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.

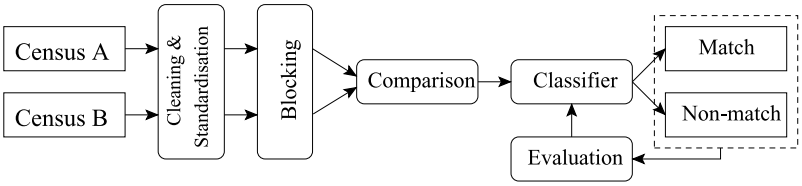

Fig. 1 Overview of record linkage system

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

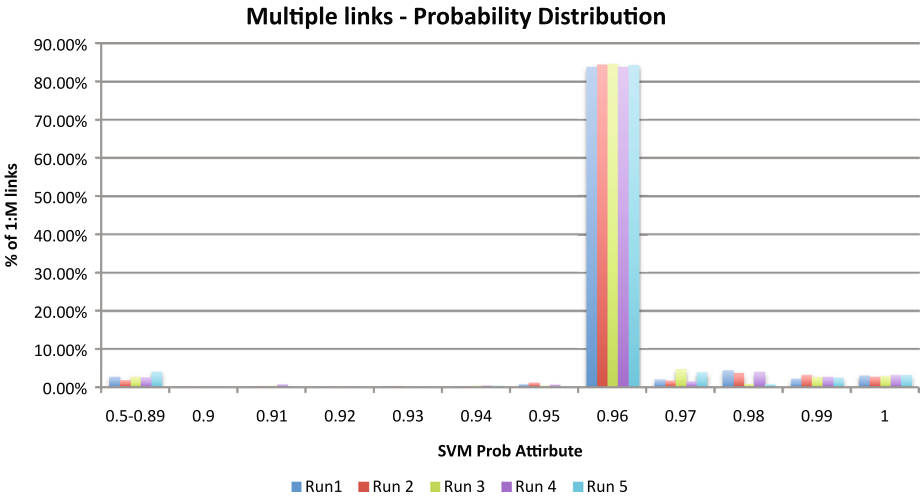

Fig. 2 Classification system probability score distribution

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.