RESEARCH REPORT SERIES
(Statistics #2007-22)

Initial Results from a Nationwide BigMatch Matching of 2000 Census Data

Michael Ikeda
Edward Porter

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

Report Issued: December 29, 2007

Initial Results from a Nationwide BigMatch Matching of 2000 Census Data

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

Abstract: A nationwide unduplication operation is being considered 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 Census Bureau's BigMatch program performed a matching of person records across all Census addresses, using data from the 2000 Census.

This report provides an overview of the matching methodology and of the results of an exploratory analysis of the matching output. As expected, most of the problem with apparent false matches 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, Across Response Matching, 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 would include a nationwide matching and modeling process to identify potential duplicates. These potential duplicates will 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 ran the matching and modeling procedures on the data from the 2000 Census and are analyzing the results.

The data we used included persons from housing units (HUs) deleted by the Housing Unit Duplication Operations (HUDO). As the name implies, HUDO was implemented to identify and remove duplicate housing units (Nash 2000a). HUDO deleted (Nash 2000b) about 3.6 million persons in about 1.4 million HUs. However, if we had done a matching and modeling operation in 2000 it would have been before HUDO, so it seems reasonable to include the HUDO deletes in our processing.

The matching and modeling process takes place in several stages. The first stage is the Across Response Matching operation, the subject of this report. This operation links person records across all responses, except when both persons are from Group Quarters (GQ). GQ person links (links where one person is from a GQ and the other person is from a HU) go directly to the GQ modeling phase where GQ person links are evaluated to identify potential duplicates. For 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.

Within Response Modeling should preferentially identify true matches since the presence of multiple links increases our confidence in the individual links. Those HU person links from Across Response matching that are *not* identified as potential duplicates in Within Response Modeling are then evaluated in the Residual Modeling operation.

This report presents an initial analysis of the results from performing the Across Response Matching procedure on the 2000 Census Data. Section 2 contains an overview of the Across Response Matching Procedure and some associated processing. Section 3 presents the results of an exploratory analysis for both GQ person links and HU person links. The main result is that false matches 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 Across Response Matching procedure matches individual persons across all Census responses, except when both persons are from GQs. For our simulation of the 2010 procedure on the 2000 Census data, each Census address is a response. The matching is performed using the BigMatch record linkage system (Yancey 2007). The BigMatch system allows for multiple sets of matching criteria to be used in a single run of the program. Each blocking pass uses one set of matching criteria. Pairs linked in one pass are removed from later passes. The blocking passes are defined to allow the results from all passes to be combined after a simple adjustment to the BigMatch matching score is made. Additional software was developed to handle the processing of the nationwide match and take advantage of parallel processing capabilities (Porter 2006).

Ten blocking passes were used in performing the Across Response Matching on the 2000 Census data. The blocking passes were basically the same as those used for the Across Response Matching in the 2006 Census Test (Lynch 2006). Each blocking pass requires the pairs of records to meet specified blocking criteria before attempting a match and then matches using some or all of: first (given) name, surname, middle initial, month of birth, day of birth, age, and gender. Most of the links that are considered in the analysis in Section 3 come from the first three blocking passes. The first blocking pass requires matching phone number (pass only used for HU person links). The second blocking pass requires matching Census block and matching first and last initials. The third blocking pass requires matching first name and surname. The next three blocking passes relax the blocking criteria, and the last four are aimed at matching records with first name and surname swapped. A "nickname file" is used to convert some common "nicknames" to the base first name (name they are a nickname of). In this simulation, the base first name is used for the first six passes and is the output first name for these passes.

The BigMatch match score for each link in each pass was adjusted to account for the blocking criteria. The adjusted score is called the adjusted across response match score (mscore). If a person was linked to more than one person in the same Census address, only one link was kept for the analysis in Section 3--the link with the highest mscore. There may have been rare cases when a GQ person link was dropped when it should have been kept or was misclassified as a HU person link.

3. Results

This analysis is based on the results of the Across Response Matching operation. The operation creates both group quarters (GQ) person links (links where one person is from a GQ and the other person is from a housing unit) and housing unit (HU) person links (links where both persons are from HUs). This analysis focuses on person links with an adjusted across response match score (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, 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 differences in first name or surname. At low geographic levels (such as block and tract) we expect that almost all of these links will be true matches. Note that 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.

The results are presented separately for HU person links and GQ person links. Many of the HU person links will go into the Within Response modeling operation, which evaluates multiple links between pairs of HUs. Within Response Modeling should preferentially identify true matches, since the presence of multiple links increases our confidence in the individual links. Since links identified in Within Response modeling do not go into Residual Modeling, the proportion of false matches going into Residual Modeling should be *larger* than suggested by this exploratory analysis of the Across Response matching. For both HU and GQ 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 there may be some effect for HU person links and there are individual first names which may have problems.

<u>Housing Unit Person Links: Surnames</u>

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)

Only links with an adjusted across response match score (mscore) of at least 9.0 are included in the distribution. Links with matching phone number were excluded from the distribution for geographic categories 3-5. There are 183,597 county links, 79,333 state links, and 38,220 national links with matching phone number and mscore of at least 9.0. 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. An upward slope suggests that more common names are being linked more often at the higher geographic level. 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, HU person links

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

natratio |
         |
      10 +
         |                                                                                                                         A
         |                             A
         |
         |                                   A
         |A
         |
       8 +                                 A
         |                                                                                           A
         |                                    A
         |
         |
         |                          A
         |                           A
       6 +A                                                                             A
         |                                                                    A A
         |                                                      A
         |A
         |               A   A  A
         |A                                                       A
         |
       4 +E        A
         |C                                  A   A
         | A   A       A                    A  A
         |C                              A    AA
         |C           A   A        A       B
         |GA                           AA
         |I     ABA BAA     B  A C BA
       2 +HAA   AA    A A  AA  CC
         |PD  AADBAAAAB AABBAA
         |QABABBEC AB BFA CA
         |ZGCFFIFCDEJFEA
         |ZHMROGRRKEDBA
         |ZZZZZZNJBA
         |ZZZZNC
       0 +ZZDA
         |
        -+----------+----------+----------+----------+----------+----------+----------+----------+----------+----------+----------+
         0        2178       4356       6534       8712      10890      13068      15246      17424      19602      21780      23958

                                            Frequency Count of Within-Block Links (COUNT)

2459 obs had missing values.  18017 obs hidden.  Only names with 11+ within-block links are included.  Four ratios>10 omitted.
```

For clarity, ratios were only plotted for names with at least eleven within-block links, and four surnames (Boswell, Whitman, Burgos, Doe) with a national ratio greater than 10 were removed from Figure 1. Boswell, Whitman, and Burgos were included in nonresponse follow-up training examples (Mule 2001). Observations with missing values in Figure 1 are names which have at least eleven within-block links but no national links.

The first thing to note in Figure 1 is a group of names that slopes sharply upwards, with between about two thousand to about seven thousand within-block links. Almost all of these names are from the group of most frequent "heavily Hispanic" surnames identified by Word and Perkins (1996). To the right of this group, we see a slower general tendency for the national ratios to increase as the number of within-block links increases. Finally, there are names on the left edge of the graph with high ratios. The majority of these names are common *first* names but there are also name variants and a couple of Asian surnames.

We next divided surnames into ten categories as given below. Name frequency information from the 2000 Census was based on tabulations by David Word (2001). The names in each category are listed in the Appendix. 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 classify it as "generally" but not "heavily" Hispanic). The abbreviation for this category is CNH3.

For categories 4-6, the number of within-block links (with mscore of at least 9.0) 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. 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. The abbreviation for this category is FIRS.

9) Four surnames that are special problem cases. Three of these names are from nonresponse follow-up 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.

Tables of surname category (snamecat) by geographic category (geocat) follow. The tables include all HU person links with an mscore of at least 9, **except** for links with matching phone number at the county level or higher. Tables 1-4 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 1: geocat by snamecat, HU person links

Controlling for truncmscore=9 exag=0

geocat	Frequency / Cell/OTHR / Row Pct / Col Pct	OTHR	ASIA	CMNH	CNH2	CNH3	FIRS	HISP	HSP2	HSP3	REMV	Total / % Total
Block	Frequency	117584	746	14302	8900	12622	335	15294	4950	3320	63	178116
	Cell/OTHR		0.01	0.12	0.08	0.11	0.00	0.13	0.04	0.03	0.00	7.82
	Row Pct	66.02	0.42	8.03	5.00	7.09	0.19	8.59	2.78	1.86	0.04	
	Col Pct	30.26	6.04	1.60	4.36	7.88	5.02	2.75	13.19	18.65	6.14	
Tract	Frequency	38252	165	5632	3594	4893	107	3819	1180	885	15	58542
	Cell/OTHR		0.00	0.15	0.09	0.13	0.00	0.10	0.03	0.02	0.00	2.57
	Row Pct	65.34	0.28	9.62	6.14	8.36	0.18	6.52	2.02	1.51	0.03	
	Col Pct	9.84	1.34	0.63	1.76	3.05	1.60	0.69	3.15	4.97	1.46	
County	Frequency	37530	764	8471	3860	4917	330	22628	2718	1529	42	82789
	Cell/OTHR		0.02	0.23	0.10	0.13	0.01	0.60	0.07	0.04	0.00	3.63
	Row Pct	45.33	0.92	10.23	4.66	5.94	0.40	27.33	3.28	1.85	0.05	
	Col Pct	9.66	6.18	0.95	1.89	3.07	4.94	4.07	7.24	8.59	4.09	
State	Frequency	27930	1543	33058	7962	7209	527	79584	5189	2408	58	165468
	Cell/OTHR		0.06	1.18	0.29	0.26	0.02	2.85	0.19	0.09	0.00	7.26
	Row Pct	16.88	0.93	19.98	4.81	4.36	0.32	48.10	3.14	1.46	0.04	
	Col Pct	7.19	12.49	3.70	3.90	4.50	7.89	14.33	13.83	13.53	5.65	
National	Frequency	167249	9140	832775	179948	130554	5378	434158	23479	9662	848	1793191
	Cell/OTHR		0.05	4.98	1.08	0.78	0.03	2.60	0.14	0.06	0.01	78.71
	Row Pct	9.33	0.51	46.44	10.04	7.28	0.30	24.21	1.31	0.54	0.05	
	Col Pct	43.04	73.96	93.13	88.10	81.50	80.55	78.16	62.58	54.27	82.65	
Total		388545	12358	894238	204264	160195	6677	555483	37516	17804	1026	2278106
% Total		17.06	0.54	39.25	8.97	7.03	0.29	24.38	1.65	0.78	0.05	100.00

Table 2: geocat by snamecat, HU person links

Controlling for truncmscore=9 exag=1

geocat snamecat

Frequency / Cell/OTHR / Row Pct / Col Pct	OTHR	ASIA	CMNH	CNH2	CNH3	FIRS	HISP	HSP2	HSP3	REMV	Total / % Total
Block	499228	1991	57461	36419	51127	1435	54348	17693	11986	224	731912
		0.00	0.12	0.07	0.10	0.00	0.11	0.04	0.02	0.00	29.00
	68.21	0.27	7.85	4.98	6.99	0.20	7.43	2.42	1.64	0.03	
	44.21	20.72	10.17	20.68	27.88	20.99	14.85	38.48	43.19	1.59	
Tract	179623	454	24270	15685	22003	492	13794	4305	3081	65	263772
		0.00	0.14	0.09	0.12	0.00	0.08	0.02	0.02	0.00	10.45
	68.10	0.17	9.20	5.95	8.34	0.19	5.23	1.63	1.17	0.02	
	15.91	4.72	4.30	8.90	12.00	7.20	3.77	9.36	11.10	0.46	
County	177467	1038	27196	15781	21864	1269	27543	6452	4384	167	283161
		0.01	0.15	0.09	0.12	0.01	0.16	0.04	0.02	0.00	11.22
	62.67	0.37	9.60	5.57	7.72	0.45	9.73	2.28	1.55	0.06	
	15.71	10.80	4.81	8.96	11.92	18.56	7.53	14.03	15.80	1.18	
State	108519	1215	27919	11255	13958	604	45918	4322	2509	538	216757
		0.01	0.26	0.10	0.13	0.01	0.42	0.04	0.02	0.00	8.59
	50.06	0.56	12.88	5.19	6.44	0.28	21.18	1.99	1.16	0.25	
	9.61	12.64	4.94	6.39	7.61	8.83	12.55	9.40	9.04	3.81	
National	164453	4912	428184	97003	74420	3037	224325	13206	5789	13110	1028439
		0.03	2.60	0.59	0.45	0.02	1.36	0.08	0.04	0.08	40.75
	15.99	0.48	41.63	9.43	7.24	0.30	21.81	1.28	0.56	1.27	
	14.56	51.11	75.78	55.07	40.58	44.42	61.30	28.72	20.86	92.95	
Total	1129290	9610	565030	176143	183372	6837	365928	45978	27749	14104	2524041
% Total	44.74	0.38	22.39	6.98	7.27	0.27	14.50	1.82	1.10	0.56	100.00

Table 3: geocat by snamecat, HU person links

Controlling for truncmscore=10 exag=0

geocat snamecat

Frequency / Cell/OTHR / Row Pct / Col Pct	OTHR	ASIA	CMNH	CNH2	CNH3	FIRS	HISP	HSP2	HSP3	REMV	Total / % Total
Block	138275	530	20622	13538	18945	316	5573	1734	1213	81	200827
		0.00	0.15	0.10	0.14	0.00	0.04	0.01	0.01	0.00	24.05
	68.85	0.26	10.27	6.74	9.43	0.16	2.78	0.86	0.60	0.04	
	43.39	12.27	6.99	16.54	24.28	17.26	11.92	33.32	38.41	29.67	
Tract	63009	141	10074	7085	9200	139	1769	579	401	24	92421
		0.00	0.16	0.11	0.15	0.00	0.03	0.01	0.01	0.00	11.07
	68.18	0.15	10.90	7.67	9.95	0.15	1.91	0.63	0.43	0.03	
	19.77	3.27	3.42	8.66	11.79	7.59	3.78	11.13	12.70	8.79	
County	38283	353	7135	4257	5708	256	3136	674	489	24	60315
		0.01	0.19	0.11	0.15	0.01	0.08	0.02	0.01	0.00	7.22
	63.47	0.59	11.83	7.06	9.46	0.42	5.20	1.12	0.81	0.04	
	12.01	8.18	2.42	5.20	7.32	13.98	6.71	12.95	15.48	8.79	
State	23711	501	12362	3969	4526	132	6189	476	287	8	52161
		0.02	0.52	0.17	0.19	0.01	0.26	0.02	0.01	0.00	6.25
	45.46	0.96	23.70	7.61	8.68	0.25	11.87	0.91	0.55	0.02	
	7.44	11.60	4.19	4.85	5.80	7.21	13.24	9.15	9.09	2.93	
National	55417	2793	244628	53007	39650	988	30084	1741	768	136	429212
		0.05	4.41	0.96	0.72	0.02	0.54	0.03	0.01	0.00	51.41
	12.91	0.65	56.99	12.35	9.24	0.23	7.01	0.41	0.18	0.03	
	17.39	64.68	82.98	64.76	50.81	53.96	64.35	33.46	24.32	49.82	
Total	318695	4318	294821	81856	78029	1831	46751	5204	3158	273	834936
% Total	38.17	0.52	35.31	9.80	9.35	0.22	5.60	0.62	0.38	0.03	100.00

Table 4: geocat by snamecat, HU person links

Controlling for truncmscore=10 exag=1

geocat	Statistic	OTHR	ASIA	CMNH	CNH2	CNH3	FIRS	HISP	HSP2	HSP3	REMV	Total / % Total
Block	Frequency	873419	1743	123153	81413	114277	1859	28718	9173	6799	372	1240926
	Cell/OTHR		0.00	0.14	0.09	0.13	0.00	0.03	0.01	0.01	0.00	41.79
	Row Pct	70.38	0.14	9.92	6.56	9.21	0.15	2.31	0.74	0.55	0.03	
	Col Pct	45.18	32.80	30.54	38.92	41.65	29.90	38.45	48.73	48.68	1.22	
Tract	Frequency	418926	499	62510	42542	57646	918	9212	2972	2214	142	597581
	Cell/OTHR		0.00	0.15	0.10	0.14	0.00	0.02	0.01	0.01	0.00	20.13
	Row Pct	70.10	0.08	10.46	7.12	9.65	0.15	1.54	0.50	0.37	0.02	
	Col Pct	21.67	9.39	15.50	20.34	21.01	14.76	12.33	15.79	15.85	0.47	
County	Frequency	293717	744	45533	28803	39929	1771	12139	3672	2829	230	429367
	Cell/OTHR		0.00	0.16	0.10	0.14	0.01	0.04	0.01	0.01	0.00	14.46
	Row Pct	68.41	0.17	10.60	6.71	9.30	0.41	2.83	0.86	0.66	0.05	
	Col Pct	15.19	14.00	11.29	13.77	14.55	28.48	16.25	19.51	20.25	0.76	
State	Frequency	189351	704	30805	18442	25525	665	7106	1420	1118	1132	276268
	Cell/OTHR		0.00	0.16	0.10	0.13	0.00	0.04	0.01	0.01	0.01	9.30
	Row Pct	68.54	0.25	11.15	6.68	9.24	0.24	2.57	0.51	0.40	0.41	
	Col Pct	9.80	13.25	7.64	8.82	9.30	10.69	9.51	7.54	8.00	3.73	
National	Frequency	157658	1624	141238	37993	36977	1005	17518	1587	1007	28513	425120
	Cell/OTHR		0.01	0.90	0.24	0.23	0.01	0.11	0.01	0.01	0.18	14.32
	Row Pct	37.09	0.38	33.22	8.94	8.70	0.24	4.12	0.37	0.24	6.71	
	Col Pct	8.16	30.56	35.03	18.16	13.48	16.16	23.45	8.43	7.21	93.83	
Total		1933071	5314	403239	209193	274354	6218	74693	18824	13967	30389	2969262
% Total		65.10	0.18	13.58	7.05	9.24	0.21	2.52	0.63	0.47	1.02	100.00

Table 5: geocat by snamecat, HU person links

geocat	Statistic	OTHR	ASIA	CMNH	CNH2	CNH3	FIRS	HISP	HSP2	HSP3	REMV	Total / % Total
Block	Frequency	1628506	5010	215538	140270	196971	3945	103933	33550	23318	740	2351781
	Cell/OTHR		0.00	0.13	0.09	0.12	0.00	0.06	0.02	0.01	0.00	27.33
	Row Pct	69.25	0.21	9.16	5.96	8.38	0.17	4.42	1.43	0.99	0.03	
	Col Pct	43.20	15.85	9.99	20.89	28.30	18.30	9.97	31.20	37.20	1.62	
Tract	Frequency	699810	1259	102486	68906	93742	1656	28594	9036	6581	246	1012316
	Cell/OTHR		0.00	0.15	0.10	0.13	0.00	0.04	0.01	0.01	0.00	11.76
	Row Pct	69.13	0.12	10.12	6.81	9.26	0.16	2.82	0.89	0.65	0.02	
	Col Pct	18.56	3.98	4.75	10.26	13.47	7.68	2.74	8.40	10.50	0.54	
County	Frequency	546997	2899	88335	52701	72418	3626	65446	13516	9231	463	855632
	Cell/OTHR		0.01	0.16	0.10	0.13	0.01	0.12	0.02	0.02	0.00	9.94
	Row Pct	63.93	0.34	10.32	6.16	8.46	0.42	7.65	1.58	1.08	0.05	
	Col Pct	14.51	9.17	4.09	7.85	10.41	16.82	6.28	12.57	14.73	1.01	
State	Frequency	349511	3963	104144	41628	51218	1928	138797	11407	6322	1736	710654
	Cell/OTHR		0.01	0.30	0.12	0.15	0.01	0.40	0.03	0.02	0.00	8.26
	Row Pct	49.18	0.56	14.65	5.86	7.21	0.27	19.53	1.61	0.89	0.24	
	Col Pct	9.27	12.54	4.83	6.20	7.36	8.94	13.31	10.61	10.09	3.79	
National	Frequency	544777	18469	1646825	367951	281601	10408	706085	40013	17226	42607	3675962
	Cell/OTHR		0.03	3.02	0.68	0.52	0.02	1.30	0.07	0.03	0.08	42.71
	Row Pct	14.82	0.50	44.80	10.01	7.66	0.28	19.21	1.09	0.47	1.16	
	Col Pct	14.45	58.45	76.34	54.80	40.46	48.27	67.71	37.21	27.48	93.04	
Total		3769601	31600	2157328	671456	695950	21563	1042855	107522	62678	45792	8606345
% Total		43.80	0.37	25.07	7.80	8.09	0.25	12.12	1.25	0.73	0.53	100.00

The surname categories can use further refinement, but some useful observations can be made. A key point is how the relationship between the percentage of names in a given name category to the percentage in the "other" name 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 three common nonhispanic categories. In the paragraphs below, "increase" is used to refer to an increase relative to the OTHR category. The REMV category is ignored below since links in this category are 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 1) when compared to the links with "perfect" mscore and an exact age match (Table 4). This may suggest some remaining problem with false matches at the national level even in the OTHR category for the links in Table 1.

- With a one-year+ age difference and a "nearly perfect" mscore (Table 1) the HISP category starts increasing for county links. The CNMH and HSP2 categories may also start increasing at the county level. The HSP3, CNH2, and CNH3 categories seem to start increasing at the state level.

- With an exact age match and a "nearly perfect" mscore (Table 2) the HISP category starts to increase at the state level. The CMNH category may also start increasing at the state level. The CNH2, CNH3, HSP2, and HSP3 categories may increase at the national level.

- For "perfect" mscore and a one-year+ age difference (Table 3) there are signs of increase in the HISP and CMNH categories at the state level. The CNH2 and CNH3 categories increase at the national level. The HSP2 category also may increase at the national level.

- For "perfect" mscore and exact age match (Table 4) the CMNH and the HISP categories increase at the national level. The CNH2 and CNH3 categories may also increase at the national level. The two smaller Hispanic categories do not increase much even at the national level.

- The overall tabulation (Table 5) 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 two smaller nonhispanic categories, although the disparity is not as large as in the Hispanic categories.

- The FIRS category is fairly small, but it may be worth separating out. It can perhaps be treated similarly to the CMNH category. And while this isn't shown in Tables 1-5, the names in the

ASIA category may be somewhat heterogeneous in behavior. Overall, it can also perhaps be treated similarly to the CMNH category.

Housing Unit Person Links: First Names

The same type of ratios for the same geographic categories were calculated for first names as were previously calculated for surnames. These ratios were calculated both for all links and for links where the surname is in the OTHR surname category. Figure 2 shows the plot of national ratios against within-block count for first names when the surname is in the OTHR surname category.

Figure 2: Plot of natratio*COUNT for first names, HU person links
Surname is in OTHR category

```
                                         Legend: A = 1 obs, B = 2 obs, etc.  (Z=26+ obs)
natratio |
         |
       7 +
         |A
         |
         |
         |
         |
       6 +
         |
         |
         |B
         |
         |A
       5 +
         |
         |C
         |
         |A
         |A
       4 +A
         |B
         |C
         |D
         |D
         |F
       3 +A
         |E
         |E
         |H
         |HBA
         |O
       2 +N                               A                                                      A    A    A                      A
         |UA    A     A                           A                                          A
         |ZA                              B  A                       A         A
         |ZA             A           A     A         A  A         B
         |ZB A A       A    A ABA   A      A       A    A   A
         |ZDAA A A AA BB   A AAAAA A C AA    A    AA
       1 +ZKAFCCEAF A DB C AABAF B BBA A   A A                   A
         |ZWNHHHFFFGCACEABBCAB       A
         |ZZZWTNKJEABBBEA  A
         |ZZZZPIFCAAA
         |ZZOBB A
         |ZK
       0 +T
         |
         -+----------+----------+----------+----------+----------+----------+----------+----------+----------+----------+----------+
          0        3212       6424       9636      12848      16060      19272      22484      25696      28908      32120    35332

                                            Frequency Count of Within-Block Links (COUNT)

   449 obs had missing values.  3526 obs hidden.  Only ratios with 11+ within-block links are plotted.  8 ratios>7 are omitted.
```

For clarity, ratios are only plotted for names with at least eleven within-block links, and eight first names with a national ratio greater than 7 were removed from Figure 2. Seven of these ratios (ranging from 7.47 to 44.09) are common surnames that appeared in the first name field. The other name is Mai. Observations with missing values in Figure 2 are first names which have at least eleven within-block links but no national links where the surname is in the OTHR category. The plot allowing all surnames is similar to Figure 2, except that the group of high ratios near the left side of the plot contains more Hispanic names.

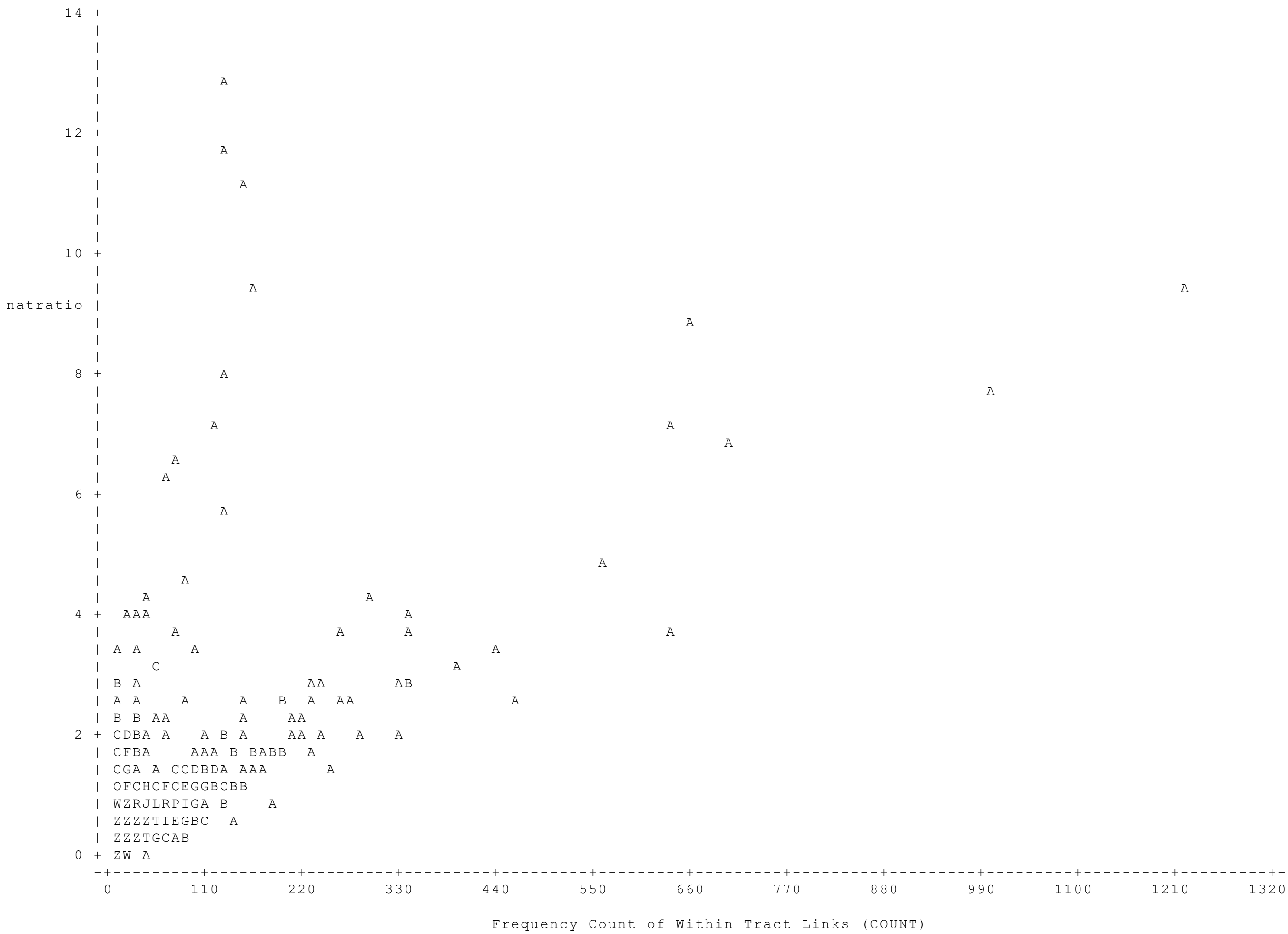
There does not appear to be much of a frequency effect for first names, although the national ratios may have a weak tendency to increase with increasing number of within-block links. There are also several different known situations which mostly affect a relatively small number of national links, at least if we consider only 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 the feast day of a patron saint (Mule 2001). Another situation is surnames in the first name field. At the national level, it may be worth separating out links with the most common surnames in the first name field. The group of high ratios near the left edge of the graph also includes what appear to be a number of Asian first names.

<u>Group Quarters Person Links: Surnames</u>

Similar ratios were calculated for GQ person links for surnames as had previously been calculated for HU person links. There are two differences between the ratios for GQ person links and the ratios for HU person links. First, the within-block and within-tract outside-block categories were merged to form a single within-tract category. The percentage distribution of surname for the within-tract category is the denominator of the GQ ratios. Second, phone number matches have been defined not to exist for GQ person links. Figure 3 shows the plot of national ratios against within-tract count for surnames for GQ person links.

Figure 3: Plot of natratio*COUNT for surnames, GQ person links

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



5 obs had missing values. 945 obs hidden. Only ratios with 11+ within-tract links are plotted.

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

Similar to the HU person links in Figure 1, two key characteristics of the plot are a general tendency for the national ratio to increase with the within-tract frequency count and a group of mostly "heavily Hispanic" names that slopes sharply upward in the left side of the plot. This suggests that similar name categories to those used for the HU person links may also be useful for the GQ person links.

We thus grouped the surnames into eight categories as outlined below:

1) The CMNH, CNH2, CNH3, HISP, HSP2, ASIA, and FIRS include the same names as they did for the HU person links.

2) The three surnames from the nonresponse follow-up training examples are housing unit examples. They are moved to the OTHR category for GQ person links.

3) The HSP3 category is not used for the GQ person links. Names in this category are moved to the OTHR category.

Tables of surname category (snamecat) by GQ geographic category (gqgeocat) follow. The tables include all GQ person links with an mscore of at least 9. Tables 6-9 break down the tabulation based on whether there is an exact age match or not and on truncated mscore (truncmscore). Exag and truncmscore are defined in the same way as they were for Tables 1-4.

Table 6: gqgeocat by snamecat, GQ person links
Controlling for truncmscore=9 exag=0

gqgeocat (Frequency / Cell/OTHR / Row Pct / Col Pct)	OTHR	ASIA	CMNH	CNH2	CNH3	FIRS	HISP	HSP2	REMV	Total / % Total
Block or Tract	7279	29	891	598	801	15	333	98	0	10044
		0.00	0.12	0.08	0.11	0.00	0.05	0.01	0.00	4.86
	72.47	0.29	8.87	5.95	7.97	0.15	3.32	0.98	0.00	
	19.28	5.21	0.90	2.93	5.06	3.19	1.13	4.89	0.00	
County	8567	51	1743	915	1151	28	1185	265	7	13912
		0.01	0.20	0.11	0.13	0.00	0.14	0.03	0.00	6.74
	61.58	0.37	12.53	6.58	8.27	0.20	8.52	1.90	0.05	
	22.69	9.16	1.75	4.49	7.28	5.96	4.01	13.22	1.37	
State	7156	96	4554	1323	1296	40	4162	354	45	19026
		0.01	0.64	0.18	0.18	0.01	0.58	0.05	0.01	9.22
	37.61	0.50	23.94	6.95	6.81	0.21	21.88	1.86	0.24	
	18.95	17.24	4.58	6.49	8.19	8.51	14.08	17.66	8.81	
National	14758	381	92204	17564	12571	387	23872	1288	459	163484
		0.03	6.25	1.19	0.85	0.03	1.62	0.09	0.03	79.18
	9.03	0.23	56.40	10.74	7.69	0.24	14.60	0.79	0.28	
	39.08	68.40	92.77	86.10	79.47	82.34	80.78	64.24	89.82	
Total	37760	557	99392	20400	15819	470	29552	2005	511	206466
% Total	18.29	0.27	48.14	9.88	7.66	0.23	14.31	0.97	0.25	100.00

Table 7: gqgeocat by snamecat, GQ person links

Controlling for truncmscore=9 exag=1

gqgeocat / snamecat

gqgeocat	Frequency / Cell/OTHR / Row Pct / Col Pct	OTHR	ASIA	CMNH	CNH2	CNH3	FIRS	HISP	HSP2	REMV	Total / % Total
Block or Tract	Frequency	52402	171	6010	3965	5891	78	1575	512	1	70605
	Cell/OTHR		0.00	0.11	0.08	0.11	0.00	0.03	0.01	0.00	18.51
	Row Pct	74.22	0.24	8.51	5.62	8.34	0.11	2.23	0.73	0.00	
	Col Pct	24.18	17.26	7.92	14.84	18.76	9.55	6.42	12.65	0.30	
County	Frequency	66529	237	10039	5844	8253	224	4624	1365	9	97124
	Cell/OTHR		0.00	0.15	0.09	0.12	0.00	0.07	0.02	0.00	25.47
	Row Pct	68.50	0.24	10.34	6.02	8.50	0.23	4.76	1.41	0.01	
	Col Pct	30.70	23.92	13.23	21.87	26.28	27.42	18.86	33.74	2.69	
State	Frequency	65146	311	11092	6077	8267	222	4962	1082	18	97177
	Cell/OTHR		0.00	0.17	0.09	0.13	0.00	0.08	0.02	0.00	25.48
	Row Pct	67.04	0.32	11.41	6.25	8.51	0.23	5.11	1.11	0.02	
	Col Pct	30.07	31.38	14.62	22.75	26.33	27.17	20.24	26.74	5.39	
National	Frequency	32595	272	48724	10830	8990	293	13354	1087	306	116451
	Cell/OTHR		0.01	1.49	0.33	0.28	0.01	0.41	0.03	0.01	30.54
	Row Pct	27.99	0.23	41.84	9.30	7.72	0.25	11.47	0.93	0.26	
	Col Pct	15.04	27.45	64.22	40.54	28.63	35.86	54.47	26.87	91.62	
Total		216672	991	75865	26716	31401	817	24515	4046	334	381357
% Total		56.82	0.26	19.89	7.01	8.23	0.21	6.43	1.06	0.09	100.00

Table 8: gqgeocat by snamecat, GQ person links

Controlling for truncmscore=10 exag=0

gqgeocat / snamecat

gqgeocat	Frequency / Cell/OTHR / Row Pct / Col Pct	OTHR	ASIA	CMNH	CNH2	CNH3	FIRS	HISP	HSP2	REMV	Total / % Total
Block or Tract	Frequency	2322	12	320	233	301	1	67	19	0	3275
	Cell/OTHR		0.01	0.14	0.10	0.13	0.00	0.03	0.01	0.00	9.35
	Row Pct	70.90	0.37	9.77	7.11	9.19	0.03	2.05	0.58	0.00	
	Col Pct	15.51	8.16	2.79	7.27	9.32	1.82	3.89	8.09	.	
County	Frequency	4518	37	893	481	660	18	226	76	0	6909
	Cell/OTHR		0.01	0.20	0.11	0.15	0.00	0.05	0.02	0.00	19.72
	Row Pct	65.39	0.54	12.93	6.96	9.55	0.26	3.27	1.10	0.00	
	Col Pct	30.18	25.17	7.79	15.01	20.44	32.73	13.12	32.34	.	
State	Frequency	4798	30	1231	570	766	17	303	55	0	7770
	Cell/OTHR		0.01	0.26	0.12	0.16	0.00	0.06	0.01	0.00	22.18
	Row Pct	61.75	0.39	15.84	7.34	9.86	0.22	3.90	0.71	0.00	
	Col Pct	32.05	20.41	10.74	17.79	23.72	30.91	17.60	23.40	.	
National	Frequency	3334	68	9022	1920	1502	19	1126	85	0	17076
	Cell/OTHR		0.02	2.71	0.58	0.45	0.01	0.34	0.03	0.00	48.75
	Row Pct	19.52	0.40	52.83	11.24	8.80	0.11	6.59	0.50	0.00	
	Col Pct	22.27	46.26	78.68	59.93	46.52	34.55	65.39	36.17	.	
Total		14972	147	11466	3204	3229	55	1722	235	0	35030
% Total		42.74	0.42	32.73	9.15	9.22	0.16	4.92	0.67	0.00	100.00

Table 9: gqgeocat by snamecat, GQ person links

Controlling for truncmscore=10 exag=1

gqgeocat / snamecat

gqgeocat	Frequency / Cell/OTHR / Row Pct / Col Pct	OTHR	ASIA	CMNH	CNH2	CNH3	FIRS	HISP	HSP2	REMV	Total / % Total
Block or Tract	Frequency	30872	62	3856	2615	3737	38	553	169	1	41903
	Cell/OTHR		0.00	0.12	0.08	0.12	0.00	0.02	0.01	0.00	14.48
	Row Pct	73.67	0.15	9.20	6.24	8.92	0.09	1.32	0.40	0.00	
	Col Pct	15.26	11.15	11.56	13.63	14.06	7.57	10.35	10.68	25.00	
County	Frequency	52894	167	8619	5202	7294	140	1913	602	2	76833
	Cell/OTHR		0.00	0.16	0.10	0.14	0.00	0.04	0.01	0.00	26.55
	Row Pct	68.84	0.22	11.22	6.77	9.49	0.18	2.49	0.78	0.00	
	Col Pct	26.14	30.04	25.84	27.12	27.45	27.89	35.79	38.03	50.00	
State	Frequency	81962	220	12066	7566	10576	206	1525	496	1	114618
	Cell/OTHR		0.00	0.15	0.09	0.13	0.00	0.02	0.01	0.00	39.60
	Row Pct	71.51	0.19	10.53	6.60	9.23	0.18	1.33	0.43	0.00	
	Col Pct	40.51	39.57	36.17	39.45	39.80	41.04	28.53	31.33	25.00	
National	Frequency	36616	107	8818	3796	4965	118	1354	316	0	56090
	Cell/OTHR		0.00	0.24	0.10	0.14	0.00	0.04	0.01	0.00	19.38
	Row Pct	65.28	0.19	15.72	6.77	8.85	0.21	2.41	0.56	0.00	
	Col Pct	18.10	19.24	26.43	19.79	18.69	23.51	25.33	19.96	0.00	
Total		202344	556	33359	19179	26572	502	5345	1583	4	289444
% Total		69.91	0.19	11.53	6.63	9.18	0.17	1.85	0.55	0.00	100.00

Table 10: gqgeocat by snamecat, GQ person links

Table of gqgeocat by snamecat

gqgeocat / snamecat

gqgeocat	Frequency / Cell/OTHR / Row Pct / Col Pct	OTHR	ASIA	CMNH	CNH2	CNH3	FIRS	HISP	HSP2	REMV	Total / % Total
Block or Tract	Frequency	92875	274	11077	7411	10730	132	2528	798	2	125827
	Cell/OTHR		0.00	0.12	0.08	0.12	0.00	0.03	0.01	0.00	13.79
	Row Pct	73.81	0.22	8.80	5.89	8.53	0.10	2.01	0.63	0.00	
	Col Pct	19.69	12.17	5.03	10.66	13.93	7.16	4.14	10.14	0.24	
County	Frequency	132508	492	21294	12442	17358	410	7948	2308	18	194778
	Cell/OTHR		0.00	0.16	0.09	0.13	0.00	0.06	0.02	0.00	21.35
	Row Pct	68.03	0.25	10.93	6.39	8.91	0.21	4.08	1.18	0.01	
	Col Pct	28.09	21.86	9.68	17.90	22.54	22.23	13.00	29.33	2.12	
State	Frequency	159062	657	28943	15536	20905	485	10952	1987	64	238591
	Cell/OTHR		0.00	0.18	0.10	0.13	0.00	0.07	0.01	0.00	26.15
	Row Pct	66.67	0.28	12.13	6.51	8.76	0.20	4.59	0.83	0.03	
	Col Pct	33.72	29.19	13.15	22.35	27.14	26.30	17.91	25.25	7.54	
National	Frequency	87303	828	158768	34110	28028	817	39706	2776	765	353101
	Cell/OTHR		0.01	1.82	0.39	0.32	0.01	0.45	0.03	0.01	38.70
	Row Pct	24.72	0.23	44.96	9.66	7.94	0.23	11.24	0.79	0.22	
	Col Pct	18.51	36.78	72.14	49.08	36.39	44.31	64.95	35.28	90.11	
Total		471748	2251	220082	69499	77021	1844	61134	7869	849	912297
% Total		51.71	0.25	24.12	7.62	8.44	0.20	6.70	0.86	0.09	100.00

Many of the patterns are similar to those for HU person links, although problems with false matches may not be quite as serious for GQ person links. As for the discussion of the tables for the HU person links, "increase" is used to refer to an increase relative to the OTHR category. The REMV category again 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 for the links with "nearly perfect" mscore and a one-year+ age difference (Table 6) when compared to the links with "perfect" mscore and an exact age match (Table 9). This may suggest some remaining problem with false matches at the national level even in the OTHR category for the links in Table 6.

- With a one-year+ age difference and a "nearly perfect" mscore (Table 6) the HISP and CMNH categories start notably increasing at the state level. The CNH2, CNH3, and HSP2 categories may also start increasing at the state level.

- With an exact age match and a "nearly perfect" mscore (Table 7) the HISP category may start increasing at the state level. The CMNH, CNH2, CNH3, and HSP2 categories increase at the national level.

- For "perfect" mscore and a one-year+ age difference (Table 8) the HISP and CMNH categories may start to increase at the state level. The CNH2, CNH3, and HSP2 categories increase at the national level.

- For "perfect" mscore and exact age match (Table 9) the HISP and CMNH categories may increase at the national level.

- The overall tabulation (Table 10) shows the sharply diminishing returns from the additional Hispanic name category. The number of national links in the HISP category is much higher than the number in the HSP2 category. The CMNH category is also considerably larger at the national level than the CNH2 and CNH3 categories, although the disparity is not as large as in the Hispanic categories.

- Both the FIRS and ASIA categories are fairly small. It might be fine to treat them similarly to the CNH2 category.

<u>GQ Person Links: First Names</u>

The same type of ratios for the same geographic categories were calculated for first names as were previously calculated for surnames. These ratios were calculated both for all links and for links where the surname is in the OTHR surname category. Figure 4 shows the plot of national ratios against within-tract count for first names when the surname is in the OTHR surname category. The plot allowing all surnames is similar, except that there is a group of high ratios near the left side of the plot that is mostly composed of Hispanic names.

For clarity, ratios were only plotted for names with at least eleven within-tract links. Missing observations in Figure 4 are names with at least eleven within-tract links but no national links.

Figure 4: Plot of natratio*COUNT for first names, GQ person links
Surname is in OTHR category

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

natratio |
         |
     5.0 +
         |
         |
         |     A
     4.5 +
         |
         | A  A
         |
     4.0 +
         |   A
         |   B
         | A
     3.5 + CA  A AA
         | B A   AA
         | A      A  C
         |  A
     3.0 + A B                   A
         | A  A       A        A
         | BBB A A                         A
         | AC  A   A A                      A                                    A
     2.5 + CAB     A  A                     A
         | CD  AA
         | IAA   AA AA              A
         | CA  A     A      A   A
     2.0 + ID A       A    A     A
         | DAB A    AAAAA   A                                                           A
         | DA CAAA  A     A              AA
         | EBBCA     A         A A          A                          A                                                         A
     1.5 + FBB ABA   A                 A
         | IDCAA  B    A         A                                                         A A
         | AGFAC      A                             A   A
         | EGA  A  B AAA  A A                                 A                    A
     1.0 + HC  DA   AA        A
         | HDAB    CC           A A      A       A
         | JAG ADA      A     A A A   A A                                           A
         | NEDE AAAA  A   A   A   A   A
     0.5 + HCFBBBA BA AA                                              A
         | SNDDEBBC A A  A    A A                                                                                          A
         | ZJEHECB  C B      A A    A                               A
         | ZQJKECDFECAACACDC BAAAABA A                          A     AA
     0.0 + FJECCBB  AA
         |
         -+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+-
          0        170       340       510       680       850      1020      1190      1360      1530      1700      1870      2040

                                             Frequency Count of Within-Tract Links (COUNT)

    26 obs had missing values.  21 obs hidden.  Only ratios with 11+ within-tract links are plotted.
```

There isn't much to say about first names for GQ person links. There isn't any sign of a frequency effect. Nor are there generally clear explanations for which names have relatively high ratios and which names don't.

Comparison of Results with Fay (2004)

Fay (2004) also did a nationwide matching of the 2000 Census, using estimated probabilities to estimate the number of duplicates at different geographic levels. His first phase of matching was an exact match on first name, surname, and full date of birth. This phase roughly corresponds to the Across Response Matching analyzed above. The three main differences are that we allowed age to differ by a year (occasionally more for phone or within-block matches), we matched on middle initial, and we included the HUDO deletes in our matching. The inclusion of the HUDO deletes should primarily affect the block and tract-level HU person links.

In the following comparison, we include or exclude entire categories from our results. This isn't completely realistic since there are true matches in excluded categories and false matches in included categories. However, it should give us a rough general-magnitude comparison between our results and those in Fay (2004).

We start by comparing Fay's Table 4 to our results for HU person links. Fay estimates 1,226,000 within-block duplicates and 322,000 tract-level duplicates for a total of 1,548,000 duplicates within-tract. He also estimates 909,000 county-level duplicates, 556,000 state-level duplicates, and 427,000 national-level duplicates. Our numbers are much higher for total within-tract duplicates, presumably because of the inclusion of the HUDO deletes. We have just about 2.35 million within-block matches and just over a million tract-level matches for a total of about 3.36 million within-tract matches, most of which we expect to be true matches. At higher geographic levels, our numbers tend to be similar to Fay's estimates. Note that in this discussion we include the phone matches and exclude the REMV category at all geographic levels. At the county level we start with about 1,039,000 links with an mscore of at least 9. The analysis following Tables 1-5 above suggests that most of them are likely to be true matches, although we lose about 35,000 links with "near perfect" mscore and a one-year age difference (we keep only the OTHR and HSP3 categories here), leaving us with just over a million. For the state level, if we assume that all of the OTHR category and all of the phone matches are true matches, that gets us to about 429,000. Adding in the links with "perfect" mscore and exact age match brings us to not quite 515,000. We pick up about another 32,000 from the CNH2, CNH3, HSP2, and HSP3 links with "near perfect" mscore and exact age match and just under 9,000 from CHN2, CNH3, and HSP3 with "perfect" mscore and one-year age difference for a total of just under 556,000. At the national level we again start with the OTHR category (except for links with "near perfect" mscore and one-year age difference) and the phone matches which totals just under 416,000. We can add in about 2,500 links from HSP2 and HSP3 with "perfect" mscore and exact age match and not quite 800 more from HSP3 with "perfect" mscore and one-year age difference, for a total of about 419,000.

Now comparing Fay's Table 3 to our results for GQ person links, Fay estimates 100,000 within-block duplicates and 41,000 tract-level duplicates for a total of 141,000 duplicates within tract. He also estimates 211,000 county-level duplicates, 224,000 state-level duplicates, and 94,000 national-level duplicates. Our numbers generally end up similar to but slightly below Fay's estimates. Excluding the REMV links, we start with about 126,000 within-tract links, about 195,000 county links, and about 239,000 state links. Most of these should be good, although we lose about 18,000 state links (everything except OTHR in Table 6, HISP in Table 7, HISP and CMNH in Table 8) and perhaps a few county links. For national links, the OTHR category (aside from those with both "near perfect" mscore and one-year age difference) and the links with "perfect" mscore and exact age match (aside from CMNH and HISP) combine for not quite 82,000 matches.

4. Summary and General Discussion

The general approach suggested is to divide surnames into categories and handle different categories differently in both the residual modeling and GQ modeling operations. 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 at least for residual modeling it may be useful to separate especially common surnames in the first name field, at least at the national level. Any suggestions relating to the residual modeling are tentative and depend on how the situation is affected by within response modeling. However, the within response modeling should preferentially identify true matches and thus the proportion of false matches in the residual modeling is likely to be *larger* than the above results suggest.

- At a minimum, we will probably want to separate the most common nonhispanic surnames and the most common Hispanic surnames. We may also want to define additional categories of common nonhispanic surnames and common Hispanic surnames. We probably also want to do something about common first names in the surname field and perhaps others. (Observation of the 1990 Puerto Rico Census by Edward Porter suggested that one thing that might be useful for Hispanic surnames is capturing a second surname.)

- We probably want to include the presence or absence of an exact age match in our conditions for when different categories of surnames are eligible for followup.

- At the national level, we likely will *not* generally be able to send HU person links with the most common nonhispanic and Hispanic surnames to followup from residual modeling. Somewhat less common nonhispanic and Hispanic surnames (as well as others) also seem questionable at the national level. For most persons with problem surnames this means that the *only* chance of finding true matches from national-level HU person links is by sending them to followup from the within response modeling. We may want to rethink our general philosophy about the within response modeling to reflect this, at least at the national level. At the national level, we can no longer count on picking up true matches in the residual modeling if we don't pick them up in within response modeling. The within response modeling will be our only chance for many person links. Note that the state level will also often be questionable for common nonhispanic and Hispanic surnames and similar comments will often apply there. Even the county level sometimes seems questionable for the most common Hispanic and nonhispanic surnames.

- We can also expect to have problems with false matches for the most common nonhispanic and Hispanic surnames in the GQ person links. The numbers are much smaller than for the HU person links, but we probably will not want to send these names to followup from the national-level GQ modeling. There are also other questionable situations for the GQ person links, but the numbers for these should be much smaller.

‣ Defining general rules for first names is trickier. There may be some weak frequency effect for first names for HU person links, although whether this is strong enough for us to want special rules is another question. There are also some identifiable 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-level HU person links to followup from residual modeling when the first name field contains any of several very common surnames.

‣ 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. One possibility is to define additional surname categories. Another is to place stricter conditions on the OTHR category at the national level. 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.

‣ We need to repeat this analysis for the HU person links after we do a reasonable run of the within response modeling. There may be additional problems apparent when we analyze the HU person links available for the residual modeling.

‣ 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. However, the calculation will be more complicated in our case since our matching procedure allows for a one-year age difference and also matches on middle initial. Note that Fay (2004) produces estimates of within-tract duplicates between HU persons that are about half of our rough estimates, presumably because we included the HUDO deletes in our matching. Our rough estimates for outside-tract duplicates and duplicates between GQ and HU persons are similar to Fay's numbers.

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.

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.

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.

Nash (2000a), "Overview of the Duplicate Housing Unit Operations," Internal Census Bureau memorandum, DMD Census 2000 Informational Memorandum No. 78, November 7, 2000.

Nash (2000b), "Results of Reinstatement Rules for the Housing Unit Duplication Operations," Internal Census Bureau memorandum, DMD Census 2000 Informational Memorandum No. 82, November 21, 2000.

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: 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) OTHR: All other surnames.