BUREAU OF THE CENSUS
STATISTICAL RESEARCH DIVISION
Statistical Research Report Series
No. RR2001/03

Record Linkage Software and Methods for Merging Administrative Lists

William E. Winkler
Statistical Research Division
Methodology and Standards Directorate
U.S. Bureau of the Census
Washington D.C. 20233

Report Issued: 07/23/2001

Record Linkage Software and Methods for Merging Administrative Lists

William E. WINKLER
Bureau of the Census, Washington, DC 20233-9100 USA
E:mail: william.e.winkler@census.gov

Abstract. National Statistical Institutes often have the need to merge administrative files from a variety of sources for which unique identifiers are not available to facilitate matching. Agencies such as Eurostat have the need to connect data sources from different countries and sources and to verify the confidentiality of microdata. To do this merging of administrative lists, agencies need fast software for cleaning up and standardizing lists and for merging the lists. The U.S. Bureau of the Census has software for name standardization, address standardization, and matching that are considered state-of-the-art. The standardization software breaks names and addresses into components that are easily compared. The matching software accounts for typographical error, automatically estimates matching parameters, and optimizes sets of assignments over large groups of pairs of records.

Keywords: matching, standardization

Introduction

Organizations often have the need to identify duplicates within large databases. In a population register, some individual entities (either persons or enterprises) may be listed under two or more registry numbers. To identify duplicates, name, address, date-of-birth, and other information may be needed. Because names do not uniquely identify, address or date-of-birth information is also needed. If names, addresses, and date-of-births contain typographical errors, then identification of duplicates can be difficult.

Duplicates can arise in a variety of situations. Duplicates can arise when a large database is updated using an external source and registry numbers are not available or are in error. Duplicates will happen when two or more databases are combined in order to obtain varying data for economic and demographic analyses.

Record linkage is the methodology of bringing together corresponding records from two or more files or finding duplicates within files. The term *record linkage* originated in the public health area. It refers to records or files of individual patients were brought together (linked) using name, date-of-birth and other information. In recent years, advances have yielded computer systems that incorporate sophisticated ideas from computer science, statistics, and operations research.

The ideas of modern record linkage originated with geneticist Howard Newcombe (Newcombe et al. 1959, see also 1988, 1992) who introduced odds ratios of frequencies and the decision rules for delineating matches and nonmatches. Newcombe’s ideas have been implemented in software that is used in many epidemiological applications. Newcombe’s methods often rely on odds-ratios of frequencies that have been computed a priori using large national health files. Fellegi and Sunter (1969) provided the formal mathematical foundations of record linkage. Their theory demonstrated the optimality of the decision rules used by Newcombe and introduced a variety of ways of estimating crucial matching probabilities (parameters) directly from the files being matched.

The outline of this paper is as follows. The second section gives more examples of record linkage applications. The third section covers methods and software for cleaning lists and putting them in a standardized form that facilitates matching. The fourth section covers matching methods. The fifth section covers the human skills needed for record linkage. The sixth section covers available software. The final section is concluding remarks.

Examples of Record Linkage

There are at least four ways that record linkage can be used.

Measuring a population by capture-recapture

A population is covered by two independent listings. Each listing covers the same set of small geographic regions. The population estimate is $s_A s_B / s_{AB}$ where s_A is size of first population, s_B is size of second population, and s_{AB} is size of overlap. A set of record linkage procedures that are of high quality can reduce the error is measuring s_{AB}. Clerical matching is too error prone and too slow. Table 1 shows the clerical resources and time needed for matching the U.S. Census with a post enumeration survey. The clerical and 1988 columns are adjusted upwards from samples to the 1990 resource counts. The computer-assisted matching procedure reduced clerical resources significantly.

Table 1. Resources for Matching

	clerical	1988	1990
# clerks	3000	600	200
# months	6	1.5	1.5
false match rate	5%	0.5%	0.2%
computer match proportion	0%	70%	75%

Updating and unduplicating a survey frame

Table 2 shows the improved use of resources in creating lists for the 1987 and 1992 Censuses of Agriculture.

Table 2. Identifying duplicates in 6 million records from 12 lists

	1987	1992
duplicates	6.6%	12.8%
potential duplicates	28.9%	19.7%
final file duplication	~10%	~2%
clerical resources	75 clerks for 3 months	6500 person hours

Bringing together two files (economic or demographic) so that analyses of microdata can be performed

Two files with common identifying information may not be very good for matching. Linkage procedures may have moderately high error. File A has y variable and file B has x variable. After linkage, we want good estimate of the true relationship $y = \beta x$. If there is linkage error, then the observed (x,y) may be significantly different from the true (x,y). Analyses need to be adjusted for linkage error. See Scheuren and Winkler (1993, 1997) and Lahiri and Larsen (2000). Somewhat related work for databases is given in Koller and Pfeffer (1998), Friedman et al. (1999), and Getoor et al. (2001).

Re-identification experiment to determine the confidentiality of analytically valid, public-use microdata

There is increased demand for public-use micro-data. Users of statistical information want more information than is available from summary tables. Because there are much better sets of micro-data (either from public files or from large privately held files such as credit agencies), it is much easier to perform re-identification experiments that can potentially attach names and addresses to records in public-use micro-data files. See Winkler (1998) and Domingo-Ferrer et al. (2001).

Cleaning and Standardizing Files

Matching is dependent on identifiers having errors and inconsistencies. Some matching uses names and addresses that are often difficult to put in comparable formats. Other types of matching can use name and full date-of-birth. Still other types can use other variables such as income, cost of housing, size of business, and number of employees. Different matching metrics are needed for comparing different types of fields such as surname (family name), house number, year-of-birth, age, and income. Record linkage must deal with the lack of unique identifiers (matching fields) and errors in the identifiers.

In the U.S., a common name is John Smith. If we match a record with the name John Smith against a large file, then we are likely to get many pairings. At most, one of the pairings will be correct. More matching information is needed. If we additionally use an address such as 123

East Main Street, Anytown, California, then we may be able to find a unique pairing. Table 3 provides examples of common variations in names that make computer comparisons more difficult.

Table 3. Errors and Inconsistencies in Names

Mr. John K. Smith	
J. K. Smith	
Margaret Helen Jones	
Peggy Jones	(nickname)
Mrs. H. Jones	
Mrs. Robert Jones	(husband name)
Margaret Helen Brown	(maiden name)
Kim Cheung	(family name first)
Cheung Kim	(family name last- English convention)
Juan Garcia-Martinez	(two last names)
Won Garsia-Marteenez	(moderate typographical variation)

Name standardization software

The name standardization is intended to separate the components of a free-form name into sub-components that can be more easily compared. The first subroutine replaces various commonly occurring words such as DOCTOR with a consistent spelling such as DR. It replaces words such as FARM with the common spelling FRM. Since these changes are based on lookup tables, it is easy for the user of the software to make modifications to the table. The second subroutine breaks up (parses) the entire name into a set of fixed sub-components that are in fixed locations. The fixed sub-components can then be more easily compared. Table 4 illustrates name standardization.

Table 4. Examples of Name Parsing

	Standardized
1.	DR John J Smith MD
2.	Smith DRY FRM
3.	Smith & Son ENTP

Parsed

	PRE	FIRST	MID	LAST	POST1	POST2	BUS1	BUS2
1.	DR	John	J	Smith	MD			
2.				Smith			DRY	FRM
3.				Smith		Son	ENTP	

Address standardization software

Address standardization software is intended to break a free-form address into components that are more easily compared. Like name standardization there is an initial subroutine that replaces various spellings of words like ROAD with a common abbreviation RD and common direction words such as EAST with a common abbreviation E. Table 5 provides examples of address standardization.

Table 5. Examples of Address Parsing

	Standardized
1.	16 W Main ST APT 16
2.	RR 2 BX 215
3.	Fuller BLDG SUITE 405
4.	14588 HWY 16 W

	Parsed									
	Pre2	Hsnm	Stnm	RR	Box	Post1	Post2	Unit1	Unit2	Bldg
1.	W	16	Main			ST		16		
2.				2	215					
3.									405	Fuller
4.		14588	HWY 16				W			

Matching Methodology

In the product space **A** × **B** of files A and B, a *match* is a pair that represents the same business entity and a *nonmatch* is a pair that represents two different entities. With a single list, a *duplicate* is a record that represents the same business entity as another record in the same list. Rather than regard all pairs in **A** × **B**, it may be necessary to consider only those pairs that agree on certain identifiers or *blocking criteria*. Blocking criteria are sometimes also called pockets or sort keys. For instance, instead of making detailed comparisons of all 90 billion pairs from two lists of 300,000 records representing all businesses in a State of the U.S., it may be sufficient to consider the set of 30 million pairs that agree on U.S. Postal ZIP code. *Missed matches* are those false nonmatches that do not agree on a set of blocking criteria.

A record linkage decision rule is a rule that designates a pair either as a link, a possible link, or a nonlink. *Possible links* are those pairs for which identifying information is not sufficient to determine whether a pair is a match or a nonmatch. Typically, clerks review possible links and decide their match status. The record linkage software uses the formal mathematical model of Fellegi and Sunter (1969). The decision rules in the Fellegi-Sunter model are optimal in the sense that, given fixed upper bounds on the rate of false matches and false nonmatches, the decision rules minimize the size of the clerical review region.

For accurate matching, it is crucial to get accurate estimates of agreement probabilities on individual fields *P(agree field i | Match)* and *P(agree field i | Nonmatch).* In many situations involving files of individuals, the EM-Algorithm can estimate agreement probabilities (Winkler 1988) that yield reasonably optimal decision rules. Due to the higher rates of varying representations of names and addresses, the EM algorithm will only yield reasonably good estimates for files of businesses. In the matching decision rules, the estimated agreement probabilities are used to obtain a likelihood ratio score (or matching weight). In most situations, the matching weights are obtained by adding up the specific agreement weights associated with individual fields.

Figure 1 illustrates a real matching situation for which true matching status is known. The left curve having lower matching weights is associated with nonmatches. The right curve having higher matching weights is associated with matches. The region between the lower cutoff L and the upper cutoff U is for clerical review. Pairs above the upper cutoff are considered matches and below the lower cutoff are considered nonmatches. In many situations (including the situation of Figure 1), the method of Belin and Rubin (1995) can be used to automatically estimate false match rates. Alternative methods for estimating error rates are described in Winkler (1994) and in Larsen and Rubin (2001).

Table 6 illustrates what pairs are brought together by the matching process. True matches having severe typographical variations in the name and address will typically get a very low matching weight. These types of variations occur more frequently with lists of enterprises than for lists of individuals. For a business, one address may refer to a physical location; another to a mailing address that is different. For example, business names for the same enterprise may take the forms ‘J K Smith and Son, Inc,’ ‘John Smith,’ and ‘J K S, Inc.’

Table 6. Record pairs by decreasing matching weight

Decreasing Weight		*What Matches Look Like*
r_1		agree exactly first, last, & addr
.	↓	
.		agree almost exactly (very minor typos)
.	↓	
.		agree closely (more drastic typos, possibly disagreements on minor fields)
.		
.		
r_k Upper		
r_l	↓	
.		first name and age often missing or in error (i.e. nickname, very severe typo)
.		
.	↓	
.	↓	
r_s		
r_t Lower	↓	
.		severe errors in name and address
.	↓	
r_n		

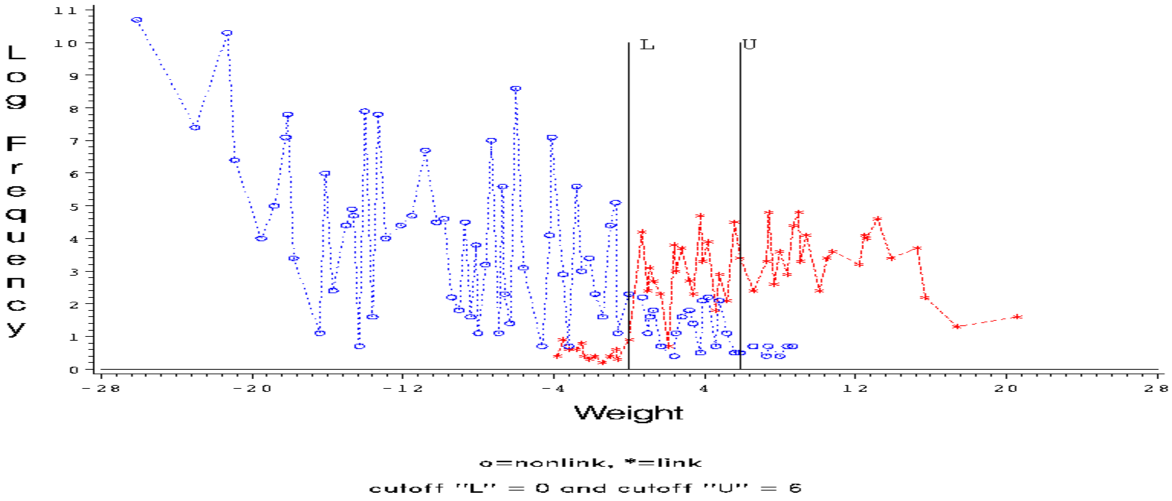

Figure 1. Log Frequency vs Weight
Links and Nonlinks Combined

Developing Skills

Record linkage is like messy-data analysis. Software can deal with general situations that have occurred repeatedly. Individuals need to recognize patterns in data. Record linkage can be straightforward to learn because it is easy to look at sets of pairs by decreasing matching weight. When unusual situations occur, then special steps may be needed. Some pairs that should be matched may have low matching weight. If the low matching weight is due to name or address standardization failure, then extra pre-processing of files may be needed. The pre-processing can be very slow because auxiliary programming to clean-up data can require substantial amounts of skill and time.

Software

U.S. Bureau of the Census Software

The standardization software breaks names and addresses into components that are easily compared. The matching software accounts for typographical error, automatically estimates matching parameters, and optimizes sets of assignments over large groups of pairs of records.

Source code and documentation are available. The software runs on all computers -- in particular IBM PCs under different versions of Windows, Unix Workstations, and VMS VAXes. Background on matching and some of the methods available in the software are described in research reports rr93/08, rr93/12, rr94/05, and rr99/04 at http:\\www.census.gov\srd\www\byyear.html.

Commercial Software

GRLS (Canlink) Statistics Canada michael.wenzowski@statcan.ca
Unix Workstation with Oracle (30 000 Canadian)
No name or address standardization, no automatic matching parameter estimation, no error rate estimation, very good documentation, free two-day training course at Statistics Canada

Integrity (http:\\www.vality.com) – formerly AutoMatch software
Most platforms (Unix workstation 195 000 USD + 15% maintenance, much more for mainframes),
Most user-friendly, good documentation, automatic matching parameter estimation, no error rate estimation

Additional software is described at http://caravel.inria.fr/~galhardas/cleaning.html and additional methods are described at http://www.niss.org/dqworkshop.html .

Data cleaning (record linkage) methods are also available for databases. Some of the current methods and extensions are described in Galhardas et. al. (2001) and in Hernandez and Stolfo (1997).

The monograph of Gill (2001) gives an extensive introduction to record linkage.

Concluding Remarks

There are now effective methods and software for matching lists. Groups undertaking matching must be aware of the large amounts of time and resources needed for developing person skills and for cleaning up lists.

This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of research and to encourage discussion. A shorter version of this paper was presented at the Exchange of Technology and Know-How ’99 in Prague, Czech Republic in October 1999.

References

Belin, T. R., and D. B. Rubin, (1995), "A Method for Calibrating False- Match Rates in Record Linkage," *Journal of the American Statistical Association*, **90**, 694-707.

Domingo-Ferrer, J., J. Mateo-Sanz, and V. Torra (2001), “Comparing SDC Methods for Microdata on the Basis of Information Loss and Disclosure Risk,” in *New Techniques and Technologies in Statistics ‘2001,* European Communities, 807-826.

Fellegi, I. P., and A. B. Sunter (1969), "A Theory for Record Linkage," *Journal of the American Statistical Association*, **64**, 1183-1210.

Friedman, N., L. Getoor, D. Koller, and A. Pfeffer (1999), Learning Probabilistic Relational Models, *Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI)*, Stockholm, Sweden, August 1999, 1300-1307.

Galhardas, H., D. Florescu, D. Shasha, E. Simon, C.-A. Saita (2001), "Declarative Data Cleaning Language, Model, and Algorithms," VLDB '2001, Rome, Italy.

Getoor, L., D. Koller, B. Taskar, and N. Friedman (2001), "Learning Probabilistic Relational Models with Uncertainty," in *Relational Data Mining* (S.Dzeroski and N. Lavrac, Eds.), Springer-Verlag, (to appear).

Gill, L. (2001), *Methods for Automatic Record Matching and Linking and their use in National Statistics,* National Statistics Methodological Series No. 25, National Statistics: London.

Hernandez, M. A., and S. J. Stolfo (1998), "Real World Data is Dirty: Data Cleansing and the Merge/Purge Problem," *Journal of Datamining and Knowledge/Discovery,* **2**, 9-37.

Koller, D. and A. Pfeffer (1998) "Probabilistic Frame-Based Systems," in Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98), 580-587.

Larsen, M. D., and D. B. Rubin (2001), "Iterative Automated Record Linkage Using Mixture Models," *Journal of the American Statistical Association*, **79**, 32-41.

Lahiri, P. and M. D. Larsen (2000), "Model-Based Analysis of Records Linked Using Mixture Models," *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 11-19.

Newcombe, H. B. (1988), *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business,* Oxford: Oxford University Press (out of print).

Newcombe, H. B., M. E. Fair, and P. Lalonde (1992), "The Use of Names for Linking Personal Records (with discussion), *Journal of the American Statistical Association*, **87**, 1193-1208.

Newcombe, H. B., J. M. Kennedy, S. J. Axford, and A. P. James (1959), "Automatic Linkage of Vital Records," *Science*, **130**, 954-959.

Scheuren, F., and W. E. Winkler (1993), "Regression analysis of data files that are computer matched," *Survey Methodology*, **19**, 39-58.

Scheuren, F., and W. E. Winkler (1997), "Regression analysis of data files that are computer matched, II," *Survey Methodology*, **23**, 157-165.

Winkler, W. E. (1988), "Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Section on Survey Research Methods*, *American Statistical Association*, 667-671 (longer version report rr00/05 available at http://www.census.gov/srd/www/byyear.html).

Winkler, W. E. (1989a), "Near Automatic Weight Computation in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Fifth Census Bureau Annual Research Conference*, 145-155.

Winkler, W. E. (1989b), "Methods for Adjusting for Lack of Independence in an Application of the Fellegi-Sunter Model of Record Linkage," *Survey Methodology*, **15**, 101-117.

Winkler, W. E. (1989c), "Frequency-based Matching in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 778-783 (longer version report rr00/06 available at http://www.census.gov/srd/www/byyear.html).

Winkler, W. E. (1990), "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 354-359.

Winkler, W. E. (1994), "Advanced Methods for Record Linkage," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 467-472 (longer version report rr94/05 available at http://www.census.gov/srd/www/byyear.html).

Winkler, W. E. (1995), "Matching and Record Linkage," in B. G. Cox *et al*. (ed.) *Business*

Survey Methods, New York: J. Wiley, 355-384.

Winkler, W. E. (1998), “Re-identification Methods for Evaluating the Confidentiality of Analytically Valid Microdata,” *Research in Official Statistics*, **1**, 87-104.

Winkler, W. E. (1999), “The State of Record Linkage and Current Research Problems,” *Statistical Society of Canada, Proceedings of the Section on Survey Methods*, 73-79 (longer version report rr94/05 available at http://www.census.gov/srd/www/byyear.html).

Winkler, W. E. (2000), “Machine Learning, Information Retrieval, and Record Linkage,” *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 20-29 (to appear at http://www.census.gov/srd/www/byyear.html).