RESEARCH REPORT SERIES
(*Statistics #2005-02*)

Approximate String Comparator Search Strategies for Very Large Administrative Lists

William E. Winkler

Statistical Research Division
U.S. Census Bureau
Washington, DC 20233

Report Issued: 03-21-05

This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the author and not necessarily those of the U.S. Census Bureau.

Approximate String Comparator Search Strategies for Very Large Administrative Lists

William E. Winkler, william.e.winkler@census.gov 1/
U.S. Bureau of the Census, Room 3000-4, Washington, DC 20233-9100

ABSTRACT

Rather than collect data from a variety of surveys, it is often more efficient to merge information from administrative lists. Matching of person files might be done using name and date-of-birth as the primary identifying information. There are obvious difficulties with entities having a commonly occurring name such as John Smith that may occur 30,000+ times (1.5 for each date-of-birth). If there are 5% typographical error in each field, then using fast character-by-character searches can miss 20% of true matches among non-commonly occurring records where name plus date-of-birth might be unique. This paper describes some existing solutions and current research directions.

Keywords: search mechanisms, approximate string comparison, computer matching

1. INTRODUCTION

Record linkage is the science of finding matches or duplicates within or across files. Matches are typically delineated using name, address, and date-of-birth information. Other identifiers such as income, education, and credit information might be used. With a pair of records, identifiers might not correspond exactly. For instance, income in one record might be compared to mortgage payment size using a crude regression function. In the computer science literature, *datacleaning or object identification* often refers to methods of finding duplicates.

If we were able to compare the corresponding fields in two records on an exact character-by-character basis, then identifying pairs of records that correspond to the same entity would be greatly simplified. For instance, if one file contains a record with name 'Zbigniew L. Zabrinsky' with date-of-birth 'January 17, 1956' and another file has a record with name 'Zabrinky, Zbigniew Laurence' and date-of birth MMDDYYYY = '01171956', then after parsing and reformatting, we could easily compare the first names 'Zbigniew,' the last names 'Zabrinsky' and the dates-of-birth '01171956' to determine that the pair is a likely or possible duplicate.

If there is typographical error, then it may be more difficult to compare 'Zbeegnief' with 'Zbigniew,' 'Zobrinksi' with 'Zabrinsky,' and '01181955' with '01171956.' Here we assume that the second set of representations is the true name and date-of-birth associated with the entity. In some situations, computer scientists have developed methods of comparing strings having typographical error. Edit distance measures the number of insertions, deletions, and substitutions to get from one string to another. The *q*-gram metric counts the number of consecutive characters of length *q* that are common across two strings. The Jaro-Winkler string comparator (see e.g., Winkler 1995) measures the similarity between strings. More advanced methods (Cohen 2003a,b), Yancey (2003), and Wei (2004) use Hidden Markov methods that adapt to different amounts of error and need training data. A number of these string comparators have been used in a variety of record linkage systems (Sarawagi et al. 2002; Bilenko and Mooney 2003a,b; Do and Rahm 2002). If typographical error is not too severe, then the string comparators can make dramatic and effective improvements in determining whether different pairs are likely to relate to the same entity (Winkler 1990; Cohen et al. 2003b).

The basic issue in this paper is how we bring together pairs in two large files when there is typographical error in individual fields that might be used in delineating a pair of records as being a *match* (referring to the same entity). This has been a subject of much research. Large survey articles (Hall and Dowling 1980; Navarro 2001) indicate the extreme difficulty of the general approximate string search problem. For this paper, we will define a *typographical error* as any error that prevents the exact character-by-character comparison of two corresponding fields in two records. In a simple situation, a last name associated with the same entity may take the forms 'Smith' and 'Smoth' where the error is in the second representation where the character 'o' is substituted for the character 'i..' In the most extreme situation, we could have name representations 'Jennifer Mary Smith' and 'Mary Jones' with date-of-birth representations '03211961' and '06151975.' The first name representation is the current correct name and the second representation is the woman's maiden name in which she uses her middle name in most situations. The second date-of-birth is completely wrong. In this type of situation, to identify correctly that two records relate to the same entity, we would need auxiliary information (possibly in a file) that associates the two name variants and that we know also has the correct date-of-birth. We could do the matching and correction of identifying information because in most situations, there would only be one individual that had the same current and maiden names.

In this paper, we are primarily concerned with the situation in which we need to search in an efficient fashion for names and other information so that we can compare two records. We will assume that any auxiliary information (possibly in an auxiliary file) can be used. In record linkage, we traditionally perform *blocking* in which we use a few characteristics such as parts of names and a geographic identifier to bring together pairs (Newcombe 1988; Gill 1999). Because there can be typographical error in all fields, we bring together sets of pairs through a set of blocking passes. With a pair of files each having 1,000,000 records, there are 10^12 pairs of records. If we perform 10 blocking passes, we may only compare 10^7-10^8 pairs of records. McCallum et al. (2000) have indicated that it is best to do an initial clustering using a computationally inexpensive comparison and then a classification with a more expensive computation. The clustering might correspond to straightforward sort/merge on blocking characteristics in record linkage. The classification might be the computation of string comparator values and associated likelihood ratios (Winkler 1990) that produce a score or weight that determines whether a pair is a match.

The obvious limitation of the methods is that we may not be able to find all pairs of records that correspond to the same entities. To estimate the number of matches (or duplicates) missed by a set of blocking criteria we apply a capture-recapture methodology suggested by Scheuren (1980) and applied by Winkler (1989) in matching business lists. Alternatively, we can guess about the amount of overlap of two lists based on experience and discussions with individuals having expertise in how the lists were created. Or we can do crude heuristic searches in which we try to guess at additional blocking criteria that may help identify more pairs that must be brought together.

The blocking problem is exceptionally difficult for two reasons. The first is that most sets of blocking criteria are found by trial-and-error based on experience. The second is that the number of pairs in successive blocking passes that are true matches can decrease at a high exponential rate. Newcombe (1988) observed that if an effective application of a set of blocking criteria was to apply, the blocking criteria in order of the proportion of pairs that were matches. Winkler (1989) observed that, as different blocking criteria were applied, the proportion of pairs that were matches in successive blocking criteria fell at a high exponential rate. This means that 40-70% of the matches ultimately found may be obtained in the first blocking pass and 1-5% of the pairs in the first blocking passes are actually matches. By the fourth blocking pass, it is possible that only 1% of the matches is actually found and that less than 0.0001% of the pairs are actually matches. In practice, it is quite typical that 10-20 blocking passes are used (Broadbent and Iwig 1999). In each individual set of blocked pairs, different classification (or decision) rules may be used to delineate pairs that might be considered to be matches.

In this paper, we provide ad hoc methods for creating sets of blocking criteria and evaluating the quality of a group of sets of blocking criteria. The quality estimate is the number (and proportion) of matches that are missed by the group. The methods assist individuals in determining additional blocking criteria that might yield additional matches. The outline of this paper is as follows. The first section consists of general background about general string comparison and search. In the second section, we provide background and examples of the kinds of typographical error that have occurred in Census files. In the third section, we give a description of the capture-recapture method for estimating the number of matches missed by a group of sets of blocking criteria. We also give a description of the empirical data files from the 2000 Decennial Census for which true matching status is known. In the fourth section, we provide results from applying the capture-recapture technology. The fifth section gives a discussion of alternative methods, limitations of the methods, and the difficulties extensions with sampling methods for estimating missed matches. The final section is concluding remarks.

2. BACKGROUND

In this section, we begin by providing an example of information in name and address files that might be used in blocking and matching. It will allow us to more precisely describe some of the issues related to general approximate string searching. Our initial descriptions are related to known characteristics of typographical error in the 1990 Census (Winkler 1994, 1995). We then apply some of the ideas to 2000 Census data.

For the 1990 Post Enumeration Survey (PES), a large sample of blocks across the U.S. were re-enumerated and matched against the main Census file. The matching was the first step in a process that allowed re-estimation of population totals through capture-recapture (see e.g., Bishop et al. 1975). Capture-recapture will be described later. Each file generally contained a geocode (block identification number for a contiguous geographic regions of approximately 70 households), first name, middle initial, last name, sex, relationship to head of household, race, house number, and street name. In some areas, a unit identifier may be given for apartment number and other locations within a fixed geographic location. Using information from the 1988 Dress Rehearsal Census, Winkler (1990) observed that more than 25% of first names and more than 15% of last names among true matches did not agree character-by-character. The true matches were delineated by two levels of clerical review and field-work and an adjudication step. Because the PES takes place within a few weeks of the Census, most individuals can still be located in the same housing unit in both lists. The truth data sets are used in evaluating the quality of matching during previous work and preparing for future matching in which true matching status is not known during production work.

PES estimation procedures of undercount and overcount only require that individuals be located in the same block or in contiguous surrounding blocks. Because only 1-2% of true matches disagree on first character of the first name, the first set of blocking criteria consists of the geocode and the first character of the last name (surname). Each set of blocking criteria is used to bring together pairs on a character-by-character basis and then all of the fields in the "blocked" pairs are used to create a matching score to delineate matches and potential matches. The matches are pairs believed to be true matches. The potential matches require clerical review that is possibly followed by field-work. The clerical review is needed when both first name and age are missing or when age is substantially in error (35 versus 44) and the first names differs significantly (Roberta versus Sissie). Most of the remaining true matches are located during the second pass. In the first pass the proportion of pairs that are matches is 1-5%. In the second pass, the proportion of pairs that are new located true matches is often less than 0.1%. In each of the approximately 500 Census-PES pairs of files there are as many as 100 million pairs (12,000 records x 15,000 records). As an example in the larger situation, only 120,000 pairs are considered in the first blocking pass and 700,000 pairs in the second blocking pass. We observe that we *miss* pairs that do not agree on geocode and neither agree on first character of last name or street name. In situations with agriculture lists that are considerably more difficult to match than person lists, Broadbent and Iwig (1999) apply ten different blocking criteria.

Table 1. Two Blocking Criteria for 1990 Census and PES Matching

Criteria	Blocking Criteria	Match Proportion Within Set of Pairs
1	geocode plus 1st character last name	1-5%
2	geocode plus 1st character street name	< 0.1%

To find a sizeable proportion of the remaining matches in the example of Table 1, we may need eight or more additional blocking criteria. We need to consider many pairs that are missed by the two blocking criteria of Table 1. The additional blocking criteria may need to consider 10 million or more pairs and can find at most 800 additional matches. The 800 matches are known to exist because they were found via steps of clerical review, field-work, and adjustication. To address the extreme number of pairs, we may need a hierarchical approach that is more efficient than the straightforward blocking plus match approach described above. If we knew the characteristics of the missed matches, then we might be able to determine additional blocking criteria. Generally, we do not have a priori information about the missed matches. This suggests a more hierarchical approach in which we block on an additional criteria, do exact character-by-character comparison on a few other fields, and finally perform the most expensive (generally string comparator) comparisons to be tentative matching scores. We make two observations. We may need a fairly sophisticated set of comparisons at the second stage with exact character-by-character and other comparisons. We will not find all pairs because some of the pairs (as noted in the introduction) will disagree on all of the fields on which we need to compare them.

The second observation suggests that we need to use additional information from auxiliary sources or from better use of the information the existing source. This would help with matching in the main two files. For instance, if both name and address are almost completely different for a pair of records that is a true match, then the actual match status cannot be determined by the information in the pair alone. To determine the true match status, we may need to use information from other individuals in the household to determine whether a pair is a match. Table 2 provides two pairs that are true matches. In the first situation 'Laura J Smith' lives in a household with a 'Robert M Smith' who is likely to be her husband and two children (not listed). The information alone and with minimal additional corroborating information may be sufficient to determine match status. In the second situation, an auxiliary file provides a married name 'Jane Smith,' a maiden name 'Laura Jane Janeway' that allows linking to 'L J Janeway' who has recently gone back to using her maiden name. Because entry 2a is from a file that is 3-4 years old, the ages need to be updated (increased) by 3-4 years so that they can be compared with entry 2b. We observe that the auxiliary information in the first situation is from some in the same household. In the second situation, it is from additional information (maiden name) in the existing list or in a list that can be easily linked (via a unique, verified identifier) with the first list.

Table 2. Pairs that are True Matches and Refer to the Same Entity

	Original Person Name	Age	Status	Other Individual Name	Age/ Status
1a.	Laura J Smith	44	bbb	Robert M Smith	48
1b.	bbb Smith	bb	spouse	Head of House	
2a.	Jane Smith	44		Laura Jane Janeway	44
2b.	L J Janeway	48		Maiden name	

With very high quality person files, we may only miss a small proportion (0.5-2%) of true matches via a set of blocking passes. With lower quality matching information in person files (Table 2), we may miss a significant proportion (10%) of matches after a moderate number of blocking passes. With business lists, we may lose a very high proportion of matches (60%+).

3. METHODS AND DATA

In this section, we cover the method and the 2000 Decennial data used in the analyses. A capture-recapture methodology is used for crudely estimating the number of matches missed by a set of blocking criteria. The ideas were introduced by Scheuren (1980) and implemented with a set of business lists by Winkler (1989). If there are *n* sets of blocking criteria, then the contingency table needed for the capture-recapture model will have 2^n cells with the 2^n cell having count 0. The inaccuracy is due to correlation bias in a set of recaptures via blocking criteria. At present, however, there are no alternative technologies for estimating the number of missed matches.

3.1. Capture-Recapture Estimation of Missed Matches

We describe two situations. The first describes a method to get approximate the number of pairs captured and not captured with different subsets within a group of blocking criteria when we do not have truth data. With this method, we need to get approximate match status flags that we can add to pairs at each blocking pass.

The methods are basically those applied by Yancey (2002) or Elfekey et al. (2002). All pairs above a cutoff are designated as matches and we estimate a proportion of matches for the set of pairs below the cutoff. We designate some of the pairs in the overlap of the sets below the different cutoffs are actual matches. This allows us to get crude estimates of the number of matched pairs captured by one set of blocking criteria and not another and vice-versa. With a single set of blocking criteria, we know (Winkler 1994, 1995) that, in many situations with person files, the estimated number of matches is low but within 1% of the true number of matches.

Alternatively, when truth data are available, then we can investigate the overlaps by directly totaling the numbers needed for the capture-recapture model. The empirical results of section 4 are based on the numbers from the truth decks. An additional advantage of the truth decks is that they provide examples of matches that are missed by a group of sets of blocking criteria.

Let A_1 and A_2 be two lists. Based on prior knowledge we may be able to guess that 80% of list A_2 can be found in list A_1. The remainder of A_2 (i.e., $A_1 \setminus A_1 \cap A_2$) might not be found because some records in list A_2 may not be in A_1 or because the weakly identifying information in A_1 and A_2 on which we match might be almost completely different. Information (variables or fields) such as first name, last name, or date-of-birth is weakly identifying if it cannot be used by itself to determine whether a pair is a match. The combination of first name, last name, and date-of-birth may determine a match. Strong identifying information such as a verified Social Security Number uniquely identifies a match. If we take *n* blocking criteria $B_1, \ldots, B_n$, then we would like to create a 2^n contingency table in which all but one of the cells has a count filled in. Each of the cells corresponds to the n-vector $(a_1, \ldots, a_n)$ where each a_i is either 1 or zero depending on whether a match was obtained by a given set of blocking criteria or not.

If we began with two lists A and B where we believe that 85% of the smaller list B can be matched against list A, we can use the number representing 85% of B as the upper bound for the estimate in the number of matches obtained by a group of sets of blocking criteria. If we believe the earlier estimate of the number of missed matches is much too low, then we need to investigate further. The further investigation may involve looking for additional blocking criteria. We note that the if A and B represent two lists of businesses from approximately the same time period, then we may not be able to match 40% or more of the records because all weakly identifying information such as names and addresses differ completely. In those situations, the only way to improve matching is to make use of auxiliary information from additional files.

3.2. Data Files

The basic file is the main 2000 Decennial Census file with approximately 300 million records. Each record in the file contains first name, middle initial, last name, address, census block id, date-of-birth, age, sex, relationship to head of household, tenure, race, and respondent phone number. There is a household identifier that identifies different individuals that are in the same housing unit. An additional file is the main ACE file of approximately 750,000 individuals representing a complete enumeration of a large sample of blocks. The ACE is the direct analog of the 1990 PES file except that the sample size is two times larger in 2000. The Census file was data-captured by a scanning process that converted hand-written information to electronic form. The ACE file was capture via a CAPI interview. In 2000, the error in census keyed data was 4.5% and, in OCR data, it was 1.1%.

The ACE is matched against the Census file by blocks in order to determine overlap of files. The overlap is in turn used to estimate undercount and overcount in the Decennial Census files. Because the true match status of each ACE record is known, some of the probabilities of P(M $\cap$ B) based on the truth where B is a set of blocking criteria. The probabilities can also be estimated for different combination of captures or non-captures of sets of blocking criteria. The probabilities (or equivalently numbers) provide the contingency table that is used in obtaining estimates of the number of matched pairs missed by a group of sets of blocking criteria.

4. RESULTS

In this section, we provide results from applying different sets of blocking criteria. During the initial phase, we investigated eleven blocking sets of criteria. There were 606,411 true matches identified from matching the ACE against the Census.

In Table 3, we can observe that most matches are obtained with all of the sets of blocking criteria except for criteria 7. With Criteria 7, we obtain the 5% of matched pairs in which first and last name are switched. As a contrast with 1990 files, matched pairs having switched first and last names represented approximately 0.5% of matches. In 1990, the keypunchers doing data capture did not correct mis-ordering. In 2000, the scanning methods were not designed to re-order first and last names. The matched pairs with switched first and last names can sometimes be brought together by blocking on part of the street address or on date-of-birth.

Table 3. Blocking Criteria and Number of Matches in the Set of Pairs

1. Zip, 1st char surname	546,648
2. 1st char surname, 1st char first name, date-of-birth	424,972
3. phone (10 digits)	461,491
4. 1st three char surname, 1st three char phone, house number	436,212
5. 1st three char first name, 1st three char ZIP, house number	485,917
6. 1st three char last name, 1st three char ZIP, 1st three char phone	471,691
7. 1st char last name = 1st char first name (2-way switch) 1st three char ZIP, 1st three char phone	31,649
8. 1st three char ZIP, day-of-birth, month-of-birth434,518	
9. ZIP, house number	514,572
10. 1st three char last name, 1st three char first name, month-of-birth	448,073
11. 1st three char last name, 1st three_char first name	522,584

With eleven blocking criteria, 1350 matches were missed. With the best four {1, 3, 11, 9}, 2766 matches were missed. With the best five {1, 3, 11, 9, 8}, 1966 matches were missed. Some of the most difficult missed matches were children in a household headed by a single or separated mother. The children were listed under two different last names, date-of-birth was missing in the ACE file, and street address was missing in the Census file. It is interesting to observe the high rate of typographical error that may, at least partially, be due to scanning error. The matching children have no 3-grams in common. Two records have a 3-gram in common if any three consecutive characters from one record can be located in another record. It is unlikely that these most difficult-to-match record pairs could be identified through any computerized procedure that uses only the information in the Census and ACE files.

Table 4. Example of missed matches (artificial data)

	Household 1		Household 2	
	First	Last	First	Last
HeadH	Julia	Smoth	Julia	Smith
Child1	Jerome	Jones	Gerone	Smlth
Child2	Shyline	Jones	Shayleene	Smith
Child3	Chrstal	Jcnes	Magret	Smith

As an additional check, we determined that only 141,846 of the matches (pairs of records) agreed exactly on exactly on first name, last name, house number, first 6 characters of street name, telephone number, and date-of-birth. This proportion (less than 25%) is well less than half the comparable proportion with 1990 Census and PES (same as ACE) files. It indicates that matching and search procedures in 2000 had to cope with more difficult typographical error than in 1990.

The best fitting of the 5-way capture-recapture models had the interactions {1-8-9-11, 3-8-9-11, 1-3-11, 1-3-8} and a χ^2 statistic of 1.9 with approximately 1 or 2 degrees of freedom. The estimated number of missed matches of 4540 represents 0.8% of the total number of matches. The lower bound on the variance (Darroch 1958) is 4690 (68^2) and the upper bound on the variance (Haberman 1974) is 170485 (413^2). The actual variance can be obtained via a complicated iterative adjustment procedure from projecting onto a series subspaces (Haberman 1974) but should be close to the upper bound given here.

No 4-way model gave a good fit. The best two had χ^2 statistics of approximately 6 and 9 with 1 or 2 degrees of freedom. In computing degrees of freedom, we use the standard method of subtracting a degree of freedom for every zero cell in the contingency table. The degree-of-freedom calculation can be in error. Haberman (1974) provided examples where degrees of freedom can be one less or one greater than the zero-cell adjustment would yield.

We observe that our estimate of 4540 might possibly be biased upward from the observed number 1966 of true matches that are identified. To check this, we used a completely different set of five sets of blocking criteria and repeated the analysis. The second estimate of missed matches was between 4500 and 5500 with comparable upper bound on variances. Surprisingly, we could still not get suitable χ^2 fits with the additional 4-way model. The additional 4-way models were chosen by trial-and-error that eliminated the zeros in the 4-way tables. We were unable to find a fifth set of criteria that did not yield zeros in the 5-way tables.

In some situations, estimates of the number of missed matches can be biased low because of the correlation between captures with two criteria (Zaslavsky and Wolfgang 1993) that can be partially corrected by using a third capture. From record linkage, we know that typographical error can be highly correlated. If a record has one typographical error in one field, it is much more likely than random to have multiple additional typographical errors in other fields. This would suggest that our capture-recapture estimates of the number of missed might be (somewhat) biased low. We can additionally observe that missed matches as given in Table 4 are exceptionally difficult to find even with very careful review and field follow-up.

5. DISCUSSION

Using sets of blocking criteria, it is intuitive that we will not find all matches within a file or across a pair of files if a proportion of matches disagrees on most of the weak identifiers such as name, address, date-of-birth because of severe typographical errors. It is also intuitive that if we could compare all pairs then we would not find all matches.

There are two alternative methods to blocking that may be able to find higher proportions of matches in some or possibly most situations. Both are considerably more difficult to implement than blocking. Both methods are intended to approximate situations where all pairs are brought together. The speed improvements are due to mechanisms for more quickly eliminating those pairs that are much more likely not to be matches. Because of their potential, we describe both. The first method provides a means of indexing information (Chaudhuri et al. 2003) from strings that can facilitate bringing together strings associated with matches more efficiently. If we were to bring to all pairs and use edit distance, then we would likely find most of the matches that agree on a number of the weak identifiers. Edit distance suffers from the difficulty that it is very computer intensive and cannot be used for bringing together pairs except in situations where all pairs are compared.

Chaudhuri et al. (2003) observed that edit distance can be approximated by *q*-gram metrics. If indexes corresponding to certain *q*-grams are created, then they can be used for searching. It is easy to create *q*-grams and use them for bringing together pairs. One of their key observations was in applying Chernoff bounds to probabilistically bound how often approximate *q*-gram indexes would be close to expected values of the pure *q*-gram indexes that, in turn, were close to the edit distances. If the relative frequency of certain *q*-grams is accounted for, then a type of frequency-based index can be created for bringing together pairs that agree on information that is relatively less frequent. If a pair has a string such as a surname that agrees with the surname of another pair on a number of *q*-grams, then it would have a greater tendency to be a match. Chaudhuri et al. (2003) provide a series of approximations that are used in creating the indexes and in bringing together pairs efficiently. With several empirical examples, they demonstrate that their method can more effectively find matches than a naïve application of edit-distance that is exceptionally slow. Because they account for the relative frequency of *q*-grams, their indexes and classification rules can more easily identify some matches that are not as easily located with edit-distance that does not account for the relative rarity of certain *q*-grams.

Jin et al. (2003) begin with a general metric distance such as edit-distance of pairs that is in R and embed the comparison of each field in a larger space R^d that can be much more easily searched. The speed improvement, in theory, reduces search time from $O(n^2)$ to $O(n)$. The first phase is an algorithm called StringMap that is analogous to the FastMap algorithm (Faloutsos et al. 1995). Each field comparison needs a separate embedding. The dimension *d* for each embedding is chosen using a sample of pairs of strings. In the embedded spaces of the form R^d, most strings that are within distance d_1 will be within distance d_2. In a review of embedding methods, Hjaltson and Samet (2003) observed that a certain unknown proportion of pairs of distances less than d_1 in the R metric would always be greater than any fixed distance d_2 in the R^d metric. As observed by Jin et al. (2003), the dimension *d* must be kept small because the amount of computation associated with the embedding grows at a rate of $O(d^2)$ times the total number of strings and a few other factors. The embedding is computationally tractable with $d = 20$ but may not be with $d = 200$. The ability to distinguish strings in the Euclidean metric of the space R^d increases as *d* increases.

Although both methods have potential, the indexing method of Chaudhuri et al. (2003) has not been applied to larger data sets with tens or hundreds of millions of records and the embedding method of Jin et al. (2003) has not be applied to files with even hundreds of thousands of records.

We also investigated possible ways of sampling within the general set of 10^{17} pairs if the 300 million record Census is matched against itself. With our group of sets of blocking criteria, we can obtain 99+% of the true matches. If we remove pairs that are matched and then consider the set of residual pairs, then only one pair in 10^{11} pairs in actually a match. Given the types of difficult-to-find pairs on Table 4, we were unable to determine a computerized search-and-compare strategy. The difficulty is that we would need to separate, say, true matches that partially agree on first name and nothing else from the overwhelming majority of truly non-matching pairs that partially agree on first name and nothing else.

The empirical results for the number of missed matches in pairs of person lists are promising because they are reasonable. In a private communication, Gill (2004) has also obtained reasonable estimates of the number of missed matches with health files of persons. We believe there are other situations where reasonable estimates could be obtained. Whether the capture-recapture methodology would give reasonable estimates with agriculture or business lists is a research problem.

6. CONCLUDING REMARKS

This paper covers methods for finding matches within and across files using weak identifiers such as name, address, date-of-birth, and other characteristics that may be subject to moderate or substantial typographical error rates. In those pairs of records for which all of the weak identifiers have substantial error, only auxiliary information from additional sources may allow delineation of true match status. In situations where a moderately high proportion of matches can be found via a group of sets of blocking criteria, we provide crude methods of estimating the number of missed matches that are not obtained by the blocking criteria.

1/ This report is released to inform interested parties of ongoing research and to encourage discussion. The views are those of the author and not necessarily those of the U.S. Census Bureau.

REFERENCES

Bilenko, M., and Mooney, R. J. (2003a), “Adaptive Duplicate Detection Using Learnable String Similarity Metrics,” *Proceedings of ACM Conference on Knowledge Discovery and Data Mi*ning, Washington, DC, August 2003, 39-48.

Bilenko, M., and Mooney, R. J. (2003b), On Evaluation and Training-Set Construction for Duplicate Detection,” *Proceedings of the ACM Workshop on Data Cleaning, Record Linkage and Object Identification,* Washington DC, August 2003.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W., (1975), *Discrete Multivariate Analysis,* Cambridge, MA: MIT Press.

Broadbent, K., and Iwig, W. (1999), “Record Linkage at NASS using AutoMatch,” FCSM Research Conference, http://www.fcsm.gov/99papers/broadbent.pdf .

Chaudhuri, S., Gamjam, K., Ganti, V., and Motwani, R. (2003), “Robust and Efficient Match for On-Line Data Cleaning,” *ACM SIGMOD* 2003, 313-324.

Cohen, W. W., Ravikumar, P., and Fienberg, S. E. (2003a), “A Comparison of String Metrics for Matching Names and Addresses,” *International Joint Conference on Artificial Intel*ligence, *Proceedings of the Workshop on Information Integration on the Web*, Acapulco, Mexico, August 2003.

Cohen, W. W., Ravikumar, P., and Fienberg, S. E. (2003b), “A Comparison of String Distance Metrics for Name-Matching Tasks,” *Proceedings of the ACM Workshop on Data Cleaning, Record Linkage and Object Identification,* Washington DC, August 2003.

Darroch, J. M. (1958), “The Multiple-Capture Census, I: Estimation of a Closed Population,” *Biometrika*, 45, 343-359.

Do, H.-H., and Rahm, E., (2002) “COMA – A system for flexible combination of schema matching approaches,” Very Large Data Bases 2002.

Elfekey, M., Vassilios, V., and Elmagarmid, A. (2002), “TAILOR: A Record Linkage Toolbox,” IEEE International Conference on Data Engineering 2002.

Faloutsos, C., and Lin, K.-I. (1995), “FastMap: A Fast Algorithm for Indexing, Data-mining and Visualization of Traditional and Multimedia Datasets,” Proceedings of the ACM SIGMOD Conference (San Jose, California), ACM: New York, 163-174.

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

Gill, L. (1999), “OX-LINK: The Oxford Medical Record Linkage System,” in *Record Linkage Techniques 1997*, Washington, DC: National Academy Press, 15-33.

Haberman, S. J. (1974), *The Analysis of Frequency Data*, Chicago: University of Chicago Press.

Hall, P. A. V., and Dowling, G. R. (1980), "Approximate String Comparison," *Association of Computing Machinery, Computing Surveys*, 12, 381-402.

Hjaltson, G., and Samet, H. (2003), “Index-Driven Similarity Search in Metric Spaces,” *ACM Transactions On Database Systems*, 28 (4), 517-580.

Jin, L., Li, C., and Mehrotra, S. (2003) Efficient Record Linkage in Large Data Sets, 8th International Conference on Database Systems for Advanced Applications (DASFAA 2003) 26 - 28 March, 2003, Kyoto, Japan, http://www.ics.uci.edu/~chenli/pub/dasfaa03.pdf .

McCallum, A., Nigam, K., and Unger, L. H. (2000, “Efficient Clustering of High-Dimensional Data Sets with Application to Reference Matching, in *Knowledge Discovery and Data Mining*, 169-178.

Navarro, G. (2001), “A Guided Tour of Approximate String Matching,” *Association of Computing Machinery Computing Surveys*, 33, 31-88.

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

Sarawagi, S., and Bhamidipaty, A. (2002), “Interactive Deduplication Using Active Learning, *Very Large Data Bases ’02*.

Scheuren, F. (1980), “Methods of Estimation for the 1973 Exact Match Study,” *Studies from Interagency Data Linkages*, (Report No. 101, U.S. Social Security Administration).

Sekar, C. C., and Deming, W. E. (1949), "On a Method of Estimating Birth and Death Rates and the Extent of Registration," *Journal of the American Statistical Association*, 44, 101-115.

Wei, J. (2004), “Markov Edit Distance,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26 (3), 311-321.

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.

Winkler, W. E. (1989), "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. (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. (1993), "Improved Decision Rules in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 274-279.

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 94/05 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. (2004), “Overview of Record Linkage and Current Research Directions,” U.S. Bureau of the Census, Statistical Research Division Report at http://www.census.gov/srd/www/byyear.html.

Yancey, W.E. (2002), “Improving EM Parameter Estimates for Record Linkage Parameters,” *Proceedings of the Section on Survey Research Methods, American Statistical Association*, CD-ROM (also at http://www.census.gov/srd/www/byyear.html).

Yancey, W.E. (2003), “An Adaptive String Comparator for Record Linkage,” *Proceedings of the Section on Survey Research Methods, American Statistical Association*, to appear (also at http://www.census.gov/srd/www/byyear.html).

Yancey, W. E. and Winkler, W. E. (2003), “BigMatch software,” computer system, documentation is in research report RRC2002/01 at http://www.census.gov/srd/www/byyear.html.

Zaslavsky, A. M., and Wolfgang, G. S. (1993), “Triple System Estimation Modeling of Census, Post Enumeration Survey, and Administration Data,” *Journal of Business and Economic Statistics*, 11, 279-288.