CARRA Working Paper Series

Working Paper *#2015-04*

Matching Addresses between Household Surveys and Commercial Data

Quentin Brummet
U.S. Census Bureau

Center for Administrative Records Research and Applications
U.S. Census Bureau
Washington, D.C. 20233

Paper Issued: July 6, 2015

Matching Addresses between Household Surveys and Commercial Data

July 1, 2015

Quentin Brummet
Center for Administrative Records Research and Applications
U.S. Census Bureau

Abstract

Matching third-party data sources to household surveys can benefit household surveys in a number of ways, but the utility of these new data sources depends critically on our ability to link units between data sets. To understand this better, this report discusses potential modifications to the existing match process that could potentially improve our matches. While many changes to the matching procedure produce marginal improvements in match rates, substantial increases in match rates can only be achieved by relaxing the definition of a successful match. In the end, the results show that the most important factor determining the success of matching procedures is the quality and composition of the data sets being matched.

1 Introduction

Matching administrative and commercial data sources to household surveys can potentially improve data quality, decrease costs, and lower respondent burden in household surveys. However, the utility of these commercial and administrative data sources depends critically on our ability to link information on specific units and buildings between administrative, commercial, and survey data sets. Therefore, improvements to matching procedures between data sets could greatly benefit existing surveys.

To that end, we investigate address matching across survey data from the American Housing Survey and commercial data. We first describe the address matching methodology of the Center for Administrative Records Research and Applications (CARRA) in the U.S. Census Bureau, then discuss potential modifications to the existing match process that lead to higher match rates between survey and commercial data. For each potential modification discussed, we show the potential improvements in match rates and examples of new matches created relative to the baseline procedure. We then discuss the potential benefits and drawbacks of these modifications and present the results of an exercise that uses new matching methodologies on a separate data set.

The results show that the most important limitation to matching is the composition of the commercial address data set. While many changes to the matching procedure produce marginal improvements in match rates, substantial increases in match rates can only be achieved by relaxing the definition of a successful match. We show that this latter change will involve relaxing either the street name or house number matching criterion, as either can lead to sizable increases in match rates. However, these increases in match rates induce a number of "false matches." This highlights the tradeoff present in any matching procedure: increasing the number of records that are successfully matched inherently means increasing the probability of false matches.

The rest of the paper is structured as follows. Section 2 provides an overview of the CARRA address matching process, while Section 3 discusses general challenges for address matching. Section 4 presents the main results of the analysis, which provide evidence on the effects of a number of potential modifications to the matching process. Section 5 then demonstrates two potential alterations to the existing match process on a different data set and Section 6 concludes by discussing implications of this research.

2 Overview of the CARRA Production Address Matching Process

The CARRA address matching process uses a probabilistic match process to link addresses between data sets. Prior to matching, the process cleans and standardizes addresses in two ways. First, the matching routine uses SAS DataFlux[1] to correct misspelled street names and update zip codes. After this, addresses are parsed using a standardizer from the U.S. Census Bureau's Geography Division. This parsing is particularly important as it extracts information such as street prefix and suffix types. The address standardizer routine parses an address string into its component parts. It also standardizes various key fields to provide consistent elements for matching (e.g. "STRE" and "STREET" converted to "ST"). Even for addresses that are already parsed, re-standardizing both files used in matching with the same standardizer assures consistency between fields and facilitates the match process. In general, the matching process uses the following fields: house number, house number suffix, street prefix, directional prefix, street name, street suffix, directional suffix, apartment number/description, and five-digit zip code. Table 1 below provides an example of these address components for a fictitious address.[2]

[1] Prior to 2012, Pitney Bowes Code1 was used to preprocess addresses.
[2] In addition to the address components presented in Table 1, some addresses include a "Street Prefix Type," which identifies the type of street (e.g., "DR," "PARKWAY," etc.), but is placed before the main street name.

Table 1. Example of Address Components

Example Address: 123 ½ E MAIN ST NW APT 1, ANYTOWN, XX 12345

Address Component Name	Address Component Value
House Number	123
House Number Suffix	½
Street Directional Prefix	E
Street Name	MAIN
Street Type Suffix	ST
Street Directional Suffix	NW
Apartment Number	1
Zip Code	12345

The match process uses a probabilistic matching algorithm to compare information from the parsed fields. Each potential match is assigned a probability based on the similarity of the fields. The matching routine is composed of two passes.[3] The first pass blocks[4] on five-digit zip code and house number, then matches based on the similarity of the other address fields. If two addresses are determined to be a match, the AHS address is considered matched and is removed from the potential matching pool for the next pass, which is designed to match rural routes and P.O. boxes that would have addresses that look slightly different from that shown in Table 1. This second pass blocks on five-digit zip code and box/route number and then matches using the rest of the address fields.

[3] In this report, we discuss a particular address match between commercial data and the AHS. While many of the basics of address matching are similar across matches between different data sets, CARRA develops different matching routines for different files that can differ. For instance, while the address matching considered in this report consists of two passes, the production CARRA Master Address File Match process consists of four passes.

[4] Blocking refers to selecting a subset of data from which possible matches can be drawn (in this case, only potential observations with the same house number and three-digit zip code). Blocking divides the comparison space into manageable pieces. It is generally infeasible to compare all records from the input file to all records in the reference file. Therefore, comparisons are usually restricted to records that agree on certain characteristics. Reducing the pairs of records to compare to the set that agree on certain characteristics (sort keys) is called blocking. All records that match on the blocking strategy are then compared using the match fields to determine the links. Each unique blocking strategy defines a pass.

In each pass, the algorithm computes a match score for each potential match within a block. First, street names are compared using a string comparator.[5] For all other fields, each potential match is checked to determine agreement or disagreement. If the two fields agree, a positive score is added to the total, and if they disagree, a negative score is given.[6] Certain fields are also assigned "missing penalties," that penalize any potential match if one of the candidate addresses has missing information in a given field.[7] The total match score is then computed as the sum of all of these individual scores for particular address fields. Then, for each record on the input file, the potential match with the highest match score is kept if it surpasses a cutoff score. In the case of tied match scores, all potential matches are retained.[8]

3 Challenges for Address Matching

CARRA's MAFMatch production process matches addresses from new data sources to an extract of the Census Bureau's Master Address File (MAF), a frame of all addresses in the U.S. This production process assigns the MAF identifier (MAFID) to input address records, which are then delivered to a data warehouse. There are a number of challenges associated with this matching. Previous research found that CARRA is able to match over 90 percent of 2009 AHS records to the MAF extract, but only a little over 60 percent of 2009 commercial data to the MAF extract using the same commercial data provider that is studied in the current report (Brummet 2014). Moreover, using the MAFIDs to link records between 2009 AHS and 2009 commercial data produces match rates of roughly 60 percent, and these match rates are particularly low for multi-unit structures.[9]

Because of these previous results regarding multi-unit structures, the remainder of this report will focus on a matching procedure that differs in two important ways from previous research using the 2009 AHS and commercial data sets. First, this report focuses on direct matching of AHS addresses to commercial data instead of using the MAF as an intermediate data set. This eliminates the potential that an address

[5] A string comparator is a piece of software that assigns a score to a given pair of strings that represents how similar the two strings are to one another.
[6] The score assigned for disagreement does not necessarily need to be negative. What is important is the relative difference between the score for agreement and score for disagreement. Hence, while the scale on which these scores are measured can change from match to match, the most important factor is the difference between agreement and disagreement scores.
[7] By default, these missing penalties are set to zero, but they can take on negative values. In the match procedure studied in the current report, three fields are assigned penalties for missing data.
[8] All potential matches were retained in the matching procedure described here. CARRA varies this process some depending on what data files are being matched.
[9] Brummet (2014) also documents that these match rates tend to differ across length of street name and geographic location.

was not included in the MAF and therefore could not be linked between AHS and commercial data. In addition, addresses are matched at the Basic Street Address (BSA) level. A basic street address is defined as the address without the apartment number field. Therefore, for a single-family residence the BSA is the same as the typical address, while there are many addresses, but only one BSA for a multi-unit structure.

Another obvious issue that affects any matching procedure is the prevalence of missing data. This can be especially problematic in some commercial data sources. Table 2 displays the prevalence of missing data for various address fields in the 2013 AHS and 2013 commercial data in one state. Note first that AHS addresses rarely contain missing information, while commercial data tends to have a higher prevalence of missing address components. However, the missing rates are overall relatively small for both data sets, implying that match rates between the two data sets have the potential to be relatively high.

Another point to take away from Table 2 is that most BSAs only contain a house number, street name, street name type and directional suffixes, and zip code. Given this fact, it is unsurprising that the analysis below will find that the treatment of street name and house number are the most important considerations in constructing a matching procedure.

Table 2. Prevalence of Missing BSA Information

Variable	Percent Missing in AHS	Percent Missing in Commercial Data
House Number	0.64%	6.36%
House Number Suffix	99.53%	99.69%
Street Directional Prefix	91.61%	94.00%
Street Type Prefix	99.94%	99.95%
Street Name	0.41%	0.97%
Street Type Suffix	2.67%	2.40%
Street Suffix Directional	99.20%	98.86%
Zip Code	0%	1.60%
N	7,044	2,041,861
N Basic Street Addresses	5,325	73,215

Source: 2013 AHS and 2013 Commercial Data, restricted to one state.

4 Potential Modifications to the Address Matching Process

When considering the quality of a match process, it important to consider not only the total number of records matched (i.e., "match rate"), but also the extent to which the match process produces false matches. It is impossible to understand the extent of false matches without truth data. Therefore, the following sections present examples of the "worst match" for each possible matching modification. While imperfect, these examples give a sense of the extent to which the matching criteria generates false matches. In particular, if the "worst match" examples appear to be very poor matches, the strictness of the matching criteria may need to be increased. However, "worst match" examples that appear correct may indicate that the match routine is producing few false matches. Note that these records do not actually represent real data. The names of the street addresses have been altered in order to avoid disclosure of units that are in the AHS sample. In addition, no zip codes are provided in the examples. All matching routines in this report require an exact match on 5-digit zip code.

The rest of this section further investigates different methods of matching addresses.[10] In particular, Section 4.1 discusses how adjustments to cutoff scores and match weights affect the match process, and Section 4.2 importantly considers an alternative match process using larger blocks. Section 4.3 then presents results of matches with various minor adjustments to address standardization and matching.

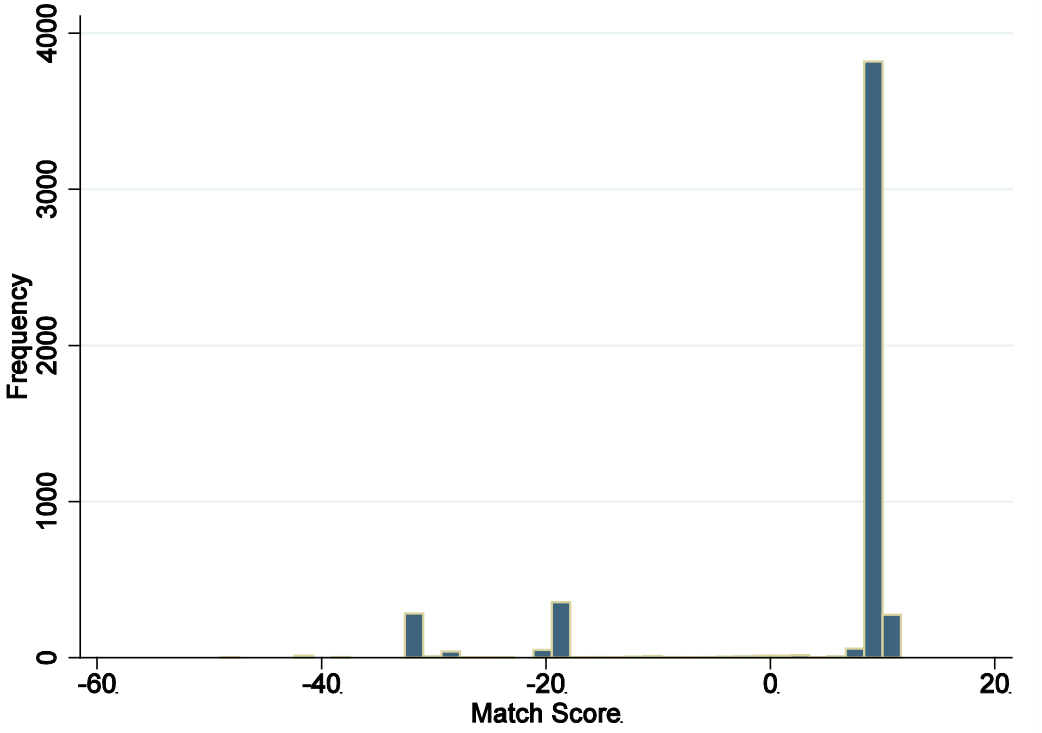
4.1 Changes to Cutoff Scores and Match Weights

As mentioned in Section 2, the current matching procedure generates scores for each pair of observations within a block, then keeps the maximum score for each unit as a potential match. Potential matches are discarded if they fall below a given cutoff score. Therefore, the decision of cutoff score can significantly alter the number of matches. To understand this further, Figure 1 presents the distribution of the maximum match scores for each AHS BSA in the first pass. As can be seen, the distribution of scores is quite lumpy. There is a large mass of high-quality matches with scores around 10, but the next masses of data fall around -20 and -30. The analysis in Table 3 shows that these latter scores represent potential matches that would likely be considered incorrect by a clerical reviewer. While there are a few potential matches with scores in between these two large spikes, they represent a relatively small number of observations and therefore do not drive the overall match rate. Note that these scores are specific to the

[10] All of these modifications will be made to the primary pass, as this is where the majority of matches are made.

particular matching procedure and data set, and changes in match weights or blocking criteria may potentially alter this distribution.

Figure 1. Distribution of Match Scores across BSAs



Source: 2013 AHS data linked to 2013 Commercial Data. N=5,325.

Based on information presented in Figure 1, the first potential modification to the match process is very straightforward: lower the match score cutoff to create more matches.[11] Table 3 presents these results. The first column displays the cutoff score, the second column the match rate for that cutoff score, and the third column presents examples of the "worst match."[12] Note that lower cutoff scores represent more lenient cutoffs that produce higher match rates.

With the most lenient cutoff of -20, it is clear that the words in the street name bear some resemblance, which caused the string comparator to award some points to the potential match. However,

[11] As mentioned previously, this applies only to the first pass in the match process. This pass is where the vast majority of matches are made.
[12] As discussed previously, no zip codes are presented because all matches require exact matches on five-digit zip code.

“ROSEWOOD” and “ELMWOOD” are most likely different streets and therefore most likely represent a poor quality match. Moving to a cutoff of -15, there is a large decrease in the match rate from 86.77 percent to 79.9 percent. This decrease corresponds to the evidence presented in Figure 1, which shows a large spike of data with scores just over -20. Moreover, the worst matches are questionable in a similar pattern to those with a cutoff score of -20, as it is unlikely that “12 WHITE OAK DR” is the same address as “12 WHITE CHERRY DR.” Therefore, -15 seems to be a particularly poor choice of cutoff score, as relative to a cutoff of -20 it produces many fewer matches with no detectable change in the quality of the “worst match.” For cutoff scores of -10 and -5, the pattern is similar and there are again only marginal decreases in the match rate with example matches that appear to be low quality.

At a cutoff of zero, the examples of “worst match” begin to become more arguable. There are four examples given, and generally they represent places where a street name might be misspelled, leading to a lower score. For example, most observers would likely think that “24 SMITH AND BLACK DR” was the same as “24 SMITH BLACK DR.” Likewise, “SEGUIN DR” and “SEQUIN DR” may be the same street. However, the example of “ELLINGTON PL” and “WELLINGTON PL” is questionable. On one hand, these may be the same address and the one data source is simply misspelled. One the other hand, “WELLINGTON” and “ELLINGTON” are reasonably common street names and these addresses could very well represent different BSAs. In either case, the match rate remains similar to the lower cutoffs, motivating the use of the stricter cutoff score of zero.

Table 3. Results of Matches at Different Cutoff Scores

Cutoff	Match Rate	Example of "Worst Match"
-20	86.72%	"824 ELMWOOD DR" = "824 ROSEWOOD DR" "234 WHITE OAK ST" = "234 WHITE MEADOW ST"
-15	79.91%	"12 WHITE OAK DR" = "12 WHITE CHERRY DR" "2 PEACHTREE ST" = "2 PINE TREE ST"
-10	79.51%	"415 PLEASANT HILL ST" = "415 PLEASANT VIEW ST"
-5	79.27%	"98 WESTARM ST" = "98 WESTMORE ST"
0	78.74%	"10 WELLINGTON PL" = "10 ELLINGTON PL" "62 EASTERLY RD" = "62 EASTERLY TER" "178 SEGUIN DR" = "178 SEQUIN DR" "24 SMITH AND BLACK DR" = "24 SMITH BLACK DR"
2	78.49%	"21 HARBOR VIEW CR" = "21 HARBORVIEW CR" "21 MAPLE D DR" = "21 MAPLE DR"
4.38*	78.14%	"21 HARBOR VIEW CR" = "21 HARBORVIEW CR" "21 MAPLE D DR" = "21 MAPLE DR"

Source: 2013 AHS data linked to 2013 Commercial Data. N=5,325.

With a cutoff score of 2, the examples of "worst match" become much clearer and would likely be considered true matches by most clerical reviewers. The cutoff score of 4.38 is denoted with a star, as it is the cutoff score currently used in the match process for the first pass. Here, we see that the match rate is 78.14 percent and the examples of "worst match" are near-perfect matches. While this match rate is lower than the higher cutoff of 2, there are only marginal differences in match rates between any cutoff score in the range of 0 to 4. Therefore, it appears to be difficult to produce substantial gains in match rates simply by lowering the cutoff score from 4.28 to two or zero.

Because the overall match score is composed of individual address field component scores, it is worth considering not just changes to the overall match cutoff scores but also changes to the way in which the individual components are scored. First, note that there is a fundamental difference in the way that street name is scored as compared to other address components. Because street names may be misspelled, street names are compared using a string comparator to attempt to account for slight misspellings. The string

comparator assigns scores on a sliding scale depending on the agreement between the street names. All other address components are either considered to "match" or "not match."[13]

Table 4 displays the results of an analysis that investigates whether these non-match penalties for individual address components (excluding house number and street name) drive the magnitude of match rates. In particular, each row represents a match that is done removing the penalty for a particular component not matching. For example, the "Street Directional Prefix" row contains the results of a match that treats "221 E MAIN ST 12345" as an identical address to "221 N MAIN ST 12345." The results in Table 4 show a compelling story – for each field, the match rate is almost identical if the penalty for non-matches is removed. This result is unsurprising given that Table 2 documented that these fields are often missing. As with all analyses, these results may change if a different data source of addresses was considered.

Table 4. Results of Matches Removing Penalty for Non-Match in a Specific Field

Variable	Match Rate
House Number Suffix	78.18%
Street Directional Prefix	78.14%
Street Prefix Type	78.14%
Street Suffix Type	78.14%
Street Suffix Directional	78.14%

Source: 2013 AHS data linked to 2013 Commercial Data. N=5,325.

Given these results, it is clear that street name and house number will have the largest effect on match rates, because Table 2 showed that these fields are contained in nearly every address on each data set. In order to investigate the important of how much weight is given to street name, Table 5 presents results that alter the amount of penalty that is given to a non-match in the street name field. Note again that street name comparisons are done using a string comparator, so there is a scale of agreement from two strings with no matching qualities at all to strings that are identical.

[13] As discussed before, missing penalties may also be assigned to particular address components. Section 4.3 considers these in greater detail.

The results in Table 5 are organized from most to least strict treatment of street name agreement. For each matching routine, all parameters except for those governing street name sensitivity are kept the same as baseline. The results show that most modest decreases in the street name mismatch penalty result in only marginal improvements in match rates. In fact, the change in penalties from -20.96 and 0 is quite substantial, but only leads to an improvement in match rates of roughly one percentage point. In addition, the examples of worst match show that many of the new matches are likely incorrect. For example, "315 EASTRIDGE ST" is probably not the same address as "315 WESTRIDGE ST." In fact, to obtain significant improvements in match rates the street name penalty has to be set so low as to allow "21 BOAT DR" to match to "21 APPLE DR," an example that is clearly a false match. Therefore, these results show that lowering the penalty for mismatched street names is a poor method of increasing match rates, because the new matches created by lowering the penalty appear to be false matches of very low quality.

Table 5. Results of Matches at Different Sensitivities to Street Name Agreement

Penalty	Match Rate	Example of "Worst Match"
-20.96*	78.14%	"21 HARBOR VIEW CR" = "21 HARBORVIEW CR" "21 MAPLE D DR" = "21 MAPLE DR"
-10	78.46%	"315 WESTRIDGE ST" = "315 EASTRIDGE ST"
-5	78.87%	"10 WORCESTER TER" = "10 DORCESTER TER"
0	79.19%	"10 MARTIN ST" = "10 MARKET ST" "413 SELLWOOD WAY" = "413 ROSEWOOD WAY"
5	87.27%	"21 BOAT DR" = "21 APPLE DR"

Source: 2013 AHS data linked to 2013 Commercial Data. N=5,325.

4.2 Relaxation of Blocking Criterion

In the current matching routine, blocks are formed based on five-digit zip code and house number. Therefore, the existing match routine requires that house number match exactly between survey and commercial data files in order to be considered a match. While this restriction is important for assuring that successful matches represent the same housing unit, it implies that any miscoding of house numbers in either data set will decrease the number of matched addresses. One way to lessen this restriction is to block on only the first character of the house number and give penalties to potential matches if the remaining characters of the house number do not match. In other words, this routine restricts every match to have the same first digit of house number, but allows the house numbers to differ thereafter. However, because matches on house number are rewarded, the routine will match "201 MAPLE ST" to "201 MAPLE ST" before matching "201 MAPLE ST" to "203 MAPLE ST."

Table 6 presents the results of matching routines based on the procedure described above, where each row represents a routine that keeps the same parameters as baseline, but blocks only on the first digit of house number. The results show small improvements in match rates for most reasonable values of the cutoff score. Relative to the baseline in Table 3, a cutoff score of 4.38 produces only a 0.26 percentage point improvement in match rates. In addition, moving to a more relaxed cutoff of 2, there is a 0.39 percentage point improvement over the results in Table 3, and the example worst matches still show matches that are potentially correct. However, with cutoff scores of -2 or less, there is an increase in match rates, and the matching routine is now assigning matches to roughly 85 percent of AHS addresses.

At the very least, this appears to be a better method of achieving substantial increases in match rates than simply relaxing the matching criteria for street name. To get in the range of 85 percent match rates simply by relaxing street name matches, one would need to treat "BOAT DR" the same as "APPLE DR." While "120 RED RIDGE RD" and "121 RED RIDGE RD" are certainly different housing units, they are arguable a better potential match than "21 BOAT DR" and "21 APPLE DR."

Table 6. Results of Matches Blocking Only on First Character of House Number

Cutoff	Match Rate	Example of "Worst Match"
-8	85.67%	"2452 WHITE WATER DR" = "2487 WHITE WATER DR"
-6	85.31%	"7818 ST JOSEPH DR" = "7818 ST JOHN DR" "1901 OLD WINCHESTER DR" = "1782 OLD WINCHESTER DR"
-4	85.24%	"201 MAPLE DR" = "202 MAPLE D" "203 MAPLE DR" = "202 MAPLE D"
-2	84.90%	"120 RED RIDGE RD" = "121 RED RIDGE RD" "135 RED RIDGE RD" = "135 RED RIDGE RD"
0	79.77%	"40 E LAKE DR" = "41 E LAKE DR"
2	78.78%	"10 WELLINGTON PL" = "10 ELLINGTON DR" "52 MAGNOLIA TER" = "52 MAGNOLIA TER"
4.38*	78.40%	"315 MAPLE ST" = "315 MAPLE PL"
6	78.18%	"21 HARBOR VIEW CR" = "21 HARBORVIEW CR" "21 MAPLE D DR" = "21 MAPLE DR"

Source: 2013 AHS data linked to 2013 Commercial Data. N=5,325.

Note that this procedure comes with one significant drawback: it leads to substantial increases in computational time. While the extent of computational burden depends on the exact make up of the data sets being matched, it is safe to say that moving from all characters of house number to only the first character of house number will substantially increase the size of each block. Given that the matching routine searches for potential matches within the set of all possible combinations of addresses between data sets within a given block, this translates to a non-trivial increase in the number of potential matches that the matching routine must consider.

Taking a step back, the above tables provide an understanding of why we fail to match 100 percent of addresses. Given that relaxing the match parameters in multiple ways creates new matches that are of questionable quality, it is clear that the commercial data file does not contain exact matches for every

address in the AHS file. For example, consider "824 APPLE ST." The results above indicate that to achieve a match rate of 85 percent or more, the analysis will be forced to use a routine that matches either "824 APPLE ST" to "824 BOAT ST" or "824 APPLE ST" to "826 APPLE ST." In other words, minor misspelling and data entry errors simply cannot account for low match rates: if "824 APPEL ST" was included in the commercial data set, it would have matched to "824 APPLE ST" instead of "824 BOAT ST" matching to "824 APPLE ST."

4.3 Editing and Missing Penalty Changes

In addition to the results presented above, the current analysis considers two other minor changes to the matching routine. First, there are a small number of addresses with "BV," "CL," or "TL" in the address that were not standardized correctly. In order to gauge the effect that this has on the matching procedure, Table 7 presents the results of baseline matching routines run on addresses with these three strings removed from all addresses. Comparing to Table 3, two facts emerge. First, the magnitude of match rates tends to be higher in Table 7, but the pattern of match rates is similar: with a cutoff at the baseline of 4.38, the match rates are 0.70 percentage points higher than those in Table 3. Moving to a cutoff of 0, the match rate is only 1.13 percentage points higher than that shown in Table 3. In addition, in order to get substantial improvements in match rates, cutoff scores must be lowered to at least -5. While there are some examples of good matches in this new group,[14] the majority of new matches are questionable (e.g., "224 STREAM ST" matching to "224 MOUNTAIN STREAM ST"). Hence, the overall message from Table 7 is that while the additional edit of removing these three troublesome strings helps to marginally improve match rates, there is still no way to substantially raise the match rate without relaxing the definition of what constitutes a match.

[14] For example, "32 APPLE ORCHARD RD" matching to "32 APPLE OCH RD."

Table 7. Results of Matches Removing BV, CL, and TL

Cutoff	Match Rate	Example of "Worst Match"
-10	85.93%	"125 ADAWIGO DR" = "125 ADEWIGO DR" "723 REDWOOD ST" = "723 ROSEWOOD ST"
-5	85.33%	"10 WELLINGTON PL" = "10 ELLINGTON DR" "32 APPLE ORCHARD RD" = "32 APPLE OCH RD" "224 STREAM ST" = "224 MOUNTAIN STREAM ST"
0	79.87%	"40 WESTRIDGE DR" = "40 WESTRIDGE DR" "105 PROSPECT ST" = "105 PROSPECT BAY ST" "213 LOYOLA NORTHWAY AVE" = "213 LOYOLA SOUTHWAY AVE"
2	78.99%	"10 WELLINGTON PL" = "10 ELLINGTON DR" "52 MAGNOLIA TER" = "52 MAGNOLIA TER"
4.38*	78.84%	"12 SWAMP HILL CT" = "12 SWAMP HILL COURT" "213 PINECONE DR" = "213 PINE COVE DR"

Source: 2013 AHS data linked to 2013 Commercial Data. N=5,325.

In addition, recall that some potential matches are penalized if one of the data sets contains missing information for a given address field. In order to understand the extent to which these missing penalties affect the match process, Table 8 presents the results of matching routines that do not use these matching penalties. The first three rows present the results from routines that remove the missing penalty for a given address component, and the final row presents results using no missing penalties. As can be seen, there is little variation across the rows in the table. In all cases, the example worst matches represent true matches, but the improvements in match rates are very small. In fact, removing all missing penalties produces a match rate that is only 0.34 percentage points higher than the baseline match rate in Table 3. As with all results, this is specific to a given data set. It very well may be that in some data sets with large number of missing observations, missing penalties play an important role in determining the success of the match process.

Table 8. Results of Matches Removing Missing Penalties

Variable with Matching Penalty Removed	Match Rate	Example of "Worst Match"
House Number Suffix	78.40%	"12 BALL HILL DR" = "12 BALL HILLDR" "500 WEST AVE" = "500 WEST AVE"
Street Directional Prefix	78.20%	"12 BALL HILL DR" = "12 BALL HILLDR" "21 HARBOR VIEW CR" = "21 HARBORVIEW CR"
Street Directional Suffix	78.16%	"12 BALL HILL DR" = "12 BALL HILLDR" "21 HARBOR VIEW CR" = "21 HARBORVIEW CR"
No Missing Penalties	78.48%	"12 BALL HILL DR" = "12 BALL HILLDR" "500 WEST AVE" = "500 WEST AVE"

Source: 2013 AHS data linked to 2013 Commercial Data. N=5,325.

5 Demonstration of Potential New Matching Procedures

In general, these previous results show that it is very difficult to generate substantial improvements in match rates while retaining high quality matches. However, these results all pertain to data from a particular state and therefore we now demonstrate the results of two potential new matching systems using data from an entirely different state. Therefore, this exercise gives us an idea of the potential generalizability of these results outside of the data that was used in Section 4.

Because most of the modifications discussed in Section 4 produce small improvements in matching outcomes, we restrict to focusing on two potential new matching systems that vary depending on the quality of matches that the analyst requires:

System 1: Remove "BV," "CL," and "TL" from all addresses and block on the first character of house number. Set the cutoff score equal to 2.

System 2: Remove "BV," "CL," and "TL" from all addresses and block on the first character of house number. Set the cutoff score equal to -2.

System 1 makes only minor changes to the baseline strategy and allows very few matches that would not be considered high quality matches by analytical review. System 2 is instead more aggressive, and is targeted to achieve higher match rates by allowing addresses such as "213 MAIN ST" to match to "215 MAIN ST." As discussed previously, if an analyst wishes to generate substantial increases in match rates, he or she must be willing to relax the matching criteria for either street name or house number. Because the results in Table 5 show that relaxing the street name criteria lead to many poor matches, we proceed by relaxing the matching criteria for house number in the case where we target high match rates.

In order to showcase the properties of these new systems, we now present the results of these new matching routines using data from a new state. This state is from another area of the country, and presents an opportunity to check the properties of these new matching procedures using data that falls outside of the sample used in the previous analysis.

Table 9 shows these results. The first row presents results of the baseline matching procedure applied to the new state's data. The match rate in the new state is 2-3 percentage points lower on baseline, which is not surprising given the noted variation in address match rates across states (Rastogi and O'Hara 2012, Brummet 2014). Moving to System 1, we see that the match rate is indeed increased by 0.73 percentage points. This is a relatively marginal increase, but comes at very little cost as the examples of worst match remain very high quality.

Importantly, the last row of Table 9 displays results from the second proposed system. Under this system, the match rate skyrockets to 90.54 percent, an increase of 14.71 percentage points. In addition, some of the new worse matches are potentially correct: "515 MAPLE NONE" is likely the same as "515 MAPLE and the "NONE" is an input error. However, many of the new matches are potentially incorrect. There are a few matches such as the "GARDWOOD ST"/"GLENWOOD ST" example that might not pass clerical review due to differing street names,[15] but the vast majority of the new matches are made by due to the relaxation of the house number match criteria (e.g., "3922 WASHINGTON ST" matching to "3986 WASHINGTON ST"). Therefore, this demonstration shows that the results presented in Section 4 hold up in the new data – while changes to address processing and match criteria can produce marginal improvements in match rates, the only way to generate substantial increases in match rates is to relax the definition of a good match.

[15] One could remove all of these false potential matches by increasing the penalty on mismatched street names. The point remains the same in either case, however, as the majority of the increase in match rates from baseline to System 2 is driven by the relaxation of the house number matching criteria.

Table 9. Matching Results in New Geographic Area

Matching System	Match Rate	Examples
Baseline	75.83%	"21 HARBOR VIEW CR" = "21 HARBORVIEW CR" "21 MAPLE D DR" = "21 MAPLE DR"
System 1	76.56%	"12 BALL HILL DR" = "12 BALL HILLDR" "21 HARBOR VIEW CR" = "21 HARBORVIEW CR"
System 2	90.54%	"515 MAPLE" = "515 MAPLE NONE" "215 GLENWOOD ST" = "215 GARDWOOD" "1924 CLEAR WATER BAY RD" = "1927 CLEAR WATER BAY RD" "3922 WASHINGTON ST" = "3986 WASHINGTON ST"

Source: 2013 AHS data linked to 2013 Commercial Data. N=3,836.

6 Conclusion

These results show that the treatment of house number and street name are the most important determinants of outcomes from matching routines. For the most part, small changes to match parameters, penalties, address standardization, and cutoff scores lead to only marginal improvements in match rates. In order to generate significant increases in match rates, the analyst must be willing to sacrifice accuracy. This can be achieved by relaxing either the street name or the house number matching criterion, but we show evidence that relaxing the house number field is likely the better method of these two methods and demonstrate with an alternate data set that relaxing the matching criteria for house number can lead to substantial increases in match rates. This comes at a cost, however, and requires allowing "221 E MAIN ST 12345" to match to "223 E MAIN ST 12345": a match that many analysts may be unwilling to accept.

As discussed earlier in this report, these results pertain only to matching between two sources of data for one state. While the results in Section 4 generalized to a new state in Section 5, there is no guarantee that these results would generalize to new sources of data. Nonetheless, these results underscore the importance of high quality survey frames and commercial data for address matching. Given the difficulties associated with address matching, it is likely that the majority of non-matches are due to one of the data sets not containing information for a given address. Future work will delve more in depth on this issue by performing similar analyses using multiple sources of addresses. In addition, the most important extension of this work would be to use "truth data" to precisely investigate the prevalence of false matches in greater detail. Understanding the true prevalence of false matches will greatly help our understanding of address matching.

References

Brummet, Quentin. 2014. "Comparison of Survey, Federal, and Commercial Address Data Quality." Center for Administrative Records Research and Applications Working Paper No. 2014-06.

Rastogi, Sonya and Amy O'Hara. 2012. "2010 Census Match Study." Center for Administrative Records Research and Applications Report, U.S. Census Bureau.