Comparing Matchers to Enhance Front-End Capture of Duplicate Addresses

by

Mark Moran,
Planning Research and Evaluation Division
U.S. Bureau of the Census

Presented in Toronto to JSM: August 9, 2004

1. INTRODUCTION

The Master Address File (MAF) provides an inventory of residential addresses throughout the U.S. for the Decennial Census and surveys. The Census Bureau adds new records to the MAF primarily using a quarterly U.S. Postal Service (USPS) file. In order only to add genuinely new records and not redundant addresses (duplicates), the Bureau first compares new addresses to existing MAF addresses. This matching process screens out duplicate addresses and screens in addresses which appear not to duplicate the inventory.

The purpose of this study was to enhance the safeguards with which the MAF screens new addresses. We want only those USPS refresh addresses not redundant with addresses already on the MAF. The basis for potential enhancement was to be a side-by-side comparison between the status quo matcher and a commercial probabilistic matcher called Automatch. The main task of each matcher is to capture duplicate addresses at the front-end, before becoming loaded onto the MAF.

The commercial matcher Automatch is programmed to link records between two user-provided files via the probabilistic record-linkage principles of Fellegi-Sunter.[1] The status quo matcher likewise scores probabilistically but its methods are more ad hoc, policy-driven, and tend to be rule-based.

2. METHODOLOGY

2.1 Methodology for Matcher Comparison

Sample Comprised of All City-Style Data from Five Counties. All addresses with street address and number (i.e., city-style) from five counties in the United States were drawn. Honolulu and Salt Lake counties were chosen for known unique addressing characteristics. Other counties were chosen for convenience.

Formatting the Sample Data into a Mock USPS File. The addresses began as actual administrative records from composite government databases: no records were fabricated. These addresses were used instead of USPS addresses because screening and linkage challenges posed by USPS addresses were expected to occur infrequently and perhaps fall below sample size minimums. The city-style addresses were combined and formatted like a USPS refresh file.

Staff operating the status quo matcher processed the mock refresh file through the existing matcher against a nonproduction copy of the MAF. Other staff processed the same file using a commercial matcher against an extract of the MAF.

Grouping of Record-by-Record Comparisons into Five Categories. A record-by-record comparison of corresponding DSF records output by the commercial and status quo matchers can yield five combinations (shown in Figure 1).

Focus on Links in Categories II, III, and IV. Category I cases underwent the same analysis as categories II, III, and IV except that they were omitted from clerical review. Cases on which both matchers agreed exactly (categories I and V) were taken as confirmed by virtue of mutual agreement.

Category V cases were also set aside from the clerical validation because category V contained no links. These counts were kept for comparison with other counts.

The status quo matcher uses a special strategy for the refresh of the MAF. Because the USPS file is considered a nearly complete inventory of delivery addresses, differences in apartment identifiers between the USPS file and the MAF could cause significant (and erroneous) pre-existent duplication in the MAF. As a result, the status quo matcher uses a conservative approach called "forced pairing" when matching units within multi-unit structures. The matcher first attempts to exact-match each unit identifier at the Basic Street Address (BSA) (e.g. Apt A matches Apt A). Then the matcher attempts to "equivocate" a match of the identifiers at the BSA (e.g. Apt A matches Apt 1, or Apt 2 matches Apt B, etc.). If there are any non-matched identifiers remaining at the BSA once the exact and equivocated matches finishes, then the matcher forces the leftover non-matched pairs to match (e.g. Apt A matches Basement Apt or Apt J matches Apt 206). If the USPS file has the same (or fewer) identifiers at a BSA as the MAF, then no new units will be added at that BSA as a result of the refresh because of the forced-pairing strategy even with quite unlike units.

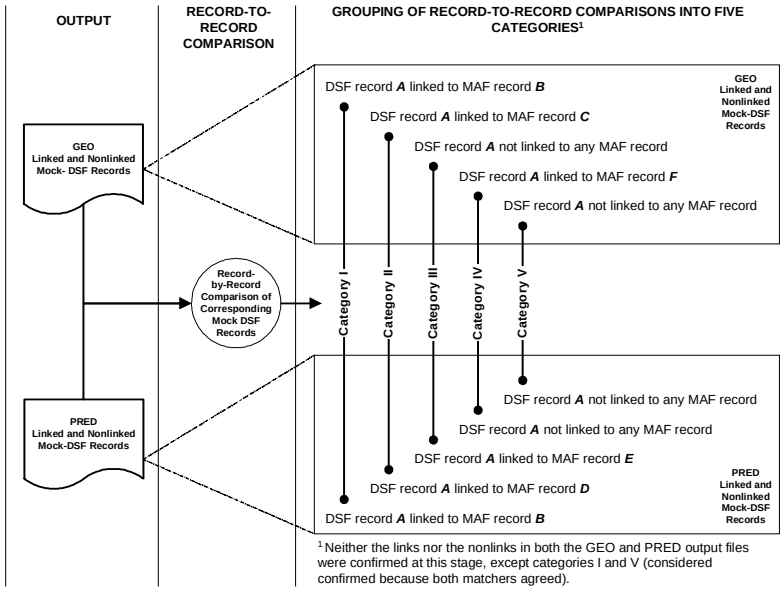

Figure 1. Record-by-Record Comparison of Corresponding DSF Records from Each Matcher's Output and Grouping of Record-to-Record Comparisons into Five Categories

To allow for this status quo feature, we had to account separately for identifiers the same across the link ("equal"), that differ across the link ("differ"), and that lack any identifier ("none"). Because of "forced pairing," the sample for clerical review was stratified, approximately Probability Proportional to Size (PPS) within county x category x stratum based on comparing identifiers. These strata were:

- cases lacking unit identifiers,
- cases whose unit identifiers were identical (e.g., Apt A and Apt A), and
- cases whose unit identifiers differed (e.g., Apt 1 and Apt A).

Clerical Review

Samples for clerical review were drawn only from the three discrepant categories (II, III, and IV). These cases were masked as to which matcher produced the link and reviewed jointly. This team confirmed, rejected, or left unresolved links. In Category II, if the team determined that both matchers linked to the same MAF record despite different indexes on each record, then they were classified as pre-existent duplicates on the MAF. In all cases, the joint clerical review team classified linkages as confirmed, as rejected, or as unresolved based on information available. Finally, analysts unmasked cases to determine the performance of each matcher.

Methodology for Regression Analysis

The purpose of our two regressions was to predict when the status quo matcher is likely to misclassify classes of addresses based on the clerical results comparing matchers. Enhancements could be based on the regression findings. One logistic regression was carried out on cases linked. Another logistic regression was carried out on cases not linked. (Table 1).

Table 1. Outcome States for Regression on Non-Linked (2 Left Columns) and Linked (2 Right Columns) Addresses

SCREEN-IN, POTENTIAL MATCHER ENHANCEMENT (FALSE MATCH)	SCREEN-OUT, NO CHANGE TO MATCHER (TRUE NONLINK)	SCREEN-OUT, POTENTIAL MATCHER ENHANCEMENT (FALSE NONLINK)	SCREEN-IN, NO CHANGE TO MATCHER (TRUE MATCH)
II Rejects II Unresolveds IV Rejects except Diffs IV Unresolveds except Diffs	I All II Confirmed GEO II Pre-existent Duplicate IV Confirmed	II Confirmed PRED III Confirmed	III Rejects III Unresolveds V All

Two Logistic Regressions

We carried out two logistic regressions, one for **cases linked by the status quo matcher**, and one for **cases not linked by the status quo matcher**. For linked cases, the predictor variables were derived from the incoming record and the MAF record. The full list of predictor variables for linked cases appears in Table 8. The response variable for status quo linked cases is an indicator of whether the linked pair should instead not have been linked (address should have been screened in).

For non-linked cases, the predictive information could only come from the incoming address. The full list of predictor variables for non-linked status quo cases appears in Table 10. The response variable for status quo non-linked cases is an indicator of whether the record should instead have been linked (i.e., address should have been screened out). We fit each of the two models on all available data.

Cross-Validating the Regressions

We carried out two series of cross-validations, one for the linked cases model and another for the nonlinked cases model. In each, we built models based on half the data and compared to the intercept-only models. We

randomly split the data into model-building and model-scoring halves twelve (12) times.

Cases Flagged as Valid and Need Extra Caution

We want to identify cases with suspicious probabilities of incorrect status quo screening where such suspicions seem reliable according to cross-validations. Therefore:

- We grouped cases into "bins" based on ranges of the continuous predictor variables.
- From the predicted probabilities for cases, we computed the weighted average probabilities for each bin.
- In a particular validation run, we judged the model validated for a bin if its prediction for the bin was 1.5 times as close to the known value as the prediction from the intercept-only model. We judged the model anti-validated for the bin if its prediction was 1.5 times as distant from the known value as the prediction from the intercept-only model.
- For linked records, we flagged bins where the probabilities of incorrect screen-out were > 0.1 and where the model for that bin was validated at least as often as anti-validated.
- For nonlinked records, we flagged bins where the probabilities of incorrect screen-in were > 0.05 and where the model for that bin was validated at least as often as anti-validated.

From predicted probabilities, we found the weighted-average probabilities across records within the bin.

3. RESULTS

3.1 Matcher Comparison Results

The results show very close agreement between the two matchers on screen-outs: Of all confirmed links by either, 98% were captured by both matchers. One percent were captured by one matcher and another one percent by the other. Note that over the entire U.S., the equivalent 1% would represent millions of addresses. For screen-ins of new records which cannot be linked conclusively to the MAF, the matchers performed in the 80% or 90% range of agreement (see Tables 6 and 7).

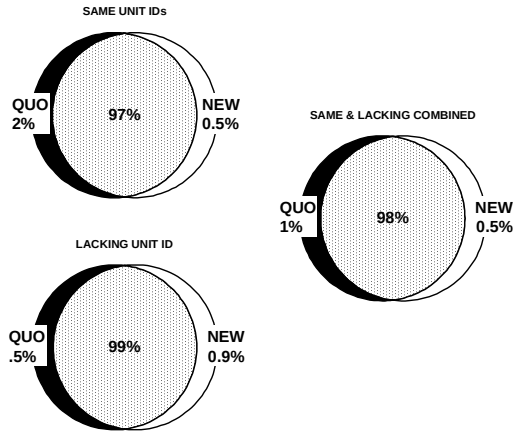

Figure 2. Differences & Commonalities Between Matchers (From Table 5)[1]

Rejected status quo links among the differing unit cases were prevalent compared to confirmed links and also prevalent compared to links of PRED's commercial probabilistic matcher among those cases. However, because of "force-pairing," such numbers were expected.

Sampling Design. In all, 911,389 cases were mutually-confirmed, Category I. In 1,422 cases, matchers provided different links (Category II). In 14,744 cases, the commercial matcher provided a link, not status quo (Category III). In 66,473 cases, status quo provided a link, but the commercial matcher did not provide a link (Category IV). Finally, 208,099 cases were considered mutually-confirmed non-links, Category V.

Clerical Review Results For analysis, the clerical review team coded the primary reasons for their judgments. One set of codes is for *rejected* links—ie., linked pairs of addresses that the team judged not to be a matched pair of addresses. The other set of codes is for *confirmed* links—i.e., links judged correctly matched. In links reviewed by the clerks, one matcher, but not the other, produced that link. The codes for confirmed links represented the team's best estimate of why one matcher failed to find a link believed correct. For each link, up to two why codes were recorded (one possible code for rejected links is "multiple faults").

[1] Numbers do not sum to 100% because subtotals represent medians between the two alternate ways to produce Table 5. The discrepancy is not due to rounding but due to the use of two calculation methods. See footnote 2 for further details on the two calculation methods

Table 2. Weighted Results of Clerical Review

CATEGORY I - BOTH NEW AND STATUS QUO PROVIDED THE SAME LINK

The 911,389 cases in this category were considered confirmed links for clerical review purposes.

CATEGORY II - NEW AND STATUS QUO PROVIDED DIFFERENT LINKS

	WEIGHTED CASES			WEIGHTED PERCENTAGES		
	CASES WITHOUT UNIT	SAME UNIT IDENTIFIER	ALL CASES UNAFFECTED BY FORCED PAIRING	CASES WITHOUT UNIT	SAME UNIT IDENTIFIER	ALL CASES UNAFFECTED BY FORCED MATCHING
Duplicate Link	59.0	18.9	77.9	86.2	23.9	52.8
Confirmed PRED	0.0	2.5	2.5	0.0	3.2	1.7
Confirmed GEO	3.2	53.6	56.8	4.7	67.8	38.5
Unresolved Link	0.0	0.0	0.0	0.0	0.0	0.0
Rejected Link	6.2	4.0	10.2	9.1	5.1	6.9
Total	68.5	79	147	100.0	100.0	100.0

CATEGORY III - NEW PROVIDED A LINK AND STATUS QUO DID NOT

	WEIGHTED CASES			WEIGHTED PERCENTAGES		
	CASES WITHOUT UNIT	SAME UNIT IDENTIFIER	ALL CASES UNAFFECTED BY FORCED PAIRING	CASES WITHOUT UNIT	SAME UNIT IDENTIFIER	ALL CASES UNAFFECTED BY FORCED MATCHING
Confirmed Link	6,052.2	1,418.0	7,470.2	52.2	65.5	54.3
Unresolved Link	0.0	0.0	0.0	0.0	0.0	0.0
Rejected Link	5,541.8	746.0	6,287.8	47.8	5.4	45.7
Total	11,594.0	2,164.0	13,758.0	100.0	100.0	100.0

CATEGORY IV - STATUS QUO PROVIDED A LINK AND NEW DID NOT

	WEIGHTED CASES			WEIGHTED PERCENTAGES		
	CASES WITHOUT UNIT	SAME UNIT IDENTIFIER	ALL CASES UNAFFECTED BY FORCED MATCHING	CASES WITHOUT UNIT	SAME UNIT IDENTIFIER	ALL CASES UNAFFECTED BY FORCED PAIRING
Confirmed Link	746.9	6411.8	7158.7	68.7	95.1	91.4
Unresolved Link	33.5	64.1	97.6	3.1	1.0	1.2
Rejected Link	306.1	268.3	574.4	28.2	4.0	7.3
Total	1,086	6,744	7,831	100.0	100.0	100.0

CATEGORY V - NEITHER NEW NOR STATUS QUO PROVIDED A LINK

The 208,099 cases in this category were considered confirmed non-links for clerical review purposes.

The new matcher did not compare House Number Suffix with Unit Identifier. Many of status quo's "Same Unit Identifier" links were absent from new's equivalent links because of this difference in strategies. Those cases (unaffected) in Category II and Category IV appear in parentheses.

Analysis of Rejected Links from the Status Quo Matcher. Categories II and IV contain the links indicated by the status quo matcher that the clerical review team examined. Here, we examine the links from the status quo matcher that were rejected. The team judged these to be false matches. These cases include the Category II cases judged "Confirmed PRED" and "Reject," and the Category IV cases judged "Reject." There were 162 Category II cases in the combined Equal Unit ID and No Unit ID groups, 80 of which were in the clerical review sample. Two of those (a weighted 2.5) were judged "Confirmed PRED," and six (a weighted 10.2) were judged "Reject." There were 8,494 Category IV cases in the combined Equal Unit ID and No Unit ID groups, of which 140 were in the clerical review sample. The clerical review team rejected 11 (a weighted 574) of the 140. Table 4.11 below contains the frequencies of why codes for the rejected status quo links, as a percentage of all weighted records.

Table 3. Why GEO Links Were Rejected (Omitting the Different Unit ID Cases)

	WHY CODE							
	DD	DN	DT	DZ	MD	MF	MU	TOTAL WEIGHTED
CATEGORY II	0	1 7.8%	1.5 11.8%	9.2 72.5%	10.2 80.4%	0	0	12.7*
CATEGORY IV	31.8 5.5%	57.4 0.9%	38.7 2.7%	93.6 16.3%	276.8 48.2%	57.4 10%	57.4 10%	574.4*
TOTAL	31.8 5.4%	58.4 9.9%	40.2 6.8%	102.8 17.5%	287.0 48.9%	57.4 9.8%	57.4 9.8%	587.1

***** The entries sum to more than the total because one link could have two why codes.

The most common reason for rejecting a status quo link was missing directional. In many of these (9 of the unweighted 11 Category IV cases), the incoming record had only one direction—a prefix direction—while the linked MAF address had both a prefix and a suffix direction: 001 E ANYSTREET vs. 001 E ANYSTREET S.

A few cases had different ZIP Codes, usually combined with other why codes. Two (unweighted) cases had different street names: in one case, the word 'OLD' was in one but not the other name; in the other, one street name contained a direction missing from the other.

Analysis of Confirmed Links Indicated by the Commercial Matcher, Not Indicated by the Status Quo Matcher. Category III contains cases where the commercial matcher linked the incoming address to a MAF address, and the status quo matcher did not link the incoming address to any MAF address. Where the link was confirmed, the clerical review team has judged that the status quo matcher should have linked (i.e., false nonmatch), and thus would have added an address to the MAF that duplicates an address already on the MAF. There were 13,758 Category III case in the combined Equal Unit ID and No Unit ID groups, of which 165 were in the clerical review team sample. Of the 165 cases, 91 cases (a weighted 7,470) were confirmed. Table 5 below contains frequencies why confirmed.

Table 4. Why Automatch Links Were Confirmed (Percentages of All Weighted Records, Omitting the Different Unit ID Cases)

WHY CODE	BOC	DTB	IH	IN	IT	MTB	P	RU	US	X	Z1T	TOTAL WEIGHTED LINKS
WEIGHTED FREQUENCY	196 2.4%	75 0.9%	266 3.2%	5009 60.9%	48 0.6%	223 2.7%	916 11.1%	75 0.9%	314 3.8%	1727 29.0%	98 1.2%	8226 *

***** The entries sum to more than the total because one link could have two why codes.

The most common reason to confirm a Category III link was "Intelligible difference in street name." These were cases where, although the street names were different, the clerical review team believed that this was due to a typographical error, misspelling, abbreviation, or other understandable difference in the representations of the street name. Most commonly, one letter was dropped or changed.

There were some "exact" matches, where the clerks could not determine why the status quo matcher had failed to link. There were a few cases where the commercial matching process treated an incoming house number bearing an embedded letter as two numbers, with the letter as a delimiter. For example, a house number such as 123D456 was treated as 123–456. The clerks judged these commercial matches correct.

Findings on the Extent of the Opportunity to Enhance the Status Quo Matcher. In the following Table, records are considered "screened out" from the the refresh of the MAF when they are linked to already existent records on the MAF. Records are "screened in" to the refresh of the MAF when they are not linked to already existent records on the MAF. Based on the three clerical review subsamples (i.e., weighted to reflect the universe of cases), the opportunities to endorse existing aspects of the matcher or to enhance the matcher are as shown in Table 6.

In Table 6, 98% of the records screened out were correctly screened out both by the commercial probabilistic matcher and by the status quo matcher. Another 8/10ths of a percent were screened out only by the status quo matcher, and another 8/10ths only screened out by the commercial matcher. If we stratify the same results into links such that the unit identifiers are the same on both ends of the link, then 97% of the records of this type were correctly screened out both by the status quo matcher and by the commercial matcher. In this stratum, the status quo matcher also correctly linked another percent or two (depending upon whether unresolved cases are discarded, i.e. the 2.5% figure, or unresolved cases are projected in proportion to the resolved cases, i.e. the 0.55% figure in square brackets). In the lacking unit identifier stratum, the results are similar although in this case the commercial matcher had more extra cases to add. The unit identifiers different stratum is completely dominated by forced pairing, as expected.

Table 5. Records Screened-Out Because They Are True Matches, Duplicating Existing MAF[2]

	SCREENED OUT BY BOTH MATCHERS	SCREENED OUT ONLY BY ONE MATCHER, CONFIRMED		MATCHER'S TOTAL
UNIT ID'S DIFFER (FORCED PAIRING)	37.9 % [39.1]	GEO	35.3% [19.0]	73.2% [58.1]
		Probabilistic	26.7% [26.2]	64.6% [65.3]
SAMEUNIT ID'S	97.0% [97.1]	GEO	2.5% [0.55]	99.5% [97.7]
		Probabilistic	0.5% [0.5]	97.5% [97.6]
LACKING UNIT ID	98.9% [99.1]	GEO	0.2% [0.9]	99.1% [99.9]
		Probabilistic	0.9% [0.9]	99.9% [99.9]
SAME & LACKING COMBINED	98.3% [98.4]	GEO	0.86% [0.81]	99.2% [99.2]
		Probabilistic	0.81% [0.81]	99.1% [99.2]

Table 6. Records Screened-In Because No Link Confirmed to Existing MAF[3]

	SCREENED IN BY BOTH MATCHERS	SCREENED IN ON BASIS OF ONE MATCHER AND WITHOUT MATCHER'S TOTAL CONFIRMED LINK BY EITHER MATCHER		
LINKS FAILING ALL 15 CONFIRM CODES TREATED AS IMPROPERLY LINKED (NO LINK = NO "DIFFER")	76.7%	GEO	2.3%	79.0%
		Probabilistic	20.6%	97.3%
LINKS STILL CONSIDERED A "DIFFER" CASE AND REMOVED (ONCE A LINK = ALWAYS A LINK)	96.5%	GEO	2.8%	99.3%
		Probabilistic	0.6%	97.1%

Table 7. Variable Predictors of Falsely Linked Records

MULT	Neither of two linked records, only one, or both have unit identifiers
DIR	Neither of two linked records, only one, or both have directionals (such as North or East)
TYPE	Neither of two linked records, only one, or both have street type suffix (such as Avenue or Boulevard)
SCRAB_DSF	Scrabble score for the original Mock Refresh Street Name field[4]
SCRAB_MAF	Scrabble score for the MAF Street Name[4]
STCOMP	String comparator score between the standardized Mock DSF Street Name, and the MAF Street Name. 6 = perfect agreement.
PROB	Estimated probability that the linked addresses are not a true match.

3.2 Regression Results

In the logistic regression for linked cases, the outcome variable is whether or not the status quo matcher gave a false match. Table 8 contains the predictor variables.

Linked and Validated Cases

We found 26 bins of linked/validated cases with suspicious probabilities of status quo false match. These 26 bins appear in Table 8:

In the logistic regression for non-linked records, the outcome variable is whether or not the status quo

[2] Numbers in brackets are the same calculation assuming the unresolved cases are ignored, rather than projected to be ultimately confirmed or rejected at the same rates as the resolved cases.

[3] Denominator is all Category V + everything rejected or unresolved; this table gives the same information twice depending upon the cut point used.

[4] Scrabble score is a number reflecting a sum of the infrequencies with which the letter occurs in the street name field. For example, A is counted 1, B is counted 3, J is counted 8, etc. Digits (from the street number) are counted 2 if 0 or 1 or 2, and counted 3 if 3, 4, 5, ... 8, 9. For a single character all by itself, the maximum scrabble score is 10.

matcher should have matched (i.e., false non-link). The predictor variables for non-linked cases appear in Table 9.

Non-linked and Validated Cases

We found 16 bins of non-linked/validated cases where the model predicts relatively high probabilities that the status quo matcher should have linked. A handful of these 16 bins appear in Table 10:

Table 8. Linked/Validated Cases with Relatively High PREDICTED Probability of Status Quo False Match

MULT	DIR	TYPE	SCRAB DSF	SCRAB MAF	STRING COMP	PROB	WTD #CASES
0	1	1	1112	1112	0	1.000	2
2	1	1	0008	1112	0	1.000	57
0	1	1	0910	1316	0	0.999	110
2	1	1	0008	1316	0	0.999	57
0	1	1	1112	1316	0	0.998	112
0	2	0	0910	1316	1	0.974	55

Table 9. Variable Predictor of Falsely Non-linked Records

MULT	Y: Mock Refresh record has a within structure identifier
DIR	Y: Mock Refresh record has a directional
HYPH	1: Mock Refresh record has a hyphen in the House Number
TYPE	Y: Mock Refresh has a street suffix type
SCRAB	Scrabble score for the Mock DSF Streetname field
PROB	Estimated probability that the address should have been linked to a MAF address

Table 10. Non-Linked/Validated Cases with Relatively High PREDICTED Probability of Status Quo False Non-match

MULT	DIR	HYPH	TYPE	SCRAB	PROB	WTD # CASES
N	Y	0	Y	1719	0.101	2107
N	Y	0	Y	1516	0.094	2060
N	Y	0	Y	1314	0.089	2193
N	Y	0	Y	1212	0.086	1365
N	Y	0	Y	1111	0.084	1635

Parameter Estimates

The string comparator score between the standardized street names strongly predicts correct status quo link. The string comparator scores range zero to one (six bins). Other things being equal, when a link has a string comparator score of 1 its odds of a false match is a factor of 10^{17} less than the odds if it had a string comparator score of 0

The scrabble scores for the Street Name are also important independent predictors of false match. The medians for both scrabble scores are 12, the 99th percentile for both is at 26. The parameter estimate for the Mock DSF Street Name is .8853. Thus, it is estimated that when the scrabble score for the Mock DSF Street Name increases by one, the odds of false match increase by a factor of $e^{.8853} = 2.42$, more than double. For nonlinked cases, suffix type presence best predicted a false screen-in (false nonmatch).

Evidence of Pre-Existent Duplication on the MAF

In Table 3, Category II shows 78 duplicates already existent on the MAF, on a weighted basis. The prevalence of pre-existing duplicates in the 1.12 million cases without disagreement between matchers is *unknown* and was *not examined* in this study. The 78 cases is certainly an underestimate of pre-existing duplicates. If an address occurs twice, we would only call it a pre-existent duplicate if our matcher did not also pick the same address from the two addresses that their matcher picked. Both matchers had to pick one of the two and not the same one of the two. We would expect the 78 pre-existing duplicates out of 82,639 to represent at most half the real rate. Such a rate across all 1.2 million records would yield at least 2,300 pre-existing duplicates in the five counties.

Table 11. Odds Ratio Estimates for Linked Cases

EFFECT	LEVELS	POINT ESTIMATE
MULT=2 vs MULT=0	2 vs 0	2.834
DIR=1 vs DIR = 0	1 vs 0	23.079
DIR=2 vs DIR = 0	2 vs 0	65.154
TYPE = 1 vs TYPE = 0	1 vs 0	4.456
TYPE = 2 vs TYPE = 0	2 vs 0	0.884
STCOMP		<0.001

Table 12. Odds Ratio Estimates for Non-linked Cases

EFFECT		POINT ESTIMATE
Directional Present	Yes or No	1.621

3.3 Limitations

The mock refresh file may differ from actual USPS administrative records. The actual administrative records file in some cases are a more complete inventory. Estimates of incorrect screenings for the nation based on these results might be inflated.

4. CONCLUSIONS

Out of 82,639 cases of matcher discrepancy (categories II, III, and IV combined), our study identified 8,044 cases which were either false status quo matches unaffected by forced pairing (574) or were confirmed false status quo non-matches that the commercial matcher captured (7,470).

The 574 false matches tended to be cases where either a directional was missing or ZIP Codes differed. From the regression we learned that the strongest predictors of false match were the string comparator and the (modified) scrabble™ score. We also learned that the best predictors of a false nonmatch were the suffix type and the (modified) scrabble™ score.

The main benefits of reducing duplicates on the MAF are (1) the costs-savings in the field and (2) improvements in the accuracy of published statistics. GAO-02-31, *2000 CENSUS Significant Increase in Cost Per Housing Unit Compared to 1990 Census* estimates the cost per housing unit for "field data collection and support systems" at $32.43 (see p. 8, Table 2 in GAO-02-31).

The 7,470 confirmed cases mentioned in Table 2 are almost certainly duplicates; therefore, existing processes would eventually flag many or all 7,470 for field resolution. These 7,470 cases were screened onto the MAF, but our study shows they should have been screened out at the front-end. Capturing these 7,470 cases at the front-end could potentially save the Government as much as $242,000 from the five researched counties.

There were about 1.05 million housing units in the five counties of this study. The total U.S. housing units for the same year was about 115,905,000. If the same rate of missed duplicates held for the entire set of all 3,141 U.S. counties, then $242,000 * (115,905,000 / 1,005,000)≈ $27,900,000 (rounded *down* at the thousands).

To be conservative, if the corrections are limited to the bins identified in the regression, we would expect 923 false matches to be identified from the regression bins (totalling all qualified bins weighted by appropriate probabilities). In the USPS refresh file, even with a misclassification occurring at half the rate found here, the dollar savings U.S.-wide for the Census Bureau from implementing the regression-based corrections would be between $1.7 million and $5.0 million.

REFERENCES

1. *A Theory for Record Linkage*, P. Fellegi and A. B. Sunter. A theory for record linkage. J. Amer. Statist. Assoc., 64:1183--1210, 1969
2. Moran, Mark, *Probabilistic MAF-Refresh Comparison: Project Specifications*, U.S. Bureau of the Census, Washington, D.C., 2003.
3. United States General Accounting Office, *GAO-02-31 2000 Census Significant Increase in Cost Per Housing Unit Compared to 1990 Census*, United States General Accounting Office, Washington, D.C., 2001.