LINKING HISTORICAL CENSUSES: A NEW APPROACH

STEVEN RUGGLES

This article describes a new initiative at the Minnesota Population Center (MPC) to create linked representative samples of individuals and family groups from the censuses of 1860, 1870, 1900 and 1910 to the 1880 census. This set of linked samples will provide new opportunities for researchers to carry out individual-level analyses of social and geographic mobility and family transitions.

This study differs from past efforts to link persons across census years in one key respect. The central goal of previous studies has been to maximize the proportion of the population linked. Our primary goals, however, are to minimize selection bias and maximize representativeness of the linked cases. To achieve these goals, we are prepared to sacrifice a significant number of demonstrably valid links.

The project takes advantage of an extraordinary new data source, a complete transcription of the 1880 census of the United States. We will also capitalize on recent advances in record linkage technology.

BACKGROUND: THE 1880 LDS DATABASE AND THE IPUMS SAMPLES

The Church of Jesus Christ of Latter-Day Saints (LDS) and the Minnesota Population Center (MPC) have produced a remarkable machine-readable database encompassing the entire population of the United States enumerated in the 1880 Census of Population. Over a thousand LDS volunteers spent 11.5 million hours over an eighteen-year period transcribing the census data from the original enumerators' manuscripts. Their goal was to create an electronic lookup system for genealogical research.

In 1999, the MPC reached an agreement with the LDS to verify and correct the census transcription in exchange for the right to disseminate the resulting database for scholarly and educational purposes. In addition to making the

History and Computing 14 (1+2) 2002 (publ. 2006), 213–224 ISSN 0 957-0144

changes required by the LDS, the project involved converting the files from a raw transcription of the census text into a numerically-coded and documented database suitable for statistical analysis.

We have published a description of the editing and coding procedures used for the 1880 project elsewhere.[1] To summarize very briefly, we corrected a range of technical errors introduced by the LDS in the course of data processing; corrected missing and incorrect geographic identifiers; identified missing cases and entered them; eliminated duplicate records; corrected flags distinguishing the breaks between households; identified and edited internal inconsistencies in family relationship, sex, marital status and age; and developed data dictionaries to classify cases according to standardized coding systems for geographic variables, group quarters type, place of birth and occupation.

Because of the large scale of the database, these tasks proved challenging. The 1880 database has fifty million cases, so even routine jobs require substantial investment of resources. For example, we entered several hundred thousand missing cases and coded over 500,000 alphabetic occupational strings into four separate classification systems. Through careful planning and extensive use of automated and semi-automated data editing tools, we have achieved unprecedented cost efficiencies in editing and coding the data. We are now on schedule to complete the project on time and within budget.[2]

The quality of the 1880 population database is good. Because the data were intended for genealogical purposes, the LDS placed a premium on transcription accuracy. LDS volunteers entered each of the fifty million cases twice so they could carry out blind verification. The technical corrections carried out by MPC have resulted in, for practical purposes, a complete transcription: in most instances, the county population totals in the database match the published statistics precisely and the population count for the country as a whole exceeds the published totals by less than 3,000 cases, or 0.006 per cent.[3]

In addition to the LDS data, the linking study will make use of census samples from the Integrated Public Use Microdata Series (IPUMS). The IPUMS is a harmonized series of U.S. census microdata samples spanning the period from 1850 to 2000. The samples range in density from one to five per cent of the population and for the period prior to 1940 they include names and addresses as well as the characteristics of each individual. Since its release in 1995, the IPUMS has attracted many users and has been used approximately 2,000 research papers and dissertations. We have coded the 1880 LDS data to be highly compatible with the IPUMS.

LINKED CENSUS SAMPLES

Perhaps the greatest limitation of the existing series of IPUMS samples is that they are cross-sectional snapshots and do not allow one to trace individuals

across time. Each sample is independent and cannot be linked to other samples to provide two observations for the same individual. It is possible, however, to link each of these samples to the new 1880 database. Using new record-linkage technology, we will construct linked samples covering pairs of census years: 1860–1880, 1870–1880, 1880–1900 and 1880–1910. Each of the linked samples will be independent, but taken together they will provide a rich longitudinal source for the nineteenth and early twentieth centuries.

The new database holds the promise of resolving some of the longest-running debates in American social history. Past studies were often inconclusive because of their exclusion of migrants and their small sample size. Scholars will be able to gauge the extent of social and geographic mobility, the interrelationship of geographic and economic movement and trends and differentials in social mobility more reliably than heretofore.[4] In addition, the linked samples will allow investigation of family formation and dissolution. For example, they will allow us to settle a debate about the formation of multigenerational households in the nineteenth century, an issue with important implications for the study of intergenerational relations and the twentieth-century transformation of the living arrangements of the aged.[5]

Historians have been linking individuals across censuses for decades, but the results have been problematic. In most cases, linked census studies have been based on local populations because no complete census for a larger area has been available. These studies generally lose between 60 and 80 per cent of the population each decade due to linkage failures.[6] The investigators attributed the high rate of linkage failure to the high migration of the mid-nineteenth century, which meant that individuals moved out of the study area.

There have been two national studies that linked individuals across census years. In 1987, NICHD funded an ambitious attempt by Thomas Pullum and Avery Guest to create a national linked 'panel' of two cohorts of men in the 1880 and 1900 censuses. Pullum and Guest linked 4,014 individuals between these two census years out of a sample of 10,252, for a linkage rate of 39.4 per cent.[7] More recently Joseph Ferrie, with funding from NSF, linked a nationally representative subset of the 1850 public use microdata sample to the 1860 manuscript census.[8] Ferrie limited his study to persons with uncommon surnames, but still achieved only a 19.3 per cent linkage rate for a total of 4,938 linked cases.

The availability of a high-quality census file including the entire 1880 population opens the door to far more sophisticated matching than has previously been possible. The Pullum/Guest and Ferrie samples had to rely on the state Soundex name indexes to locate individuals in the census. Interstate migrants were for the most part lost. The process was labour-intensive, expensive and involved large potential for human error and selection biases.

Using the new 1880 database in combination with recent advances in record-

matching technology, however, the entire country can be searched using such characteristics as age, sex, birthplace, birthplace of mother and birthplace of father, as well as name. The new database will be able to provide samples many times larger than the Pullum/Guest and the Ferrie studies. Just as important, however, we can design the new linked samples to minimize major problems of selection bias that were an inevitable concomitant of previous linking methods.

RECORD LINKAGE PROCEDURES

Overview

We will exploit new record-linkage and data-mining technology to create linked representative samples of individuals and family groups from the censuses of 1860, 1870, 1900 and 1910 to the 1880 census. These representative linked samples will provide unprecedented opportunities for researchers to carry out individual-level analyses of social and geographic mobility and family transitions in the early stages of industrial development.

For over half a century, social scientists have linked records from different sources to create longitudinal historical datasets.[9] During the past fifteen years, however, technological developments have opened new opportunities to create more powerful linked historical datasets than were previously possible.[10] This new technology derives from two main sources. First, central statistical agencies in North America and Europe have invested in the development of techniques for matching information from censuses, vital statistics and administrative records. Second, work on data-mining techniques – carried out by both academics and commercial software developers – has contributed to methods of data cleaning and probabilistic linkage

Our procedures will build on these innovations. Our goals, however, differ significantly from those of most recent researchers. The primary goal of virtually all the work on record linkage has been to maximize the number of valid links. A typical data-mining application, for example, would involve linking membership records to address lists to identify potential sales prospects. The goal of such an application is not to create a statistically valid representative sample, but simply to generate the largest possible number of customers. The most important linking application for statistical agencies is the estimation of undercount through the capture-recapture method, so they also aim for the largest possible number of reliable links.

We will not focus on maximizing the number of accurate links. Instead, our procedures will be designed to maximize the *representativeness* of the linked cases. This means that we must pay close attention to sources of selection bias and ignore much of the information routinely used by other record linkage procedures.

The principal applications of the samples will be the study of social mobility, migration, family change and life-course transitions. We therefore must avoid using any information that could bias the sample with respect to those changes in characteristics. For example, record linkage algorithms ordinarily make use of place of residence as a linking variable. This greatly increases the potential for reliable links: if we identify an individual in the 1870 sample who partially matches the name and other characteristics of a person in the 1880 database, our confidence that the two records refer to the same person would be improved if we knew that they both reside in Poughkeepsie. If we use place of current residence in the linking algorithm, however, we will inevitably bias the sample in favour of non-migrants. Likewise, if we use spouse's name in the algorithm we will bias the sample in favour of those who remain married and if we use occupation we will favour cases with low social mobility.

Planned samples

We plan three categories of linked samples, each with a different universe: all males, females who do not marry in the census interval and married couples. All three samples will be further restricted to the population old enough to have been alive in both census years. Although none of these groups is representative of the entire population, our goal is to make each category representative of its defined universe. The male individual sample will be general purpose, useful for studying economic and geographic mobility, transitions to adulthood, changes in family composition and retirement. The female sample will be useful for studying many of the same topics, but will apply to the subset of women who do not change their surname between censuses and therefore will be inappropriate for some topics. The married-couple samples will offer the greatest reliability, since it will allow us to link on characteristics of both husband and wife and will be especially useful for topics relating to fertility, child mortality and age of leaving home. Because it is restricted to the continuously married population, however, it will be less useful for population-wide generalizations about social and geographic mobility. Although we are linking individuals or couples, we will also capture all characteristics of all co-resident household members.

For each sample, we will start by identifying a subset of individuals in the IPUMS one-per cent samples (1860, 1870, 1900, or 1910). We will then search for these individuals in the complete 1880 census database. We will create three linked samples for each pair of census years, for a total of twelve samples. Half of the samples use forward links (1860 and 1870 to 1880) and half rely on backward links (1900 and 1910 to 1880). Forward-linked samples are more challenging than the backward-linked ones because mortality and emigration substantially reduce the potential for links. Moreover, the 1860 and 1870 censuses are missing two key linking variables, birthplace of mother and

birthplace of father. Nevertheless, we believe that the substantive importance of linked samples in the earlier period justifies linking in both directions. The 1870 census offers the earliest potential to trace the bulk of the black population and the geographic and occupational mobility of blacks after the Civil War is a subject of enormous historical interest and importance. Linking to the 1860 census, though restricted to the free population, will help us to gauge the demographic consequences of the war itself.

Backward linkage is simplified by the availability of retrospective information in the 1900 and 1910 censuses about year of immigration for the foreign-born population. Together with age, this question will allow us to define a universe that includes only persons who were alive and resident in the United States in 1880. Although this universe will be imperfect because of errors in enumeration and transcription, it will allow more aggressive linking strategies by reducing uncertainty. The 1900 and 1910 samples also include a richer set of census questions than any censuses before the mid-twentieth century, including retrospective inquiries about marriage, children born and surviving, immigration and naturalization, which will augment the longitudinal dimension of the linked samples.

Linking characteristics

Our algorithm will rely exclusively on characteristics that would not change over time if there were no enumerator errors, transcription errors, or name changes. The linking characteristics we plan to use for each census year are given in Table 1. For the married-couple samples, these characteristics are available for both husband and wife.

Genealogists and data miners make use of a considerably broader range of characteristics to confirm links and resolve ambiguities. We believe, however, that knowledge of any additional characteristics would introduce biases that would severely damage the samples. The chief problem posed by our approach is that this limited set of variables is insufficient to identify individuals uniquely.

Table 1. Variables available for record linkage, 1860–1910.

1860 to 1880	*1870 to 1880*	*1900–1910 to 1880*
First name	First name	First name
Last name	Last name	Last name
Birth year	Birth year	Birth year
Sex	Sex	Sex
Race	Race	Race
State or country of birth	State or country of birth	State or country of birth
	Father foreign-born	Father's state or country of birth
	Mother foreign-born	Mother's state or country of birth

For example, the 1880 census includes seventeen white men aged thirty-three who were named John Smith and born in New York State. This is, of course, a worst-case scenario, since John Smith was the most common male name and New York was the largest state. Nationally, we estimate that over three-fourths of the population can be uniquely identified by the limited variable set available in 1860. For the married-couple universe, virtually every case would be uniquely identified.

In practice, however, those numbers are optimistic. Because of errors in enumeration and transcription, a high proportion of matches are imperfect: linking must be carried out on a probabilistic basis, allowing for imperfect correspondence of names and ages. Allowing for such near matches, the proportion of uniquely identified individuals would decline significantly.

To reduce the potential for ambiguity, we will follow the precedent of Ferrie and eliminate names that identify multiple persons of the same age and birthplace.[11] Ferrie found that this procedure creates little bias with respect to ethnicity or other characteristics, but we will weight the remaining cases to eliminate any significant biases with respect to birthplace and parental birthplace, state of residence, or occupation.

Name cleaning and metaphones

Record linkage begins with software for parsing and standardizing names. Names are by far the most important piece of information available for record linkage, but they are the most problematic. Errors in naming can arise from respondent error (as when, for example, a farm wife responding to an enumerator misstates the name of a farm hand), enumerator error, or transcription error. Moreover, names often change over time, sometimes did not have standard spellings and in some cases people will be enumerated under a nickname or middle name in one census and under their formal first name in the other.

To minimize error from these sources, we plan a comprehensive program of name cleaning, accounting for common typographical transpositions, handwriting recognition errors and common nicknames. This work will draw on the rich body of research on name cleaning carried out during the past decade.[12]

We will also employ phonetic name coding, a standard tool for record linkage since the 1930s. The most commonly used systems are Soundex, NYSIIS and Phonex. All of these systems lose much of the phonetic detail, however. Although we have not yet finalized our phonetic coding plans, we prefer the more subtle Double-Metaphone system, which returns two encoded strings corresponding to variant pronunciations.[13]

Linking algorithm

Because there are multiple opportunities for errors to be introduced, it is essential that the linking algorithm accommodate approximate matches on a probabilistic basis. Planning and design of the linking algorithm is therefore a significant component of the project. The design must consider not only optimization of links, but also computational efficiency: some techniques are extraordinarily computationally intensive and would be unfeasible for a project of this scale.[14]

The theoretical framework of record linkage originates from Fellegi and Sunter, who demonstrated that it is possible to define an optimal linkage rule that minimizes the number of false links.[15] In addition, Fellegi and Sunter derived a test statistic for evaluating error rates and specified the assumptions necessary for estimating the matching probabilities used to calculate the test statistic. Extensions and refinements of record linkage theory were contributed by Jaro, Winkler, Belin and Rubin and Larson and Rubin.[16]

All these models assume that every pair of records drawn from two files are either matches referring to a single individual or non-matches describing two different persons; optimal matching requires that every individual be compared with every possible match. It is not computationally feasible to implement every potential match; for example, implementation of such a linking algorithm for the full 1880 database and the 1900 sample would involve over fifteen trillion comparisons. To reduce the computational requirements, we will introduce 'blocking factors' – such as state of birth, race and sex—and limit comparisons to persons who share the same blocking factors. If necessary, we will make an additional blocking pass based on metaphone. The computational problem will nevertheless be large and we will carefully explore various methods that have been proposed to improve efficiency.[17]

Our linking algorithm will depart from current practice in several respects. As noted, we will ignore information that can change over time for reasons other than misreporting, data-entry error, or deliberate name changes. The elimination of common names will reduce the number of multiple matches, but it will not eliminate them. Wherever there is no a clear favourite, we will drop the case. Some linking strategies make use of the relative frequencies of different linking variables.[18] Thus, for example, individuals with uncommon characteristics, such as persons aged ninety-five born in Delaware, receive a higher linking score than persons with common characteristics, such as five-year-olds born in New York, simply because of the differential probability of those characteristics occurring by chance. We will avoid this approach since it will introduce selection bias favouring persons with uncommon characteristics.

Because our linking strategy must rely heavily on names, identification of the optimal approximate string comparison algorithm is of paramount importance. Many algorithms have been proposed. For example, the Jaro string comparator

as modified by Winkler computes a similarity measure between 0.0 and 1.0 based on the number of common characters in two strings, the lengths of both strings and the number of transpositions, accounting for the increased probability of typographical errors towards the end of words.[19] Since developments in this field are proceeding rapidly, however, a superior algorithm may appear during the course of the project.

The other linking variables – birthplace, parental birthplaces, age, sex and race – pose few string comparison problems because those variables are already classified and numerically coded according to the IPUMS coding system. Thus, for example, we will not have to cope with the innumerable spelling variations of Massachusetts. We will, however, need to develop an algorithm for age misreporting that can account for digit preferences: inconsistencies in age between two census years should be partly discounted if age is rounded to a five or zero in one or both census years.

Whenever possible, we plan to build on open-source software for both data cleaning and record linkage. In particular, we are making extensive use of the 'Freely extensible biomedical record linkage' (Febrl) software created by Peter Christen and Tim Churches at the Australian National University.[20]

Training data

To estimate the matching parameters and error rate of the linking algorithm and to refine the linking strategy, we need a set of training data. Training data consist of cases where the true links are known. Ordinarily, training data are compiled by hand-coding a subset of cases and we will follow that procedure. In addition, however, we will capitalize on a very large set of training data we already have in hand: the one-per cent sample of the 1880 census created by the MPC.

We can divide the potential sources of failure in record linkage across census years into seven categories:

1. Departure from universe through death or emigration. (This applies only to the forward links, 1860 and 1870 to 1880.)
2. Name changes due to marriage, Anglicization, etc.
3. Enumerator error in recording names or other characteristics.
4. Census under-enumeration.
5. Multiple valid links: two or more persons exist with similar or identical linking characteristics.
6. Transcription error, either by MPC staff or by the LDS.
7. Omission of records from the LDS file.

Of these seven sources of linkage failure, the first four are only relevant for links across census years. The last three sources of error, however, also apply to links between the one-per cent 1880 sample produced by MPC in the early 1990s

and the complete 1880 database entered by the LDS. Since both files include locater information (microfilm reel number and census page number) it is straightforward to identify true links. Using only the variables that will be available for cross-census linking, we will be able to tune the algorithm to maximize accuracy for the 500,000 persons in the one-per cent file. We will then apply the algorithm to a much smaller set of hand-coded data from the other census years and make further adjustments needed to account for the other four sources of linkage failure.

Discussion

By deliberately ignoring much of the information available for linking in order to minimize selection bias, it is possible that our linked samples will include a higher percentage of false matches than do previous linked samples. This would bias the results through a different mechanism: true links are likely to have lower geographic and social mobility than would random pairs of individuals who are incorrectly linked. This means that although conventional linking procedures almost certainly understate geographic and social mobility, if our samples have a higher frequency of false matches they could actually *overstate* such mobility. We will take several steps to evaluate the potential for this problem and to minimize its impact.

We will have three main sources of information about false-positive matches. The first is the training data described above, including both hand-linked samples and the 1880 IPUMS sample. The second will be the cases rejected because of multiple competing matches. As described above, where there are multiple competing matches that all have a reasonable probability of being correct, we will not make any match. We will, however, evaluate such cases by examining additional characteristics and determine where possible which match is correct. Even though the corrected data will not be added to the linked samples, it provides an excellent source for the study of false matches. The third source of information about false links will be the married couple samples. We will carry out analyses of the married-couple linked samples to identify cases in which the extra information on spouses yields results that differ from what would be attained with individual links only.

Each of these three sources has technical limitations, but taken together they will provide a rich body of evidence on the success of our linking strategy and will allow us to develop estimates of unobserved error rates in the linked samples. Because these sources will allow direct comparison of transitions for false links and true links, they will help us understand the ways in which false-positive links bias the results. Moreover, information from rejected multiple matches, married-couple samples and training data will help us further tune the algorithm to specify thresholds for matches that minimize false positives.

We do not contend that our approach – using a minimal set of matching variables and a fully-automated algorithm – will result in perfectly unbiased samples; given the available information, that is probably impossible. We do expect, however, that the new samples will be far more representative of the population than are hand-linked samples that make use of all the information on the record.

ENDNOTES

[1] R. Goeken, C. Nguyen, S. Ruggles and W. Sargent, 'The 1880 United States population database', *Historical Methods*, 32 (2003), 27–34.

[2] We completed the work required by the LDS on schedule and they released the database on a set of fifty-five CD-ROMs and through an on-line lookup system (http://www.familysearch.org). We released a preliminary, partially coded version of the database to academic researchers in 2001 and a revised, coded, version in 2003.

[3] Most of the extra cases represent persons who were determined to be deceased on census day through comparison with the mortality schedules. These cases were transcribed by the LDS, but were not counted by the census. See Goeken *et al*., 'The 1880 United States population database'.

[4] J. P. Ferrie, 'The end of American exceptionalism?: mobility in the U.S. since 1850', *Journal of Economic Perspectives*, 19 (2005), 199–215.

[5] S. Ruggles, 'The transformation of American family structure', *American Historical Review* 99 (1994), 103–28; S. Ruggles, 'Living arrangements and economic well-being of the aged in the past', *Population Bulletin of the United Nations* 42/43 (2001), 111–61; S. Ruggles, 'Multigenerational families in nineteenth-century America', *Continuity and Change* 17 (2003), 139–65.

[6] M. B. Katz, *The people of Hamilton, Canada West: family and class in a mid-nineteenth-century city* (Cambridge, 1975); P. Knights, *Yankee destinies: the lives of ordinary nineteenth-century Bostonians* (Chapel Hill, 1991); S. Thernstrom, *Poverty and progress; social mobility in a nineteenth century city* (Cambridge, 1964).

[7] A. Guest, 'Notes from the National Panel Study: linkage and migration in the late-nineteenth century', *Historical Methods*, 20 (1987), 63–77.

[8] J. P. Ferrie, 'A new sample of males linked from the Public-Use-Microdata-Sample of the 1850 U.S. federal census of population to the 1860 U.S. federal census manuscript schedules', *Historical Methods*, 29 (1996), 141–56.

[9] P. Rosenthal, 'Thirteen years of debate: from population history to French historical demography (1945–1958)', *Population: An English Selection*, 9 (1997), 215–41.

[10] Committee on Applied and Theoretical Statistics, National Research Council, *Record Linkage Techniques – 1997: Proceedings of an International Workshop and Exposition* (Washington, D.C, 1999).

[11] Ferrie, 'A new sample of Males'; J. P. Ferrie, *How ya gonna keep 'em down on the farm [when they've seen Schenechtady]?: rural to urban migration in nineteenth-century America 1850–1870*. Working paper, Northwestern University (Evanston, 1999).

[12] E. H. Porter and W. E. Winkler. 'Approximate string comparison and its effect on an advanced record linkage system', Census Bureau Research Report RR97/02 (Washington D.C., 1997); P. Christen, T. Churches and J. X. Zhu, 'Probabilistic name and address cleaning and standardisation', in *Proceedings of the Australasian Data Mining Workshop* (Canberra, December 2002); W. E. Winkler, 'String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage', *American Statistical Association 1990 Proceedings of the Section of Survey Research Methods* (1990), 354–59; L. Nygaard, 'Name standardization in record linkage: an improved algorithmic strategy', *History and Computing*, 4 (1992), 63–74; J. I. Maletic and A. Marcus, 'Data cleansing: beyond integrity analysis' in *Proceedings of the*

Conference on Information Quality (Boston, 2000), 200–9.

[13] L. Philips, 'The double-metaphone search algorithm', *C/C++ User's Journal*, 18 (2000); A. J. Lait and B. Randell. 'An assessment of name matching algorithms', Department Technical Report Series No. 550, Department of Computing Science, University of Newcastle upon Tyne, UK, 1993.

[14] P. Christen *et al*., 'High-performance computing techniques for record linkage', *Proceedings of the Australian Health Outcomes Conference* (AHOC-2002) (Canberra, 2002).

[15] I. P. Fellegi and A. B. Sunter, 'A theory for record linkage', *Journal of the American Statistical Association*, 40 (1969), 1183–1210.

[16] M. A. Jaro, 'Advances in record linking methodology as applied to matching the 1985 census of Tampa, Florida', *Journal of the American Statistical Association*, 84 (1989), 414–20; W. E. Winkler, 'Improved decision rules in the Fellegi-Sunter model of record linkage', *American Statistical Association 1993 Proceedings of the Section of Survey Research Methods*, 274–79; T. R. Belin and D. B. Rubin, 'A method for calibrating false-match rates in record linkage', *Journal of the American Statistical Association*, 90 (1995), 694–707; M. Larson and D. B. Rubin, 'Iterative automated record linkage using mixture models', *Journal of the American Statistical Association*, 96 (2001), 32–41.

[17] L. Jin, C .L. Liang and S. Mehrotra, 'Efficient Record Linkage in Large Data Sets', paper presented at the 8th International Conference on Database Systems for Advanced Applications (Kyoto, 2003); V. S. Verykios, A. K. Elmagarmid and E. N. Houstis, 'Automating the approximate record matching process', *Journal of Information Sciences*, 126 (2000), 83–98; M. A. Hernádez and S. J. Stolfo, 'Real-world data is dirty: data cleansing and the merge/purge problem', *Data Mining and Knowledge Discovery*, 2 (1998), 9–37; A. E. Monge and C. Elkan, 'An efficient domain-independent algorithm for detecting approximately duplicate database Records', paper presented at SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery (Tucson 1997).

[18] M. B. Fortini, A. Liseo, A. Nuccitelli and M. Scanu, 'On Bayesian record linkage', in E. I. George, ed., *Bayesian methods with applications to science, policy and official statistics* (Pittsburgh, 2000), 159–64; W. E. Winkler, 'Methods for record linkage and Bayesian networks' Census Bureau Research Report RRS2002–05 (Washington D.C., 2002); H. B. Newcombe *et al*., 'Automatic linkage of vital records', *Science*, 130 (1959), 954–959.

[19] E. H. Porter and W. E. Winkler, 'Approximate string comparison'.

[20] Febrl is distributed by the Australian National University at http://datamining.anu.edu.au/software/febrl/febrldoc/.