RESEARCH REPORT SERIES
(Statistics #2008-2)

Additional Results from a Nationwide Matching of 2000 Census Data

Michael Ikeda
Edward Porter

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

Report Issued: March 5, 2008

Additional Results from a Nationwide Matching of 2000 Census Data

Michael Ikeda and Edward Porter, Statistical Research Division, U.S. Census Bureau

Abstract: A nationwide unduplication procedure is being considred for the 2010 Census. One potential problem is the possibility of finding large numbers of false positives, especially when matching above the county level. To help evaluate the extent of this problem, the matching and modeling procedures are being run on the data from the 2000 Census.

This report provides an overview of the results from Within Response Modeling, which evaluates households with multiple links, and of an analysis of the resulting Residual Person links. As expected, name frequency does not seem to have much effect for links accepted in Within Response Modeling, while most of the problem with apparent false matches in the Residual Person links seems to be concentrated in the most common surnames and the most common Hispanic surnames, especially for matches outside the state. In contrast, for given names there does not appear to be a strong effect of name frequency on false matches.

Key Words: Census Unduplication, Within Response Modeling, Record Linkage

1. Introduction

One important goal for the 2010 Census is the reduction of person duplication. One possibility for reducing duplication is conducting a nationwide person unduplication operation. This operation includes a nationwide matching and modeling process to identify potential duplicates. These potential duplicates would then be resolved by a followup operation. One problem is that nationwide matching, even under very strict matching rules, is likely to find large numbers of coincidental matches (Fay 2004). Sending large numbers or high proportions of false matches to followup is likely to produce undesirable results. To evaluate the extent of the problem of false matches and develop suggestions for dealing with it, we are analyzing the results from matching and modeling procedures on the data from the 2000 Census. The data we used included persons in units deleted by the Housing Unit Duplication Operations (HUDO). As explained in Ikeda and Porter (2007), if a matching and modeling operation had been done in 2000, it would have taken place before before HUDO.

The matching and modeling process has several stages. The first stage is the Across Response Matching operation. This operation links person records across all responses, except when both persons are from Group Quarters (GQ). Results from this operation are summarized in Ikeda and Porter (2007).

For housing unit (HU) person links (links where both persons are from HUs), there is an additional matching step, the Within Response Matching, which tries to find additional person links between HUs linked by the Across Response Matching. The Within Response Modeling operation then evaluates the results of Within Response Matching in pairs of HUs with two or more person links between them. Those HU person links from Across Response matching that are *not* identified as potential duplicates in Within Response Modeling are called Residual Person links and are evaluated in the Residual Person Modeling operation.

This report presents an initial analysis of the results from performing the Within Response Modeling procedure on the 2000 Census Data and of the resulting Residual Person links. Section 2 contains an overview of methodology. Section 3 presents the results of an exploratory analysis for the links selected in Within Response Modeling and for the Residual Person links. Name frequency does not appear to play an important role in false matches from Within Response Modeling. In contrast, false matches in the Residual Person links appear to be concentrated in the most common surnames and the most common Hispanic surnames, especially for matches outside the state. Section 4 provides a summary and general discussion of the results.

2. Methodology

The starting point for Within Response Matching is the person links found in Across Response Matching. Across Response Matching matches individual persons across all Census responses, except when both persons are from GQs. For our simulation on the 2000 Census data each Census address is a response. Across Response Matching can find multiple links to a person and is performed using the BigMatch record linkage system (Yancey 2007). The blocking passes used are basically those used for the Across Response Matching in the 2006 Census Test (Lynch 2005). A "nickname file" is used to convert some common "nicknames" to their base first name (the name that they are the nickname of). In this simulation, the first name used in matching is also the output first name. Most of the links come from blocking passes where the base first name is used in matching. The BigMatch match score for each link in each pass is adjusted to account for the blocking criteria for the pass (variables that must agree before a match is attempted). The adjusted score is called the adjusted across response match score (mscore).

Within Response Matching tries to find additional person links between HUs linked during Across Response Matching. In this simulation, links with low mscores (less than 6.0 for phone number matches and within-block links, less than 7.0 for other links) are dropped before Within Response Matching. Within Response Matching uses the SRD matcher for one-to-one matching. The matcher calculates a "Within matching score" for each person link that it finds. This may not include all of the links found in Across Response Matching. The variables used to calculate the Within matching score are first name, surname, middle initial, month and day of birth, age, gender, and phone number (Lynch 2005). A person link with matching first name, last name, month and day of birth, and age, and noncontradictory middle initial will have a Within matching score in the range 1.774-1.9846. HU pairs with two or more person links between them in Within Response Matching are sent to Within Response Modeling. The average Within matching score is calculated for each HU pair sent to Within Response Modeling.

Cutoff scores are assigned for each of five geographic distance categories. We set the following cutoffs based on our examination of the output from Within Response Matching: within-block 0.8513, outside-block within-tract 1.07311, outside-tract within-county 1.205136, outside-county within-state 1.36271, outside-state 1.44709. Person links in HU pairs where the average Within matching score is above the relevant cutoff are accepted as potential duplicates.

HU person links from Across Response matching that are *not* identified as potential duplicates in Within Response Modeling are called Residual Person links and are evaluated in the Residual Person Modeling operation. If a person is linked to more than one person in the same Census address, only one link becomes a Residual Person link--the link with the highest mscore (unless one of the links was accepted in Within Response Modeling, then none of the links become Residual Person links). Some links are automatically eliminated because of missing or contradictory data. Additional information on the Within Response Modeling and Residual Person Modeling can be found in Lynch, Ikeda, and Porter (2005). An overview of how the different groups of links in this report are connected can be found in Appendix 2.

3. Results

There are two analyses in this section: the first looks at the person links accepted in Within Response Modeling, the second is based on the Residual Person links. Name frequency does not appear to be an important factor in false matches for Within Response Modeling.

The analysis of the Residual Person links focuses on person links with an mscore of at least 9.0 (maximum mscore is just over 10). Most of these links fall into one of two groups: links with maximum agreement scores on all matching variables (first name, surname, middle initial, month and day of birth, age, and gender), or links where one or both persons has a missing middle initial but all other matching variables have maximum agreement scores. Most of the rest are cases where the two persons have minor spelling differences in first name or surname. At the block and tract levels we expect that almost all of these links will be true matches. Note that age can differ by one year and still have a maximum agreement score. The matching procedure calls one of the persons in a link the "A" person and the other person the "B" person. Except where specified otherwise, the first name and surname in this analysis are the names of the "A" person. Certain common nicknames are converted to the base name for most matching purposes. The names used in this analysis are the names used for matching.

For the Residual Person links, most of the problem with apparent false matches seems to be concentrated in the most common surnames and in the most common Hispanic surnames. There does not seem to be a strong effect of name frequency for first names, although some individual first names may have problems. The results are similar to the corresponding results in Ikeda and Porter (2007), except that the problems with false matches appear to be somewhat worse.

Within-Response Modeling: Surnames

An exploratory graphical procedure was used to help identify general patterns. The percent distribution of surnames was calculated for links in the following five geographic categories:

1) Within-block links
2) Links within the tract but outside the block (tract links)
3) Links within the same county but outside the tract (county links)

4) Links within the same state but outside the county (state links)
5) Links outside the state (national links)

Links with matching phone number were excluded from the distribution for geographic categories 3-5. There are 258,841 county links, 98,116 state links, and 42,858 national links with matching phone number. For each name for a given higher geographic level, the ratio of the percentage of links with that name at that level to the percentage at the within-block level was calculated. The ratios were then plotted against the count of within-block links (anything of interest generally shows up in the national plot). Figure 1 shows the plot of national ratios (ratios of national percentage to within-block percentage) against within-block count for surnames.

Figure 1: Plot of natratio*COUNT for surnames, Person links accepted in Within Response Modeling

15,465 obs had missing values. 12,130 obs hidden. Only names with 11+ within-block links are included. Eight ratios>20 omitted.

For clarity, ratios were only plotted for names with at least eleven within-block links, and eight

surnames with a national ratio greater than 20 were removed from Figure 1. Six of these surnames seem to be related to nonresponse follow-up (NRFU) training examples (see Appendix I in Mule (2001) for a list of training examples). One of the other names is "Na" which could be an abbreviation for "Not available". Observations with missing values in Figure 1 are names which have at least eleven within-block links but no national links.

There does not seem to be much of a frequency effect for links accepted in Within Response Modeling. Apart from those surnames related to NRFU training examples, the high ratios that exist may simply be due to random fluctuation. The number of accepted within-block links is nearly 20 times the number of accepted national links, which means that even a small number of national links can produce large ratios for names with a small number of within-block links.

For an additional analysis, we divided the surnames into ten categories as given below. The categories are the same as those used for HU person links in Ikeda and Porter (2007). Name frequency information from the 2000 Census is based on tabulations by David Word (2001). The names in each category are listed in Appendix 1. Person links are included in a category if the surname of either the "A" person or the "B" person is in that category. The assignment procedure checks categories in the following order: 4, 1, 5, 2, 3, 6, 7, 8, 9, 10.

1) The 25 most common nonhispanic surnames. The abbreviation for this category is CMNH.

2) Nonhispanic surnames not included in CMNH with at least 200,000 occurrences in Census 2000. Nguyen is excluded because it is placed in category 7 below. The abbreviation for this category is CNH2.

3) Nonhispanic surnames with more than 100,000 but fewer than 200,000 occurrences in Census 2000. Kim and Tran are excluded because they are placed in category 7 below. Silva is included even though it is often Hispanic (Word and Perkins (1996) classify it as "generally" but not "heavily" Hispanic). The abbreviation for this category is CNH3.

For categories 4-6, the number of within-block HU person links (with mscore of at least 9.0) in Across Response Matching is calculated for the surnames in the top 175 positions of the Word and Perkins (1996) list of most common heavily Hispanic surnames. The tabulation is based on the surnames of the "A" person.

4) The 45 names with the most within-block links in the above tabulation are placed in this category. The abbreviation for this category is HISP.

5) Names ranked 46-100 in the number of within-block links in the above tabulation are placed in this category. The abbreviation for this category is HSP2.

6) The 75 remaining names in the above tabulation are placed in this category. The abbreviation for this category is HSP3.

7) Eight mostly Asian surnames with a national ratio of 1.0 or more in the HU person links in Across Response Matching. The abbreviation for this category is ASIA.

8) Common *first* names that appeared in the surname field with a national ratio of 1.0 or more in the HU person links in Across Response Matching. The abbreviation for this category is FIRS.

9) Four surnames that are special problem cases. Three of these names are from NRFU training examples. The fourth is Doe, which appears to show up as a substitute for unknown surname. The abbreviation for this category is REMV.

10) All other surnames. The abbreviation for this category is OTHR.

Table 1 gives a tabulation of surname category (snamecat) by geographic distance category (geocat) for the person links accepted in Within Response Modeling. Phone number matches were removed from the tabulation for the county level and above. At the block and tract levels most of the links sent to Within Response Modeling were accepted, while only a small proportion of the national links that were not phone matches were accepted (187,837 out of 8,486,735).

Table 1: geocat by snamecat, Person links accepted in Within Response Modeling

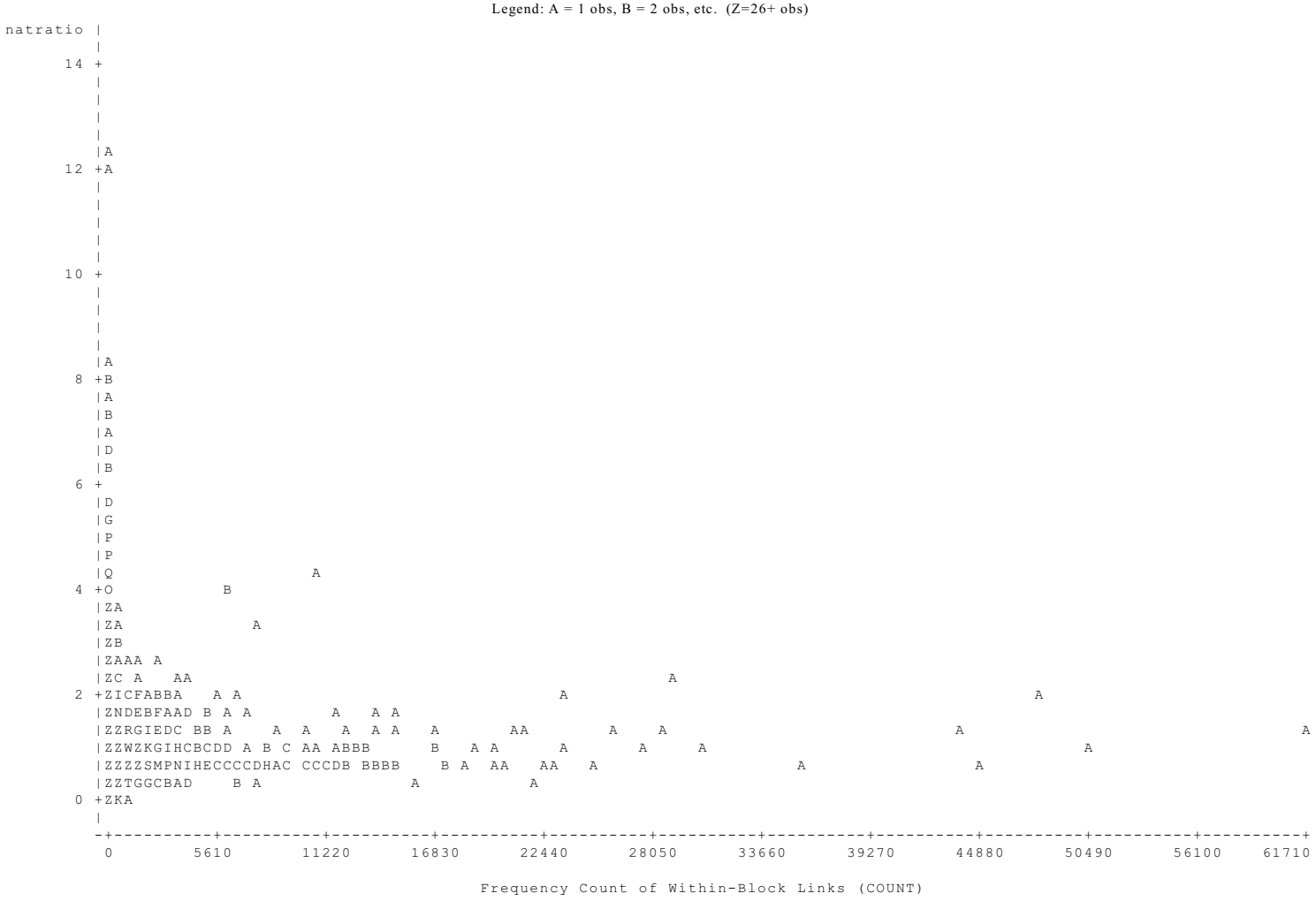
geocat	snamecat									
Frequency| Cell/OTHR| Row Pct | Col Pct	OTHR	ASIA	CMNH	CNH2	CNH3	FIRS	HISP	HSP2	HSP3	REMV	Total % Total
Block	2399902	10055	304396	196237	274800	8078	204967	66429	46019	1360	3512243
		0.00	0.13	0.08	0.11	0.00	0.09	0.03	0.02	0.00	62.09
	68.33	0.29	8.67	5.59	7.82	0.23	5.84	1.89	1.31	0.04	
	62.40	68.46	59.82	59.45	59.52	63.98	69.77	70.75	69.31	4.69	
Tract	863410	2212	122949	82573	112840	2718	46680	14789	10953	338	1259462
		0.00	0.14	0.10	0.13	0.00	0.05	0.02	0.01	0.00	22.26
	68.55	0.18	9.76	6.56	8.96	0.22	3.71	1.17	0.87	0.03	
	22.45	15.06	24.16	25.01	24.44	21.53	15.89	15.75	16.50	1.17	
County	320466	1675	48206	30214	41081	1058	31367	9631	7049	267	491014
		0.01	0.15	0.09	0.13	0.00	0.10	0.03	0.02	0.00	8.68
	65.27	0.34	9.82	6.15	8.37	0.22	6.39	1.96	1.44	0.05	
	8.33	11.40	9.47	9.15	8.90	8.38	10.68	10.26	10.62	0.92	
State	144772	529	19128	12380	18048	467	6565	1855	1433	1149	206326
		0.00	0.13	0.09	0.12	0.00	0.05	0.01	0.01	0.01	3.65
	70.17	0.26	9.27	6.00	8.75	0.23	3.18	0.90	0.69	0.56	
	3.76	3.60	3.76	3.75	3.91	3.70	2.23	1.98	2.16	3.96	
National	117301	216	14166	8690	14953	305	4187	1195	946	25878	187837
		0.00	0.12	0.07	0.13	0.00	0.04	0.01	0.01	0.22	3.32
	62.45	0.11	7.54	4.63	7.96	0.16	2.23	0.64	0.50	13.78	
	3.05	1.47	2.78	2.63	3.24	2.42	1.43	1.27	1.42	89.26	
Total	3845851	14687	508845	330094	461722	12626	293766	93899	66400	28992	5656882
% Total	67.99	0.26	9.00	5.84	8.16	0.22	5.19	1.66	1.17	0.51	100.00

The one notable point is the sharp increase in the proportion in the REMV category at the national level. This is mostly due to matching of NRFU training examples.

Within Response Modeling: First Names

The same type of ratios for the same geographic categories were calculated for first names as were previously calculated for surnames. Figure 2 shows the plot of national ratios against within-block count for first names. Phone number matches were excluded from the national distribution.

Figure 2: Plot of natratio*COUNT for first names, Person links accepted in Within Response Modeling



4,786 obs had missing values. 5,447 obs hidden. Only ratios with 11+ within-block links are plotted. Ten ratios>20 are omitted.

For clarity, ratios are only plotted for names with at least eleven within-block links, and ten first names with a national ratio greater than 20 were removed from Figure 2. Nine of these names appear to be related to NRFU training examples. The remaining name is "Na" which could be an abbreviation for "Not available". Observations with missing values in Figure 2 are first names which have at least eleven within-block links but no national links.

There does not seem to be much of a frequency effect for links accepted in Within Response Modeling. Apart from those first names related to NRFU training examples, the high ratios that exist may simply be due to random fluctuation.

Residual Person Links: Surnames

Ikeda and Porter (2007) found what appeared to be a serious problem with false matches for the most common surnames and the most common Hispanic surnames, especially at the national level for the links from Across Response matching. Since the Within Response modeling should tend to remove true matches, the resulting Residual Person Links should tend to have a higher proportion of false matches. This suggests that we may want to find special situations where we can have confidence in even the national links. One such situation is matching phone number, another such situation is multiple links between a HU pair. Table 2 gives a tabulation of surname category (snamecat) by geographic distance category (geocat) for Residual Person Links with matching phone number and adjusted across response match score (mscore) of at least 9. Table 3 gives a tabulation of snamecat by geocat for Residual Person Links (with mscore of at least 7) from HU pairs with two or more links with mscore of at least seven. Phone number matches are removed from Table 3 for links at the county level or above. The surname categories in Tables 2 and 3 are the same as used above for the links from Within Response Modeling.

Table 2: geocat by snamecat, Residual Person Links with Matching Phone Number
mscore ge 9

geocat snamecat

Frequency| Cell/OTHR| Row Pct | Col Pct	OTHR	ASIA	CMNH	CNH2	CNH3	FIRS	HISP	HSP2	HSP3	REMV	Total % Total
Block	119776	133	15339	10256	14526	332	2857	881	632	39	164771
		0.00	0.13	0.09	0.12	0.00	0.02	0.01	0.01	0.00	54.31
	72.69	0.08	9.31	6.22	8.82	0.20	1.73	0.53	0.38	0.02	
	54.42	42.49	54.30	54.17	54.08	51.00	53.38	52.66	51.47	57.35	
Tract	42258	30	6126	4070	5575	125	676	242	178	14	59294
		0.00	0.14	0.10	0.13	0.00	0.02	0.01	0.00	0.00	19.54
	71.27	0.05	10.33	6.86	9.40	0.21	1.14	0.41	0.30	0.02	
	19.20	9.58	21.69	21.50	20.76	19.20	12.63	14.47	14.50	20.59	
County	20892	85	2780	1814	2534	73	1208	343	275	8	30012
		0.00	0.13	0.09	0.12	0.00	0.06	0.02	0.01	0.00	9.89
	69.61	0.28	9.26	6.04	8.44	0.24	4.03	1.14	0.92	0.03	
	9.49	27.16	9.84	9.58	9.43	11.21	22.57	20.50	22.39	11.76	
State	22564	47	2552	1754	2514	85	477	175	112	2	30282
		0.00	0.11	0.08	0.11	0.00	0.02	0.01	0.00	0.00	9.98
	74.51	0.16	8.43	5.79	8.30	0.28	1.58	0.58	0.37	0.01	
	10.25	15.02	9.03	9.26	9.36	13.06	8.91	10.46	9.12	2.94	
National	14589	18	1451	1039	1709	36	134	32	31	5	19044
		0.00	0.10	0.07	0.12	0.00	0.01	0.00	0.00	0.00	6.28
	76.61	0.09	7.62	5.46	8.97	0.19	0.70	0.17	0.16	0.03	
	6.63	5.75	5.14	5.49	6.36	5.53	2.50	1.91	2.52	7.35	
Total	220079	313	28248	18933	26858	651	5352	1673	1228	68	303403
% Total	72.54	0.10	9.31	6.24	8.85	0.21	1.76	0.55	0.40	0.02	100.00

There is not much to say about these results. The proportion in the OTHR category remains fairly constant for all geographic levels and there is no indication that any of the remaining categories becomes more prevalent at the national level. We are therefore willing to accept these

matches as potential duplicates, with the possibility of some minor exceptions.

Table 3: geocat by snamecat, Residual Person Links from Multiple Link HU pairs
mscore ge 7

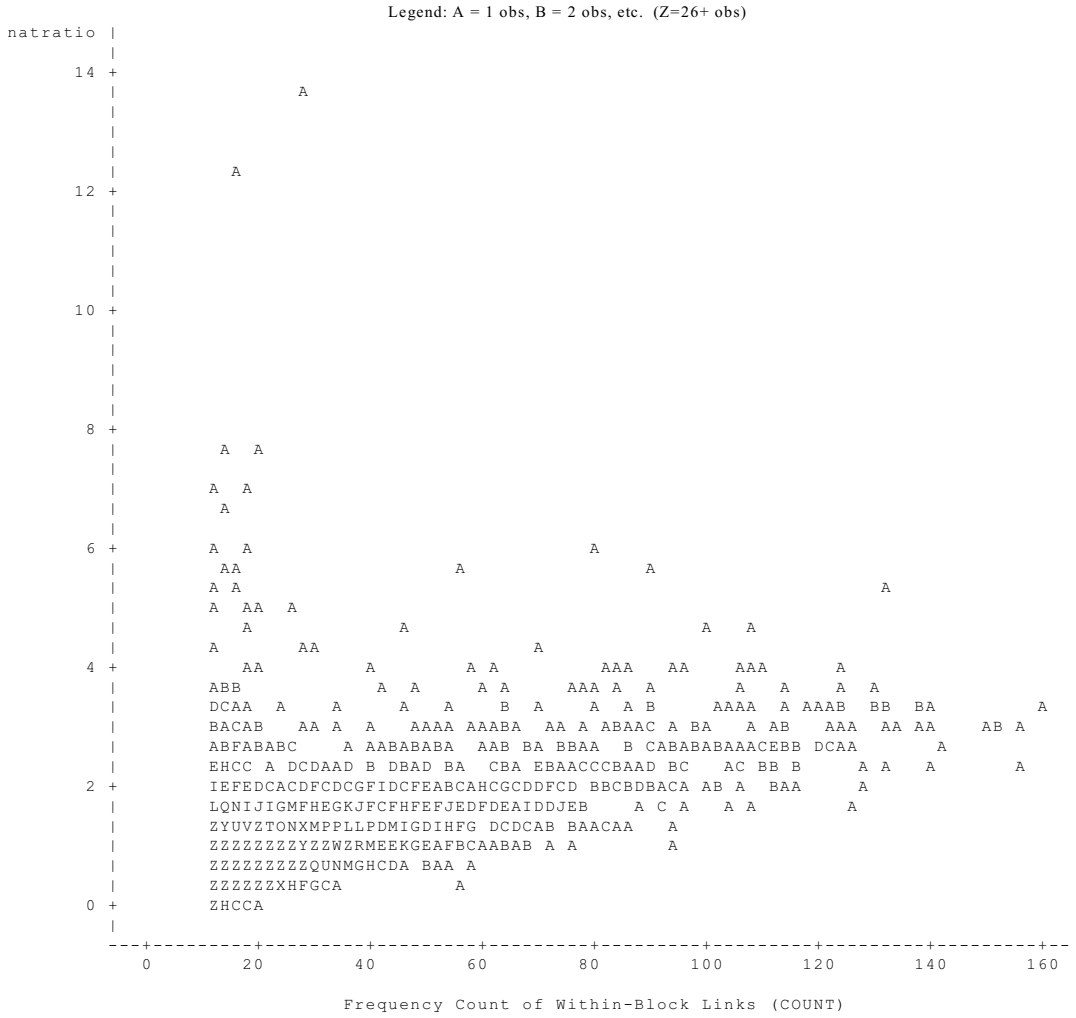
geocat	snamecat									
Frequency| Cell/OTHR| Row Pct | Col Pct	OTHR	ASIA	CMNH	CNH2	CNH3	FIRS	HISP	HSP2	HSP3	REMV	Total % Total
Block	6901	15	864	552	872	29	392	118	74	2	9819
		0.00	0.13	0.08	0.13	0.00	0.06	0.02	0.01	0.00	3.11
	70.28	0.15	8.80	5.62	8.88	0.30	3.99	1.20	0.75	0.02	
	3.22	2.36	3.20	3.22	3.44	2.75	3.95	3.88	3.18	0.01	
Tract	21759	19	2978	2160	2920	83	750	258	198	9	31134
		0.00	0.14	0.10	0.13	0.00	0.03	0.01	0.01	0.00	9.86
	69.89	0.06	9.57	6.94	9.38	0.27	2.41	0.83	0.64	0.03	
	10.16	2.99	11.03	12.59	11.51	7.87	7.55	8.48	8.52	0.06	
County	48396	247	7267	4305	5973	426	4202	1199	951	64	73030
		0.01	0.15	0.09	0.12	0.01	0.09	0.02	0.02	0.00	23.12
	66.27	0.34	9.95	5.89	8.18	0.58	5.75	1.64	1.30	0.09	
	22.59	38.84	26.92	25.09	23.55	40.38	42.29	39.40	40.90	0.42	
State	65723	209	7830	4969	7344	260	2683	789	570	627	91004
		0.00	0.12	0.08	0.11	0.00	0.04	0.01	0.01	0.01	28.81
	72.22	0.23	8.60	5.46	8.07	0.29	2.95	0.87	0.63	0.69	
	30.68	32.86	29.01	28.96	28.95	24.64	27.00	25.93	24.52	4.14	
National	71476	146	8051	5170	8257	257	1909	679	532	14439	110916
		0.00	0.11	0.07	0.12	0.00	0.03	0.01	0.01	0.20	35.11
	64.44	0.13	7.26	4.66	7.44	0.23	1.72	0.61	0.48	13.02	
	33.36	22.96	29.83	30.14	32.55	24.36	19.21	22.31	22.88	95.36	
Total	214255	636	26990	17156	25366	1055	9936	3043	2325	15141	315903
% Total	67.82	0.20	8.54	5.43	8.03	0.33	3.15	0.96	0.74	4.79	100.00

Again, the proportion in the OTHR category remains fairly constant for the all geographic levels. The REMV category does become more prevalent at the national level. In general, except for the REMV category (and perhaps other exceptions) we again would be willing to accept these matches as potential duplicates. Note that Table 3 includes links with mscore of 7 or more. Also note that any HU pair with one link in the REMV category and one link in a different category was *not* counted as having multiple links.

We now look at the remaining Residual Person Links. Similar ratios are calculated for surnames for the Residual Person links as were previously calculated for the links from Within Response Matching. Only links with an mscore of at least 9 are included in the distribution. Links with matching phone number and links from HU pairs with multiple links (two or more links with mscore of 7+) are excluded from the distribution for the county level and higher. There are 59,175 county links (52 from REMV category), 77,979 state links (503 from REMV category), and 92,949 national links (11,191 from REMV category) with mscore of at least 9 from HU pairs with multiple links 7+.

Figure 3: Plot of natratio*COUNT for surnames, Residual Person Links

Surname is in OTHR category, mscore ge 9



5 obs had missing values. 1,113 obs hidden. Only ratios with 11+ within-block links are plotted.

For clarity, ratios are only plotted for names with at least eleven within-block links. Observations with missing values in Figure 3 are names with at least eleven within-block links but no national links.

The upward slope seems to suggest some tendency for the national ratio to increase as the number of within-block links increases. We therefore added an addtional common nonhispanic name category called CNH4. CNH4 includes all surnames not already assigned to another category that have more than 60,000 occurences in Census 2000 based on tabulations by David Word (2001). The names in the CNH4 category are listed in Appendix 1. A link is included in the CNH4 category if it has not already been assigned to a category (except for the OTHR category) and the surname of either the "A" person or the "B" person is in the CNH4 category. The CMNH, CNH2, CNH3, and CNH4 categories include a total of 380 surnames (Fay (2004) had 353 surnames in his "Highly common non-Hispanic surnames" category).

Tables of surname category (snamecat) by geographic category (geocat) follow. The tables include all Residual Person links with an mscore of at least 9, **except** for links at the county level or higher in the following two categories: links with matching phone number, links from HU pairs that have two or more person links with an mscore of at least 7. Tables 4-7 break down the tabulation based on whether there is an exact age match or not and on truncated mscore (truncmscore). Exag=1 indicates that age matches exactly, exag=0 indicates there is a one-year (occasionally 2-5 years for phone or within-block matches) difference in age. Truncmscore is mscore truncated to the integer portion. For example, truncmscore of 9 indicates an mscore of at least 9 but less than 10. Truncmscore=10 indicates a "perfect" match (maximum agreement on all matching variables). Truncmscore=9 usually indicates a match that is "perfect" except that one or both persons are missing middle initial.

Table 4: geocat by snamecat, Residual Person Links
Controlling for truncmscore=9 exag=0

geocat / snamecat

Frequency / Cell/OTHR / Row Pct / Col Pct	OTHR	ASIA	CMNH	CNH2	CNH3	CNH4	FIRS	HISP	HSP2	HSP3	REMV	Total / % Total
Block	16203	39	2483	1588	2173	1464	35	1028	298	222	10	25543
	0.00	0.15	0.10	0.13	0.09	0.00	0.06	0.02	0.01	0.00	1.25
	63.43	0.15	9.72	6.22	8.51	5.73	0.14	4.02	1.17	0.87	0.04	
	9.16	0.35	0.28	0.83	1.51	2.47	0.57	0.19	0.96	1.66	1.61	
Tract	10162	32	1763	1205	1590	996	22	756	241	154	4	16925
	0.00	0.17	0.12	0.16	0.10	0.00	0.07	0.02	0.02	0.00	0.83
	60.04	0.19	10.42	7.12	9.39	5.88	0.13	4.47	1.42	0.91	0.02	
	5.74	0.28	0.20	0.63	1.11	1.68	0.36	0.14	0.78	1.15	0.64	
County	21262	609	6424	2713	3305	2151	206	20224	1991	1043	31	59959
	0.03	0.30	0.13	0.16	0.10	0.01	0.95	0.09	0.05	0.00	2.93
	35.46	1.02	10.71	4.52	5.51	3.59	0.34	33.73	3.32	1.74	0.05	
	12.02	5.40	0.73	1.41	2.30	3.62	3.38	3.78	6.43	7.82	4.98	
State	18921	1493	32339	7526	6596	3538	506	79123	5045	2324	47	157458
	0.08	1.71	0.40	0.35	0.19	0.03	4.18	0.27	0.12	0.00	7.70
	12.02	0.95	20.54	4.78	4.19	2.25	0.32	50.25	3.20	1.48	0.03	
	10.69	13.23	3.70	3.91	4.60	5.96	8.30	14.79	16.30	17.43	7.56	
National	110403	9109	831498	179443	129883	51242	5329	433786	23376	9593	530	1784192
	0.08	7.53	1.63	1.18	0.46	0.05	3.93	0.21	0.09	0.00	87.29
	6.19	0.51	46.60	10.06	7.28	2.87	0.30	24.31	1.31	0.54	0.03	
	62.39	80.74	95.08	93.23	90.48	86.28	87.39	81.09	75.53	71.93	85.21	
Total	176951	11282	874507	192475	143547	59391	6098	534917	30951	13336	622	2044077
% Total	8.66	0.55	42.78	9.42	7.02	2.91	0.30	26.17	1.51	0.65	0.03	100.00

Table 5: geocat by snamecat, Residual Person Links
Controlling for truncmscore=9 exag=1

geocat | snamecat

Frequency| Cell/OTHR| Row Pct | Col Pct	OTHR	ASIA	CMNH	CNH2	CNH3	CNH4	FIRS	HISP	HSP2	HSP3	REMV	Total % Total
Block	71527	106	9820	6268	8943	6150	200	3419	1023	730	19	108205
	0.00	0.14	0.09	0.13	0.09	0.00	0.05	0.01	0.01	0.00	7.20
	66.10	0.10	9.08	5.79	8.26	5.68	0.18	3.16	0.95	0.67	0.02	
	18.60	1.57	2.03	5.03	8.07	10.89	4.51	1.17	4.65	6.47	0.27	
Tract	48378	85	7701	4959	7005	4519	111	2512	761	578	19	76628
	0.00	0.16	0.10	0.14	0.09	0.00	0.05	0.02	0.01	0.00	5.10
	63.13	0.11	10.05	6.47	9.14	5.90	0.14	3.28	0.99	0.75	0.02	
	12.58	1.26	1.59	3.98	6.32	8.01	2.50	0.86	3.46	5.13	0.27	
County	98531	626	17530	9857	13751	9157	701	18820	3724	2445	105	175247
	0.01	0.18	0.10	0.14	0.09	0.01	0.19	0.04	0.02	0.00	11.65
	56.22	0.36	10.00	5.62	7.85	5.23	0.40	10.74	2.13	1.40	0.06	
	25.62	9.27	3.62	7.90	12.40	16.22	15.80	6.45	16.93	21.68	1.48	
State	67600	1095	24066	8822	10384	6843	502	43958	3714	2061	303	169348
	0.02	0.36	0.13	0.15	0.10	0.01	0.65	0.05	0.03	0.00	11.26
	39.92	0.65	14.21	5.21	6.13	4.04	0.30	25.96	2.19	1.22	0.18	
	17.58	16.21	4.97	7.07	9.37	12.12	11.31	15.07	16.89	18.28	4.27	
National	98568	4844	424640	94829	70782	29779	2924	223026	12773	5462	6652	974279
	0.05	4.31	0.96	0.72	0.30	0.03	2.26	0.13	0.06	0.07	64.79	
	10.12	0.50	43.59	9.73	7.27	3.06	0.30	22.89	1.31	0.56	0.68	
	25.63	71.70	87.78	76.02	63.85	52.75	65.89	76.45	58.07	48.44	93.72	
Total	384604	6756	483757	124735	110865	56448	4438	291735	21995	11276	7098	1503707
% Total	25.58	0.45	32.17	8.30	7.37	3.75	0.30	19.40	1.46	0.75	0.47	100.00

Table 6: geocat by snamecat, Residual Person Links
Controlling for truncmscore=10 exag=0

geocat | snamecat

Frequency| Cell/OTHR| Row Pct | Col Pct	OTHR	ASIA	CMNH	CNH2	CNH3	CNH4	FIRS	HISP	HSP2	HSP3	REMV	Total % Total
Block	13717	32	2402	1579	2113	1351	31	337	111	64	7	21744
	0.00	0.18	0.12	0.15	0.10	0.00	0.02	0.01	0.00	0.00	4.10
	63.08	0.15	11.05	7.26	9.72	6.21	0.14	1.55	0.51	0.29	0.03	
	15.41	0.91	0.91	2.59	4.31	6.44	2.51	0.87	4.12	4.71	14.58	
Tract	9438	13	1856	1213	1675	1025	19	238	83	65	4	15629
	0.00	0.20	0.13	0.18	0.11	0.00	0.03	0.01	0.01	0.00	2.95
	60.39	0.08	11.88	7.76	10.72	6.56	0.12	1.52	0.53	0.42	0.03	
	10.60	0.37	0.70	1.99	3.42	4.88	1.54	0.62	3.08	4.78	8.33	
County	18814	232	4224	2486	3156	1917	121	2124	386	257	14	33731
	0.01	0.22	0.13	0.17	0.10	0.01	0.11	0.02	0.01	0.00	6.36
	55.78	0.69	12.52	7.37	9.36	5.68	0.36	6.30	1.14	0.76	0.04	
	21.14	6.61	1.60	4.08	6.43	9.13	9.79	5.51	14.33	18.91	29.17	
State	13762	462	11131	3201	3381	1914	99	5937	420	239	4	40550
	0.03	0.81	0.23	0.25	0.14	0.01	0.43	0.03	0.02	0.00	7.64
	33.94	1.14	27.45	7.89	8.34	4.72	0.24	14.64	1.04	0.59	0.01	
	15.46	13.17	4.23	5.26	6.89	9.12	8.01	15.40	15.59	17.59	8.33	
National	33276	2769	243701	52416	38721	14780	966	29917	1694	734	19	418993
	0.08	7.32	1.58	1.16	0.44	0.03	0.90	0.05	0.02	0.00	78.96	
	7.94	0.66	58.16	12.51	9.24	3.53	0.23	7.14	0.40	0.18	0.00	
	37.39	78.93	92.55	86.08	78.95	70.42	78.16	77.60	62.88	54.01	39.58	
Total	89007	3508	263314	60895	49046	20987	1236	38553	2694	1359	48	530647
% Total	16.77	0.66	49.62	11.48	9.24	3.95	0.23	7.27	0.51	0.26	0.01	100.00

Table 7: geocat by snamecat, Residual Person Links
Controlling for truncmscore=10 exag=1

geocat	snamecat										
Frequency| Cell/OTHR| Row Pct | Col Pct	OTHR	ASIA	CMNH	CNH2	CNH3	CNH4	FIRS	HISP	HSP2	HSP3	REMV	Total % Total
Block	90756	92	14201	9557	13386	8908	198	1622	504	404	32	139660
		0.00	0.16	0.11	0.15	0.10	0.00	0.02	0.01	0.00	0.00	14.38
	64.98	0.07	10.17	6.84	9.58	6.38	0.14	1.16	0.36	0.29	0.02	
	18.69	3.48	6.93	12.34	14.88	16.61	8.22	5.18	9.91	11.26	0.21	
Tract	65002	44	11229	7525	10077	6691	125	1251	390	325	28	102687
		0.00	0.17	0.12	0.16	0.10	0.00	0.02	0.01	0.00	0.00	10.57
	63.30	0.04	10.94	7.33	9.81	6.52	0.12	1.22	0.38	0.32	0.03	
	13.38	1.66	5.48	9.71	11.20	12.48	5.19	3.99	7.67	9.06	0.19	
County	145954	422	25616	15917	22064	14626	819	6388	1967	1377	124	235274
		0.00	0.18	0.11	0.15	0.10	0.01	0.04	0.01	0.01	0.00	24.22
	62.04	0.18	10.89	6.77	9.38	6.22	0.35	2.72	0.84	0.59	0.05	
	30.05	15.94	12.51	20.55	24.52	27.27	33.98	20.40	38.67	38.37	0.83	
State	105505	539	20699	11760	15868	10321	438	5460	938	741	564	172833
		0.01	0.20	0.11	0.15	0.10	0.00	0.05	0.01	0.01	0.01	17.79
	61.04	0.31	11.98	6.80	9.18	5.97	0.25	3.16	0.54	0.43	0.33	
	21.72	20.36	10.11	15.18	17.64	19.24	18.17	17.44	18.44	20.65	3.77	
National	78437	1550	133078	32704	28572	13085	830	16595	1287	742	14194	321074
		0.02	1.70	0.42	0.36	0.17	0.01	0.21	0.02	0.01	0.18	33.05
	24.43	0.48	41.45	10.19	8.90	4.08	0.26	5.17	0.40	0.23	4.42	
	16.15	58.56	64.97	42.22	31.76	24.40	34.44	52.99	25.30	20.67	94.99	
Total	485654	2647	204823	77463	89967	53631	2410	31316	5086	3589	14942	971528
% Total	49.99	0.27	21.08	7.97	9.26	5.52	0.25	3.22	0.52	0.37	1.54	100.00

Table 8: geocat by snamecat, Residual Person Links
mscore ge 9

geocat	snamecat										
Frequency| Cell/OTHR| Row Pct | Col Pct	OTHR	ASIA	CMNH	CNH2	CNH3	CNH4	FIRS	HISP	HSP2	HSP3	REMV	Total % Total
Block	192203	269	28906	18992	26615	17873	464	6406	1936	1420	68	295152
		0.00	0.15	0.10	0.14	0.09	0.00	0.03	0.01	0.01	0.00	5.84
	65.12	0.09	9.79	6.43	9.02	6.06	0.16	2.17	0.66	0.48	0.02	
	16.92	1.11	1.58	4.17	6.76	9.38	3.27	0.71	3.19	4.80	0.30	
Tract	132980	174	22549	14902	20347	13231	277	4757	1475	1122	55	211869
		0.00	0.17	0.11	0.15	0.10	0.00	0.04	0.01	0.01	0.00	4.20
	62.77	0.08	10.64	7.03	9.60	6.24	0.13	2.25	0.70	0.53	0.03	
	11.70	0.72	1.23	3.27	5.17	6.95	1.95	0.53	2.43	3.80	0.24	
County	284561	1889	53794	30973	42276	27851	1847	47556	8068	5122	274	504211
		0.01	0.19	0.11	0.15	0.10	0.01	0.17	0.03	0.02	0.00	9.98
	56.44	0.37	10.67	6.14	8.38	5.52	0.37	9.43	1.60	1.02	0.05	
	25.04	7.81	2.95	6.80	10.75	14.62	13.02	5.30	13.29	17.33	1.21	
State	205788	3589	88235	31309	36229	22616	1545	134478	10117	5365	918	540189
		0.02	0.43	0.15	0.18	0.11	0.01	0.65	0.05	0.03	0.00	10.70
	38.10	0.66	16.33	5.80	6.71	4.19	0.29	24.89	1.87	0.99	0.17	
	18.11	14.83	4.83	6.87	9.21	11.87	10.89	15.00	16.66	18.15	4.04	
National	320684	18272	1632917	359392	267958	108886	10049	703324	39130	16531	21395	3498538
		0.06	5.09	1.12	0.84	0.34	0.03	2.19	0.12	0.05	0.07	69.28
	9.17	0.52	46.67	10.27	7.66	3.11	0.29	20.10	1.12	0.47	0.61	
	28.22	75.53	89.41	78.89	68.11	57.17	70.86	78.45	64.44	55.92	94.21	
Total	1136216	24193	1826401	455568	393425	190457	14182	896521	60726	29560	22710	5049959
% Total	22.50	0.48	36.17	9.02	7.79	3.77	0.28	17.75	1.20	0.59	0.45	100.00

The surname categories can use more refinement, but some useful observations can be made. A key

point is how the relationship between the percentage of names in a given category to the percentage of names in the OTHR category changes as one moves to higher geographic levels. A substantial increase suggests a problem with false matches since we expect the OTHR category to be less affected by any tendency for false matches to become more prevalent at higher geographic levels. The focus will be on the three hispanic categories and the four common nonhispanic categories. In the paragraphs below, "increase" is used to refer to an increase relative to the OTHR category. The REMV category is expected to be mostly false matches at all geographic levels.

‣ It does appear to make an important difference whether there is an exact age match. It also appears to make an important difference whether there is a "perfect" mscore (truncmscore=10) or a "nearly perfect" mscore (truncmscore=9).

‣ Looking at the OTHR category, the proportion of national links is substantially higher and the proportion of block and tract links are lower for the links with "nearly perfect" mscore and a one-year+ age difference (Table 4) when compared to the links with "perfect" mscore and an exact age match (Table 7). This may suggest some remaining problem with false matches at the national level even in the OTHR category for the links in Table 4.

‣ With a one-year+ age difference and a "nearly perfect" mscore (Table 4) the HISP, HSP2, and CMNH categories start notably increasing at the county level. The CNH2, CNH3, HSP3 and perhaps the CNH4 categories start increasing at the state level.

‣ With an exact age match and a "nearly perfect" mscore (Table 5) the HISP category starts increasing at the state (or even county) level. The CMNH and HSP2 categories start increasing at the state level. The CNH2, CNH3, CNH4, and HSP3 categories increase at the national level.

‣ For "perfect" mscore and a one-year+ age difference (Table 6) the HISP category starts to increase at the state (or even county) level. The CMNH and HSP2 categories also may start to increase at the state level. The CNH2, CNH3, CNH4, and HSP3 categories increase at the national level.

‣ For "perfect" mscore and exact age match (Table 7) all of the common Hispanic and nonhispanic categories may increase at the national level.

‣ The overall tabulation (Table 8) shows the sharply diminishing returns from the additional Hispanic name categories. The number of national links in the HISP category is much higher than the number in the HSP2 and HSP3 categories. The CMNH category is also considerably larger at the national level than the CNH2, CNH3, and CNH4 categories, although the disparity is not as large as in the Hispanic categories. Another point about the overall tabulation is that the Within Response Modeling accepted (thus removing from the Residual links) most of the within-block links with mscore 9+, but not many of the national links. If we exclude the phone number matches, fewer than 100,000 national links with mscore 9+ were accepted during Within Response Modeling.

‣ The FIRS and ASIA categories are relatively small categories.. It might be fine to treat the ASIA

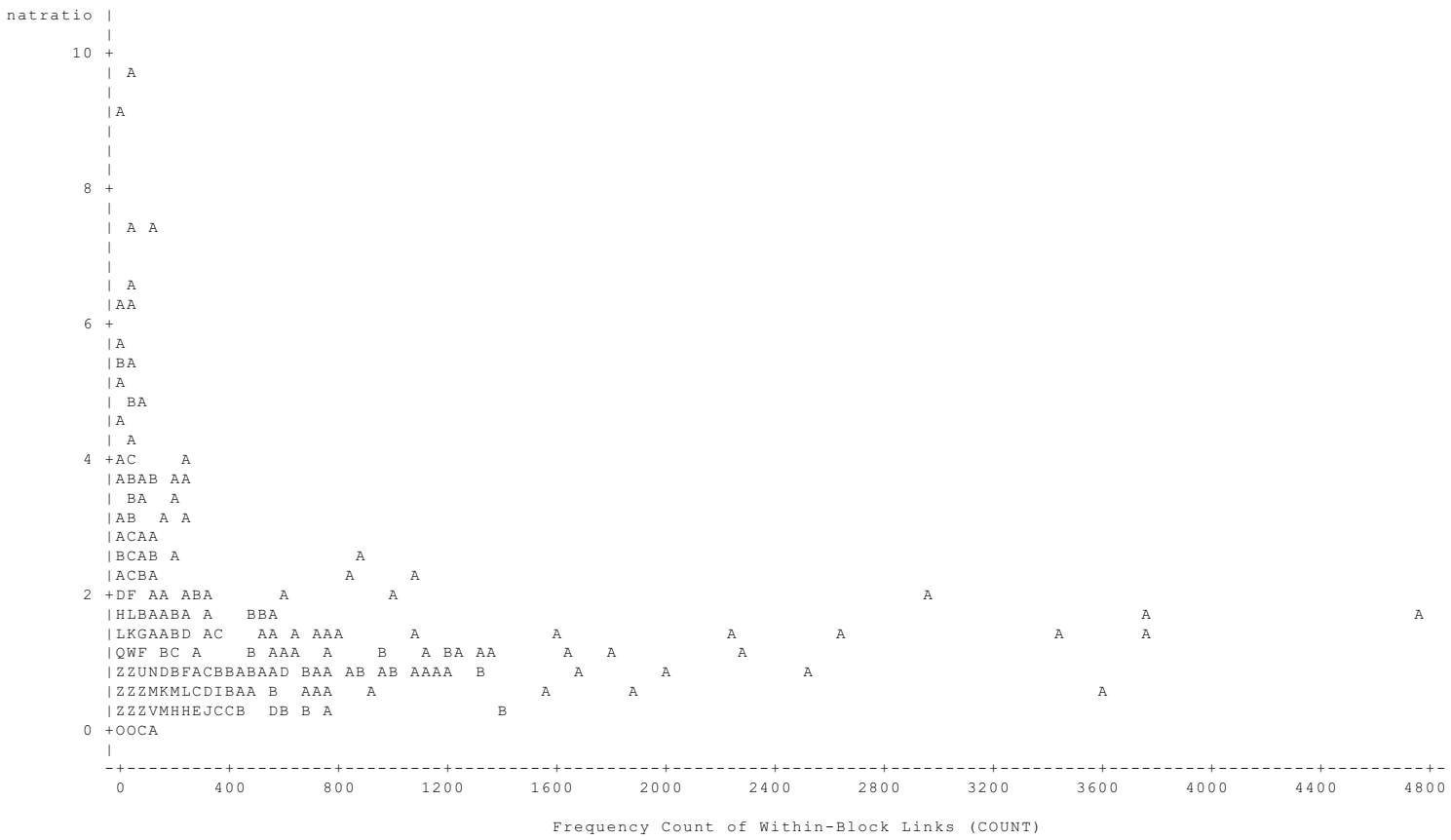
category similarly to the CNMH category and the FIRS category similarly to the CMNH or CNH2 categories.

Links From Residual Modeling: First Names

The same type of ratios for the same geographic categories.were calculated for first names as were previously calculated for surnames. Figure 4 shows the plot of national ratios against within-block count for first names when the surname is in the OTHR surname category. The within-block distribution included all Residual Person links in the OTHR category with mscore 9+. The national distribution excluded phone number matches and links from HU pairs with multiple links. It also required either an exact age match with mscore 9+ or an mscore greater than 10.

Figure 4: Plot of natratio*COUNT for first names, Residual Person Links
Surname is in OTHR category (Four Common Nonhispanic Categories)
Mscore 10+ or Exact Age Match Required For National Links

Legend: A = 1 obs, B = 2 obs, etc. (Z=26+ obs)



8 obs had missing values. 354 obs hidden. Only ratios with 11+ within-block links are plotted.

For clarity, ratios were only plotted for names with at least eleven within-block links. Observations with missing values in Figure 4 are names with at least eleven within-block links

but no national links.

There isn't much to say about first names for Residual Person links. There isn't much sign of a frequency effect. Nor are there generally clear explanations for which names have relatively high ratios and which names don't. There are some potential problem situations which affect a relatively small number of national links, at least if we only consider those links where the surname is in the OTHR category. One such situation is names where the reported birthday is heavily concentrated on specific days, such as "saint days" (the feast day of a patron saint) (Mule 2001). Another situation that showed up in the analysis of the Across Response matching is common surnames in the first name field (Ikeda and Porter 2007).

4. Summary and General Discussion

For Within Response Modeling there may be problems with false matches in specific situations such as names related to NRFU training examples, and "names" that are basically substitutes for unknown name. Name frequency does not seem to play a major role in false match rate for either first name or surname.

For Residual modeling, the general approach suggested is to divide surnames into categories and handle different categories differently. Conditions will be defined (e.g. mscore, exact age match or not) under which individual name categories would be ineligible for followup. The precise conditions would be affected by the desired tradeoff between not wanting to follow up false matches and not wanting to exclude true matches. The situation is less clear for first names, although it may be useful to separate especially common surnames in the first name field and certain "saint days" associated with first names, at least at the national level. Further discussion of the residual modeling is given below. The discussion below is similar to the corresponding discussion in Ikeda and Porter (2007) for the HU person links in Across Response Matching. However, the problems with false matches appear to be somewhat worse in the Residual Person links.

- We probably don't have to worry about name frequency for the phone number matches or for multiple-link (with mscore 7+) HU pairs. We do still need to worry about situations such as names related to NRFU training examples and names that are substitutes for unknown name, especially for the multiple-link HU pairs.

- For the remaining links, we want to at least separate the most common nonhispanic surnames and the most common Hispanic surnames. We probably also want to define additional categories of common nonhispanic surnames and common Hispanic surnames and to do something about common first names in the surname field and some other situations. Observation of the 1990 Puerto Rico Census by Edward Porter suggested that one thing that might be useful for matching Hispanic surnames is capturing a second surname.

- We want to include the presence or absence of an exact age match in our conditions that

determine when different categories of surnames are eligible for followup.

‣ At the national level, we likely will *not* generally be able to send person links with the most common nonhispanic and Hispanic surnames to followup from residual modeling due to the undesirability of sending large numbers or high proportions of false matches to followup. Somewhat less common nonhispanic and Hispanic surnames (as well as others) also seem questionable at the national level. Note that the state level will also often be questionable for common nonhispanic and Hispanic surnames. Even the county level sometimes seems questionable for the most common Hispanic and nonhispanic surnames.

‣ Defining general rules for first names is trickier. There are some potential problem conditions, but they mostly affect a relatively small number of links once the problems in the surnames are taken care of. We may want to avoid sending national links to followup from residual modeling when the first name field contains any of several very common surnames. Another situation we might want to avoid for national links is when the reported birthday is a "saint day" or other situations where reported birthdays are heavily concentrated on specific days.

‣ There may still be problems with false matches even within the OTHR category at the national level, especially for links with a "nearly perfect" mscore and a one-year age difference. We may also want to think about surnames that are especially common in a county or state.

‣ We will want to have some procedure in place for handling the equivalent of the REMV cases. This is especially important at the national level.

‣ Finally, this analysis has been entirely heuristic and exploratory. Probabilities are not assigned to the links. Research should continue. In particular, Fay (2002, 2004) outlines an approach which may be applicable with appropriate modifications for the Residual Person links. However, the calculation will be more complicated in our case. For example, our matching procedure allows for a one-year age difference and also matches on middle initial. Other complications include the removal of the within-response matches and the special treatment of phone number matches and multiple-link HU pairs.

5. References

Fay, R.E. (2002), "Probabilistic Models for Detecting Census Person Duplication," *2002 Proceedings of the Joint Statistical Meetings on CD-ROM*, American Statistical Association, Alexandria, VA, pp. 969-974.

Fay, R.E. (2004), "An Analysis of Person Duplication in Census 2000," *2004 Proceedings of the Joint Statistical Meetings on CD-ROM*, American Statistical Association, Alexandria, VA.

Ikeda, M. and Porter, E. (2007), "Initial Results from a Nationwide BigMatch Matching of 2000 Census Data," Research Report Series RRS2007/22, Statistical Research Division,U.S. Census

Bureau, December 2007.

Lynch, M. (2005), "2006 Coverage Followup and Census Coverage Measurement Person Matching Parameter Software Requirements Specification," Internal Census Bureau memorandum, DSSD 2006 Census Test Memorandum Series I-05, December 22, 2005.

Lynch, M., Ikeda, M., and Porter, E. (2005), "2006 Coverage Followup and Census Coverage Measurement Person Match Modeling Software Requirements Specification," Internal Census Bureau memorandum, DSSD 2006 Census Test Memorandum Series I-06, December 22, 2005

Mule, T. (2001), ESCAP II: Person Duplication in Census 2000, ESCAP II Report 20, Decennial Statistical Studies Division, U.S. Census Bureau, October 11, 2001.

Porter, E. (2006), "Using Meta-Programs to Take Advantage of Multiple Processors," Unpublished.

Word, D.L.(2001?), Tabulations of Name Frequencies in 2000 Decennial Census, Excel Spreadsheets.

Word, D.L. and Perkins, R.C. Jr. (1996), "Building a Spanish Surname List for the 1990's--A New Approach to an Old Problem," Technical Working Paper No. 13, Population Division, U.S. Census Bureau, March 1996.

Yancey, W. (2007), "BigMatch: A Program for Extracting Probable Matches from a Large File," Research Report Series RRC2007/01, Statistical Research Division, U.S. Census Bureau, June 2007.

Appendix 1: Surname Categories

Below are lists of the surnames in the ten name categories used in the analysis of the housing unit person links. Category CMNH is in descending order of name frequency in the 2000 Census, other categories are in alphabetical order.

1) CMNH: Smith, Johnson, Williams, Brown, Jones, Miller, Davis, Wilson, Anderson, Taylor, Thomas, Moore, Martin, Jackson, Thompson, White, Lee, Harris, Clark, Lewis, Robinson, Walker, Young, Allen, Hall.

2) CNH2: Adams, Bailey, Baker, Bell, Bennett, Brooks, Campbell, Carter, Collins, Cook, Cooper, Cox, Edwards, Evans, Foster, Gray, Green, Hill, Howard, Hughes, James, Kelly, King, Long, Mitchell, Morgan, Morris, Murphy, Myers, Nelson, Parker, Peterson, Phillips, Price, Reed, Richardson, Roberts, Rogers, Ross, Sanders, Scott, Stewart, Turner, Ward, Watson, Wood, Wright.

3) CNH3: Alexander, Andrews, Armstrong, Arnold, Austin, Barnes, Berry, Bishop, Black, Boyd, Bradley, Bryant, Burke, Burns, Butler, Carlson, Carpenter, Carr, Carroll, Chapman, Cole, Coleman, Crawford, Cunningham, Daniels, Dean, Dixon, Duncan, Dunn, Elliott, Ellis, Ferguson, Fisher, Ford, Fox, Franklin, Freeman, Gardner, George, Gibson, Gilbert, Gordon, Graham, Grant, Greene, Griffin, Hamilton, Hansen, Hanson, Harper, Harrison, Hart, Harvey, Hawkins, Hayes, Henderson, Henry, Hicks, Hoffman, Holmes, Howell, Hudson, Hunt, Hunter, Jacobs, Jenkins, Jensen, Johnston, Jordan, Kelley, Kennedy, Knight, Lane, Larson, Lawrence, Lawson, Lynch, Marshall, Mason, Matthews, Mcdonald, Meyer, Mills, Montgomery, Morrison, Murray, Nichols, Obrien, Oliver, Olson, Owens, Palmer, Patel, Patterson, Payne, Perkins, Perry, Peters, Pierce, Porter, Powell, Ray, Reynolds, Rice, Richards, Riley, Robertson, Rose, Russell, Ryan, Schmidt, Shaw, Silva, Simmons, Simpson, Snyder, Spencer, Stephens, Stevens, Stone, Sullivan, Tucker, Wagner, Wallace, Warren, Washington, Watkins, Weaver, Webb, Weber,Wells, West, Wheeler, Williamson, Willis, Woods.

4) HISP: Aguilar, Alvarez, Castillo, Castro, Chavez, Cruz, Delgado, Diaz, Fernandez, Flores, Garcia, Garza, Gomez, Gonzales, Gonzalez, Gutierrez, Guzman, Hernandez, Herrera, Jimenez, Lopez, Martinez, Medina, Mendez, Mendoza, Morales, Moreno, Munoz, Ortiz, Pena, Perez, Ramirez, Ramos, Reyes, Rivera, Rodriguez, Romero, Ruiz, Salazar, Sanchez, Santiago, Soto, Torres, Vargas, Vasquez.

5) HSP2: Acosta, Aguirre, Alvarado, Arroyo, Avila, Ayala, Cabrera, Calderon, Campos, Cardenas, Carrillo, Colon, Contreras, Cortez, Deleon, Dominguez, Duran, Espinoza, Estrada, Figueroa, Franco, Fuentes, Guerrero, Juarez, Lara, Leon, Luna, Maldonado, Marquez, Mejia, Mercado, Miranda, Molina, Navarro, Nunez, Ochoa, Ortega, Pacheco, Padilla, Rios, Robles, Rojas, Rosales, Rosario, Salinas, Sandoval, Santana, Serrano, Solis, Suarez, Trujillo, Valdez, Vazquez, Vega, Velez.

6) HSP3: Acevedo, Arias, Baca, Barrera, Beltran, Benitez, Bernal, Blanco, Bonilla, Camacho, Cano, Cantu, Castaneda, Cervantes, Cisneros, Cordova, Correa, Cortes, Davila, Dejesus, Delacruz, Delarosa, Enriquez, Escobar, Esparza, Espinosa, Gallegos, Galvan, Guerra, Ibarra, Jaramillo, Lozano, Lucero, Lugo, Macias, Mata, Melendez, Meza, Montes, Montoya, Mora, Muniz, Nieves, Orozco, Otero, Pagan, Pineda, Ponce, Quinones, Quintana, Quintero, Rangel, Reyna, Rivas, Rocha, Rodriquez, Rosado, Rosas, Rubio, Salas, Sosa, Tapia, Trevino, Valencia, Valenzuela, Velasquez, Velazquez, Vigil, Villa, Villanueva, Villarreal, Villegas, Zamora, Zavala, Zuniga.

7) ASIA: Kim, Le, Mohamed, Nguyen, Thao, Tran, Vang, Xiong.

8) FIRS: Amanda, Angela, Barbara, Brenda, Brian, Brittany, Carol, Christina, Christine, Christopher, Crystal, Cynthia, David, Deborah, Denise, Diane, Donna, Dorothy, Elizabeth, Enrique, Eric, Fernando, Francisco, Guadalupe, Heather, Helen, Jamie, Jason, Jennifer, Jerry, Jessica, Jesus, John, Jorge, Jose, Joshua, Juan, Julie, Kathleen, Karen, Kenneth, Kevin, Kimberly, Laura, Linda, Lisa, Luis, Margaret, Maria, Mary, Matthew, Melissa, Michael, Michelle, Miguel, Nancy, Nicole, Pamela, Patricia, Rafael, Ricardo, Robert, Ronald, Samantha, Sandra, Sarah, Stephanie, Steven, Susan, Teresa, William.

9) REMV: Boswell, Burgos, Doe, Whitman.

10) CNH4: Adkins, Baldwin, Ball, Banks, Barber, Barker, Barnett, Barrett, Barton, Bates, Beck, Becker, Benson, Blair, Blake, Bowen, Bowman, Brady, Brewer, Burgess, Burton, Bush, Byrd, Caldwell, Cannon, Casey, Chambers, Chan, Chandler, Chang, Chen, Christensen, Cohen, Craig, Cross, Cummings, Curry, Curtis, Daniel, Davidson, Dawson, Day, Dennis, Douglas, Doyle, Erickson, Farmer, Fields, Fischer, Fitzgerald, Fleming, Fletcher, Fowler, Francis, Frank, Frazier, Fuller, Gallagher, Garner, Garrett, Gill, Goodman, Goodwin, Graves, Gregory, Griffith, Gross, Hale, Hammond, Hampton, Hardy, Harmon, Haynes, Higgins, Hines, Hodges, Holland, Holt, Hopkins, Horton, Hubbard, Ingram, Jennings, Joseph, Keller, Klein, Lambert, Leonard, Little, Love, Lowe, Lucas, Lyons, Mack, Malone, Mann, Manning, Maxwell, May, Mccarthy, Mccoy, Mcdaniel, Mcgee, Mckinney, Mclaughlin, Miles, Moran, Moss, Mullins, Neal, Newman, Newton, Norris, Oconnor, Osborne, Page, Park, Parks, Parsons, Paul, Pearson, Potter, Powers, Quinn, Ramsey, Reese, Reeves, Reid, Rhodes, Robbins, Rodgers, Rowe, Santos, Saunders, Schneider, Schroeder, Schultz, Schwartz, Sharp, Shelton, Sherman, Simon, Sims, Singh, Stanley, Steele, Stevenson, Strickland, Sutton, Swanson, Tate, Terry, Thornton, Todd, Townsend, Vaughn, Wade, Walsh, Walters, Walton, Wang, Warner, Waters, Watts, Webster, Welch, Wise, Wolf, Wolfe, Wong, Yang, Zimmerman.

11) OTHR: All other surnames. When the CNH4 category is not used, names in the CNH4 category are placed in the OTHR category.

Appendix 2: Conceptual Flow for Resolving HU Person Links