A first look at longitudinal data from the Canadian censuses of 1871 and 1881

Luiza Antonie, University of Guelph
Kris Inwood, University of Guelph
J. Andrew Ross, University of Guelph

This paper reports, assesses and illustrates the use of a recent machine learning strategy to link Canadian 1871 and 1881 census records. The method generates life-course information for large numbers of individuals with a low (3%) false positive error rate. The linked data are broadly representative of the population with some under-representation of particular groups. The new data reveal considerably more movement in and out of occupational groups than is suggested by aggregate tabulations. The beginnings of diversification away from agriculture was led by younger workers.

A first look at longitudinal data from the Canadian censuses of 1871 and 1881[1]

Introduction and overview of the record-linking system

The considerable importance of the North American census for historical research derives from its rich systematic detail and a paucity of alternate sources describing the population.[2] In the absence of other sources Canadian and American scholars turn to the census for population profiles and to track individuals through time. 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; Long and Ferrie 2007) has been followed by machine learning approaches (Ruggles 2006; Christen 2008; Goeken et al 2011; Fu et al 2014). The new methods are capable of generating in a near-automatic way large representative samples of longitudinal and even multi-generational data.

In this paper we consider a recent application of support vector machine (SVM) classification to historical Canadian historical census records (Antonie et al. 2014). The linked data are reasonably representative of the population although admittedly some groups are harder to link. In spite of these and other idiosyncrasies the longitudinal data provide a fresh window into occupational change during the 1870s, a decade which otherwise is difficult to interpret.

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 and 4.3 million records of the 1881 census. We construct records that follow individuals over time by comparing every 1871 record with each 1881 record, and then classifying each comparison as a match or non-match. If a particular pair of records (one from 1871 and one from 1881) points 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 ignore information about occupation, location and household composition o avoid any bias to people who persist in the same area, in the same job or in the same family. There are 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. We classify support vector machine (SVM) methodology that delivers superior results for this kind of data (Christen 2008; Richards 2013).[3] The classification software 'learns' from matches already confirmed 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.

[1] The research reported in this paper would not have been possible without grants from the Canadian Foundation

[2] Nowhere in North America was there an established church with a commitment to public vital registration. Only in Utah and Quebec was there a dominant church whose records might serve that purpose, and over time even their records became less comprehensive. Consistent birth, marriage and death registration emerged in most states and provinces during the early decades of the twentieth century.

[3] More generally on SVM and classification methods see Vapnik (1995).

We have four sets of true links: 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 ten years old in 1871.[4] The chief criterion for a match is census co-residence of other family members; additional information where available (eg church records in Toronto and Quebec City) is used. We confirm 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 there is no other likely match in the 1881 Canadian census or the 1880 U.S. census.[5]

The 'true links' 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 who tend to cohabit with the same family members over a decade. Single people and those who became single over the decade (for example children leaving home) are under-represented. Fortunately, even if the true links are not demographically representative, they still reflect the imprecision of 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. Calculating similarities between millions of 1871 records and millions of 1881 records would require almost one year of continuous operation by a single processor.[6] Even running the system in parallel, 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.[7] We block by birthplace, marital status (allowing for obvious changes), first letter of surname and first name grouping that recognize 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). We identify similarity between birth years with a log-linear decay function. A description of the features used for linking is given in Appendix A; Antonie et al (2014) present the system formally.

[4] The proprietors were linked in preparation for Inwood and Reid (2001). The Logan records were linked in preparation for Baskerville (2015). Andrew Hinson generated the St. James links or 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.

[5] We check the United States census as well, because in this period many Canadians migrated to the United States.

[6] 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 an additional task.

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

Of course, many 1871 records cannot be matched because the individual died before 1881, left the country, or reported information differently in the two years. Nevertheless, the most common reason for failing to identify a match is *not* an inability to find someone with the same characteristics ten 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 B.

The severity of the problem of multiples is clear from the distribution of outcomes for 1871 records, as reported in Table 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.[8] 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, 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.

Table 1 about here

The level and sources of error among the 1871-1881 linked records

The first thing to ask 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%.[9] 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 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 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

[8] 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 1 reports the outcome of records submitted to the classification system.

[9] 3% is the false positive rate on Ontario true links using a five-fold cross validation method.

enumerators.[10] Thus the maximum possible unique link rate that we might hope to achieve is 40-50%.[11]

The reason we achieve 21% rather than the maximum possible 40-50% is related to the reasons why there are any mistakes at all. Every time we do not 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 (ie 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. Imprecision further aggravates the problem of multiples links since it forces a broadening of the tolerance for declaring a match.[12] 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 (ie 3% of the 21%) reflects a successful balance of tolerances for linking characteristics.

Table 2 about here

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 population. From the outset, we can anticipate reasons why linked records may be 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, of course, they correlate with other biases of greater analytical import.

[10] Under-enumeration in the 19th century U.S. censuses is estimated to be about 5% (Hacker 2013).

[11] Even 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).

[12] For example, we might choose to 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.

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 (www.census1871/ca). One effect is immediately apparent in Table 2: 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. The record linking process is most 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.

Interestingly we are no more likely to link the native-born than immigrants (Table 3).[13] The same is largely true for individual countries of birth. Admittedly those born in England are over-represented in the linked sample. 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 3 indicates a distribution of ethnicities roughly matching that of the population, with two important exceptions: fewer French-origin people are linked while people of English-origin are linked more successfully. The under-representation of people who report a French origin is notable.

Table 3 about here
Table 4 about here

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.[14] The hazard or odds ratios reported in Table 4 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 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 confirm that men and married people are more likely to be linked after controlling for multiple influences. These patterns are consistent with our expectation that 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

[13] 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.

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

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 vs singles and men vs women allows more precise estimation of the age, ethnicity and literacy effects (columns 2-5 in Table 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).[15] 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.[16] We can estimate odds ratios for these effects separately (Tables 5 and 6). 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 under-linked (Table 5) but they (especially single women) suffer a great deal from the problem of multiples (Table 6). 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 7 reports the odds of finding at least one link. The pattern of odds ratios is very close to that of the general population (Table 5) with one exception. The impact of illiteracy on the odds ratio disappears for married men and becomes slightly stronger for married women. 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 under-linking 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, although one explanation is departure from the population (emigration of the French and death of older people). 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 5). Indeed, the problem is that we find too many good matches and in consequence cannot discriminate amongst them (Table 6). Any strategy for disambiguation of multiples might be especially helpful for the young adults.

[15] 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.

[16] 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.

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. 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.[17]

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 5).

Table 5 about here
Table 6 about here
Table 7 about here

Error rates among movers versus stayers

Linked or longitudinal census data are often used to describe and analyze mobility – both social and geographical. Here we consider error rates among those who change location by checking if the individual links are ‘credible’ (as indicated by the continued co-residence of family members). This differs from the earlier evaluation insofar as we do not begin with secure knowledge acquired independently of the linkage system. Checking the credibility of links 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 checking process is more restricted and much less costly. We ask if the 1881 match recommended by the system has co-resident family members

[17] 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.

who resemble those of the 1871 record using structured criteria (see Appendix C). There might be a number of 1881 records with similar co-residents, but these are not checked. Rather, we assess the 'credibility' of the one record selected by our linkage system.

The checking process is imprecise to the extent that we ignore other possible matches that, if examined, might reduce confidence in our results. There are other limitations. For example, we use the co-residence in 1881 of people who would not be expected to be absent (given what we know from the 1871 family) as evidence undermining credibility.[18] 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 co-resident 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 idiosyncratic the checking process provides an economical but plausible check. 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 Eight. 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 Nine.[19] 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.[20] The difference is dramatic, and invites explanation.

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.[21] 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 co-resident 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

[18] Category D in Appendix C.

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

[20] 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.

[21] 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).

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'.[22]

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 inter-provincial 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 10 illustrates that a 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%.[23] In practice, of course, the severity of this complication depends a great deal on particular circumstances, as illustrated in Table 10.

Table 8 about here
Table 9 about here
Table 10 about here

Reflections on the quality of linked data

The nature of the source and underlying population does not allow us to link every record. Nevertheless, 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. The links slightly over-represent immigrants born in the British Isles but replicate reasonably well the proportions of the population born in Canada versus immigrants and in one province versus another. People who were unable to read are 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

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

[23] 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.

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. Quebec intellectuals headed the Census Bureau in 1871, and most enumerators in French-speaking areas were themselves francophone. A more likely influence is that the relatively small pool of French names makes it harder to isolate a unique match or that emigration during the 1870s was greater from heavily Francophone districts. The lower level of literacy in Quebec might have reduced the reporting precision, although 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.[24] Analysis of the odds of linking shows that the under-representation of French-origin population is pronounced only for married people (and is especially large for married women).

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 is low, 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.

Our assessment of the linked records identifies specific limitations notwithstanding their excellent quality overall. The ultimate test of the value of the linked records, however, is whether or not they tell us something interesting about behaviour in the past. For the remainder of the paper we use the linked records to examine occupational change during the 1870s.

The 1870s: A paradoxical decade

The third quarter of the nineteenth century marks a significant turning point in the evolution of the north Atlantic societies. Technological advances, new management structures and organizational change fuelled a second phase of industrialization that slowly but surely

[24] 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.

transformed work and everyday life in the United States, Britain and increasingly in the rest of Europe (Allen 2009). Within Britain the working classes finally began to share in some of the gains of the industrial revolution. Trade expanded enormously as long-distance transportation and communication costs fell (Harley 1988; Williamson 1999). Governments responded with industrial policies designed to take advantage of new opportunities and mitigate the more disruptive consequences of market realignment (Inwood and Keay 2013; O'Rourke and Williamson 1999). A steady growth of incomes contributed to fertility decline, greater longevity and, increasingly, rural-urban and long-distance migration. Within the United States settlement spread across the continent with lightning speed following the abolition of slavery and cessation of the Civil War.

Canada was changing as well. Incomes rose, and the population expanded westward and northward. And yet, the aggregate evidence of economic and demographic transformation is muted. National income estimates suggest an image of 'balanced' rather than revolutionary growth, with modest structural change (Green and Urquhart 1987; Urquhart 1993). Agriculture's contribution to GDP fell only slightly (35.2% to 34.6%) and the manufacturing share remained steady at 21% during the 1870s (Urquhart 1993: table 1.1). Aggregate data reveal little evidence of change in the labour force. Nearly half (47.5%) of those reporting an occupation in 1871 were in the agricultural class; one-fifth (21.1%) were in the industrial class. Ten years later the picture is remarkably similar: 47.7% of the reported occupations were agricultural and 20.7% were in industry (Table 11).[25]

Table 11 about here

This picture of aggregate stability is at odds with the evidence of numerous community and institutional case studies, the severity of the business cycle and broader transformation of the north Atlantic economy. At a local, or disaggregate, level scholars see evidence of considerable change and transformation but at the aggregate level we see continuity and very limited change. Our goal in this paper is to reconcile the conflicting evidence of continuity and change in the labour market during the 1870s. By following a large number of individuals from census to census we can identify slow-moving but powerful processes of change consistent with the evidence of structural change and industrialization.

The 1870s was the first full decade for the new Canadian Confederation. Over the decade Canada's population grew 17%, spilled into three new provinces and a territory and reached 4.3 million in 1881. Fertility and mortality declined, emigration increased and family reproduction strategies adapted to changing circumstance (Darroch and Ornstein 1984; Dillon 2008; McInnis 2000; Olson and Thornton 2011). New national policies responded to powerful economic cycles of boom, bust and recovery (Chambers 1964; Forster 1986). Farmers successfully re-oriented their production as western farm products invaded eastern markets (Harley 1978; Marr 1981). Entirely new manufacturing industries appeared, and output and productivity increased within established industries in a process of successful industrialization (Forster and Inwood 2003; Inwood and Keay 2012).

[25] Change was limited even within individual provinces, with the exception of New Brunswick which lost much of its manufacturing sector in the devastating Saint John fire of 1871.

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 (Gaffield 1987; Gagan 1982; Widdis 1989). Contemporary observers worried that farming and the mechanic arts were declining in opportunity and status, an anxiety enhanced by the vicissitudes of the business cycle, a tendency for young people to leave their farm families and, of course, change within individual occupations (Prentice 1977: 95). In contrast, Drummond argues that farmers adapted well to new pressures by moving into mixed farming, and that this freed up labour for other sectors such as manufacturing, which in Ontario grew by one-third in the 1870s (Drummond 1987: 29-30). Crowley (1995: 44) also questions the severity of the rural crisis and explains in Ontario “the movement into farming remained the predominant agricultural occupational trend… during the third quarter of the nineteenth century.”

One way to reconcile divergent impressions is to imagine that recent immigrants took up occupations in a way that exactly offset any transformation among the native-born. In fact, there was relatively little immigration to Canada during the 1870s and net emigration was still at modest levels (Emory, Inwood and Thille 2007; McInnis 2000). Alternately, a synthetic cohort strategy might expose if individual subgroups within the population changed in ways that happened to offset each other (eg Inwood, MacKinnon and Minns 2014). And yet, knowing the characteristics of a group and knowing the characteristics of individuals within a group are two different things. The latter may have changed in ways that remain invisible to the extent that they offset each other. It is preferable to follow exactly the same people in both years. This is now possible using the linked census records reported above. Here we restrict our attention to linked records found in the 5% sample of the 1871 census, since only these records have full information on occupations.

Table 12 about here

We summarize in Table 12 the linked microdata with occupations in both years. The new data are not strictly comparable to published census tabulations because we examine only a subset of the 1871 and 1881 populations and the three broad classes are defined differently.[26] Nevertheless, both sources describe the same broad pattern of *no change* in the broad distribution of occupations. In this respect the microdata replicate the feature of published tables that we wish to investigate: the impression that the Canadian labour market did not change during the 1870s.

The patterns of occupational change

Not everyone reported an occupation in both years (Table 13). The youngest males did not have an occupation in 1871 although many moved into an occupation during the following decade. By 1881 some older men ceased to work. In this paper we examine only those who reported an occupation in both years. We are interested in occupational changes and not the

[26] We examine only those people who reported an occupation in both years while the Census Bureau tallied occupations without regard for continuity. We aggregate manuscript descriptions to occupational groups and the three broad classes follows a well-defined set of rules within the framework of the NAPP-HISCO coding scheme (Roberts et al 2003). Unfortunately, procedures used by the Census Bureau 140 years ago are unknown.

choice of occupation for those entering the labour market for the first time or the age at which people retire in different occupations. Consequently, we lose a few of the records for linked males and a large majority of the linked records for females.

Occupational reporting by women reflects both the extent to which women worked outside the home and a well-known gender bias in occupational reporting (Abel and Folbre 1990; Carter and Sutch 1996; Inwood and Reid 2001). Many women, even if they worked outside the home, did not report an occupation. The bias was most extreme for married women. Those women who did report an occupation in both years, and are linked from 1871 to 1881, are summarized in Table 13. We organize the occupations into six broad categories: farming, manufacturing (all industrial occupations), commerce (merchants, retail and wholesale), labour and construction (labour and trades), and other services (professionals and miscellaneous).[27]

Table 13 about here
Table 14 about here
Table 15 about here

Most women reporting an occupation identified themselves in a service job (eg servants, members of religious orders, and teachers), and they remained in the service sector ten years later. For example, in 1871 Margaret Boutilier was working as a servant in the household of William Gammell, a retired merchant in Sydney Mines, Nova Scotia. Ten years later she had moved out to the local farm of Peter Jackson in Ball's Creek, again being employed as a servant. A relatively smaller number of women clustered in manufacturing, primarily dressmakers and seamstresses, but also food and other products.[28] One married woman, Catherine Brennan of Tyendinaga Township, in Hastings Country, Ontario, was listed as a cheesemaker in 1871, and also in 1881, by which time she was widowed and supporting her seven children. However, she was an exception insofar as the information in Table 14 mainly describes the subset of women who did *not* marry; most women who married ceased to report an occupation and therefore drop out of our analysis.

A much higher proportion of men reported occupations.[29] In Table 15 we report men by province of residence in 1871. The provincial comparison points to distinctive and offsetting regional patterns of change. Ontario saw a net movement of men out of farming and into commerce. Quebec men also moved into commerce and, importantly, into manufacturing. New Brunswick and Nova Scotia, in contrast, saw a sizeable net movement *into* farming. Divergent regional trends in occupations are consistent with the long-run pattern of structural change in which the Atlantic region apart from its small coal belt conspicuously failed to industrialize (Acheson 1972; Inwood 1991).

We organize the same data in Table 16 to highlight differences by birth cohort rather than province. We distinguish younger from older men using their age in 1871; the two cohorts were

[27] See Roberts et al., "Occupational Classification," for the classification scheme.
[28] Dunae (2009) finds that in some cases "dressmaker" was a coded way of identifying sex workers. Unfortunately, we cannot distinguish such cases. McDevitt, Irwin, and Inwood (2009) analyze gender differences in productivity and pay for Canadian clothing makers.
[29] Among men aged 15 to 55, 87% reported an occupation in 1871 and 85% in 1881. Enumerators had the option to record one or more occupations, but rarely did so. We use the first-mentioned occupation.

born 1816-1845 and 1846-1855. A desire to maintain a sufficient number of observations in cells throughout the table recommends that we drop the provincial distinctions. At the level of the entire country the younger group, adolescents and young adults, had a net movement into commerce and manufacturing and out of farming and other sectors (comparing the shaded bands in the table). The reverse is true for middle-aged and older men: there was a net movement into farming. There was *no* net movement of older men into commerce or manufacturing.

Admittedly, the interior diagonal of the table for each age group indicates substantial occupational persistence. Most men, even the younger ones, stayed in their occupational classes. Labour and construction occupations showed a strong preference for turning to farming, a reflection of the large component of farm labourers in this group. Farming itself shows the highest degree of persistence, although age was a significant factor: 86% of farmers aged 26–55 who farmed in 1871 continued to do so in 1881, compared to 74% of those aged 15–25. Figure 6.1, which includes both age groups, shows the relative size of the transitions in proportion, as well as the continued dominance of farming.

Table 16 about here
Figure 1 about here

The younger men switched out of farming at double the rate of the older generation: either into manufacturing (7% of young men compared to 3% of older men), commerce (3% compared to 2%), or labour/construction (12% to 6%). Among the labour/construction occupations, which often served as transitional jobs for younger men, we see that the younger cohort was more likely than the older to enter manufacturing (15% to 9%). Nonetheless, persistence of almost three-quarters of young men in farming attests to its pull, as does the significant numbers of young men entered farming from other sectors over the decade. The most important feature of Table 16 is the shift towards commerce and industrial work and away from agricultural work for younger men.

In Ontario, at least, it has been argued that farming was largely the preserve of the Canadian-born (Drummond 1987: 30). Our data confirm this pattern for the country as a whole (Table 17). In both cohorts more than half the Canadian-born are farmers; the foreign-born were more likely to be involved in industrial or labour/construction pursuits. We do see an uptick in the number of younger and older foreign-born men entering the agricultural sector across the decade, suggesting farming still had an attraction for foreigners.

Table 17 about here
Table 18 about here
Table 19 about here

Information on ethnic origin sharpens the picture. In Table 18 we show the proportions by ethnicity in each occupational sector in 1881 and the change from 1871 (in italics). While there is stability across the sectors, there is also some intriguing variation. Fewer than half of the younger men in each group were likely to choose farming, except for those of European heritage (mostly German descendants, many of them born in Canada), who showed a marked preference for farming in both cohorts. Over the decade, all groups of young men except those of Irish ethnicity shifted out of farming, labour/construction and other services, and moved increasingly to commercial and industrial pursuits. Those of Scottish origin led the way in these trends (especially the shift to industry).

There was also a marked generational divergence within the French and Irish origin groups. In 1881, 9% of the older French cohort but 15% of the younger men were involved in manufacturing. The gap was similar for Irish-descendants: 10% and 16%. The younger French also had the highest proportion of commercial workers. These transitions highlight the role of young French-origin and Irish-origin men in powering industrial growth, both in Quebec and in Ontario (Charpentier 1990; Heron 1995).

The existing literature on the Irish and their descendants in Canada points to the importance of church or denomination; Darroch and Ornstein, and Akenson report that Roman Catholics of Irish descent were significantly less likely to be farmers than the Protestants (Darroch and Ornstein 1980, 1984; Akenson 1988: 94-95). We replicate this pattern in Table 19. Irish-origin Protestants were more likely to farm, although older Roman Catholics led the movement of older men of Irish descent into farming over the decade. Protestants were more active in industry at a younger age (though losing some ground over the decade), and yet the reverse was true in the older cohort, where Roman Catholics had a greater proportion in industry. Our longitudinal profile is broadly consistent with Akenson's suggestion that denominational differences among the Irish and their descendants arise in part from the timing of immigration; the Irish Protestant advantage of early arrival in Canada diminished as Roman Catholics were catching up during the 1870s.

Conclusion: Change amid continuity

The linked census data allow us to follow a large number of Canadians from 1871 to 1881 and to generate a number of insights that would not otherwise be available. The most important is that the impression of a stasis in occupational choice and patterns of work, suggested by Table 11 and Table 12, is misleading. Patterns of work did change, and the changes are broadly consistent with previous literatures. To a very large extent young people were the agents of change. Adolescents and young adults were more likely to change occupation than older workers; the younger men increasingly entered manufacturing, commerce and services.

And yet there was considerable continuity insofar as a significant proportion of people moved into farming during the 1870s, and relatively few left. Farming remained the preferred alternative 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.

Our findings are consistent with the experience of an earlier decade, the1860s, as interpreted by Darroch and Ornstein (1984) and Darroch (2015). There were no radical breaks in the relative importance of agricultural employment during the 1860s or in the 1870s. Agriculture remained attractive, but our linked individuals moved in and out of the sector more frequently in the 1870s. Darroch and Ornstein found ~90% of farmers remained in the occupation from 1861 to 1871. In contrast, we find that only 86% of the older farmers and a still smaller proportion

(74%) of the young farmers remained in the sector.[30] Significant numbers of young workers entered industry and commerce, which together expanded during the 1870s from 16% to 22% of all young workers (Table 15).

Differences by province and ethnic origin are visible. There was a net movement into farming by workers in Nova Scotia and New Brunswick, and a net movement out of farming by workers in Ontario (Table 14). Men of central European origin (largely German ethnicity) continued to cluster in farming, while the Irish and their descendants (especially Roman Catholic Irish) stand out for a net movement into farming. Those of Scottish origin and to a lesser extent the English shifted into manufacturing and commerce (Table 18 and Table 19). Immigrants in general were more likely to enter farming (Table 17).

An identification of broad patterns does not deny the diversity of individual trajectories, as revealed in a separate examination of a small set of men who were eighteen years old in 1871. Some of them followed the pattern of Daniel Lank of Colchester County, Nova Scotia, who was a labourer in 1871 but ten years later was farming in his own right. Samuel Hiltz, a young farmer of German origin from Lunenburg, was able to stay on the family farm a decade later, but many others moved from farming to other occupational classes, particularly manufacturing: Nathaniel Haverstock moved into cooperage, Peter Frayne into harness-making, and Paul Belisle into shoemaking. In Ontario, Thomas Ripley and Swan Dean, turned from farming to cheese-making, and James Woodland became a butcher, These Ontario men left farming although they still contributed to the reorientation of Ontario agriculture towards mixed farming and animal products.

In this paper we have not considered residential change, even though relocation or migration was important for many individuals. For example, Jean Baptiste Robert was enumerated in 1871 as a farmer in Maskinongé, Quebec, on the north shore of the St Lawrence, but by 1881 he had moved 100km downstream to work as a clerk in Montreal. Eighteen-year-old Julius Galbraith was enumerated as a printer in Orangeville, Ontario in 1871; ten years later he was a publisher in a small town on the Manitoba frontier, and was married with two small children. Another small-town printer, Elias Saunders from Goderich, Ontario, also stayed in the trade; he moved to the nearby regional centre of London. His upward mobility appears to have been less rapid: on enumeration day in 1881 he was still living in a rooming house. Additional examples add to the diversity. Joseph Allan of Portland NB, went from being a clerk to an umbrella mender and moved into Saint John proper. George Scarlett in 1871 was a teacher at eighteen in Cramahe Township, but ten years later he was a travelling salesman in nearby Cobourg.

These individual trajectories are part of the broader collective experience revealed by the linked data. In this paper we report on the construction of longitudinal data and use them to refine our understanding of economic and demographic change in Canada during the 1870s. The dominant impression created by the new source is one of change amid continuity. There was considerable occupational persistence and, yet, some weakening in the case of agriculture and a tendency for younger men to switch into manufacturing and commerce. Admittedly, change was slow and gradual, and largely invisible published tabulations that, until recently, comprised our only evidence. The longitudinal micro-data, however, document complex patterns of change especially among the young that cumulatively, over several decades, would transform Canada profoundly.

[30] Admittedly, we examine the entire country while Darroch and Ornstein analyze the central regions of Ontario.

Figure 1. Occupational Transitions of Linked Males (15-55) from 1871 to 1881

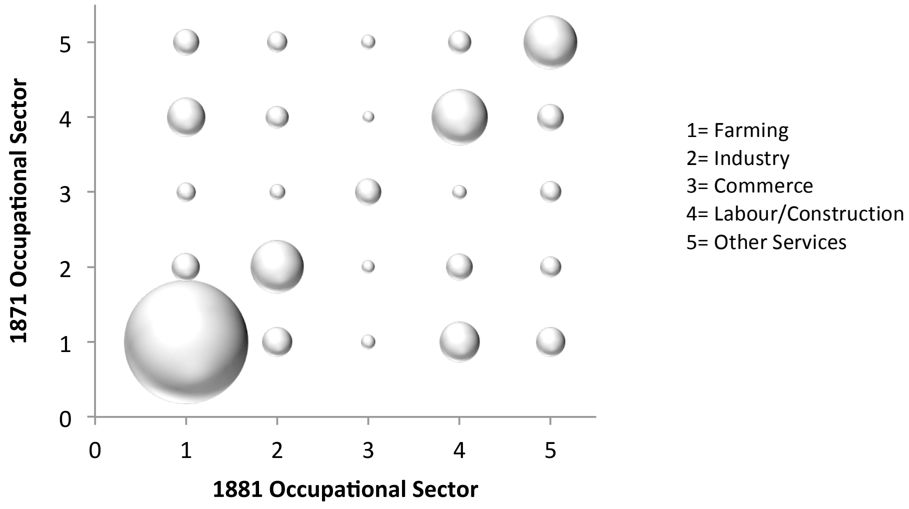

Table 1: Outcome for 1871 census records in the classification system

	No. of Records	Share
One to one links	550,726	0.22
No links returned	611,702	0.24
Many-to-one-and one-to-many links	1,397,915	0.54

Table 2: Age distribution of 1871 population and linked women and men

	Women		Men	
	Pop.	Linked	Pop.	Linked
Age in 1871				
0 to 14	.38	.40	.39	.38
15 to 25	.24	.16	.22	.19
26 to 55	.31	.37	.30	.35
56 and over	.08	.07	.09	.08

Source: Canada, Census, 1871, 5% microdata sample constructed at the University of Guelph http://census1871.ca (ignoring records for which age is missing).

Table 3. Distribution by nativity and ethnicity in the population and in linked records

	Pop.	Linked
Birthplace		
Foreign-born	.19	.20
England	.04	.06
Scotland	.04	.03
Ireland	.06	.06
Germany	.01	.01
U.S.	.02	.03
Canadian-born	.81	.80
Ontario	.33	.29
Quebec	.29	.30
Origin or ethnicity		
French	.32	.27
English/Welsh	.20	.27
Irish	.25	.23
Scottish	.14	.12
Continental Euro.	.06	.09
North American	.01	.003
African	.01	.005
Other	.01	.01

Source: as Table 2.

Table 4: 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

* indicates that the co-efficient differs significantly from 1.0 at 10% confidence level
** indicates that the co-efficient differs significantly from 1.0 at 5% confidence level,
*** indicates that the co-efficient differs significantly from 1.0 at 1% confidence level.
Note: Full regression detail is available from the authors.

Table 5: Odds ratios for finding at least one link for each record

	Married		Single/widowed	
	Women	Men	women	men
21-25 yrs	0.97	0.99	1.11***	1.16***
>55 yrs	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

Table 6: Odds ratios for finding only one link among the linked records

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

Table 7: 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

Table 8: Distribution of 1871-1881 linked pairs by gender and inter-provincial movers vs stayers

	Female	Male	All
No. of linked pairs	247,663	303,030	550,726
No. of links with change in province	8,037	9,848	17,910
Apparent movers as a share of all links	.032	.032	.033

Table 9: Individual assessment of linked pairs implying movement between provinces from 1871 to 1881

	Movers	Stayers
Number of records checked	39	1,787
Share assessed highly credible (A)	.46	.76
Share assessed credible (B)	.05	.09
Share that cannot be confirmed (C)	.15	.10
Share assessed likely incorrect (D)	.33	.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 C.

Table 10: Likely 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
2 possible states	0.47	0.34	0.21	0.12	0.08	0.05
3 possible states	0.54	0.41	0.26	0.15	0.10	0.07
4 possible states	0.57	0.44	0.29	0.17	0.12	0.07
5 possible states	0.58	0.46	0.30	0.17	0.12	0.08

Notes: We simulate the observed pattern of state change assuming records are distributed equally across all possible states and 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 apparent/phantom movers under these simple assumptions.

Table 11: Distribution of Occupations from published tabulations, Canada, 1871 and 1881

	Agricultural	Industrial	Other	N
1871				
Ontario	.49	.20	.31	463,424
Quebec	.47	.19	.34	341,291
New Brunswick	.47	.22	.31	86,488
Nova Scotia	.42	.29	.29	118,645
Total	**.475**	**.211**	**.314**	1,009,848
1881				
Ontario	.48	.21	.31	630,762
Quebec	.47	.19	.34	433,264
New Brunswick	.52	.18	.30	105,459
Nova Scotia	.45	.28	.27	141,695
Total (old Canada)	***.477***	***.207***	***.316***	*1,311,180*
Prince Edward Island	.60	.19	.21	34,132
British Columbia	.15	.38	.47	18,027
Manitoba	.58	.11	.31	23,261
Northwest Territories	.26	.07	.67	4,004
Total (all Canada)	**.477**	**.207**	**.316**	1,390,604

Source: *Census of Canada, 1870-71*, Vol. II, Table XIII (Occupations of the People), pg. 245 and *1880-81*, Vol. IV, Table J (Occupations of the People and their Ratios), pg. 72

Table 12: Distribution of occupations from published tabulations and linked microdata, old Canada, 1871 and 1881

	Agricultural	Industrial	Other
Published tabulations			
1871	.50	.14	.37
1881	.50	.14	.36
Linked microdata			
1871	.48	.21	.32
1881	.48	.21	.32

Source: Table 1 and the People in Motion record linking system http://www.people-in-motion.ca as described in the text and in Antonie *et al.* (submitted).

Table 13: Share of linked records with occupation, by age and sex

	Men		Women	
1871 Age cohort	**1871**	**1881**	**1871**	**1881**
0 to 14	0.02	0.47	0.01	0.09
15 to 25	0.77	0.92	0.12	0.12
26 to 55	0.96	0.94	0.07	0.07
56 and over	0.91	0.75	0.08	0.05

Table 14: Occupational transitions 1871-1881 of linked women with occupations, by age

		Occupation in 1881				
	N	**Farming**	**Industry**	**Commerce**	**Labour and Construction**	**Other Services**
Occupation in 1871						
Farmer	8	6	1	0	0	1
Manufacturer	40	3	27	3	2	5
Merchant	5	0	1	3	0	1
Labour/Construction	5	0	0	0	2	2
Other Service	147	8	8	1	6	124
Total	211	17	38	7	10	139

Table 15: Occupational Sectors of Linked Males in 1881 by Province of Residence in 1871, with change from 1871

	N	**Farming**	**Industry**	**Commerce**	**Labour and Construction**	**Other Services**
Ontario	4415	.56	.13	.06	.16	.09
Change from 1871		*-1.3%*	*0.0%*	*2.0%*	*-0.6%*	*-0.1%*
Quebec	2786	.49	.13	.06	.19	.14
Change from 1871		*1.4%*	*1.1%*	*1.4%*	*0.6%*	*-4.5%*
New Brunswick	865	.60	.12	.02	.15	.11
Change from 1871		*5.7%*	*-0.7%*	*0.0%*	*-3.4%*	*-1.5%*
Nova Scotia	1175	.46	.16	.03	.12	.24
Change from 1871		*3.3%*	*-1.6%*	*0.6%*	*-1.7%*	*-0.6%*

Table 16: Occupational transitions 1871-1881 of linked men with occupations, by age, Canada

		Distribution of occupations in 1881, by 1871 occupation					
		Farming	Industry	Commerce	Labour and Construction	Other Services	
Age 15-25 years in 1871 (n=2499)							
	Share in 1871						
Share in 1881		.46	.16	.06	.18	.14	
Farming	.47	**.74**	.07	.03	.12	.05	1.00
Industry	.14	.15	**.57**	.05	.12	.11	1.00
Commerce	.02	.09	.14	**.54**	.12	.12	1.00
Labour/Const	.20	.31	.15	.04	**.40**	.10	1.00
Other Service	.17	.17	.10	.12	.15	**.46**	1.00
Age 26 to 55 years in 1871 (n=5690)							
	Share in 1871						
Share in 1881		.54	.13	.05	.16	.13	
Farming	*.52*	**.86**	.03	.02	.06	.03	1.00
Industry	*.13*	.18	**.61**	.05	.09	.07	1.00
Commerce	*.05*	.15	.11	**.50**	.09	.02	1.00
Labour/Const	*.16*	.24	.09	.03	**.56**	.08	1.00
Other Service	*.14*	.16	.06	.06	.12	**.60**	1.00

Table 17: Occupations of Linked Males in 1881 by Birthplace and Age, with change from 1871 to 1881

		Farming	**Industry**	**Commerce**	**Labour/ Construction**	**Other Service**
15 to 25 years	**Canadian-born**	47.9% *-2.3%*	15.3% *2.0%*	5.5% *3.7%*	17.9% *-1.4%*	13.3% *-2.2%*
	Foreign-born	36.1% *5.7%*	21.3% *0.8%*	6.2% *4.6%*	20.0% *-4.4%*	16.4% *-6.7%*
26 to 55 years	**Canadian-born**	58.0% *1.1%*	10.7% *-0.3%*	4.2% *0.1%*	14.0% *-0.6%*	13.0% *-0.4%*
	Foreign-born	45.6% *2.1%*	16.4% *-1.1%*	6.9% *1.4%*	18.7% *0.2%*	12.4% *-2.6%*

Table 18: Occupational Sectors of Linked Males in 1881 by Ethnic Origin & Age Cohort, with change from 1871 to 1881

	Origin	N	Farming	Industry	Commerce	Labour/ Construction	Other Service
Age 15 to 25 years	French	671	44.4	14.6	6.9	22.4	11.8
			-1.7%	*1.6%*	*4.7%*	*2.1%*	*-6.7%*
	English/Welsh	675	45.5	18.7	4.6	16.3	15
			-0.6%	*2.7%*	*3.3%*	*-2.5%*	*-2.8%*
	Irish	583	45.3	16.5	5	18	15.3
			1.6%	*-0.1%*	*3.5%*	*-4.8%*	*0.0%*
	Scottish	320	44.4	18.1	6.6	16.9	14.1
			-3.4%	*5.0%*	*4.4%*	*-3.4%*	*-2.5%*
	Continental European	226	56.2	11.5	5.3	15.5	11.5
	(excluding French)		*-4.4%*	*0.4%*	*4.0%*	*0.0%*	*0.0%*
Age 26 to 55 years	French	1353	52.8	8.9	4.4	19.2	14.8
			0.7%	*-0.4%*	*0.4%*	*0.2%*	*-0.8%*
	English/Welsh	1680	48	16.1	5.6	15.5	14.7
			0.4%	*-0.3%*	*0.6%*	*0.5%*	*-1.3%*
	Irish	1311	56.6	10.5	5	17.3	10.6
			4.0%	*-1.3%*	*0.6%*	*-0.5%*	*-2.7%*
	Scottish	790	56.8	15.2	7.3	9.9	10.8
			1.9%	*-0.6%*	*1.1%*	*-2.1%*	*-0.2%*
	Continental European	494	65.8	11.9	1.8	10.1	10.3
	(excluding French)		*-0.4%*	*-0.2%*	*-0.2%*	*-0.8%*	*1.6%*

Table 19: Occupational Sectors of Linked Irish-Origin Males in 1881 by Birthplace & Age Cohort, with change from 1871

	Religion (1871)	N	Farming	Industry	Commerce	Labour/ Construction	Other Service
Age 15 to 25 years	Roman Catholic	227	38.8%	13.7%	3.5%	24.7%	19.4%
			0.9%	*-2.2%*	*2.2%*	*-3.1%*	*2.2%*
	Protestant	352	49.4%	18.5%	5.7%	13.9%	12.5%
			1.7%	*1.2%*	*4.0%*	*-5.4%*	*-1.4%*
Age 26 to 55 years	Roman Catholic	437	43.5%	12.6%	5.3%	24.7%	14.0%
			5.5%	*-0.2%*	*1.4%*	*-0.2%*	*-6.4%*
	Protestant	855	62.9%	9.4%	5.0%	13.6%	9.1%
			3.1%	*-2.1%*	*0.3%*	*-0.7%*	*-0.7%*

Appendix A: 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 ∈ 0, 1, 2 F(x) = 1 – 1/x if x∈[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.

Appendix B: An example of census records with similar attributes

1871 Census

Surname	Forename	Age	BPL	Marital status
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

Surname	Forename	Age	BPL	Marital status
Barns	Mary	20	15030	Single
Barns	Mary	22	15030	Single

BPL = birthplace

Appendix C: Protocol for Checking Automatically Generated Links

We check the reliability of links in order to prepare Table 9 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?

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.

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.

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.

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. inter-provincial) 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.

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, neighbouring 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 neighbours/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 neighbouring 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 neighbouring 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 neighbours.
    - 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.

Evaluation/Arbitration

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

References

Abel, Marjorie, and Nancy Folbre (1990) “A Methodology for Revising Estimates: Female Market Participation in the U.S. Before 1940.” *Historical Methods* 23, no. 4 (Fall): 167–76.

T.W. Acheson (1972) “The National Policy and the Industrialization of the Maritimes, 1880-1910.” *Acadiensis* 1, no. 2: 3–28.

Donald H. Akenson (1988) *Small Differences: Irish Catholics and Irish Protestants, 1815-1922: An International Perspective*. Montreal & Kingston: McGill-Queen’s University Press.

Donald H. Akenson (1999) *The Irish in Ontario: A Study in Rural History*. 2nd ed. Montreal & Kingston: McGill-Queen’s University Press.

Robert C. Allen (2009) *The British Industrial Revolution in Global Perspective*. Cambridge: Cambridge University Press.

Luiza Antonie, Kris Inwood, Dan Lizotte, and J. Andrew Ross (2014) “Tracking People over Time in 19th Century Canada.” *Machine Learning* 96: 129-146

Peter Baskerville (2015) “Wilson Benson Revisited: Movement and Persistence in Rural Perth County, Ontario, 1871-1881.” Forthcoming in Peter Baskerville and Kris Inwood, eds., *Lives in Transition: Longitudinal Research from Historical Sources*. Montreal & Kingston: McGill–Queen’s University Press.

Peter Baskerville (2008) *A Silent Revolution? Gender and Wealth in English Canada*. Montreal & Kingston: McGill-Queen’s University Press.

Susan B. Carter and Richard Sutch (1996) “Fixing the Facts: Editing of the 1880 U.S. Census of Occupations with Implications for Long-Term Labor Force Trends and the Sociology of Official Statistics.” *Historical Methods* 29, no. 1 (Winter): 1-35.

Edward J. Chambers (1964) “Late Nineteenth Century Business Cycles in Canada.” *Canadian Journal of Economics and Political Science* 30, no. 3: 391–412.

Louise Charpentier, René Durocher, Christian Laville, and Paul-André Linteau (1990) *Nouvelle histoire du Québec et du Canada*. Anjou : Centre éducatif et Culturel.

Terry Crowley (1995) “Rural Labour.” In Paul Craven, ed., *Labouring Lives: Work and Workers in Nineteenth-century Ontario*, 13–102. Toronto: University of Toronto Press.

Peter Christen (2008) “Automatic record linkage using seeded nearest neighbour and support vector machine classification”. Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining: 151–159.

Bruce Curtis (2000) *The Politics of Population: State Formation, Statistics, and the Census of Canada, 1840–1875.* Toronto: University of Toronto Press.

Gordon Darroch and Michael Ornstein (1980) "Ethnicity and Occupational Structure in Canada in 1871" *Canadian Historical Review* 61, no. 3: 305-333.

Gordon Darroch and Michael Ornstein (1984) "Ethnicity and Class, Transitions over a Decade: Ontario, 1861-1871." *Canadian Historical Association, Historical Papers,* 111–137.

Gordon Darroch (2015) Lives In Motion: Revisiting the 'Agricultural Ladder' in 1860s Ontario, A Study of Linked Microdata. Forthcoming in Peter Baskerville and Kris Inwood, eds., *Lives in Transition: Longitudinal Research from Historical Sources*. Montreal & Kingston: McGill–Queen's University Press.

Lisa Dillon (2006) "Challenges and Opportunities for Census Linkage in the French and English Canadian Context." *History and Computing* 14 (1-2): 185-212.

Lisa Dillon (2008) *The Shady Side of Fifty: Age and Old Age in Late Victorian Canada and the United States.* Montreal & Kingston: McGill-Queen's University Press.

Ian Drummond (1987) *Progress Without Planning: The Economic History of Ontario from Confederation to the Second World War*. Toronto: University of Toronto Press.

Patrick Dunae (2009) "Sex, Charades, and Census Records: Locating Female Sex Trade Workers in a Victorian City." *Histoire Sociale/Social History* 42, no. 84: 267–97.

Herb Emery, Kris Inwood, and Henry Thille (2007) "Hecksher-Ohlin in Canada: New Estimates of Regional Wages and Land Prices." *Australian Economic History Review* 47 (1): 22-48.

Joseph P. Ferrie (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 (Fall): 141-156.

Joseph P. Ferrie (1999) *'Yankees Now': European Immigrants in the Antebellum U.S., 1840-1860.* New York: Oxford University Press.

Ben Forster (1986) *A Conjunction of Interests: Business, Politics, and Tariffs, 1825-1879.* Toronto: University of Toronto Press.

Forster, Ben, and Kris Inwood (2003) "The Diversity of Industrial Experience: Cabinet and Furniture Manufacture in Late Nineteenth-Century Ontario." *Enterprise and Society* 4, no. 2: 326-371.

Allegra Fryxell, Kris Inwood and Aaron Van Tassel (2015) "Aboriginal and Mixed Race Men in the Canadian Expeditionary Force 1914-1918." Forthcoming in Peter Baskerville and Kris Inwood, eds., *Lives in Transition: Longitudinal Research from Historical Sources*. Montreal & Kingston: McGill–Queen's University Press.

Fu Zhichun, Mac Boot, Peter Christen and Jun Zhou (2014) "Automatic Record Linkage of Individuals and Households in Historical Census Data" *International Journal of Humanities and Arts Computing*: forthcoming.

Chad Gaffield (1987) *Language, Schooling, and Cultural Conflict: The Origins of the French-Language Controversy in Ontario*. Montreal & Kingston: McGill-Queen's University Press.

David Gagan, David (1982) *Hopeful Travellers: Families, Land, and Social Change in Mid-Victorian Peel County, Canada West*. Toronto: University of Toronto Press.

Ron Goeken, Lap Huynh, Thomas Lenius, and Rebecca Vick (2011) "New Methods of Census Record Linking." *Historical Methods* 44 (1): 7-14

Alan G. Green and M. C. Urquhart (1987) "New Estimates of Output Growth in Canada: Measurement and Interpretation." In D. McCalla, ed., *Perspectives on Canadian Economic History*, 182-99. Toronto: Copp Clark Pitman.

David Hacker (2013) "New Estimates of Census Coverage in the United States, 1850–1930." *Social Science History* 37(1): 71-101.

C. Knick Harley (1978) "Western Settlement and the Price of Wheat, 1872–1913." *Journal of Economic History* 38, no. 4: 865-878.

C. Knick Harley (1988) "Ocean Freight Rates and Productivity, 1740–1913: The Primacy of Mechanical Invention Reaffirmed." *The Journal of Economic History* 48, no. 4: 851–76.

Timothy J. Hatton and Jeffrey G. Williamson (1994) *Migration and the international labor market, 1850-1939*. London: Routledge.

Craig Heron (1995) "Factory Workers." In Paul Craven, ed., *Labouring Lives: Work and Workers in Nineteenth-century Ontario*, 479–590. Toronto: University of Toronto Press.

Andrew Hinson (2010) *Migrant Scots in a British City: Toronto's Scottish Community, 1881-1911*. PhD diss. University of Guelph.

Kris Inwood (1991) "Maritime Industrialization from 1870 to 1910: A Review of the Evidence and Its Interpretation", *Acadiensis* v21 n1: 132-55.

Kris Inwood and Ian Keay (2012) "Diverse paths to industrial development: Evidence from late nineteenth century Canada." *European Review of Economic History* 16, no. 3: 311–33.

Kris Inwood and Ian Keay (2013) "Trade Policy and Industrial Development: Iron and Steel in a Small Open Economy, 1870-1913." *Canadian Journal of Economics,* 46 n4: 1265-1294.

Kris Inwood, Mary MacKinnon, and Chris Minns (2013) “Labour Market Dynamics in Canada, 1891–1911”, in G. Darroch ed., *The Dawn of “Canada’s Century”: The Hidden Histories*. Montreal and Kingston: McGill-Queen’s University Press.

Kris Inwood and Gregory Kennedy (2012) “A New Prosopography: The Enumerators of the 1891 Census in Ontario.” *Historical Methods* 45: 65-77.

Kris Inwood and Richard Reid (2001) “Gender and Occupational Identity in a Canadian Census.” *Historical Methods* 32, no. 2: 57–70.

Peter R. Knights (1969) “A Method for Estimating Census Under-Enumeration.” *Historical Methods Newsletter*, 3:1, 5-8.

Peter Knights (1991) *Yankee Destinies: The Lives of Ordinary Nineteenth-Century Bostonians*. Chapel Hill: University of North Carolina Press.

Jason Long and Joseph Ferrie (2007) “The Path to Convergence: Intergenerational Occupational Mobility In Britain and the U.S. in Three Eras.” *The Economic Journal* 117: C61-C71.

William Marr (1981) “The Wheat Economy in Reverse: Ontario Wheat Production 1887-1917." *Canadian Journal of Economics* 14: 136-45.

Marvin McInnis (2000) “The Population of Canada in the Nineteenth Century.” In *A Population History of North America*, 371-432. New York: Cambridge University Press.

Cathy McDevitt, Jim Irwin, and Kris Inwood (2009) “Gender Pay Gap, Productivity Gap and Discrimination in Canadian Clothing Manufacturing in 1870.” *Eastern Economic Journal* 35: 24-36.

Sherry Olson. 2015. “Ladders of Mobility in a Fast-growing Industrial City: Two by Two, and Twenty Years Later.” Forthcoming in Peter Baskerville and Kris Inwood, eds., *Lives in Transition: Longitudinal Research from Historical Sources*. Montreal and Kingston: McGill–Queen’s University Press.

Sherry H. Olson and Patricia Thornton (2011) *Peopling the North American City: Montreal, 1840-1900.* Montreal: McGill-Queen's University Press.

Donald Parkerson (1991) “Comments on the Underenumeration of the U.S. Census, 1850-1880.” *Social Science History* 15 (4): 509-515.

Lawrence Philips (2000) “The double metaphone search algorithm.” *C/C++ Users Journal.*

Alison Prentice (1977) *The School Promoters: Education and Social Class in Mid-Nineteenth Century Upper Canada*. Toronto: McClelland and Stewart.

Richard Reid (1995) “The 1871 United States Census and Black Underenumeration.” *Histoire sociale/Social History* 28: 487-499.

Laura Richards (2013) *Disambiguating Multiple Links in Historical Record Linkage.* MA diss., University of Guelph.

Evan Roberts, Matthew Woollard, Chad Ronnander, Lisa Dillon, and Gunnar Thorvaldsen (2003) "Occupational Classification in the North Atlantic Population Project." *Historical Methods* 36 (2, Part 2): 89-96.

Steven Ruggles (2006) "Linking Historical Censuses: A New Approach." *History and Computing* 14 (1-2): 213-224.

Andrew Smith (2008) *British Businessmen and Canadian Confederation: Constitution Making in an Era of Anglo-Globalization*. Montreal and Kingston: McGill-Queen's University Press.

Richard H. Steckel (1988) "The Health and Mortality of Women and Children, 1850–1860." *Journal of Economic History* 48 (2): 333-345.

Richard H. Steckel (2008) "Biological Measures of the Standard of Living." *Journal of Economic* Perspectives 22 no. 1 (Winter): 129–152.

Stephen Thernstrom (1964) *Poverty and Progress: Social Mobility in a Nineteenth Century City*. Cambridge: Harvard University Press.

Stephen Thernstrom (1973) *The Other Bostonians: Poverty and Progress in the American Metropolis, 1880-1970*. Cambridge: Harvard University Press.

Malcolm C. Urquhart (1986) "New Estimates of Gross National Product Canada, 1870-1926." In *Long Term Factors in American Economic Growth*, edited by S.L. Engerman and R.E. Gallman, 9-94. Chicago: University of Chicago Press.

Malcolm C. Urquhart (1993) *Gross National Product, Canada, 1870-1926: The derivation of the Estimates*. Montreal and Kingston: McGill-Queen's University Press.

Vladimir N. Vapnik (1995) *The Nature of Statistical Learning Theory*. Springer Verlag: Heidelberg.

Michael Wayne (1995) "The Black Population of Canada West on the Eve of the Civil War." *Histoire sociale/Social History* 28: 465-485.

Randy W. Widdis, (1989) "Tracing Eastern Ontarian Emigrants to New York State, 1880-1910." *Ontario History* LXXXI, no. 3: 201–23.

Kevin H. O'Rourke and Jeffrey G. Williamson (1999) *Globalization and History: The Evolution of a 19th Century Atlantic Economy*. Cambridge, MA: MIT Press, 1999.

William E. Winkler. 2006. Overview of record linkage and current research directions. *Statistical Research Division Report*. U.S. Census.