Digging into Data White Paper

Mining Microdata: Economic Opportunity and Spatial Mobility in Britain, Canada and the United States, 1850-1911

Baskerville, Peter, Professor of History and Classics/Humanities Computing, University of Alberta, Edmonton, Alberta, Canada.

Dillon, Lisa Y., Associate Professor and Co-Director, Programme de recherche en démographie historique, Département de Démographie, Université de Montréal, Montréal, Québec, Canada.

Inwood, Kris, Professor of Economics, University of Guelph, Guelph, Ontario, Canada.

Roberts, Evan, Assistant Professor of Population Studies, University of Minnesota, Minneapolis, Minnesota, United States of America.

Ruggles, Steven, Regents Professor of History and Population Studies, and Director of the Minnesota Population Center, University of Minnesota, Minneapolis, Minnesota, United States of America.

Schürer, Professor Kevin, Professor of History and Pro Vice Chancellor, University of Leicester, Leicester, United Kingdom.

Warren, John Robert, Professor of Sociology, University of Minnesota.

Project progression

Mining Microdata: economic opportunity and spatial mobility in Britain, Canada and the United States was funded in Round 2 of the Digging into Data Challenge in 2011. The team learned of the award in October 2011, enabling a preliminary meeting of the project principals at the annual meetings of the Social Science History Association in November 2011.

Although initially scheduled for two years, the grant was extended in two no-cost extensions because of significant delays in obtaining key data. A contract to obtain the British 1911 data in a format suitable for data record linkage was not agreed until nearly the end of the initial grant period, and data was not delivered until late-2014. The revision to the data delivery schedule anticipated in the grant required revision to the project's analysis and publication timeline. We have devoted more time than initially anticipated to describing the theory and practice of record linkage in large sets of individual-level records.

Despite these challenges we have presented initial analytical results on the first cohort (1850/1-1880/1) proposed for analysis, and developed a suite of analytical programs that will be run on the datasets constructed for the second cohort. Our white paper includes a publication on social mobility in the first cohort; as well as extensive discussion of our methods in published papers.

The project principals met to discuss and develop the project on several occasions during the grant period

- May 2012 in Guelph, Ontario. This meeting was hosted by the Canadian team with local arrangements made by Kris Inwood, and held in conjunction with an international meeting about longitudinal data analysis in historical social science.
- April 2013 in Leicester, United Kingdom. This meeting was hosted by the British team, with local arrangements made by Kevin Schürer. We met for three days to outline plans for data construction and the eventual analytical papers. Our publications have followed the plans developed at this meeting, modified to accommodate the delayed production of the second cohort (1880/1-1910/11) of data.
- April 2014 in Vienna, Austria. The Canadian and British teams and Evan Roberts from the U.S. team met in Vienna in conjunction with the European Social Science History Conference, at which we presented a jointly authored paper.
- November 2014 in Toronto, Ontario. The Canadian and U.S. teams met in conjunction with the annual meetings of the Social Science History Association conference.

Project management

The team for this Digging Into Data project had been working together for a decade on developing historical census data into a consistent international dataset as part of the North Atlantic Population Project (http://www.nappdata.org). This was our first joint foray into substantive research, though many of the team members had worked on substantive research together in different groups of 2-3 people.

We continued the style of project management that had been developed in the North Atlantic Population Project, and worked successfully. We agreed on a schedule of in-person meetings early in the project, and established goals to be achieved in advance of those meetings. At the meetings we allocated by mutual agreement responsibility for different aspects of the project to be primarily led by particular institutions or individuals.

Key data creation, analysis and writing components were allocated as follows

- Theoretical development of record linkage algorithms: Guelph team led by Kris Inwood with computer scientist Luiza Antonie.
- Evaluation of household relationships (father-son links) in Canadian census data which did not explicitly enumerate household relationships: Montréal team led by Lisa Dillon.
- Development of methods for analysis of small area geographic context: British team led by Kevin Schürer.
- Supervision of record linkage production: Minnesota team led by Evan Roberts
- Drafting of article on record linkage techniques and methods: Canadian team (Inwood, Dillon, Baskerville)
- First draft of article on comparative social mobility: Evan Roberts.

By allocating key leadership roles for particular tasks to different members of the team we ensured that the project would make progress in multiple areas, and that we would not duplicate effort. The work produced by each lead individual or team was reviewed by the other members of the collaboration prior to submission to conferences or journals.

Throughout the project we remained in regular contact between in-person meetings by phone and email.

Project challenges

Our project faced two major, and inter-related challenges. At the time of grant submission we anticipated that we would take delivery of the 1911 British census dataset shortly after the grant began in early 2012, and complete linking of the 1881-1911 panel within the 2012 year; allowing us to focus on data analysis in Year 2 of the grant. In fact, we did not receive the data until late-2014, primarily owing to delays in obtaining a license from Findmypast — the genealogical company that produced the data after their corporate ownership changed. Without these necessary permissions we were unable to make progress on a critical component of the grant: creating a panel of men aged 0-20 in 1881, and linked to their adult observations in 1911.

After obtaining the data, we ran into a second set of challenges: the methods we had derived to create linked datasets with smaller pairs of datasets were not computationally efficient for linking complete enumerations at both ends of the observation period. For all of our other cohort panels we had just one complete count enumeration, as summarized in the following table.

Table 1. Sample densities in Mining Microdata cohorts

Country	1850/1-1880/1 sample densities	1880/1-1910/11 sample densities
Canada	20% to 100%	100% to 5%
Great Britain	2% to 100%	100% to 100%
United States	1% to 100%	100% to 1%

Our approach to record linkage relies on comparisons of names within blocks defined by the intersection of age, birthplace, sex (and race in the United States). These are characteristics that should not change over time, and allow identification of a set of individuals where links are not biased by changes in social outcomes. Within each block defined by a single year of birth (surrounded by a small window to account for inaccuracies in enumeration of ages), birthplace, and sex we compare the names of all individuals who appear in both datasets. Thus for all individuals born in 1875 in a particular county (province or state) we compare the similarity of names between the two datasets. If there are 2000 people in the first dataset and 1000 people in the second dataset we must make 2,000,000 (1000 x 2000) comparisons of the similarity of names. Links are made from the pool of people for whom there is no closely competing person with a similar name.

Individuals with a common name in large entities are unlikely to be matched. John Smith born in Ontario, Yorkshire, or New York is never going to be matched because so many other individuals have the same representation in the data. Men with names that are genuinely rare or unique, but have a close similarity to another name in the dataset will also not be matched. Thus "Jahn Smithson" from Ontario, Yorkshire, or New York may be the only man with that name in the few years surrounding his birth, but his name's similarity to a more common name means it is possible he really was John Smith, and his name was spelt incorrectly. Our record linkage procedure is designed to ensure that people are not linked because they are erroneously unique. If there is a close competitor in age or spelling from the same birthplace, a match is not made.

Our record linking procedures built on a significant existing literature. Our goals, however, differed significantly from those of most data mining applications of record linkage. The primary goal of most data mining has been to maximize the number of valid links. Our objective is different: we do not focus on maximizing the linkage rate. Instead, our procedures are designed to maximize the *representativeness* of the linked cases and the *accuracy* of the links. This means we pay close attention to potential sources of selection bias, and ignore information routinely used by other record-linkage procedures. Although we cannot eliminate selection bias for unobserved characteristics, we can adopt procedures that greatly reduce the potential for bias compared with previous approaches.

Our algorithm relies exclusively on characteristics that should not change over time. At minimum, these variables are first name, last name (for men and for women who do not marry between observations), birth year, sex, and place of birth. Most record linkage software makes use of a broader range of characteristics to confirm links and resolve ambiguities, but that

approach introduces bias. For example, if we used spouse's characteristics to confirm linkages, we would bias the sample in favor of persons who remained married to the same person for multiple decades, and such persons are not representative with respect to either occupational or geographic mobility.

A challenge posed by our approach is that the limited set of variables we use cannot uniquely identify all individuals. To take the worst-case scenario—the most common male name with the most common birthplace—the 1880 U.S. census has 17 white men aged 33, named John Smith, and born in New York. Even this example understates the problem, because it assumes an exact match of name and age. Errors in enumeration and transcription cause a significant proportion of matches to be imperfect: linking must be carried out probabilistically, allowing for imperfect correspondence of name and age. Whenever there is more than one possible match, we must exclude all potential matches. This eliminates many true matches, but is necessary to minimize false matches. False matches would lead to systematic upward bias for transition rates—such as migration and occupational mobility—and therefore must be avoided at all cost.

Because our linking strategy must rely heavily on names, we needed an approximate string comparison algorithm. We used the Jaro string comparator as modified by Winkler. This algorithm computes a similarity measure between 0.0 and 1.0 based on the number of common characters in two strings, the lengths of both strings, and the number of transpositions, accounting for the increased probability of typographical errors towards the end of words. In addition to using a string comparator, we standardize given names to account for diminutives and abbreviations (e.g., "Willie" and "Wm." are transformed into "William.") Such name-cleaning techniques are language-specific and must be customized for each language of enumeration. This work draws on the rich body of research on name cleaning. Finally, we used both NYSIIS and Double-Metaphone phonetic name coding, which provide multiple encoded strings corresponding to variant pronunciations.

These procedures worked efficiently in the datasets which paired one 100% dataset with a smaller sample (1% to 20% of the population). In attempting to construct the 1881-1911 British dataset our approach proved infeasible, as the block sizes were very large. For example, in 1881 there are 774,611 men between the ages of 0 and 20 born in Lancashire, and 508,196 from the same birth cohort and county of birth still alive in 1911. Within the 5 year age windows that we make comparisons there are a smaller number of pairs, but still an infeasible number. For example in 1911 there are 149,743 Lancashire men between the ages of 31 and 35, and a corresponding 228,674 men from the same birth cohort (1876-1880) and birth county in 1881. This would require a total of 34,242,330,782 (34 billion) comparisons of name similarity centered only around the birth cohort of 1878 in one county. While this is a worst case example, it suggests the scale of the problem in record linkage blocks that are by necessity of the original data source defined broadly. We lack, for example, any more detailed specification of birthplace that would reliably differentiate people born in different parts of the provinces or states. Even in Britain where parish of birth is asked in the census, it is not reliable as many individuals simply do not recall the parish of birth accurately or consistently, or provide parishes that do not exist. Parish of birth, in other words, is not a salient identifier for individuals. We note that in modern sources with numerate populations, by comparison, exact date of birth (day, month, and year) is a powerful tool for identifying people in different sources even without unique identifiers (such as identification numbers). The combination of day, month, and year sub-divides the population

even more finely, and while people make frequent errors with their integer ages (and year of birth) birthdays within the year are typically remembered accurately.

We persisted in attempting to create the linked files with large block sizes for much of 2015; and despite having access through the Minnesota Supercomputer Institute to a very powerful computing suite, the problem is not amenable to a brute force solution. It was clear that we needed to refine our blocking strategy. In doing so, we used the linked samples we had already created from 1 to 5% samples of the United States census to the 1880 complete count dataset. These datasets were constructed with blocks of age (+/- 2 years), sex, race and state of birth. With limited other variables available for blocking we decided to investigate blocks based on the first letter of the last name.

Table 2. Proportion of male sample linked to 1880 with last name initials agreement

1850	1860	1870	1900	1910	1920	1930
99.32	98.76	98.79	99.03	97.89	99.46	99.95

The United States' samples suggested this approach would significantly reduce the size of the blocks, and lead to only a modest reduction in the chances of finding the right person across long time periods. In panels constructed with no restrictions on the name comparisons (every name within a birthplace/age/race block was compared to every other name) more than 99% of our matches agreed on the first letter of the last name. Comparing "John Smith" to "Edward Jones" is computationally inefficient. Moreover, we were able to realize significant additional computational improvements by not comparing "John Smith" to "Johannes Smith". In other words, if there are multiple individuals with the same exact characteristics in one of the datasets we remove them from the comparisons because they can never be definitively matched to anyone.

Our revised approach to constructing panels of links in large datasets retains the approach we began with at the beginning of the project: we aim to create representative links that are weighted for the differential chances of selection into the linked panel; rather than the maximal size of the linked dataset. However, whereas our comparisons within age/birthplace blocks were un-restricted and done in a single pass on smaller datasets, we have refined this approach to deal with the challenges of working with much larger datasets. We have now constructed a working version of a panel of men aged 0-20 in 1881 linked to adult observations in 1911 – from a terminal population of 4.8 million men aged 30-50 we have a dataset of nearly 1 million men linked to childhood observations in 1881. As we anticipated in our proposal, a complete database of the 1850 United States census would become available shortly after the end of the grant period. We plan to implement the record linkage procedures we have developed for the British complete-count pairs on these U.S census pairs to give us a similarly large sample for the first analysis period in the United States.

Evaluation of project

Unanticipated delays in data receipt required us to change our plans for data analysis and publication during the course of the grant. As discussed in the project management section, we devolved responsibility for developing publications to small groups within the project in order to work efficiently. Our cross-sectional datasets for each census, and the linked datasets for each pair of census years are constructed in identical fashion; so that programs developed for initial publications will be able to be run on the six linked datasets to complete the analysis projected in the initial proposal.

Our project had three important components

1. Refining methods for historical census record linkage methods
2. Developing methods to analyze the social conditions in which people originated to explain differences between and within countries in social mobility
3. Measurement of levels of social mobility, and comparison between time periods and countries.

We have published multiple papers discussing the development and evaluation of methods for census record linkage, and two papers advancing methods for analyzing social conditions. The Canadian team took the lead on writing about our strategies for record linkage, with papers by Inwood and Dillon, and co-authors. The British team developed an innovative approach to measuring social conditions in childhood by using a principal components method to identify clusters of types of small areas, reflecting a combination of household structure and economic conditions in parishes. We will be applying these methods to analysis of the United States and Canadian datasets in future work. Using the first pairs of linked datasets that we created, we have published an initial analysis of social mobility focusing on the United States and Britain between 1850/1 and 1880/1. In undertaking this analysis we developed a set of programs to classify occupations into a small number of social classes, and then measure the degree of association between fathers' and son's occupations. We found that the United States was a significantly more mobile society for sons than Britain.

Overall we regard the project as successful despite the need to re-schedule key parts of the analysis to deal with the delays in data receipt. We have published 7 peer reviewed papers, and developed software that will allow us to extend the data analysis to the remaining datasets within the next year. Scholarly interest in the project has been strong, with conference presentations at the European Social Science History Conference, Social Science History Association, IEEE Big Data Humanities workshop, Population Association of America, and International Union for the Scientific Study of Population workshops. An invited presentation on the project was delivered at the University of Colorado.

Methods papers

Luiza Antonie, Kris Inwood, Daniel J. Lizotte, and J. Andrew Ross. "Tracking people over time in 19th century Canada for longitudinal analysis." *Machine Learning*, 95:129– 146, 2013

Luiza Antonie, Kris Inwood, and J. Andrew Ross. Dancing with dirty data: Problems in the extraction of life-course evidence from historical censuses. In *Population Reconstruction*. Springer International Publishing, 2015: 217-242.

Catalina Torres and Lisa Dillon. "Using the Canadian Censuses of 1852 and 1881 for Automatic Data Linkage: A Case Study of Intergenerational Social Mobility". In *Population Reconstruction*. Springer International Publishing, 2015: 243-261.

Kevin Schürer, Tatiana Penkova & Yanshan Shi (2015) "Standardising and Coding Birthplace Strings and Occupational Titles in the British Censuses of 1851 to 1911" *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 48:4, 195-213.

Kevin Schürer and Tatiana Penkova (2015) "Creating a typology of parishes in England and Wales: Mining 1881 census data". *Historical Life Course Studies* 2: 38-57.

Analytical papers

Peter Baskerville, Lisa Dillon, Kris Inwood, Evan Roberts, Steven Ruggles, Kevin Schürer, and Rob Warren (2014) "Economic Opportunity and Spatial Mobility in Canada, Great Britain, and the United States, 1850-1881" *Proceedings of the IEEE Big Data Humanities Workshop*. DOI: 10.1109/BigData.2014.7004446. Link to PDF

Mach Learn (2014) 95:129–146
DOI 10.1007/s10994-013-5421-0

Tracking people over time in 19th century Canada for longitudinal analysis

Luiza Antonie · Kris Inwood · Daniel J. Lizotte · J. Andrew Ross

Received: 21 November 2012 / Accepted: 26 September 2013 / Published online: 1 November 2013

Abstract Linking multiple databases to create longitudinal data is an important research problem with multiple applications. Longitudinal data allows analysts to perform studies that would be unfeasible otherwise. We have linked historical census databases to create longitudinal data that allow tracking people over time. These longitudinal data have already been used by social scientists and historians to investigate historical trends and to address questions about society, history and economy, and this comparative, systematic research would not be possible without the linked data. The goal of the linking is to identify the same person in multiple census collections. Data imprecision in historical census data and the lack of unique personal identifiers make this task a challenging one. In this paper we design and employ a record linkage system that incorporates a supervised learning module for classifying pairs of records as matches and non-matches. We show that our system performs large scale linkage producing high quality links and generating sufficient longitudinal data to allow meaningful social science studies. We demonstrate the impact of the longitudinal data through a study of the economic changes in 19th century Canada.

Keywords Record linkage · Classification · Historical census

Editors: Kiri Wagstaff and Cynthia Rudin.

L. Antonie (✉)
Historical Data Research Unit, University of Guelph, Guelph, Canada
e-mail: lantonie@uoguelph.ca

K. Inwood
Department of Economics and Finance, University of Guelph, Guelph, Canada
e-mail: kinwood@uoguelph.ca

D.J. Lizotte
David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, Canada
e-mail: dlizotte@uwaterloo.ca

J. Andrew Ross
Department of History, University of Guelph, Guelph, Canada
e-mail: jaross@uoguelph.com

1 Introduction

The impact of industrialization, one of the most important topics in history and the social sciences, remains uncertain until we have information that follows individual people through their lives. Millions of records from census, church and military data sources are available from the 19th century, but they must be linked together in order to reconstruct the life-courses of individual people. Computer scientists are collaborating with historians and social scientists to adapt machine-learning strategies for this purpose in a number of countries. In Canada, we are linking millions of records from Canadian censuses taken every ten years (1852–1911) in order to construct life-course or *longitudinal* data. We describe a successful linkage between the 1871 and 1881 Canadian censuses, which span a particularly interesting historical period.

Record linkage is the process of identifying and linking records that refer to the same entities across several databases. If unique identifiers exist for the entities, this is easily done using a database join. Without unique identifiers, one must use attributes common to all of the databases and compare their values to determine whether two records refer to the same entity. The problem of record linkage has been studied in the statistics community for more than five decades (Fellegi and Sunter 1969; Newcombe 1988; Newcombe et al. 1959), and advances in databases, machine learning and data mining have led to a variety of sophisticated methods (Christen 2008; Elfeky et al. 2002). Winkler (2006) and Elmagarmid et al. (2007) offer a detailed discussion of the field. The record linkage process is also referred to as data cleaning (Rahm and Do 2000), de-duplication (within a database) (Bilgic et al. 2006), object identification, approximate matching, approximate joining, fuzzy matching, data integration and entity resolution (Kang et al. 2008). This is a challenging problem. Frequently, common attributes are in different formats in different databases, and they contain typographical and other clerical errors that make naïve rule-based matching ineffective. Furthermore, even in very well-curated databases, it is computationally too costly to evaluate every potential match.

In the context of creating longitudinal data from census data, record linkage refers to finding the same person across several censuses. The recent emergence of 100 percent national census collections enables a systematic identification and linking of the same individuals across censuses in order to create a new database of individual life-course information. A record linkage system for census data relies on attributes describing individuals (name, age, marital status, birthplace, etc.) to determine whether two records describe the same person. Difficulties are presented by different database formats, typographical errors, missing data and ill-reported data (both intentional and inadvertent). Furthermore, not everyone in a census is present in the next one because death and emigration remove people from the population, while births and immigration add new people who were not present in the previous census but who may have characteristics similar to those who were present. Finally, processing the millions of records in a Canadian census requires significant computation. Besides these common challenges, in order to be of scientific value we must ensure that the linked records we produce are representative of the population as a whole, that is, we must avoid any *bias* toward linking one sub-population more than another.

We present solutions to these and other challenges in the first part of the paper, in which we describe a linkage system that incorporates a supervised learning module for classifying pairs of entities as matches or non-matches in order to automatically link records from the 1871 Canadian census to the 1881 Canadian census. In the second part, we evaluate the performance of the linkage system and discuss the results. Our approach follows most closely the pioneering efforts of the North Atlantic Population Project (NAPP) on comparable US

data for 1870 and 1880, where tens of thousands of links were generated (Goeken et al. 2011).

2 Link quality, bias, and variance

The end goal of our record linkage task is to produce datasets that are useful for social scientists. These end-users wish to know how the lives of individuals in Canada changed over time between 1871 and 1881. Ideally they would like to know at the population level, for example, what proportion of farmers became manufacturers. Unfortunately, the entire population cannot be linked, so this quantity must be estimated from the sub-sample of links that our system generates. In order for this estimate to be useful, it is crucial that it have both low bias and have low variance. Low variance can be achieved simply by producing a large enough set of links; we will see in Sect. 5 that this is not a difficult problem. Achieving low bias, however, requires a very thoughtful approach and induces us to make design decisions that are atypical for many machine learning settings.

Bias can occur when the individuals in the recovered links are not representative of the entire population. This in turn occurs when the probability of being linked is influenced by the quantity we are studying. For example, if we use occupation information to produce links, we may disproportionately form links for people who remain in the same occupation, thus biasing our results. To avoid this problem, and to make our links as broadly useful as possible, we endeavour to use as little information as possible to find links. Furthermore, bias can be caused by false negatives (i.e. true links that are omitted by our system) and by false positives (i.e. recovered links that should not be present). If bias is induced by false negatives only, we can view our set of links as a subset of the entire population of true links, and we can reduce bias by using stratified sampling or re-weighting to ensure that among our links, relevant variables (e.g. gender, occupation, age, etc.) have the same distribution as they do in the census overall. Even if we do not make such adjustments, if we have only false negatives, summary statistics based on our links are lower bounds on corresponding population quantities. If we have bias induced by false positives this argument does not necessarily hold; thus we endeavour to produce as few false positives as possible even if we must incur more false negatives. In addition, certain historical questions to be studied revolve around particular people, families or communities. For this kind of research it is especially important to avoid false positives.

3 Data

We use the 1871 and 1881 Canadian censuses, which were transcribed by the Church of Jesus Christ of Latter-Day Saints and cleaned (but not linked; see Sect. 3.1) at the University of Ottawa (1881) and University of Guelph (1871). The 1871 census has 3,466,427 records and the 1881 census has 4,277,807 records. We know of no other classification analysis of historical data on this scale. Our classification is also challenged by a unique combination of (i) imprecise recording and (ii) extensive duplication of attributes. A third challenge is that we restrict linking criteria to characteristics that do not change over time[1] or change in predictable ways (last name, first name, gender, birthplace, age, marital status) in order

[1]Note that misspelling of names and data imprecision still occur.

to be able to analyze attributes such as occupation, location etc. that change over the life course. Last name and first name are strings, gender is binary, age is numerical, birthplace and marital status are categorical. Social science and historical (SSH) research typically seeks to analyze the determinants of the attributes that change. Therefore it is inappropriate to use time-varying attributes to establish links. For example, taking occupation or location as a linking attribute would bias or, in the extreme, restrict links to those who did not change. The rate of successful linkage might increase but at a cost of significant bias to SSH analysis of change versus persistence (Hall and Ruggles 2004; Ruggles 2006). Linkage with time-varying attributes might be less damaging for other research purposes; if so, there is potential to adapt the linking strategy to meet different needs.

To train and evaluate our record linkage system, we use a set of true links that human experts have identified between records in 1871 and records in 1881. We have four sets of true links matched to unique identifiers[2] in the 1871 and 1881 censuses:

1. 8331 family members of 1871 Ontario industrial proprietors (Ontario_Props)
2. 1759 residents of Logan Township, Ontario (Logan)
3. 223 family members of communicants of St. James Presbyterian Church in Toronto, Ontario (St_James)
4. 1403 family members of 300 Quebec City boys who were ten years old in 1871. (Les_Boys)

The 11,716 total records were linked using *family-context matching*, which allows a high degree of certainty (i.e. generates very few false positives) but biases the links toward those who co-habit with family members. Family-context matching is accomplished by searching for an individual whose vital information (name, age, sex, birthplace, marital status) matches in two census databases (e.g. 1871 and 1881), and confirming it is the same individual by: (1) finding at least one other household member (and preferably two or more) with matching vital information and (2) making sure there is no significant contradictory information that makes a link improbable (for example, when one family member matches, but three others do not). Other data on geography, occupation, religion, name prevalence etc., may also be considered, but the primacy is on the matching of family spouse and children.

Although this approach should generate very few (or perhaps no) false links, it produces a set that is not demographically representative. It generates links only for people living in families within a single household; thus single people will not be matched. It also generates relatively fewer links for children who were around the age of fifteen in 1871 due to difficulty in matching children who left home and young women who got married and changed their last names during that timespan. There is therefore a bias toward young children and established adults.

Fortunately, even if our population of true links is not demographically representative, they can still capture issues such as imprecision of information and name duplication that are needed to train the linkage system. Thus our system will take this biased set of links and use it to produce a new set of links that is less biased, more demographically representative, and therefore more scientifically valuable.

[2]These unique identifiers do not exist in the original censuses, but they are created during digitization to keep track of the records.

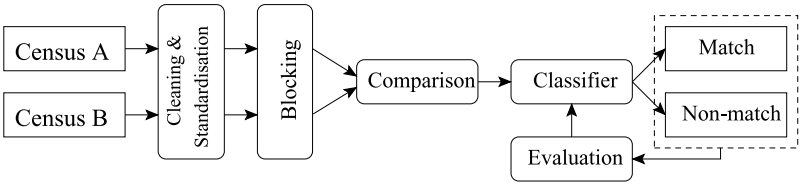


Fig. 1 Overview of record linkage system

3.1 Data cleaning

The first step in any linkage process involves cleaning and standardization of data. This step is needed to effectively compare records from different databases. Each string in 1871 for the sex, age and marital status attributes has been cleaned to match the 1881 database using a standard format across the databases. We removed all non-alphanumerical characters from the strings representing names, as well as all titles (e.g., Rev., Dr.). For all attributes, we cleaned and standardised all the English and French enumerated information (e.g., 5 months, 3 jours, married, marié(e)). We removed duplicate records appearing in 1871, since several census pages had been digitized and entered into the database twice, and we removed the records of people who died in 1870/1871. Originally, the 1871 collection had 3,601,663 records. This was reduced to 3,466,427 records when duplicates and deceased individuals were removed.

As part of the data cleaning process, we also undertook the laborious task of coding all the first names in the census (e.g. Elizabeth, Beth, Liz would be given the same code). 1871 census has 106,759 distinct first names and 1881 census has 152,880 distinct first names. This process was semi-automatic and it was a joint effort between a team of computer scientists and a team of historians. More details about how we use these codes are given in Sect. 4.1.3.

4 The record linkage system

We wish to link records from one data collection $\mathcal{A}$ to another, $\mathcal{B}$. A record a in $\mathcal{A}$ (viz. b in $\mathcal{B}$) consists of all the information pertaining to a particular entity; in our case the entity is a person, and the information includes all answers collected in the census, e.g. first name, last name, date of birth, birth place, and so on. Our goal is to find all pairs $(a, b),\ a \in \mathcal{A},\ b \in \mathcal{B}$ such that a matches b, that is, such that a and b refer to the same entity. In this case we write $a \simeq b$.

The record linkage process has two main steps. First, for each pair, a feature vector $\phi_{(a,b)}$ is constructed that contains information about the similarity between a and b. In the second step, a classifier is used to label the pairs of records as matches or non-matches based on their feature vectors. We learn this classifier from a training set derived from the data described in Sect. 3. An overview of the system is shown in Fig. 1.

Sections 4.1 and 4.2 describe in detail the two main steps of the system.

4.1 Feature construction, blocking, and thresholding

During the feature construction step, the attributes in each pair (a, b) of records are used to compute a set of similarity measures which are used as features. We use the following attributes to generate features that reflect record-pair similarity:

Tag	L	F	GD	AGE	BP	MS
Attr.	Last name	First name	Gender	Age	Birthplace	Marital status
Type	String	String	Binary	Integer	Categorical	Categorical

We will refer to specific attributes using subscripted tags, for example a_{F} represents the first name associated with record a.

In the feature construction step, there are two challenges that we address. First, the similarity measures must be tailored to the different attribute types. We therefore select specialized similarity measures for each attribute. Second, we must avoid explicitly evaluating $\phi_{(a,b)}$ for all possible pairs, as this quickly becomes intractable as the size of $\mathcal{A}$ and $\mathcal{B}$ increases. We accomplish this by *blocking*, described below.

4.1.1 String comparison and processing

To compare names (last and first names) we use two character-based similarity measures (Winkler 2006) that are well-suited to comparing names: *edit distance* and *Jaro-Winkler score*. In addition, we make use of two different phonetic representations of the original string using the *double metaphone* algorithm (Philips 2000).

The *edit distance* between two strings S_1 and S_2, which we denote by $\mathrm{Edit}(S_1, S_2)$, is the minimum number of edit operations (insert, delete and replace) on single characters needed to transform the string S_1 into S_2, divided by $\max(|S_1|, |S_2|)$ where $|.|$ denotes the length of a string.

The *Jaro-Winkler score* is a string similarity measure[3] developed for comparing names in the U.S. census (Winkler 2006). It is based on the Jaro similarity score given by

$$\mathrm{Jaro}(S_1, S_2) = \frac{1}{3}\left(\frac{c}{|S_1|} + \frac{c}{|S_2|} + \frac{c-t}{c}\right)$$

where c is the number of *common* characters and t is the number of *transpositions* of the common characters. A character at position i in S_1 has a *common* character in position j of S_2 if the characters are the same and $|i - j| \leq \lfloor \max(|S_1|, |S_2|)/2 \rfloor$. Let C_1 and C_2 be the subsequences of common characters in S_1, S_2. Then t is the number of transpositions we must apply within C_1 so that $C_1 = C_2$. Note that $0 \leq \mathrm{Jaro}(S_1, S_2) \leq 1$. The Jaro-Winkler score is a modification based on the idea that fewer errors typically occur at the beginning of names. It takes the Jaro score and increases it if there is agreement on initial characters (up to four) so that

$$\mathrm{JW}(S_1, S_2) = \mathrm{Jaro}(S_1, S_2) + 0.1 \cdot \min(s, 4)\big(1 - \mathrm{Jaro}(S_1, S_2)\big)$$

[3]Unfortunately, the term "Jaro-Winkler *distance*" is commonly used to describe this quantity, even though larger values are associated with greater similarity. We use the term "score" throughout when describing features that positively correlate with similarity.

where s is the length of the longest common prefix of S_1 and S_2.

The *double metaphone* algorithm takes a string S and produces two *codes* DM1(S) and DM2(S) for the string. Each of the two codes are themselves strings over a reduced 21-character alphabet, and they are both designed to represent the phonetic pronunciation of S.

4.1.2 Feature construction

Name comparison features We use a total of eight features derived from the first and last names in the records. They are given by

$$\begin{aligned}
\phi_{(a,b)}^{\text{L-ED}} &= \text{Edit}(a_{\text{L}}, b_{\text{L}}) & \phi_{(a,b)}^{\text{F-ED}} &= \text{Edit}(a_{\text{F}}, b_{\text{F}}) \\
\phi_{(a,b)}^{\text{L-JW}} &= \text{JW}(a_{\text{L}}, b_{\text{L}}) & \phi_{(a,b)}^{\text{F-JW}} &= \text{JW}(a_{\text{F}}, b_{\text{F}}) \\
\phi_{(a,b)}^{\text{L-DM1}} &= \text{Edit}\big(\text{DM1}(a_{\text{L}}), \text{DM1}(b_{\text{L}})\big) & \phi_{(a,b)}^{\text{F-DM1}} &= \text{Edit}\big(\text{DM1}(a_{\text{F}}), \text{DM1}(b_{\text{F}})\big) \\
\phi_{(a,b)}^{\text{L-DM2}} &= \text{Edit}\big(\text{DM2}(a_{\text{L}}), \text{DM2}(b_{\text{L}})\big) & \phi_{(a,b)}^{\text{F-DM2}} &= \text{Edit}\big(\text{DM2}(a_{\text{F}}), \text{DM2}(b_{\text{F}})\big).
\end{aligned}$$

Age comparison feature Let a_{AGE} be the age in years from a record in the 1871 census, and b_{AGE} be the age in years from a record in the 1881 census. We construct a binary feature indicating whether the ages match given by

$$\phi_{(a,b)}^{\text{AGE}} = \mathbb{1}\big\{8 \leq |b_{\text{AGE}} - a_{\text{AGE}}| \leq 12\big\} \tag{1}$$

where $\mathbb{1}$ is the indicator function. Since the two censuses are 10 years apart, if in fact $a \simeq b$, we would expect that in most cases $b_{\text{AGE}} - a_{\text{AGE}} = 10$. We allow a 20 % error in the age difference, as census experts consider this window when performing manual linking.

Gender, birthplace, and marital status comparison features For the *gender* and *birthplace code* attributes we perform an exact match comparison, giving two features

$$\phi_{(a,b)}^{\text{GD}} = \mathbb{1}\{a_{\text{GD}} = b_{\text{GD}}\}, \qquad \phi_{(a,b)}^{\text{BP}} = \mathbb{1}\{a_{\text{BP}} = b_{\text{BP}}\}.$$

For the *marital status* attribute, we construct a feature that is 1 if a valid marital status change appears (e.g. single to married) and 0 otherwise.

$$\phi_{(a,b)}^{\text{MS}} = \text{is-valid}(a_{\text{MS}}, b_{\text{MS}}).$$

Feature vector Our feature vector for a pair of records (a, b) is given by

$$\begin{aligned}
\phi_{(a,b)} = \big(&\phi_{(a,b)}^{\text{L-ED}}, \phi_{(a,b)}^{\text{F-ED}}, \phi_{(a,b)}^{\text{L-JW}}, \phi_{(a,b)}^{\text{F-JW}}, \\
&\phi_{(a,b)}^{\text{L-DM1}}, \phi_{(a,b)}^{\text{F-DM1}}, \phi_{(a,b)}^{\text{L-DM2}}, \phi_{(a,b)}^{\text{F-DM2}}, \phi_{(a,b)}^{\text{GD}}, \phi_{(a,b)}^{\text{BP}}, \phi_{(a,b)}^{\text{MS}}\big).
\end{aligned}$$

4.1.3 Blocking and thresholding

The most straightforward way to approach the record linkage problem is to apply a classifier to all possible pairs of records $(a, b) \in \mathcal{A} \times \mathcal{B}$, that is, the entire Cartesian product of the two sets of records. There are two problems with this approach.

First, there are certain rules that experts use when matching that should eliminate certain record pairs as candidates for a match. While these rules eliminate some pairs that are true

matches, this is viewed as an acceptable cost because the quality of SSH analyses is degraded much more by false positives than by false negatives, as we discussed in Sect. 2.

Second, computing feature vectors for all possible pairs is impractical as there would be $3,446,427 \times 4,277,807 \approx 14.8 \cdot 10^{12}$ feature vector computations. Our system is written in C to be efficient in the calculation of similarity between census records. Benchmarking indicates that our system calculates string comparisons at a rate of approximately 4 million per second. Although at first glance this throughput might seem sufficiently fast, it is actually not fast enough to run on a single machine for our application in a reasonable time. Assume for the moment that we would run our record linkage system on a single processor. Computing similarity between all $14.8 \cdot 10^{12}$ pairs would give us a run-time estimate of close to a CPU-year: $(14.8 \cdot 10^{12}$ pairs $\times$ 8 string-based features$)/(4 \cdot 10^6$ comparisons/s$)/(86400$ s/day$) =$ 342.6 days. This does not include the cost of classifying each pair.

To mitigate these two problems, we use *blocking* and *thresholding* to reduce the number of candidate pairs. Blocking is the process of dividing the databases into a set of mutually exclusive blocks under the assumption that no matches occur across different blocks. Thresholding allows us to abort the computation of a feature vector if, based on a subset of the features, it appears no match will result.

In our system, we block by the first name code (recall that "Beth" and "Liz" would be within the same block, for example) and within that block we block again by the first letter of the last name. Experts have empirically noted that fewer mistakes are found in the beginning of a name, thus by choosing to block on the first letter only, we reduce the probability of eliminating a true match. Based on this blocking, "Eliza Jones" and "Beth Jonze" are a candidate match, but "Eliza Jones" and "Eliza Phair" are not. Thus, women who change their last name between 1871 and 1881 are not matched by our system. This source of false negatives is also present in our hand-labeled data, and is extremely difficult to correct without inducing false positives given the data we have. Social scientists who study this group are well aware of this problem. Many analyses, including the one in Sect. 7, are unaffected by it and where it is an issue, statistical social science techniques to treat selection bias are used.

Note that we block by the name code, but when we perform the similarity calculations we do so on the original string. This allows us to better link persons who were consistent in reporting their name in a certain way (e.g. someone named Beth is part of the Elizabeth block, but will be more similar to those named Beth than Eliza). After name blocking, we require that records in a candidate pair must have the same birthplace, an attribute known to have few errors.

Within blocks, we apply thresholds on the similarity of last name: For a pair (a, b) to be a candidate, it must satisfy

$$\phi_{(a,b)}^{\text{L-ED}} < 0.15, \qquad \phi_{(a,b)}^{\text{L-JW}} > 0.85, \qquad \phi_{(a,b)}^{\text{L-DM1}} < 0.15, \qquad \phi_{(a,b)}^{\text{L-DM2}} < 0.15.$$

By applying these thresholds, we further eliminate dissimilar pairs that are unlikely to be linked by the classifier. These thresholds were selected based on expert evaluation of the last-name similarities we observed on our training data.

4.2 Pair classification

Now that we have defined our feature vectors, we can cast our matching problem as a binary classification problem. We construct a training set based on the true matches described in Sect. 3, and we learn a Support Vector Machine (SVM) with a Radial Basis Function (RBF)

kernel. We use LIBSVM (Chang and Lin 2001) as the classifier implementation, and we make use of the LIBSVM facility for producing class probability estimates based on work by Wu et al. (2004). The probability estimate scores allow us to see how confident the system is in each prediction, and they can be used to select the most confident matches. These estimates are used for manual verification of links; we discuss this in Sect. 6.

4.2.1 Training set and class imbalance

Our training set is based on the 11,716 true links described in Sect. 3. These pairs of records represent the *match* class. To create examples for the *non-match* class, we generate all of the $11{,}716 \cdot (11{,}716 - 1) \approx 1.4 \cdot 10^8$ incorrect pairs of records. To produce our training set, we apply our similarity thresholds to the total $11{,}716^2$ pairs, resulting in a training set of size 81,281, with 8,543 matches (positive class) and 72,738 non-matches (negative class). Note that the number of matches has considerably decreased when the similarity thresholds are applied. This shows the imprecision of the data and that dissimilar records could in fact be matches. However, when building the training set, we consider it better to build our classification model from pairs of records that are less likely to produce errors.

In many applications, it is important to "correct" class imbalance by one of several mechanisms, e.g. over-sampling, under-sampling, sample re-weighting, etc. This is most commonly done because class imbalance can cause learning machines to place much more emphasis on false negative rate than false positive rate, or vice versa. As we discussed in Sect. 2, in our application, false positives are much more damaging than false negatives, so the ambient class balance of our training set with its abundance of negative examples biases our classifier in a desirable way—it emphasizes getting the negative examples right. We therefore do not try to achieve class balance in the training set, and we will show in Sect. 5 that the resulting classifier has the properties we want.

4.2.2 Classification and linking

Once we have learned our classifier, in order to produce links we take a record a from 1871, we find all records in 1881 that fall within the same block, compute the feature vector from each pair while removing vectors that do not meet our thresholds. We then classify each pair. If all pairs are negative, we produce no link for record a. If exactly one pair (a, b) is labeled positive for a record b in 1881, *and* if there is no other 1871 record c for which (c, b) is labeled positive, then we produce the link (a, b). For any other result, we view the output as ambiguous, and we produce no link for record a. This linking rule, like many of our other design choices, aims to minimize the chance of generating false positive links. We examine other potential rules in Sect. 5.

5 Empirical evaluation

This section evaluates the linkage system we propose and shows the results for linking the Canadian census of 1871 to the Canadian census of 1881. We begin with a standard evaluation of our SVM-based classifier in terms of cross-validation estimates of relevant error rates. We illustrate that we can produce a classifier that has the properties we require: our system has an adequate true positive rate and a very low false positive rate. We then describe the challenges associated with the application of our system to the full censuses, and we discuss the bias present in our links, which we can measure using the full, unlabeled data sets.

Table 1 Classification system evaluation—5 fold cross validation—mean (std. dev.)

	Positives	Negatives	TP	FP	FN	TN	AUC
Mean	1708.6	16256.2	1427.2	70.2	281.4	14477.4	0.9662
Std. Dev.	45.09	0.45	30.46	8.23	19.96	43.71	0.0004

Table 2 Types of candidate links generated by the system

Type	Number	Percentage
One to One	596,284	24.22 %
One to Many	831,145	33.76 %
Many to One	240,482	9.77 %
No Link	793,501	32.23 %

5.1 Classification system evaluation

We perform 5 fold cross validation on the training data to evaluate the proposed classification system. We report the true positives, false positives, false negatives, true negatives and the area under the ROC curve. Averages and standard deviation over the 5 folds are presented in Table 1.

We can see that our classifier achieves a very low number of false positives, and a reasonably low number of false negatives. It therefore meets the criteria we set out in Sect. 2. However, this evaluation does not illustrate the biases incurred when we apply the system to link the full censuses. This is discussed in detail in the next section.

5.2 Full Canadian census linkage results

As we discussed in Sect. 4.2.2, not every pair labeled "positive" by our classifier becomes a link. In effect, we end up with three types of potential links after pair classification. The number and type of potential links generated by the classifier are shown in Table 2. We consider a link successful (a match) if the classification system found only a one-to-one link between a person in 1871 and a person in 1881. One-to-many (a record in 1871 is linked to two or more records in 1881) and many-to-one links (several records in 1871 are linked to the same record in 1881) are removed. We consider these links ambiguous; thus we do not consider them for evaluation and we do not present them to the user.

The 'no link' proportion of 32.23 % is consistent with expectations. We know from other sources that roughly 10 % of the population died between 1871 and 1881 (Bourbeau et al. 1997); another 10 % emigrated largely to the United States (Emery et al. 2007); a majority of young single women changed their surname after marriage; some people were missed in the enumeration and others inadvertently or deliberately misreported their characteristics in one census year or the other. None of these records can be confidently linked using the data we have available. Table 2 also indicates that roughly 45 % of the links were many-to-one or one-to-many. Again, this is not surprising because of considerable duplication of names, the limited number of fields with which to link and, equally important, the imprecision with which name and age were reported (Goeken et al. 2011). We cannot use these ambiguous links for social science analysis. We interpret a 'one-to-one' link, a single 1871 record connected to a single 1881 record, as providing information about the same person at different points in his or her life. This group accounts for 24.22 % of all links. The number of links, nearly 600,000, is sufficient to support a wide range of social science and historical studies.

Table 3 Full linkage system evaluation estimates—5 fold cross validation—mean (std. dev.)

True Links	TP	FP	FN	TPR	FPR
1,708.6 (45.1)	684.8 (38.4)	36.0 (9.6)	1,023.8 (24.1)	40.1 % (1.5 %)	5.0 % (1.3 %)

5.2.1 False positives and bias

In this section we present and discuss evaluation of the true links in the context of linking the full census data. Note that in our problem, we cannot evaluate all the generated links because we do not know their correct class. We perform this evaluation on the positive examples in the 5 folds used in Sect. 5.1. This evaluation is different from the one done in the previous section due to considering all the pairs of records classified. Under these circumstances, some of the people may have been linked to multiple other persons and vice versa. Such cases would not be presented to the user due to their ambiguity; thus they are not part of this evaluation. We consider only the one-to-one links for evaluation.

For evaluation, we calculate the following: true positives (TP): pairs of records that have been labelled as a match by both the classification system and the human expert; false positives (FP): pairs of records that have been labelled as a match by the classification system, but have not been labelled as a match by the human expert; false negatives (FN): pairs of records that have been labelled as a non-match by the classification system but have been labelled as a match by the human expert.

We are interested only in the positive examples (matches), thus the evaluation for our application is slightly different than a standard classification evaluation. The calculation of true positives is straightforward: a pair of records in our testing set that is also found in the matches produced by the classifier is a true positive. To calculate the false positives we search for records in our testing sets that were incorrectly linked by the classifier (e.g. (a, b) is a pair labelled as a match by the expert, we find (a, c) as a pair labelled by the classifier as a match; given that we know that the correct link would have been (a, b), we can conclude that (a, c) is a false positive). We count as a false negative all the pairs from the testing set that were not found. Note that for this particular application, we are most interested in finding high quality links that would allow us to build reliable longitudinal databases; thus the true positive and false positive values are key to our evaluation. For this reason we calculate how many of the true links were recovered (true positive rate) by the system as well as how many of the generated links were false. The true and false positive rates on one-to-one links are defined in (2) and (3), respectively. Table 3 presents the evaluation for our testing sets based on these measures.

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{TP} + \mathrm{FP}}. \tag{3}$$

One should note that it is very difficult to recover all true links with the limited number of attributes we use for linking, and that when links are manually created by experts, they use more information such as family context and location. Table 4 shows the distribution of the attribute values for the created links in comparison with the distribution of records in 1871. We see that while many of the proportions match well, we are under-linking females, persons between 15 and 25 years of age, and single persons. This can be attributed in part

Table 4 Attribute distribution

Attribute	1871	Links
Female	49.35 %	44.47 %
Male	50.61 %	55.53 %
0–15 years	41.61 %	41.64 %
15–25 years	20.39 %	15.85 %
25–50 years	26.40 %	30.71 %
50+ years	11.60 %	11.80 %
Married	30.75 %	37.67 %
Widowed	3.26 %	2.44 %
Single	66.00 %	59.88 %
Birthplace	**1871**	**Links**
Ontario	32.68 %	32.90 %
Quebec	28.74 %	28.00 %
England	4.21 %	5.96 %
Scotland	3.54 %	3.54 %
Ireland	6.39 %	5.57 %
Germany	0.65 %	0.71 %
USA	1.83 %	1.89 %

Table 5 Distribution of false negatives

Multiples	Blocking	Classifier
66.14 %	6.36 %	27.48 %

to the difficulty of linking females who marry and change their last name—there is often no way of being sure that a married woman in 1881 should link to the record of a single woman in 1871. It is very important to minimize these biases and to ensure that end users are aware of them so that they can decide if the data are useful, and what correction methods, if any, they will want to use for their analyses.

In addition, we are interested to explore why we have such a large number of false negatives. There are three categories that generate false negatives: pairs of records missed due to the blocking technique, records being part of one to many and many to one links, and false negatives generated by the classifier. Table 5 shows the distribution of the false negatives in these categories. It can be observed that most false negatives (66.14 %) are coming from the one-to-many and many-to-one links. The cases where the classifier incorrectly classifies the true links represent a considerably smaller percentage of the total number of false negatives.

Our team of historians is able to verify about 20 links per hour. To make a complete analysis of all the generated links (596,284) would require close to 30,000 hours of manual verification. This shows the unfeasibility of manually checking all the produced links and it also shows how costly and difficult it is to create even training and evaluation data.

The data generated with the system presented in this paper is available from http://hdru.ca/.

6 Implications for machine learning

In our pursuit of a useful set of social science data, the most important lesson we have learned is that in this setting, standard performance measures for classifiers in ML—even more "comprehensive" ones like area under the ROC curve—are not sufficiently descriptive measures of the quality of the data we produce. To convince ourselves and our collaborators of the quality of our results, we investigated how the confidence asserted by our system corresponded with human confidence in the links produced, and we took time to understand biases in the data by examining the attribute distributions of different subsets of links. These investigations facilitated a dialogue between the ML practitioners and social scientists in our group, and we anticipate that our approach will be useful in other areas where machine learning methods are used to produce "new data" for applied fields. Here we briefly summarize our findings.

High-confidence versus low-confidence links As we mentioned, we use an SVM that produces a confidence in its classification; these were examined in two different ways. First, these confidences were used to see how well the classifier matched what the human labellers were doing. We pulled the most-confident links and, upon discussion with our labellers, we found that they did indeed appear most "obvious" to a human. This was an important sanity check, and we recommend that practitioners use this approach to facilitate discussions of system performance and reliability with subject-area collaborators. We also investigated whether we could reduce the false-positive rate by carefully selecting a threshold confidence for links. We found that the distribution of confidences among the TP and FP links was similar; thus we do not believe the current system could be improved by using a carefully-selected confidence threshold for distinguishing positives from negatives. This was in line with our expectations given the limited amount of personal characteristics used in the linking process.

TP, TN, FP, FN links We examined the attribute similarity distributions of these different categories of links in the training/validation data to investigate whether there were obvious biases, for example, whether certain types of links were much easier for our system to recover. We did not find any such biases.

Discarded many-to-one and one-to-many links For the current application, we discard all the one-to-many and many-to-one links. This is due to the fact that we can not disambiguate them given the information we use for linking. One approach to disambiguate some of these links would be to consider the classifier probabilities distribution and to find a threshold that would resolve some of these links. We have investigated this avenue and we were unable to find a good threshold because the resulting one-to-one links introduce more false positives which is unacceptable for our application. Figure 2 shows the distribution of the classifier probabilities for those links that belong to one-to-many and many-to-one groups. It can be observed that the distribution is very skewed with more than 80 % of the links having the same probability score. This is expected since many records share very similar personal characteristics. This is especially true for people with common names.

7 Impact to historical census linkage

The classification system identifies a large number of people, each of whom is observed in 1871 and again in 1881. We have used the linked data, generated with the system described in this paper, to resolve a long-standing puzzle in the historical literature. The later

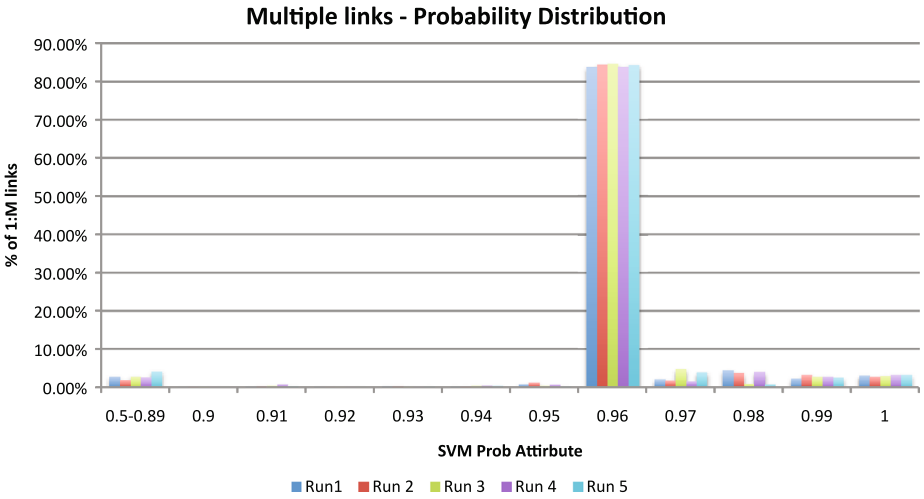


Fig. 2 Classification system probability score distribution

nineteenth century was a period of rapid social and economic change in the North Atlantic world. Numerous community and institutional case studies, extensive bankruptcies and reconfiguration of companies, and qualitative evidence of personal anxieties indicate that economic change in this period was rapid and disruptive (Chambers 1964; Drummond 1987; Gagan 1982; Inwood and Keay 2012; Kealey 1980). And yet the standard aggregate indicators, GNP and workforce composition, show little or no change (Urquhart 1986; Green and Urquhart 1987). In another paper we reconcile the conflicting micro and macro evidence using longitudinal data created with the linkage system described in this paper (Antonie et al. 2014). In that paper, we analyze the work transitions for large numbers of individuals in order to demonstrate that many people changed jobs, but that the changes partially offset each other and are thus hidden if we examine only the unlinked data. This fact, which is not visible in aggregate data but can be seen in the linked data (see Table 6), is one step toward a reconciliation of micro and macro evidence. The linked records allow us to determine, for the first time, how individuals moved between different occupations.

Canada at this time had a largely agricultural economy. Farming was still the largest source of employment; the availability of inexpensive farmland continued to attract European immigrants. But the decisions of young people to leave and enter particular sectors would determine the future shape of the economy. Already in the 1870s significant numbers of young people were beginning to leave farming. Based on the occupational distribution in 1871 (47 % in farming) we can calculate that 12.6 % of the entire young working population left farming as opposed to 11.4 % who entered. Other sectors experienced a net gain; for example 1 % exited and 5 % entered commerce while 6 % left and 8 % entered industry.

Individual-level linked data reveal the complexity of job changing even at this high level of aggregation that reduces a myriad of jobs to five broadly-defined sectors. The linked data also demonstrate that the patterns of job change were different among younger and older people (Table 7). During the decade a higher proportion of the 15–25 year olds changed sectors (41 % against 27 % of the 26–55 year olds). The older group showed a net movement out of industry (0.7 %) and into farming (1.9 %), in contrast to the younger group which had a net flow out of farming (1.3 %) and into industry (2.2 %). Moreover the younger group shifted more decisively into commerce (3.0 % against 1.2 % among older workers).

Table 6 Individual occupational transitions, by sector

Occupations 1871	Occupations 1881				
	15–25 year olds in 1871				
	Farming (46 %)	Industry (16 %)	Commerce (6 %)	Labour (18 %)	Other services (14 %)
Farming (47 %)	**74 %**	7 %	3 %	12 %	5 %
Industry (14 %)	15 %	**57 %**	5 %	12 %	11 %
Commerce (2 %)	9 %	14 %	**54 %**	12 %	12 %
Labour (20 %)	31 %	15 %	4 %	**40 %**	10 %
Other services (17 %)	17 %	10 %	12 %	15 %	**46 %**
	26–55 year olds in 1871				
	Farming (54 %)	Industry (13 %)	Commerce (5 %)	Labour (16 %)	Other services (13 %)
Farming (52 %)	**86 %**	3 %	2 %	6 %	3 %
Industry (13 %)	18 %	**61 %**	5 %	9 %	7 %
Commerce (5 %)	15 %	11 %	**50 %**	9 %	2 %
Labour (16 %)	24 %	9 %	3 %	**56 %**	8 %
Other services (14 %)	16 %	6 %	6 %	12 %	**60 %**

Table 7 Net flow of workers, by age and sector

Occupation	15–25 year olds in 1871		26–55 year olds in 1871	
	Out of	Into	Out of	Into
Farming	12.69 %	11.37 %	7.28 %	9.17 %
Industry	6.02 %	8.27 %	5.07 %	4.39 %
Commerce	0.94 %	4.95 %	1.85 %	3.01 %
Labour	12 %	10.11 %	7.04 %	6.42 %
Other services	9.18 %	6.13 %	5.6 %	3.85 %
Total	40.83 %	40.83 %	26.84 %	26.84 %

The generational differences are not large but they identify a slow but powerful historical movement that eventually, in the long-run, would fundamentally change the character of economic activity. The net loss of young people from agriculture is especially notable because it signals a fading of the appeal of a sector that once had been the most desirable in the entire economy.[4]

There has been some uncertainty about how to interpret change in the agriculture sector, the single largest economic area, at this time. Regional and community micro-studies have pointed to "a genuine crisis" in 1860s agriculture, especially in Ontario, and with it substantial economic instability and social mobility. Farming remained the preferred alternative

[4]Two other sectors, labour/construction and other services, also experienced a net loss of young people. Many young men began their working lives in these sectors and then, after gaining experience, moved into farming, industry or commerce. We do not dwell on this movement because it reflects a familiar life-cycle process rather than structural change in the economy.

choice for all occupation groups (suggesting it was a default occupation), although individual trajectories provide evidence of a decline in appeal for the young. And when the linked data are viewed in combination with cohort data in 1881 for the youngest and oldest males, we can anticipate the longer-term shift out of agricultural occupations that took place over ensuing decades.

Of course, the beginnings of a shift out of agriculture and into industry and commerce is unsurprising to the extent that a similar process of macro change had been visible in Europe for several decades. A familiar label for this important process is industrialization. The most important contribution of the linked individual-level data is to reveal the beginnings of industrial transformation even in a classic primary product exporting economy such as Canada.

These arguments have been presented at conferences in London, Chicago, Toronto and Victoria and are now forthcoming in a book from a prestigious university press (Baskerville and Inwood 2014).

Another paper in the same volume uses our longitudinal data to improve our understanding of rural adjustment to economic stress (Baskerville 2014). Our collaborator Peter Baskerville demonstrates that previous estimates of rural residential persistence were seriously flawed because in the absence of machine learning techniques the research was based on linking records within the local area only. The linkage system provides much more accurate data used by Baskerville to analyze who moved and who stayed. He finds surprising differences by ethnicity; farmers of German origin were much less likely to move. Another paper in the same volume by Gordon Darroch uses a smaller set of census data from different years (Canada 1861 and 1871), linked with a semi-automatic method to analyze the choices made by young men as they first entered the labour market (Darroch 2014). Two other papers in the volume use data linked deterministically and on a smaller scale between World War One enlistment records and the 1901 census. One of these papers exploits linked data to show that early life family circumstance was an important influence on adult health (Cranfield and Inwood 2014) and that child socio-economic circumstance explains only a small part of the difference between French and English Canadians. The other paper identifies Canadian soldiers of aboriginal origin and analyzes the different patterns of education, occupation and language for pure-blood and mixed race Indians (Fryxell et al. 2014). None of these important research findings would have been possible without methodology for linking historical records.

The importance of machine learning applications to historical data is reflected in broad international participation in a series of annual workshops on the topic at the University of Guelph since 2007. Machine learning principles provide the basis for a prestigious 'Digging in Data' award (http://www.diggingintodata.org) in which the People in Motion classification system is being used. The People in Motion project has attracted the attention of the Ontario Genealogical Society, which recently opened a collaboration with the University of Guelph. Another indicator of impact is the use of our linked historical data by seven graduate students to date as part of their degrees (in History, Economics, Demography and Computing Science) at four Canadian universities and at Cambridge.

Longitudinal data derived from the application of machine learning to historical data comprise key data infrastructure for the next generation of historical and social scientific research. The broader public impact will be felt after specialized domain research find its way into textbooks and is synthesized in meta-review publications read by policy-makers. The knowledge of occupational change in the 19th century, for example, will provide long-term context and perspective for modern analysis of labour market mobility. Five years from first journal publication is a plausible timescale for this distribution of knowledge.

8 Conclusions

In this paper we presented and discussed the implementation of a record linkage system for historical census data. The goal of the system is to produce longitudinal data tracking people in 19th century Canada. We described how, for this application, we must pay careful attention to the false positive rate of our system and to demographic biases that may be introduced by our classifier. In our experimental study, our cross-validation analysis showed that our system produces very few false positives. At the same time, it is capable of successfully linking nearly 600,000 records that are, for the most part, demographically representative. Because the discrepancies in demographics between the links and the full census are relatively small, stratified sampling or re-weighting can be used to correct the difference prior to analysis. We have therefore created high-quality longitudinal data that will be used to investigate important historical trends.

Future directions of this research include incorporating more census collections for building longitudinal data over multiple decades. In this case, we will want to recover n-tuples that represent an individual over the course of n censuses; this will make the computational challenges even greater. We are also planning to include United States and British census data to be able to track those Canadians who emigrated and immigrated in that time frame. The challenges associated with bringing in other census collections will present themselves both at the data cleaning phase and the feature construction phase—the census was conducted differently in different countries, thus making the data more difficult to compare.

Acknowledgements The authors are grateful for financial support from the Canadian Foundation for Innovation, Ontario Ministry of Research and Innovation, Social Sciences and Humanities Research Council, Google and the University of Guelph. We would also like to thank our genealogical collaborators, the Ontario Genealogical Society, Ontario GenWeb and Family Search. Detailed comments from the reviewers and editors of this journal have helped us substantially to improve our work.

References

Antonie, L., Baskerville, P., Inwood, K., & Ross, J. A. (2014, forthcoming). Change amid continuity in Canadian work patterns during the 1870s. In *Lives in transition: longitudinal perspectives from historical sources*.

Baskerville, P.: (2014, forthcoming). Wilson Benson revisited: movement and persistence in rural Perth County, Ontario, 1871–1881. In *Lives in transition: longitudinal perspectives from historical sources*.

Baskerville, P. & Inwood, K. (Eds.) (2014, forthcoming). *Lives in transition: longitudinal perspectives from historical sources*. Kingston and Montreal: McGill-Queen's University Press.

Bilgic, M., Licamele, L., Getoor, L., & Shneiderman, B. (2006). D-dupe: an interactive tool for entity resolution in social networks. In *Visual analytics science and technology (VAST)*. Baltimore.

Bourbeau, R., Légaré, J., & Édmond, V. (1997). *New birth cohort life tables for Canada and Quebec, 1801–1991*.

Chambers, E. J. (1964). Late nineteenth century business cycles in Canada. *Canadian Journal of Economics and Political Science*, *3*, 391–412.

Chang, C. C., & Lin, C. J. (2001). Libsvm: a library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Christen, P. (2008). Automatic record linkage using seeded nearest neighbour and support vector machine classification. In *Proceeding of the 14th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '08* (pp. 151–159).

Cranfield, J., & Inwood, K. (2014, forthcoming). Genes, class or culture? French–English height differences in Canada. In *Lives in transition: longitudinal perspectives from historical sources*.

Darroch, G. (2014, forthcoming). Lives in motion: revisiting the 'agricultural ladder' in 1860s Ontario, a study of linked microdata. In *Lives in transition: longitudinal perspectives from historical sources*.

Drummond, I. (1987). *Progress without planning: the economic history of Ontario from confederation to the Second World War*. Toronto: University of Toronto Press.

Elfeky, M. G., Elmagarmid, A. K., & Verykios, V. S. (2002). Tailor: a record linkage tool box. In *Proceedings of the 18th international conference on data engineering, ICDE '02* (pp. 17–28).

Elmagarmid, A. K., Ipeirotis, P. G., & Verykios, V. S. (2007). Duplicate record detection: a survey. *IEEE Transactions on Knowledge and Data Engineering*, *19*, 1–16.

Emery, J., Inwood, K., & Thille, H. (2007). Hecksher–Ohlin in Canada: new estimates of regional wages and land price. *Australian Economic History Review*, *47*(1), 22–48.

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, *64*, 1183–1210.

Fryxell, A., Inwood, K., & van Tassel, A. (2014, forthcoming). Aboriginal and mixed race men in the Canadian expeditionary force 1914–1918. In *Lives in transition: longitudinal perspectives from historical sources*.

Gagan, D. (1982). *Hopeful travellers families, land, and social change in Mid-Victorian Peel County, Canada West*. Toronto: University of Toronto Press.

Goeken, R., Huynh, L., Lenius, T., & Vick, R. (2011). New methods of census record linking. *Historical Methods*, *44*(1), 7–14.

Green, A., & Urquhart, M. (1987). New estimates of output growth in Canada: measurement and interpretation. In *Perspectives on Canadian economic history* (pp. 182–199).

Hall, P. K., & Ruggles, S. (2004). Restless in the midst of their prosperity: new evidence of the internal migration patterns of Americans, 1850–1990. *Journal of American History*, *91*, 829–846.

Inwood, K., & Keay, I. (2012). Diverse paths to industrial development: evidence from late nineteenth century Canada. *European Review of Economic History*, *16*, 311–333.

Kang, H., Getoor, L., Shneiderman, B., Bilgic, M., & Licamele, L. (2008). Interactive entity resolution in relational data: a visual analytic tool and its evaluation. *IEEE Transactions on Visualization and Computer Graphics*, *14*(5), 999–1014.

Kealey, G. (1980). *Toronto workers respond to industrial capitalism* (pp. 1867–1892). Toronto: University of Toronto Press.

Newcombe, H. B. (1988). *Handbook of record linkage: methods for health and statistical studies, administration, and business*. New York: Oxford University Press

Newcombe, H., Kennedy, J., Axford, S., & James, A. (1959). Automatic linkage of vital records. *Science*, *130*, 954–959.

Philips, L. (2000). The double metaphone search algorithm. *C/C++ Users Journal*.

Rahm, E., & Do, H. H. (2000). Data cleaning: problems and current approaches. *IEEE Data Engineering Bulletin*, *23*, 2000.

Ruggles, S. (2006). Linking historical censuses: a new approach. *History and Computing*, *14*, 213–224.

Urquhart, M. C. (1986). New estimates of gross national product, Canada, 1870–1926: some implications for Canadian development. In *Long term factors in American economic growth* (pp. 9–94). Chicago: University of Chicago Press.

Winkler, W. E. (2006). *Overview of record linkage and current research directions*. Statistical Research Division Report.

Wu, T. F., Lin, C. J., & Weng, R. C. (2004). Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, *5*, 975–1005.

Chapter 11
Dancing with Dirty Data: Problems in the Extraction of Life-Course Evidence from Historical Censuses

Luiza Antonie, Kris Inwood and J. Andrew Ross

Abstract This chapter builds on a recent use of SVM classification to create linked sets of Canadian 1871 and 1881 census records. The census data are imprecise and have limited granularity; many records share identical detail. In spite of these challenges, the SVM generates life-course information for large numbers of individuals with a low (3 %) false positive error rate. However, there is a higher incidence of error among apparent migrants when the true rate of migration is small. The linked data are broadly representative of the population with some underrepresentation of illiterates, young adults (especially unmarried women), older people (especially men), and married people of French origin. The new longitudinal data are of considerable research value but users must take into account these weaknesses.

11.1 Introduction

In this chapter, we explore strengths and weaknesses of a recent application of support vector machine (SVM) classification to Canadian historical census records. The classification identifies matched pairs of records from the 1871 and 1881 census. Each matched pair describes the same person and thus provides insight into the change in individual circumstances from one year to the next.

L. Antonie (✉)
School of Computer Science and Department of Economics,
University of Guelph, Guelph, ON, Canada
e-mail: luiza.antonie@gmail.com

K. Inwood · J.A. Ross
Department of History and Department of Economics,
University of Guelph, Guelph, ON, Canada
e-mail: kinwood@uoguelph.ca

J.A. Ross
e-mail: jaross@uoguelph.ca

G. Bloothooft et al. (eds.), *Population Reconstruction*,
DOI 10.1007/978-3-319-19884-2_11

The considerable importance of the North American census for historical research derives from its rich systematic detail and a paucity of alternate sources providing a comprehensive description of the population. Nowhere in North America was there an established church with a commitment to public vital registration. There was not even a dominant church whose records might serve that purpose, except perhaps in Utah or Quebec, and over time even their records became less comprehensive. Individual states and provinces gradually developed effective systems of vital registration but in both the United States and Canada the consistent national registration of births, marriages, and deaths emerged only in the twentieth century. Thus, it is in the absence of other sources that Canadian and American scholars turn to the nineteenth century censuses for population profiles and for the construction of longitudinal data that tracks individuals from census to census.

Since the 1980s, there have been significant advances in the method of linking records between censuses. A first wave of studies using manual techniques (Steckel 1988; Knights 1991; Ferrie 1996, 1999) has been followed by the use of machine-learning methodology (Ruggles 2006; Christen 2008; Goeken et al. 2011; Fu et al. 2014). The new approach is capable of generating in a near-automatic way large representative samples of longitudinal and even multigenerational data. The value of the new methodology for understanding historical populations makes it important to assess its strengths and weaknesses (Wisselgren et al. 2014).

In this chapter, we take the example of a recent application of SVM classification to historical Canadian historical census records (Antonie et al. 2013). A principal challenge for any attempt to track individuals from census to census is the relative imprecision of the 1871 and 1881 data. This requires careful adaptation of the classification methodology and some assessment of the quality of linking. We find that the linked data that we generate are reasonably representative of the population although care is needed, depending on research application, because some groups are harder to link: adolescents and young adults especially unmarried women, older people especially men, and married people of French origin.

An additional complication is that, while the overall error (i.e., false positive) rate is only 3 %, the records describing people who apparently migrate, or change categories in some other way, are difficult to interpret if the proportion migrating is small. These idiosyncrasies recommend some care in the research use of these valuable data.

11.2 Overview 1871–1881

We begin with an overview of the record-linking system described in greater detail elsewhere (Antonie et al. 2013). Our objective is to identify pairs of records that describe the same person in two different bodies of data: the 3.4 million records of the 1871 Canadian census (www.census1871.ca) and 4.3 million records of the 1881 census. The Church of Jesus Christ of Latter-day Saints created both databases;

the latter is housed at the Université de Montréal (www.genealogie.umontreal.ca/en). We construct records that follow individuals over time by comparing every 1871 record with each 1881 record, and then classify each comparison as a match or non-match. If we think a particular pair of records (one from 1871 and one from 1881) point to the same person, we accept them as a match.

The process requires us to compare, literally, millions of records in 1871 with millions of records in 1881 in order to establish which pairs are identical, i.e., describe the same person. The comparison is made using four personal attributes that should not change over time (last name, first name, gender, and birthplace) and two others that change in a predictable way (age and marital status). We do not use information about occupation, location, and household composition in order to avoid any bias to people who persist in the same area, in the same job or in the same family. The decision not to link with these characteristics reflects the sensitivity of hypothesis testing in history and the social science to bias (Ruggles 2006).[1]

The process has two computationally demanding steps. The first is to calculate how similar each 1871 record is to each 1881 record on each of the six characteristics. Then the system classifies each possible pairs of records as a match or non-match based on a score for their overall similarity. The classification is accomplished with a methodology, the SVM, used in a number of other classifications of historical census data (Christen 2008; Goeken et al. 2011; Richards et al. 2014). The classification software "learns" from a number of matches already confirmed as reliable on a case-by-case basis by expert genealogists. Without these "training data" the software would be unable to learn how to classify new pairs of records. We also use the individually prepared matches, or "true links", to assess accuracy.

An overview of the system is shown in Fig. 11.1. There are three main steps in the record linkage process. Step one consists of partitioning each census into smaller blocks to reduce the number of record pairs produced between the two censuses. Step two consists of comparing the records in each record-pair and creating a feature vector that contains information about how similar the records in the record-pair are to each other. In step three, the constructed record-pair feature vectors are labeled as matches or non-matches using a classifier that has learned from a training set constructed from both the 1871 and 1881 Canadian census data sets. During the comparison step, feature vectors are constructed for each record-pair (a, b) by comparing how similar the records attributes are to each other using various similarity measures. During the classification step, each feature vector is labeled as a match or non-match. The classification algorithm used in the classification step is a SVM classifier (Vapnik 1995). The SVM is trained on a labeled set of record-pair feature vectors constructed from the true links.

[1]The convention to understand and if possible avoid selection bias is part of the motivation for this paper. The consensus among social scientists on this point is sufficiently broad to recommend some reduction in linking accuracy in order to minimize bias, providing we also achieve a sufficiently small false positive link rate and a size of linked sample sufficiently large for hypothesis testing.

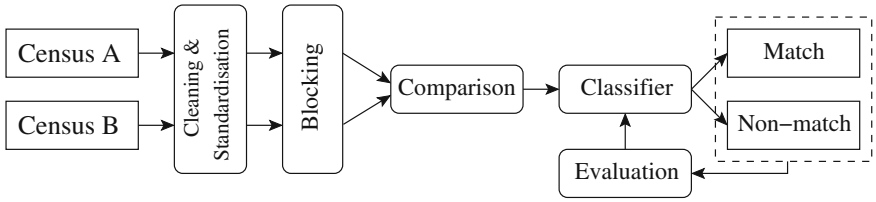


Fig. 11.1 Record linkage system

The record-pair feature vectors that are produced in the comparison step are given to the trained SVM classifier, from which they are labeled as positive or negative links. If the label for a feature vector is negative, the record-pair is seen as a non-match. If the label for a feature vector is positive, the record-pair is only seen as a match if each record is not found in another positive record-pair.

Four sets of true links are available to us: 8331 members of Ontario industrial proprietor families, 1759 residents of Logan Township, Ontario; 223 family members at St. James Presbyterian Church in Toronto and 1403 families of 300 Quebec City boys who were 10 years old in 1871.[2] The pairs of 1871 and 1881 records were established with additional information where available (e.g., church records in Toronto and Quebec City) although the chief criterion in all four collections was the census record of coresidence of other family members. Reliance on family context permits a high degree of confidence but biases the links toward those who persistently cohabit with the same family members. For example, we confirm the Logan and proprietor true links by (1) finding in both censuses at least one other household member (preferably two or more) with matching vital information, (2) making sure there is no significant contradictory information that makes a link improbable (for example, when one family member matches, but three others do not), and (3) determining that there is no other likely match in the 1881 Canadian census or the 1880 U.S. census.[3]

We have considerable confidence in the accuracy of the true links. They represent a useful diversity of population although, admittedly, they are not demographically representative insofar as they describe people living in the same family, or part of the same family in both years. This creates a bias to young children and married couples. Single people and those who became single over the decade (for example children leaving home) are underrepresented. Fortunately, even if the true links are not demographically representative, they still reflect the imprecision of

[2]The proprietors were linked in preparation for Inwood and Reid (2001). The Logan records were linked in preparation for Baskerville (2015). The St. James links were generated by Andrew Hinson for his doctoral dissertation (2010). The Quebec City links were made by the project *Population et histoire sociale de la ville de Québec* (www.phsvq.cieq.ulaval.ca) and kindly provided to us by Marc St-Hilaire.

[3]We check the United States census as well, because in this period Canadians could and did migrate to the United States.

information and name duplication needed to train the linkage system. Thus, our system will take this biased set of links and use it to produce new links that are less biased, more demographically representative and therefore more useful.

We use Ontario's high-performance computing grid SHARCNET (www.sharcnet.ca) because hundreds of millions of calculations are needed to compare name, age, place of birth, etc., and then to classify each pair of records as a match or non-match. Simply calculating similarities between millions of 1871 records and millions of 1881 records would require almost one year of continuous operation by a single processor.[4] Even running the system in parallel, however, a single run of the linkage system would be impractical without efficient code written in C, blocking to reduce the number of similarity comparisons and thresholding to remove some records from consideration.[5] We block by birthplace, marital status (allowing for obvious changes), first letter of surname, and our own first name groups (designed to allow for nicknames, diminutives, and unusual spelling variation). Similarity between pairs of names is assessed using the edit distance, Jaro-Winkler and double metaphone algorithms (Philips 2000; Winkler 2006). Similarity between birth years is assessed using a log-linear decay function. A description of the features used for linking and their similarity measures is given in Appendix Table 11.12.

Of course, many 1871 records cannot be matched because the individual died before 1881, left the country, or reported information differently in the 2 years. Nevertheless, the most common reason for failing to identify a match is not an inability to find someone with the same characteristics 10 years later. Rather, the biggest problem is that too many 1881 records have more or less the same characteristics as an 1871 record, and so produce multiple links. In such cases, we cannot identify which of the multiple links is correct. An example of records afflicted by the problem of "multiples" is given in Appendix Table 11.13.

The severity of the problem of multiples is clear from the distribution of outcomes for 1871 records, as reported in Table 11.1. About one-quarter are successfully linked in the sense that one 1871 record is classified as a match to only one 1881 record, and the 1881 record is matched to only one 1871 record. Another group comprising about one-quarter of the records cannot be linked with sufficient confidence to any 1881 record.[6] The largest group, 54 % of all 1871 records, consists of multiples. A multiple is an 1871 record that is either linked to more than one 1881 record or is part of a group of 1871 records linked to a single 1881 record,

[4]Computing similarity between all possible pairs of the 3 million and 4 million records on 8 string-based features with a single processor would require 343 days. Classifying each pair is additional.

[5]Blocking reduces the number of calculations. For example, we do not compare similarities between surnames beginning with different letters. Thresholding sets aside pairs of records that are sufficiently dissimilar that there is no prospect of being classified as a match.

[6]Thresholding and blocking remove 28 % of the 1871 records from consideration. A genealogical expert would be able to link some of these records but our automated system is less flexible. Table 11.1 reports the outcome of records submitted to the classification system.

Table 11.1 Outcome for 1871 census records in the classification system

	No. of records	Share
One-to-one links	550,726	0.215
No links returned	611,702	0.238
Many-to-one-and One-to-many links (multiples)	1,397,915	0.545

or both. Nevertheless, the system does report unique links for 550,000 people enumerated in 1871. This scale of longitudinal data is more than sufficient for most analysis providing these links are of sufficient quality. A careful assessment of these links is therefore needed.

11.3 The Level and Sources of Error Among the 1871–1881 Linked Records

In this section, we assess the level and sources of error among linked records. We begin with a general discussion of census data and their characteristics that make it difficult to link a substantial share of the records. This provides some context for the outcomes reported in the previous section. Next we assess the representativeness of the linked records using tabular descriptions and logistic analysis of the relative likelihood of linking different kinds of records. Finally, we point out that regardless of the overall error rate being low, a high proportion of the errors manifest themselves as individuals who have changed their location. This inflates the number of people who appear to have moved and complicates use of the linked data for migration analysis.

Our first question is if the system pairs up the right 1871 and 1881 records. Two kinds of mistakes are possible: an 1871 record can be linked to the wrong 1881 record, and an 1881 record can be paired to the wrong 1871 record. We assess the propensity for both errors by examining if the classification system has managed to identify correctly our true links, pairs of 1871–1881 records already linked with care by experts independent of the classification system. The fate of true links in the classification system indicates a combined incidence for both kinds of error of 3 %.[7] Is this a large or small number? We know that census data are in general somewhat imprecise. 3 % is similar to the rate for other sources of error in the North American historical censuses (Hacker 2013; Knights 1969; Parkerson 1991).

We might ask the same question of the 21 % rate of unique linking (Table 11.1). Is that high or low? Here is it useful to recognize that 30 % of our true links have surnames that differ by one or more letters and 20 % of the true links have name differences so large (edit distance > 0.15) that our classifier cannot find them. If the pattern of surname reporting in population is the same as in our true links, a full

[7] 3 % is the false positive rate on Ontario true links using a fivefold cross-validation method.

Table 11.2 Summary of probable limitations to potential link success

Records available	Loss of records	Reason and authority
100 %		
	20 %	Surname imprecision (true link analysis)
80 %		
	10 %	Age, birthplace, forename imprecision (true links)
70 %		
	5 %	Underenumeration estimate (Hacker 2013)
65 %		
	10 %	Emigration estimate (Emery et al. 2007)
55 %		
	10 %	Death estimate (Bourbeau et al. 1997)
45 %		Estimate of records available to be linked

20 % of the 1871 records cannot be linked for this reason alone. Imprecision in age, birth place, and first name reporting likely raises the "cannot link" share to at least 30 %. We also know that 10 % or more of the population would have died during the 1870s, and another 10 % would have emigrated. Another, smaller proportion may have been missed by enumerators.[8] Thus, we estimate, admittedly very roughly, that we are unlikely to be able to link more than 40–50 % of the records. We summarize the likely limitations to link success in Table 11.2. These estimates are of necessity somewhat speculative approximations.

The reason we achieve 21 % rather than 40–50 % is related to the reasons why there are any mistakes at all. Every time we cannot find the right person (for whatever reason), we are at risk of identifying the wrong person because of the widespread repetition of names, even among people with the same age, birthplace, and marital status. Multiple people who share a common set of characteristics are challenging in complicated ways. First, if a number of people have roughly similar characteristics (i.e., similar name, age, and birthplace), the system cannot distinguish among them, since a link cannot be accepted unless it is unique. Second, if the correct person reports age or name imprecisely, or if a woman changes her name at marriage, an incorrect person with similar characteristics might be selected in place of the correct one. In the first case no link is identified; in the second an incorrect link is made. A related problem arises if the correct person dies or emigrates before the next census, and therefore is not present in 1881. In this case, again, we are at risk of mistakenly selecting someone else with a similar combination of name, age, and birthplace.

Problems of this nature are more severe to the extent that names are common or that some kinds of people report their characteristics imprecisely. The imprecision means that occasionally we will connect together the wrong pair of records.

[8]Underenumeration in the nineteenth century the U.S. censuses is estimated to be about 5 % (Hacker 2013).

A second and perhaps more pervasive effect of imprecise reporting is to force a broadening of the tolerance for declaring a match. For example, we may accept any 1881 age between 28 and 32 for someone who reported 20 years in 1871 because someone is as likely to be 1–2 years off as to be exact in both years. Broadening tolerance, however, aggravates the problem of multiple links.

Classifying any data must strike a balance between broadening tolerance to avoid mistakes from a presumption of undue precision and; on the other hand, diminishing unique links by expanding the pool of multiples. It is particularly challenging to strike the right balance with our data because of their intrinsic imprecision. Many people did not remember their age or even their birth place correctly. The spelling of names varied a great deal. Enumerators who record information on the census manuscript page and volunteers who transcribe that information into a digital framework also made mistakes. In the face of this data imprecision, a combination of 21 % unique links and 3 % false positive errors (i.e., 3 % of the 21 %) reflects a successful balance of tolerances for linking characteristics. More importantly, the linked data are sufficient to identify and test hypotheses about broad patterns at the level of an entire population or large subpopulations.

11.3.1 Representativeness of the 1871–1881-Linked Records

Another way to assess our linked or matched data is to consider if they were broadly representative of the broader population. From the outset, we can anticipate reasons why linked records may be slightly atypical. We are more likely to link people with less common names and people who report their personal detail with greater precision and consistency. These biases are trivial unless they lead to other biases of greater analytical import.

In order to assess the implication of these and other biases, we compare the age and ethnicity of linked 1871–1881 records with the entire population in 1871. Here we use a subset of the linked records for which additional characteristics are available because they are part of a specially constructed 5 % representative sample. One effect is immediately apparent in Table 11.3: we link a much lower proportion of adolescents and young adults (15–25 years) than other groups. Young people are harder to link because they were of an age to move away from the family home, to start a new life, and to some extent reinvent themselves by reporting different characteristics. A propensity for women to change surname as they marry, of course, is an extreme example that leaves us with a noticeably smaller number of linked records for women aged 15–25 years.[9] The record-linking process is most

[9]We estimate, for example, that 40–45 % of single 15-year-old women in 1871 entered marriage during the following 10 years (comparing the number of single 15-year olds in 1871 with the number of married 25-year olds in 1881). The estimate is an approximation for only one birth year, however it suffices to indicate the scale of difficulty in linking young women. Only 2–3 % of women married someone with the same surname or retained their own surname in marriage.

Table 11.3 Age distribution of 1871 population and linked women and men

Age in 1871	Women		Men	
	Pop.	Linked	Pop.	Linked
0–14	0.38	0.40	0.39	0.38
15–25	0.24	0.16	0.22	0.19
26–55	0.31	0.37	0.30	0.35
56 and over	0.08	0.07	0.09	0.08

Source Canada, Census, 1871, 5 % microdata sample constructed at the University of Guelph http://census1871.ca (ignoring records for which age is missing). The linked records are generated by the People-in-Motion record-linking system (www.people-in-motion.ca) as described in Antonie et al. (2013)

successful for young children and the middle-aged, presumably because their information was reported more consistently over time. People over the age of 55 in 1871 are more difficult to identify in 1881 for a different reason—they were less likely to be alive in the latter year.

Interestingly, there is no bias to a more effective linking of the native born than of immigrants (Table 11.4). The foreign-born share of linked records is exactly the same as the foreign-born share of the population in 1871. The same is true for individual countries of birth (admittedly those born in England are overrepresented in the linked sample). This implies, unexpectedly, the linkage rate for immigrants is

Table 11.4 Distribution by nativity and ethnicity in the population and in linked records

	Pop.	Linked
Birthplace		
Foreign-born	0.19	0.20
England	0.04	0.06
Scotland	0.04	0.03
Ireland	0.06	0.06
Germany	0.01	0.01
U.S.	0.02	0.03
Canadian-born	0.81	0.80
Ontario	0.33	0.29
Quebec	0.29	0.30
Origin or ethnicity		
French	0.32	0.27
English/Welsh	0.20	0.27
Irish	0.25	0.23
Scottish	0.14	0.12
Continental Euro.	0.06	0.09
North American	0.01	0.003
African	0.01	0.005
Other	0.01	0.01

Source as Table 11.3

Table 11.5 Logit analysis (odds ratio) of 1871 records being linked uniquely, i.e., to a single 1881 record

		Married		Single/widowed	
		Women	Men	Women	Men
Male	1.18***				
Single	0.60***				
21–25	0.86***	0.85***	0.85***	0.71***	0.91**
>55	0.81***	0.87***	0.79***	0.99	0.65***
Fr. orig.	0.82***	0.72***	0.85***	0.91*	0.98
Illiterate	0.79***	0.67***	0.84***	0.81***	0.92
N	95,760	29,372	30,581	18,341	17,466

Note Full regression detail is available from the authors
*indicates that the coefficient differs significantly from 1.0 at 10 % confidence level
**indicates that the coefficient differs significantly from 1.0 at 5 % confidence level
***indicates that the coefficient differs significantly from 1.0 at 1 % confidence level

comparable to that of Canadian-born.[10] The linked records also mimic the population share of those born in the two largest provinces Quebec and Ontario.

There is some variance, however, with different ethnicities. Here we use the Canadian census category of "origin" as a measure of ethnicity. The information in Table 11.4 indicates a distribution of ethnicities roughly matching that of the population, with two important exceptions: fewer French-origin people are linked while the English origin are linked more successfully. The underrepresentation of people who report a French origin is notable.

11.3.2 The Likelihood of Establishing a Unique Link for Different Kinds of Records

We further investigate sources of linking bias in a logistic regression that considers the influence on being linked on age, sex, marital status, literacy, and if the individual reports a French origin.[11] The hazard, or odds ratios, reported in Table 11.5 indicate the contribution of each characteristic to the likelihood of being linked after controlling for other influences. A deviation from 1.0 indicates the size and direction of the effect; a number less/more than 1.0 indicates the odds of being linked for this category is less/greater than average. For example, in the first column

[10]This is unexpected because place of birth is reported more precisely for native born, to the level of province, in contrast to immigrants who simply report a country of birth. As well, immigrants or anyone moving a long distance has more scope for imprecise reporting of age, name, etc., than does someone living in the same location as his parents and family friends.

[11]We restrict the age categories being considered in this section because literacy is only available for people aged 21 or more years.

1.18 for men indicates they are 18 % more likely to be linked. The 0.60 reported for singles implies that they are 40 % less likely to be linked.

The odds ratios reported in the first column add to what we can learn from the previous tables. Men and married people are more likely to be linked; young adults and older people are less likely. These patterns conform to expectations. Singles are harder to link because they were more likely to change circumstances as they married (and of course most women changed their names). Younger adults were more likely to reinvent themselves as they left their parents' home. Some of them left the population entirely through emigration elsewhere in North America. Older adults were more likely to leave the population through death. We also see that people unable to read were less likely to be linked, as also for those reporting a French origin. The former is unsurprising. People lacking an ability to read probably reported their information with reduced precision. The French effect is more difficult to explain.

Partitioning the sample into married versus singles and men versus women allows more precise estimation of the age, ethnicity, and literacy effects (columns 2–5 in Table 11.4). For all groups, the youngest and oldest were less likely to be linked, but the effect was greatest for younger women (because of name-changing) and older men (because their 10-year survival rate was lower).[12] The French and illiteracy disadvantage is larger for women and for married people; the reason for these differentials is not immediately obvious. We do learn that the French effect is independent of literacy levels and age structure.

Records that are not linked fall into one of two groups: (i) we do not find even one good match in 1881 or (ii) we cannot identify the correct link because there are too many close possible matches.[13] We can estimate odds ratios for these effects separately (Tables 11.6 and 11.7). The odds of not finding of any match at all are large for older adults but this is offset by a smaller risk of losing sight of the correct match in a sea of multiple possibilities. In contrast, the younger adults are not at risk of being underlinked (Table 11.6) but they (especially single women) suffer a great deal from the problem of multiples (Table 11.7). For people reporting a French origin, the bias against finding a unique link arises primarily because of the failure to find even one possible link (similar to the older adults).

The challenge of finding unique links for the French-origin population leads us to estimate the odds of linking within this population. Table 11.8 reports the odds of finding at least one link. The pattern of odds ratios is very close to that of the general population (Table 11.6) with one exception. The impact of illiteracy on the odds ratio disappears for married men and becomes slightly stronger for married women.

[12]Similar patterns are observed if we abandon the restriction to people with 21 years of age or more. Literacy is unavailable for those under 21 but other effects are robust to the age restriction.

[13]For clarity, we highlight that the failure to find even one potential match can occur two ways: if the 1871 record is removed during the initial filtering or if it survives the filter but the classifier does not recognize any 1881 records with sufficient similarity.

Table 11.6 Odds ratios for finding at least one link for each record

	Married		Single/widowed	
	Women	Men	Women	Men
21–25 years	0.97	0.99	1.11***	1.16***
>55 years	0.60***	0.68***	0.51***	0.42***
Fr. origin	0.75***	0.78***	0.71***	0.74***
Illiterate	0.85***	0.91***	1.11**	1.01
N	29,372	30,581	18,341	17,466

Significance levels as in Table 11.5

Table 11.7 Odds ratios for finding only one link among the linked records

	Married		Single/widowed	
	Women	Men	Women	Men
21–25 years	0.83***	0.83***	0.62***	0.80***
>55 years	1.29***	1.11**	1.83***	1.22**
Fr. origin	0.82***	0.98	1.19***	1.22***
Iliterate	0.69***	0.87***	0.73***	0.88*
N	15,561	16,402	7,718	8,835

Significance levels as in Table 11.5

Table 11.8 Odds ratios for finding at least one link, French origin only

	Married		Single/widowed	
	Women	Men	Women	Men
21–25	1.08	1.13*	1.25***	1.16**
>55	0.65***	0.60***	0.49***	0.41***
Illiterate	0.83***	1.02	1.10*	1.11
N	9172	9670	6440	4527

Significance levels as in Table 11.5

Interestingly, although levels of illiteracy were higher in the French-origin population, and they are less likely to be linked, literacy patterns apparently did not contribute to the link bias (with the exception of married women).

Decomposing the link bias into two stages has not helped a great deal to understand the underlinking of older adults, people of French origin, and married people who cannot read. For these groups, we know only that we are less likely to find even one good link. Why that is the case remains unclear. The two-stage approach does help, however, with younger adults and singles who cannot read. We learn that there is a better than average prospect of finding a match for these groups (Table 11.6). Indeed, the problem is that we find too many good matches and in consequence cannot discriminate amongst them (Table 11.7). Any strategy for disambiguation of multiples might be especially helpful for the young adults.

We conclude that although the linked records are roughly representative of the 1871 population by birthplace and by major age and sex categories, there is some bias.

It is easy to see why younger single women and older men are less likely to be linked. Reweighting the linked sample by demographic category is an easy way to limit the impact of this bias in any analysis of the linked records.

There is a small but noticeable effect of illiteracy on the odds of being linked. The few people who described themselves as being unable to read were less likely to be linked. This must be kept in mind for any social or economic analysis using the linked sample. Fortunately, only a small share of the population was unable to read (about 10 % of young adults and 20 % of those aged 55 years or more).

There remains a mystery about the difficulty of linking people of French origin. This group comprises nearly one-third of the population. One possible explanation is that the quality of enumeration was influenced by language. Lower quality enumeration of the French-descended population might imply less precise or less consistent information that, in turn, would be more difficult to link. There is no reason, however, to think the census was undertaken less carefully in Francophone districts. A Quebec intellectual headed the Census Bureau in 1871, regional directors were drawn from the respective jurisdictions and most enumerators in French-speaking areas were themselves Francophone (Curtis 2000; Inwood and Kennedy 2012). Admittedly, any francophones relocating to English Canada were at greater risk of name misspelling.[14]

Dillon (2006) suggests (a) that the relatively small pool of French names increases the incidence of multiple links and makes it harder to isolate a unique match and (b) that the transcription of the 1881 census was weaker for French names. Both effects are plausible. Another possible influence is faster emigration of the French-descended population during the 1870s (Emery et al. 2007). Differential emigration and perhaps mortality would explain at least some part of the 25 % lower odds of finding at least one link for French-origin men and women (penultimate row of Table 11.5).

11.3.3 Error Rates Among Movers Versus Stayers

Linked or longitudinal census data are often used to describe and analyze mobility —both social and geographical. Elsewhere, we consider the broad patterns of occupational mobility during the 1870s (Antonie et al. 2015). Here we consider error rates among those who change location in order to assess the usefulness of these data for the study of migration. Since there are insufficient "movers" within our true links to support a comprehensive assessment along the lines reported in Sect. 11.3, we revert to a simpler strategy of checking if the individual links are

[14]The Jaro-Winkler and edit distance similarity measures are not phonetic and carry no obvious bias against recognizing similarities in the French language. Our third similarity measure, double metaphone, is phonetic but has been designed to minimize bias against languages other than English.

"credible" or not. This differs from the earlier evaluation insofar as we do not begin with secure knowledge acquired independently of the linkage system. Rather, we assess select links produced by the system in a way that relies to a large extent on the continued coresidence of other family members.

This process differs in principle from the generation of true links (above). Here we do not attempt to identify which 1881 record, if any, represents the same person as the 1871 record. That would require a broad investigation of all possible 1881 matches. The current process is more restricted and much less costly. We ask if the 1881 match recommended by the system has coresident family members who resemble those of the 1871 record using structured criteria (see Appendix 1). There might be a number of 1881 records with similar coresidents, but these are not checked. Rather, we assess the "credibility" of the one record selected by our linkage system.

This process is imprecise to the extent that we ignore other possible matches that, if examined, might reduce confidence in our results. Clearly, this implies a bias in favor of accepting matches recommended by the system. There are other sources of imprecision. For example, we use the coresidence in 1881 of people who would not be expected to be absent (given what we know from the 1871 family) as evidence undermining credibility.[15] And yet, families change for good reason; it is entirely plausible that family configuration changes and thereby creates the appearance of contradictory information. In these situations we may have a bias against acceptance of the correct match. Another complication is that we can assess the credibility of only those matches who have coresident family members in both years. We can say little about the credibility of links involving people who live alone or with non-family members in one or other year.

Although the process is particular in these ways, it provides an economical but plausible check on all kinds of linked pairs, with no obvious bias between different kinds of records. We use the method to compare people who appear to have changed provinces and those who do not. The distribution of linked pairs between interprovincial movers and stayers is reported in Table 11.9. Two verification assistants, independently, have checked each linked pair. Any differences are adjudicated; we report only those pairs on which there is consensus after adjudication.

We report our assessment of a random selection of links in Table 11.10.[16] The summary indicates a large difference between the movers and the stayers. Links for those people who stayed in place are highly credible; 83 % of the stayers are deemed credible (A and B categories) and only 5 % look to be incorrect (category D). In contrast, nearly half (45 %) of the linked pairs involving a change of province are incorrect.[17] The difference is dramatic, and invites explanation.

[15]Category D in Appendix 1.

[16]We examine a random selection of linked pairs for both movers and stayers.

[17]Although this example has only 39 movers, a larger sample of 1363 movers checked with a slightly different method had a comparable 42 % being deemed unlikely links.

Table 11.9 Distribution of 1871–1881 linked pairs by gender and interprovincial movers versus stayers

	Female	Male	All
No. of linked pairs	247,663	303,030	550,726
No. of links with change in province	8037	9848	17,910
Apparent movers as a share of all links	0.032	0.032	0.033

Table 11.10 Individual assessment of linked pairs implying movement between provinces from 1871 to 1881

	Movers	Stayers
Number of records checked	39	1787
Share assessed highly credible (A)	0.46	0.76
Share assessed credible (B)	0.05	0.09
Share that cannot be confirmed (C)	0.15	0.10
Share assessed likely incorrect (D)	0.33	0.05

NB: Here we report linked 1871–1881 pairs for which two independent assessments agree after adjudication. "Movers" are records that imply a change in province of residence. The assessment categories are described in Appendix 1.

One reason for errors among the reported movers is that some proportion of 1871 records cannot be linked properly. For example some individuals died or left the country before 1881, or were present and overlooked by enumerators, or were enumerated in 1881 with some misstatement of personal information.[18] Situations like these prevent the system from making a correct link.[19] Further, as noted above, when the correct link is not available, the system may identify incorrectly someone else with similar personal characteristics. For example, a 48-year-old woman named Joanna Munroe who in 1871 was enumerated in Southampton, New Brunswick, was linked in 1881 to Jane Munroe, a 58-year-old from Lingan, Nova Scotia. While all the linkage criteria match very well (only the first name is off), different coresident families make it clear they are different people. The linkage error is attributable to the 1881 Jane being enumerated as Jessie (a Scottish nickname for Jane) in 1871, and the fact that the 1871 Joanne Munro had likely died by 1881. Because location is not used for linking, mistaken 1881 links like these will have a wide geographical distribution. If the mistaken link is in another province the system can generate a "phantom mover".[20]

[18]Socially marginal groups such as aboriginal, African-descendants or Chinese are more likely to be enumerated with substantial imprecision (Reid 1995; Fryxell et al. 2015).

[19]The most careful, genealogical-like researchers seldom manage to surpass an 80 % rate of linking from one Canadian census to another, for exactly these reasons. See Darroch (2015), Baskerville (2015) and Olson (2015).

[20]Ron Goeken at the Minnesota Population Center first suggested this interpretation of the relationship between geography and errors in linking.

Table 11.11 Simulated share of observed state changes that are correct

True rate of state changes						
	0.03	0.05	0.1	0.2	0.3	0.5
# possible states						
2	0.47	0.34	0.21	0.12	0.08	0.05
3	0.54	0.41	0.26	0.15	0.10	0.07
4	0.57	0.44	0.29	0.17	0.12	0.07
5	0.58	0.46	0.30	0.17	0.12	0.08

Notes The simulation is based on the idea that a characteristic for a record has a number of possible states, e.g., for the locational characteristic, there might be two locations, or three locations, etc. For a consideration of interprovincial movement, there are four possible states in 1871, corresponding to the four provinces. We assume records are distributed equally across all possible states, i.e., if there are 2 states, they are 50 and 50 %. If 4 states, each has 25 %. This is like assuming the four provinces in 1871 are of equal population size. Further assume that any mistake in linking is random with respect to states/locations (e.g., if there only two locations, any "mistake" will be in the same place in 1881 as in 1871 half of the time). The other half of the time the mistake will register as a change of state. If there are three states, the mistakes will appear as a change of state two-thirds of the time. Some correct links also appear as a change of state since some people really do change provinces. We predict the likely number of true and phantom movers under these simple assumptions, and report the phantom share of reported movers in Table 11.11. Formally, the table is generated as

WM (wrong movers) = P (population size) * FPR (false positive rate) * $(S - 1)/S$; S is number of states
CM (correct movers) = P (population size) * TPR (true positive rate) * TRS (true rate of state change)
Phantom movers rate = WM/(WM+CM)
True movers rate = CM/(WM+CM)

This phenomenon complicates use of the data because people who really did not move but were linked incorrectly contaminate the evidence of movement. Indeed, if there are few genuine interprovincial movers, as in the 1870s (Baskerville 2015) then a large share of the apparent movers may be mistaken, and the overall level of mobility is exaggerated significantly. The overall error rate is still 5 % or less, but *among the reported movers* the proportion of mistakes can be much higher.

A simple simulation in Table 11.11 illustrates that an uncomfortably large proportion of the apparent movers will be mistakes if the true extent of movement is less than 15 %. Changes in religion, occupation, etc., will have a similar problem. The implication is that analysis of change by a small proportion of the population will be subject to more uncertainty than is suggested by the overall error rate of 5 %.[21] In practice, of course, the severity of this complication depends a great deal on particular circumstances, as illustrated in Table 11.11.

[21]This is independent of how well the system links people who really did move; the problem is not the quality of data describing true movers. That said, movers were disproportionately young adults who generally are more challenging to link. For this reason, the system may generate a higher rate of error among true movers. The only way to assess this possibility would be to generate more true links than currently are available.

11.4 Summary and Observations

The application of machine-learning systems to historical censuses generates useful data describing people at different points in their lives (Ruggles 2006). The method is especially important for jurisdictions that lack comprehensive church or public vital registration and must depend on the census for understandings of population-wide experience. The new longitudinal source provides, for the first time, large-scale and near-representative life-course information about nineteenth century Canadians. This is an important and very welcome development.

The nature of the source and underlying population does not allow us to link every record. Nevertheless, as we demonstrate with the 1871 and 1881 Canadian censuses, it is possible to generate samples large enough for most historical and social science research. The overall quality of the data, as reflected in a low rate of false positive links, is excellent. A carefully designed system brings the false positive rate down to an acceptable range, circa 3 % on independently verified links.

We assess the extent of bias or representativeness of the linked pairs by examining unconditional means and with logistic analysis of the propensity to link. We find that birthplaces are reasonably representative of the population. The linking method is slightly more successful for immigrants born in the British Isles but otherwise it roughly replicates the proportions of the population born in Canada versus immigrants and in one province versus another. People who were unable to read are noticeably more difficult to link, but they account for a small share of the population. Older men and young adults are more difficult to link than people at other ages. The former reflects differential mortality at advanced ages; the latter probably reflects change accompanying the departure of children from a family home. A near universal tendency for women to change their surname at marriage is the largest single complication in this vein.

A lower rate of linking people who report a French origin in 1871 is more puzzling. There is no reason to think that the enumeration of Francophone communities was in any way inferior. Logistic analysis rejects the hypothesis that ethnic differences in literacy are responsible. Literacy matters, but it does not explain the ethnic differential in linking. Breaking the process into two stages, identification of at least one promising match and discrimination among multiple possibilities, points to the first stage as especially challenging for the French-origin records. Again, however, there is no reason to think blocking or the use of similarity measures in the first stage carries a bias against French language names. Further investigation of similarity algorithms for French names may prove useful.[22]

[22]It is worth noting that the Canadian census category of 'origin' is itself obscure. People were asked their 'origin' in the sense of ancestry or ethnicity, but as best we know no instructions were made available about how to identify in the event of mixed ancestry. There is likely to have been some discretion in the self-identification of origin. An improved understanding of this process may help us to understand why French origin Canadians are more difficult to link.

Analysis of the odds of linking shows that the underrepresentation of French-origin population is pronounced only for married people (and is especially large for married women). Until this problem is better understood, it would be prudent in most research to reweight observations to correct the underrepresentation of French-origin married couples.

One final problem, a higher rate of mistaken links among those who appear to move between provinces, is easier to understand. The bias arises because a linking error for any reason is likely to generate the appearance of geographic relocation. The problem of phantom migrants looms large when the true rate of moving is low. In principle we might mitigate the effect by adjusting standard errors for hypothesis testing, but in practice this is difficult because we do not know the true rate of moving independent of the analysis. As a practical matter, therefore, when the reported rate of changing category falls below 15 %, it would be prudent to verify the intrinsic credibility of linked pairs implying a change of state. Admittedly, verification is only possible for those who continued to live with the same family members. Thus, even after a process of verification, linked data cannot be used to analyze the relationship between family evolution and migration if the rate of reported movement is small.

Our assessment of the linked records identifies specific limitations notwithstanding their excellent quality overall. Some problems are small enough to ignore (impact of illiteracy, small deviations in birthplace composition). Others require a simple reweighting to compensate for underrepresentation (younger single women, older men, French-origin married couples). The clustering of errors among movers when only a small proportion appears to move requires more caution and where possible manual verification. These are manageable problems, which further research and improvements to the record-linking system may reduce further.

Our experience linking historical census data indicates that, for this case at least, an optimal application of machine-learning methodology takes account of the quality of underlying data. Of course, this is only one case study. Nevertheless, if our experience were to be replicated elsewhere, it would be useful practice for computing science researchers to take data characteristics into account in their application of otherwise standard machine-learning methods. There is a comparable lesson for social science and historical users. If the issues encountered with linking the Canadian data were to obtain elsewhere, social scientists and historians would find it useful to assess and accommodate data quality issues that arise from the intersection of sophisticated machine-learning methodology and sometimes messy historical data.

Acknowledgments The authors are grateful for financial support from the Canadian Foundation for Innovation, Ontario Ministry of Research and Innovation, Social Sciences and Humanities Research Council, Google, Sharcnet and the University of Guelph. We would also like to thank our genealogical collaborators, the Ontario Genealogical Society, Ontario GenWeb and Family Search.

Appendix 1: Protocol for Checking Automatically Generated Links

We check the reliability of links in order to prepare Table 11.10 and assess the relative "movers" and "stayers." This process differs from that of determining true links insofar as (i) we do not rule out the possibility of other, equally plausible matches and (ii) we cannot bring to bear any insight from the independent study of some community or subset of the population. Checking involves two independent experts assessing a link without reference to each other's decision (blind double-checking). Each link is assessed based on the household information in the two census years, as well as the consistency of information, and then assessed with a quality letter grade. The basic question being asked and answered is the common genealogical query: Is this the same person in both records? (Tables 11.12 and 11.13)

Table 11.12 Description of linking features

Original attribute	Type	Similarity measure(s)	Feature score
Last name	String	Edit distance (ED): the minimum number of single letter edit operations needed to convert string A into string B	Float [0–1]
		Jaro-Winkler (JW): calculated based on the number of common characters, character transpositions and string length between two strings, giving preference to strings that share a common prefix	
		Double metaphone (DM1, DM2): transforms strings into their corresponding phonetic representation, creating a primary and secondary representation on which edit distance is applied	
First name	String	See above	Float [0–1]
Age	Integer	1 if $x \in 0, 1, 2$ $F(x) = 1–1/x$ if $x \in [3,10]$ 0 otherwise	Float [0–1]
Gender	Binary	Exact match	Binary (0,1)
Birthplace	Categorical	Exact match	Binary (0,1)
Marital Status	Categorical	Rule based 1 if valid status change (ex. single to married) 0 otherwise	Binary (0,1)

Note Blocking techniques are applied on three different attributes to reduce the number of record-pairs being compared. These attributes are a name-code based on the first name, the first letter of the last name and birthplace. This means that a record-pair is considered for comparison only if the two records reside in the same name-code and last name block of their respective censuses, and their birthplaces match

Table 11.13 An example of census records with similar attributes

Surname	Forename	Age	BPL	Marital status
1871 Census				
Barns	Mary	11	15030	Single
Barns	Mary	9	15030	Single
Barns	Mary	8	15030	Single
Barns	Mary	12	15030	Single
Barns	Mary	10	15030	Single
Barns	Mary	10	15030	Single
1881 Census				
Barns	Mary	20	15030	Single
Barns	Mary	22	15030	Single

BPL = birthplace

In addition to the grading, experts provide reasons for their decisions by recording the answers to certain questions. This information (a) helps us refine our linkage system, (b) allows us to compare decision-making between coders and ensure consistency, and (c) possibly change quality grades in future without having to revisit the links manually.

1.1 Links: Primary and Subsidiary

A primary link is the one that the system linked using six linking variables (First Name (FN), Last Name (LN), Age, Marital Status (MS), Birthplace (BPL), and Sex), and this kind of link is the one that we are interested in giving a link quality assessment. In the course of checking the primary links, we may also see other people who link up. These we call a subsidiary link and are usually a household member of the primary link whom we have determined with good confidence is the same person in both years. It may be a spouse, sibling, or child or parent, or even servant.

1.2 Deciding on Quality

The six linking variables used by the automated linkage system to generate the primary link are likely to be very consistent, and so not very useful for distinguishing false positives by themselves (although commonness of surnames could be a consideration). Accordingly, in order to verify a link, checkers consider the household/family context as well as other personal fields (of primary and provisional subsidiary links) and also to assess whether or not they appear consistent.

1.3 The Questions to Ask

- Household/family Context—does the family have some of the same members in both years? Are family member details (FN, LN, Age, MS, BPL, Sex, Origin, Religion, etc.) consistent?
  - Does the spouse match? (Could the linked person have remarried?)
  - How many children match by name and age? (exclude those who were born in census period, i.e., those aged under 10)
  - Are there any other family members that are the same? (e.g., parents, servants)
  - Does the household transition make sense—deaths, leaving family to start new family, etc.
  - Are children on the same age ladder?
- Is birthplace consistent?
- Is ethnic Origin consistent? (children's origins sometimes change to follow one or other of the parents)
- Is Religion consistent, or show a likely transition (i.e., more likely between Protestant denominations than between Catholic and Protestant)?
- In some circumstances, contradictions in other fields may also give a reason to look more closely at a link (e.g., an unlikely occupational change, or an eastward (as opposed to westward) long distance (i.e., interprovincial) move).

The link checking protocol may include one or more of the above questions in explicit form as fields to be filled out, and these are usually designed to require notation only in cases where the answer is unexpected. In addition, there is a Comments field, in which checkers can indicate other difference in the information given for the same person in the two censuses.

1.4 The Link Quality Typology

The assessment of link quality is a holistic summary of the answers to these questions, with the primary consideration being the matching of family members, although contradiction of information is considered. The qualities are:

A = Two or more family members match

- With no major contradictory information (such as children appearing in one census that were not there ten years before)
- In some cases, neighboring families can be used to make an A (see CM and CF below)
- e.g., Spouse and child

- In very rare cases an A can be achieved with fewer than two subsidiary links if there is certainty it is the same person (e.g., in a case where a man has no family by next census (wife dead and children grown up) and he has moved in with neighbors/other family that are evident in both census but whose records not linkable because they are not in the link set in the first census year.)

B = One family member matches

- e.g., spouse or child
- With no major contradictory information

C = Possible match but no family in one year to confirm against

- e.g., single man in rooming house in 1871, or a man in barracks in 1881
- Information otherwise very consistent

CM = Single to married man with new family by next census

- MS will change from single to married, and children (if any) will be below the age of 10.
- e.g., a single man in 1871 (on his own or in a family) got married and started his own family by 1881.
- When possible, we check if family members are in neighboring households—in this case CM might be upgraded to an A.
- If the man is a widow in the first year, and married the next, then it is a CB.

CF = Single to Married woman with new family by next census (Rare)

- MS will change from single to married, and all children will be below the age of 10.
- e.g., a single woman in 1871 (on her own or in a family) got married and started a own family by 1881.
- Some women did keep their own names in some cases (French and Scottish), but in most cases single to married women with the same surname will be bad matches (D) (linking criteria may even prevent a link in the first place). Therefore, this code is used only when there is very good evidence it is the same person (e.g., she retains her maiden name in the married family (husband has different surname); or there is evidence she married a man with the same name (possibly a neighbor); or she has been enumerated with the same (birth) family in both years).
- When possible, we check if family members are in neighboring households—in this case CF might be upgraded to an A.

CB = Possible match, but for less common reasons

- These are possible matches where families do not match, but links may be possible. Examples of these are:
  - Widow/Widowers—A older married man or woman with family is alone by the next census and marital status has changed to widowed/divorced/separated (or possible married/spouse absent)
  - A single person who has joined a different family to work as a servant in 1881
  - Spinster/Bachelors who change families
  - A man with only a wife in both years, but wife's name might change, however all other information about her stays the same, and they still have the same neighbors.
  - If the man is a widow in the first year, and (re)married the next.

D = evidence of wrong match

- e.g., families are different, and/or there is significant contradictory information.

1.5 Evaluation/Arbitration

When checkers disagree on the quality of a link or whether a newID should be assigned, the records are either reevaluated by the checkers or arbitrated by a third party for a final decision.

References

Antonie, L., Inwood, K., Lizotte, D., & Ross, J. A. (2013). Tracking people over time in 19th century Canada. *Machine Learning, 96*(1) (S1), 129–146.

Antonie, L., Baskerville, P., Inwood, K., & Ross, J. A. (2015). Change amid continuity in Canadian work patterns during the 1870s. In P. Baskerville & K. Inwood (Eds.), *Lives in transition: longitudinal research from historical sources* (pp. 120–140). Kingston: McGill-Queen's University Press.

Baskerville, P. (2015). Wilson benson revisited: Movement and persistence in rural Perth County, Ontario, 1871–1881. In P. Baskerville & K. Inwood (Eds.), *Lives in transition: longitudinal research from historical sources* (pp. 141–164). Kingston: McGill–Queen's University Press.

Bourbeau, R., Légaré, J., & Édmond, V. (1997). *New birth cohort life tables for Canada and Quebec, 1801–1991*. Ottawa: Statistics Canada.

Christen, P. (2008). Automatic record linkage using seeded nearest neighbour and support vector machine classification. *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 151–159.

Curtis, B. (2000). *The politics of population: State formation, statistics, and the census of Canada, 1840–1875*. Toronto: University of Toronto Press.

Darroch. G. (2015). Lives in motion: Revisiting the 'Agricultural Ladder' in 1860s Ontario, a study of linked microdata. In P. Baskerville & K. Inwood (Eds.), *Lives in transition: longitudinal research from historical sources* (pp. 93–119). Kingston: McGill–Queen's University Press.

Dillon, L. (2006). Challenges and opportunities for census linkage in the French and English Canadian context. *History and Computing, 14*(1–2), 185–212.

Emery, H., Inwood, K., & Thille, H. (2007). Hecksher-Ohlin in Canada: New estimates of regional wages and land prices. *Australian Economic History Review, 47*(1), 22–48.

Ferrie, J. 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*, 141–156.

Ferrie, J. P. (1999). *'Yamkees Now': European immigrants in the antebellum U.S., 1840–1860*. New York: Oxford University Press.

Fryxell, A., Inwood, K., & Van Tassel, A. (2015). Aboriginal and mixed race men in the Canadian expeditionary force 1914–1918. In P. Baskerville & K. Inwood (Eds.), *Lives in transition: Longitudinal research from historical sources* (pp. 254–273). Kingston: McGill–Queen's University Press.

Fu, Z., Boot, M., Christen, P., & Zhou, J. (2014). Automatic record linkage of individuals and households in historical census data. *International Journal of Humanities and Arts Computing, 8*(2), 204–225.

Goeken, R., Huynh, L., Lenius, T., & Vick, R. (2011). New methods of census record linking. *Historical Methods, 44*(1), 7–14.

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

Hinson, A. (2010). Migrant scots in a British City: Toronto's scottish community, 1881–1911. Ph. D. Dissertation University of Guelph.

Inwood, K., & Kennedy, G. (2012). A new prosopography: The enumerators of the 1891 census in Ontario. *Historical Methods, 45*, 65–77.

Inwood, K., & Reid, R. (2001). Gender and occupational identity in a Canadian census. *Historical Methods, 32*(2), 57–70.

Knights, P. R. (1969). A method for estimating census under-enumeration. *Historical Methods Newsletter, 3*(1), 5–8.

Knights, P. R. (1991). *Yankee destinies: The lives of ordinary nineteenth-century bostonians*. Chapel Hill: University of North Carolina Press.

Olson, S. (2015). Ladders of mobility in a fast-growing industrial city: Two by two, and twenty years later. In P. Baskerville & K. Inwood (Eds.), *Lives in transition: Longitudinal research from historical sources* (pp. 189–210). Kingston: McGill–Queen's University Press.

Parkerson, D. (1991). Comments on the underenumeration of the U.S. census, 1850–1880. *Social Science History, 15*(4), 509–515.

Philips, L. (2000). The double metaphone search algorithm. *C/C ++ Users Journal, 18*, 38–43.

Reid, R. (1995). The 1871 United States census and black underenumeration. *Histoire sociale/Social History, 28*, 487–499.

Richards, L., Antonie, L., Areibi, S., Grewal, G., Inwood, K., & Ross, J. A. (2014). Comparing classifiers in historical census linkage. *Data Integration and Applications Workshop, in Conjunction with IEEE ICDM 2014.*

Ruggles, S. (2006). Linking historical censuses: A new approach. *History and Computing, 14*(1–2), 213–224.

Steckel, R. H. (1988). The health and mortality of women and children, 1850–1860. *The Journal of Economic History, 48*(2), 333–345.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. Heidelberg: Springer.
Winkler, W. E. (2006). Overview of record linkage and current research directions. *Statistical Research Division Report*. U.S. Census.
Wisselgren, M. J., Edvinsson, S., Berggren, M., & Larsson, M. (2014). Testing methods of record linkage on swedish censuses. *Historical Methods, 47*(3), 138–151.

Chapter 12
Using the Canadian Censuses of 1852 and 1881 for Automatic Data Linkage: A Case Study of Intergenerational Social Mobility

Catalina Torres and Lisa Y. Dillon

Abstract This chapter discusses the issues of missing and uncertain data in the Canadian census sample of 1852 within the context of automatic linkage with the complete census of 1881. The resulting linked sample from these two censuses was created to provide an opportunity to study intergenerational social mobility in Canada between fathers (in 1852) and sons (1881). We discuss the accuracy and representativeness of the automatically generated links and show how the use of marriage registers can be helpful in order to verify the results of the automatic linkage. Our verifications suggest that most of the links are accurate. However, the linked sample is not representative of some subgroups of the studied population, since some attributes favoured while others hindered the fact of being automatically linked from 1852 to 1881. Finally, based on our efforts of manual linkage between the BALSAC marriage registers and the automatically linked census sample for the verification of the latter, we present some considerations about the great research potential of linking census and parish register data in Quebec.

12.1 Introduction

Between 2004 and 2006, the Programme de recherche en démographie historique (PRDH)[1] created a 20 % sample of the first nominal census of Canada in the nineteenth century: the census of 1852. The quality of this census has been criticized by some researchers. For example, Gagan (1974) described the "lack of

[1]Research programme in historical demography, Université de Montréal.

C. Torres (✉) · L.Y. Dillon
Programme de recherche en démographie historique (PRDH), Université de Montréal, Montreal, Canada
e-mail: catalina.torres@umontreal.ca

L.Y. Dillon
e-mail: ly.dillon@umontreal.ca

G. Bloothooft et al. (eds.), *Population Reconstruction*,
DOI 10.1007/978-3-319-19884-2_12

consistency" of this census in making reference to the irregular quality of the 1852 census manuscripts. A more recent critique made by Curtis (2001) concerns the combination of the approaches *de jure* and *de facto* in the taking of the 1852 census. By this combination of approaches, the Canadian population of 1852 could be overestimated. Dillon and Joubert (2012), who have examined the 20 % sample of the 1852 census in the light of those critiques, suggest that the remarks made by Gagan and Curtis regarding the quality of this census concern a minority of the observations. Thus, the 20 % sample of the 1852 census offers unique opportunities to broaden our knowledge about the Canadian population of the mid-nineteenth century, particularly the rural population (Dillon and Joubert 2012).

Both the census sample of 1852 and the 100 % database of the 1881 Canadian census constitute rich sources of information about the Canadian population of the mid- and late-nineteenth century. For example, both sources contain valuable socio-economic variables and provide information at the individual level, making these data suitable for record linkage.[2] For instance, by linking individuals (e.g. the boys of a certain age) from 1852 to 1881, phenomena such as the intergenerational social mobility can be studied through a comparison of the occupation of the fathers (in 1852) and the sons (in 1881).[3]

This chapter analyzes an automatically linked sample from the Canadian censuses of 1852 (20 % sample) and 1881 (complete census). Since the aim of this sample is to provide opportunities to study the intergenerational social mobility between fathers (in 1852) and sons (in 1881), the linkage efforts were concentrated on a limited subgroup of the population, namely the boys aged from 0 to 15 years, living mainly in a rural area in the provinces of Ontario or Quebec in 1852. In total, our linked sample contains information about 4226 individuals linked from 1852 to 1881. This linked sample was created for exploratory purposes in the framework of the international project *Mining Microdata: Economic Opportunity and Spatial Mobility in Britain, Canada and the United States, 1850–1911*.[4] The two census data sources were provided by the PRDH. The linkage between both data sets was performed in the Historical Data Research Unit (HDRU) at the University of Guelph, while the Mining Microdata project is pursuing a parallel linkage effort of the 1852 and 1881 censuses headquartered at the Minnesota Population Center.

In order to analyze this linked sample, we start with a brief description of the linkage technique, followed by a discussion of the results of the automatic linkage. This discussion includes some considerations of the factors that affect the linkage

[2]The PRDH has lengthy experience with the record linkage of Quebec parish registers, and more recently undertook linkage of a sample of the 1871 Canadian census to the 1881 census. Our current effort to link the 1852 and 1881 Canadian censuses is funded by the international project *Mining Microdata: Economic Opportunity and Spatial Mobility in Britain, Canada and the United States, 1850–1911,* Digging Into Data Challenge, ESRC/NSF/CRSH.

[3]In both censuses, the only information about the socio-economic status of the individuals is the occupation.

[4]This project aims to contribute to the discussion about the social and geographical mobility in North America and in Great Britain in the late nineteenth and early twentieth centuries.

success, such as mortality and emigration. Indeed, linking individuals from 1852 to 1881 is a substantial challenge, since the larger the interval of time between the two observations, the more the individuals could be lost and become impossible to link due to those factors. Following this discussion, we present some analyses of the accuracy and representativeness of the automatically generated links. In this stage, we will show how the use of the BALSAC marriage registers (Balsac fichier de population 2013) can help assessing the validity of those links. Finally, we will discuss the utility of linking census data in order to study intergenerational social mobility between fathers and sons from 1852 to 1881.

12.2 Linkage Technique and Results

The automatic linkage procedure employed to generate our linked sample from 1852 to 1881 is very similar to that explained by Antonie et al. (2015, this volume) regarding the linkage between the Canadian censuses of 1871 and 1881. This procedure is based on the individual attributes that should not change over time, such as first and last name,[5] gender and place of birth. Other characteristics that change over time but in a predictable way, namely age and the marital status, are also used in the linkage procedure. We limited our record linkage criteria to this range of attributes to avoid biasing the sample in favour of stable individuals; this kind of bias could occur if variables which change over time, such place of residence and occupation, were used to link cases.

The linkage procedure—explained in more detail by Antonie et al. (2015)—starts with data cleaning and standardization. Once the variables previously mentioned are cleaned and standardized, the observations of the two data sets can be compared in order to find the 1881 record that corresponds to each individual in 1852. In order to reduce the number of comparisons, blocking by some characteristics is useful, for example by birthplace at the country or provincial level. This means, for instance, that boys born in Canada according to the 1852 census are compared only with men born in Canada according to the 1881 census. Similarity scores for the comparison of each attribute between two individuals as well as global score are generated. Based on the similarity scores, the link is accepted when there is only one candidate who passes a certain threshold (one-to-one approach). The candidates are chosen by a support vector machine (SVM) programme, which uses training data—a previously generated set of manual links made by genealogist experts—as a guide for the acceptable links.

[5]Since for our purposes we were interested in linking men only, we are not faced to the problem of changing name at marriage. This was usual among women and more frequent among some subgroups of the population than others.

Table 12.1 Survivors by age group, males aged 0–19 living in rural Quebec or rural Ontario in 1852

Age group (1852)	N (1852)	Survivors in 1881 (%)
0–4	18,492	73.6
5–9	18,015	80.6
10–14	15,425	78.9
15–19	13,638	75.6

Sources Canadian census of 1852 (20 % sample)

Despite the fact that some differences in name spelling between 1852 and 1881 are tolerated, as well as some discrepancies between the expected and the observed age, the linkage process briefly described above advantages individuals with more accurate information on both censuses. However, since the one-to-one approach reduces the quantity of false links, this linkage procedure should favour the precision and the representativeness of the linked sample (Roberts 2012).

With 4226 individuals linked from 1852 to 1881 and taking into account mortality, emigration and imprecise data, we estimate that the linkage rate is about 15 %. First, based on the period life tables of Boubeau et al. (1997), we estimate that around 25 % of the 57,023 boys who composed the initial population in 1852 died before 1881 (Table 12.1).

The method to calculate the proportion of survivors by age group in 1881 is based on the methods presented by Boubeau et al. (1997). The mortality quotients by age group and year for the male population of Canada come from their period life tables.

Regarding emigration, we consider that around 15 % of our initial subpopulation of boys could not be linked because of emigration to the United States. Contrary to the mortality calculations, the latter estimate is not based on quotients because there are no emigration quotients by age and sex for the Canadian population during the period 1852–1881. In automatically linking individuals from 1871 to 1881, Antonie et al. (2015) estimated that the percentage of linkage failure due to emigration is about 10 %. In our case, we can reasonably establish that the percentage of individuals who could not be linked because of emigration is higher because the American censuses of 1860, 1870 and 1880 suggest that immigration of men born in Canada between 1835 and 1854 (i.e. the approximate birth cohort of our subpopulation of interest) was particularly important during the 1860s. Table 12.2 shows that the immigration of males born in Canada between 1835 and 1854 seems

Table 12.2 Number of males born in Canada between 1835 and 1854 enumerated on at least one U.S. census between 1860 and 1880 by age cohort. For 1860 and 1870: samples (weights applied); 1880: complete census

Census year	Age cohort	N
1860	6–25	59,600
1870	16–35	135,600
1880	26–45	157,200

Source IPUMS U.S. censuses of 1860, 1870 and 1880 (Ruggles et al. 2010)

to have been particularly important during the 1860s: in the 1870 census, the number of those Canadians is more than two times higher than the corresponding number in the census of 1860. In the light of these numbers, our estimate of 15 % is probably conservative.

Finally, regarding the accuracy of the information stated in the census manuscripts, Antonie et al. (2015) estimate that around 10 % of the observations could not be linked between 1871 and 1881 because of imprecise information regarding age, first name and place of birth. In our case, we can expect a higher percentage of linkage failure due to imprecise information, since persons in 1852 likely had lower literacy than persons in 1871, suggesting a greater possibility of imprecise declarations in 1852 compared to 1871. Moreover, we have to consider not only the imprecision of the information stated on the census manuscripts, but also the inaccuracies arising from the transcription of the names from the original manuscripts. Our estimate of linkage failure due to inaccurate information is 12 %.

Thus, with the three estimates regarding linkage failure due to mortality, emigration and imprecise information, we can establish that the percentage of boys in the initial subpopulation who it would be impossible to link to the 1881 census might be as high as 50 %. Our final linkage rate was 15 %, yielding 4226 linked cases. This rate must be interpreted in the light of this linkage failure estimation. Although a linkage rate of 15 % is low compared to other studies using similar linkage techniques (e.g. Antonie et al. 2015; Long 2005), one has to consider that an intervening span of 29 years between observations is quite long, which increases the chances of linkage failure due to mortality and emigration. In order to enhance the linkage rate, we could have used an intermediate census between 1852 and 1881, e.g. the 1871 census sample: because of the shorter time between observations, a linkage between 1852–1871–1881 would probably have resulted in a higher linkage rate. However, as the source for 1852, the 1871 data source is a census sample. Linking from sample to sample implies more uncertainty in the accuracy of the links, since the true link for an individual could be someone excluded from the sample. For this reason, "each linked pair of censuses must include at least one complete enumeration" (Roberts 2012, p 7). In addition, over and above the number of the links, it is the accuracy and the representativeness of the links that we are interested in.

12.3 Accuracy and Representativeness of the Linked Sample

An assessment of the accuracy and representativeness of the 4226 1852–1881 linked cases based solely on the 1852 and 1881 censuses is challenging. The main difficulty arises from the fact that we do not have a complete enumeration of the population in 1852 (the 1852 source is a sample of 20 %) as we do for 1881. A 100 % index of the 1852 census does exist, which includes first and last names,

age, gender, place of birth and place of residence. However, unlike the 1881 microdata, this index is not clean for the moment and cannot yet be used for linkage projects. Furthermore, this index omits occupations, making the study of intergenerational social mobility impossible. The fact that our data source for 1852 is a sample that hinders the validation of the automatically generated links. For a given linked individual from 1852 to 1881, one potential match in 1881 for the corresponding individual in 1852 was found, and vice versa. However, the correct link for the 1881 individual may not be the one found in the 20 % sample but someone else who is not included in this sample. The linkage with the wrong 1852 individual might have been accepted because both the correct (not included in the 1852 sample) and the false (included) links have very similar personal information. Had the other "true", individual been included in the 20 % sample, the linkage had probably not been accepted because there would be more than one single candidate.

We can still verify the quality of the automatic links by drawing upon other historical data sources. Thanks to the availability of marriage registers for a subgroup of the population—namely Catholics who celebrated their marriage(s) in a Quebec parish between 1852 and 1881—, we could verify the accuracy of some of the automatically generated links via manual linkage at the individual level between the automatically linked sample (1852–1881) and the corresponding marriage registers.

Figure 12.1 provides an example of the basic principle used to assess the validity of a link between the censuses of 1852 and 1881: the automatically linked sons (Alfred and Isidore Poitras) lived with their father and mother or stepmother in 1852 (Narcisse Poitras and Thérèse Pelletier) and with their respective spouses in 1881 (Philomène Etu and Virginie Jalbert). In the marriage registers, we find the names of the new couple (the same names we would expect to find in the 1881 census) along with the names of the parents of the linked son (also the same names that we would expect to find in the 1852 census). As clearly indicated in Fig. 12.1, the only information available to establish the validity of this link are the first and last names of the son, his parents (in the 1852 census and in the marriage registers) and his wife (in the 1881 census and in the marriage registers).

We looked for the marriage certificates of 533 individuals who had been automatically linked from 1852 to 1881. However, among these individuals, 130 (24 %) could not be verified for one of the reasons indicated in Table 12.3. Thus, we could check the validity of the linkage for 403 individuals. Our verification of the automatically generated links using the marriage registers as illustrated in Fig. 12.1 suggests that the automatic linkage has produced satisfactory results: among the sample of 403 verifiable links, 73 % are accurate (Table 12.3). However, the percentage of false links is also considerable (27 % or 20 % when considering all 533 verified cases). Moreover, considering that the majority of the population in Quebec married—i.e. that a marriage certificate should exist for most individuals—, the fact that 107 (20 %) out of 533 cases were not verifiable suggests the need of further research in order to understand the high percentage of missing records.

Table 12.3 shows that to validate a link it is necessary to find information that demonstrates the connection between the 1852 and the 1881 censuses. This proof

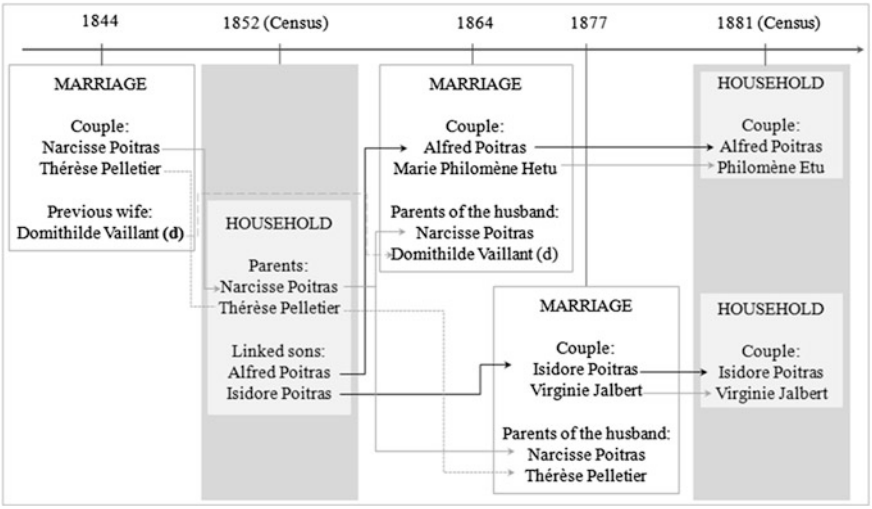


Fig. 12.1 Example of link validation via manual linkage between the 1852–1881 linked panel and the BALSAC mariage registers for Quebec. *Sources* BALSAC marriage registers and 1852–1881 linked sample. The letter "(*d*)" for "deceased" indicates the survival status of an individual at the moment of the marriage. In our example, Domithilde Vaillant died before the marriage of her son, Alfred. She is indicated as deceased on the marriage register of her son. We located the marriage certificate of the second marriage of Narcisse, Alfred's father, where we see the name of his second wife, Thérèse, who is observed in 1852. This same certificate also confirms the name of Alfred's mother, Domithilde

Table 12.3 Validation of automatically generated links between 1852 and 1881 using the BALSAC marriage registers, Catholic population of Quebec

Results	*N*	%[a]
Failure	*109*	**27.05**
Different spouse according to the 1881 census (1852 ⟶ M ⟶̸ 1881)	62	
Another individual with the same name is married with the woman that appears as spouse in 1881 (1852 ⟶̸ M ⟶ 1881)	47	
Success	*294*	**72.95**
The linked individual has the same spouse in 1881 (1852 ⟶ M ⟶ 1881)	264	
The linked individual remained single or became a widower between 1852 and 1881. Some family members who were present in 1852 are still present in 1881 (1852 ⟶ 1881)	30	
Non-verifiable	*130*	
No marriage register found and no other family members linked	107	
The linked individual did not live with his parents in 1852	18	
More than one marriage register found (common names)	5	
Total	*533*	
Total (verifiable)	*403*	

Sources 1852–1881 linked sample and BALSAC marriage registers
[a]% of verifiable cases; *M*: marriage register; 1852 and 1881: censuses

could be via the marriage register, as shown in Fig. 12.1, or through the presence of other family members in both census years, as shown in the "success" section of Table 12.3. Failures can be identified when the chain 1852-M-1881 is broken. Finally, in some cases, there is not enough information to establish whether a link is a failure or a success; these are the "non-verifiable" cases. This verification exercise suggests that most of the automatic links from 1852 to 1881 are accurate. However, as previously pointed out, our verification of the automatic links using the marriage registers is limited to Catholic individuals who married in a Quebec parish.

To explore the representativeness of the 4226 automatically generated links more generally, we present Table 12.4, which shows the results of a logistic regression analysis on the predictors of being automatically linked from 1852 to 1881. For this regression, the independent variables are: having a common name, living in a frontier district in 1852 (i.e. close to the border with the United States), being a farmer's son, going to school in 1852, the age group in 1852, the place of birth, the type of residence in 1852 and the fact of living in a household where the head was a labourer in 1852.

The results of the logistic regression suggest that the linkage was favoured by some characteristics, such as being older than 5 years in 1852, being a farmer's son, attend to school in 1852 and being born in Quebec or in England. Apart from the results concerning the place of birth, the results relating to each of the other attributes are as expected. Indeed, considering that in the context of our study mortality is considerably high in the first years of life, the proportion of survivors in 1881 (among the boys aged 0–15 in 1852) should be the lowest among those aged 0–5 years in 1852. This is what we have observed in our mortality and survival estimations (Table 12.1): 73.6 % of the boys aged 0–5 years in 1852 are expected to have survived until 1881 whereas the corresponding percentage for the boys who were in older age groups in 1852 is higher.

As to being a farmer's son, some studies have stressed the fact that the geographical stability of farmers is favourable for data linkage (e.g. Gagnon and Bohnert 2012; Dillon 2002). Moreover, farmers might have had lower mortality levels compared to individuals in other socio-economic groups (Gagnon et al. 2011). For our purposes, both the lower mortality and the geographic stability suggest that the likelihood of finding the corresponding record in 1881 is higher among farmers than among individuals in other socio-economic groups.

Concerning school attendance in 1852, it is possible that school-attending boys lived in households with other educated household members who in turn were more likely to provide accurate information on the census, favouring the linkage of that boy. We note, however, that the information about school attendance provided in the 1852 census manuscripts is quite limited: according to the 1852 census enumerator instructions, "By the words "attending school", not only those actually attending school at the time, but those who usually attend during some or any portion of the year, are meant to be included" (Gagan 1974, p. 360). Thus, our education variable identifies individuals who went to school at any time of the year, regardless of the amount of time spent at school. Questions also remain about the subgroup of boys who did not report school attendance. School attendance might

Table 12.4 Logistic regression: probability of being automatically linked, boys aged 0–15 years in the 20 % sample of the 1852 Canadian census

Variables	Odds ratio	
Common name[a]	0.315	***
Residence in a frontier district	0.925	*
Age group		
0–5 (*ref*)		
6–10	1.084	†
11–15	1.180	***
Farmer's son	1.272	***
School attendance in 1852	1.133	**
Birth place		
Canada (province not specified) (*ref*)		
Quebec	1.230	***
Ontario	1.052	
England	1.547	***
Ireland	0.787	**
Scotland	0.828	
Other	1.127	
Unknown	0.127	***
Type of place of residence		
Rural (*ref*)		
Village	1.052	
City	1.088	
Labourer household head	0.935	
N = 57023		

Source 1852–1881 linked sample

***p < 0.001, **p < 0.01, *p < 0.05, †p < 0.1

[a]We identified the 10 most frequent family names in each province among boys aged 0–15 years in the 20 % sample of the 1852 census. The top ten family names among these boys are (1) in Quebec: Coté, Tremblay, Gagnon, Roy, Morin, Ouellet, Gauthier, Boucher, Belanger and Demers and (2) in Ontario: Smith, McDonald, Campbell, Brown, Miller, Johnson, Scott, Wilson, Thompson and Taylor. In Quebec, 6.2 % of the boys of the subpopulation of interest have one of these common names whereas the corresponding percentage for the boys of Ontario is 5.7 %

have been lower among boys whose parents were labourers (Thernstrom 1973). However, Table 12.4 indicates that living in a household where the head was a labourer in 1852 does not significantly affect the chances of being automatically linked.

Regarding the place of birth, we observe that the reference category is being born in "Canada" (province not specified). In a separate exercise (not presented here) we looked at the distribution of the subpopulation of interest and of the linked sample by birthplace. We noticed that the share of individuals born in Canada (whether in Quebec, in Ontario, in "Canada" or in other parts of Canada) is higher among the latter (90 %) than among the former (84 %). This implies that the fact of

being native Canadian increased the chances of being automatically linked. In particular—and surprisingly—, the probability of being automatically linked seems to have increased significantly with the fact of being born in Quebec (odds ratio significant at the 0.1 % level): the proportion of individuals born in Quebec is indeed higher in the linked sample (34 %) than in the corresponding group of boys in the population in 1852 (28 %). As previously mentioned, this result is rather surprising, since the greater homogeneity of last names among French-Canadians compared to Canadians of other origins should have diminished the chances of linking individuals born in Quebec (who were mainly of French-Canadian origin). In Canada, the stock of French surnames is indeed more limited than the stock of English origin surnames, since the French-Canadian population is descended from basically 10,000 French immigrants who arrived in Quebec before the 1760s. In contrast, the regular immigration to Canada of people from the British Isles during the nineteenth century nourished the pool of English last names (Charbonneau et al. 2000). This implies that the likelihood of finding more than one individual with the same name might be higher among French than among English Canadians. In our case, the share of boys of the subpopulation of interest in 1852 who had one of the common last names indicated in Table 12.4 is higher among those who lived in Quebec (6.2 %) than among those who lived in Ontario (5.7 %). Moreover, it has been suggested that, compared to individuals of English origin, French-Canadians were more often in lower socio-economic strata and had a lower school participation—for example in the city of Montreal (Gauvreau and Olson 2008). Thus, if French-Canadians were less educated and if they had more homogeneous last names compared to other Canadians, one would expect them to be under-represented on the linked sample. In their analysis of an automatically linked sample between the Canadian censuses of 1871 and 1881, Antonie et al. (2015) mention that married French-Canadians are among the under-represented groups. The linkage technique used in their study is the same as the one employed in the creation of our linked sample from 1852 to 1881. Thus, the reason why boys born in Quebec—who are mostly of French-Canadian origin—were favoured in the automatic linkage procedure between 1852 and 1881 is not completely clear yet. One possible explanation that needs further research is that, due to language and cultural barriers, French-Canadians in our subpopulation of interest might have been less likely to emigrate to the United States compared with their English counterparts.

An additional consideration about the place of birth concerning the accuracy of the data is worth mentioning here. As stated earlier, the reference category of this variable in Table 12.4 is being born in “Canada” (province not specified). In the 1852 census, most native Canadians provided rather vague information about their place of birth, since they indicated only their country of birth without specifying the province. Moreover, contrary to the 1881 census, the 1852 census did not include a specific question about the origin of the individuals. In our case, 42 % of the linked individuals from 1852 to 1881 were born in “Canada” (province not specified). The corresponding percentage within the subpopulation of interest is 42.4 %. By

visualizing the 1852 census manuscripts, we could identify that some enumerators wrote a letter, "b" or "f", next to the mention "Canada" as place of birth. In the PRDH, we recently discovered that these letters could be an indicator of the origin of the individuals born in Canada: the letter "f" indicates the French-Canadian origin of an individual whether a "b" might indicate the British (or English) origin of a person. Indeed, in almost all cases, the letters "f" and "b" that accompany the mention "Canada" as place of birth correspond to individuals whose last name is of French or of English origin, respectively. Moreover, the content of several census pages suggests that the enumerators did sometimes fill the column Place of Birth with mentions relative to the cultural origin of the individual, such as "French-Canadian", "Br. Canadian", "Irish", etc.

In order to take the previous considerations into account, we created two new birthplace codes in the 20 % sample of the 1852 census, namely "Canada French" and "Canada English" (province of birth not specified in both cases). In the entire 20 % sample of the 1852 census, the proportion of individuals identified with "f" or with "b" as well as with the corresponding birthplace codes is about 10 % and 1 %, respectively. In particular, the code "Canada English"—which corresponds to mentions where there is a "b" accompanying the place of birth or where the place of birth includes an indication of the British or English origin of an individual—aims to correct a previous interpretation of the letter "b" regarding the place of birth. Initially, strings such as "Canada b" and "b Canada" were coded as born in Quebec, as the "b" was probably associated with "Bas" (prior to 1841, Quebec was known as *Bas Canada*, which means Lower Canada). However, we have considered the possibility that the letter "b" may not always mean "Bas", e.g. when the enumerator was Anglophone. We believe that the new codes better document the mentions inscribed on the census manuscripts, indicating the cases in which a specific province of birth cannot be attributed, and the cases in which "f" and "b" suggest the cultural origin of an individual rather than a place of birth. For linkage purposes, these codes could be useful in the stage of manual verification of automatically linked individuals. For example, as previously mentioned, the 1881 census contains a direct question about the cultural origin of the individuals.[6] Thus, during the stage of manual verification of the links from 1852 to 1881, the indicator of cultural origin based on the letters "f" and "b" could be compared with the corresponding answer in the 1881 census.

Back to the discussion about the representativeness of the linked sample, we have so far treated of the factors that might have increased the chances of being automatically linked from 1852 to 1881. However, some other characteristics seem to have diminished the chances of being automatically linked. Such characteristics are the fact of having a common name, of living in a frontier district and of being

[6]According to the instructions to the enumerators of the 1881 census "Origin is to be scrupulously entered, as given by the person questioned; in the manner shown in the specimen schedule, by the words English, Irish, Scotch, African, Indian, German, French, and so forth" (Department of Agriculture (Census Branch), 1881, p. 30).

born in Ireland. These attributes have an odds ratio significantly lower than 1 (Table 12.4). Regarding the fact of having a common name, we did a separate analysis (not presented here) in order to know whether having a common name was associated with some especial socio-economic characteristics. This analysis suggests that the main differences between the boys who had a common name (i.e. one of the surnames indicated in Table 12.4) and those who did not are the fact of living in a household where the head was a farmer in 1852 and the fact of being born in Quebec: having a common name is more frequent among farmers and among individuals born in Quebec. Since individuals born in Quebec and farmers are overrepresented in our linked sample, we can say that the fact of having a common name did not introduce bias in our linked sample. This bias would have occurred if the individuals with more common names had been under-represented in the linked sample, which is not our case.

As to being born in Ireland, this characteristic seems to have significantly diminished the chances of being linked: the percentage of individuals born in Ireland is indeed higher among the subgroup of boys in the population in 1852 (4.4 %) than in our linked sample (3.5 %). Some studies suggest that the Irish living in North America during the second half of the nineteenth century were overrepresented among the manual labourers, who constituted the lowest and more vulnerable socio-economic group, especially in the cities (Gaurvreau and Olson 2008; Katz 1975; Thernstrom 1973). In our case, more than 25 % of the boys born in Ireland in our subpopulation of interest lived in a household where the head was a labourer in 1852. The corresponding percentages for the boys of other origins vary between 10 and 20 %. Despite that in our subpopulation of interest in 1852 boys born in Ireland lived more frequently in the house of a labourer compared to boys of other origins, we observe in Table 12.4 that the negative impact of being born in Ireland on the chances of being automatically linked persists even after controlling for the fact of living in a household where the head was a labourer in 1852 (which is not significant). We note that among the boys born in Ireland who were recorded in the Canadian census of 1852, some might have been migrants who fled from their native country due to the potato famine. These immigrants were particularly vulnerable and lived in unstable conditions in North America (Crowley et al. 2012). Thus, it is possible that mortality was higher among these boys, diminishing their probability of being present in 1881.

In short, despite that the linkage technique aims to increase the validity and the representativeness of the links, the previous analyses suggest that our linked sample from 1852 to 1881 is not representative of some subgroups of the population of interest. On the one hand, being an immigrant (particularly from Ireland), having a common name and living in a district close to the border with the United States are characteristics that decreased the chances of being automatically linked. On the other hand, being older than 5 years in 1852, being native (especially form Quebec), being a farmer's son and attending to school are characteristics that increased the chances of being linked.

One very important issue about representativeness which we have not yet discussed concerns the population by type of place of residence in 1852. Table 12.4

shows that, for the purposes of the automatic linkage, there is no significant difference between living in a rural place, in a village or in a city in 1852. Indeed, the distribution of the linked individuals by type of place of residence in 1852 is very similar to that of the boys in the subpopulation of interest in 1852: both lived mainly in a rural area (92 %), 2.5 % lived in a small village (up to 2999 inhabitants), 1.5 % lived in a big village (3000 or more inhabitants) and 4 % lived in a city.[7] Thus, our linked sample is representative of the subpopulation of interest regarding the type of the place of residence in 1852. However, the subpopulation of interest is composed by individuals who are included in the 20 % sample of the 1852 census, which is affected by the absence of one third of the records: the census manuscripts covering 34 % of the population disappeared before being microfilmed (Dillon and Joubert 2012). Most of those missing manuscripts contained the records of the urban population: in Ontario (Upper Canada), the records of the cities of Toronto, Kingston, London and the big district of Simcoe are missing; in Quebec (Lower Canada), the records of Montreal are lost, except for those of the neighbourhood of St. Louis.

According to Dillon and Joubert (2012), 9.3 % of the population enumerated in the provinces of Ontario and Quebec in 1852 lived in the cities of Montreal, Toronto, Quebec, Hamilton, Kingston, Bytown and London. Due to the loss of most of the urban manuscripts mentioned above, the corresponding percentage in the 20 % sample of the 1852 census is of 4.8 %. Fortunately, the data of some cities, namely Bytown (Ottawa), Hamilton and Quebec, has been preserved. This data can be used in order to increase the representation of the urban population in the 20 % sample of the 1852 census: by attributing weights to the population living in those cities in 1852, some aspects of the urban population of mid-nineteenth century Canada can be analyzed. A weight variable is already available on the 20 % sample. This variable is based on the distribution of the population by type of place of residence according to the volume of aggregated statistics of the 1852 census (Board of Registration and Statistics 1853). It gives more weight to the population living in the cities of Quebec, Hamilton and Bytown in 1852, so that they constitute 9.3 % (instead of 4.8 %) of the total population of the two provinces. For example, each one of the individuals included in the 20 % sample who lived in Hamilton or in Bytown in 1852 has a weight of 2.8. The corresponding weight for their counterparts living in Quebec city in 1852 has a value of 1.7.

Despite the availability of some urban data, and even if the majority of the Canadian population in the mid-nineteenth century was rural, the missing records are problematic for any linkage procedure at the individual level, since most of the people living in the big cities in 1852 will be excluded. Attributing weights is not an optimal solution to this problem, since the population living in the biggest cities, i.e. Toronto and Montreal, very likely differed in some aspects from the population living in smaller cities, such as Hamilton, Bytown or Quebec. For example, the

[7]According to the type of place indicated in the aggregated volume of the 1852 census (Board of Registration and Statistics 1853).

ethnocultural composition and the economic opportunities—which are factors that have an impact on the intergenerational social mobility—seem to have been different in the biggest cities, on the one hand, and in the smaller cities, on the other hand. According to the preserved volume of aggregated statistics of the 1852 census (Board of Registration and Statistics 1853) the share of individuals of French-Canadian origin was higher in Quebec city (58.3 %) than in Montreal (45 %) in 1852. In turn, the share of the people from Ireland and from Scotland was more important in Montreal (25.8 %) than in Quebec city (16.6 %). In the Ontarian cities, the share of French-Canadians was minimal in Toronto and Hamilton, whereas in Bytown one-quarter of the population was of French-Canadian origin. Regarding the economic opportunities, looking at the occupational distribution of men aged 18–65 years in the complete census database of 1881 gives us an idea of the economic opportunities in the same five cities. In 1881, the share of merchants, manufacturers and professionals was more important in Montreal and Toronto (around 23 %) than in Ottawa, Hamilton and Quebec city (around 17 %). The opportunity to have an occupation on the manual skilled sector seems to have been the highest in the city of Hamilton, where 44 % of the men aged 18–65 years were skilled workers in 1881 (the corresponding share in Montreal and Toronto was around 35 and 38 %, respectively). In Ottawa, the share of men employed in white collar occupations was the highest (around 17 %), whereas the corresponding part in the other four cities was around 10 %. Though these observations are based on the occupational distribution of 1881, they provide an idea of the differences that might have existed among the five cities compared regarding the development of certain economic sectors as well as the occupational opportunities for the individuals living in those cities.

The previous considerations suggest that Hamilton, Bytown and Quebec city differed in some aspects from Toronto and Montreal. For this reason, using the weights provided in the 20 % sample of the 1852 census in order to increase the representation of the urban population needs caution in our interpretations. Moreover, for automatic linkage purposes, the absence of most of the data of the population living in the biggest cities in 1852 means that the urban individuals available for linkage will represent only certain selected cities, leaving Toronto and Montreal under-represented. Thus, when using the 20 % sample of the 1852 census, one should consider whether the missing urban data constitutes a problem to the analysis of a phenomenon of interest. For example, the missing urban data should not be problematic in the analysis of the rural exodus, which could be studied by linking the census sample of 1852 with the complete enumeration of 1881. It should be kept in mind that the urban population of Canada in 1852 constituted only 10 % of the population, and many of the missing manuscripts are distributed across communities of varying sizes.

12.4 The Use of the 1852 Census to Study the Intergenerational Social Mobility

Despite the difficulties mentioned above regarding the use of the Canadian census of 1852, this source of data includes valuable information about the population who lived in the provinces of Ontario or Quebec at the mid-nineteenth century. The 20 % sample can be used to analyze several aspects of this population, since it includes information about family composition and coresidence[8] as well as about the socio-economic conditions of individuals and households. For instance, the 1852 census includes questions about the occupation and the type of dwelling (e.g. log house, stone house, shanty, etc.). Moreover, compared to other data sources (e.g. the parish registers), the census data provides more details about the composition and the socio-economic characteristics of households. In our case, the aim of linking individuals from 1852 to 1881 was to have an opportunity to study the intergenerational social mobility between fathers (in 1852) and sons (in 1881). Using linked census data is favourable to study this phenomenon, since the information about the socio-economic characteristics of the family of origin in 1852 is of great interest when one wants to analyze the intergenerational social mobility and the occupational attainment of the sons in 1881. Moreover, despite the long intervening time span of 29 years between the linked censuses, linking the sons (aged 0–15 years in 1852) from 1852 to 1881 increases the chances of observing the father and the son living together in 1852. This coresidence is essential in order to have information about the father in 1852.

The use of the 1852 census is also supported by the subject of study. If one is interested in studying the Canadian population of the mid-nineteenth century, the use of the 20 % sample of the 1852 census is appropriate when the difficulties associated with the data do not hinder the subject of study. For instance, if one wants to study the rural exodus, the 20 % sample is appropriate, whereas it would not be the case if the subject of study were the urban life in Canada in the mid-nineteenth century. In our case, we are interested in the intergenerational social mobility at the beginning of the industrialisation in Canada. For this purpose, we need to observe the father (at the mid-nineteenth century) and the son (some decades later) when they were adults in order to compare their occupations at similar points in their careers. Some researchers (e.g. Prandy and Bottero 2000; de Sève and Bouchard 1998; Delger and Kok 1998; Van Poppel et al. 1998) have criticized

[8]The 20 % sample of the 1852 census includes the variable "Household number" but not "Family number". Thus, we can identify who lived with whom (in the same household) but not who belonged to which family in 1852. In order to have an idea of the different families that lived together, the 20 % sample includes some variables that aim to identify the relationship between individuals living in the same household. For example, the constructed variable CANREL indicates the relationship with respect to the household head (e.g. "wife of head", "child of head", "parent of head", "other kin of head", and "undetermined"). The variables MOMLOC and POPLOC indicate, within each household, the position (in order of enumeration) of the mother and the father of an individual, respectively.

the use of marriage registers as only source of data in the study of the intergenerational social mobility: "Studies using marriage records are obviously comparing, for the most part, fathers (and/or fathers in law) who are nearing the end of their working lives with sons who are at a fairly early stage in theirs" (Prandy and Bottero 2000, p. 4). This difference of age between fathers and sons increases the risk of overestimating downwards the social mobility. Thus, by linking individuals from different censuses, this risk can be reduced, since the comparison of fathers and sons can be made at more similar points in their respective occupational careers.

However, for the purposes of studying the intergenerational social mobility, the use of the 1852 census implies that we have to take care with our interpretations, since the mobility observed does not include the sons of the fathers who lived in the biggest cities (Toronto and Montreal) in 1852. This means that the observed mobility would concern mostly the rural population, which constituted the majority (90 %) of the Canadian population in 1852 (Dillon and Joubert 2012). Moreover, the occupation is the only variable available in the 1852 census that can be used in order to have an idea of the social status of an individual. The Canadian censuses started to include more questions about employment—e.g. the employment status (employer/employee) and the fact of earning a salary—only in 1891 (Baskerville 2000). Thus, before the 1891 census, the only information available in the personal censuses regarding the social status of the individuals is their occupation. Other types of census questionnaires containing economic information existed already before 1891. Some examples are the agricultural schedule of 1852 and the industrial return of 1871 (LAC 2014). However, most of the information contained in those questionnaires is not immediately available for research, since the manuscripts that have been preserved have not been transcribed yet (with the exception of a sample of the agricultural and the industrial return of 1871).

In short, despite the difficulties associated with the use of the Canadian census of 1852, this source of data provides valuable information about the socio-economic conditions as well as the household composition of the population living in Ontario or in Quebec at the mid-nineteenth century. This data is suitable for the study of certain phenomena as well as for data linking with other nominative sources of data, such as the complete 1881 Canadian census.

12.5 Conclusion and Discussion

This chapter has briefly described and discussed some aspects about an automatically generated linked sample from the Canadian censuses of 1852 and 1881. The linked sample, composed by males aged 0–15 years in 1852, was created with the specific purpose of studying the intergenerational social mobility between fathers and sons at the beginning of the industrialisation in Canada. In order to analyze this linked sample, we have briefly described the technique by which it was generated. This technique of automatic linkage, based on the individual attributes that should

not change over time (or that should change in a predictable way), aims to favour the representativeness and the accuracy of the linked sample by reducing the number of false links. We have presented some analyses about the accuracy and the representativeness of our linked sample from 1852 to 1881: on the one hand, our verification of the automatic linkage via the use of marriage registers suggests that most of the links are accurate. On the other hand, our analyses regarding the representativeness of the linked sample suggest that some attributes favoured while others hindered the fact of being automatically linked from 1852 to 1881: immigrants (particularly from Ireland), individuals with common names and those living close to the border with the United States had fewer chances to be automatically linked, while the native (especially from Quebec), older than 5 years in 1852, farmer's sons and attending to school had more chances to be linked. Thus, the linked sample is not representative of some subgroups of the population (here, the "population" is composed by the boys aged 0–15 years in 1852 who are included in the 20 % sample). The identification of the under-represented groups is important in order to be careful in the interpretations of a study using this linked sample.

As a final consideration, we would like to put emphasis on the great research potential of linking censuses and parish registers. In our case, we used the BALSAC marriage registers in order to verify the accuracy of the automatically generated links between the censuses of 1852 and 1881. Besides being appropriate for this purpose, the marriage registers could be used to add more information about the linked individuals from 1852 to 1881. Indeed, though the automatically linked sample 1852–1881 provides valuable information about the composition and the socio-economic conditions of individuals and households in 1852 and 1881, it does not include information about the demographic events, such as the age at marriage of the individuals. This information could be added by linking the marriage registers to the linked sample between censuses, such as illustrated in Fig. 12.1. Such an approach would not exclude individuals who did not marry during a determined interval of time, since their information would be available on the censuses. We note, however, that in our case, the BALSAC marriage registers are limited to the catholic individuals who married in a parish in the province of Quebec while the linked sample from 1852 to 1881 includes individuals with other religious affiliations who married outside the province of Quebec between 1852 and 1881.

Our exercise of manual linkage between the automatically linked sample (1852–1881) and the BALSAC marriage registers is not currently connected to any other project in Quebec. However, researchers could profit from the linkage between these registers and the Canadian censuses, since both data sources contain information at the individual level and provide information about complementary aspects such as family composition and socio-economic conditions (the censuses) and the demographic events (the parish registers).

References

Antonie, L., Inwood, K., & Ross, A. J. (2015). Dancing with dirty data : Problems in the extraction of life-course evidence from historical censuses. *This book*, chapter 10.

Balsac, fichier de population (accessed 2013) [online]. http://balsac.uqac.ca/. The BALSAC marriage registers were accessed via the website of the *Registre de la Population du Québec Ancien* (RPQA): http://www.prdh.umontreal.ca/rpqa/.

Baskerville, P. (2000). Displaying the working class: The 1901 census of canada. *Historical Methods, 33*(4), 229–234.

Bord of registration and statistics. (1853). *Census of the Canadas* (pp. 1851–1352). Quebec: J. Lovell.

Boubeau, R., Légaré, J., & Émond, V. (1997). *Nouvelles tables de mortalité par génération au Canada et au Québec, 1801–1991*. Ottawa: Statistics Canada, Demography division.

Charbonneau, H., Desjardins, B., Légaré, J., & Denis, H. (2000). The population of the St. Lawrence Valley, 1608–1760. In M. R. Haines & R. H. Steckel (Eds.), *A population history of North America* (pp. 99–142). New York: Cambridge University Press.

Crowley, J., Murphy, M., Roche, C., & Smyth, W. (Eds.). (2012). *The scattering, in Atlas of the Great Irish Famine*. New York: New York University Press.

Curtis, B. (2001). *The politics of population: State formation, statistics, and the census of Canada, 1840–1875*. Toronto: University of Toronto Press.

De Sève, M., & Bouchard, G. (1998). Long-term intergenerational mobility in Quebec (1851–1951): The emergence of a new social fluidity regime. *Canadian Journal of Sociology/Cahiers Canadiens de Sociologie, 23*(4), 349–367.

Delger, H., & Kok, J. (1998). Success or selection? The effect of migration on occupational mobility in a Dutch province, 1840–1950. *Historical Methods, 13*(3–4), 289–322.

Dillon, L. Y. (2002). Challenges and opportunities for census record linkage in the French and English Canadian context. *History and Computing, 14*(1–2), 185–212.

Dillon, L. Y., & Joubert, K. (2012). Dans les pas des recenseurs: une analyse critique des dimensions géographiques et familiales du recensement canadien de 1852. *Cahiers québécois de démographie, 41*(2), 299–339.

Gagan, D. (1974). Enumerator's instructions for the Census of Canada, 1852 and 1861. *Social History, 7*(14), 355–365.

Gagnon, A., & Bohnert, N. (2012). Early life socioeconomic conditions in rural areas and old-age mortality in twentieth-century Quebec. *Social Science and Medicine, 75*, 1497–1504.

Gagnon, A., Tremblay, M., Vézina, H., & Seabrook, J. A. (2011). Once were farmers: Occupation, social mobility, and mortality during industrialization in Saguenay-Lac-Saint-Jean, Quebec 1840–1971. *Socioeconomic Inequalities and Mortality, 48*(3), 429–440.

Gauvreau, D., & Olson, S. (2008). Mobilité sociale dans une ville industrielle nord-americaine: Montréal, 1880–1900. *Annales de démographie historique, 115*(1), 89–114.

Katz, M. B. (1975). Transiency and Social Mobility. *The people of Hamilton, Canada West: Family and class in a mid-nineteenth-century city* (pp. 94–175). Cambridge: Harvard University Press.

Library and Archives Canada (LAC). (2014). Censuses. Online: http://www.bac-lac.gc.ca/eng/census/Pages/census.aspx.

Long, J. (2005). Rural-urban migration and socioeconomic mobility in Victorian Britain. *The Journal of Economic History, 65*(1), 1–35.

Prandy, K., & Bottero, W. (2000). Reproduction within and between generations. *Historical Methods, 33*(1), 4–15.

Roberts, E. (2012). Mining microdata: Economic opportunity and spatial mobility in Britain, Canada and the United States, 1850–1911. In *Digging into data challenge, National Science Foundation, Economic and Social Research Council (UK), and Social Science and Humanities Research Council (Canada).* Project description.

Ruggles, S., Alexander, J. T., Genadek, K., Goeken, R., Schroeder, M. B., & Sobek, M. (2010). Integrated public use microdata series: Version 5.0 [machine-readable database]. Minneapolis : University of Minnesota.

Thernstrom, S. (1973). *Poverty and progress: Social mobility in a nineteenth century city*. New York: Atheneum.

Van Poppel, F., de Jong, J., & Liefbroer, A. C. (1998). The effects of paternal mortality on sons' social mobility: A nineteenth-century example. *Historical Methods, 31*(3), 101–112.

Routledge
Taylor & Francis Group

Historical Methods: A Journal of Quantitative and Interdisciplinary History

ISSN: 0161-5440 (Print) 1940-1906 (Online) Journal homepage: http://www.tandfonline.com/loi/vhim20

Standardising and Coding Birthplace Strings and Occupational Titles in the British Censuses of 1851 to 1911

Kevin Schürer, Tatiana Penkova & Yanshan Shi

To cite this article: Kevin Schürer, Tatiana Penkova & Yanshan Shi (2015) Standardising and Coding Birthplace Strings and Occupational Titles in the British Censuses of 1851 to 1911, Historical Methods: A Journal of Quantitative and Interdisciplinary History, 48:4, 195-213, DOI: 10.1080/01615440.2015.1010028

To link to this article: http://dx.doi.org/10.1080/01615440.2015.1010028

HISTORICAL METHODS, October–December 2015, Volume 48, Number 4

Routledge
Taylor & Francis Group

Standardising and Coding Birthplace Strings and Occupational Titles in the British Censuses of 1851 to 1911

KEVIN SCHÜRER
Centre for English Local History
University of Leicester

TATIANA PENKOVA
Institute of Computational Modelling
SB RAS

YANSHAN SHI
Department of Mathematics
University of Leicester

Abstract. This article presents a technique of standardising and coding textual birthplace and occupation strings in the censuses of England and Wales and Scotland, 1851–1911. While the approaches for the two text strings are different, they are both based upon the integration of computer technologies, mathematical methods, and expert knowledge. Both processes are described formally using Structured Analysis and Design Technique methodology. The classification of occupations is defined by two algorithms based on statistical decision theory in order to allocate codes from the original occupation strings. The method of standardising parishes is based on the comparison of original birthplace strings and reference data.

Keywords: birthplace standardisation, historic census data occupation coding

Introduction

Historic census data have long been used for examining a wide variety of themes, especially in relation to socioeconomic development and migration (e.g., Higgs 2005; Lawton 1978; Mills and Schürer 1996; Wrigley 1972). Despite the richness of nominal- and household-level census data as a source, a fundamental problem applies to the analysis of all historical census enumeration data, regardless of time or place: namely that the data were collected as textual responses. While some standardisation of the responses from householders may have been undertaken as part of the enumeration process, the resulting textual strings require some form of preprocessing before they can be interpreted and analysed. This point is clearly illustrated by the example of the decennial censuses undertaken in Great Britain between 1851 and 1911. The material from these censuses form the basis of the work described in this report,[1] which in turn forms part of a wider project to clean, standardize, and disseminate the nominal-level census data for wider academic use (Higgs et al. 2013; Schürer and Higgs 2014). Accumulatively, this collection of census data relates to 183,470,969 person records documented across twelve separate census enumerations. In total, these combined records gave rise to 7,304,708 different occupational descriptions or strings, and 6,703,779 unique birthplace strings. The lack of standardised responses in the nineteenth-century censuses meant that the variety of

Address correspondence to Kevin Schürer, c/o VC's Office, Centre for English Local History, University of Leicester, University Road, Leicester LE1 7RH, UK. E-mail: ks291@le.ac.uk
Color versions of one or more of the figures in this article can be found online at www.tandfonline.com/vhim.

responses for essentially the same answer were many and varied. Thus, a relatively straightforward occupation such as "watchmaker" was expressed as, for example, together with multiple other forms:

WATCH MACKER
WATCH MAKER & GREEN GROCER
WATCH MAKER (REPAIRER)
WATCH MAKER EMPLOYING 4 MEN & 3 BOYS
WATCH MAKER IN ALL BRANCHES
WATCH MAKER IN GENERAL
WATCH REPAIRER & MAKER
WATCH- MAKER
WATCHMAAKER
WATCHMAER
WATCHMAK
WATCHMAKER (MAS)
WATCHMAKER CLOCK
WATCHMAKER EMP 1 ASSIST 1 APPRENT
WATCHMAKER ETC
WATCHMAKER GENERAL
WATCHMAKER MASTER EMP 1 MAN + 1 BOY
WATCHMAKER MASTER EMPLOY 1 MAN
WATCHMAKER OUT OF EMPLOY
WATCHMAKER REPAIR
WATCHMAKER SPRINGER
WATCHMAKR
WATCHMALER
WATCHMEKER
WATCHMENDER
WATCHV MAKER
WATCHWORK
WATCJMAKER
WATCKMAKER
WATCMAKER
WATHCHMAKER
WATHCMAKER
WATHMAKER
WORKING WATCHMAKER
WORKING WATCHMAKER AND JEWELLER
WQTCHMAAKER
WTAHCMAKER
WTCHMAKER
WTCHMKR
WWATCH MAKER
WWATCHMAKER

Likewise, the parish of birth, Husband's Bosworth, in the county of Leicestershire, was expressed in multiple ways, including the following:[2]

H BOSWORTH|[BLANK]|[BLANK]
H.BOSTH|LEICESTER|ENGLAND
HBDS BOSWORTH|LEICESTERSHIRE|[BLANK]
HUSBANDS BODWORTH|LEICESTERSHIRE|[BLANK]
HUSBANDS BOSWORTH|[BLANK]|LEISTER
HUSBANDS BOSWORTH|[BLANK]|[BLANK]
HUSBANDS BOSWOR|LEICESTERSHIRE|[BLANK]
HUSBANDS BOSWTH|LEICESTER|ENGLAND
HUSBANDS BSWORTH|LEICESTERSHIRE|[BLANK]
HUSBANDS B|LEICESTER|ENGLAND
HUSBANS B|LEICESTERSHIRE|[BLANK]
HUSBARDS BOSWORHTH|LEICESTERSHIRE|[BLANK]
HUSBARDS BOSWORTH|LEICESTERSHIRE|[BLANK]
HUSBORNE BORWORTH|LEICESTERSHIRE|[BLANK]
HUSBOS BOSWORTH|LEICESTER|ENGLAND
HUSDS BOSWORTH|LEICESTER|ENGLAND
HUSLANDS HOSWORTH|LEICESTERSHIRE|[BLANK]
LEIC HUSBANDS BOSWORTH|[BLANK]|[BLANK]
LENTER HUSBANDS BOSWORTH|[BLANK]|[BLANK]
LESTER HBS BOSWORTH|[BLANK]|[BLANK]
LESTER HUSBANDS BOSWORTH|[BLANK]|[BLANK]
LESTERSHIRE HDS BOSWORTH|[BLANK]|[BLANK]
LEUICESTERSHIRE HUSBANDS RESIDENT BOS WOTH|LEICESTERSHIRE|[BLANK]
LICESTER HUSBANDS BOSWORTH|[BLANK]|[BLANK]
LTDS BOSWORTH|LEICESTERSHIRE|[BLANK]
LUSHAND BAWSWORTH|LEICESTERSHIRE|[BLANK]
RUSBAND BASWORTH|LEICESTERSHIRE|[BLANK]
RUSBUNDS BASWORTH|LEICESTERSHIRE|[BLANK]
[BLANK]|LEICESTERSHIRE|HASBORD BOSWORTH

The pattern of variability is incredibly heterogeneous. In the case of occupation titles, the original 115,247,642 individual census records with an entry for occupation[3] gave rise to 7,304,708 unique strings, of which 77.7% had a frequency of one. At the other end of the distribution curve, as illustrated by Figure 1, the five most common strings accounted for 22% of all individuals returning an occupation; the top ten strings for 27%; the top 100 for 47%; the top 1,000 for 66%; and the top 10,000 for 80%. Although, as one might expect, the very most common birthplace strings accounted for a smaller proportion of the population than in the case of occupations, a similar situation is found with the responses to the place of birth question (see Figure 2). Of the 6,703,779 separate birthplace strings, 70.2% had a frequency of one, while the five most common strings accounted for 4% of all individuals;[4] the top ten strings for 7%; the top 100 for 20%; the top 1,000 for 42%; and the top 10,000 for 72%.

With this volume of strings needing to be coded, full manual coding was simply not an option within the timeframe of the project. It is also the case that the high proportion of low frequency strings suggests that these are rather idiosyncratic in nature, thus making the task of coding all the more difficult. Consequently, automatic/semi-automatic processes for coding must be devised. For occupations, the task required was allocate each string to one of 797 predefined categories of an occupational

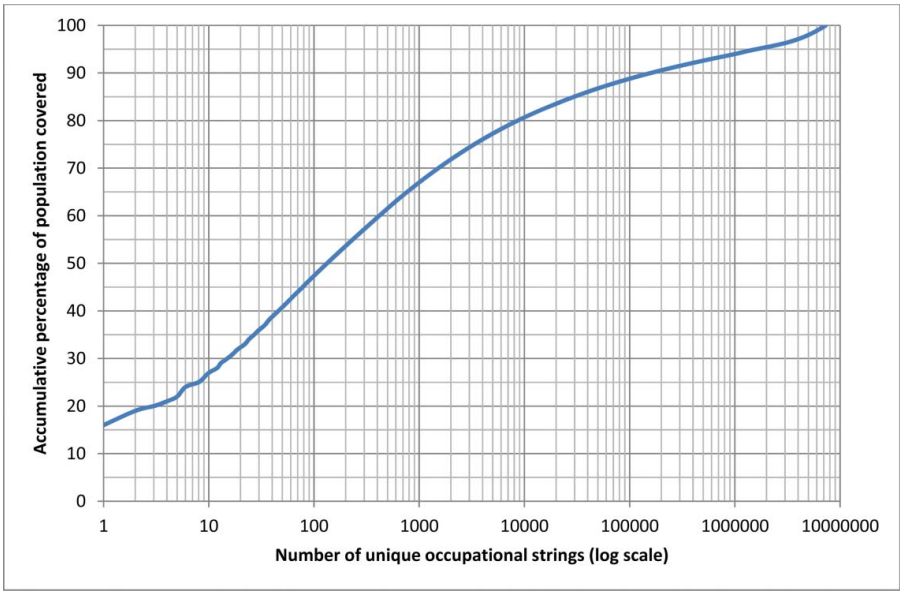


FIGURE 1. Distribution of occupational strings.

scheme that had been specially devised (referred to as the Matrix Classification) to enable cross-comparison between the various coding schema adopted by the census offices of England and Wales (a joint office) and Scotland between 1851 and 1911 (see http://www.essex.ac.uk/history/research/icem/documentation.html). Due to the nature of birthplaces, in this case the standardisation of the string presented a different set of problems and challenges in which the principal task was to assign each string initially to either an administrative county of England, Scotland, Wales, or Ireland OR to a country other than these OR to an indeterminate category. Secondly, those allocated to a British county were then assigned, where possible, to one of 17,453 predefined "standard parishes." Thus, for example, all the strings except two given above for Husband's Bosworth would have had the county code (CNTI) for Leicestershire (LEI) assigned to them. The string "H BOSWORTH|[BLANK]|[BLANK]" would have had a CNTI value of "UNK" assigned to it given no county level information was recorded. The string "MUSBANA BOSWORTH|NORTHAMPTONSHIRE|[BLANK]" would have been given the CNTI code of "NTH" for Northampton even though it is incorrect, since at this stage it would impossible to determine this without considering the parish level information. A string such as "NY|[BLANK]|UNITED STATES" would have been assigned the country code (CTRY) of "USA" and not considered further since standard parishes were only assigned for England, Wales, and Scotland.[5] Because of the different nature of the tasks, interrelated yet separate solutions were adopted for occupation titles and birthplace strings.

Processing Occupational Titles

The strategy for coding occupations was built around two preexisting factors. First, a comprehensive occupation coding dictionary created by Dr. Matthew Woollard for an earlier related project already existed for some 1.4 million occupational strings for 1881 (Schürer and Woollard 2000; Woollard 1999). Second, in the case of the occupational titles of England and Wales for 1911, a related Hollerith code[6] had been transcribed for most of the occupations for that year (Anderson 1988; Austrian 1982; Eames and Eames 1990; Higgs 1996, 2004; Schürer 1991). This meant that a number of strings could be mapped to the occupation coding scheme being applied by using the Matrix Classification. In addition, strings remaining uncoded with a frequency of fifteen or greater were manually coded in order to provide a reference authority list of 57,780 high frequency strings with which to compare the remaining 5,915,852 uncoded strings.

The technique adopted to automatically code the remaining strings was based on the integration and combination of computing technologies, mathematical methods, and expert knowledge. Figure 3 presents processes and stages in the form of an IDEF0 diagram. IDEF0 methodology provides a modelling function based on the language syntax of the Structured Analysis and Design Technique (SADT) (Davis 1994; Marca and McGowan 1987). An IDEF0 model describes the functions as a series of activities, actions, processes, or operations. In this case, the inputs and outputs can be seen as the data needed to perform the function and the data that is produced as a result of the function,

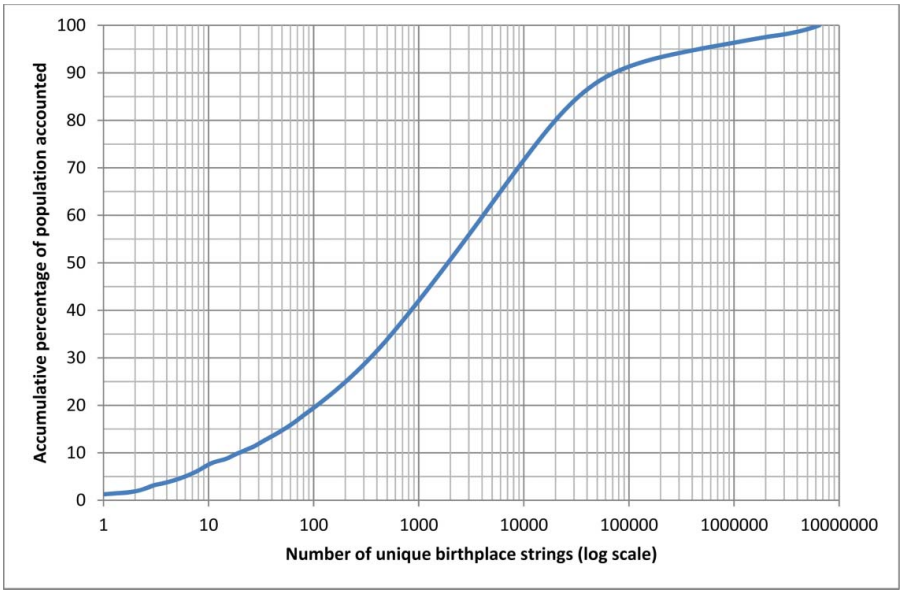


FIGURE 2. Distribution of birthplace strings.

respectively. Likewise, controls which constrain or govern the function can be thought of as the person or device which performs the function.

The occupation coding processes consisted of three basic steps: Step O1: Creating the authority lists; Step O2: Coding the occupation strings; and Step O3: Valuation of coding results. The first of these, O1: *Creating the authority lists*, is the process (already mentioned) of developing and modifying a dictionary to be used as an authority list for coding occupation the strings. Step O2: *Coding the occupation strings* is the process of allocating codes to the original occupation strings. This process included several substeps that were performed by specially written algorithms, and which forms the bulk of the work described here. The various substeps are discussed below. Step O3: *Evaluation of coding results* is the final process in which samples of the data were manually checked and tested in order that amendments to the algorithms could be made as necessary. This was an iterative process.

The various separate substeps of Step O2: *Coding the occupation strings*, are represented in detail in the IDEF0 diagram shown in Figure 4. These are as follows: O2.1: Cleaning the occupation strings; O2.2: Spell checking the occupation strings; O2.3: Identifying "non-work" strings; and O2.4: Coding the occupation strings. The first preprocessing stage (O2.1: *Cleaning the occupation strings*) was undertaken in order to delete characters other than A–Z, a–z (symbols, numbers, etc.) from a copy of the original string, together with what were defined as "null" words. These words were words within a given string which were not seen as adding any meaning or clarification relation to the occupation in question. These, typically, would be prepositions, institutional names, personal names, and geographical locations. For example, the title "A FARMER OF 500 ACRES" would be redacted to "FARMER ACRES," while "A PAINTER OF HOUSES" would be redacted to "PAINTER HOUSES," and "A TAILOR, LEICESTER" would become simply "TAILOR." In order to enable this stage, a list of key high frequency "null" words was constructed manually from a complete list of all words, by frequency, contained within the input strings.

The second preprocessing stage (O2.2: *Spell checking the occupation strings*) was aimed to check the composite words of the cleaned strings against those of the authority list and to "correct" for spelling variation. This process of spelling correction was based on combination of two well-know and highly used name-matching algorithms: SPEEDCOP and the Levenshtein algorithm. The SPEEDCOP algorithm was initially developed as a way of automatically correcting spelling errors (predominantly typing errors) in a very large database of scientific abstracts (Pollock and Zamora 1984). The correction algorithm uses a similarity key (the so-called skeleton key) that is constructed by concatenating the following features of the string (word or misspelling): the first letter, the remaining unique consonants in order of occurrence, and the unique vowels in order of occurrence. The rationale for this key is as follows: (a) the first letter keyed is likely to be correct; (b) consonants carry more information than vowels; (c) the original consonant order is mostly preserved; and (d) the key is not altered by the doubling or undoubting of letters, or most transpositions. Some examples of string/key pairs are given in Table 1.

The Levenshtein algorithm was used as it is one of the best known algorithms for measuring the similarity between

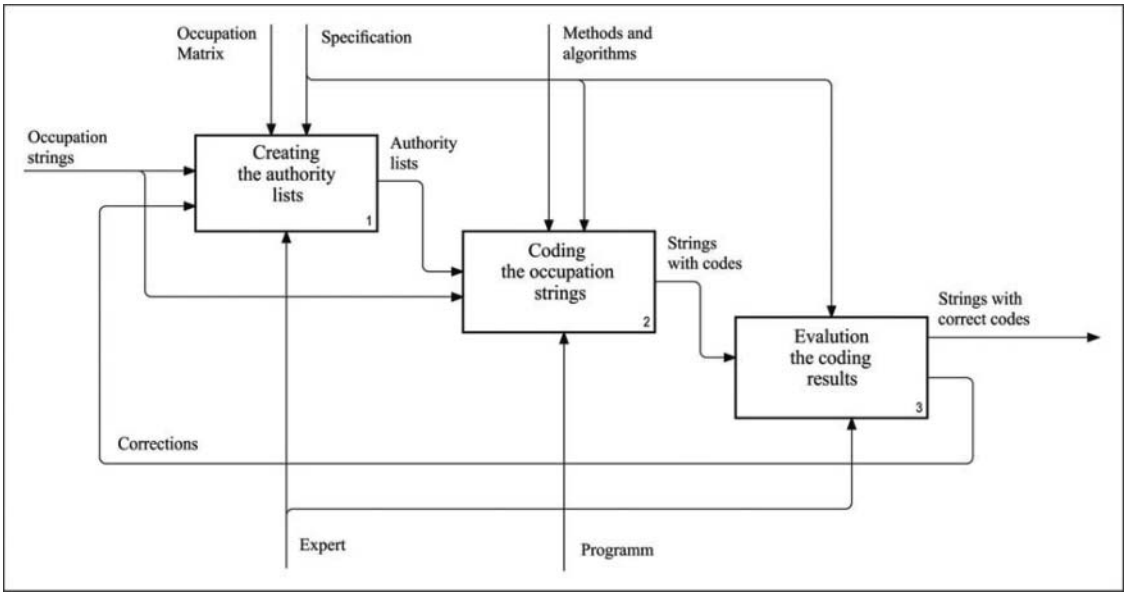


FIGURE 3. IDEF0 diagram of occupation coding processes.

two strings, and one which (unlike other existing algorithms that provide a measure of similarity) is embedded within several programming languages (Schierle, Schulz, and Ackermann 2008). In comparing two strings, the algorithm calculates an index of similarity (the edit distance), which is the minimum-cost sequence of operations on individual characters which are required to transform one string into another, in other words, the number of character substitutions, insertions, or deletions which are required to transform the first string into the second. The algorithm for constructing string-1 (the output string) from string-2 (the input string) and for computing the sum of the costs

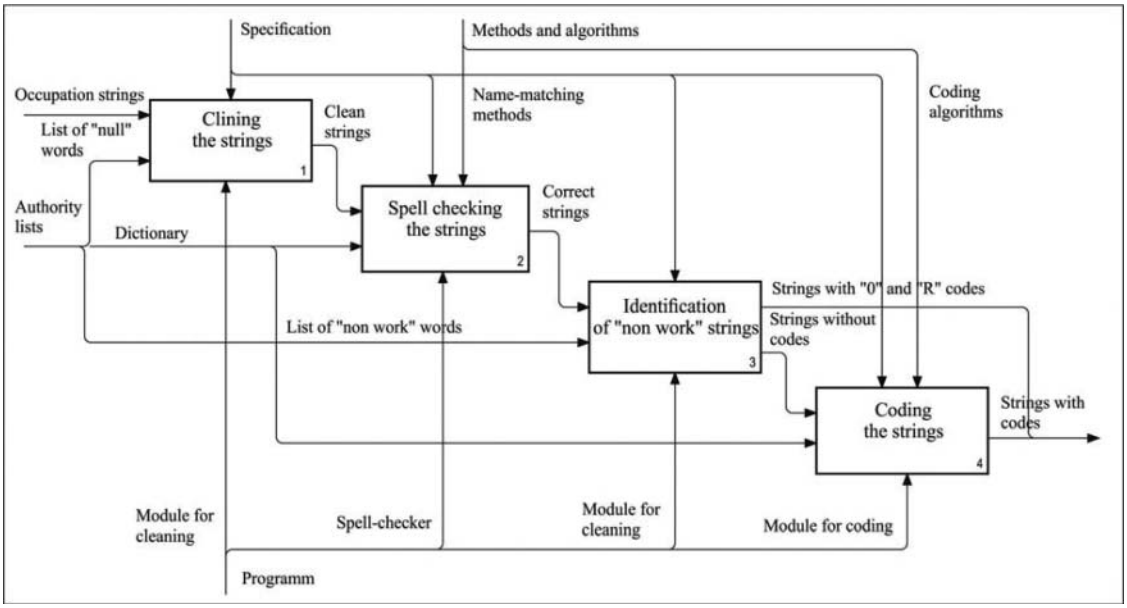


FIGURE 4. IDEF0 diagram of stage 2 of the occupation coding process.

TABLE 1. Example of SPEEDCOP Word Transformations

String	Skeleton key
WORKER	WRKOE
WORKEKR	WRKOE
WORKEER	WRKOE
WORKERP	WRKPOE
VORK	VRKO
WOKER	WKROE

involves a pointer that points to a character in string-2. An output string is constructed by a sequence of operations that might advance the pointer, add one or more characters to the output string, or both. Initially, the pointer points to the first character in the input string, and the output string is empty. The operations and their associated costs which were used are given in Table 2, and examples of the application of the Edit distance calculation are provided in Table 3. This algorithm demonstrated the good machining results with minimum computational burden.

Substep O2.3 (*Identification of* "*non-work*" *strings*) is aimed at identifying words within the occupational strings (in addition to the "null" words, see O2.1) which are of no direct relevance in the classification of the raw strings. These fall into one of two basic types: "Non work": The string contains a suggestion of current unemployed, such as UNEMPLOYED, FORMERLY, RETIRED, and so on; or "Relative": The string indicates that a familial relationship is present, such as WIFE, SISTER, DAUGHTER, MOTHER, WIDOW, NIECE, and so on. This step is important in the case of the first group since in the historic censuses being processed, some individuals would give their former employment even though they were no longer employed ("FORMERLY A BLACKSMITH," "BLACKSMITH, RETIRED"). In this exercise, we wished to classify any such string as though they were still working, adding a

TABLE 2. Levenshtein Edit Distance Operations

Operation	Cost	Operation description
APPEND	50	When the output string is longer than the input string, add any one character to the end of the output string without moving the pointer.
BLANK	10	Do any of the following: Add one space character to the end of the output string without moving the pointer. When the character at the pointer is a space character, advance the pointer by one position without changing the output string. When the character at the pointer is a space character, add one space character to the end of the output string, and advance the pointer by one position.
DELETE	100	Advance the pointer by one position without changing the output string.
DOUBLE	20	Add the character at the pointer to the end of the output string without moving the pointer.
FDELETE	200	When the output string is empty, advance the pointer by one position without changing the output string.
FINSERT	200	When the pointer is in position one, add any one character to the end of the output string without moving the pointer.
FREPLACE	200	When the pointer is in position one and the output string is empty, add any one character to the end of the output string, and advance the pointer by one position.
INSERT	100	Add any one character to the end of the output string without moving the pointer.
REPLACE	100	Add any one character to the end of the output string, and advance the pointer by one position.
SINGLE	20	When the character at the pointer is the same as the character that follows in the input string, advance the pointer by one position without changing the output string.
SWAP	20	Copy the character that follows the pointer from the input string to the output string. Then copy the character at the pointer from the input string to the output string. Advance the pointer two positions.
TRUNCATE	10	When the output string is shorter than the input string, advance the pointer by one position without changing the output string.

supplementary variable to indicate that they were no longer active (Higgs et al. 2013, 246). Equally, other individuals designated their occupation by referencing themselves via a third person ("WIFE OF A BLACKSMITH," "BLACKSMITH'S DAUGHTER"). It was important to identify such titles so as to ensure that they were not misleading classified to the occupational group of the third person. As with substep O2.1, these processes were undertaken with recourse to authority lists of "not working" and "relative" words constructed manually from a complete list of all words, by frequency, contained within the input strings. The net result was that strings falling into either of these two types were not passed onto the next coding stage.

Substep O2.4 (*Coding the occupation strings*) is a basic stage where the coding of the amended strings is carried out. This process is realised by two algorithms developed through iterative evaluation and testing based on statistical decision theory (Berger 1985). The outcome of these coding processes depends essentially on the statistical properties of the authority list of known coded occupational strings (the Dictionary). In addition to the raw strings, this Dictionary consists of the composite words with make up the string, the codes of occupation classification, and number of words in each code. The description of these two occupation coding algorithms is set out in the following sections.

First Occupation Coding Algorithm

The purpose of the first algorithm is to identify the significance of each word taken from the occupation strings associated with each code and in turn, to identify the most likely code for related to each word. This first coding algorithm consists of seven steps.

(1) Extract words from occupation string and form the set of words for each string: String $= \{w_1, w_2, \ldots, w_i, \ldots, w_n\}$.

(2) Calculate the weighting coefficient for each word:

$$\varpi_i = \sum_{k=1}^{K} \frac{n_{ik}}{N_k}, \tag{1}$$

TABLE 3. Examples of Edit Distance Calculation

String-1	Cost	String-2
WORKER	0	WORKER
WORKEKR	100	WORKER
WORKEER	20	WORKER
WORKERP	50	WORKER
VORK	220	WORKER
WOKER	100	WORKER

ϖ_i is a weighting coefficient of i-th word
n_{ik} is a number of i-th word in k-th classification code in the Dictionary
N_k is a total number of words in k-th classification code in the Dictionary
$k = \overline{1, K}$, K is a number of classification codes.

(3) *Calculate the coefficient of popularity for each word (less popular word is more important):*

$$\alpha_i = \frac{1}{n_i/N}, \tag{2}$$

α_i is a coefficient of popularity of i-th word
n_i is a number of i-th word in the Dictionary
N is a total number of words in the Dictionary.

(4) *Calculate the coefficient of order for each word (the first word in string is more important)*:

$$\beta_i = \frac{1/(i+k)}{\sum_{i=1}^{n} 1/(i+k)}, \tag{3}$$

β_i is a coefficient of order of i-th word, $\sum_{i=1}^{n} \beta_i = 1$
i is an order number of word in the string, $i = 1, 2, 3, \ldots, n$
n is a total number of words in the string
k is a rate of differences of the first word in the string from others, $k = 1$.

(5) *Calculate the Index for each word:*
Variant 1:

$$Index_i^1 = \varpi_i \cdot \alpha_i \cdot \beta_i, \tag{4}$$

Variant 2:

$$Index_i^2 = \varpi_i \cdot \alpha_i, \tag{5}$$

$Index_i$ is an Index of i-th word,
ϖ_i is a weighting coefficient,
α_i is a coefficient of popularity, and
β_i is a coefficient of order.

(6) *Identify the most significant word of string*:

$$w_R = \text{argmax}(Index), \tag{6}$$

w_R is the most significant word of string,
arg max is a word with maximum value of Index, and
Index is an Index of word.

(7) *Identify the classification code of string by significant word*:

$$C^* = \text{argmax}(n_{w_R}), \tag{7}$$

C^* is a classification code of string,
arg max is a classification code where the number of the most significant word is maximum, and
n_{w_R} is a number of the most significant word in the classification code.

Applying this initial algorithm, results in each string are assigned two occupation codes.

Second Occupation Coding Algorithm

The second algorithm was developed to find statistically the most appropriate code using all remaining valid words from the original occupation string based on the combination classified codes and their associated words. This second algorithm consists of seven steps.

(1) *Extract words from occupation string and form the set of words for each string:*

$String = \{w_1, w_2, \ldots, w_i, \ldots, w_n\}$.

(2) *Calculate the frequency of word for each classification code*:

$$v_{ik} = \frac{n_{ik}}{N_k}, \tag{8}$$

v_{ik} is a frequency of i-th word in k-th classification code,
n_{ik} is a number of i-th word in k-th classification code in the Dictionary, and
N_k is a total number of words in k-th classification code in the Dictionary.

(3) *Calculate the frequency of word including error:*

$$v^*_{ik} = v_{ik} + P_0, \tag{9}$$

v^*_{ik} is a frequency of i-th word in k-th classification code with probability of error P_0,
v_{ik} is a frequency of i-th word in k-th classification code, and
P_0 is a probability of error, $P_0 = 0.001$

(4) *Calculate the coefficient of popularity for each classification code:*

$$w_k = \frac{n_k}{N}, \tag{10}$$

w_k is a coefficient of popularity of *k-th* classification code (popularity of occupation),
n_k is a number of words in *k-th* classification code in the Dictionary, and
N is a total number of words in the Dictionary.

(5) *Calculate the significance coefficient of classification code:*

Variant 1:

$$Q^1_k = \prod_i v^*_{ik}, \tag{11}$$

Variant 2:

$$Q^2_k = \prod_i v^*_{ik} \cdot (w_k), \tag{12}$$

Q_k is a significance coefficient of *k-th* classification code,
v^*_{ik} is a frequency of i-th word in k-th classification code with probability of error P_0, and
w_k is a coefficient of popularity of *k-th* classification code.

(6) *Calculate the Index for each classification code:*

Variant 1:

$$Index^1_k = \frac{Q^1_k}{\sum_k Q^1_k}, \tag{13}$$

Variant 2:

$$Index^2_k = \frac{Q^2_k}{\sum_k Q^2_k}, \tag{14}$$

$Index_k$ is an Index of *k-th* classification code, and
Q_k is a significance coefficient of *k-th* classification code.

(7) *Identify the classification code of string by Index of code:*

$$C^* = \arg\max(Index), \tag{15}$$

C^* is a classification code of string,
arg max is a classification code with maximum value of Index, and
Inde*x* is an Index of classification code.

The outcome from applying this second algorithm was (again) that each string is assigned two codes.

A worked example of the application of the two algorithms can be considered taking the occupation string "COTTON RING SPINNER OVERLOOKER." (Ring spinning is a method of spinning fibres to make a yarn.) According to the first algorithm, at the first step, the set of composite words is formed, as follows:

$w_1 =$ COTTON
$w_2 =$ RING

$w_3 =$ SPINNER
$w_4 =$ OVERLOOKER.

At the second step, for each word of string the weighting coefficient is calculated according to formula (1):

$$\varpi_1 = \frac{1173271}{2763615} + \frac{383881}{972759} + \frac{341070}{853174} + \frac{223748}{603840} + \frac{138974}{599958} + \ldots = 3.9029.$$

The first five elements of the sum are shown for the most significant codes: "551," "549," "550," "555," and "548" where the number of "COTTON" word is a maximum value:

$$\varpi_2 = \frac{25018}{972759} + \frac{81}{362432} + \frac{75}{91300} + \frac{54}{691820} + \frac{53}{362026} + \ldots = 0.0278.$$

The first five elements of the sum are shown for the most significant codes: "549," "608," "382," "606," and "560" where the number of "RING" word is a maximum value:

$$\varpi_3 = \frac{232811}{972759} + \frac{105110}{362026} + \frac{44325}{178115} + \frac{32965}{187026} + \frac{22976}{344594} + \ldots = 2.0625.$$

The first five elements of the sum are shown for the most significant codes: "549," "560," "559," "607," and "580" where the number of "SPINNER" word is a maximum value:

$$\varpi_4 = \frac{12536}{603840} + \frac{9324}{2763615} + \frac{5697}{132254} + \frac{4409}{362026} + \frac{3951}{92807} + \ldots = 0.2341.$$

The first five elements of the sum are shown for the most significant codes: "555," "551," "761," "560," and "572" where the number of "OVERLOOKER" word is a maximum value:

At the third step, the coefficient of word popularity is calculated according to formula (2).

$$\alpha_1 = \frac{1}{2346263/150000514} = 63.9317;$$

$$\alpha_2 = \frac{1}{25379/150000514} = 5910.4190;$$

$$\alpha_3 = \frac{1}{495256/150000514} = 302.8747;$$

$$\alpha_4 = \frac{1}{54756/150000514} = 2739.4350.$$

Here, the least popular word has a higher value of coefficient. Thus, "RING" has the highest value, "OVERLOOKER" and "SPINNER" and the "COTTON" has the lowest value of coefficient.

At the fourth step, the coefficient of order is calculated according to formula (3). The words of string have a follow values of coefficient: $\beta_1 = 0.40$; $\beta_2 = 0.25$; $\beta_3 = 0.19$; $\beta_4 = 0.16$.

At the fifth step, the two variants of Index are calculated according to formulas (4) and (5).

For Variant 1, the words have the follow values of Index:

$Index_1^1 = 3.9029 \cdot 63.9317 \cdot 0.40 = 99.8076$
$Index_2^1 = 0.0278 \cdot 5910.4190 \cdot 0.25 = 41.0774$
$Index_3^1 = 2.0625 \cdot 302.8747 \cdot 0.19 = 118.6890$
$Index_4^1 = 0.2341 \cdot 2739.4350 \cdot 0.16 = 102.6083.$

For Variant 2, the words have the follow values of Index:

$Index_1^2 = 3.9029 \cdot 63.9317 = 249.5190$
$Index_2^2 = 0.0278 \cdot 5910.4190 = 164.3096$
$Index_3^2 = 2.0625 \cdot 302.8747 = 624.6791$
$Index_4^2 = 0.2341 \cdot 2739.4350 = 641.3017.$

At the sixth step, the most significant word of string is identified. As can be seen above, for Variant 1 the most significant word is a "SPINNER," which has a maximum value of Index $Index_3^1 = 118.6890$, and for Variant 2, the most significant word is an "OVERLOOKER," which has a maximum value of Index $Index_4^1 = 641.3017$. Also, it should be noted that in Variant 2 there is the slight difference of Index between "OVERLOOKER" and "SPINNER."

At the seventh step, the classification code for the string is identified. In accordance with identified significant words, for Variant 1 the classification code of string is a "549" where the number of "SPINNER" word is a maximum value $n_{549} = 232811$, and for Variant 2 the resulting code of string is a "555" where the number of "OVERLOOKER" word is a maximum value $n_{555} = 12536$.

As a result of the first algorithm, the example string is assigned the two codes "549": "Cotton and cotton goods manufacture spinning processes" and "555": "Cotton and cotton goods manufacture undefined."

Moving to the second algorithm, at the first step, the set of words again formed (as with algorithm one) as below:

$w_1 = \text{COTTON}$
$w_2 = \text{RING}$
$w_3 = \text{SPINNER}$
$w_4 = \text{OVERLOOKER}.$

At the second step, for each word of string the frequency is calculated according to formula (8). This identifies the five most significant codes for each word of original string.

The number of "COTTON" word is a maximum value in "551," "549," "550," "555," and "548" codes. The values of frequency for these codes are:

$$\nu_{1,551} = \tfrac{1173271}{2763615} = 0.4245; \quad \nu_{1,549} = \tfrac{383881}{972759} = 0.3946;$$
$$\nu_{1,550} = \tfrac{341070}{853174} = 0.3998; \quad \nu_{1,555} = \tfrac{223748}{603840} = 0.3705;$$
$$\nu_{1,548} = \tfrac{138974}{599958} = 0.2316.$$

The number of "RING" word is a maximum value in "549," "608," "382," "606," and "560" codes. The values of frequency for these codes are:

$$\nu_{2,549} = \tfrac{25018}{972759} = 0.0257; \; \nu_{2,608} = \tfrac{81}{36432} = 0.000223;$$
$$\nu_{2,382} = \tfrac{75}{91300} = 0.000821; \quad \nu_{2,606} = \tfrac{54}{691820} = 0.000078;$$
$$\nu_{2,560} = \tfrac{53}{362026} = 0.00015.$$

The number of "SPINNER" word is a maximum value in "549," "560," "559," "607," and "580" codes. The values of frequency for these codes are:

$$\nu_{3,549} = \tfrac{232811}{972759} = 0.2393; \quad \nu_{3,560} = \tfrac{105110}{362026} = 0.2903;$$
$$\nu_{3,559} = \tfrac{44325}{178115} = 0.2489; \quad \nu_{3,607} = \tfrac{32965}{187026} = 0.1763;$$
$$\nu_{3,580} = \tfrac{22976}{344594} = 0.0667.$$

The number of "OVERLOOKER" word is a maximum value in "555," "551," "761," "560," and "572" codes. The values of frequency for these codes are:

$$\nu_{3,555} = \tfrac{12536}{603840} = 0.0208; \quad \nu_{3,551} = \tfrac{9324}{2763615} = 0.0034;$$
$$\nu_{3,761} = \tfrac{5697}{132254} = 0.0431; \quad \nu_{3,560} = \tfrac{4409}{362026} = 0.0122;$$
$$\nu_{3,572} = \tfrac{3951}{92807} = 0.0426.$$

In reality, the frequency was calculated for all existing codes, yet due to practical reasons, only the top five are shown here.

At the third step, for each word of string the frequency with error is calculated according to formula (9). Considered codes have the follow values of frequency with probability of error:

$$\nu^*_{1,551} = 0.4255; \quad \nu^*_{1,549} = 0.3956; \quad \nu^*_{1,550} = 0.4008;$$
$$\nu^*_{1,555} = 0.3715; \; \nu^*_{1,548} = 0.2326$$
$$\nu^*_{2,549} = 0.0267; \quad \nu^*_{2,608} = 0.0012; \quad \nu^*_{2,382} = 0.0018;$$
$$\nu^*_{2,606} = 0.0011; \; \nu^*_{2,560} = 0.0012$$
$$\nu^*_{3,549} = 0.2403; \quad \nu^*_{3,560} = 0.2913; \quad \nu^*_{3,559} = 0.2499;$$
$$\nu^*_{3,607} = 0.1773; \; \nu^*_{3,580} = 0.0677$$
$$\nu^*_{4,555} = 0.0218; \quad \nu^*_{4,551} = 0.0044; \quad \nu^*_{4,761} = 0.0441;$$
$$\nu^*_{4,560} = 0.0132; \; \nu^*_{4,572} = 0.0436.$$

In cases where the code does not contain word, $\nu^* = 0.001$. Analogically, the frequency is calculated for all existing codes.

At the fourth step, the coefficient of code popularity is calculated according to formula (10). Considered codes have the follow values of coefficient:

$$w_{382} = 0.000609; \; w_{548} = 0.003999; \; w_{549} = 0.006485;$$
$$w_{550} = 0.005688; \; w_{551} = 0.018424; \; w_{555} = 0.004026;$$
$$w_{559} = 0.001187; \; w_{560} = 0.002420; \; w_{572} = 0.000619;$$
$$w_{580} = 0.002297; \; w_{606} = 0.004612; \; w_{607} = 0.001247;$$
$$w_{608} = 0.002416; \; w_{761} = 0.000882.$$

Analogically, the coefficient of popularity is calculated for all existing codes.

At the fifth step, the two variants of significance coefficient of the code are calculated according to formulas (11) and (12). The result of this step for considered codes is represented in Table 4. Analogically, the significance coefficient of the code is calculated for all existing codes.

At the sixth step, the two variants of Index of the code are calculated according to formulas (13) and (14). The normalisation is performed based on calculated values of significance coefficient for all existing codes. Analogically, the Index of code is calculated for all existing codes. For Variant 1, the considered codes have the follow values of Index:

$$Index^1_{382} = 6.2{\cdot}10^{-8}; \; Index^1_{548} = 7.8{\cdot}10^{-6}; \; Index^1_{549} = 0.771213; \; Index^1_{550} = 1.4{\cdot}10^{-5}; \; Index^1_{551} = 6.3{\cdot}10^{-5};$$
$$Index^1_{555} = 0.002121; \; Index^1_{559} = 8.4{\cdot}10^{-6}; \; Index^1_{560} = 0.000155; \; Index^1_{572} = 5.4{\cdot}10^{-5}; \; Index^1_{580} = 1.9{\cdot}10^{-5};$$
$$Index^1_{606} = 3.6{\cdot}10^{-8}; \; Index^1_{607} = 5.9{\cdot}10^{-6}; \; Index^1_{608} = 4.1{\cdot}10^{-8}; \; Index^1_{761} = 1.5{\cdot}10^{-6}.$$

For Variant 2, the considered codes have the follow values of Index:

$Index^2_{382} = 7.5{\cdot}10^{-9}$; $Index^2_{548} = 6.3{\cdot}10^{-6}$; $Index^2_{549} = 0.997784$; $Index^2_{550} = 1.5{\cdot}10^{-5}$; $Index^2_{551} = 0.000231$;

$Index^2_{555} = 0.001703$; $Index^2_{559} = 2.0{\cdot}10^{-6}$; $Index^2_{560} = 7.5{\cdot}10^{-5}$; $Index^2_{572} = 6.6{\cdot}10^{-6}$; $Index^2_{580} = 8.6{\cdot}10^{-6}$;

$Index^2_{606} = 3.4{\cdot}10^{-8}$; $Index^2_{607} = 1.5{\cdot}10^{-6}$; $Index^2_{608} = 1.9{\cdot}10^{-8}$; $Index^2_{761} = 2.6{\cdot}10^{-7}$.

At the seventh step, the classification codes for the string are identified. Taking into account the results of all existing codes, the five codes with maximum value of Index are following:

Variant 1: "549": $Index^1_{549} = 0.771213$; "555": $Index^1_{555} = 0.002121$; "520": $Index^1_{520} = 0.000273$;

"560": $Index^1_{560} = 0.000155$; "525": $Index^1_{525} = 0.000149$; "571": $Index^1_{571} = 0.000098$.

Variant 2: "549": $Index^2_{549} = 0.997784$; "555": $Index^2_{555} = 0.001703$; "551": $Index^2_{5551} = 0.000231$;

"560": $Index^2_{560} = 0.000075$; "550": $Index^2_{550} = 0.000015$; "571": $Index^2_{571} = 0.000012$.

As can be seen from the figures, the value of code "549" is significantly higher than the values of other candidate codes. As a result, in this example, the original string gets the same code, "549": "Cotton and cotton goods manufacture spinning processes" for both variants.

Thus, following the application of the two algorithms, each of the 5,915,852 uncoded occupational strings was allocated four potential coded values. The next problem was to decide which of the four candidate codes was "correct." Post-processing of the results revealed that 1% of the strings had been allocated the same code across all four variants; 20% shared three codes across the variants; 63% had the same code assigned for both the variants of algorithm one; 5.5% had the same code assigned for both the variants of algorithm two; 4% also had two codes the same, but split across algorithms one and two; and for 5.5% the suggested code was different for all four variants. However, an analysis of the outcomes suggested that in spite of the fact that assigned codes are different for all variants, their meanings are very similar or ambiguous. Therefore, in the many cases when result of algorithm does not match with the human-coded result, it can be considered as a "correct" code. Applying the algorithms for the initial human-coded dataset (57,780 strings) produced correct matching for 90% of strings. Of the rejected strings, 70% were identified as acceptable, confirming the overall efficiency of the proposed algorithms.

Birthplace Data

In considering the standardisation of birthplace strings, it is important to realise that in the historic census, enumerators' books birthplace information is essentially recorded at three levels: parish of birth, county of birth, and country of birth (mainly for those living outside of their country of birth), translating to three variables within the database, BPCMTY, BPCNTY, and BPCTRY. It is also important to realise that while the data had been transcribed according to these three levels, the order in which the information is recorded in the enumerators' books must not necessarily conform to these three levels. Thus, parishes may be recorded in the county or country fields, and vice versa. It was decided to code according to geographical hierarchy: county first, then county parish. This was helped by the fact that in the case of the 1911 data for England and Wales, a Hollerith code has

TABLE 4. Example of Code Probability Calculation

Code (k)	v^*_{1k}	v^*_{2k}	v^*_{3k}	v^*_{4k}	w_k	Q^1_k	Q^2_k
382	0.001	0.0018	0.001	0.001	0.000609	1.8·10–12	1.1·10–15
548	0.2326	0.001	0.001	0.001	0.003999	2.3·10–10	9.3·10–13
549	0.3956	0.0267	0.2403	0.001	0.006485	2.2·10–5	1.4·10–7
550	0.4008	0.001	0.001	0.001	0.005688	4.0·10–10	2.2·10–12
551	0.4255	0.001	0.001	0.0044	0.018424	1.8·10–9	3.4·10–11
555	0.3715	0.001	0.001	0.0218	0.004026	6.3·10–8	2.5·10–10
559	0.001	0.001	0.2499	0.001	0.001187	2.5·10–10	2.9·10–13
560	0.001	0.0012	0.2913	0.0132	0.002420	4.6·10–9	1.1·10–11
572	0.001	0.001	0.001	0.0436	0.000619	1.6·10–9	9.8·10–13
580	0.001	0.001	0.0677	0.001	0.002297	5.5·10–10	1.3·10–13
606	0.001	0.0011	0.001	0.001	0.004612	1.1·10–12	4.9·10–15
607	0.001	0.001	0.1773	0.001	0.001247	1.8·10–10	2.2·10–13
608	0.001	0.0012	0.001	0.001	0.002416	1.2·10–12	2.9·10–15
761	0.001	0.001	0.001	0.0441	0.000882	4.4·10–11	3.9·10–14

been transcribed for birthplace which assigned a county code to English, Welsh, Scottish, and Irish birthplaces, and a country code to some but not all those born overseas (the focus was on the countries of the British Empire). Building upon this and the previous work undertaken on the 1881 census, it was relatively easy to assign a county code (CNTI) to each birthplace string (combining the three birthplace variables). Having achieved this, those born within the counties of England, Scotland, and Wales could be readily identified as candidates for parish-level coding. As a consequence, parishes could be standardised within county groups. This was supported by the creation of a parish-level authority list or dictionary covering each (ancient) county in England, Wales, and Scotland. In creating this dictionary, it was important to include not only all civil parishes but a subparish-level field linking to the parent parish, as it is clear that the information recorded in the censuses does not always relate to the parent civil parish. For example, the parish of Hatfield Board Oak in Essex contains two distinct hamlets, Bush End and Hatfield Heath. While not technically parishes, these designations may (and do) occur in the census returns and therefore must be associated with the parent parish. This issue of subparish-level information is particularly relevant in a number of northern counties where the ancient civil parish may cover a wider geographical area with several distinct settlements. In order to address this problem, a gazetteer or dictionary of parishes with associated place names was constructed from a variety of sources, including the 1911 census report, which lists a large number of subparish settlements and relates them to civil parishes in the numerous footnotes to the parish population totals (Census of England and Wales 1911, 8–373), the Office of Population Censuses and Surveys Gazetteer of Place Names and the Ordnance Survey Gazetteer (OPCS 1977; Ordnance Survey 1985). In addition, within the dictionary, each entry was assigned a weigh variable which indicated the relative size of the places in terms of population (based on the size at the end of the period, 1911). This weight variable was used in subsequent steps within the automatic processing algorithm.

Two additional problems also must be taken into consideration and incorporated into the dictionaries. The first of these is name variation and change. To continue the example of Hatfield Board Oak, this was historically known also as Hatfield Regis due to the royal forest which historically made up a large part of the parish. Linked to this is the related issue of name standardisation, which is a particular feature for Welsh parishes in the nineteenth century and to lesser extent Scottish parishes as well. The second problem relates to agglomerations of parishes, especially in the case of urban areas. To take another Essex example, the town of Southend-on-Sea is not a single parish but rather an amalgam of the four civil parishes of Prittlewell (where it had its origin in the "south end" of the parish), Leigh, Southchurch, and Eastwood. Obviously, the same is true of many large urban conurbations throughout England, Wales, and Scotland, and the dictionary must include these as well as parishes.

Once country and county had been appropriately allocated, then a similar approach to the coding of occupations was applied to birthplaces: Strings were compared to those in the compiled dictionary, initially as strings and then as word combinations, in order to predict the most likely candidate parish.

Technique for Birthplace Standardisation

As with occupational string, the technique applied in the case of birthplace standardisation was based upon a SADT approach. The processes of the birthplace standardisation are shown by the IDEF0 diagram in Figure 5. This identifies three basic stages: B1: *Creating the authority lists*; B2: *Identification of standard parish*; and B3: *Evaluation of program results*. The first of these (B1) is the process of developing and modification of the necessary dictionaries and authority lists, as outlined in the previous section. The second process (B2) allocated a standard parish to each of the original birthplace strings. This process involves several substages and steps, described below. The final process (B3) is the verification of the identified parishes. This stage was an iterative process in which initial results were checked, and where necessary, amendments to the authority lists or the algorithms were made. Figure 6 presents a detailed IDEF0 diagram of these processes.

Stage B2 (*Identification of standard parish*) consists of four substages, as follows: B2.1 Cleaning the birthplace strings; B2.2 Identification of "UNK" parish; B2.3 Identification of standard parish; and B2.4 Spell checking the birthplace strings. The first of these (B2.1) *Cleaning the birthplace strings* is a preprocessing stage. This process carried out a number of discrete tasks, as follows: (1) conversion of key abbreviations to full words (e.g., GT to GREAT, ST to SAINT, LT to LITTLE); (2) deletion of non-alpha characters, numbers, single letters, and names of countries (e.g., ENGLAND, BRITISH, SCOTLAND), names of counties (e.g., ARGYLLSHIRE, BEDFORDSHIRE, CAMBRIDGESHIRE), and other redundant words which do not contain information about a place or parish (e.g., PLACE, FROM, RESIDENT, NEAR); (3) deletion of any element of the birthplace string which contains the a subparish level address using key word pointers (e.g., ROAD, STREET, LANE, RD, and ST). All these transformations of the original string were made utilising related authority lists created manually from a full list of composite words derived from the birthplace strings.

In the historic British censuses, reflecting a period when official reporting was significantly less than it is today, it was sometimes the case that individuals did not know in detail where they had been born. Because of this, it proved important to identify these and exclude

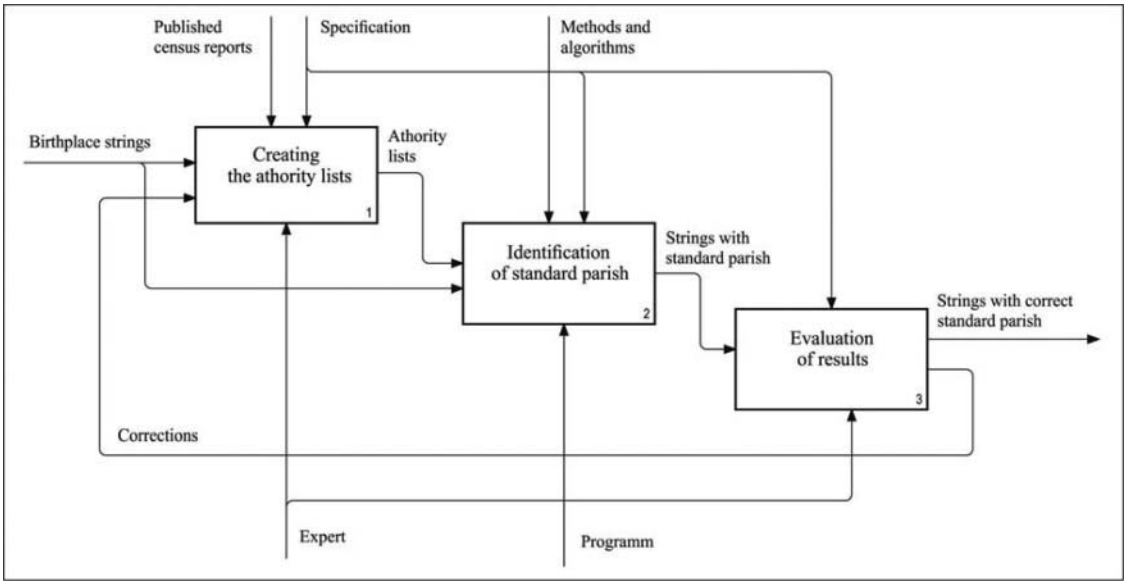


FIGURE 5. IDEF0 diagram of the birthplace string standardisation process.

them prior to trying to assign standard parishes. This was done in substep B2.2 whereby such entries were allocated "UNK" as the standard parish (variable name = STD_PAR) via recourse to an authority list created manually (containing words as UNKNOWN, NOT KNOWN, NOT NAME, NK, and BLANK).

Substage B2.3 (*Identification of standard parish)* is the key stage in the standardisation of birthplace strings and is discussed in more detail in the section that follows. As mentioned previously, this is underpinned by the use of pre-prepared authority lists or dictionaries, giving valid parishes and subparish levels places, by county. The basic

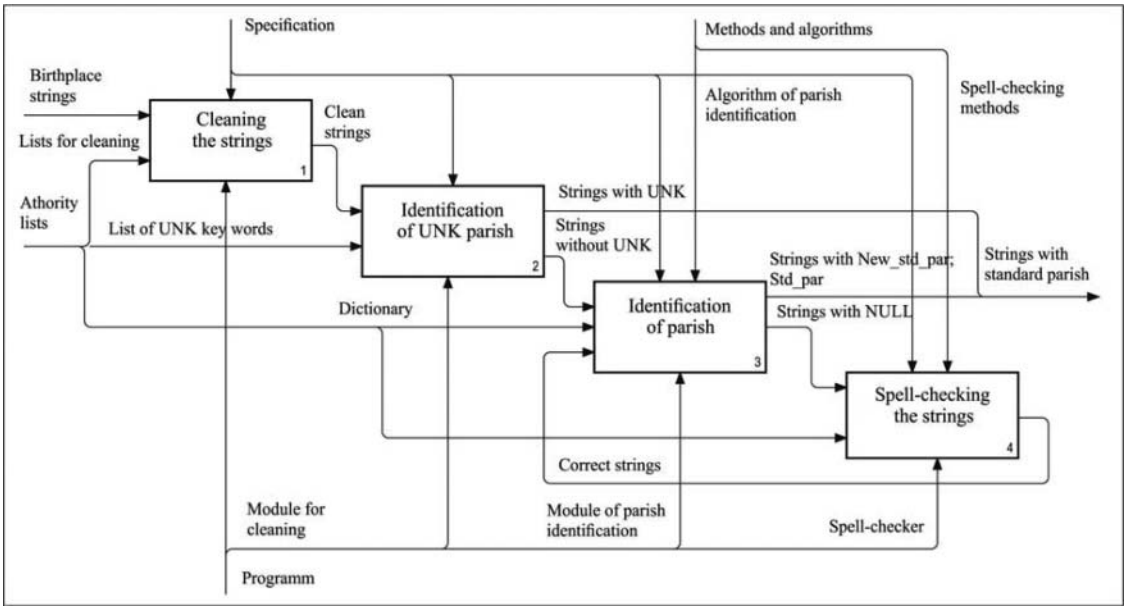


FIGURE 6. IDEF0 diagram of step B2 of the birthplace string standardisation process: identification of standard parish.

initial output of stage B2.3 is one of three outcomes: (a) where a match is made and a standard parish name assigned to the variable STD_PAR; (b) where a proxy standard name is assigned to the dummy variable alt_STD_PAR (see Figure 8); (c) where no match is made and a NULL value is assigned. In this stage, as with the coding of occupations, this process applied the spellchecking algorithm. It also examines initially the candidate string as a whole, and then if a match cannot be made, by checking the separate composite words extracted from birthplace string. The final substage, B2.4 (*Spell checking the birthplace strings*), takes those values falling into category (c) above, checks the composite strings, and attempts to correct the spelling. The strings are then passed back to substage B2.3. Similar to the coding of occupations, stage B2.4 applied the SPEEDCOP and Levenshtein algorithms. Those strings falling into category (b) above are first examined to assign an appropriate threshold (based on a proxy measure of distances

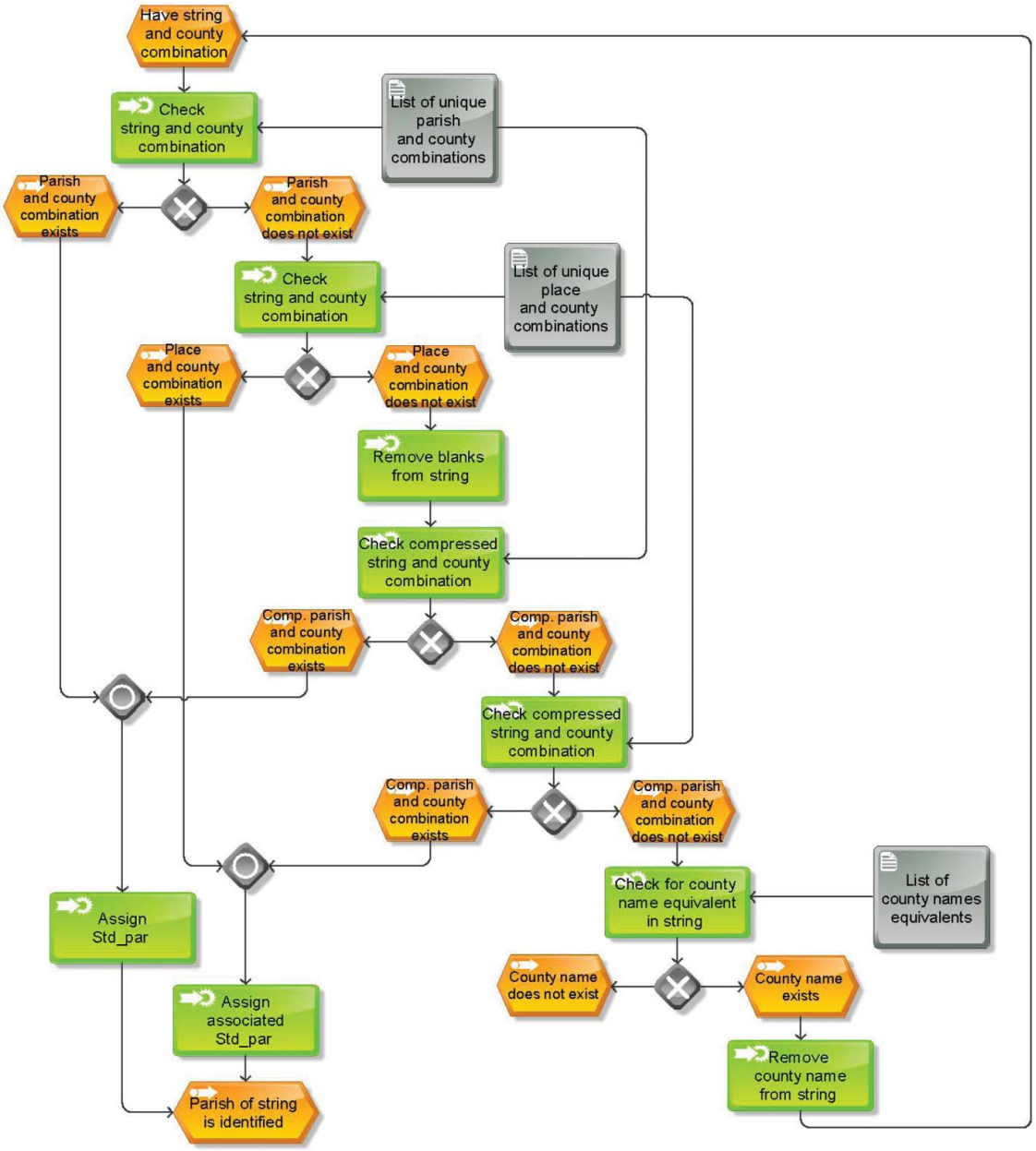


FIGURE 7. ARIS-model of parish standardisation process: checking the birthplace string.

between alternative counties). Those within the threshold have the following operations performed: The related standard parish name is assigned to the variable STD_PAR; a new value for CNTI is assigned linked to standard parish; and the original county value is written to the variable ALT_CNTI. Those strings falling outside of the threshold have a NULL value assigned and are subsequently treated as strings falling into category (b) above.

The key steps of stage B2.3 (checking the birthplace string and allocation of standard parish via checking the separate words of the candidate birthplace string) are represented by Figures 7 and 8, which illustrate the decision making processes in the form of ARIS-models. ARIS (ARchitecture of integrated Information Systems) is an approach to business process modelling which uses the Event-driven Process Chains (EPC) modelling language (Whitten, Barlow, and Bentley 1997; Hommes 2004, 137). The EPC graph describes the dynamics of the process as a sequence of events, activities, and other objects. Events (represented as a hexagon) describe under what circumstances a function or a process works; activities (represented as a rounded rectangle) describe transformations from an initial state to a resulting state. Logic rules (e.g., AND, OR, and XOR (exclusive OR)) are used for specifying the logical connections between events and activities.

Consider first Figure 7, as the ARIS-model indicates that this stage is split into six steps, as follows:

Step 1.1. Checking unique parish name in string. This step is carried out using an authority list giving unique standard parish and county combinations (in this regard, it must be realised that a number of parishes names in England and Wales occur in more than one county. This list contains variants of both parish and place (subparish), by CNTI (county name). If the birthplace string and county match a unique parish and CNTI combination in the list, then STD_PAR is assigned.

Step 1.2. Checking unique place name in string. This step uses the same authority list as step 1.1. If the birthplace string and county match a unique place and CNTI combination, then STD_PAR is assigned.

Step 1.3. Removing blanks from strings and creating compressed birthplace string.

Step 1.4. Checking unique parish name in compressed string. If the compressed birthplace string and county match a unique parish and CNTI combination in the list, then STD_PAR is assigned.

Step 1.5. Checking unique place name in compressed string. If the compressed birthplace string and county match a unique place and CNTI combination in the list, then STD_PAR is assigned.

Step 1.6. Identification of county name equivalents in string. This step is undertaken using an authority list which contains county name equivalents where the county name in abbreviated form could also be a valid place (city) name (e.g., BUCKINGHAM, BEDFORD, CAMBRIDGE, DERBY, LEICESTER, OXFORD). If the birthplace string contains a name from the list, then this word is removed, and process of checking the string is started again. If the string does not contain a name from the list, then this stage of checking is finished, and next stage (below) started.

Moving to Figure 8, the ARIS-model indicates that this stage is split into four steps, as follows:

Step 2.1. Extracting words from birthplace string and forming array of single words.

Step 2.2. Checking unique parish name in word. This step also used the authority list of unique parish and county combinations. If the extracted word from the string and county match a unique parish and CNTI combination in the list, then STD_PAR is assigned.

Step 2.3. Parish identification for word. This step uses the same authority list but utilises a place (subparish level) string in addition to the parish string. There are several conditions within the step, as shown in Figure 8:

- If an extracted word together with the county match a unique place and CNTI combination, then the associated STD_PAR is assigned.
- If an extracted word together with the county do not match a unique place and CNTI combination in the list, then the value of STD_PAR with the greatest weight is assigned.
- If an extracted word matches a unique place in the authority list yet the county does not match a CNTI, then CNTI is reassigned together with the alt_STD_PAR variable.
- If an extracted word does not match a unique place in the list and the county does not match a CNTI but matches a non-unique place and county combination, then (a) the CNTI value of the geographically nearest county is identified using a matrix of county distances, and (b) the associated alt_STD_PAR is assigned.
- If no match between the extracted birthplace words and place in the list is found, then the value "null" is assigned to STD_PAR.

Step 2.4. Resolution of possible outcomes for string. The value for STD_PAR and the associated CNTI

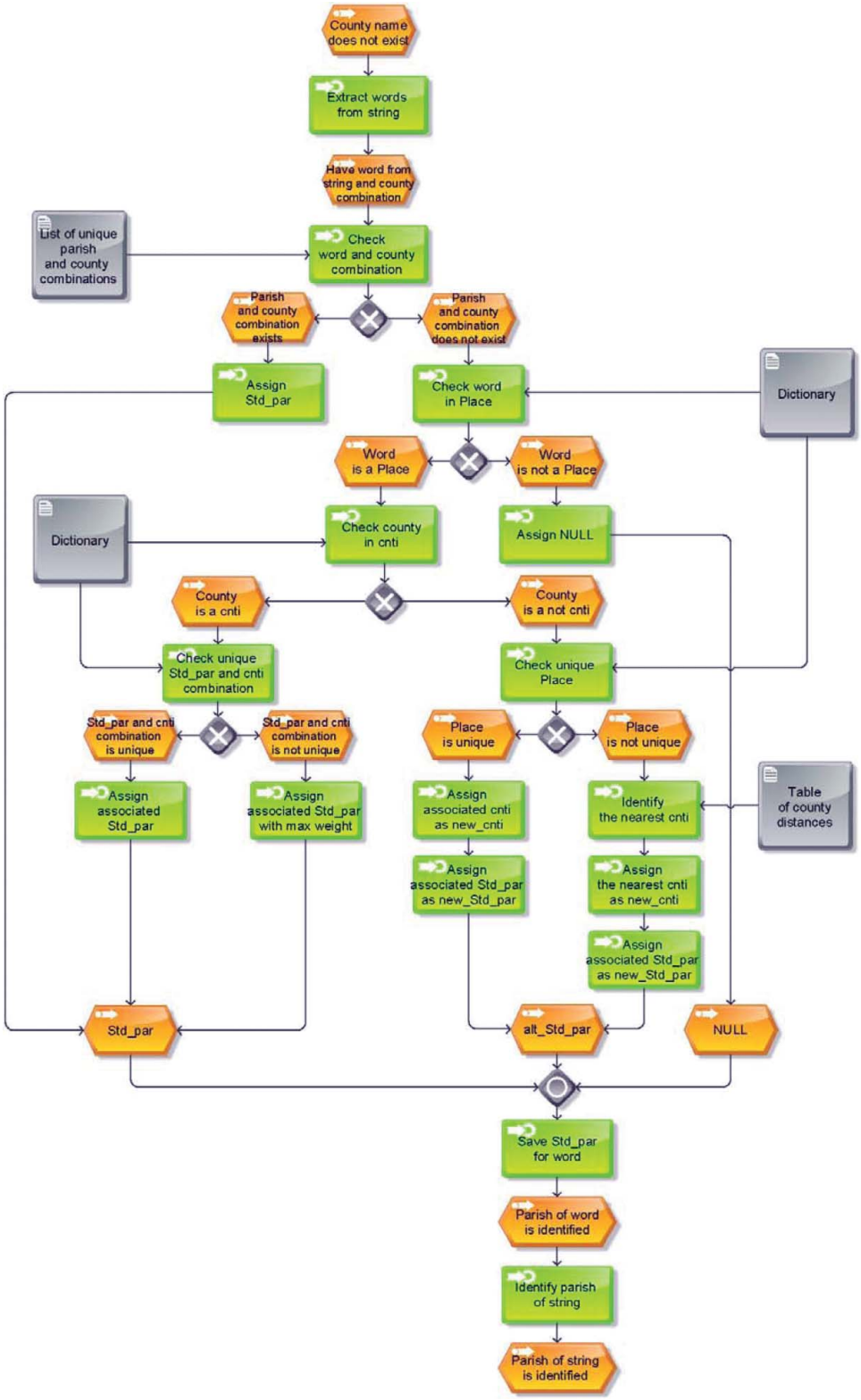


FIGURE 8. ARIS-model of parish standardisation process: checking composite words of a birthplace string.

are assigned (or reassigned in the case of CNTI) to the resulting STD_PAR with the lowest weight, or alternatively, if no STD_PAR has been allocated, to the alt_STD_PAR with the lowest weight, and if neither STD_PAR or alt_STD_PAR have been allocated, to a value of NULL. The lowest weight was used (rather than the highest) on the basis that these tended to indicate specific places located within or connected with a higher level or larger place, and thus probably provide a greater level of accuracy.

This algorithm of identification of standard parish was created by trial-and-error method taking into consideration the new detectable aspects of the task and features of the data. In order to illustrate these steps, the following example raw birthplace and CNTI combinations can be considered:

(1) HURDSFIELD | CHS
(2) GREAT HASELEY | OXF
(3) CAMBRIDGE RAMPTON | CAM
(4) LINCOLN TATTESHALL | LIN
(5) MIDDLETON | SAL
(6) CLIFTON | SOM
(7) LLANFAIR FERNDALE | GLA.

Step 1.1: String 2 has a unique parish and county combination, and thus is assigned the standard parish name GREAT HASELEY.

Step 1.2: String 1 has a unique place and county combination, and this is assigned the associated standard parish name MACCLESFIELD, since HURDSFIELD is a subparish-level place within MACCLESFIELD.

The result of steps 1.1 and 1.2 is the following:

(1) HURDSFIELD | CHS – MACCLESFIELD
(2) GREAT HASELEY | OXF – GREAT HASELEY
(3) CAMBRIDGE RAMPTON | CAM
(4) LINCOLN TATTESHALL | LIN
(5) MIDDLETON | SAL
(6) CLIFTON | SOM
(7) LLANFAIR FERNDALE | GLA.

Step 1.3: Blanks from strings are removed, with the result as follows:

(3) CAMBRIDGERAMPTON | CAM
(4) LINCOLNTATTESHALL | LIN
(4) MIDDLETON | SAL
(6) CLIFTON | SOM
(7) LLANFAIRFERNDALE | GLA.

Step 1.4: The compressed string is matched against the authority list for parish names, yet no matches are found.

Step 1.5: The compressed string is matched against the authority list for place names, yet no matches are found.

Step 1.6: The occurrence of county name equivalents in the strings is searched for CAMBRIDGE in String 3 and LINCOLN in String 4 are found, and these words are removed from strings. The result of this step is the following:

(3) RAMPTON | CAM
(4) TATTESHALL | LIN
(5) MIDDLETON | SAL
(6) CLIFTON | SOM
(7) LLANFAIR FERNDALE | GLA.

Repeating step 1.1 after removing county name equivalents, String 3 now has a unique parish and county combination, and thus is assigned the standard parish name RAMPTON.

The result of first stage is the following:

(1) HURDSFIELD | CHS – MACCLESFIELD
(2) GREAT HASELEY | OXF – GREAT HASELEY
(3) RAMPTON | CAM – RAMPTON
(4) TATTESHALL | LIN
(5) MIDDLETON | SAL
(6) CLIFTON | SOM
(7) LLANFAIR FERNDALE | GLA.

With no further matches, Strings 4, 5, 6, and 7 move to the second stage, checking the composite words of the birthplace string separately.

Step 2.1: The result of extracting words from string provides the following word/county pairs:

(4) TATTERSHALL | LIN
(5) MIDDLETON | SAL
(6) CLIFTON | SOM
(7a) LLANFAIR | GLA
(7b) FERNDALE | GLA.

Step 2.2: Checking for unique parish name and county combination assigns the standard parish LLANFAIR to String 7a.

Step 2.3: Checking for unique place name and county combination assigns the standard parish RHONDDA to String 7b based on FERNDALE being a subparish-level place in RHONDDA. The MIDDLETON | SAL pair of String 5 does not match a unique place/county combination since MIDDLETON exists as a place in two parishes within Shropshire (SAL) – BITTERLEY (weight = 168) and OSWESTRY (weight = 1307). Consequently, string 5 is assigned the standard parish OSWESTRY as it has a higher weight (indicating that it has a bigger population than BITTERLEY and therefore

has a higher probability of an individual being both there). String 6 CLIFTON does not have a unique parish/county combination. The authority list contains several parishes called CLIFTON, yet non in Somerset (SOM). There are seven other candidate counties with a parish called CLIFTON (WES, WOR, BDF, DBY, DEV, GLS, LAN), of which the nearest geographically is Gloucester (GLS). Consequently, an alternative CNTI for string 7 is allocated the value GLS the variable alt_STD_PAR is assigned the value of CLIFTON. Lastly, string 4 fails to match any parish/place name in the authority list and therefore the NULL variable is set. The result of steps 2.2 and 2.3 is as follows:

(4) TATTESHALL – | LIN
(5) MIDDLETON – OSWESTRY | SAL
(6) CLIFTON – CLIFTON| GLS
(7) LLANFAIR – LLANFAIR | GLA (weight is 16)
(8) FERNDALE – RHONDDA | GLA (weight is 1,307).

Step 2.4: The two variants for string 7 is resolved using the associated weights. The result after the two completed stages is as follows:

(1) HURDSFIELD | CHS – MACCLESFIELD | CHS
(2) GREAT HASELEY | OXF – GREAT HASELEY | OXF
(3) CAMBRIDGE RAMPTON | CAM – RAMPTON | CAM
(4) LINCOLN TATTESHALL | LIN – (null) NULL| LIN
(5) MIDDLETON | SAL – OSWESTRY | SAL
(6) CLIFTON | SOM – (alternative) CLIFTON | GLS
(7) LLANFAIR FERNDALE | GLA –LLANFAIR | GLA.

This example leaves strings 4 and 6 unresolved. String 4 is spellchecked (using the SPEEDCOP and the Levenshtein algorithms) and processed through the various steps again. This then assigns the standard parish of TATTERSHALL. The result for strings 4 is ATTERSHALL | LIN. String 6 is then examined in order to determine what is a suitable threshold to accept, given that a proxy measure of distance is used to determine between "correct" and "incorrect" county. Those strings for which the substituted county is taken as correct have their CNTI value replaced by the new county code, and ALT_CNTI is set to the old county code. STD_PAR is then assigned accordingly. Those strings outside the threshold are assigned a NULL value and processed the same as string 4 in the example, being spellchecked and processed through the various steps again. In the example given, the alternative county code is accepted as being within the threshold, and the substitution is made, the result being CLIFTON | GLS (alternative variant).

Conclusion

The assignment of coded values to occupational title strings and standardised parish names (and associated county values) to birthplace strings is not a new topic. Usually, such tasks are performed manually or semi-automatically, such as by searching, sorting, and rearranging the data in order to make coding easier, as well as basic comparison of strings for "likeness" of match. The size of the problem in this particular project required a different approach. The balance of work must be on fully automatic coding rather than manual coding. Given the distribution of the strings to be coded in terms of their frequency within the underlying data (see Figures 1 and 2), with a small minority of strings representing a numeric majority of the population as a whole, in the case of birthplaces and occupations, it was decided to solve the problem by developing algorithms based on statistical decision theory and Bayesian analysis. This is a novel approach: Other large scale projects standardising textual strings have tended to employ a mix of manual coding and semi-automatic solutions relying more heavily on phonetic and/or pattern matching algorithms in isolation, without a statistical probability component. Using expert knowledge to initially code (manually) a small but significantly representative subset of the data, the algorithms were then used to allocate codes to the remaining strings in the data. In so doing, string comparison algorithms were used, together with word transformation algorithms within strings (spellchecking and word substitution), and the calculation of "popularity" indices for occupations and other indices based on proximity and population size to determine differential probabilities for birthplaces. This complex set of decision-making stages enabled a large body of data to be coded. Due to the nature of the variations between and inherent complexity of the underlying raw strings, mistakes will undoubtedly have been made, which only detailed local geographical knowledge in the case of birthplaces or specific expertise about historical occupational terms in their regional context could resolve. However, the result is a large database for which the majority of decisions made can be taken as "correct," together with a methodology which we believe could be applied successfully to related standardisation problems.

NOTES

1. Further details on the I-CeM project are available from http://www.essex.ac.uk/history/research/icem. The data are available to users registered with the UK Data Service from http://icem.data-archive.ac.uk/#step1.

2. In the historic censuses, place of birth was essentially recorded at three levels: parish, county, and country, in that order. However, in the actual census records, the order might be switched and levels missed out (e.g., Scotland, Glasgow). Likewise, it was not required to state the country of birth if one was living in that country. Also, the lowest level recorded

might not be a parish per se, but rather a "place" which could be a subunit (village, hamlet) within a parish, or an amalgam of parishes (city or town). See section 5 below.

3. Many children would have no occupation recorded, as is also the case for some women.

4. Unlike occupation where some individuals did not provide a response, by definition, all individuals were liable to provide a place of birth, even if it was "unknown."

5. Only those born in England, Scotland, and Wales were standardised at parish level since the censuses invariably do not contain enough information to determine parish level for those born elsewhere.

6. Hollerith punch card tabulators were first used in Great Britain for the censuses for 1911. Concerned by the lengthy process of producing reports (the final report of the 1880 U.S. Census was not published until 1888), the U.S. Census Bureau held a contest in order to select a new system. Hollerith was the clear winner, and his tabulating system was used by the U.S. Census Bureau for the 1890 census, as well as for analysing the Austrian and Canadian censuses of the same year. Following the success, in 1896 Hollerith founded the Tabulating Machine Company, one of the four companies which later joined together to form the Computing-Tabulating-Recording Company, better known in its reorganised form as IBM (see *BPP 1911*, CVII, *General Report with Appendices*, Appendix B). A report to the Treasury Committee in 1890 had earlier recommended the introduction of "mechanical appliances to aid the work of tabulation," but the process was not implemented for another twenty years (see *BPP 18901*, LVIII, *Report of the committee appointed by the Treasury to enquire into certain questions concerned with the taking of the census with evidence and appendices, and the Treasury minute appointing the committee*). Tabulation was done in England and Wales directly from the household schedules, and as a result the latter were no longer copied by enumerators into enumeration books for dispatch to the London Census Office as in previous years.

REFERENCES

Anderson, M. J. 1988. *The American census. A social history*. New Haven: Yale University Press.

Austrian, G. 1982. *Herman Hollerith: Forgotten giant of information processing*. New York: Columbia University Press.

Berger, J.O. 1985. *Statistical decision theory and Bayesian analysis*. New York: Springer-Verlag.

Census of England and Wales. 1911. *Area, families or separate occupiers, and population*. Vol. II. Registration areas. BPP 1912-13 CXI 679-[Cd. 6259]. Table 5. Registration counties, districts and sub-districts with their constituent civil parishes. Urban or rural district in which each parish is situated; area; families or separate occupiers, and population, 1901 and 1911; and population enumerated in institutions, large establishments, and on vessels, &c.

Davis, W. S. 1994. *Business systems analysis and design. Business & economics*. Wadsworth: University of Michigan.

Eames, C., and R. Eames. 1990. *A computer perspective. Background to the computer age*. Cambridge: Harvard University Press.

Higgs, E. 1996. The statistical Big Bang of 1911: Ideology, technological innovation and the production of medical statistics. *Social History of Medicine* 9:409–26.

Higgs, E. 2004. *The information state in England: The central collection of information on citizens, 1500–2000*. London: Palgrave.

Higgs, E. 2005. *Making sense of the census revisited. Census records for England and Wales, 1801–1901—A handbook for historical researchers*. London: The National Archives and Institute of Historical Research.

Higgs, E., C. Jones, K. Schürer, and A. Wilkinson. 2013. *The integrated census microdata (I-CeM) guide*. Colchester: UK Data Archive. http://www.essex.ac.uk/history/research/icem/documents/icem_guide.pdf.

Hommes, B.-J. 2004. *The evaluation of business process modeling techniques*. Delft: University of Technology.

Lawton, R. (Ed.). 1978. *The census and social structure. An interpretive guide to nineteenth century censuses for England and Wales*. London: Cass.

Marca, D. A., and C. L. McGowan. 1987. *SADT: Structured analysis and design technique*. New York: McGraw-Hill.

Mills, D. R., and K. Schürer (Eds.). 1996. *Local communities in the Victorian census enumerators' books*. Oxford: Leopold's Head.

Office of Population Censuses and Surveys. 1977. *Census 1971: England and Wales, Index of place names. Great Britain*. London: HMSO.

Ordnance Survey. 1985. *Ordnance survey Landranger gazetteer: A gazetteer of all names shown on ordnance survey 1:50,000 scale Landranger maps. Great Britain*. Southampton: Author.

Pollock, J., and A. Zamora. 1984. Automatic spelling correction in scientific and scholarly text. *Communications of the ACM* 27:358–68.

Schierle, M., S. Schulz, and M. Ackermann. 2008. From spelling correction to text cleaning—Using context information. *Data Analysis, Machine Learning and Applications* (series Studies in Classification, Data Analysis, and Knowledge Organization) 5:397–404.

Schürer, K. 1991. The 1891 census and local population studies. *Local Population Studies* 47:16–29.

Schürer, K., and E. Higgs. 2014. *Integrated census microdata (I-CeM); 1851–1911* [computer file]. Colchester: UK Data Archive [distributor], SN: 7481, http://dx.doi.org/10.5255/UKDA-SN-7481-1.

Schürer, K., and M. Woollard. 2000. *1881 census for England and Wales, the Channel Islands and the Isle of Man (Enhanced Version) [computer file]*. Genealogical Society of Utah, Federation of Family History Societies, [original data producers]. Colchester: UK Data Archive [distributor], SN: 4177, http://dx.doi.org/10.5255/UKDA-SN-4177-1.

Whitten, J. L., V. M. Barlow, and L. Bentley. 1997. *Systems analysis and design methods*. New York: McGraw-Hill Professional.

Woollard, M. 1999. *The classification of occupations in the 1881 census of England and Wales*. Colchester: University of Essex.

Wrigley, E. A. (Ed.). 1972. *Nineteenth-century society. Essays in the use of quantitative methods for the study of social data*. Cambridge: Cambridge University Press.

HISTORICAL LIFE COURSE STUDIES

VOLUME 2
2015

EHPS NETWORK

MISSION STATEMENT

HISTORICAL LIFE COURSE STUDIES

Historical Life Course Studies is the electronic journal of the *European Historical Population Samples Network* (EHPS-Net). The journal is the primary publishing outlet for research involved in the conversion of existing European and non-European large historical demographic databases into a common format, the Intermediate Data Structure, and for studies based on these databases. The journal publishes both methodological and substantive research articles.

Methodological Articles
This section includes methodological articles that describe all forms of data handling involving large historical databases, including extensive descriptions of new or existing databases, syntax, algorithms and extraction programs. Authors are encouraged to share their syntaxes, applications and other forms of software presented in their article, if pertinent, on the EHPS-Net website.

Research articles
This section includes substantive articles reporting the results of comparative longitudinal studies that are demographic and historical in nature, and that are based on micro-data from large historical databases.

Historical Life Course Studies is a no-fee double-blind, peer-reviewed open-access journal supported by the European Science Foundation (ESF, http://www.esf.org), the Scientific Research Network of Historical Demography (FWO Flanders, http://www.historicaldemography.be) and the International Institute of Social History Amsterdam (IISH, http://socialhistory.org/). Manuscripts are reviewed by the editors, members of the editorial and scientific boards, and by external reviewers. All journal content is freely available on the internet at http://www.ehps-net.eu/journal.

Editors: Koen Matthijs & Paul Puschmann
Family and Population Studies
KU Leuven, Belgium
hislives@kuleuven.be

The European Science Foundation (ESF) provides a platform for its Member Organisations to advance science and explore new directions for research at the European level. Established in 1974 as an independent non-governmental organisation, the ESF currently serves 78 Member Organisations across 30 countries. EHPS-Net is an ESF Research Networking Programme.

The European Historical Population Samples Network (EHPS-net) brings together scholars to create a common format for databases containing non-aggregated information on persons, families and households. The aim is to form an integrated and joint interface between many European and non-European databases to stimulate comparative research on the micro-level.
Visit: http://www.ehps-net.eu.

HISTORICAL LIFE COURSE STUDIES
VOLUME 2 (2015), 38-57, published 29-09-2015

Creating a typology of parishes in England and Wales: Mining 1881 census data

Kevin Schürer
University of Leicester

Tatiana Penkova
Institute of Computational Modelling SB RAS

ABSTRACT

The paper presents the application of principal component analysis and cluster analysis to historical individual level census data in order to explore social and economic variations and patterns in household structure across mid-Victorian England and Wales. Principal component analysis is used in order to identify and eliminate unimportant attributes within the data and the aggregation of the remaining attributes. By combining Kaiser's rule and the Broken-stick model, four principal components are selected for subsequent data modelling. Cluster analysis is used in order to identify associations and structure within the data. A hierarchy of cluster structures is constructed with two, three, four and five clusters in 21-dimensional data space. The main differences between clusters are described in this paper.

Keywords: Principal Component Analysis, Cluster Analysis, Census Data, Household Structures

e-ISSN: 2352-6343
PID article: http://hdl.handle.net/10622/23526343-2015-0004?locatt=view:master

The article can be downloaded from here.

1 INTRODUCTION

The opportunities to explore household and family patterns in new ways as a result of the emergence of new data resources providing large amounts of individual level historical microdata, sometimes covering entire countries, has been commented upon by Steven Ruggles (2012). One approach advocated by Ruggles is to undertake analyses of spatial variation, using the greater and finer geographical coverage of these new data resources to illustrate complexities and differences that single place studies cannot.[1] As one strand of a larger multi-national JISC-funded project,[2] this paper does exactly that. It explores spatial variations and patterns in household structure across mid-Victorian England and Wales in terms of socio-economic indicators, by applying multi-dimensional analysis techniques to historical geo-referenced census data. However, in so doing, it specifically does not address the decline of patriarchal family forms in Europe and beyond, a topic that Ruggles specifically suggests that these new data resources be used to address (Ruggles 2012). In part, this is because it is an analysis of just a single census year, and thus change over time cannot be detected. Moreover, this is a study of variations in *household* form rather than a study of evolving *family* systems. The two are rather different. Thus, while this research includes co-residential kinship structures as part of its analysis, it paints with a much broader brush. Moving the focus from *family* to *household* and then to *parish*, this study marshals a wide range of indices, familial and non-familial alike, in order to try and understand how the composite households and their inhabitants within one locality or place (in this case the parish) are similar or different from those in the places which surround them. In this sense, the goal is to better understand how variations at the household and parish levels contribute to broader regional differences and variations. Are households in the north, south, east or west essentially the same in mid-Victorian England and Wales, or can we detect differences at a regional level between them?

To date, there have been relatively few studies of geographical variations in historical household structure in England and Wales. Those that have been attempted have been relatively inconclusive due to a basic lack of detailed data in order to fully investigate the subject, mainly because they have had to resort to the use of aggregated census data resulting in a lack of spatial granularity and detail, or partial sources for pre-census periods (Wall 1977; Schürer 1992). Since the publication of *Household and Family in Past Time* in 1972, the common orthodoxy which has developed is that the households of the past in England and Wales were predominately nuclear in terms of family form and varied little over space (and time) (Laslett 1972; Laslett 1983).[3] This was summarised by Wall in 1983 as follows: "The basic structure of English households in the pre-industrial era is now well known. Households were small. The majority contained fewer than five persons and membership was customarily confined to parents and their unmarried children" (Wall 1983). However, despite this bold statement, any systematic attempt to consider regional variation has been mainly absent. Curiously, when Peter Laslett presented his initial findings on English historical household structure in the journal *Population Studies* in 1969 the article was entitled 'Part I'.[4] The second instalment, to be published later in the same journal, was to "describe and analyse variations in mean household size by region and by period" (Laslett 1969). But 'Part II' never appeared, it seems, primarily because there was no story to tell.

The conventional view that household structure varied little historically, has in part been re-enforced by a number of demographic studies that have emphasised the homogeneity of England's demographic experience rather than its variance – especially in comparison with other European countries (Wrigley & Schofield 1983; Wrigley 1985). Reviewing Teitelbaum's study of fertility decline in England and Wales, Laslett commented that it portrayed the demographic experience of the English like "the red coats on parade in front of Buckingham Palace, every unit in step with every other, and all changing direction at the same time" (Laslett 1985; Teitelbaum 1984; cf. Garrett, Reid, Schürer & Szreter 2001). However, we are still left with two basic problems: how much of this seemingly homogeneity is a factor of either, first, the size of the units under observation; or second, the range of variables under consideration. Teitelbaum's

1 Ruggles (2012) proposes that studies using the newly available large data sources should use demographically appropriate measures, study spatial variation in families and households and study long-run historical changes. (p.424).

2 The title of the JISC-funded project was "Mining Microdata: economic opportunity and spatial mobility in Britain, Canada and The United States, 1850-1911". This was undertaken jointly with the University of Alberta, University of Montreal, University of Guelph (all Canada), the Minnesota Population Center at the University of Minnesota (USA). Details are available at http://www.miningmicrodata.org/. Details on the Digging into Data research programme are at: http://www.diggingintodata.org/.

3 The traditional picture for England and Wales varies dramatically to that recently portrayed by Szołtysek, Gruber, Klüsener & Goldstein (2014) in which they suggest a distinct north/south division with greater household complexity in the north and with disparities being explained by agriculture, fertility and differences in age structures.

4 The sub-title, rarely cited is 'Part I. Mean Household Size in England since the Sixteenth Century'. See Laslett (1969).

study of fertility decline geographically focused on the 50 or so administrative historic county units of England and Wales. Widening the scope to 614 'registration' districts used in England and Wales in the nineteenth-century, Woods has demonstrated considerably more geographic variation in relation to mortality (Woods & Shelton 1997; Woods 2000). What would the situation look like if the telescope lens was amplified not just 10-fold, from 50 to 600 units, but over 3000-fold, to 17,000 units? And, from a household perspective, would homogeneity persist if we broadened our focus beyond size and the presence or otherwise of co-resident kin, to include summary measures on servants, lodgers, occupational concentration, isonomy, migration and so on? By significantly changing the focus of the investigation, in terms of both the geographic scope and the thematic range, this research will test the notion of homogeneity in household structure and produce a new typology of parish-based regional variation.

In order to do this, this paper will examine variations in household structure by using complete count, individual level, census data for 1881. In all, some 25 million person records have been aggregated at household and parish levels and then examined applying principal component analysis and cluster analysis. Principal component analysis (PCA) is one of the most common techniques used to describe patterns of variation in multi-dimensional data (Gorban & Zinovyev 2009). Moreover, PCA is recognised as one of the more robust ways to identify and carry out dimensionality reduction, which in turn, allows the selection of the most informative features (Abdi & Williams 2010). Cluster analysis is a tool for discovering key associations and structures within the data and typology development (MacQueen1967). Within this research, the analysis and visualisation of multi-dimensional data has been conducted using the ViDaExpert application (Zinovyev 2000). This software allows users to construct simple visual representations of the dataset in order to explore its intrinsic patterns and regularities.

The paper is structured as follows: section 2 presents a description of the data; section 3 considers the results of the PCA including the elimination of unimportant features and the aggregation of attributes, the selection of the number of principal components, the contribution of the data attributes to the principal components and data visualisation; section 4 presents the results of the cluster analysis with visualisation of two-, three-, four- and five- cluster structures within the data; section 5 presents conclusions.

2 DATA DESCRIPTION

The dataset is derived from the individual level census data from the 1881 of England and Wales (Schürer & Woollard 2000; Schürer & Woollard 2002). From this some 25 million person records were aggregated at household and then parish level. The resulting dataset used in this analysis contains 13,390 objects, essentially discrete parish-level geographical entities, each with 45 measured attributes. The set of attributes includes two basic types: the first are those providing a range of socio-economic summary measures derived from the underlying data relating to the respective parishes; the second are additional locational reference characteristics, used for data interpretation and visualisation. The dataset contains 33 main attributes and 12 additional attributes. These are listed in Table 1.

Table 1 *List of the data attributes*

Main attributes		
1	*HHsize*	Mean household size
2	*SolitaryMHH*	% of households headed by a solitary male
3	*SolitaryFHH*	% of households headed by a solitary female
4	*HH_with_kin*	% of households with residential kin
5	*HH_with_servt*	% of households with residential servants
6	*HH_with_inmates*	% of households with non-family members
7	*WorkingF25+*	% of females aged 25+ who are working
8	*Working20+*	% aged 20+ who are working
9	*Working<=14*	% aged 14 and less who are working

Main attributes		
10	*Working55+*	% of males aged 55+ who are working
11	*Males_in_agric*	% of males aged 25+ working in agriculture
12	*Native*	% who are native (born in same county)
13	*Foreign*	% who are born overseas
14	*Scottish*	% born in Scotland
15	*Irish*	% born in Ireland
16	*HHsize6+*	% of households with 6 or more offspring
17	*No_par<=5*	% aged 5 or less living without parents
18	*Sing_Par<=5*	% aged 5 or less living with a single parent
19	*Live_with_par15-16*	% aged 15-16 living in the parental home
20	*Live_with_par17-18*	% aged 17-18 living in the parental home
21	*Live_with_par19-20*	% aged 19-20 living in the parental home
22	*Live_with_par21-22*	% aged 21-22 living in the parental home
23	*With_older_sibs*	% aged 25+ living with siblings aged 25+
24	*Aunt/uncle*	% living with aunts or uncles
25	*Nephew/niece*	% living with nieces or nephews
26	*Cousins*	% living with cousins
27	*Grandparents*	% living with grandparents
28	*Grandchildren*	% living with grandchildren
29	*Occ_similarity*	Measure of occupation concentration
30	*Name_similarity*	Measure of surname heterogeneity
31	*Blind*	% blind
32	*Deaf*	% deaf
33	*Mental*	% with mental disability
Additional attributes		
34	*Standardparish*	Name of place
35	*Country*	Country
36	*Division*	Census Division
37	*RC*	Census Registration County
38	*RC_ref*	Census Registration County ref
39	*RD*	Census Registration District
40	*RD_ref*	Census Registration District ref
41	*Area*	Area of parish unit
42	*Aggpop*	Population size of parish unit
43	*Density*	Population density of parish unit
44	*X_centroid*	X coordinate of parish unit
45	*Y_centroid*	Y coordinate of parish unit

3 DATA AGGREGATION

The data aggregation process includes both the aggregation of attributes and the elimination of unimportant features. The aggregation of attributes is based on PCA techniques and correlation analysis. To identify attributes with similarities, the contribution of the data attributes to the four principal components was analysed. This suggested that there are six groups of attributes with equal signatures:

1 – *SolitaryMHH* and *SolitaryFHH*
2 – *WorkingF25+* and *Working20+*
3 – *Live_with_par17-18*; *Live_with_par19-20* and *Live_with_par21-22*
4 – *Aunt/uncle, Nephew/niece* and *Cousins*
5 – *Grandchildren* and *Grandparents*
6 – *Occ_similarity* and *Name_similarity*.

The results demonstrate a strong correlation between attributes. Taking into account the contribution of the data attributes to the principal components and correlation coefficients between attributes, it was possible to create the following aggregate attributes:

1 – *SolitaryHH*
2 – *Working20+*
3 – *Live_with_par17-22*
4 – *Distant_relatives*
5 – *Gdchildren/Gdparents*
6 – *Occ/names_similarity*

As a result of data aggregation, the number of attributes was reduced to 25 (from 33).

The elimination of unimportant features is based upon a PCA definition of unimportant attributes. The criterion for the definition of unimportant attributes is Kaiser's rule for eigenvector of the principal components: $z_i^2 < \frac{1}{n}\sum_{i=1}^{n} z_i^2$, where $i = \overline{1,n}$; n – is a number of attributes; λ_i – is a value of i-th attribute in eigenvector. The attributes that have values less than the average value for all principal components are excluded. The analysis of the principal components showed that there were four attributes which could be deleted from further analysis: *P5singpar, Blind, Deaf* and *Mental*. Consequently, after elimination of unimportant features, the dataset contained only 21 attributes.

4 PRINCIPAL COMPONENT ANALYSIS

PCA is one of the most common techniques used to describe patterns of variation within a multi-dimensional dataset, and is one of the simplest and robust ways of doing dimensionality reduction. PCA is a mathematical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components (Peres-Neto, Jackson & Somers 2005). The number of principal components is always less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has the largest possible variance and each subsequent component, respectively, has the highest variance possible under the constraint that it be orthogonal to the preceding components.

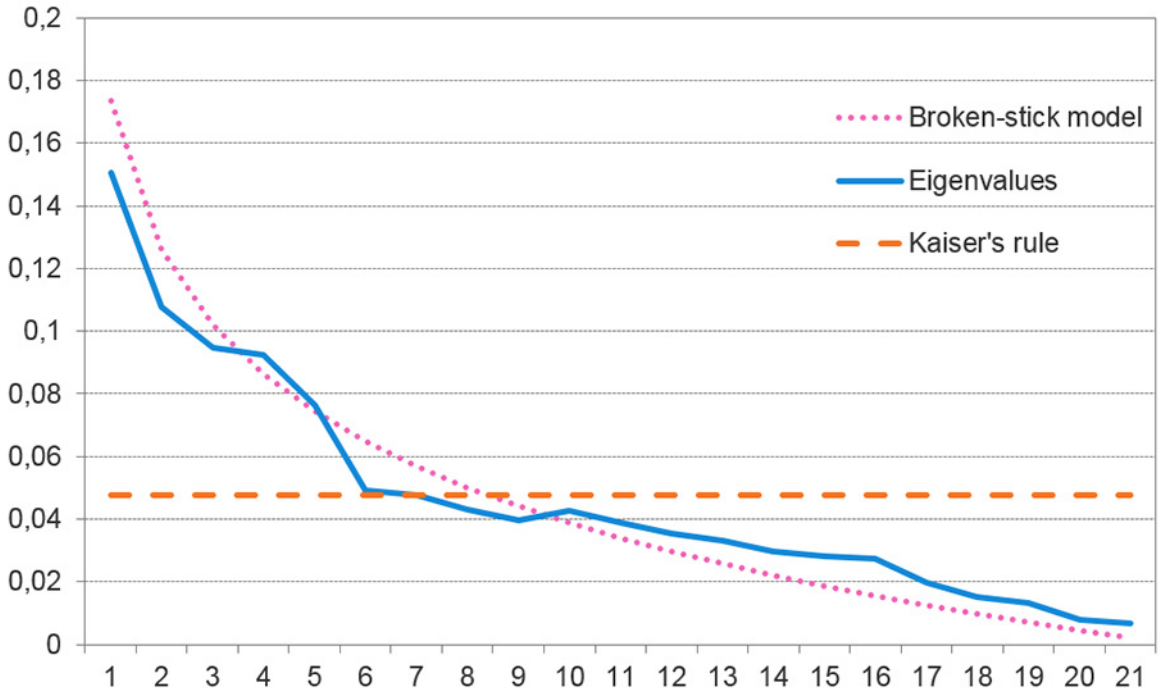
4.1 SELECTION OF THE NUMBER OF PRINCIPAL COMPONENTS

One of the greatest challenges in providing a meaningful interpretation of multi-dimensional data using PCA is determining the number of principal components. There are various methods and stopping rules used to identify the number of principal components. In selecting the number of principal components we applied the most commonly used method, namely Kaiser's rule and the Broken-stick model based on eigenvalues of components. According to Kaiser's rule, the components that have eigen values greater than the average value are retained for interpretation: $\lambda_i > \frac{1}{n}\sum_{i=1}^{n}\lambda_i$, where $i = \overline{1,n}$; n – is a number of components; λ_i – is a eigenvalue of i-th component. The concept underlying the Broken-stick model is that if a stick is randomly broken into n pieces, l_1 would be the average size of the largest piece in each set of broken sticks; l_2 would be the average size of the second largest piece, and so on. The number n equals the number of components and the total amount of variation across all components. The proportion of total variance associated with the eigenvalue for i-th component under the broken-stick model is obtained by: $l_i = \frac{1}{n}\sum_{j=i}^{n}\frac{1}{j}$. If the i-th component has an eigenvalue larger than l_i, then the component is retained. Initially, four principal components were identified.

Principal components for a reduced number of data attributes were selected based on combination of Kaiser's rule and the Broken-stick model. Figure 1 illustrates the eigenvalues of components.

As can be seen from Figure 1, Kaiser's rule determines five principal components – eigenvalues of first five components are significantly greater than the average value. The Broken-stick model gives three principal components – the line of Broken-stick model cuts the eigenvalues of first three components. In addition, the spectral gap (i.e. the distance between eigenvalues) separates the first component significantly, and the second, third, fourth and fifth components from other components. Consequently, for reduced data attributes four principal components were identified: PC1, PC2, PC3 and PC4.

Figure 1 *Eigenvalues of components for reduced data attributes*



4.2 CONTRIBUTION OF THE DATA ATTRIBUTES TO THE PRINCIPAL COMPONENTS

The contribution of the reduced data attributes to principal components is represented in Figures 2-5.

The first principal component (PC1, Figure 2) is characterised by the following attributes: moderately large *household size*; high proportions of both *households with residential kin* and *households with residential servants*; a high percentage of *males working in agriculture*; a strong negative correlation with the percentage of *households with six or more offspring* and also *children (ages from 15 to 22) living in the parental home*; a high proportion of *children aged 5 or less living without parents*; high proportions *living with siblings, aunts or uncles, nieces or nephews, cousins, grandparents* and *grandchildren*; and high levels of *occupation concentration* and *surname concentration* (i.e. relatively low surname heterogeneity). In combination, these components suggest rural parishes dominated by a single source of employment (agriculture) with large families, but where residential (extended) kin and servants are an important element of overall household size pro rata to offspring. Strong surname concentration may also indicate a less mobile population.

The second principal component (PC2, Figure 3) is characterised by the following attributes: relatively small *household size*; a high percentage of *households with residential servants* and *households with non-family members*; low proportions of *males working in agriculture*; relatively low proportions *native born* and high percentages *born overseas, born in Scotland* and *born in Ireland*; low proportions of households with *six or more offspring*; high proportions of *children aged 5 or less living without parents*; low proportions of *children (ages from 15 to 22) living in the parental home*; high proportions of households with members *living with siblings, aunts or uncles, nieces or nephews* and *cousins*; together with high levels of *occupation concentration* and *surname concentration*. In combination, these components suggest mainly inner urban parishes with a mobile population and varied economy/occupation structure, with relatively small households, but where residential (extended) kin, boarders. lodgers and servants are an important element of overall household size pro rata to offspring.

The third principal component (PC3, Figure 4) is characterised by the following attributes: moderately large *household size*; high proportion of *households with residential kin*; low proportions of *households with residential servants*; low percentage of *males working in agriculture*; high proportion of households with *six or more offspring;* high percentages of *children (ages from 15 to 22) living in the parental home*; high percentages l*iving with siblings, aunts or uncles, nieces or nephews, cousins, grandparents* and *grandchildren*; and low levels of *occupation concentration* and *surname concentration*. In combination, these components suggest parishes with a fairly mixed economy/occupational structure, yet which are not urban areas with a high migrant component – maybe smaller market towns – with large families where both residential kin and the retention of children in the household are important, yet servants less so.

The fourth principal component (PC4, Figure 5) is characterised by the following attributes: *large household size*; low proportions of *households with residential kin*; high proportions of *households with residential servants*; low percentages of *males working in agriculture*; high proportions of *households with six or more offspring*; low proportions of *children (ages from 17 to 22) living in the parental home;* low proportions *living with siblings, aunts or uncles, nieces or nephews, cousins, grandparents* and *grandchildren*; and relatively high levels of *occupation concentration* and *surname concentration*. In combination, these components suggest non-agricultural parishes yet with relatively little variation in the local economy/occupational structure and a fairly 'stable' non-migratory population, with large households in which young children are a key element (suggesting maybe higher fertility). These characteristics could indicate mining and similar 'mono-culture' communities.

Figure 2 *Contribution of the reduced data attributes to PC1*

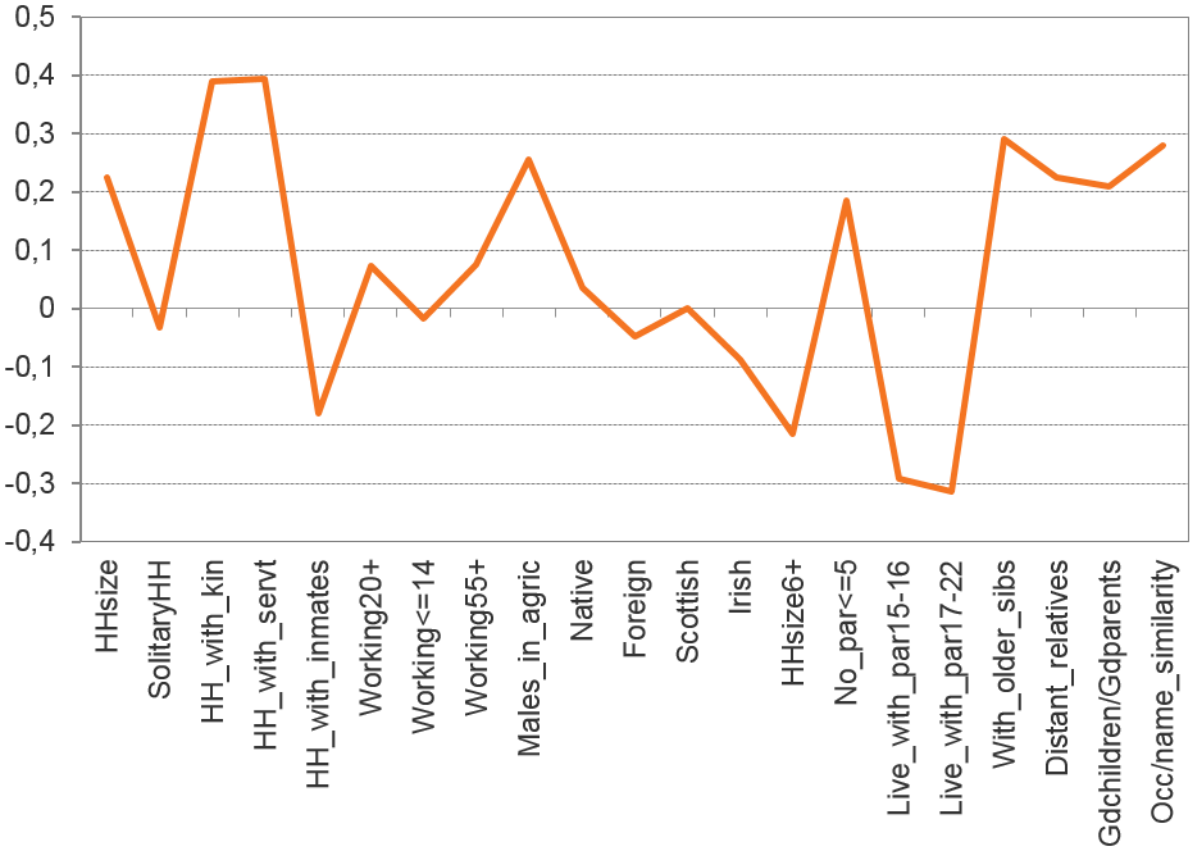


Figure 3 *Contribution of the reduced data attributes to PC2*

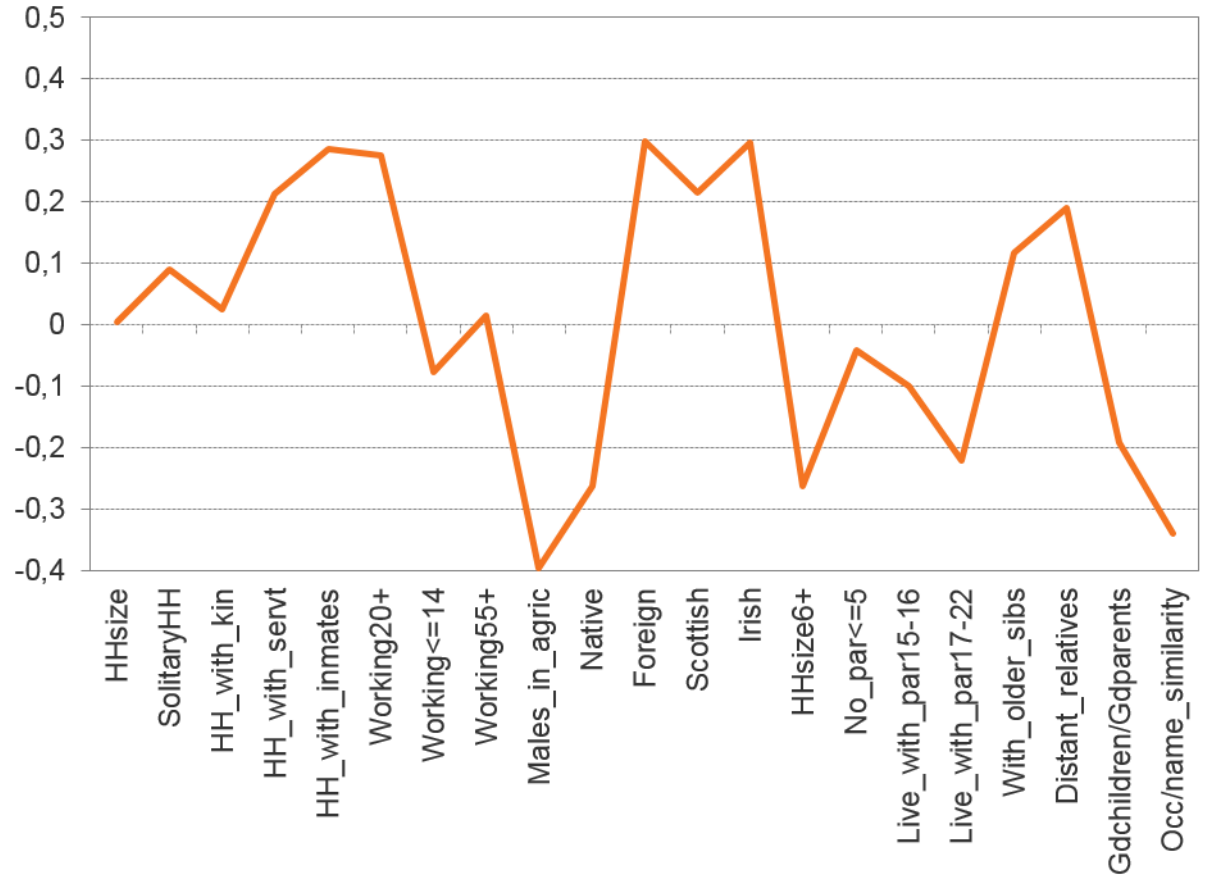


Figure 4 *Contribution of the reduced data attributes to PC3*

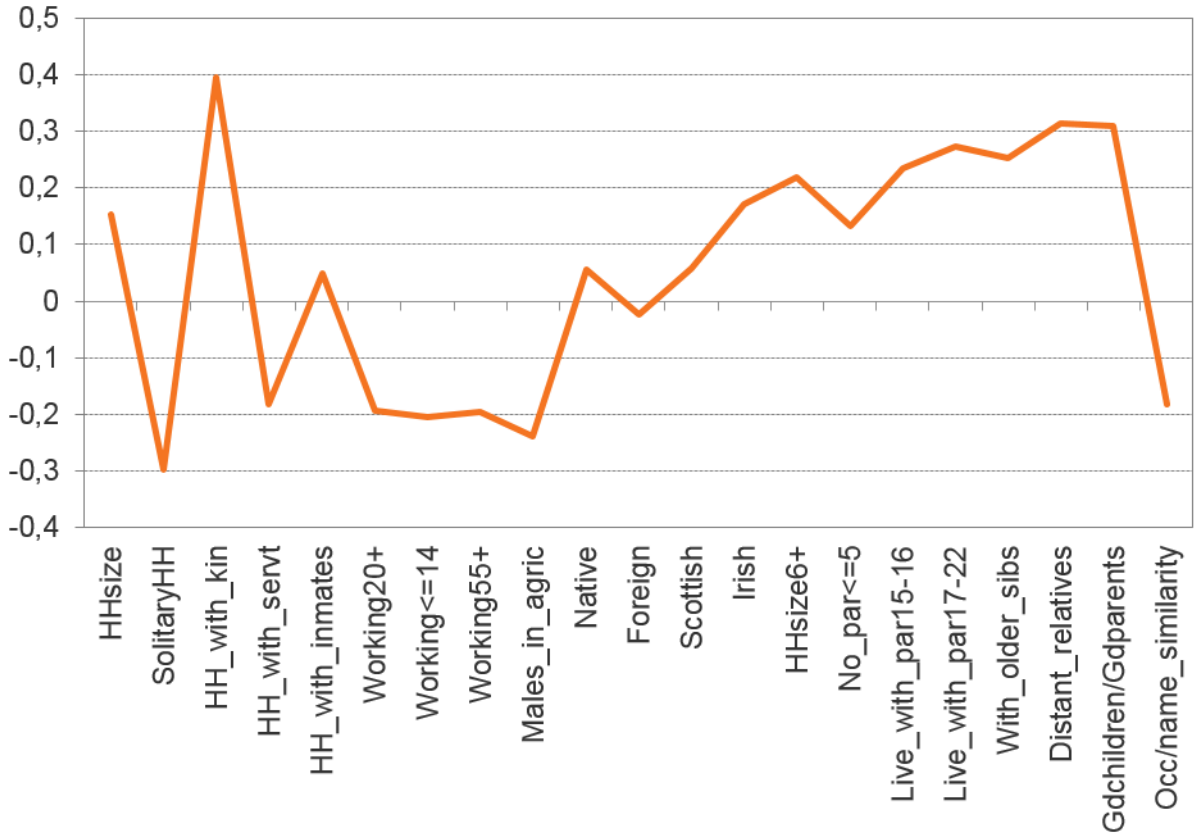
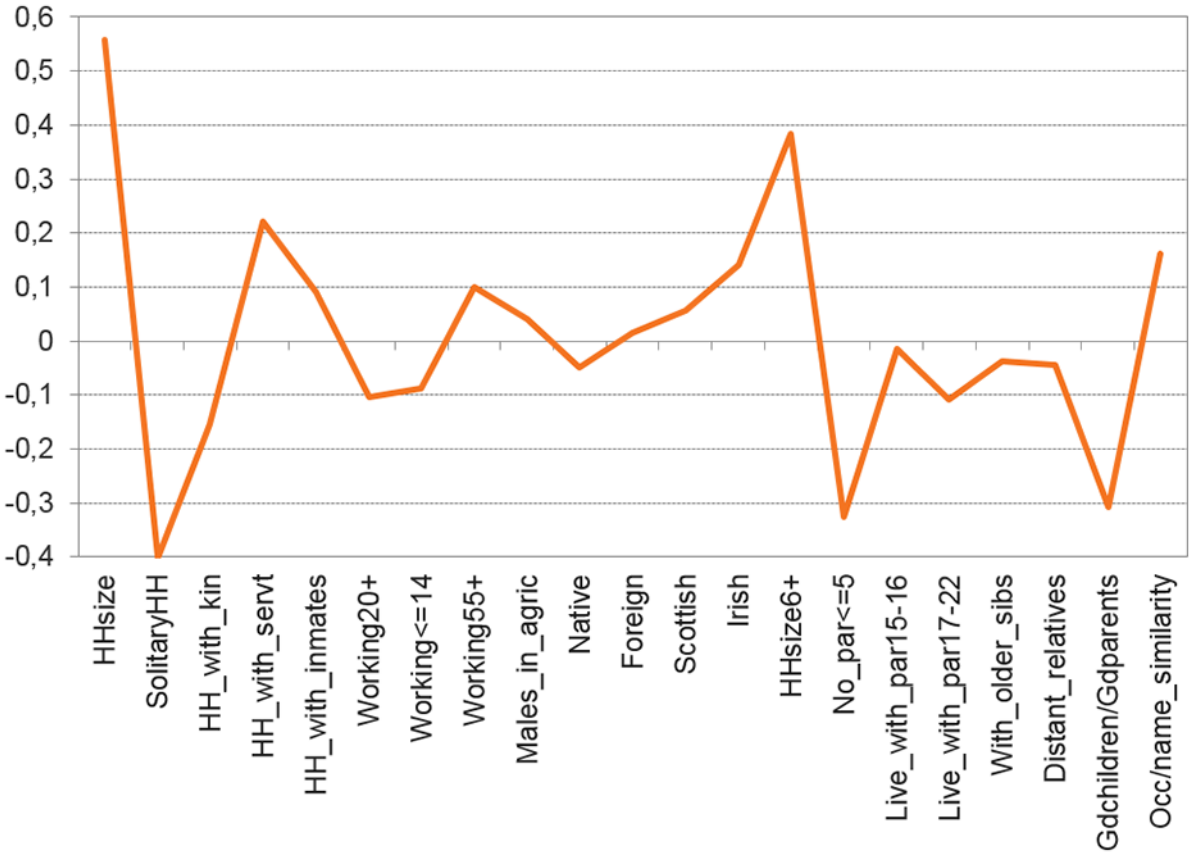


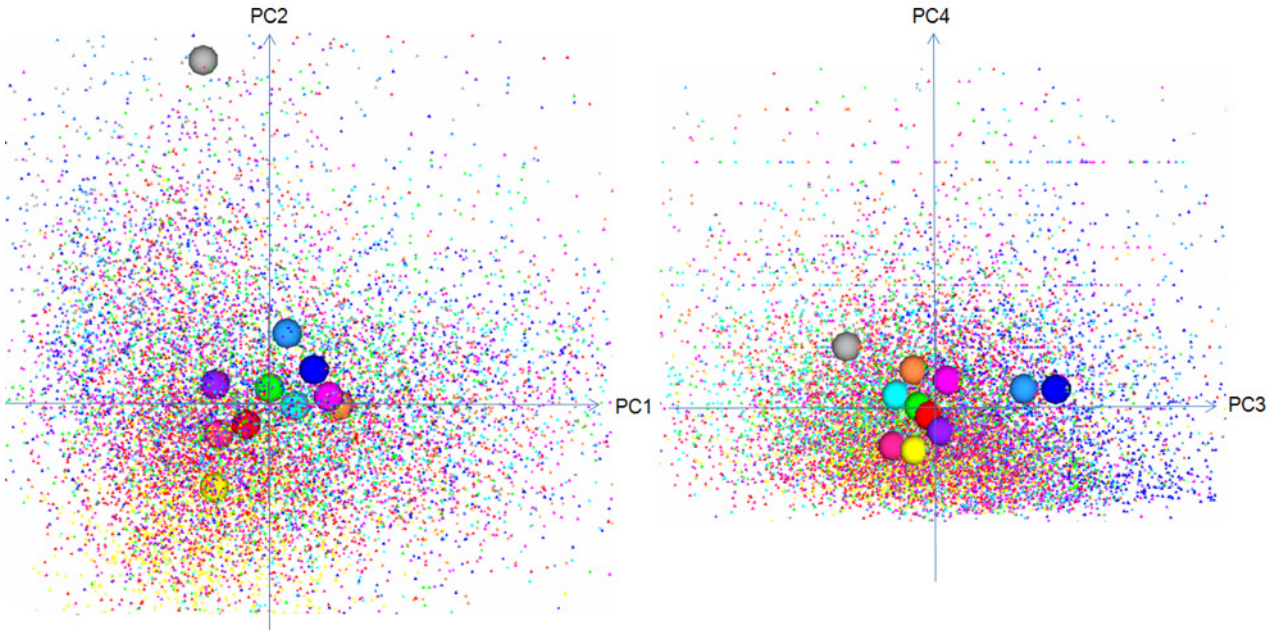
Figure 5 *Contribution of the reduced data attributes to PC4*



4.3 DATA DISTRIBUTION ON THE PRINCIPAL COMPONENTS

The data can be divided into eleven groups according to where the objects (parishes) are located in terms of Standard Regions. These are: group 1 (blue) – North, 941 objects; group 2 (rose) – Yorkshire, 1407 objects; group 3 (light blue) – North Western, 874 objects; group 4 (turquoise) – North Midland, 1546 objects; group 5 (brown) – Monmouth/Wales, 1093 objects; group 6 (green) – West Midland, 1515 objects; group 7 (red) – South Western, 1696 objects; group 8 (crimson) – South Midland, 1318 objects; group 9 (purple) – South East, 1371 objects; group 10 (yellow) – Eastern, 1473 objects; group 11 (grey) – London, 156 objects. Figure 6 shows the visualisation of these eleven standard geographic regions on the PCA plot.

Figure 6 *Visualisation of geographic regions on the PCA plot*



As can be seen from Figure 6, regions such as Wales, Yorkshire, North, North Midland, West Midland, South Western, South Midland and South East are mainly distributed along the first principal component (PC1), while Eastern and London are associated with the second principal component (PC2). The North, North Western and London differ from other regions by the third principal component (PC3), while Wales, Yorkshire, North Midland, West Midland, South Western, South Midland, South East and Eastern are distributed along the fourth principal component (PC4).

According to values of principal component projections, the data were divided into five groups. The results of data grouping are represented in Table 2.

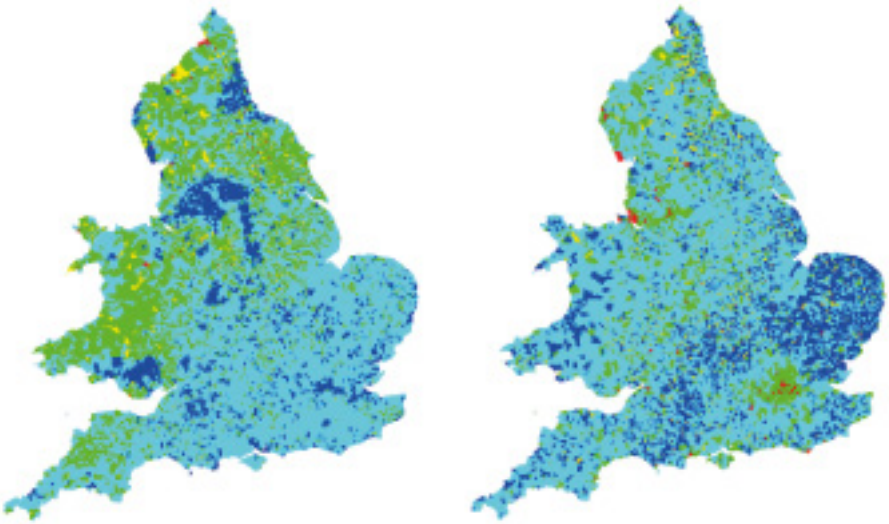
Table 2 *Data grouping according to values of principal component projections*

GROUPS	PC1	PC2	PC3	PC4
Group 1 (blue)	-4.482, -1.803	-5.948, -1.004	-11.681,-3.888	-8.334, -2.927
n. of objects	8,185	3,168	106	153
Group 2 (light blue)	-1.803, 0.874	-1.004, 1.467	-3.888, -1.342	-2.927, -0.232
n. of objects	1,679	8,417	1,848	6,029
Group 3 (green)	0.874, 3.551	1.467, 3.943	-1.342, 1.243	-0.232, 2.474
n. of objects	3,014	1,541	9,208	6,567
Group 4 (yellow)	0.551, 6.245	3.943, 6.426	1.243, 3.828	2.474, 5.217
n. of objects	440	225	2,143	586
Group 5 (red)	6.245, 14.261	6.426, 16.298	3.828, 8.997	5.217, 13.271
n. of objects	72	39	85	55

Figure 7 displays the visualisation of the projections on the first and second principal components based on geographic coordinates. As can be seen from the visualisation of the first component (Figure 7, left), the low values of projections (light blue points) are dominant in the southern part of England; the high values of projections (green, yellow and red points) dominate in the northern part of England and in Wales. Besides, the lowest values (blue points) are concentrated as big cities across the country.

The visualisation of the second component (Figure 7, right) demonstrates that the low values of projections (blue points) also are dominant in the southern part of England; the high values of projections (light blue points) dominate in northern part of England and in Wales. However, the highest values (yellow and red points) are concentrated in larger towns and cities across the country in the southern part of England; the high values of projections (green, yellow and red points) dominate in the northern part of England and in Wales. Besides, the lowest values (blue points) are concentrated as big cities across the country.

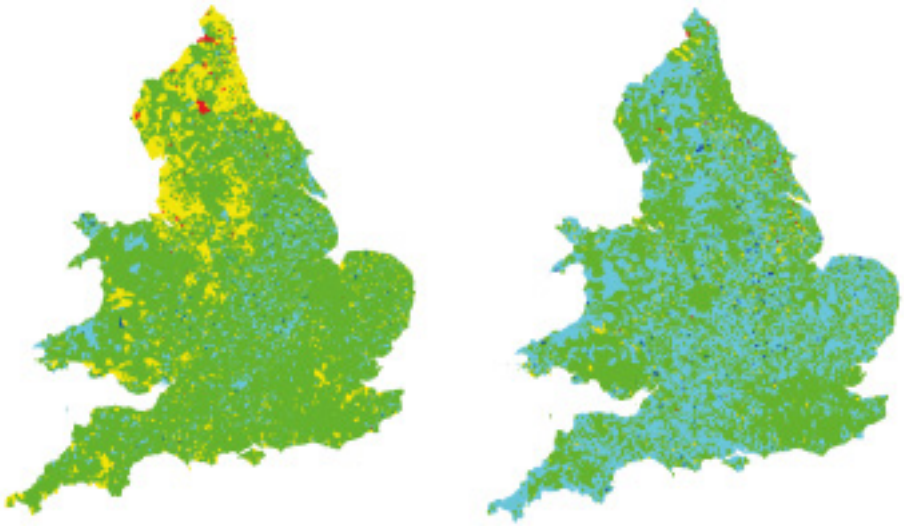
Figure 7 *Visualisation of the projections on the first and second principal components on the geographic coordinates*



Note: *see Table 2 and text for the colour assignment*

Figure 8 displays the visualisation of the projections on the third and fourth principal components on the geographic coordinates. The visualisation of the third component (Figure 8, left) illustrates that objects with high values of projections (yellow and red points) are observed primarily in the north of the country, while objects with low values of projections are occur mainly in central England (light blue and blue points) and the south (green points). Also, we can notice that objects with high values are concentrated in large towns and cities. The visualisation of the fourth component (Figure 8, right) illustrates that objects with high values of projections (yellow and red points) occur mainly in the north of the country, while objects with low values of projections are observed in the central and southern regions. We also can see objects with higher values in the larger towns and cities across the country.

Figure 8 *Visualisation of the projections on the third and fourth principal components on the geographic coordinates*



Note: see Table 2 and text for the colour assignment

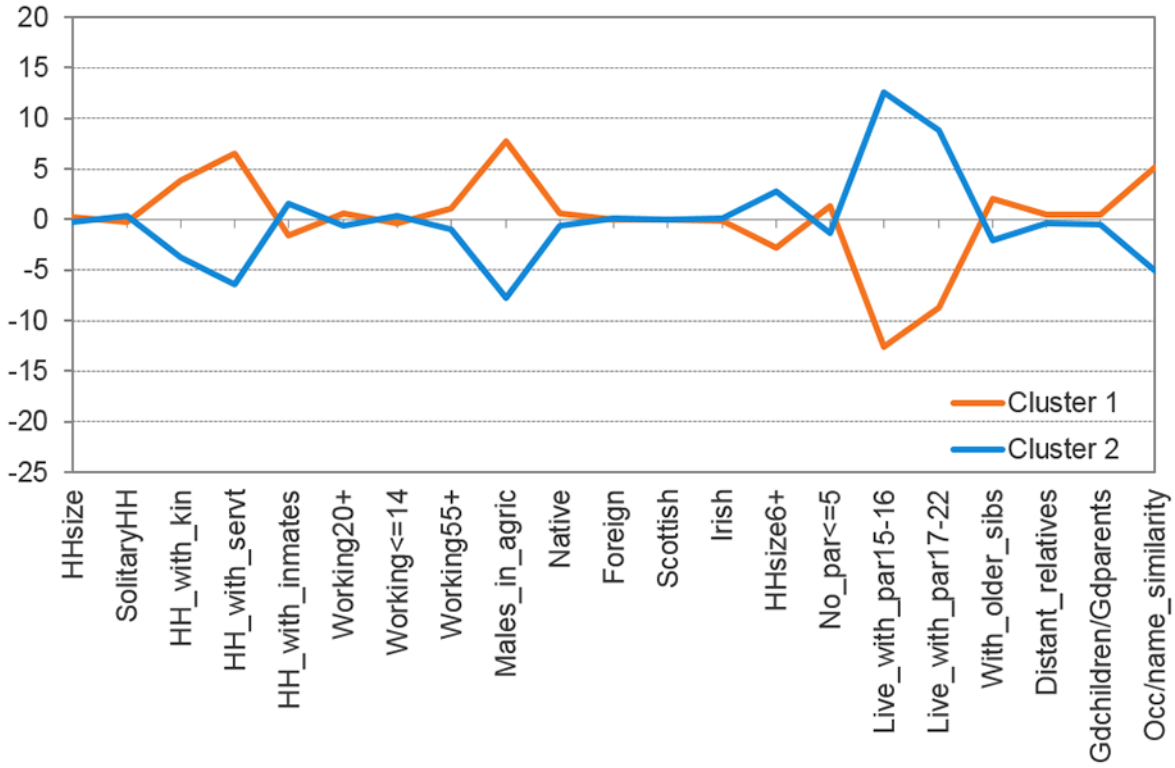
5 CLUSTER ANALYSIS

Cluster analysis is a tool for discovering and identifying associations and structure within the data and typology development (MacQueen 1967). Cluster analysis provides insight into the data by dividing the dataset of objects into groups (clusters) of objects, such that objects in a cluster are more similar to each other than to objects in other clusters. At present, there are many various clustering algorithms which are categorized based on their cluster model (Jain & Dubes 1988). In this research, for cluster analysis of census data the centroid-based clustering method is used. *K*-means is a well-known and widely used clustering method which aims to partition objects based on attributes into *k* clusters. The *k*-means clustering is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid. The centroid can be interpreted as a prototypical point for this cluster. The *K*-means method has two key features: 1) Euclidean distance is used as a metric and variance is used as a measure of cluster scatter; 2) the number of clusters (*k*) is an input parameter which should be specified in advance. For the *k*-means clustering method the most important and difficult question is the identification of the number of clusters that should be considered. In this case, in order to determine the number of clusters the PCA technique was used: the number of clusters being dependent upon the number of principal components. Thus, referring back to the previous discussion, the first component forms two clusters, second component forms three clusters, and so on. According to the eigenvalues of components (Figure 1 above) there are 1-4 principal components. This means that the data has 2-5-cluster structures, where k=5, is the maximum number of informative (significant) clusters.

5.1 TWO-CLUSTER STRUCTURE

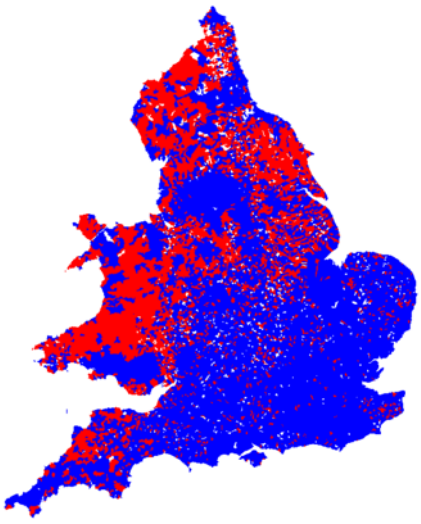
In the two-cluster structure (k=2) cluster 1 (blue) has 9,118 objects and cluster 2 (orange) has 4,272 objects.The difference between clusters is identified by the standard deviation of cluster averages of attributes. Figure 9 shows the distribution of the clustered data on the attributes in two-cluster structure.

Figure 9 *Distribution of the clustered data on the attributes in two-cluster structure*



As can be seen, the two clusters differ significantly on such characteristics as: *households with residential kin*; *households with residential servants*; *males working in agriculture; households with six or more offspring*; *children living in the parental home* and *occupation/surname concentration*. Cluster 1 is characterized by *high proportions of households with residential kin, households with residential servants* and *males working in agriculture*; low proportions of *children living in the parental home*; slightly lower values of *households with six or more offspring* and moderate proportions of *occupation/surname concentration*. Cluster 2 is the mirror image of this pattern. The distribution of the clustered data on the regions in two-cluster structure is represented in Figure 10. As can be seen, the elements of Cluster 1 dominate southern England, and run through the midlands, while the elements of cluster 2 dominate in east and north Yorkshire, the north-west around Cumbria, south Lancashire, Wales, and curiously, Devon.

Figure 10 *Two-cluster structure on the geographic coordinates*

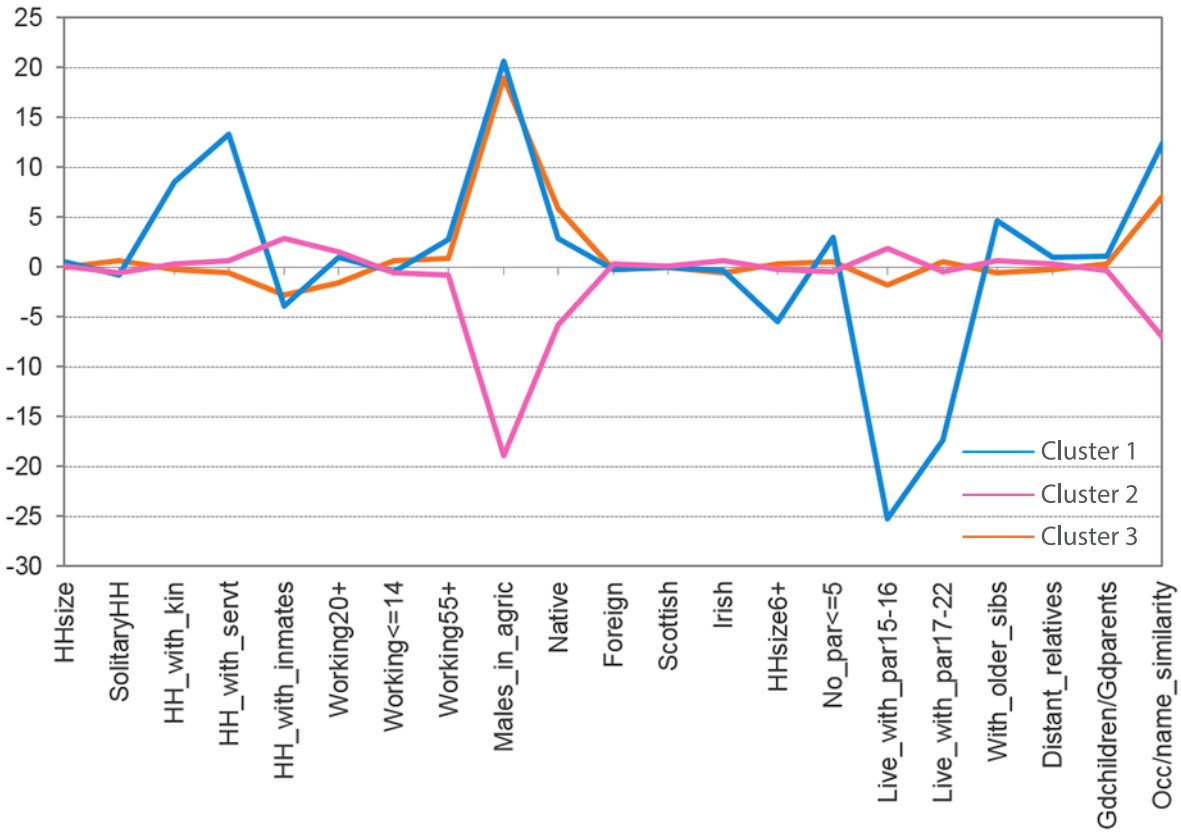


Note: Cluster 1 = blue, Cluster 2=red

5.2 THREE-CLUSTER STRUCTURE

The clusters within the three-cluster structure (*k*=3) are: cluster 1 (blue) with 6,662 objects, cluster 2 (pink) with 3,353 objects and cluster 3 (orange) with 3,375 objects. Figure 11 shows the distribution of the clustered data on the attributes in three-cluster structure.

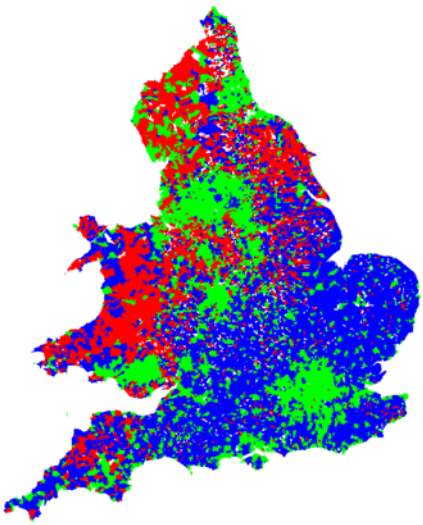
Figure 11 *Distribution of the clustered data on the attributes in three-cluster structure*



As can be seen, the three clusters are significantly different on such characteristics as: *households with residential kin*; *households with residential servants*; *males working in agriculture*; *children living in the parental home*. Each cluster is different and cluster 3 is dramatically different from the other two clusters. Cluster 3 is characterized by high proportions of *households with residential kin*, *households with residential servants* and *males working in agriculture*; and low proportions of *children living in the parental home*; *households with six or more offspring*; as well as a high value for *occupation concentration* and *surname similarity*.

In contrast, clusters 1 and 2 tend to differ from cluster 3 on all the key attributes mentioned above, with the exception of cluster 1 having similar experience in *males working in agriculture*. In contrast, cluster 2 stands out as having low levels of *males working in agriculture* and a correspondingly low value for *occupation concentration* and *surname similarity*, which in combination would suggest that this cluster is mainly urban. Figure 12 illustrates the geographical distribution of the separate clusters. As can be seen, in combination these nuance the two cluster model described earlier. Cluster 2 in the three cluster structure essentially removes the predominantly urban places from cluster 1 of the two cluster structure discussed earlier, leaving a basic north-south divide represented by clusters 1 and 3 – roughly diagonal Severn-Wash line – although north Devon again stands out.

Figure 12 *Three-cluster structure on the geographic coordinates*

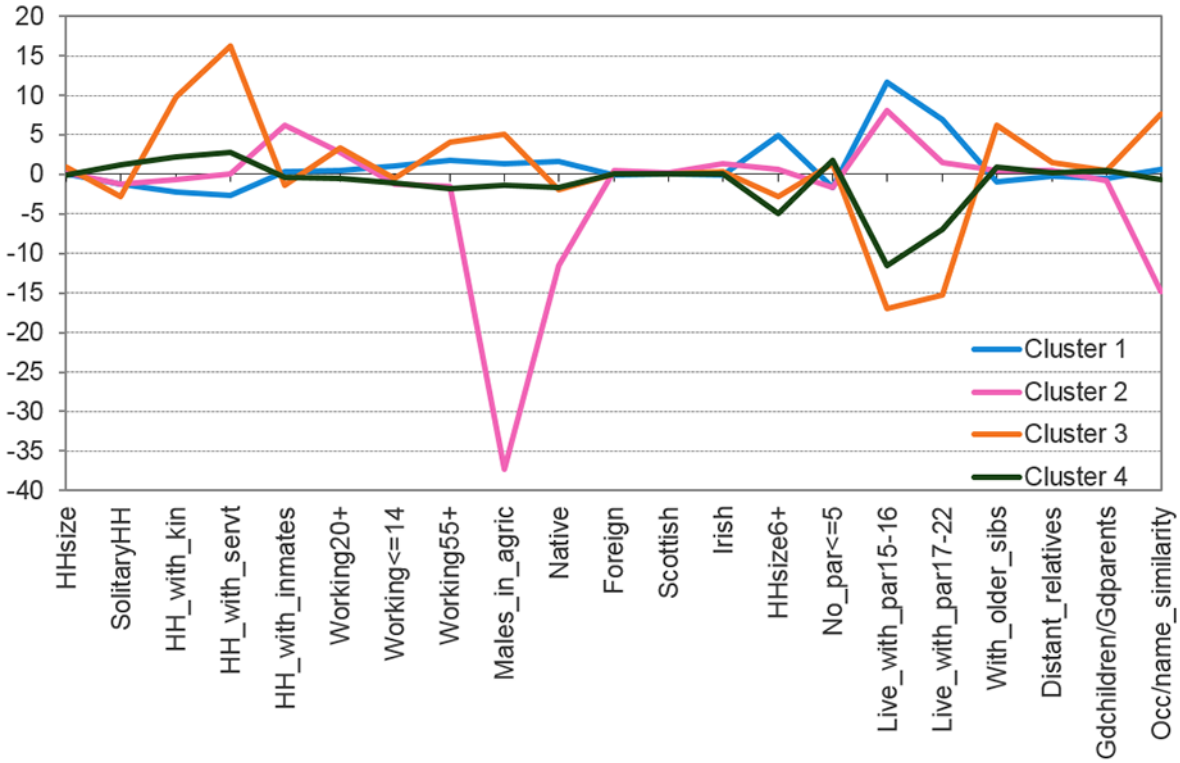


Note: *Cluster 1 = blue, Cluster 2 = green, Cluster 3 = red*

5.3 FOUR-CLUSTER STRUCTURE

The four-cluster structure (k=4) is as follows: cluster 1 (blue) with 4,350 objects, cluster 2 (pink) with 2,992 objects, cluster 3 (green) with 4,096 objects, and cluster 4 (orange) with 1,952 objects. Figure 13 shows the distribution of the clustered data in relation to the attributes within the four-cluster structure.

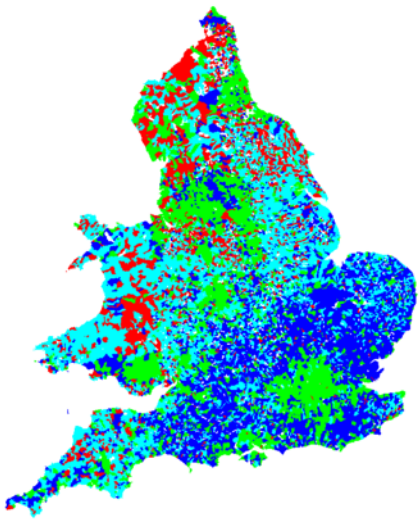
Figure 13 *Distribution of the clustered data on the attributes in four-cluster structure*



As can be seen, the four clusters differ considerably around the following key characteristics: *households with residential kin*; *households with residential servants*; *households with unrelated persons*; *males working in agriculture*; *households with 6 or more offspring*; *children living in the parental home* and *occupation concentration* and *surname similarity*. Cluster 1 is characterised by high proportions of *households with six or more offspring* and high retention of children within the parental home, and conversely, low proportions of *households with servants*. Cluster 2 shows what might be seen as common characteristics of urban populations: significantly low proportions of *males working in agriculture* together with a low value for *natives* and very low value for *occupation concentration* and *surname similarity*. Also this cluster displays relatively high proportions of *households with unrelated persons* (boarders and lodgers) and *children living in the parental home* (aged 15-16). Cluster 3 is in some respects the mirror image of Cluster 1. It has low proportions of *households with six or more offspring*, relatively low retention of *children living in the parental home*, together with relatively high proportions of *households with residential kin* and *servants*. Lastly, cluster 4 is conversely characterised by high proportions of *households with residential kin* and *servants*, together with relatively high proportions of *males working in agriculture* and *elderly workers*. Equally, the proportions of *households with six or more offspring* and *children living in the parental home* is low, while the proportion *households with elderly siblings* living together is relatively higher and the value for *occupation concentration* and *surname similarity* is comparatively very high. These characteristics suggest rural places dominated by mono-cultures.

Figure 14 maps the geographic distribution of the 4 clusters. This illustrates, as already indicated, that cluster 2 within the four-cluster structure is primarily composed of larger urban communities, distributed across the country. In contrast, cluster 1 features mainly in southern rural England, but interestingly, moving from the three to four cluster structure suggests a split between the south-west (Cornwall and Devon), and the rest of southern England (south of the Severn-Wash) line. The south-west joins cluster 3 in this model, in a mainly rural northern England/Wale grouping, but within which parts of East Anglia are also represented. Lastly cluster 4 parishes are located mainly in the north of England, with especially heavy concentrations in south Lancashire, Northumberland, Durham and East Yorkshire. In part, it is tempting to suggest that this cluster could be influenced by the existence of mining industries, but Figure 14 indicates that this is not exclusively mining.

Figure 14 *Four-cluster structure on the geographic coordinates*

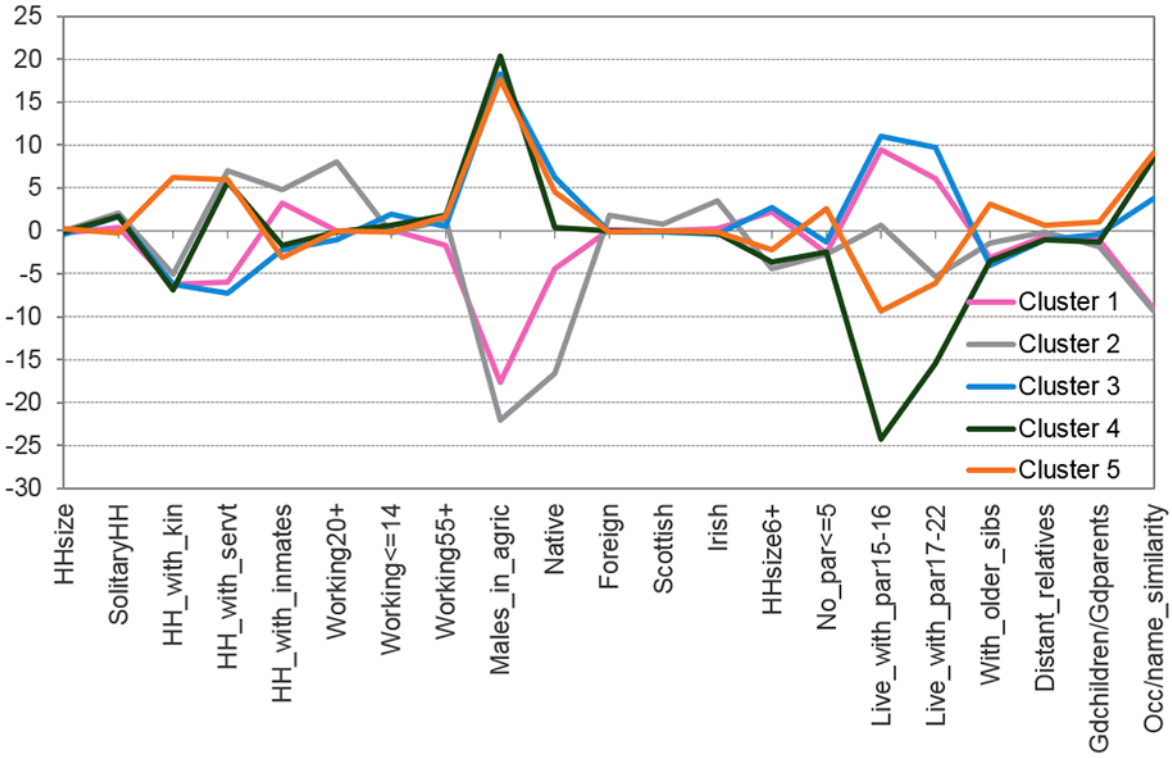


Note: Cluster 1 = dark blue, Cluster 2 = green, Cluster 3 = light blue, Cluster 4 = red

5.4 FIVE-CLUSTER STRUCTURE

The breakdown of the five-cluster structure (*k*=5) is as follows: cluster 1 (blue) with 4,789 objects, cluster 2 (pink) with 3,462 objects, cluster 3 (gray) with 543 objects, cluster 4 (green) with 2,511, and cluster 5 (orange) with 2,085 objects. Figure 15 shows the distribution of the clustered data in relation to the attributes in five-cluster structure. As can be seen, the five clusters different significantly around the following characteristics: *households with residential kin*; *households with residential servants*; *households with unrelated persons*; *population ages 20 are working*; *males working in agriculture*; *native*; *households with 6 or more offspring*; *children living in the parental home* and *occupation concentration* and *surname similarity*.

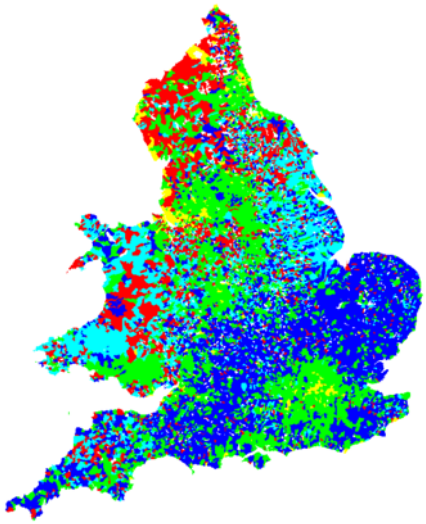
Figure 15 *Distribution of the clustered data on the attributes in five-cluster structure*



Cluster 1 is characterised by moderately low proportions of *households with residential kin* and *servants*; low proportions of *households with unrelated persons*; high proportions of *males working in agriculture* and *children living in the parental home*; moderately high proportions for *occupation concentration* and *surname similarity*; and slightly higher proportions of *households with 6 or more offspring.* In contrast, cluster 2 is characterized by very low proportions of both *males working in agriculture* and *natives*; moderately low proportions of *households with residential kin*; *servants* and for *occupation concentration* and *surname similarity*, together with high proportions of *children (15-22) living in the parental home*. Cluster 5 is virtually the mirror of the cluster 2 experience. Cluster 3, like cluster 2, has very low proportions of both *males working in agriculture* and *natives*, more so than cluster 2 especially in relation to *natives*; moderately low proportions of *households with residential kin* and for *occupation concentration* and *surname similarity;* slightly low proportion of *children (17-22) living in the parental home* and *households with 6 or more offspring*; yet moderately high proportions *households with residential servants*; *unrelated persons* and *those aged 20+ working*. Lastly, cluster 4, likes clusters 1 and 5, has high proportions of *males working in agriculture*; moderately high proportions of *households with residential servants* and *occupation concentration* and surname similarity; together with very low proportions of children (15-16) *living in the parental home* and to a lesser extent aged 17-22; and moderately low proportions of *households with residential kin*.

The five cluster map (Figure 16) still has cluster 1 (blue) dominating in the south of England, in a north-south divide running from the Severn to the Wash, except for the extreme south-west. In comparison to the four cluster model, the extra-metropolitan area around London falling into cluster 2 (green) is even more pronounced, especially around Surrey and Middlesex, while parts of inner London fall with cluster 3 (yellow) characterised mainly by large urban city centres, yet not exclusively so. In addition to extra-metropolitan London, cluster 4 also links to the northern counties or Durham, south Lancashire, west Yorkshire, Cheshire and down to Derby and parts of the West Midlands, as well as, Glamorgan in south Wales. This cluster would appear to represent mixed, mainly urbanised industrial economies. This is partly shadowed by cluster 5 (red) which is less urban, less industrial but is mainly northern, predominating in Cumbria, Northumbria, north Yorkshire and north Lancashire, yet with few clear concentrations. Lastly, cluster 4 (light blue) would also appear to be predominantly rural, being focused in Wales, the east of England north of the Wash, especially Lincolnshire and east Yorkshire, as well as the south-west.

Figure 16 *Five-cluster structure on the geographic coordinates*



Note: Cluster1 = dark blue, Cluster 2 = green, Cluster 3 = yellow, Cluster 4 = light blue, Cluster 5 = red

6 CONCLUSION

So what do all of these statistics and these maps tell us? Turning first to Wall's analysis of 1851 census data, aggregated by standard regions, which focused primarily on household complexity in terms of kinship, this revealed few clear patterns. However, in general terms Wales and south-west of England had the lowest levels of complexity, northern England the highest, and with eastern England being roughly in the middle (Wall 1977). Again focusing on household structural complexity, by 1981 this changed significantly, reversing in some instances. A basic dividing line could be seen running east from the Bristol Channel to the Cotswolds then turning northwards along the spine of the Pennies before heading east towards the Irish Sea below Cumbria (Wall 1982). West of this line household complexity was generally higher than to the east of the line: East Anglia recorded the lowest levels of complexity, but breaking away from this general dichotomy, London was associated with high levels of household complexity. A regional analysis of Marriage Duty Act data for the late seventeenth century, which has only patchy national coverage, revealed little in terms of clear regional variations, yet did demonstrate the distinctiveness of London and the importance of rural/urban of a potential dichotomy (Schürer 1992). Moving, to more recent trends, analysis of the 1991 census data suggests that the percentage of one person households was generally low across the western and central counties of England, slightly higher north of a line running from the rivers Mersey to Humber, including Wales, with higher percentages also recorded for London, east Sussex, Devon and Dorset. Likewise, the proportions of lone parent households were lower in the eastern counties and higher for a belt running down the centre of England, from Lancashire to Kent, as well as being high in south Wales (Champion, Wong, Rooke, Dorling, Coombes & Brunsdon 1996). More recently, Dor-

ling and Thomas, mapping the 2001 census data for the UK suggest a growing trend towards what they term 'London and the Archipelago'. They argue that the UK is becoming more and more divided, with an imaginary line running from the rivers Severn to Humber separating a growing London metropolis from the rest of the UK. Within the London core, population is more densely concentrated, increasingly becoming younger. To the north of the line within the archipelago, are numerous centres each with their own outer areas and remoter edges. Essentially, the archipelago is an amalgam of places which have most in common in not being in the London metropolis - where, in general, population is less concentrated, often decreasing in numbers, becoming older and focusing on industries that have died or are dying (Dorling & Thomas 2004).

The analysis presented here both confirms elements of the previous work outlined above, yet adds also a much greater level of clarity. It shows that the pattern of regional variation in household structure varies in detail as different levels of complexity are considered. In part, this is like viewing a landscape through the lens of a telescope whilst gradually focusing. At a basic level, the geography of household structure is defined by a two-fold division, with a noticeable north-south divide running diagonally across the country, from the river Severn to the Wash, but taking in parts of the south midlands as well. This is not characterised by a simple urban/rural divide, as both sides of the line contain each of these elements. However, as one focuses further, urbanisation (and industrialisation) does become more of a defining feature. London, and as one focuses further, its extra-metropolitan surroundings, becomes a distinct 'region' – illustrating that the process described by Dorling and Thomas has long historical roots. North of the Severn-Wash divide, rural areas begin to segregate, with the more northerly rural areas showing distinct differences from those in the east and in Wales, with residential kin, in particular, being a key difference between these two rural types, as is the retention of children within the parental home. The evidence of this research suggests that regional variations in the patterns of residential kinship, children at home, the keeping of servants and addition of unrelated household members, such as boarders and lodgers, did exist in nineteenth-century England and Wales independent of urban and industrial drivers.

ACKNOWLEDGEMENTS

The research reported in this paper formed part of the Integrated Census Microdata (I-CeM) project, based at the Department of History, University of Essex. I-CeM was funded by the (UK) Economic and Social Research Council (award RES-061-23-1629) and supported by BrightSolid, who kindly supplied the underlying raw data. Additional work on the data was supported as part of the (Joint Information Systems Committee) JISC-funded project "Mining Microdata: economic opportunity and spatial mobility in Britain, Canada and The United States, 1850-1911". This was undertaken jointly with the University of Alberta, University of Montreal, University of Guelph (all Canada), the Minnesota Population Center at the University of Minnesota (USA). Details are available at http://www.miningmicrodata.org/. The authors would also like to thank Professor Alexander Gorban of the Department of Mathematics, University of Leicester and Andrey Zinoviev of the Institut Curie, Paris.

REFERENCES

Abdi, H. & Williams, L. (2010). Principal Components Analysis. *Wiley Interdisciplinary Reviews: Computational Statistics,* 2(4), (pp 439-459). DOI: 10.1002/wics.101

Champion, T., Wong, C., Rooke, A., Dorling, D., Coombes, M. & Brunsdon, C. (1996). *The Population of Britain in the 1990s. A social and economic atlas.* Oxford: Clarendon Press.

Dorling, D. & Thomas, B. (2004). *People and places. A 2001 Census atlas of the UK*. Bristol: Policy Press

Garrett, E., Reid, A., Schürer, K. & Szreter, S. (2001). *Changing Family Size in England and Wales. Place, Class and Demography, 1891-1911*. Cambridge: Cambridge University Press.

Gorban, A. N. & Zinovyev, A. Y. (2009). Principal Graphs and Manifolds, In: E.S. Olivas, J.D.M. Guererro, M.M. Sober, J.R.M. Benedito & A.J.S. Lopes (Eds.) *Handbook of Research on Machine Learning Applicat ions and Trends: Algorithms, Methods and Techniques*, (pp 28-59) IGI Global: Hershey,

PA, USA.
DOI: 10.4018/978-1-60566-766-9.
Jain A. & Dubes R. (1988). *Algorithms for Clustering Data*. Michigan State University: Prentice Hall.
Laslett, P. (1969). Size and Structure of the Household in England Over Three Centuries. *Population Studies,* 23(2), 199-223.
DOI:10.1080/00324728.1969.10405278
Laslett, P. (1972). Introduction. In: Laslett, P. with the assistance of Wall, R. (Eds.), *Household and family in past time. Comparative studies in the size and structure of the domestic group over the last three centuries in England, France, Serbia, Japan and colonial North America, with further materials from Western Europe*, (pp.1-89). Cambridge: Cambridge University Press.
Laslett, P. (1983). Family and household as work group and kin group: areas of traditional Europe com pared. In: Wall, R. in collaboration with Robin, J. and Laslett, P. (Eds.) *Family forms in historic Europe,* (pp 513-563). Cambridge: Cambridge University Press.
Laslett, P. (1985). Review. *Population and Development Review*, 11(3), 534-537.
MacQueen J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, Vol. I, Statistics, (pp 281–297). Berkeley: University of California Press.
Peres-Neto, P., Jackson, D. & Somers, K. (2005). How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics & Data Analysis*, 49(4), 974-997.
DOI: 10.1016/j.csda.2004.06.015
Ruggles, S. (2012). The Future of Historical Family Demography. *Annual Review of Sociology*, 38, 423-441.
DOI: 10.1145/annurev-soc-071811-145533.
Schürer, K. (1992). Variations in household structure in the late seventeenth century: towards a regional analysis. In: K. Schürer and T. Arkell, (Eds.) S*urveying the People. The interpretation and use of document sources for the study of population in the later seventeenth century*, (pp 253-278) Oxford: Leopard's Head.
Schürer, K. & Woollard, M. (2000). *1881 Census for England and Wales, the Channel Islands and the Isle of Man (Enhanced Version)* [computer file]. Genealogical Society of Utah, Federation of Family History Societies, [original data producer(s)]. Colchester, Essex: UK Data Archive [distributor].
DOI: 10.5255/UKDA-SN-4177-1.
Schürer, K. & Woollard, M. (2002). *National Sample from the 1881 Census of Great Britain 5% Random Sample: working documentation version 1.1.* Colchester: University of Essex, Historical Censuses and Social Surveys Research Group.
Szołtysek, M., Gruber, S., Klüsener, S. & Goldstein, J. R. (2014). Spatial Variation in Household Structures in Nineteenth-Century Germany. *Population-E,* 69(1) 55-80.
Teitelbaum, M. S. (1984). *The British fertility decline: demographic transition in the crucible of the Industrial Revolution*. Princeton: Princeton University Press.
Wall, R. (1977). Regional and temporal variations in household structure from 1650. In: J. Hobcraft and P. Rees, (Eds.) *Regional demographic development,* 89-113) London.
Wall, R. (1982). Regional and temporal variations in the structure of the British household since 1851. In: T. Barker and M. Drake, (Eds.) *Population and society in Britain 1850-1980,* (pp 62-99). London.
Wall, R. (1983). The household: demographic and economic change in England, 1650-1970. In: Wall, R. in collaboration with Robin, J. and Laslett, P. (Eds.) *Family forms in historic Europe,* (pp 493-512). Cambridge: Cambridge University Press.
Woods, R. & Shelton, N. (1997). *An Atlas of Victorian Mortality.* Liverpool University Press: Liverpool.
Woods, R. (2000). *The Demography of Victorian England and Wales*. Cambridge: Cambridge University Press.
Wrigley, E. A. (1985). The fall of marital fertility in nineteenth-century France: Exemplar or exception? (Part II). *European Journal of Population*, 1, 141-177.
DOI: 10.1007/BF01796931
Wrigley, E. A. & Schofield, R. S. (1983). English population history from family reconstitution: summary results 1600-1799. *Population Studies,* 37, 157-184.
DOI: 10.1080/00324728.1983.10408745
Zinovyev A. (2000). ViDaExpert – multidimensional data visualization tool. Institute Curie, Paris.

Mining Microdata: Economic Opportunity and Spatial Mobility in Britain and the United States, 1850-1881

Peter Baskerville
Department of History
University of Alberta
Edmonton, Canada
pab@uvic.ca

Steven Ruggles
Department of History & Minnesota Population Center
University of Minnesota
Minneapolis, United States of America
ruggles@umn.edu

Lisa Dillon
Department of Demography
Université de Montréal
Montréal, Canada
ly.dillon@umontreal.ca

Kevin Schürer
Department of History
University of Leicester
Leicester, United Kingdom
ks291@leicester.ac.uk

Kris Inwood
Departments of Economics and History
University of Guelph
Guelph, Canada
kinwood@uoguelph.ca

John Robert Warren
Department of History & Minnesota Population Center
University of Minnesota
Minneapolis, United States of America
warre046@umn.edu

Evan Roberts
Department of History & Minnesota Population Center
University of Minnesota
Minneapolis, United States of America
eroberts@umn.edu

***Abstract*— For almost two centuries social theorists have argued that the fundamental difference in social structure between Europe and North America arises from greater economic and geographic mobility in North America. We study social mobility in three countries across two generations using machine learning techniques to create panels of individuals linked between censuses thirty years apart (1850-1880, 1880-1910). This paper reports on a preliminary analysis of social mobility between 1850 and 1880, finding that mobility was markedly higher in the United States and Canada, compared to Great Britain.**

Keywords—machine learning; social mobility; census

I. Introduction

For almost two centuries, social theorists have argued that differences in economic opportunity and geographic mobility on the two sides of the Atlantic led to fundamental differences in social structure. In the opening line of *Democracy in America,* de Tocqueville stated that "no novelty in the United States struck me more vividly during my stay there than the equality of conditions"[1]. When he visited Canada, de Tocqueville found "the spirit of equality and democracy alive there as in the United States"[2]. Explaining why Americans were "restless in the midst of their prosperity," de Tocqueville expressed amazement at their rootless mobility, claiming that "a man will carefully construct a home in which to spend his old age and sell it before the roof is on . . . He will settle in one place only to go off elsewhere shortly afterwards with a new set of desires" [1]. Nineteenth-century commentators from de Tocqueville to the historian Frederick Jackson Turner maintained that the exceptional level of North American economic mobility was closely tied to geographic mobility: the availability of cheap land in North America allowed economic advancement and promoted high migration [3]. Westward expansion created a "safety valve," which many observers saw as the chief explanation for the failure of the socialist movement in North America [4-13].

In the twentieth century, Canadian and U.S. historians challenged this interpretation. Using linked censuses of more than a dozen communities, historians in both countries argued that despite high geographic mobility, nineteenth-century North America had a rigid class structure with comparatively little upward mobility [14-26]. Some suggested that migrants constituted a "floating proletariat" of declining fortune [17]. In recent years, however, a few studies using national data have argued that the nineteenth-century United States was extremely

fluid compared with nineteenth-century England [27]. The new results suggest that there has been a dramatic decline in the United States in both economic and geographic mobility over the past 150 years. If confirmed, these results would have profound implications for our understanding of social structure and social change on both sides of the Atlantic.

In a project funded by the 2011 application round of the Digging into Data initiative, we apply new data-mining technology to massive new census microdata collections in Britain, Canada, and the United States to address four key questions:

1. What were the relative levels of economic and geographic mobility in Britain, Canada, and the United States in the late nineteenth century?
2. What were the mobility trends in each country?
3. How were economic opportunity and geographic mobility interrelated in each country?
4. What individual and community characteristics were associated with economic and geographic mobility?

II. Data

This project is based on one of the largest microdata collections in existence, the North Atlantic Population Project (NAPP) [28-30]. The NAPP database includes complete enumerations of the populations of Britain, Canada, the United States, and several other countries between 1850 and 1911. The data consist of numerically coded transcriptions of historical censuses for Britain, Canada, and the United States. The files have a hierarchical format, with individuals nested into families and households; within each family and household, the interrelationships of the members are known. The numeric coding system is consistent across countries. Most of the data we intend to use was already incorporated into the NAPP data access system (http://www.nappdata.org) at the inception of the project. The data from which we draw our samples are freely available on the Internet [29]. In addition to the existing NAPP data, during the course of the project, we incorporated new complete-count datasets for Britain in 1911 and a large new sample for Canada in 1852.

Censuses in the United States were conducted every 10 years after 1790. In Canada and Great Britain censuses have been scheduled every 10 years on the '1' years, though Canada's scheduled 1851 census was taken in 1852. Thus our comparison of social mobility over similar generations will be of slightly different years in each country: 1850-1880 and 1880-1910 in the United States, 1852-1881 and 1881-1911 in Canada, and 1851-1881 and 1881-1911 in Great Britain. In the remainder of the text we abbreviate these thirty year intervals as 1850/1-1880/1 and 1880/1-1910/1.

III. Project Goals

The project aims to create representative longitudinal panels of census data in a comparable manner in three countries, and contribute to a long-standing debate on social structure and opportunity in Britain and North America. Given the recent availability of large-scale census databases the challenge *now* in constructing panel data from censuses is the adapting of machine learning teachniques to replace case by case linking pioneered by genealogists. The principal challenge is *not* to find sufficient cases, but ensuring that the panels are representative, unbiased and accurate. False links lead to artifactual social mobility, so it is important to ensure high levels of accuracy. We do this in a similar way across Canada, Great Britain, and the United States taking account of differences in census enumeration methods and questions.

IV. Record linkage approach

Our linkage strategies build on recent research in data mining and machine-learning [31]. The theoretical framework for probabilistic record linkage derives from Fellegi and Sunter, who demonstrated that it is possible to define an optimal linkage rule that minimizes the number of false links [32]. Major extensions and refinements of record-linkage theory were contributed by Jaro, Winkler, Belin, Rubin, and Larsen [33-36]. Recent research has focused on using machine-learning techniques instead of fixed linkage rules [37]

Our record linking procedures build on these innovations. Our goals, however, differ significantly from those of most data mining applications of record linkage. The primary goal of most data mining has been to maximize the number of valid links. Our objective is different: we do not focus on maximizing the linkage rate. Instead, our procedures are designed to maximize the *representativeness* of the linked cases and the *accuracy* of the links. This means we pay close attention to potential sources of selection bias, and ignore information routinely used by other record-linkage procedures. Although we cannot eliminate selection bias for unobserved characteristics, we can adopt procedures that greatly reduce the potential for bias compared with previous approaches.

Our algorithm relies exclusively on characteristics that should not change over time. At minimum, these variables are first name, last name (for men and for women who do not marry between observations), birth year, sex, and place of birth. Most record linkage software makes use of a broader range of characteristics to confirm links and resolve ambiguities, but that approach introduces bias. For example, if we use spouse's characteristics to confirm linkages, we would bias the sample in favor of persons who remained married to the same person for multiple decades, and such persons are not representative with respect to either occupational or geographic mobility. Wisselgren et al provide a recent discussion and evaluation of these issues in historical census record linkage [38].

A challenge posed by our approach is that the limited set of variables we use cannot uniquely identify all individuals. To take the worst-case scenario—the most common male name with the most common birthplace—the 1880 U.S. census has 17 white men aged 33, named John Smith, and born in New York. Even this example understates the problem, because it assumes an exact match of name and age. Errors in enumeration and transcription cause a significant proportion of matches to be imperfect: linking must be carried out probabilistically, allowing for imperfect correspondence of name and age. Whenever there is more than one possible

match, we must exclude all potential matches. This eliminates many true matches, but is necessary to minimize false matches. False matches would lead to systematic upward bias for transition rates—such as migration and occupational mobility—and therefore must be avoided at all cost.

Because our linking strategy relies heavily on names, we need an approximate string comparison algorithm. We use the Jaro string comparator as modified by Winkler [39, 40]. This algorithm computes a similarity measure between 0.0 and 1.0 based on the number of common characters in two strings, the lengths of both strings, and the number of transpositions, accounting for the increased probability of typographical errors towards the end of words. In addition to using a string comparator, we standardize given names to account for diminutives and abbreviations (e.g., "Willie" and "Wm." are transformed into "William.") Such name-cleaning techniques are language-specific and must be customized for each language of enumeration. This work draws on the rich body of research on name cleaning [39-44]. Finally, we use both NYSIIS and Double-Metaphone phonetic name coding, which provide multiple encoded strings corresponding to variant pronunciations [45, 46].

We use two approaches to calculate similarity measures, including Jaro/Winkler indices and age similarity scores. We use both the open-source "Freely Extensible Biomedical Record Linkage" (FEBRL) software [47-50] and a new implementation of distance function routines written by Guelph post-doctoral researcher Luiza Antonie customized for large historical datasets [51-53]. Other linking variables—such as birthplace and sex—do not pose string comparison problems because they are numerically coded to eliminate spelling variation. Thus, for example, we do not worry about the innumerable spelling variations of Aberystwyth, or variant names for the same location.

We assume every pair of records drawn from two files are either matches referring to a single individual or non-matches describing two different persons. Optimal matching requires every individual be compared with every possible match. It is not computationally feasible, however, to assess every potential match. For example, using such a linking algorithm for the full U.S. 1880 census and 1900 U.S. sample would involve over 15 trillion comparisons. To reduce the computational load, we use "blocking factors"—such as birthplace, sex, and race—limiting comparisons to people sharing blocking factors.

To estimate parameters for the record linkage algorithm, we need training data. Training data are cases where true links are known. We obtain training data by having multiple research assistants hand-link the same sets of data, and combine the results to obtain a set of highly-reliable links. We use the training data to implement a Support Vector Machine (SVM) on the full set of unlinked census data to classify each potential match [54-56]. We implement the SVM using the open-source library of tools developed by Chang and Lin [57]. We include in the SVM: Jaro-Winkler scores for the first and last names separately; a Jaro-Winkler score for a standardized first name; indicator values for NYSIIS, Double Metaphone and matching first letter; indicators for middle names; and a measure of age similarity. Comparisons are performed within sex-race-birthplace blocks, so equality on those measures is assumed. Based on the training data, the SVM calculates a confidence score for every potential match; when one and no more than one potential match exceeds the threshold, we establish a link. We have extensively tested our procedures against known links, and we estimate that the false link rate averages less than 3%. Once we have established the full set of links, we weight the cases to represent the potentially linkable population with respect to age, sex, birthplace, whether related to head, occupational group, and size of place in the terminal year.

V. Measuring social mobility

We adopted the Historical International Standard Classification of Occupations (HISCO) as our basic framework for occupational classification.[58-60] The HISCO system is a modification of the 1968 United Nations occupational classification system with extensions to accommodate historical occupations. HISCO was developed by an international committee with representatives from Belgium, Canada, England, France, Germany, the Netherlands, Norway, Sweden and the United States. We modified and extended the HISCO system to accommodate the additional detail available in the North Atlantic database.[61] To ensure that we coded the millions of occupations comparably across each country, we traded random samples of the occupation dictionary across countries, so that part of each country's occupations were independently coded by researchers in each other participating country. We then reconciled all differences of interpretation, which sometimes involved lengthy discussion and debate.

Our measure of social background outcomes is occupation in early adulthood, measured for the subjects' fathers when the subjects are 0-19 years old, and for the subjects at age 30-49. Occupations are the only measure of social and economic status collected in a consistent manner across time and space in pre-World War II statistical sources. While earnings varied within occupations, there is a relatively stable ordering of earnings across occupations over time [62]. We classify our occupations initially into a modified version of the Historical International Standard Classification of Occupations coding scheme and then aggregate occupations into four categories to measure social class [58, 59, 61]. In this paper we combine occupations into four broader groupings for analysis: (1) white collar workers: a broad group encompassing professionals, clerical workers, and sales people, (2) farmers (3) skilled workers or supervisory workers, such as foremen or overseers, and (4) unskilled workers, encompassing various industrial sectors from service work to farming to manufacturing. Our classification mirrors that in Ferrie and Long's recent analysis of social mobility in the same countries [63].

VI. Results

In this paper we report on an initial analysis of social mobility between 1850/1 and 1880/1 in Great Britain and the United States. Our sample for analysis is boys aged 0-19 in 1850/1, who were living with a co-resident father. In both countries we

obtain a sample of slightly under 4000 young men, who we are able to follow into their own adult lives thirty years later. The demographic characteristics of the panels are fairly similar (Great Britain, Table 1; United States, Table 2).

Several demographic aspects of the two samples are interesting. Family size at a comparable stage of the life-course dropped significantly between generations in both countries. In the second generation family size in 1880/1 averaged 5 (prototypically, a husband, wife and three children). Yet in Great Britain these men had come from families with, on average, 1.4 more children in 1851. In the United States, these men had hailed from families with an average family size of 7.2. Thus, the family context of these men became more similar in the second generation. In many other respects the demographic characteristics of the two samples are remarkably similar: compare for example the average ages of fathers and sons, and the fertility of the second generation by 1880/1.

TABLE I. DEMOGRAPHIC CHARACTERISTICS OF BRITISH SAMPLE

Variable	Mean	StdDev	CV	Min	Max	N
Age, 1850/1	8.4	5.7	0.68	0	20	3919
Age, 1880/1	38	5.7	0.15	30	51	3919
Family size, 1850/1	6.4	2	0.32	2	16	3919
Family size, 1880/1	5	2.7	0.54	1	61	3919
Siblings, 1850/1	3.3	2	0.61	0	13	3919
Working kids	0.58	0.93	1.6	0	6	3919
Has kids, 1880/1	0.71	0.45	0.63	0	1	3919
Num. Children, 1880/1	2.6	2.3	0.89	0	9	3919
if has kids	3.7	1.9	0.53	1	9	2800
Youngest child	4	4.6	1.2	0	27	2800
Eldest child	12	5.9	0.51	0	37	2800
Number < 5	0.83	0.99	1.2	0	5	3919
Father's age, 1850/1	42	18	0.43	20	999	3919
Married, 1880/1	0.82	0.39	0.47	0	1	3919

TABLE II. DEMOGRAPHIC CHARACTERISTICS OF AMERICAN SAMPLE

Variable	Mean	StdDev	CV	Min	Max	N
Age, 1850/1	8.8	5.8	0.65	0	20	3715
Age, 1880/1	39	5.8	0.15	30	51	3715
Family size, 1850/1	7.2	2.4	0.33	2	17	3715
Family size, 1880/1	5	2.3	0.46	1	16	3715
Siblings, 1850/1	3.9	2.3	0.59	0	9	3715
Working kids	0.48	0.84	1.7	0	5	3715
Has kids, 1880/1	0.77	0.42	0.55	0	1	3715
Num. Children, 1880/1	2.5	2.2	0.86	0	9	3715
if has kids	3.3	1.9	0.58	1	9	2854
Youngest child	4.6	4.8	1	0	25	2854
Eldest child	12	5.8	0.5	0	31	2854
Number < 5	0.72	0.88	1.2	0	6	3715
Father's age, 1850/1	42	9.6	0.23	17	81	3715
Married, 1880/1	0.86	0.35	0.41	0	1	3715

A. Geographic mobility over thirty years

Particularly in the nineteenth century geographic and social mobility were strongly related. Young men often significant distances to seek new work. Indeed, the restlessness that de Tocqueville and other observers noted about North America was a geographic one. Just over half (52%) of the American sample moved counties between 1850 and 1880. In Britain 36% of men moved counties between 1851 and 1881. Yet this overstates movement in Britain relative to the United States, since the geographic size of British counties was substantially smaller. The pattern of moves was dispersed. No origin-destination pair of states accounted for more than 2.2% of all those who moved. Yet, there was a consistent pattern to geographic mobility in the United States: nearly everyone who moved headed west. Thus, by 1880 the population of this sample had spread widely across the contiguous United States (Fig. 1, Fig. 2). The most common moves in Britain were to adjacent counties, whereas in the United States many movers had skipped entire adjacent states.

Fig. 1. Residence of U.S. sample in 1850

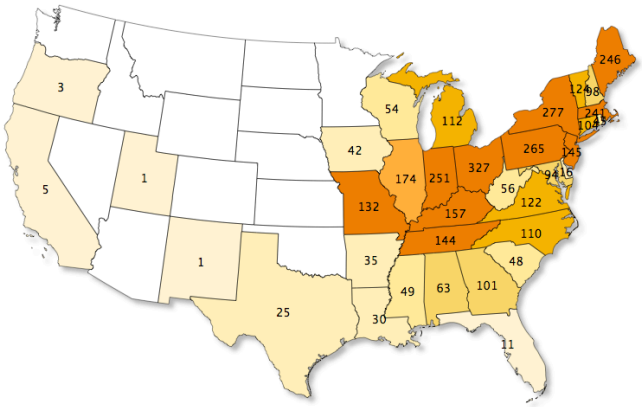
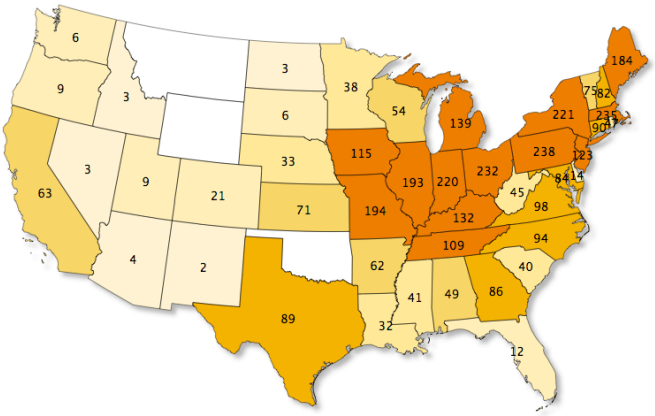


Fig. 2. Residence of U.S. sample in 1880



VII. Occupational Status and Transitions

In both countries, the information on occupations of fathers in 1850/1 and sons in 1880/1 allows rich description and analysis of the pattern of occupational change across generation. As well as the occupation, both countries's censuses also included other information indicating social and economic status. In the United States census of 1880 information on literacy and unemployment was also collected. Again, the sample is broadly representative of American white men of this era, who had achieved nearly universal literacy. Unemployment was also low, with just 5.5% of this sample having experienced unemployment in the year preceding the census.

A major difference in the occupational structure of the two countries is the radically different proportion of the workforce, and of these representative samples, in farming (Table III). In the United States, 62% of the fathers were farmers in 1850, declining only to 47% among their sons in 1880. Britain's occupational structure was quite different, with the industrial revolution much further advanced. In Britain just 9.4% of fathers were farmers and 4.6% of their sons in 1881. In both countries this mirrored broader trends in the changing occupational structure. The proportion of farmers among American men was not below 10% until well into the 1920s, showing the dramatic differences in occupational structure between the two countries. Despite a large drop in the proportion of American men farming, nearly half the sons in 1880 were still farmers. Though not all were the sons of farmers, many were. Thus, in the United States a far greater proportion (42%) of sons had the exact same occupation as their father than in Britain (23%). Yet this highlights a limitation of the occupational information in the census. Although both countries supported a diversity of farming, the census recorded nearly all as "Farmer," omitting to record the crop or animal farmed.

TABLE III. Occupational Status of Sons and Fathers

Great Britain						
Variable	**Mean**	**StdDev**	**CV**	**Min**	**Max**	**N**
Exact occ as dad	0.23	0.42	1.8	0	1	3919
Same major group	0.4	0.49	1.2	0	1	3919
Father farmer, 1850/1	0.094	0.29	3.1	0	1	3919
Father, acres farmed	132	158	1.2	1	1141	325
Son farmer, 1880/1	0.046	0.21	4.5	0	1	3919
Son, acres farmed	184	209	1.1	1	1400	156
United States						
Can read and write	0.95	0.22	0.23	0	1	3715
Unemployed in 1879/80	0.055	0.23	4.1	0	1	3715
Sick on 1880 census day	0.014	0.12	8.4	0	1	3715
Exact occ as dad	0.42	0.49	1.2	0	1	3715
Same major group	0.5	0.5	1	0	1	3715
Father farmer, 1850/1	0.62	0.49	0.78	0	1	3715
Son farmer, 1880/1	0.47	0.5	1.1	0	1	3715

Particularly for farmers, sharing the exact same occupational description is likely to overstate the extent to which sons were actually doing the exact same work as their father. A broader measure of occupational inheritance between generations is the proportion of men who had an occupation in the same "major group" as their father. The HISCO occupational codes identify nine major groups of occupations: Professionals, Managers, Clerical workers, Sales workers, Service workers, Agricultural workers, Manufacturing workers, Transport workers, and Laborers. Sons who had jobs in the same major group as their father were likely to be doing something similar, either in terms of what they were producing, or the level of education and skill brought to the job. To make the concept more concrete, a father who was a carpenter and a son who was a painter would both be in the same major group. Both might have worked in the construction industry. Similarly, a father who was a lawyer and a son who was a doctor are both professionals, both occupations requiring a high level of education and thus similar in that respect.

Occupations provide a great deal of detail on what fathers and sons were doing, but this very detail can inhibit understanding of how father's occupations influenced son's occupations. In order to make sense of how closely a father's occupation influenced his son's occupation, we need to aggregate occupations into a smaller number of categories. To assess occupational mobility between generations we combine occupations into four broader groupings for analysis: (1) white collar workers: a broad group encompassing professionals, clerical workers, and sales people, (2) farmers (3) skilled workers or supervisory workers, such as foremen or overseers, and (4) unskilled workers, encompassing various industrial sectors from service work to farming to manufacturing. Our classification mirrors that in Ferrie and Long's recent analysis of social mobility in the same countries [63].

Our results are summarized in Table IV, describing occupational mobility from 1850/1 to 1880/1 in both countries. The layout of the panel for the two countries is identical. Occupations of the father are described in the columns, and of sons in each row. The classification of occupations is the same for fathers and sons. For each cell we list the number of sons of fathers in that occupational group who end up in a given occupational group. Percentages are calculated within columns for each country. For example, in Britain, 484 fathers had white collar occupations, and 274 of their sons (56.6%) also had a white collar occupation.

An assessment of occupational mobility requires us to measure how closely associated son's occupations were the occupation of their father's. In a symmetrical table the natural measure of association is a cross-product. However, as discussed earlier the occupational structure of the two countries differed significantly. We follow Long and Ferrie in calculating the Altham statistic for the tables of fathers' and sons' occupations [63, 64]. By multiplying one of the tables by a series of arbitrary constants the marginal frequencies are made identical, allowing us to compare only the degree to which the rows and columns are associated, i.e. the extent to which fathers occupations influence their son's occupations.

TABLE IV. INTERGENERATIONAL OCCUPATIONAL MOBILITY IN GREAT BRITAIN AND THE UNITED STATES, 1850/1-1880/1

	Father's occupations (1850/1)				
Great Britain	**White collar**	**Farmer**	**Semi/skilled**	**Unskilled**	**Total**
Son's occupation (1881)					
White collar	274	57	368	83	782
	56.61	15.24	18.1	8.07	19.95
Farmer	9	134	29	18	190
	1.86	35.83	1.43	1.75	4.85
Semi/skilled	158	109	1,438	472	2,177
	32.64	29.14	70.73	45.91	55.55
Un-skilled	43	74	198	455	770
	8.88	19.79	9.74	44.26	19.65
Total	484	374	2,033	1,028	3,919
	100	100	100	100	100
United States					
Son's occupation (1880)					
White collar	150	298	183	33	664
	48.86	12.78	23.4	11.22	17.87
Farmer	71	1,439	186	92	1,788
	23.13	61.71	23.79	31.29	48.13
Semi/skilled	66	358	323	90	837
	21.5	15.35	41.3	30.61	22.53
Un-skilled	20	237	90	79	426
	6.51	10.16	11.51	26.87	11.47
Total	307	2,332	782	294	3,715
	100	100	100	100	100

Note: Each cell reports frequency (e.g. 274) and column percent (e.g. 56.61)

Some aspects of the different occupational structure and transitions can be seen just from Table IV. In Great Britain, 44% of sons of unskilled workers remained in the same unskilled class, whereas in the United States just 27% of sons of the unskilled remained in that class. Thus, upward mobility for the sons of the lowest skilled was approximately half as likely again in the United States.

In both countries occupational inheritance was strong, with high percentages along many of the diagonals of the table. The exceptions to this are relatively low inheritance of farming in Britain, and the higher upward mobility of the unskilled in the United States. While occupational inheritance of farming occupations was high in the United States—61% of farmers' sons were farmers—more than 20% of the sons of other occupational classes also ended up in farming. The most similar aspect of the two countries occupational structure was entry into white collar work. In both countries occupational inheritance was relatively high, with around half of the sons of white collar workers being white collar workers themselves thirty years later. The proportion of sons of other occupational classes who ended up as white collar workers was relatively similar in both countries (compare the top row of each panel of Table IV).

TABLE V. ASSOCIATION BETWEEN FATHERS' AND SONS' OCCUPATION IN GREAT BRITAIN & THE UNITED STATES, 1850/1-1880/1

(1) Comparison	(2) M	(3) M'	(4) d(P,J)	(5) d(Q,J)	(6) d(P,Q)	(7) d^i(P,Q)
Ferrie/Long GB 1881 (P)	42.6	35.5	22.7 ***		13.2 ***	4.5
Ferrie/Long US 1880 (Q)	45.4	47.9		11.9 ***		
This paper GB 1881 (P)	41.2	33.8	25.2 ***		12.2 ***	2.5
This paper US 1880 (Q)	46.4	50.4		14.9 ***		

Note: *** indicates statistical significance at p=0.01.

Table V summarizes occupational mobility in Britain and the United States over a similar period of thirty year. We compare our results to Long and Ferrie, who created samples over a similar time period using alternative linkage methods. Column M reports the proportion of off-diagonal entries in each country, sons who ended up in a different occupational group than their father. Overall levels of mobility are similar, with the higher occupational inheritance of farming in the United States being balanced out by higher occupational inheritance in other categories in Britain.

However, the occupational structure differed in the two countries over time. Thus Column M' reports adjusted mobility statistics where the American marginal totals have been adjusted to match the British, and vice-versa. This comparison shows mobility in the United States to have been substantially greater than in Britain: son's occupations were not as tightly related to their father's occupations in the United States.

The underlying association between fathers' and sons' occupations is measured by the Altham statistic, which calculates the distance from independence of the occupational structure. In a simple 2 x 2 matrix the Altham statistic is the familiar cross-product ratio (ac/bd). If the rows and columns are independent, then the cross product ratio is 1. A matrix where all elements are ones satisfies these conditions, or indeed any matrix of constants. Matrices with more than 2 rows and columns have multiple cross product ratios, and the Altham statistic incorporates all the cross-product ratios into a single statistic.

$$d(\mathbf{P},\mathbf{Q}) = \left[\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{l=1}^{r}\sum_{m=1}^{s}\left|log\left(\frac{p_{ij}p_{lm}q_{im}q_{lj}}{p_{im}p_{lj}q_{ij}q_{lm}}\right)^2\right|\right]^{1/2} \quad (1)$$

The statistic has a chi-squared distribution, and the statistical significance of the metric can be calculated. The Altham statistics for Britain and the United States are presented in Columns 4 and 5 of Table V. In both countries the occupations of fathers and sons were strongly related, as the Altham statistic are significantly different from 0 in both cases. That is, comparing the frequencies for each country to a matrix of

identical constants (independent occupations) shows that both countries father-to-son occupational transitions differed significantly from the baseline of independence. However, the Altham statistics for Britain were 2/3 as large again as in the United States. Just as we can calculate the difference between each country's matrix and the null hypothesis of independence, we can also calculate the difference between the Altham statistics for each country, and whether it is statistically significant. This statistic is displayed in Column 6: d(**P,Q**), and we compare our results with Ferrie and Long's prior work on the same time period.

Ferrie and Long's matching method relied to a greater extent on exact similarity in the spelling of names, and a more rigid treatment of age discrepancies between censuses. Our linking methodology allows slightly greater tolerance for discrepancies in names and ages, particularly when there are no other potential matches that could be made. Ferrie and Long's method is slightly more likely to lead to false positive matches, and a higher degree of mobility. The differences in the Altham statistics between our results and theirs lie consistently in this direction (Columns 4 and 5). We find that both Great Britain and the United States were further from independence than Ferrie and Long do: in our results fathers' occupations exerted a slightly greater constraint on their sons' occupations than Ferrie and Long found. However, as can be observed the differences are relatively small, and do not attain statistical significance. Indeed, the difference that we find between Great Britain and the United States is very similar to what Ferrie and Long found (Column 6).

Finally, looking at the off-diagonal elements only (Column 7), we find only small differences in the overall degree of association between the countries. Thus, the differences in mobility between the two countries are mostly due to differences in occupational inheritance within the same occupational groups. In only one case (white collar to white collar) are the diagonal elements similar across the two countries, and the differences along the diagonals are fundamental to the differences between the two countries.

VIII. Agricultural Inheritance in Great Britain

Although relatively few men farmed in late nineteenth century Britain, compared to the United States, the transition of sons out of farming was socially significant. Many people in late nineteenth century British society were concerned about concentrated wealth holding, and the continuing control of farms by a landed elite. Data on overall patterns of land inheritance within British farming are scarce, yet the census returns contain information that allows much greater exploration of these questions than in the existing literature.

Instructions to British census enumerators asked them to record the acreage of farms, and the number of employees that a farmer had. Thus, farmers in the British census typically have occupational responses of the following form:

Farmer of x_1 acres employing y

Farmer of x_2 acres employing y_1 men and y_2 girls

The expressions are regular, with the number of acres almost always preceding the word acres, or a limited number of spelling variations. There is slightly less regularity of the expressions describing employees, but the number of variants of ways to describe employees is finite and straightforward to identify. Our linked sample is small, matching a 2% sample of the 1851 census with a 100% database of the 1881 census. Thus, we have 367 fathers who are farmers, and 182 sons. Complete databases of all British censuses from 1851-1911 will soon be available with occupational information transcribed, and it will be feasible to parse out information on acres farmed and employees on farms from occupational descriptions.

To do this, we first identify variants of the word "acres" that are found in the data, such as "ac", "acr", "acers", "acres)", "acs", "acre", "acrs", and "a". The program then reads each occupational description and extracts the word before acres to place in a new variable measuring acres farmed. We do this for both fathers' and sons' occupational descriptions. This new data allows us to examine how acres farmed by the father affected son's occupational chances using a simple probit model. For sons who remained in farming, we can compare the acres farmed between generations.

Fig. 3. Relationship of son staying in farming to fathers' acres farmed

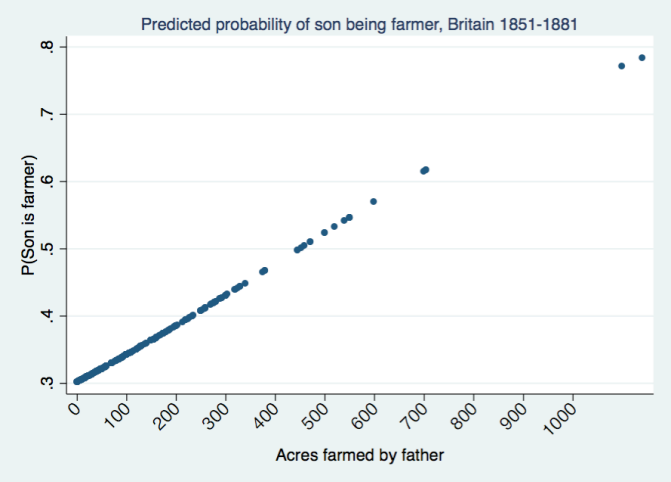
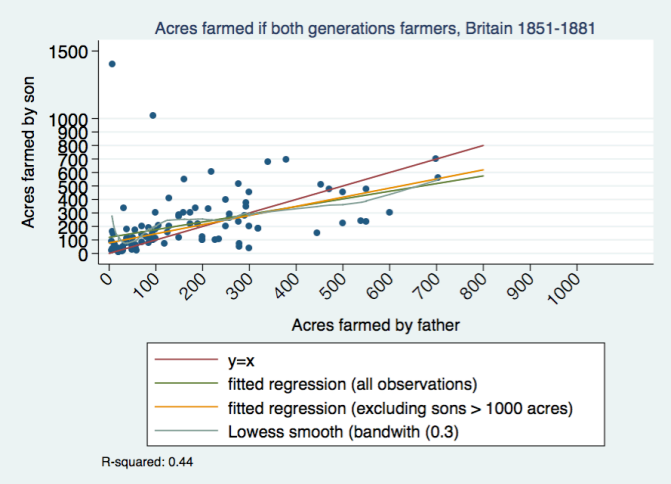


Fig. 4. Relationship of sons' and fathers' acres farmed



Several conclusions are apparent from this analysis, which we emphasize is more suggestive of the potential for application to pending complete-count databases of the British census, than a definitive analysis of agricultural inheritance in nineteenth century Britain. First, the chances that a son stayed in farming was strongly related to how many acres a father farmed. It was not until a father farmed more than 400 acres that his son had a greater than even chance of remaining in farming (Figure 3). Among sons who remained in farming, however, most ended up farming more acres than their father (Figure 4). This pattern is suggestive of selection into farming, where sons with the best chances of acquiring land stayed in the occupation; and also confirms that land ownership became more concentrated in late nineteenth century Britain. Finally, it is notable that the relationship between fathers' and sons' acres farmed is relatively consistent across the range of observations, with the fitted OLS regression line and a locally weighted regression line remaining close to each other over the range of the data.

IX. Conclusion

We began this paper by noting the fundamental historical question dating back to de Tocqueville, if not earlier, that motivates our research: was (is) America a more mobile society than Europe. Social mobility is an issue of special significance to the humanities, reflecting the extent to which societies organize themselves to allow either many or few of their citizens to exercise the full extent of their talents. We apply linking methods new to the historical literature. Historical linking has either been done by hand without specified rules, or by machine with exact matching or with rigid criteria for deviations from exact matching.

In this paper we use samples of the American and British 1850/1 censuses linked to complete databases of the 1880/1 censuses, but the methodology is scalable to forthcoming complete databases of these populations that will increase the number of potential and achieved matches significantly.

Despite the differences in our linking methodology to the research of Ferrie and Long [63] we find relatively small differences in the substantive conclusion that late nineteenth century America was a significantly more mobile society than Britain at the same time. That this finding is robust to alternative methods of constructing linked census samples only strengthens the conclusion about social differences across the Atlantic.

The meaning of those differences in mobility is complicated by large differences in the economic structure of the two countries. American mobility comes largely, but not entirely, from the escape valve of farming on the frontier. The more urban and industrial British economy offered a greater diversity of occupational opportunities. The specific responses to the census enumeration highlight these differences with a greater range of job descriptions in Britain. On the other hand, recent evidence suggests that per-capita income in the United States exceeded Britain's throughout the era.

Our research also highlights the importance of large samples for investigating questions of social mobility, and indeed other historical questions. While we summarize the overall differences between occupational mobility in Britain and the United States in a single statistic, the statistic can be decomposed into a smaller number of component statistics that show more precisely where the two countries diverged. In the late nineteenth century, those differences lay largely in greater American persistence in farming across generations, and a significantly greater chance for sons of unskilled men to end up in farming, white collar work or skilled occupations. Moreover, in the United States sons of farmers who left farming were much more likely to avoid ending up in unskilled work than their peers in Britain. Taken together, these results suggest that young men in the late nineteenth century United States had significantly better life chances than their British peers. Were these differences the result of institutions—such as government and educational opportunities—or environments—with more abundant land in the United States? The next phase of our research will incorporate Canadian data for the same time period, and for all three countries for a subsequent generation (1880/1 – 1910/1) to address these questions.

References

[1] A. de Tocqueville, *Democracy in America*. Paris, 1831.

[2] A. de Tocqueville, *Regards sur le Bas-Canada*. Montréal: Typo, 2003 (1836).

[3] F. J. Turner, "The Significance of the Frontier in American History," *Proceedings of the State Historical Society of Wisconsin,* vol. 41, pp. 79-112, 1893.

[4] R. Archer, *Why is there no labor party in the United States?* Princeton: Princeton University Press, 2007.

[5] J. Heffer and J. Rovet, Eds., *Why Is There No Socialism in the United States?* Paris, 1988, p.^pp. Pages.

[6] A. Bosch, "Why Is There No Labor Party In The United States? A Comparative New World Case Study: Australia And The U.S., 1783-1914," *Radical History Review,* vol. 67, pp. 35-78, 1997.

[7] W. Sombart, *Why Is There No Socialism in America?* New York, 1976 (1906).

[8] S. M. Lipset, *Continental divide : the values and institutions of the United States and Canada*. New York: Routledge, 1990.

[9] S. M. Lipset and G. W. Marks, *It didn't happen here : why socialism failed in the United States*, 1st ed. New York: W.W. Norton & Co., 2000.

[10] R. Archer, "Labour Politics in the New World: Werner Sombart and the United States," *Journal of Industrial Relations,* vol. 49, pp. 459-482, 2007.

[11] L. Cox, "Review Essay: Revisiting the Labour Question in the United States," *Thesis Eleven,* vol. 100, pp. 168-178, 2010.

[12] H. D. Forbes, "Hartz-Horowitz at Twenty: Nationalism, Toryism and Socialism in Canada and the United States," *Canadian Journal of Political Science,* vol. 20, pp. 287-315, 1987.

[13] S. M. Lipset, *Agrarian socialism*. Berkeley,: University of California Press, 1950.

[14] S. Blumin, "Mobility and Change in Ante-Bellum Philadelphia," in *Nineteenth-Century Cities*, S. Thernstrom and R. Sennett, Eds., ed New Haven: Yale University Press, 1969, pp. 165-208.

[15] P. R. Knights, *The plain people of Boston, 1830-1860: A study in city growth*. New York: Oxford University Press, 1971.

[16] J. Modell, "The Peopling of a Working-Class Ward: Reading, Pennsylvania, 1850," *Journal of Social History,* vol. 5, pp. 71-95, 1971.

[17] S. Thernstrom, *The other Bostonians; poverty and progress in the American metropolis, 1880-1970*. Cambridge: Harvard University Press, 1973.

[18] H. M. Gitelman, *Workingmen of Waltham: Mobility in American urban industrial development, 1850-1890*. Baltimore: Johns Hopkins University Press, 1974.

[19] M. B. Katz, *The people of Hamilton, Canada West: Family and class in a mid-nineteenth-century city*. Cambridge: Harvard University Press, 1975.

[20] D. Gagan, "Geographical and social mobility in nineteenth-century Ontario: a microstudy*," *Canadian Review of Sociology/Revue canadienne de sociologie,* vol. 13, pp. 152-164, 1976.

[21] M. B. Katz, M. J. Doucet, and M. J. Stern, "Migration and the Social Order in Erie County, New York: 1855," *The Journal of Interdisciplinary History,* vol. 8, pp. 669-701, 1978.

[22] L. Glasco, "Migration and Adjustment in the Nineteenth-Century City: Occupation, Property, and Household Structure of Native-Born Whites, Buffalo, New York, 1855," in *Family and Population in Nineteenth Century America*, T. Hareven and M. A. Vinovkis, Eds., ed Princeton: Princeton University Press 1978, pp. 154-178.

[23] C. Griffen and S. Griffen, *Natives and newcomers : the ordering of opportunity in mid-nineteenth-century Poughkeepsie*. Cambridge: Harvard University Press, 1978.

[24] G. Darroch, "Migrants in the Nineteenth Century: Fugitives or Families in Motion?," *Journal of Family History,* vol. 6, pp. 257-277, 1981.

[25] D. Gagan, *Hopeful Travelers: Families, Land and Social Change in Mid-Victorian Peel County, Canada West*. Toronto: University of Toronto Press, 1981.

[26] T. Dublin, "Rural-Urban Migrants in Industrial New England: The Case of Lynn, Massachusetts, in the Mid-Nineteenth Century," *The Journal of American History,* vol. 73, pp. 623-644, 1986.

[27] J. Long and J. Ferrie, "The path to convergence: intergenerational occupational mobility in Britain and the US in three eras," *Economic Journal,* vol. 117, pp. C61-C71, 2007.

[28] S. Ruggles, E. Roberts, S. Sarkar, and M. Sobek, "The North Atlantic Population Project: Progress and Prospects " *Historical Methods,* vol. 44, pp. 1-6, 2011.

[29] Minnesota Population Center, *North Atlantic Population Project: Complete Count Microdata. Version 2.0*. [machine readable database]. Minneapolis, MN: Minnesota Population Center [distributor], 2008.

[30] E. Roberts, S. Ruggles, Lisa Y. Dillon, Ó. Garðarsdóttir, J. Oldervoll, G. Thorvaldsen*, et al.*, "The North Atlantic Population Project: An Overview," *Historical Methods,* vol. 36, pp. 80-88, 2003.

[31] L. Gu, R. Baxter, D. Vickers, and C. Rainsford, "Record linkage: Current practice and future directions," Canberra2003.

[32] I. P. Fellegi and A. B. Sunter, "A Theory for Record Linkage," *Journal of the American Statistical Association,* vol. 64, pp. 1183-1210, 1969.

[33] M. A. Jaro, "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida," *Journal of the American Statistical Association,* vol. 84, pp. 414-420, 1989.

[34] T. R. Belin and D. B. Rubin, "A Method for Calibrating False-Match Rates in Record Linkage," *Journal of the American Statistical Association,* vol. 90, pp. 694-707, 1995.

[35] M. D. Larsen and D. B. Rubin, "Iterative Automated Record Linkage Using Mixture Models," *Journal of the American Statistical Association,* vol. 96, pp. 32-41, 2001.

[36] W. E. Winkler, "Improved Decision Rules in the Fellegi-Sunter Model of Record Linkage," presented at the American Statistical Association Proceedings of the Section of Survey Research Methods, 1993.

[37] P. Christen, "Automatic Record Linkage Using Seeded Nearest Neighbour And Support Vector Machine Classification," presented at the Proceedings of the 14th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining, Las Vegas (NV), 2008.

[38] M. J. Wisselgren, S. Edvinsson, M. Berggren, and M. Larsson, "Testing Methods of Record Linkage on Swedish Censuses," *Historical Methods: A Journal of Quantitative and Interdisciplinary History,* vol. 47, pp. 138-151, 2014.

[39] E. H. Porter and W. E. Winkler, "Approximate String Comparison and its Effect on an Advanced Record Linkage System," U.S. Bureau of the Census, Washington D.C.1997.

[40] W. E. Winkler, "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage," presented at the American Statistical Association Proceedings of the Section of Survey Research Methods, 1990.

[41] R. Vick and L. Huynh, "The Effects of Standardizing Names for Record Linkage: Evidence from the United States and Norway," *Historical Methods: A Journal of Quantitative and Interdisciplinary History,* vol. 44, pp. 15 - 24, 2011.

[42] P. Christen, T. Churches, and J. X. Zhu, "Probabilistic Name and Address Cleaning and Standardisation," presented at the Proceedings of the Australasian Data Mining Workshop, Canberra, 2002.

[43] J. I. Maletic and A. Marcus, "Data cleansing: Beyond integrity analysis," presented at the Proceedings of the Conference on Information Quality, Boston, 2000.

[44] L. Nygaard, "Name Standardization in Record Linkage: an Improved Algorithmic Strategy," *History and Computing,* vol. 4, pp. 63-74., 1992.

[45] L. Philips, "The Double-Metaphone Search Algorithm," *C/C++ User's Journal,* vol. 18, 2000.

[46] A. J. Lait and B. Randell, "An Assessment of Name Matching Algorithms," Department of Computing Science, University of Newcastle upon Tyne, Newcastle1993.

[47] P. Christen, "Development and user experiences of an open source data cleaning, deduplication and record linkage system," presented at the SIGKDD, 2009.

[48] P. Christen, *Data matching*. Berlin: Springer, 2012.

[49] Z. Fu, P. Christen, and J. Zhou, "A Graph Matching Method for Historical Census Household Linkage," in *Advances in Knowledge Discovery and Data Mining*, ed: Springer, 2014, pp. 485-496.

[50] Z. Fu, M. Boot, P. Christen, and J. Zhou, "Automatic Record Linkage of Individuals and Households in Historical Census Data," *International Journal of Humanities and Arts Computing,* 2014.

[51] L. Antonie, P. Baskerville, K. Inwood, and A. Ross, "Creating Longitudinal Data from Canadian Historical Censuses," Department of Economics, University of Guelph, Guelph2011.

[52] L. Antonie, P. Baskerville, K. Inwood, and A. Ross, "An Automated Record Linkage System – Linking 1871 Canadian census to 1881 Canadian Census," Department of Economics, University of Guelph, Guelph2010.

[53] L. Antonie, K. Inwood, D. J. Lizotte, and J. A. Ross, "Tracking people over time in 19th century Canada for longitudinal analysis," *Machine Learning,* vol. 95, pp. 129-146, 2014.

[54] S. Abe, *Support vector machines for pattern classification*, 1st ed. New York: Springer, 2010.

[55] N. Cristianini and J. Shawe-Taylor, *An introduction to support vector machines and other kernel-based learning methods*. Cambridge: Cambridge University Press, 2000.

[56] V. N. Vapnik, *Statistical learning theory*. New York: Wiley, 1998.

[57] C.-C. Chang and C.-J. Lin, "LIBSVM: A Library for Support Vector Machines," Department of Computer Science, National Taiwan University, Taipei2007.

[58] M. H. D. van Leeuwen, I. Maas, and A. Miles, "Creating a Historical International Standard Classification of Occupations," *Historical Methods,* vol. 37, pp. 186-197, 2004.

[59] M. H. D. van Leeuwen, I. Maas, and A. Miles, *Historical International Standard Classification of Occupations*. Leuven: Leuven University Press, 2002.

[60] S. Edvinsson and J. Karlsson, "Recoding occupations in the Demographic Data Base into HISCO," HISMA Berlin1998.

[61] E. Roberts, M. Woollard, C. Ronnander, L. Y. Dillon, and G. Thorvaldsen, "Occupational Classification in the the North Atlantic Population Project," *Historical Methods,* vol. 36, pp. 89-96, 2003.

[62] M. Sobek, "Work, Status, and Income: Men in the American Occupational Structure since the Late Nineteenth Century," *Social Science History,* vol. 20, pp. 169-207, 1996.

[63] J. Long and J. Ferrie, "Intergenerational Occupational Mobility in Britain and the U.S. Since 1850," *American Economic Review,* vol. 103, pp. 1109-1137, 2013.

[64] P. M. E. Altham and J. E. Ferrie, "Comparing Contingency Tables," *Historical Methods,* vol. 40, pp. 3-16, 2007.