matching steps among individuals within the same registration district, I first remove matches that have a high probability of being correct. Reducing the comparison set in this manner prevents it from becoming too unwieldy when comparisons are done over a larger geographic area. The matches I

al. select the household in the second census with the highest similarity. Yet the converse may not be true (i.e. a different pairing may arise if for each household in the second census, the household in the first census with the highest similarity is selected). To circumvent this problem, I require household pairs to have the maximum similarity score in both directions.

Table 2: Matching Stages by Residential Location

	Type		
Location	A	B	C
Civil Parish	1		
Registration District	2	1	1
RD Small Group	3	2	
RD Large Group	4	3	
Cornwall	5	4	2

specifically remove are those where at least two individuals have been matched in a household pair during the multi-level group linking step. For each of the three matching types, I repeat this process, gradually increasing the geographic coverage until the remaining set of records are compared across the entirety of Cornwall.

As a final note on methodology, I used Febrl to calculate the similarity scores (Christen, 2008), and the R package e1071 to conduct the SVM analysis (this package implements the libsvm program by Chang and Lin, 2011).

4. Results

Combining the multi-level group linking with the nested geographic matching, there are as many as 15 different matching stages. To highlight the marginal contribution of each, Table 3 shows the number of pairwise comparisons to be classified as a match at each stage for the 1871–81 census pair. I report two columns for each of the three matching types: (i) the number of initial matches as given by the algorithm; (ii) a refined set of matches, based on stricter criteria that will be explained shortly. By far, the greatest number of matches occur within the smallest geographic divisions. This should not come as a surprise since a large portion of the population remains settled over time; and if migration does occur, it is presumably easier for individuals to move than for families. The multi-level matching does pick up a fair number of individuals who leave their former household—such as when multiple families or extended family members are living together in one census but are in different households by the next. The multi-level matching also frequently identifies individuals who are the sole match between two households. However, I group these

Table 3: Number of Matches Identified at Each Stage: 1871–1881

		Type A		Type B		Type C	
	ML	Init.	Ref.	Init.	Ref.	Init.	Ref.
Civ Par	1	90,785	90,348				
	2	1,266	1,061				
	3	77	61				
Reg Dist	1	11,797	11,731	102,033	101,558	102,033	101,558
	2	228	190	2,088	1,661	2,088	1,661
	3	31	18	232	160	232	160
RD Sm	1	1,776	1,751	1,774	1,754		
	2	52	40	62	52		
	3	8	8	18	16		
RD Lg	1	2,818	2,784	2,812	2,785		
	2	74	47	76	47		
	3	26	18	28	18		
Cornwall	1	2,841	2,821	2,832	2,816	7,391	7,330
	2	117	103	119	103	263	218
	3	50	44	50	46	106	86
Solo Matches		65,487	39,274	63,684	39,537	57,081	37,736
Total		177,433	150,299	175,808	150,553	169,194	148,749

solo matches separately in Table 3, regardless of the stage in which they are found.[23]

Another feature of Table 3 is that the number of matches within a geographic region is largely unaffected by whether or not the matching is first conducted within a smaller geographic area. This is especially the case among non-solo matches. For instance, Type A has a total of 104,184 non-solo matches across both the civil parish and registration district levels—less than 200 off of the number generated when the matching starts at the registration district level (104,353 matches for Types B and C). Likewise, Type C jumps straight from the registration district level to comparing all of Cornwall, while Types A and B take two additional geographic steps.

[23]Since records in solo matches remain eligible for the next stage of matching, many of these solo matches are identified multiple times. Table 3 gives the number of unique pairwise solo matches.

Table 4: Overlap of Match-Pairs Across Matching Types: 1871–1881

Match Types	Initial	After Refinements
A, B, C	167,329	148,435
A, B	6,817	1,824
A, C	4	1
B, C	1,550	294
A	3,271	38
B	106	0
C	305	19
Total	179,382	150,611

Yet the number of non-solo matches identified by each type are practically identical (Type A: 7,762; Type B: 7,771; Type C: 7,760). Although the numbers of matches remains similar, there is still a question as to whether they represent the same pairwise matches. Table 4 examines the overlap of pairwise matches between the three different matching types. Interestingly, the vast majority of pairwise matches found by any of the matching types were found by all matching types. This holds true both before (93.3%) and after (98.6%) the refinements and suggests that geographic divisions may be chosen to suit computational needs. Looking at the matches that were not identified in all three cases, Types A and C had the most dissimilarities—each conforming well with Type B, but not with each other. Type A has the most unique contributions while Type B has the least. Many of the dissimilarities across matching types center on the solo matches.

The veracity of the solo matches is the most questionable. Yet blanketly removing them would severely compromise the construction of a panel dataset designed to follow individuals over their life cycle: the connection between childhood and adulthood would largely vanish as individuals leave home. I implement a five step refinement process in order to remove solo matches that have a low level of credibility, as well as to hold non-solo matches to a higher standard. The number of matches remaining after each stage of the refinement process is shown in Table 5. This refinement process also synthesizes the three different matching types. Thus, before beginning the refinement process, I combine all of the unique pairwise matches from the three different matching types into a single dataset, noting where each pairwise match was identified and the associated number of matches in its household pair.

Table 5: Unique Match-Pairs After Each Refinement

	1841–51	1851–61	1861–71	1871–81	1881–91
All Match Types	203,165	191,451	193,452	179,382	167,529
Refinement 1	189,812	185,878	188,861	175,839	165,108
2	167,254	175,317	179,289	168,261	158,438
3	115,498	155,408	159,738	151,797	145,285
4	108,904	153,204	157,995	150,688	144,595
5	108,624	153,062	157,870	150,611	144,448

The first refinement drops all solo matches that were identified by only one of the three matching types. While some of these may be legitimate matches that take advantage of the unique geographic structure of a particular matching types, I reject them for lack of a second witness (either in terms of additional household members or by being acknowledged by a second matching type). In the second refinement, I remove all solo matches where the individual is listed as a wife in the latter of the two censuses. Frequently, when a wife is matched, additional members of her household are also matched; so a solo match in this case is likely comparing a woman whose married name happens to coincide with another woman's maiden name.

The final three refinements address duplicates where a record in one census is matched to two or more records in another census. For the third refinement, duplicates are identified, and any duplicate that is a solo match is discarded. The fourth refinement is similar, but is done separately because of the large number of solo duplicates. It first entails re-identifying duplicates among the remaining matches. Then, any match belonging to a household pair comprised of either two or three matches is removed if there is a duplicate anywhere in the household pair. For example, suppose that three members of one household have been matched with three members of another. If two of the household members have also been matched elsewhere, the credibility of the entire household pair as a match comes into question. The fifth and final refinement stage addresses duplicates in household pairs that have at least four matches. By this point, duplicate matches in the remaining data are primarily limited to family members with similar names and attributes. The number of these duplicates is small enough that it would be feasible (although still many hours of work) to manually determine the correct matching status of each. For now, I simply drop the remaining duplicates, leaving the remainder of the household pair intact. I use the refined data from here on out.

Table 6: Number of Matches by the Size of Household Pairs

Matches/HH	1841–51	1851–61	1861–71	1871–81	1881–91
1	32,148	42,549	41,496	39,538	38,573
2	24,358	28,436	30,510	29,616	28,798
3	18,798	26,475	29,109	27,657	26,652
4	15,219	22,802	24,331	22,959	21,891
5	9,651	16,108	16,315	15,572	14,177
6	5,197	9,759	9,366	8,964	7,754
7	2,273	4,331	4,417	4,056	4,050
8	700	1,769	1,612	1,617	1,498
9	205	583	495	468	522
10+	75	250	219	164	533

Table 6 breaks down the refined data by the size of household pairs. For example, between the 1861 and 1871 censuses, 29,109 matches were part of household pairs that consisted of exactly three SVM matches between the households. Across the board, the number of matches steadily declines as the size of household pairs increases. Another feature of Table 6 is that the number of matches within each size remains remarkably consistent from one census pair to another between 1851 and 1891. The matching rates are likewise similar.

Forward matching rates, or the percent of the population in one census that can be identified in the next, are given for each census pair in Table 7. These rates are presented for all of Cornwall, as well as at the registration district level.[24] It is somewhat intriguing how little variation there is in the forward matching rates at both the aggregate and registration district levels from the 1851 census onward. At the aggregate, the matching rates are all within 2.2 percentage points of each other with an average of 43.1%. The average spread at the registration district level across census pairs is 3.7 percentage points. There is also no clear pattern between the forward matching rates and the populations of the registration districts.[25] The

[24]Tavistock, Holsworthy, and the 1861 shipping records are not shown individually, but are included in the totals.

[25]Pearson correlation coefficients between the forward matching rates and the average population of the registration districts range from -0.32 to 0.23 for census pairs between 1851 and 1891. The correlation is stronger for 1841–51 at 0.57. Most of the registration districts in Cornwall had fairly constant populations over this time period.

consistency of the matching rates over time suggests a homogeneity in the population structure throughout Victorian Cornwall.

The lower matching rate between the 1841 and 1851 censuses (32%) largely stems from the lack of precision in the 1841 census. As was previously noted, the 1841 census only recorded the place of birth as either in or out of the county. Given that the overwhelming majority were born in the county, I conducted the SVM matching for this census pair without birthplace information. Another known issue with the 1841 census is that ages above 15 were often rounded in five-year increments. In future work, it may be beneficial to specially train an SVM model that is based on a sample of 1841 and 1851 census records (as opposed to the 1861 and 1871 sample used here). Many of the matches that have been identified could be used in such a sample. For that matter, it may also be useful to train an SVM model for each census pair, taking advantage of matches which have been identified here. Larger and more specific training samples may allow the algorithm to better identify the defining traits of matches.

There are additional ways in which the matching rate can be considerably improved. Marriage records provide the vital link for tracing women as their surnames change. Based on the volume of marriage records, approximately 7% of the 1871 Cornwall population changed their surname before the 1881 census.[26] Not only does the current algorithm miss these brides, but the bias toward selecting matches with multiple members in the household is limiting the number of grooms that are matched as they move away from home. With marriage records in hand, an additional training sample could be developed for tracking brides and grooms at the point of matrimony.

Death records are also pertinent to improving the matching rate. Although these individuals will not be found in the next census, they can at least be accounted for—removing the ambiguity as to whether the person died, emigrated, or cannot be found for some other reason. For monitoring attrition, the value of death records is appreciable. Around 12% of the 1871 Cornwall population died before the 1881 census.[27] Hence, incorporating marriage and death records has the potential to

[26]There are 53,444 marriage records in Cornwall (half from brides, half from grooms) between the second quarter of 1871 to the first quarter of 1881 (search performed at http://www.freebmd.org.uk/ on 3 March 2014). A small percentage of these records are likely duplicates.

[27]FreeBMD records 66,967 deaths in Cornwall between the second quarter of 1871 and the first quarter of 1881. However, this number includes duplicates, as well as infants who lived and died between censuses. The organization, A Vision of Britian Through Time, reports total deaths and infant deaths during census years. Non-infant deaths at Vision of Britain consistently

Table 7: Forward Matching Rates by Registration District

		Forward Matching Rates (%)				
RD	Ave Pop	1841–51	1851–61	1861–71	1871–81	1881–91
Bodmin	19,600	29.3	40.2	44.1	44.1	45.2
Camelford	8,000	31.4	39.6	45.3	41.6	44.6
Falmouth	23,300	29.8	40.8	40.4	39.7	39.1
Helston	27,600	33.5	47.3	46.4	44.9	47.6
Launceston	16,300	29.0	39.6	42.7	42.0	43.3
Liskeard	29,100	31.1	42.3	43.8	41.0	44.6
Penzance	51,700	33.6	45.8	47.1	44.9	47.0
Redruth	51,100	33.7	44.3	41.2	39.0	43.9
Scilly Islands	2,200	32.1	51.1	48.9	54.4	47.8
St Austell	31,500	32.4	44.2	42.5	43.8	45.6
St Columb	16,500	32.0	43.5	43.3	44.5	44.6
St Germans	16,900	25.8	36.3	39.0	39.7	39.3
Stratton	8,000	28.1	38.6	39.7	40.7	43.3
Truro	39,900	33.2	43.3	43.5	41.9	44.9
TOTAL	344,600	31.9	43.1	42.4	42.3	44.5

boost the percentage of records that are accounted for from one census to the next to well over 60%.[28]

In order to construct a panel dataset, I merely link the individual record numbers from the matches in each of the census pairs. There are 714,615 matches across the five census pairs after the refinements (see Table 5). Most of these (58.5%) can be strung together in chains which follow individuals over three to six censuses. Put together, the panel dataset contains 461,053 individuals that are matched in two or more censuses; 164,588 that are matched in three or more censuses; 59,976 in four or

account for about 65% of the number of death records at FreeBMD for the years 1871, 1881, and 1891. So the 12% death rate between the 1871 and 1881 census is based on 65% of 66,967. See http://www.freebmd.org.uk/ and http://www.visionofbritain.org.uk/ (accessed 3 March 2014).

[28] Ship manifests, census records from neighboring counties, and other emmigration records could boost the matching rate even more; and the demographic picture could be further fleshed out with probate records, obituaries, and local directories.

Table 8: Distribution of Start and End Years in Linked Panel Dataset

		End Year				
		1851	1861	1871	1881	1891
Start Year	1841	62,489	24,604	9,896	5,515	6,120
	1851	-	60,476	25,253	9,955	11,243
	1861	-	-	53,274	19,367	17,247
	1871	-	-	-	45,776	35,388
	1881	-	-	-	-	74,450

more censuses; 22,878 in five or more; and 6,120 individuals that can be indentified in all six censuses. While there is still ample room for improvement, as it stands, this is quite possibly the largest panel dataset for Victorian England at present. Table 8 shows the distribution of the first census and last census where an individual has been identified.[29] The panel could potentially be improved by matching between non-adjacent censuses. If a person is absent or difficult to identify in one census, it may still be possible to find them in the next.

5. Application: Occupational Following

It is not uncommon to learn that a person is in the same occupation as their father or another family member—a pattern which has been named occupational following.[30] Indeed, occupations tend to be perpetuated along family lines, and a vast number of Western surnames (Baker, Cooper, Tanner, etc.) allude to what in many cases was probably a multigenerational family trade. The Victorian era provides an interesting backdrop for examining occupational following. Formal schooling standards were steadily growing during this era, propelled in part by a series of educational legislations that began in 1870.[31] However, basic education in literacy and arithmetic did not of itself prepare a child for a career as a stonecutter

[29]It is likely that many people are represented as two or more individuals in the panel dataset. For instance, two censuses may match a girl in her childhood. Then two or more subsequent censuses may match the same girl as a married woman.

[30]The term occupational following at least dates back to Laband and Lentz (1983). Those authors have since used it in a long list of papers, as have other scholars.

[31]Schooling for children ages five to ten became compulsory in Britain in 1880. Information about these educational laws can be found on the UK Parliament's website: http://www.parliament.uk/about/living-heritage/transformingsociety/livinglearning/school/ (accessed 21 June 2014).

or a shoemaker. Those were skills that were often acquired through some form of apprenticeship—frequently with a father teaching his son.

From the new panel data set, I have identified four cohorts of father-son pairs. Each pair gives the father's occupation when the son is between the ages of 7 and 16. In a skills based economy with little formal education, these are formative years for exposing a child to an occupation. The second part of the father-son pair is the son's occupation when he is between the ages of 21 and 30. Given that censuses occur at ten-year intervals, the father's observation is either ten or twenty years before the son's observation. The cohorts are defined by the census in which the son's occupation is recorded (1861, 1871, 1881, and 1891). For the sons, each of these cohorts represent roughly one-fourth of Cornwall's working male population between the ages of 21 and 30.[32]

For analyzing occupational patterns, I have divided the economy into 18 sectors. The occupational groupings for these sectors, described in Table 9, are based in part on the industry divisions employed by the occupational coding scheme for the 1911 UK Census. I have, however, separated and regrouped many occupations based on their rate of occurrence in the data and on the proximity of skills required for different occupations. For instance, due to the overwhelming size of the agricultural industry, I have separated it into six sectors. Other sectors, such as Letters or Fine Crafts, are a miscellany of occupations that are at best loosely related in terms of the type of the product or the requisite training.

The distribution of the workforce across the 18 sectors can be seen in Table 10. The table first lists the distribution for all males in Cornwall between the ages of 21 and 30 in each census year from 1861 to 1891. Distributions for the sons and fathers in each cohort make up the remainder of the table. By far, mining is the largest sector, accounting for 41% of the sons in the 1861 cohort (36% among all males 21–30 and 35% of fathers). Over the coming decades, however, the mining sector in Cornwall collapsed so that by the 1891 cohort it employed roughly half as many people. The depletion of the mining sector was a major factor in the decline of Cornwall's population over the later half of the nineteenth century (see Table 1).[33] The three primary agricultural sectors—farmers, relatives of farmers, and agricultural labourers—together make up a sizeable block of the economy, and the largest

[32]Many more father-son pairs can be identified. The prominent constraint here is having an occupation listed for both the father and the son during the correct time intervals.

[33]The Wikipedia article "Cornish Diaspora" estimates that roughly a quarter of a million Cornish people, most of them miners, left the UK between 1861 and 1901.

block once mining goes into decline. Sons were more likely to be working on their father's farm than to have their own farm by the time they were 21–30 (and only a smattering of fathers were working on a relative's farm when they had a son aged 7–16). Fishing and boating, woodworking, and the metal and machine sectors each comprises about 5% to 10% of the workforce from decade to decade. Across each of the sectors (mining being the primary exception) there are only minor fluctuations in the distributions from one census to another.[34]

The persistence of occupational following varies considerably across sectors. Combining all cohorts, Table 11 provides the transition probabilities from a father's occupation to that of his son's. Percents sum across the rows so that, for instance, of those who had a father in the meat and cereal sector, 46% entered the same sector, 10% went into mining, 5% became woodworkers, etc. Children of miners and of those in the fishing and boating sector have the highest propensity to follow in their father's footsteps (67%). Masonry has a 60% continuance rate, and if farmers and their relatives are combined, then farming comes in at 62%. More commonly, however, the rate at which son's enter their father's profession tends to fall between a quarter and a half. Sons of tavern and livestock workers typically scout out more fertile ground in other professions—a mere 12% maintaining their father's trade in each of these sectors.

There is a great deal of transitioning among and into low-skilled occupations. Children of livestock workers and common labourers will commonly switch to the mines or to other labourer positions. Since these sectors are dominated by manual hired labour, cultivating a skill set is likely not as pertinent as simply finding employment. Even for those who had a father in a more skilled or capital intensive sector, large numbers fell back on the security of low-skilled employment. Mining in particular attracted workers from each background in large numbers.

Another feature of Table 11 is that, conditional on a son switching away from his father's occupation, there is a wide dispersement of destination occupations. The son of a shoemaker may very well decide to pick up masonry or woodworking, or even be a tailor. The same can also be said for the son of a common labourer. Such transitions necessarily involve the cost of skill and capital acquisition—two principal barriers to entering a new occupation. The full opportunity cost is likely highest for those who have already learned profitable skills at the tutelage of their fathers. While complete reasons for changing occupations may be unclear, the cen-

[34]The Labourer—Other sector does have a sizeable jump between 1861 and 1871 which may well be catching some of the displaced miners.

Table 9: Occupational Sectors

	Industry	Description	Abbr
1	Farmer	Principal farmer on a farm of any size	Farm
2	Farmer—Relative	Workers on a relative's farm (e.g. farmer's children)	FRel
3	Cereal & Meat	Millers, butchers, other dealers in meat and grain	Meat
4	Grocer & Baker	Dealers of baked goods, produce, and other groceries	Groc
5	Livestock, etc.	Employed with horses, cows, other livestock, or in sundry agricultural trades	Live
6	Labourer—Agr.	Low-skilled, hired agricultural labourers	LabA
7	Labourer—Other	Domestic, railway, dockyard, and other low-skilled labourers	LabO
8	Mining	Workers and supervisors in various mines (tin, copper, stone, etc.)	Mine
9	Military & Police	All branches of the military, police, municipal offices	Govt
10	Metal & Machines	Metal manufacturers, blacksmiths, engine operators, etc.	Metl
11	Fine Crafts	Watchmakers, chemists, building workers, and various fine arts	Fine
12	Letters	Medicine, education, law, clergy, printing, clerking, etc.	Ltrs
13	Clothing	Tailors, drapers, milliners, and other cloth, cord, and clothing workers	Clth
14	Fishing & Boating	Fishermen, seamen, pilots, and bargemen	Boat
15	Masonry	Masons and other stonecutters	Masn
16	Shoes & Leather	Shoe and boot makers, saddlers, other workers in leather and skins	Shoe
17	Woodworkers	Carpenters, shipwrights, sawyers, coopers, wheelwrights, etc.	Wood
18	Taverns & Sales	Innkeepers, shopkeepers, brewers, street vendors, etc.	Tvrn

Table 10: Percent of Workforce in Each Sector

		All Males 21–30				Sons				Fathers			
	Ind	1861	1871	1881	1891	1861	1871	1881	1891	1861	1871	1881	1891
1	Farm	3.2	3.4	3.9	5.2	3.7	4.5	4.5	6.7	16.0	16.5	17.4	18.8
2	FRel	3.4	4.8	6.4	4.7	6.6	7.7	10.9	8.3	0.1	0.1	0.1	0.4
3	Meat	1.8	2.0	2.1	2.3	2.0	2.0	2.6	2.7	2.0	2.3	2.4	2.1
4	Groc	0.9	1.1	1.7	2.0	0.8	0.9	1.4	1.8	0.8	1.2	1.7	1.5
5	Live	3.5	1.9	2.4	2.7	1.7	1.6	1.9	2.2	1.2	1.4	1.7	2.1
6	LabA	11.6	12.2	13	15.6	7.7	11.5	12.7	15.1	14.1	15.1	14.6	16.2
7	LabO	5.6	9.1	10.7	10.3	3.6	6.1	7.7	8.2	3.7	3.6	4.8	6.9
8	Mine	35.7	26.7	19.1	18.1	41.4	29.2	20.8	18.4	34.8	29.1	25.5	20.9
9	Govt	1.0	1.0	1.0	1.2	0.5	0.4	0.4	0.6	0.3	0.4	0.3	0.4
10	Metl	5.3	6.1	5.1	6.1	5.8	6.6	5.5	6.3	4.3	5.3	5.2	5.3
11	Fine	1.1	1.5	2.0	2.7	1.1	1.1	1.8	2.3	0.7	0.9	1.1	1.4
12	Ltrs	2.9	3.5	4.7	5.0	2.2	2.5	3.6	3.6	1.3	1.4	1.3	1.1
13	Clth	2.3	2.3	2.0	2.3	2.2	2.3	2.0	2.4	1.7	2.1	1.8	1.6
14	Boat	8.7	8.5	12.1	8.4	4.5	5.4	9.1	7.4	3.2	4.7	7.8	8.4
15	Masn	3.2	4.8	4.7	4.9	4.1	5.5	5.2	5.3	4.5	4.5	4.1	4.2
16	Shoe	3.1	3.1	2.3	1.8	3.9	3.6	2.5	1.8	3.6	4.2	3.4	2.6
17	Wood	6.0	7.3	5.9	5.6	7.6	8.4	6.8	6.1	5.9	6.2	5.8	5.1
18	Tvrn	0.7	0.9	0.8	1.1	0.6	0.7	0.6	0.8	1.8	1.1	1.1	1.2
Employed		22,559	19,624	18,953	19,533	4,358	4,942	4,929	5,323	4,358	4,942	4,929	5,323

sus provides additional information that at least allows for the identification of key correlates.

The linked census dataset makes it possible to tell whether the son is living in a different civil parish as an adult than he did as a child. If he did move, the distance can be calculated from the geocodes. Birth order within the son's family can also be approximated since the ages and relations of each household member are included in the census. A caution here is that the oldest son and heir, if that custom is followed, may not be living at home in a given census. Grown children typically leave home, and younger children on occasion can also be found living away from their parents (such as in apprenticeships or as servants). To control for primogeniture, I construct a variable that flags those who are clearly not the firstborn son. Another variable of note is the son's position within the household in the census where he is between 21 and 30. If the son is the head of the household that implies a different set of responsibilities than if he is listed as a son.

Given these variables, I implement a multinomial logistic regression for predicting a son's propensity to switch from one occupation to another. In doing so, I streamline the 18 occupational sectors into just four: farming (Farm and FRel); labour (Live, LabA, LabO, and Govt); mining (Mine); and trade (all other sectors).[35] The question then is, given that a father works in either farming, labour, mining, or trade, what is his son's probability of being employed in each of these industries, conditional on his characteristics in the census. Specifically, whether he moved from one civil parish to another (*Moved*), and if so how far (*Distance* in kilometers); whether he has older brothers (*Not.1st*); if he is the head of his own household (*rel.Head*), living as a son in his father's household, or has some other relationship (*rel.Oth*); and also the decade of his cohort.

Table 12 shows the regression results for sons who have fathers in the farming or labour categories. For those with fathers in mining or trade, the results are in Table 13. In each of the four multinomial regressions, the coefficients are in comparison to the son going into his father's occupation. Since the trade category contains a large number of sectors with diverse specialized skills, I further specified whether the son went into the exact same sector as his father or a different sector in the trade category; so the coefficients in that regression are in reference to the son following his father's footsteps exactly.

[35]The inclusion Govt (Military & Police) as a labour sector is based on the assumption that enrollment in the military required little more than being healthy and fit.

Table 11: Given a Father's Occupation, Percent of Sons Working in Each Sector (All Cohorts)

		Son's Occupation																		
		Farm	FRel	Meat	Groc	Live	LabA	LabO	Mine	Govt	Metl	Fine	Ltrs	Clth	Boat	Masn	Shoe	Wood	Tavn	**Count**
Father's Occupation	Farm	20	42	2	1	1	12	3	9		2		2	1	1	1	1	3		3,373
	FRel	24	41	3			15		3				3			6		6		34
	Meat	3	3	46	2	3	6	6	10	1	4	2	1	2	3	2	1	5		429
	Groc	3	2	3	25	1	6	6	8	1	5	4	9	5	4	2	3	10	2	256
	Live	4	4	2	2	12	14	11	12		6	3	5	2	4	3	4	11	1	318
	LabA	3	3	1		3	37	11	17	1	4	1	1	2	3	4	3	6	1	2,940
	LabO	2	2	2	1	3	14	21	15	1	7	2	5	3	7	4	4	8	1	947
	Mine	1	1	1	1	1	6	4	67		6	1	2	1	2	2	1	3		5,319
	Govt	3			1	1	4	15	10	3	7		10	6	19	3	3	12	1	68
	Metl	1	1	1	1	2	6	6	24		37	2	3	2	2	4	2	5	1	986
	Fine	2	1		4	2	4	4	9		7	30	8	2	3	11	3	13	2	198
	Ltrs	3	2	1	2	2	4	9	15		9	5	37	2	3	3	2	3		249
	Clth	1		1	1	2	6	5	9	1	7	3	6	35	4	3	6	7	1	348
	Boat	1		1	1	1	4	5	3		2	1	2	2	67	1	2	6		1,203
	Masn	1		1	1	2	3	4	8		3	4	2	2	2	60	2	5		844
	Shoe	1		3	2	3	5	7	9		5	4	5	4	5	6	30	9	1	674
	Wood	2	1	2	2	1	6	6	9		5	3	4	2	4	3	2	46		1,120
	Tavn	5	7	6	2	1	8	7	15		5		9	3	5	2	2	12	11	246

Table 12: Multinomial Logistic Regressions for Predicting a Son's Occupation Given His Father's Occupation (Part 1)

	Father: Farm			Father: Labour		
	Labour	Mine	Trade	Farm	Mine	Trade
Intercept	-1.833^{***}	-1.739^{***}	-2.049^{***}	-1.194^{***}	0.064	0.369^{***}
	(0.135)	(0.147)	(0.144)	(0.191)	(0.115)	(0.103)
Moved	0.211	0.334	0.602^{***}	0.193	-0.193	-0.147
	(0.136)	(0.179)	(0.143)	(0.178)	(0.114)	(0.089)
Distance	0.017^{**}	0.001	0.019^{***}	0.000	0.010^{*}	0.012^{***}
	(0.005)	(0.008)	(0.005)	(0.008)	(0.004)	(0.003)
Not.1st	0.156	0.206	0.335^{**}	-0.264	0.076	-0.059
	(0.104)	(0.133)	(0.112)	(0.157)	(0.097)	(0.077)
rel.Head	1.317^{***}	1.205^{***}	0.847^{***}	-0.917^{***}	0.157	-0.315^{***}
	(0.112)	(0.143)	(0.127)	(0.156)	(0.103)	(0.083)
rel.Oth	1.727^{***}	1.104^{***}	1.649^{***}	-2.754^{***}	-0.599^{***}	-0.229^{*}
	(0.176)	(0.257)	(0.181)	(0.465)	(0.160)	(0.108)
1871	-0.210	-0.698^{***}	-0.168	-0.354	-1.198^{***}	-0.638^{***}
	(0.149)	(0.165)	(0.157)	(0.224)	(0.128)	(0.113)
1881	-0.447^{**}	-1.768^{***}	-0.628^{***}	-0.468^{*}	-1.710^{***}	-0.764^{***}
	(0.150)	(0.212)	(0.165)	(0.220)	(0.138)	(0.112)
1891	-0.248	-1.197^{***}	-0.399^{**}	-0.513^{*}	-1.571^{***}	-0.804^{***}
	(0.142)	(0.174)	(0.154)	(0.210)	(0.125)	(0.107)

Standard errors in parentheses. Significance levels: * 5%, ** 1%, *** 0.1%.

Table 13: Multinomial Logistic Regressions for Predicting a Son's Occupation Given His Father's Occupation (Part 2)

	Father: Mine			Father: Trade			
	Farm	Labour	Trade	Farm	Labr	Mine	Trade
Intercept	-3.910^{***}	-3.201^{***}	-2.162^{***}	-2.658^{***}	-2.064^{***}	-1.415^{***}	-0.663^{***}
	(0.252)	(0.140)	(0.100)	(0.205)	(0.120)	(0.106)	(0.081)
Moved	0.188	0.489^{***}	0.294^{**}	0.806^{***}	0.746^{***}	0.579^{***}	0.587^{***}
	(0.251)	(0.115)	(0.097)	(0.208)	(0.106)	(0.115)	(0.085)
Distance	0.025^{***}	0.024^{***}	0.022^{***}	0.011	0.013^{***}	0.011^{**}	0.007^{*}
	(0.007)	(0.003)	(0.003)	(0.007)	(0.003)	(0.004)	(0.003)
Not.1st	-0.104	-0.130	-0.052	-0.423^{*}	0.000	0.032	0.009
	(0.198)	(0.096)	(0.077)	(0.178)	(0.086)	(0.090)	(0.063)
rel.Head	-0.440^{*}	0.134	-0.015	-0.434^{*}	0.493^{***}	0.550^{***}	-0.019
	(0.210)	(0.105)	(0.084)	(0.177)	(0.092)	(0.095)	(0.068)
rel.Oth	-0.994^{*}	1.025^{***}	0.773^{***}	-1.776^{***}	0.716^{***}	0.032	0.421^{***}
	(0.479)	(0.140)	(0.119)	(0.475)	(0.129)	(0.161)	(0.101)
1871	0.417	0.680^{***}	0.538^{***}	-0.099	0.122	-0.388^{**}	-0.005
	(0.286)	(0.148)	(0.105)	(0.241)	(0.128)	(0.115)	(0.090)
1881	1.079^{***}	1.422^{***}	0.897^{***}	-0.028	0.156	-0.792^{***}	-0.081
	(0.268)	(0.141)	(0.107)	(0.232)	(0.125)	(0.123)	(0.089)
1891	0.829^{**}	1.652^{***}	1.045^{***}	-0.097	0.410^{**}	-0.803^{***}	-0.030
	(0.293)	(0.142)	(0.110)	(0.234)	(0.122)	(0.123)	(0.089)

Standard errors in parentheses. Significance levels: * 5%, ** 1%, *** 0.1%.

The intercepts reveal that—all else equal—sons of labourers are more likely to switch to a trade than to become labourers themselves. This, however, is the exception since, based on the intercepts alone, sons of farmers, miners, and those in trade are highly prone to remain in their father's occupation. Moving from one parish to another is an important indicator of changing occupations. For those growing up on farms, there is a strong correlation between moving and switching to one of the trade sectors. If a son's father is already in one of the trade sectors, then transitioning to any other occupation is typically associated with a move. For instance, both the coefficients for *Moved* and *Distance* are particularly strong for those switching from trade to labour.

In terms of both magnitude and significance, the only time that primogeniture inheritance appears to play a role is in farming. Sons of fathers who are not the firstborn son have a significantly higher occurrence of going into a trade than of continuing on in farming. Not being the firstborn son also has a positive effect on switching from farming to labour or mining, although the coefficients are not significant. Even for those whose father was not in farming when they were young, having older brothers still reduces the probability of transitioning into farming. This can be seen in each of the other three regressions, although the effect is largest (and significant) for those whose father worked in trade.

The collapse of the mining industry over the decades is evident in each regression. The 1871 cohort is significantly less likely than the 1861 cohort to switch into the mining industry, and sons of miners are more likely to obtain jobs outside of mining. The magnitude of the coefficients only amplify this pattern over the 1881 and 1891 cohorts.

Whether a son still lives in his father's household as a young adult is another telltale sign of the son's occupation. The exact meaning of the sign, however, is specific to the different industries. A son who, like his father, is in farming is likely listed in the census as a son in his father's home. Even if these sons eventually go on to inherit the farm, at the age of 21 to 30 their father is likely still alive and working. Being listed as the head of the household is in many other instances a strong correlate of following a father's occupation. Sons of labourers, for example, are more likely to be labourers themselves than to go into farming or trade if they are the head of the household. Being listed as some other position within the household (perhaps as a brother, son-in-law, hired hand, etc.) produces coefficients that have a similar effect in the different regressions to being listed as the head of the household.

6. Conclusion

With over 160,000 individuals linked across three or more censuses, this paper has detailed the construction of what is quite possibly the largest panel dataset for Victorian England at present. The matching algorithm utilizes SVM technology to classify matches; a multi-level group linking technique for monitoring changes in household composition; and a nested matching process for tracking migration. Yet, there are far taller mountains to climb, and the methodologies presented here can certainly be applied on a much larger scale. The course of an entire nation, with the comings and goings of each of its citizens, could be captured over generations in a way never before seen by linking all of its available censuses.

As an example, I have presented a brief analysis of father-son occupational patterns across sectors. This includes assessing the connection between birth order, moving, and household roles on whether a son remains in his father's occupation. The question of primogeniture is one that can certainly be probed in future work.

Acknowledgments

I express gratitude to Michael McCormick of the Cornwall Online Population Project for his generous permission to use the organization's collection of more than 2 million census record transcriptions.

Data Sources

GB Historical GIS / University of Portsmouth, Cornwall RegC through time | Life and Death Statistics, *A Vision of Britain through Time*. This work is based on data provided through www.VisionofBritain.org.uk and uses historical material which is copyright of the Great Britain Historical GIS Project and the University of Portsmouth.

McCormick, M., *Cornwall Online Census Project*, 1841 [computer file]. Colchester, Essex: UK Data Archive [distributor], November 2005. SN: 5221, http://dx.doi.org/10.5255/UKDA-SN-5221-1.

McCormick, M., *Cornwall Census Returns*, 1851 [computer file]. Colchester, Essex: UK Data Archive [distributor], March 2011. SN: 6738, http://dx.doi.org/10.5255/UKDA-SN-6738-1.

McCormick, M., *Cornwall Online Census Project*, 1891 [computer file]. Colchester, Essex: UK Data Archive [distributor], July 2004. SN: 4978, http://dx.doi.org/10.5255/UKDA-SN-4978-1.

UK Ordnance Survey, *1:50 000 Scale Gazetteer*. Contains Ordnance Survey data © Crown copyright and database right 2013.

Bibliography

Chang, C.-C., Lin, C.-J., 2011. LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm

Christen, P., 2008. Febrl: An open source data cleaning, deduplication and record linkage system with a graphical user interface. In: 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1065–1068.

Cortes, C., Vapnik, V., 1995. Support-vector network. Machine Learning, 20, 273–297.

Goeken, R., Huynh, L., Lynch, T.A., Vick, R., 2011. New methods of census record linking. Historical Methods: A Journal of Quantitative and Interdisciplinary History, 44:1, 7–14.

Fu, Z., Christen, P., Boot, M., 2011a. Automatic cleaning and linking of historical census data using household information. In: Workshop on Domain Driven Data Mining, held at IEEE ICDM. Vancouver.

Fu, Z., Christen, P., Boot, M., 2011b. A supervised learning and group linking method for historical census household linkage. In: AusDM, CRPIT, vol 125. Ballarat, Australia.

Laband, D.N., Lentz, B.F., 1983. Like Father, like son: Toward an economic theory of occupational following. South. Econ. J. 50 (2), 474–493.

Long, J., Ferrie, J., 2013. Intergenerational Occupational Mobility in Great Britain and the United States since 1850. Amer. Econ. Rev., 103 (4), 1109–1137.