FINAL REPORT

Assessment of the U.S. Census Bureau's Person Identification Validation System

PRESENTED TO:
U.S. Census Bureau
4600 Silver Hill Road
Washington, DC 20233-4400

MARCH 31, 2011

PRESENTED BY:
NORC at the
University of Chicago
4350 East West Highway,
Suite 800
Bethesda, MD 20814
(301) 634-9300
(301) 634-9301 – Fax

NORC PROJECT TEAM:
Edward Mulrow Ph.D, PStat® (Principal Investigator)
Ali Mushtaq MS (Co-investigator)
Santanu Pramanik PhD (Co-investigator)
Angela Fontes PhD (Project Manager)

NORC
at the UNIVERSITY of CHICAGO

Table of Contents

List of Exhibits

Report Summary

This report presents the results of an assessment by NORC at the University of Chicago of the Person Identification Validation System (PVS) currently used by the U.S. Census Bureau. The PVS is the Census Bureau's production capability to verify and search for Social Security Numbers (SSNs) or Protected Identification Keys (PIKs) for person records in demographic surveys, censuses, or administrative records. The assessment reviewed the Census Bureau's record linkage methods, and focused on the efficiency of the matching algorithm, reviewing the quality of the input file, and reviewing the coverage of the reference files. Analyses and results include:

- **Comparison of GeoSearch and NameSearch Modules**

  Using the ACS 2009 file as the incoming file, match and agreement rates of the PVS GeoSearch and NameSearch modules were compared. Results indicate a general positive correlation between the match rates of the two modules. A substantial geographic relationship is also present in the matched proportions and the disagree proportions; Southwest states have lower matched proportions than Midwest states, and Northeast, most mid-Atlantic, and Midwest states were above the median state (Illinois).

- **Unmatched Record Analysis**

  NORC reviewed the ACS 2009 unmatched records to understand what may be causing the failure-to-match in three ways:

  - Cut and blocking strategy effects: For this analysis, records that failed to match within either the GeoSearch or NameSearch were run through the PVS system without blocking within module cuts. Results indicate very few additional matches can be found outside *both* the geo- and name-cuts.
  - Socioeconomic/Demographic profile of unmatched records: This analysis investigated whether unmatched records were associated with social, economic, or demographic factors of interest to data users. Results indicate differences in the composition of unmatched records, when compared to all records, on characteristics such as reported income, employment status, race/ethnic identity, and US citizenship.
  - Blocking/Matching variable missingness analysis: In this analysis, the level of missingness in unmatched records in variables such as Date of Birth (DOB), Geokeys (streetname,

streetname prefix and suffix, house number, rural route and box, and ZIP code), and Name was examined. The percent of missingness of DOB information appears to be correlated with high rates of unmatched records. For name data, when fake or incomplete names are considered equivalent to missing information, a correlation with the unmatched rate exists as well. It is less clear that Geokey missingness is as important a factor.

- **Reference File Coverage Assessment**

  Two methods were used to assess the coverage of the current PVS reference file:

  - Comparison of unmatched records between incoming files – ACS 2009 vs. Census 2010: The unmatched ACS 2009 records were compared with the unmatched Census 2010 records (used as the reference file). Results indicate some degree of under coverage in the reference files, but the substantial number of duplicate or unresolved matches present could point to quality issues with the records in both files.
  - Association between socioeconomic/demographic/geographic factors and missingness in unmatched records: The final investigation explores the association between the social, economic, demographic and geographic characteristics and the missingness of key blocking and matching variables in the unmatched ACS 2009 records. Results indicate that there are a number of dependencies between the missingness factors and the socioeconomic, demographic and geographic characteristics. Given this association, it will be difficult to increase the PVS match rates without addressing the quality of DOB and name variables in the incoming file. Addressing under-cover of certain groups within the reference file will help to increase PVS match rate, but the benefits will be dampened because of missing DOB and fake/incomplete name information in the incoming file records.

The Report concludes with a comprehensive set of Recommendations based on the above analyses which include:

- Recommended additional research based on the investigation undertaken in our PVS assessment in the following areas:
  - Cut and blocking strategies
  - Relationship between social, economic and demographic factors and the likelihood of a PVS match
  - The effect of incoming record data quality on matching

  - Matching cause and effect research
  - Reference file assessments
- Recommend research based on best practice concepts voiced by others who have used or reviewed the PVS, as well as the application of record linkage best practice concepts.
- Recommendation to consider creating a research and evaluation environment for PVS so that on-going research will not interfere or jeopardize PVS production runs.

Study Background and Purpose

The Person Identification Validation System (PVS) is the Census Bureau's production capability to verify and search for Social Security Numbers (SSNs) or Protected Identification Keys (PIKs) for person records in demographic surveys, censuses, or administrative records. PIK's are internal Census identifiers that correspond one-to-one with the set of nine-digit numbers from 000000000 to 999999999. Thus, a Social Security Number (SSN), which is a nine-digit number, corresponds one-to-one with a PIK and represents a unique individual. The PIK is assigned independently and randomly to protect the privacy of the individual person. Used as unique person identifiers, PIKs facilitate record linkage across files while enhancing data confidentiality and privacy. The quality of the PVS research files depends on the technical ability to assign the correct person identifier across linked files.

As part of the Person Identification Validation System Assessment engagement with the Census Bureau, NORC at the University of Chicago (NORC) has conducted a review of the Census Bureau's record linkage methods associated with the PVS, as well as an environmental scan of record linkage methods used by other government agencies—both within and outside of the U.S.—and private enterprises. This report provides NORC's assessment of the PVS to assign correct PIKs to a set of input records, as well as the PVS methods in the context of methods used by other public and private organizations.

This report has two primary sections and four appendices. The first section, **Review of the Person Identification Validation System**, provides the details of the NORC's review of PVS documentation, software programs, input files and system output. The second section, **Recommendations**, provides NORC's recommendations for possible PVS enhancement, and suggestions for PVS research projects. **Appendix A: Environmental Scan of Record Linkage Methods**, provides a summary of NORC's review of over 300 papers, conference presentations, and books that describe record linkage and entity resolution methods and applications. **Appendix B: List of Fake and Incomplete Names**, provides a list of first names and last names that we suspect are fake names used to fill-in the survey name field.[1] Such names are almost the same as blank names and need to be accounted for in an assessment of record linkage. The appendix also includes the list of fake or incomplete names that the PVS name-edit program tries to find and remove in the PVS initial edit step. **Appendix C: Loglinear Model SAS Code and Output**, provides the SAS code for the loglinear model that was fit to unmatched ACS 2009 data in order

[1] The lists of fake first and last names were extracted from the PVS unmatched records of the ACS 2009 incoming file. The Census Bureau has a list of fake or incomplete names that is used in a preprocessing step to blank-out incoming file records that have both first and last fake names. Because records with both first and last names blank are out-of-scope, such records are not processed in PVS, and are therefore not part of this assessment.

to test for independence between certain socioeconomic/demographic characteristics and the missingness of key blocking and matching variables. **Appendix D: Glossary**, is a glossary of terms and acronyms used in this report.

1 Review of the Person Identification Validation System

1.1 Introduction

1.1.1 PVS Background

The Person Identification Validation System (PVS) verifies SSNs and assigns PIKs by comparing person characteristics from an incoming file to the characteristics of records in the PVS reference files. The PVS uses three reference files containing Numident[2] data to verify and search for SSNs:

- The Census Numident – all Social Security Administration (SSA) Numident SSN records are edited (collapsed) to produce a Census Numident file that contains "one best-data record" for each SSN. All variants of name information for each SSN are retained in the Alternate Name Numident file, while all variants of date of birth data are retained in the Alternate DOB Numident. The SSN-PIK crosswalk file[3] is used to attach a corresponding unique PIK value for each SSN value in the Census Numident file.
- GeoBase Reference File – addresses are attached to Numident data from U.S. government administrative records,[4] including all possible combinations of alternate names and dates of birth for each SSN.
- Name Reference File – all possible combinations of alternate names and dates of birth for each SSN.

The PVS ensures the name and DOB information for an SSN matches the Numident information for that SSN and only returns the PIK corresponding to that SSN. The standard PVS methodology consists of an initial edit process, plus any or all of three modules – Verification, GeoSearch, and NameSearch.

---

[2] The Social Security Administration's (SSA) Numerical Identification (Numident) file contains all transactions ever recorded against any single SSN.

[3] The SSN-PIK crosswalk file is comprised of the output from the algorithm to randomly generate PIK values for every possible number between 1 and 999,999,999. This crosswalk file is created once and is used in creating the Census Numident files.

[4] Addresses from the IRS Individual Master File and Returns Transaction file (1040), IRS Information Returns file (1099), HUD assisted renter files, CMS Medicare file, Indian Health Service Registration file, and Selective Service Registration File are linked to Census Numident using SSNs. The vintage of the source data for PVS determines which administrative records addresses are used.

- Initial Edit – Perform name and address edits. Exclude from further processing any incoming records flagged as SSN refusals, and any records lacking first and last name data.
- Verification – When an SSN is provided on an incoming record, the verification step attempts to verify that the SSN/name/date of birth elements exist in the reference file.
- GeoSearch – When an incoming record does not have an SSN, or when an existing SSN is not verified, the GeoSearch module attempts to use address information to locate the appropriate SSN/name/date of birth record in the reference file, and outputs the PIK associated with the matched reference file record onto the incoming record. The GeoSearch capability is enhanced by the addition of an address (Geokey) to the reference file records using administrative records address information.
- NameSearch – When an incoming record is not verified or not matched in GeoSearch, or an incoming record has no SSN and no address information, a NameSearch step is used. NameSearch uses name and date of birth components of an incoming record to attempt to locate the appropriate record in the reference file, and output the PIK associated with the matched reference file record onto the incoming record.

The output of the PVS is a validated file containing all records from the incoming file. In PVS parlance, the term "**validated**" refers to the output file as well as to all records assigned a validated PIK, whether verified during the verification module, or assigned through one of the search processes. The term "**verified**" will refer only to those records validated through the verification module.

1.1.2 PVS Match Rates

The Census Bureau runs a number of survey datasets through the PVS, as well as all acquired administrative records. It has also run both Census 2000 and Census 2010 through the PVS. In general it appears that about 90 – 93 percent of survey records are matched to the PVS reference files and assigned PIKs. A similar percentage of Census records are assigned PIKs. A much higher percentage, approximately 98 percent, of federal administrative records are assigned PIKs. This should not be surprising because these federal administrative records are of generally high quality, and often include SSNs. **Exhibit 1** is a summary of match percentages that were obtained from reports provided to NORC by the Census Bureau for this PVS assessment. The match percentages are calculated relative to the number of records submitted to the module, whereas the validated percentage in the last column is related to all records in the incoming file.

Exhibit 1: Match Percentages for Census Bureau PVS Projects

Incoming Data	Matched in Verification	Matched in GeoSearch	Matched in NameSearch	Validated All Incoming
Survey Records				
ACS 2001	N/A	86.30	58.12	93.49
ACS 2002	N/A	86.27	57.57	93.12
ACS 2003	N/A	87.05	54.15	92.39
ACS 2004	N/A	88.16	53.63	92.60
ACS 2005	N/A	89.93	44.77	92.90
ACS 2006	N/A	87.87	47.53	92.03
ACS 2007	N/A	89.06	41.76	91.65
ACS 2008	N/A	88.08	46.07	91.71
ACS 2009	N/A	84.02	52.23	90.82
SIPP 2001*	93.74	69.57	33.19	93.06†
CPS 2001*	94.07	82.20	32.28	76.53
Census Records				
Census 2010	N/A	83.04	57.57	91.14
Federal Administrative Records (2009)				
HUD Public and Indian Housing Information Center File	99.27	42.05	43.53	99.54
IRS Individual Master File and Returns Transaction File (1040)	96.61	7.97	0.30	96.73
IRS Information Returns (1099)	97.28	50.61	0.46	98.66
CMS Active Medicare Enrollment Database	99.92	17.42	30.60	99.89
Indian Health Services Patient Registration File	97.17	29.41	67.23	97.43
Selective Service System Registration File	98.72	46.03	60.01	98.82
HUD Tenant Rental Assistance Certification System File	96.98	55.82	70.19	99.43

ACS yearly results were obtained from "ACS PVS Results All Years for Groves Briefing.xls"
CPS and SIPP results were obtained from "PVS Final Evaluation Report 10242006.doc"
Census 2010 Decennial Response File (DRF) results were obtained from "2010 Char Imp Results by State Table.rtf"
Federal Administrative Records results were obtained from "StARS 2009 PVS Results.doc"
*Results shown are for PVS reruns that occurred after improvements to the system where implemented during the 2004 timeframe.
† The refusals for SIPP 2001 were removed before the file was sent for PVS. Had they been in the file—as they were for the CPS 2001 file—the percent validated of all incoming records would have been much lower.

NORC understands that PVS match rates for records from commercial databases are lower, even when SSNs are present. This is likely due to a lower quality of data, and the commercial data providers' inability to verify that the SSNs are correct.

It is important to keep in mind that records from surveys such as the ACS do not include SSNs, and this is one of the reasons match rates for survey records are lower than those of federal administrative records. However, it may also be the case that the PVS reference file, which is built from administrative records, may not contain records for people that surveys sometimes capture—people who are "off-the-grid," which may include undocumented people, and other segments of society that may not have found their way into government agency records. Additionally, a person record in a survey database may contain an incomplete or bad name, address, or date of birth, which makes it unlikely for the record to get matched to a reference file record even if the person represented by the survey record also has a corresponding record in the PVS reference file.

Without the benefit of SSN matching within the Verification module, the PVS is essentially the two probabilistic record linkage modules GeoSearch and NameSearch. The match rates for these modules represent matches that are highly likely based on a probability linkage model. There may be false matches between the incoming and reference files, and there may be unmatched records from the incoming that do have a record in the reference file. The Census Bureau has investigated these issues and issues related to the reference file coverage as part of past PVS evaluation research projects.

1.1.3 Past PVS Evaluations

NORC understands that the PVS has undergone two past evaluations, which are documented in the following reports:

- *A Review of the Social Security Number Verification and Search Process (PVS) of the Planning, Research, and Evaluation Division*, March 21, 2003, Marc Roemer and Martha Stinson.
- *PRED Social Security Number Validation System Research Project*, August 24, 2004, Planning, Research and Evaluation Division (PRED) Social Security Number Validation System (PVS) Research Team.

The 2003 evaluation tested the PVS using the 1997 Current Population Survey (CPS) that had previously been verified using the SSA's Enumeration Verification System (EVS). The evaluation report summary lists the following results from the study. All statements regarding both PVS and EVS are for the systems in use circa 2003.

- Both EVS and PVS processes exhibit a high degree of accuracy. However, PVS was more accurate and more effective than EVS.
- The PVS provided validated SSNs for more CPS people than the EVS, while reducing error. The PVS identified 80,230 SSNs while the EVS identified 78,218. Two methods of review produced the same accuracy rates among permitted cases (individuals aged 15 and older who did not refuse provision of their SSNs in the CPS instrument). The PVS process had a high degree of certainty 99% of the time, while the EVS was similarly certain only 93% of the time.
- The major enhancement over the EVS process is that the PVS process uses address information to increase accuracy and the number of successes.
- The PVS might be further improved by:
  - treating a missing address differently from a non-matching address
  - relaxing the importance of birth date when name and address match exactly.
- If SSNs were not collected in the CPS, the PVS would identify (search/validate) SSNs for 90 – 92 percent of all CPS adults. This prediction derives from applying success rates among non-refusals lacking a CPS SSN to all adults by completeness of name and birth date. Excluding records for respondents who refused to provide an SSN limits success to 80% in the 1997 CPS.
- Between 1.6 percent and 1.7 percent of the CPS population does not have an SSN. Foreign-born citizens, non-citizens, young people, and females are less likely to have an SSN.
- Identifying the SSNs of children would expand the scope of longitudinal analyses. Census attempted to identify SSNs of all CPS people, while SSA considered only people at least 15 years old. Identifying SSNs for young CPS people creates the ability to link more CPS people to longitudinal administrative data. People younger than 15 in 1997 will appear in administrative wage records in later years. Finding them in these administrative databases will depend on whether their SSN is available.
- A discernible but small amount of income bias appears in the availability of SSNs for people in the March CPS. The results of the EVS and PVS systems are indistinguishable in this regard.

The 2004 evaluation report focused on the methodology and results of the PVS Improvement Project, which identified and researched improvements in the system. The goals and results of the PVS Improvement Project, as described in the project report, were:

- Evaluate the benefit of additional addresses to the GeoSearch phase.
  - Three 2001 surveys, the American Community Survey (ACS), the Survey of Income and Program Participation (SIPP), and CPS were run through PVS before and after additional

addresses were incorporated and changes were made to the match thresholds and duplicate post-processing of the system. The overall match rates increased with the new PVS process; the increase was greatest for the ACS, which did not have the verification step because the ACS does not collect SSN.[5]

- More matches were made in the GeoSearch step, and the matches were more likely to be correct. Thus, the quality of the matches increased for all three surveys.
- The net effect of incorporating additional address sources, implementing tighter comparisons, raising cutoffs, and applying post-processing rules to delete certain assignments was positive.

- Evaluate the quality of the search phase by processing ACS 2001 records assigned an SSN using the PVS through the SSA EVS system and comparing the results.
  - The percentage of records with the same outcome between the two processes was 97.46 percent.
  - A review of the records assigned an SSN by PVS but not assigned an SSN by EVS (2.33 percent of all records) resulted in a recommendation to change the match cutoff parameters for GeoSearch and NameSearch. Additionally, it was found that results were optimized by dropping assignments where the Numident first or last name is a single letter.
  - Reviews of records assigned an SSN by EVS but not assigned an SSN by PVS (0.15 percent of all records) and records assigned different SSNs by each system (0.06 percent) did not result in any recommendations for PVS changes.[6]
- Evaluate the PVS in an environment without SSN.
  - The CPS and SIPP 2001 surveys were processed through the improved PVS with and without the use of respondents' SSNs to simulate the effect of an SSN-less survey environment. The results show a drop of 1 percentage point for the CPS overall match rate and a drop of 6 percentage points for the SIPP overall match rate when only the search phase of PVS was used. The SIPP data used in the 2001 PVS did not include the expected number of within-structure identifiers, which hindered GeoSearch. Disclosure protections perturbed some other required data, hindering NameSearch.

[5] At the time of the study CPS and SIPP asked survey participants for SSNs, but this is no longer the case.
[6] Some exploration of changing matching rules with NameSearch was conducted, but changes that would result in assigning SSNs to records that PVS passed would likely result in too many additional false matches for other records. Furthermore, research on other topics resulted in recommendations to *tighten* the name criteria in NameSearch.

- The SSN-less PVS process nearly always assigned the same SSNs as the SSN-laden process. For the respondents where an SSN was assigned in both PVS processes, the SSN was the same for about 99 percent of the cases for both CPS and SIPP.

- Evaluate the PVS false-match and failed-match probabilities for the search process using a truth deck.
  - The CPS 2001 PVS verified records[7] were considered true matches and formed a "truth deck" of records that were run through the PVS GeoSearch/NameSearch process without SSNs.
    - A review of the validated records from search procedures showed that 0.34 percent were false-matches, that is, the search process assigned a different SSN than the verification process. A clerical review of these records attempted to determine whether the assigned SSN or the verified SSN was correct for each case. Some records matched in GeoSearch were found to be correct, resulting in a revised false-match rate of 0.31 percent. It was difficult to resolve records matched by NameSearch, so the false-match rate from this analysis may be lower.
    - A number of records were not assigned an SSN at all during the PVS search stages. Of all the records in the truth deck, 2.0 percent were failed-matches.
- Seek resolution of the follow two situations: 1) duplicate set – the same SSN is assigned to more than one incoming record; and 2) multiple set – multiple SSNs are found for one incoming record.[8]
  - Duplicate person records may exist in an incoming file, so duplicate sets may be completely legitimate. An algorithm was developed for a post-processing review of duplicate set records to determine which initial set of links to retain. However, results from the verification phase are left as is, and any duplicates created between the search modules, i.e., a source record receives an SSN during Geosearch and another source record receives the sam SSN during the NameSearch, will also remain in the final output file.
  - For multiple sets, the PVS now contains a post-processing algorithm to attempt to select one record from the set.

[7] Recall that for PVS the term verified applies to only those records that have SSNs and are assigned a PIK in the Verification module.

[8] PVS uses different processing rules based on the each customer's needs. For example, the PVS survey version contains algorithms to seek resolution of duplicate and multiple sets, while the PVS federal administrative records version does not. The handling of multiple sets may differ as well. A customer that may ask for all SSNs found for one input record (the multiple set), so that algorithm can be turned off as needed.

The two past evaluations both led to improvements in the PVS matching algorithms, and the Census Bureau Center for Administrative Records Research and Applications (CARRA) staff continues to review and update the system. The system originally used the commercial software product AutoMatch for completing the probabilistic record linkage processes within the search modules. The current system, however, uses a set of SAS® programs, developed by CARRA, for the record linkage process. The Census Bureau has engaged a contractor to add additional modules to the PVS process in order to increase the match rates of the system. NORC's assessment of the system concentrates on PVS effectiveness and quality issues.

1.1.4 Current Assessment's Focus

Given the focus of past evaluations, the current review of the system focused on issues related to the efficiency of the matching algorithm, the quality of the input file, and the coverage of the reference files. Current incoming survey records rarely include SSNs, and neither did the Census 2010 records. Hence, the Verification module tends to be used on administrative record files, many of which contain high quality data, that is, SSN, name and date of birth information that is complete and with few errors. Therefore, NORC focused the PVS assessment on the modules that deal with incoming records that may be harder to match to the reference files: the GeoSearch and NameSearch modules. These modules rely on personal identification information such as name, date of birth and address—some of which may be missing—to determine record matches between two files based on probability models.

In its assessment, NORC conducted the following investigations using the ACS 2009 file as the incoming file. In some instances, Census 2010 records were used as well.[9]

- Match Rate Comparison of GeoSearch and NameSearch Modules
- Unmatched Record Analysis
  - Cut and Blocking Strategy Effects
  - Social, Economic and Demographic Profiles of Unmatched Records
  - Blocking and Matching Variable Missingness Analysis
- Reference File Coverage Assessment
  - Comparison of Unmatched Records Between Incoming Files – ACS 2009 vs. Census 2010
  - Association Between Social, Economic and Demographic Factors and Missingness in Unmatched Records

[9] As previously mentioned (see footnote 8), PVS uses different processing rules for surveys as it does for processing federal administrative records. Therefore, the NORC assessment only pertains to the PVS version used for surveys.

1.2 Comparison of GeoSearch and NameSearch Modules

Using the ACS 2009 file as the incoming file, match and agreement rates of the two PVS search modules were compared. Because ACS is a survey that does not request an SSN from respondents, the PVS process only runs through GeoSearch and NameSearch. Normally, NameSearch matches only the unmatched records coming from GeoSearch, and the match rates for NameSearch, such as those in **Exhibit 1**, are relative to the number of records passed to it from GeoSearch. In order to have more comparable performance metrics, NORC ran the complete incoming file through both modules, providing matching metrics that are both relative to the full set of incoming records.

There are six passes through GeoSearch defined currently for an ACS PVS run. These passes use the first three digits of an address ZIP code (ZIP3) as a database "cutting" strategy. All GeoSearch geographic blocking variables define a subarea of a ZIP3 geographic area for all passes. The GeoSearch matching variables include name and DOB, but also several variables derived from the Geokey (street name, house number, etc).

The NameSearch module, by contrast, does not use any geographic variables for matching. Only the Name and DOB are used to match. There are four NameSearch passes defined for the ACS. All passes use the first characters of the First and Last names to define cuts. All NameSearch blocking variables define a subgroup of these cuts for all passes. The NameSearch uses fewer variables for matching than the GeoSearch, and therefore runs the risk of higher false match rates. To compensate, a higher cut-off threshold for matches is used in NameSearch. But in addition to reducing false matches, the stringent matching criteria also make it more difficult for true matches to pair up in the NameSearch module.

The details of how the PVS uses the two modules, including the fact that NameSearch is intended to work specifically with records unmatched in GeoSearch, is important to keep in mind when comparing records run independently through both modules. **Exhibit 2** shows the matching rates and agreement rates for 4,408,507 ACS 2009 records[10] run through both search modules. An incoming record that matches more than one reference record, called a "multiple set" in PVS documentation, is considered matched for this tabulation. We will use the terminology that a PIK is assigned by PVS, as well as the term "matched" to describe when an incoming record is linked to a PVS reference file record.

[10] The ACS 2009 incoming file contained 4,483,528 records, but 75,021 records were excluded by the initial edit process.

Exhibit 2: ACS 2009 Records Matched by GeoSearch and NameSearch

<table>
<tr><td colspan="3">Matched by Both Modules: 3,330,089 – 75.5%

Modules Agree Completely: 3,310,567 – 71.7%</td></tr>
<tr><td colspan="3">Modules Agree Partially: 150,210 – 3.4%</td></tr>
<tr><td colspan="3">Modules Disagree Completely: 19,522 – 0.4%</td></tr>
<tr><td>Matched by GeoSearch Only:
376,580 – 8.5%</td><td>Matched by NameSearch Only:
375,661 – 8.5%</td><td>Not Matched by Either Module:
326,177 – 7.4%</td></tr>
</table>

We can see each module captures an equivalent percentage of the incoming file that was not captured by the other: 8.5 percent. There are three matched module subgroups—modules agree completely, modules agree partially, and modules completely disagree—because sometimes an ACS record is assigned more than one PIK by one or both of the search modules. If either or both search module contains multiple PIK assignments for an ACS record, then they are considered to agree completely only if they both match all the same PIKs. Otherwise, they agree partially, or if they do not overlap at all, they disagree completely. To clarify, most complete agreements and complete disagreements are not multiple sets. The category "partially agrees" contains most of the multiple sets, which is typically where one search module has multiple PIK assignments for an incoming record, but the other has only one match—and here, the one match will generally agree with one of the multiple matches. **Exhibit 3** provides more detail on the overlap between the matches found in both modules.

Exhibit 3: Records Matched by GeoSearch and NameSearch

Category	ACS Records	Percent
Only one PIK assigned per search module	3,176,220	72.0
Search modules agree	3,160,005	71.7
Search modules disagree	16,215	0.4
More than one PIK assigned by GeoSearch, but only one of the PIKs assigned by NameSearch	24,138	0.5
Modules agree – PIK assigned by NameSearch is at least one of the GeoSearch PIKs	23,989	0.5
Modules disagree – PIK assigned by NameSearch does not match any GeoSearch PIKs	149	0.0
More than on PIK assigned by NameSearch, but only one of the PIKs assigned by GeoSearch	128,806	2.9
Modules agree – PIK assigned by GeoSearch is at least one of the NameSearch PIKs	125,670	2.9
Modules disagree – PIK assigned by GeoSearch does not match any NameSearch PIKs	3,136	0.1
More than one PIK matched in both modules	925	0.0
Complete agreement – All PIKs assigned agree	352	0.0
Partial agreement	551	0.0
Complete disagreement	22	0.0

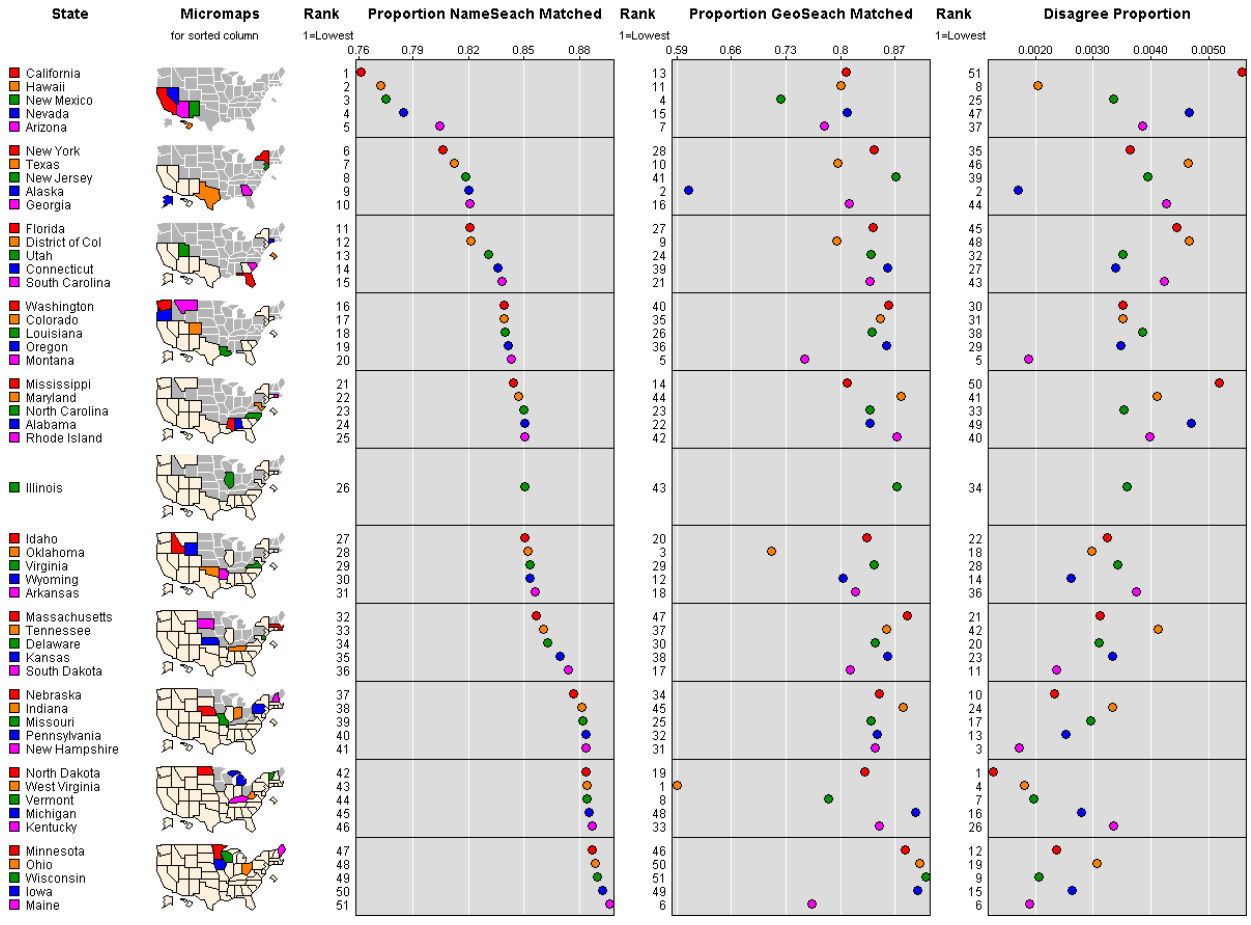
We see in **Exhibit 3** that there are far more multiple matches in NameSearch than in GeoSearch (2.9 percent vs. 0.5 percent). A multiple match is an indication of at least one false match, suggesting the NameSearch module has a higher false match rate than the GeoSearch module. This agrees with past assessments, which have resulted in higher cut-off values for the NameSearch module for precisely this reason.

The "Draft PVS Technical Documentation" (Wagner, 2007) states that PVS survey version contains a post-processing algorithm to attempt to select one record from a multiple set. But any duplicate sets created between the search modules, i.e., an incoming record receives a PIK from GeoSearch and another incoming record receives the same PIK from NameSearch, will remain in the final output file. As a practical matter, we note that most of the multiple sets from the ACS 2009 PVS run end up in the set of unmatched records. The multiple match sets are run through the post-processing algorithm, but this analysis fails to assign one PIK most of the time. Hence, the record is left unmatched.

In order to drill down further into this matter, we look at how each search module performed by geographic and name cuts. **Exhibit 4** is a linked micromap (Linked Micromaps, 2009; Carr et al. 1998)

showing the NameSearch and GeoSearch matched proportion and the disagreement proportion for each of the 50 states and the District of Columbia, sorted by NameSearch matched proportion from lowest-to-highest. There is some positive correlation between the search module match proportions, i.e., as the NameSearch matched proportion increases the corresponding state-level GeoSearch matched proportion generally increases with some exceptions. There is also a slight negative correlation between the match proportions and the disagree proportion, i.e., the disagree proportion tends to decrease as the matched proportion increases. A geographic relationship is also apparent—Southwest states have lower matched proportions than Midwest states. The Western, Southwest, and Southern states are below the median (Illinois) while Northeast, most Mid-Atlantic, and Midwest states are above the median. A further drill-down is needed to understand this better.

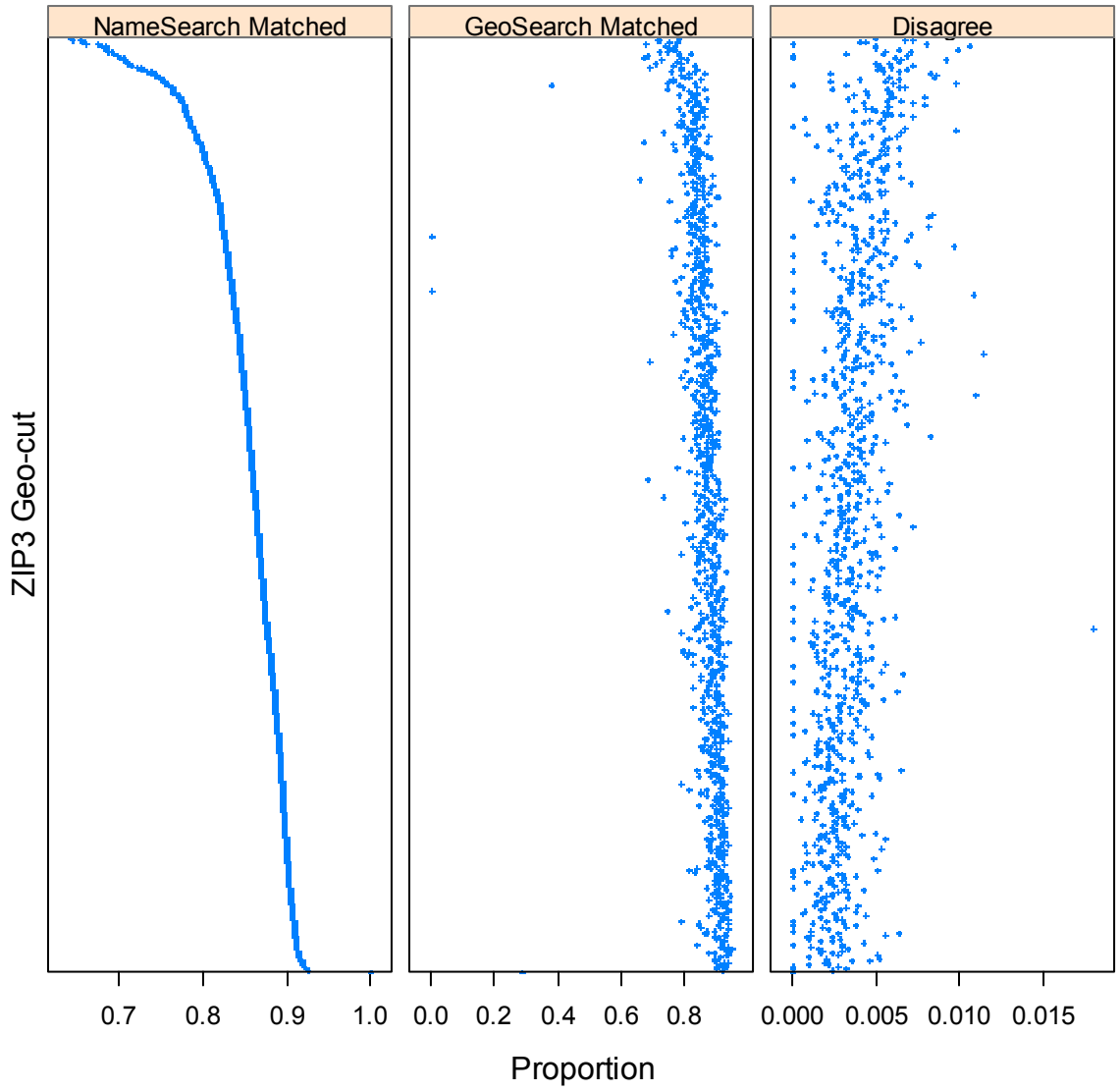
Exhibit 4: ACS 2009 PVS Match Rates and Disagreement Rates by State Sorted by NameSearch Matched Proportion



A similar and perhaps stronger relationship can be seen using a ZIP3 geo-cut comparison level. Because there are approximately 900 populated ZIP3 levels represented in the ACS 2009 data, a linked micromap

view is not feasible. Exhibit 5 is a similar type of graphic in which the plotted points are the matched and disagree proportions for an individual ZIP3 geo-cut. Again, we see that the NameSearch matched proportion and the GeoSeach matched proportion are positively correlated, along with a slight negative correlation with the disagee proportion.

Exhibit 5: ACS 2009 PVS Match Rates and Disagreement Rates by ZIP3 Geo-cut Sorted by NameSearch Matched Proportion



To get a better idea of the states associated with the ZIP3 geo-cuts, the plots in Exhibit 6 show only the lowest 25 and highest 25 ZIP3 geo-cuts based on the NameSearch matched proportion. The plotting symbols are the state abbreviation for the state associated with the ZIP3 geo-cut. The Southwest states

California, Arizona and New Mexico are all represented in the lowest 25 grouping with California ZIP3 regions appearing ten times. This corresponds with what is observed in state-level **Exhibit 4**. However, the Mid-Atlantic States New York and New Jersey also appear several times each in the lowest 25, and Illinois appears once. This partly explains why these states have lower NameSearch matched proportions than their neighboring states. The group of the 25 largest NameSearch match proportions includes Northeast and Midwest states along with one instance each for Kentucky and Wyoming. For Wyoming ZIP3 821, a small ACS sample was selected and all records were assigned a PIK in NameSearch, which resulted in a matched proportion of one. Pennsylvania appears ten times in the highest NameSearch proportion group. This may explain why it has one of the highest proportions for the Mid-Atlantic States.

Exhibit 6: ACS 2009 PVS Match Rates and Disagreement Rates by ZIP3 Geo-cut for 25 Lowest and Highest NameSearch Matched Proportions

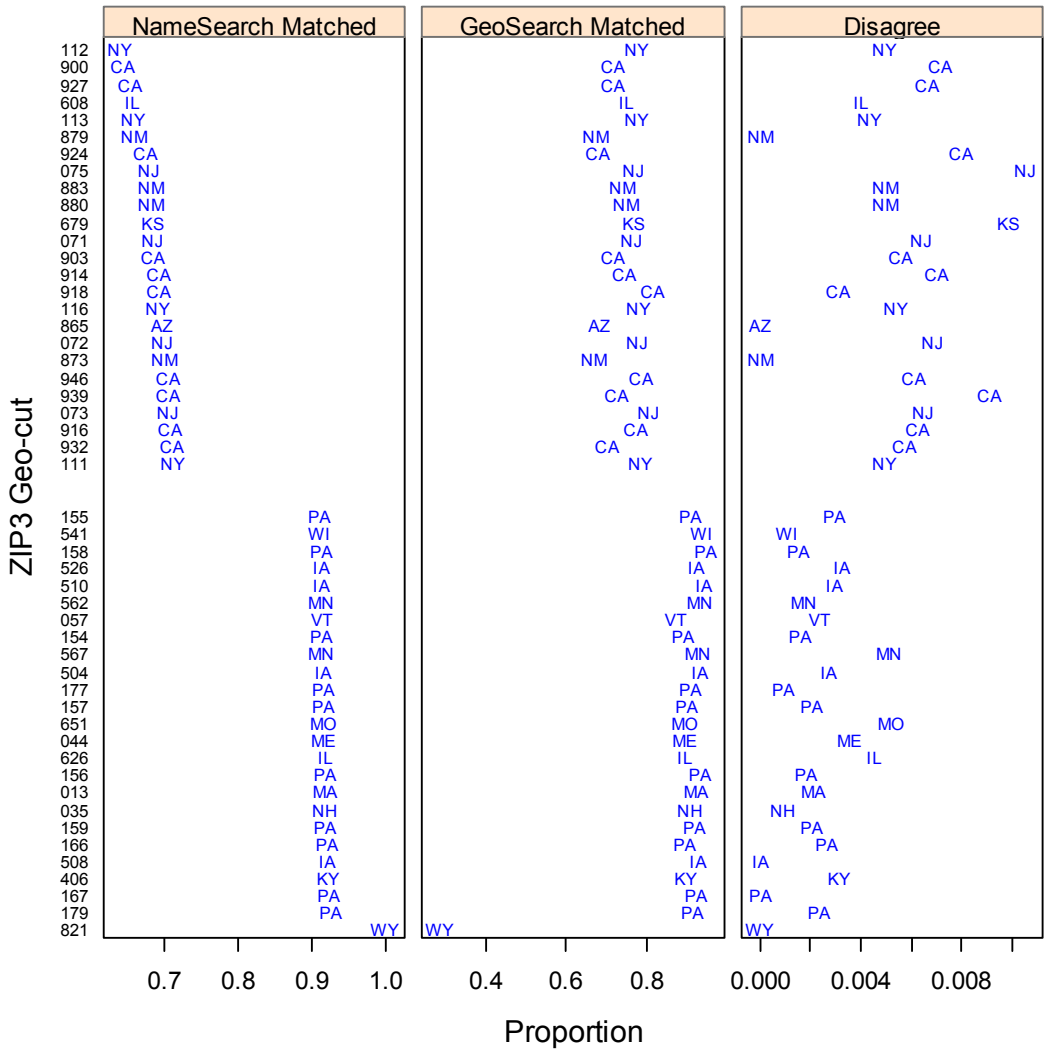
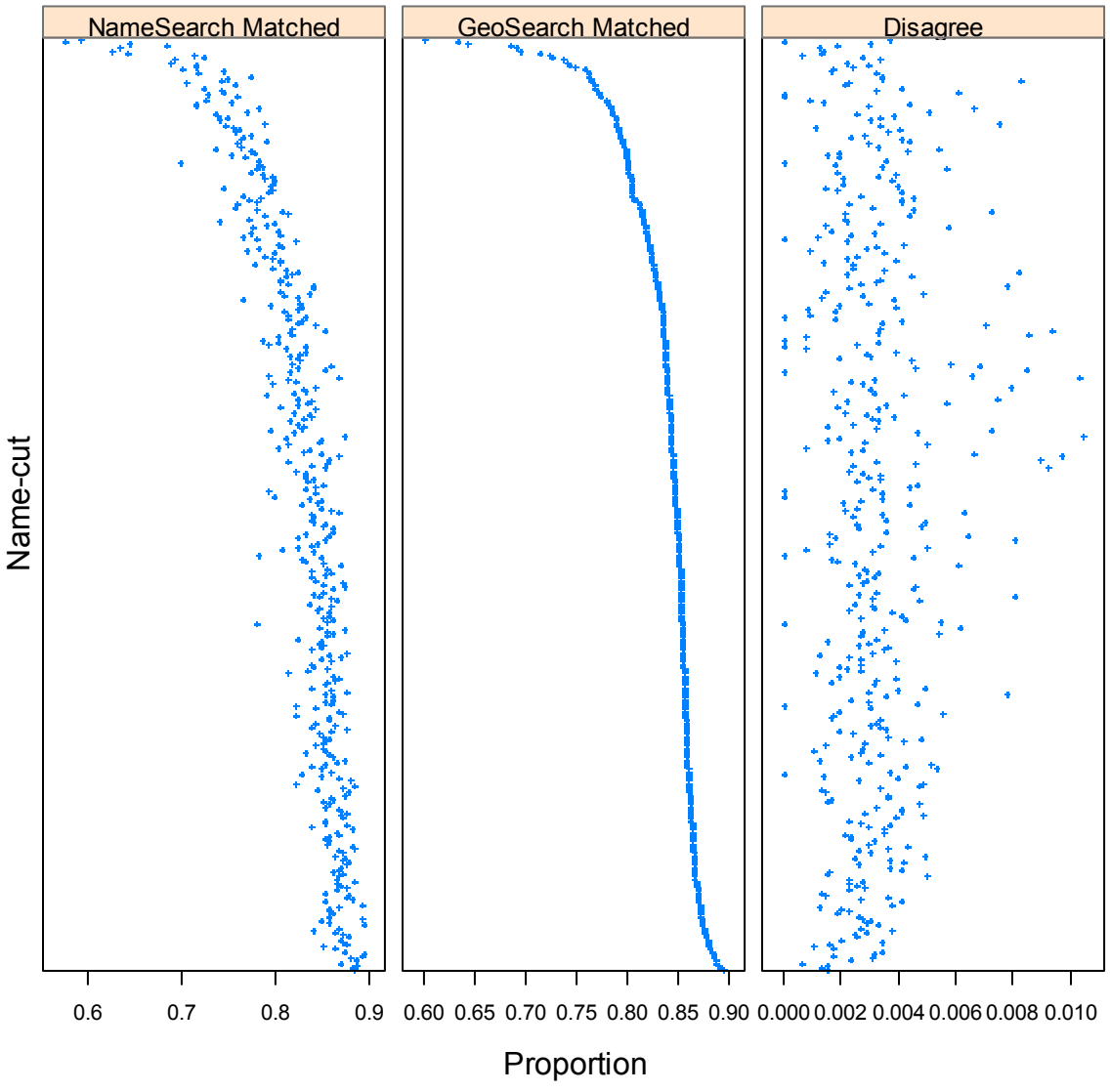


Exhibit 7 is similar to Exhibit 5, but the match proportions are calculated within NameSearch cuts, and these name-cuts are sorted by the GeoSearch matched proportion of ACS records in the cut. Name-cuts are defined by combinations of the first characters of the first and last names. The twenty letter groupings for the first character are: A-or-blank, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, and U-Z. Thus, there are 400 name-cuts used in NameSearch. Again, a positive correlation between NameSearch and GeoSearch matched proportions is noticeable. Negative correlation with the disagree proportions is not as noticeable as it is in the geo-cut plots.

Exhibit 7: ACS 2009 PVS Match Rates and Disagreement Rates by Name-cut Sorted by GeoSearch Matched Proportion



To get a better focus on the name-cut categories with the lowest and highest GeoSearch match proportion, we use a variant of the micromap plots where the geographic map of the U.S. is replaced by the name-cut map shown in **Exhibit 8**. The map is a matrix grid where each row represents a category for the first character of the first name and each column represents a category for the first character of the last name. The cell in the first row and first column is the name-cut with "A-or-blank" first names and "A-or-blank" last names.[11]

Exhibit 8: Name-cut Map

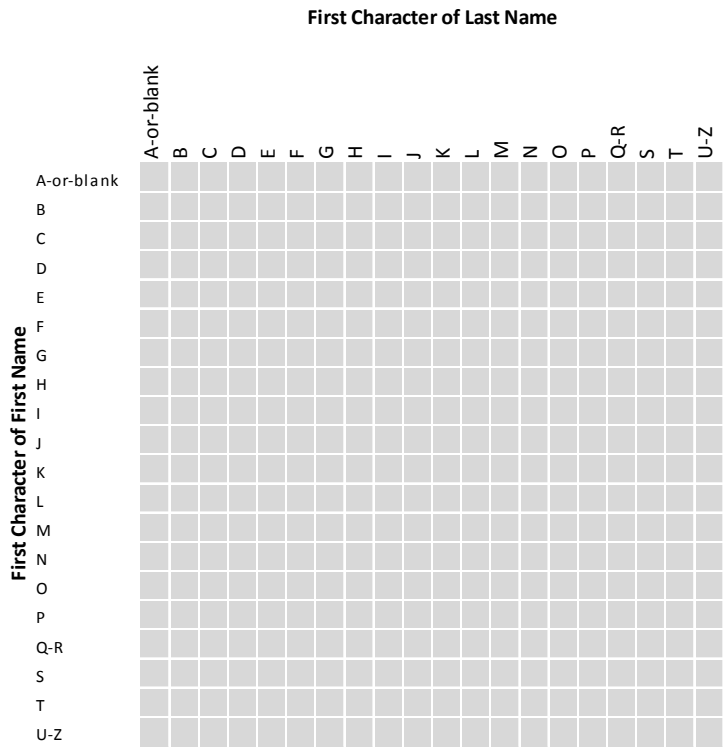
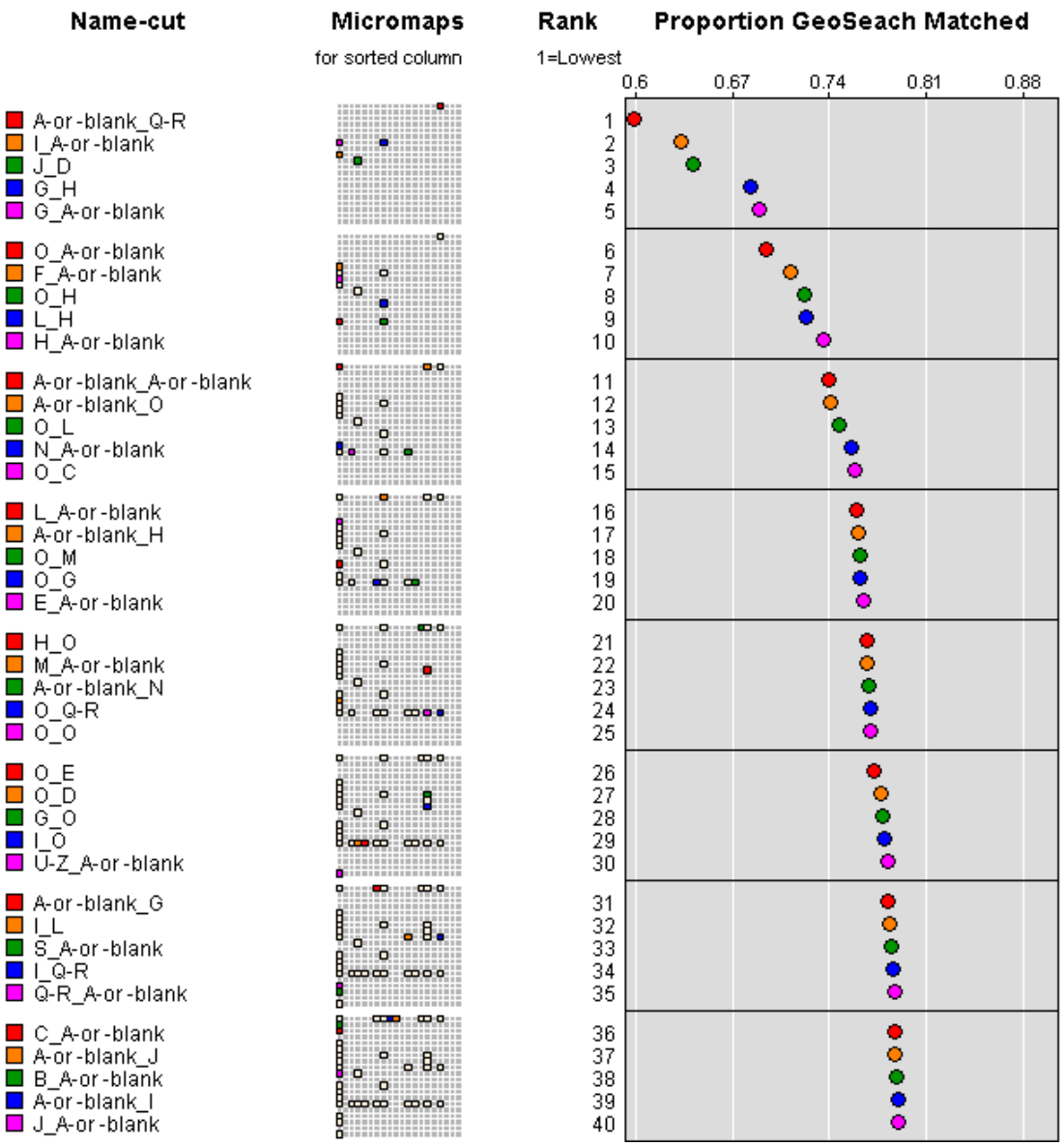


Exhibit 9 is a linked name-cut micromap for the name-cuts with the 40 lowest GeoSearch matched proportions. Only the GeoSearched matched proportion is shown to allow better focus on the name-cut maps. The labels for the name-cuts are constructed by concatenating the letter group name for the first and last name with and underscore, "_" in-between: name-cut "B_C" includes people with a first name starting with a B and a last name starting with a C. This particular name-cut is the cell at the intersection of the second row and third column in the name-cut map.

[11] The initial edit process, described in the **Introduction: PVS Background** section, removes from consideration incoming records that have no name data. Therefore, no record that is processed in PVS has blank first and last names. The name-cut "A-or-blank_A-or-blank" would only include records where the first name is blank and the last name begins with an A, or where the first name begins with an A and the last name is blank.

Exhibit 9: ACS 2009 PVS GeoSearch Matched Proportions Micromap by Name-cut for the 40 Lowest GeoSearch Matched Proportions



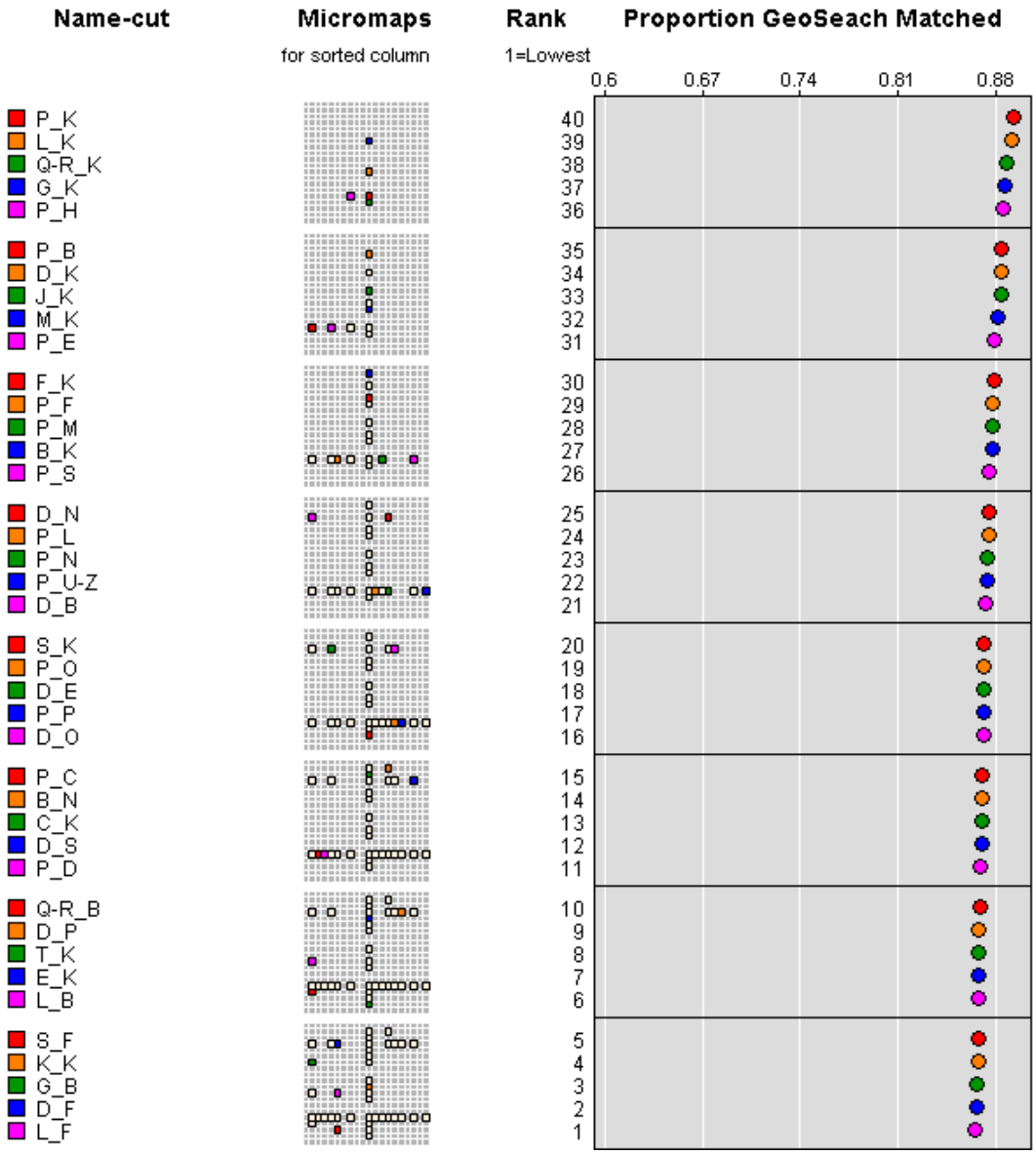
Patterns in the name-cut maps indicate that the name-cuts with some of the lowest GeoSearch match proportions are those with the first character of the last name "A-or-blank." Those name-cuts with the first character of the first name "A-or-blank" also are among the 40 lowest matched proportions, but almost all name-cuts where this is true for last names are part of the 40 lowest GeoSearch matched proportions. This would also be generally true for NameSearch matched proportions because of the positive correlation between the two as indicated in **Exhibit 7**.

It is understandable that names with a blank first or last name would be associated with lower match proportions. **Exhibit 9** also reveals that name-cuts with a first name starting with the letter "O" have a lower GeoSearch matched proportion than most other name-cuts. Those name-cuts with a last name starting with the letter "O" also begin to appear in the second half of the display. While the cause of this is unknown, we suspect it occurs because of the use of filler fake or incomplete names for survey responses. Sometimes a person's name might be filled in as "Occupant" or "Owner" and this would very likely not match a record in the reference file. Variants of the expression "of the house" may appear in a fake name, and the name standardization software used in the initial edit process might parse this into the first or last name. More investigation is needed to understand why name-cuts with first or last names starting with "O" have low match proportions.

Exhibit 10 is similar to **Exhibit 9**, but shows the name-cuts with the 40 highest GeoSearch matched proportions. The patterns in the name-cut maps indicated that last names starting with the letter "K" and first names starting with the letter "P" have the highest match rates within GeoSearch. Again, this would also be generally true for NameSearch match proportions because of the positive correlation between the two (see **Exhibit 7**). First names beginning with the letter "D" also have high GeoSearch matched proportions.

The comparison of GeoSearch and NameSearch analysis suggests further research into why some regions of the country and some first/last name name-cut combinations have lower match proportions than others. In subsequent sections, we look at issues related to the poor quality of the name information of an incoming record, including fake/incomplete first and last names. We suspect that this may be one of the issues related to the difference in matched proportions, but more research is needed to form conclusions.

Exhibit 10: ACS 2009 PVS GeoSearch Matched Proportions Micromap by Name-cut for the 40 Highest GeoSearch Matched Proportions



We mentioned above that some of the conclusions reached as a result this exercise may be related to false-match and failed-match probabilities. Winkler (2010) points out that "…the general problem of error rate estimation (both false match and false nonmatch rates) is likely impossible in situations without training data and exceptionally difficult even in the extremely rare situations when training data are available." Consequently, without a truth deck for each incoming file, as was used in the 2004 PVS

Improvement Project, it would appear to be exceptionally difficult work to determine error rate estimates for every PVS run.

However, the PVS program is very "rich" in information about record linkage conducted over a number of years. A model based on a number of factors—the type of incoming file (survey, census, and administrative records), data collection year, search module matched proportions, disagree proportion, name field missingness measures, etc.—and the false-match/failed-match rates from incoming files that can produce a truth deck—because an SSN is available—could be constructed to help estimate error rates of PVS output. A Bayesian hierarchical model for estimating error rates may be possible to construct for PVS error rates. We discuss this more in the **Recommendations** section.

1.3 Unmatched Record Analysis

NORC reviewed the ACS 2009 unmatched records to understand what may be causing the failure-to-match. This section describes three investigations. The first is focused on investigating the cutting and blocking strategy as a source of unmatched records. The second investigation is a descriptive statistical review of unmatched ACS 2009 records in terms of certain social, economic, and demographic factors. By profiling the unmatched records in this way, we gain a better understanding of the characteristics of records that do not get an assigned PIK. The third investigation considers the quality of the incoming records used in the matching process. That is, we investigate the "missingness" within the blocking and matching variables, and we explore the contribution to the unmatched percentage of the degree of missingness of an incoming record.

1.3.1 Cut and Blocking Strategy Effects

NORC examined removing the "cut" as a blocking variable for the unmatched ACS 2009 records. For this analysis, the ACS records which failed to match within either the GeoSearch or NameSearch were run through the PVS without blocking by module cuts. The 326,177 records which failed to match in either search module (see **Exhibit 2**) were run through the PVS matching against all 1,000 geo-cuts and 400 name-cuts. This analysis is intended to determine the number of matching opportunities lost due to the module cut definitions.

The results indicate that very few additional matches can be found outside *both* the geo- and name-cuts. Specifically, an additional 4,054 matches were found in GeoSearch and 470 in NameSearch.

Many of the additional matched found in GeoSearch are not good. The parameter file[12] for GeoSearch is set up assuming the first three characters of a ZIP code match. The cutoffs are set under this assumption, so when matching outside a geo-cut, the cutoffs should be raised to reduce the number of bad matches.

Regardless, there are some good matches. Most commonly, among the GeoSearch matches that appear to be good, when either the incoming or the reference record has a ZIP3 of "000." This indicates that the ZIP code was either missing or not good. There are 1,003 matches for which the geo-cut for one source (either the ACS file or the reference files) is defined as "000," but the geo-cut for the other source is based on a valid ZIP code. Of these, 797 match exactly for the first 53 characters of the Geokey (all characters prior to the ZIP code). All address information is run through the commercial address standardizer CODE1 for both the reference file and incoming file records. Therefore, it appears that a ZIP code was not assigned during this process in one of the sources for these 797 records.

In fact, of the 127,246 ACS records that are in Geo cut "000," only 81 were matched in the original PVS run to records in the "000" reference file geo-cut. Most ACS records in this cut are matched in the NameSearch module. It is unlikely that the ZIP code is missing for both the reference file and the incoming file. The match for an ACS record in geo-cut "000" is likely found in a reference file geo-cut for valid ZIP codes. Close to 3 percent of all ACS records are in geo-cut "000". In contrast, less than one-tenth of one percent of the records in the reference files is in geo-cut "000."[13] Missing or poor zip codes for a portion of addresses appear to be an incoming survey file quality issue that may not be correctable for incoming files such as the ACS.

Additionally, there are some good matches from GeoSearch in what appear to be ZIP code keying errors—for example, "773" instead of "778". If ZIP code is verified for the reference and/or incoming files for a given address, then some are apparently missed.

For the NameSearch module, cuts are defined using the first character of the first name coupled with the first letter of the last name. When an error occurs for the first letter of either part of the name, records are placed in the wrong cut and fail to match. William Kaplin and William Caplin is a typical (but fabricated) example. The 470 additional matches found in the name module all appear to be good—all match exactly on date of birth. Most, 453, are interchanges between C and K (denote as C↔K), which put the name in

[12] A file that includes a number of parameters needed for running PVS. This includes the cut-off for match scores. A PIK is assigned when the match score between an incoming record and a reference file record is above the cut-off.
[13] The percentage of reference file records in geo-cut "000" is based on the reference files found in the directory /geokey2/stars09_ssr09_cn09.

different incoming and reference file cuts. The other problem letter interchanges are W↔R, K↔N, and P↔F.

Overall, running the ACS unmatched records against all cuts in both modules yielded about 2,000 good matches, or two-thirds of one percent of the unmatched file. This is a measure of matches lost due to the cutting scheme. This is a very small percentage of the unmatched records. Understandable, given that to fall in this category, a record has to coincidentally fail to be in the right cut for both modules. We see this happens, though, in a few set of predictable cases—when one file has a ZIP3 of "000", and/or when the first letter is a common C↔K interchange. NORC believes that the PVS, as it currently stands, could be adjusted for these two common ACS 2009 cases with a small amount of effort. However, before a system change is made, examination of other incoming files is needed to see if this situation is found elsewhere.

1.3.2 Social/Economic/Demographic Profile of Unmatched Records

NORC investigated whether the unmatched records from the ACS 2009 are associated with social, economic, or demographic factors that are important in social, economic and public policy research. For example, in a study examining the misreporting of Food Stamp Program (FSP) benefits in Maryland and Illinois, Meyer and Goerge (2010) linked administrative data with ACS and CPS 2001 data using PVS assigned PIKs. They found a need to correct for possible bias due to the fact that PIK assignment rates for the ACS and CPS records were lower for those who are likely food stamp recipients. This type of under-coverage has implications for research that relies on linking data between administrative records and survey information using PIKs assigned by PVS.

To better understand the issue, we considered the distribution of unmatched records[14] across states to see if there is an association with geographic location, which is depicted with a linked micromap in **Exhibit 11**. The unmatched percentages vary from 4.0 percent to 14.3 percent. The overall percentage of all ACS 2009 unmatched records is approximately 7 percent. The five states with the highest percentage of unmatched records are the Southwest states New Mexico, California, Nevada, and Arizona, along with Alaska. Midwest states Michigan, Minnesota, Wisconsin, Iowa, and Ohio have the smallest unmatched percentages, suggesting a regional effect. The regional pattern is similar to that seen in **Exhibit 4** for the NameSearch matched proportion. For that exhibit, all ACS 2009 records were run through NameSearch, whereas now we are looking at the unmatched rates for the PVS production process—GeoSearch

[14] Here, unmatched records are defined as those assigned a Verification and Search Flag (VERFLG) values of "A." A record with this flag went through the GeoSearch and NameSearch modules but was not assigned a PIK. Records that match to multiple reference file records are not included in this group even though they can be considered unmatched.

followed by NameSearch of only those records unmatched by GeoSearch. The fact that the regional effect is similar in both situations suggests that name characteristics play a big role in the overall PVS probabilistic matching process.

We also included the PVS unmatched proportion for Census 2010 Decennial Response File (DRF) records in **Exhibit 11**. There is a positive correlation between the two sets of unmatched proportions based on state. The set of five states with the highest percentages for the Census 2010 output retains the four Southwest states as for the ACS 2009, with the inclusion of the District of Columbia instead of Alaska.

Exhibit 11: PVS Unmatched Proportion by State: ACS 2009 and Census 2010 DRF Sorted by ACS Unmatched Proportion

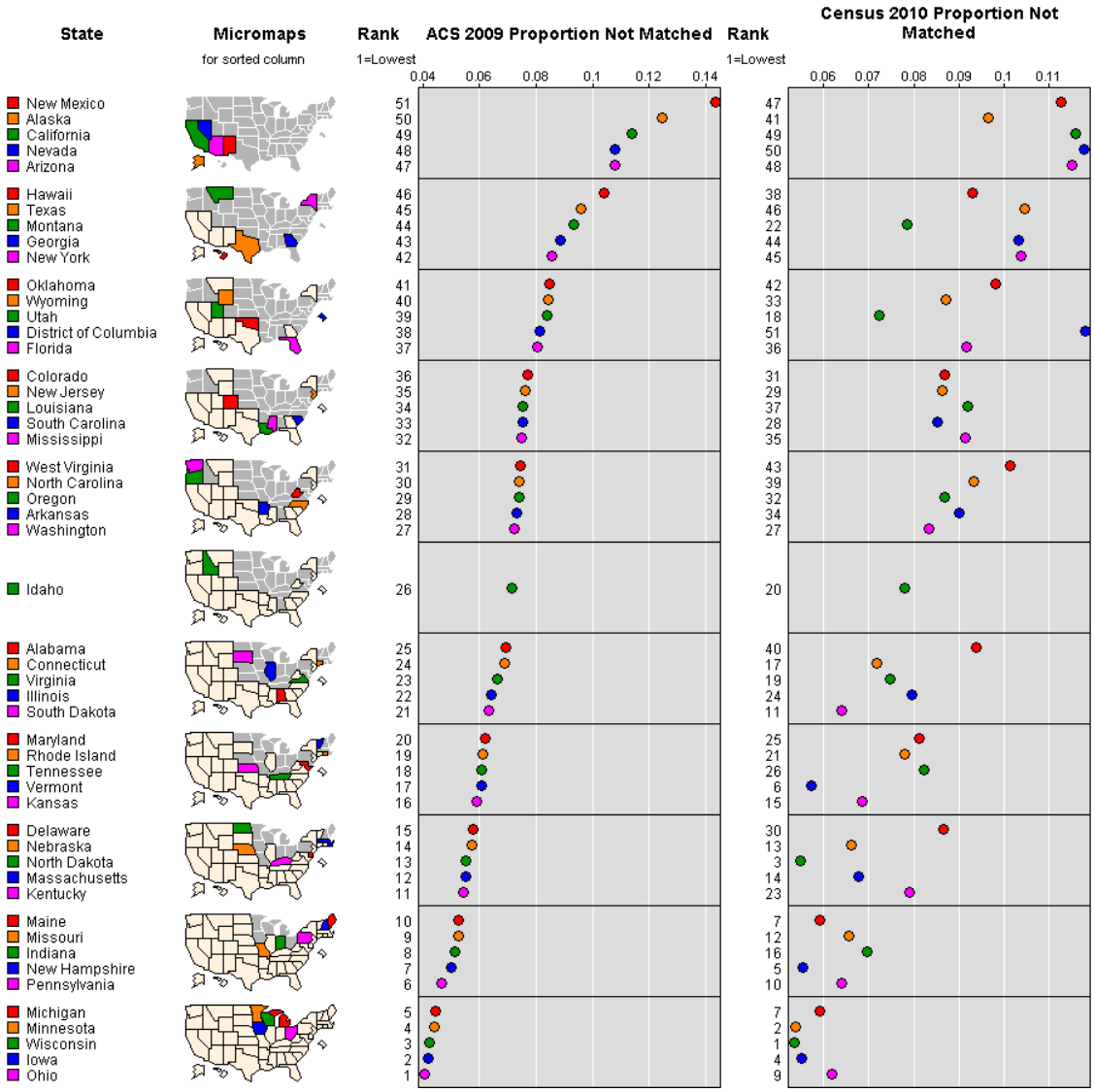


Exhibit 12 compares the percentages of selected social, economic, and demographic characteristics based on all the ACS 2009 records and only unmatched records.

Exhibit 12: ACS 2009 Social, Economic, and Demographic Characteristics†

Characteristics	All Records	Unmatched Records
Selected Social Characteristics		
U.S. Citizenship Status		
Not a U.S. Citizen	5.1%	25.3%
Language Spoken at Home		
Language other than English	15.8%	35.0%
Educational Attainment*		
High School Graduate or Higher	84.15%	62.21%
Bachelor's Degree or Higher	27.91%	16.50%
Selected Economic Characteristics		
Employment Status		
Unemployed**	9.09%	12.5%
Income and Benefits		
Median Household Income	$ 61,276	$ 47,318
Food Stamp Recipients	10.8%	16.3%
Health Insurance Coverage		
No Health Insurance Coverage	12.6%	30.4%
Poverty Status		
Below the Poverty Level	11.7%	23.9%
Selected Demographic Characteristics		
Age		
Less than 35	42.4%	59.7%
35 years and above	46.8%	36.9%
Hispanic Origin		
Hispanic	12.1%	32.3%
Race		
Non-White	19.9%	33.8%

† Results are based on unedited ACS 2009 records. Therefore, no imputed values were used and percentages are based on records with nonmissing values for the characteristic of interest.
* Educational attainment percentages are calculated as a percentage of persons aged 25 years and above.
** Unemployment percentages are calculated as a percentage of total civilian labor force.

The numbers suggest that the unmatched records are different than the full set of ACS records in terms of socioeconomic and demographic composition. We can see that the percentages for non-US citizens,

people that speak a language at home other than English, the unemployed,[15] the uninsured, those below the poverty level, and food stamp recipients are higher for the unmatched records as compared to the overall set of records. Also, the percentage of people with at least a high school education and people with at least an undergraduate college education (bachelor's degree or higher) are lower for the unmatched records.[16] The demographic composition of the unmatched records is also different than that of all records; the unmatched group has a higher percentage of those less than 35 years of age, those of Hispanic origin, and Non-white people. These results are similar to those noticed by Meyer and Goerge for the 2001 ACS and, as noted, may bias research that uses PVS PIKs to link together databases of interest.

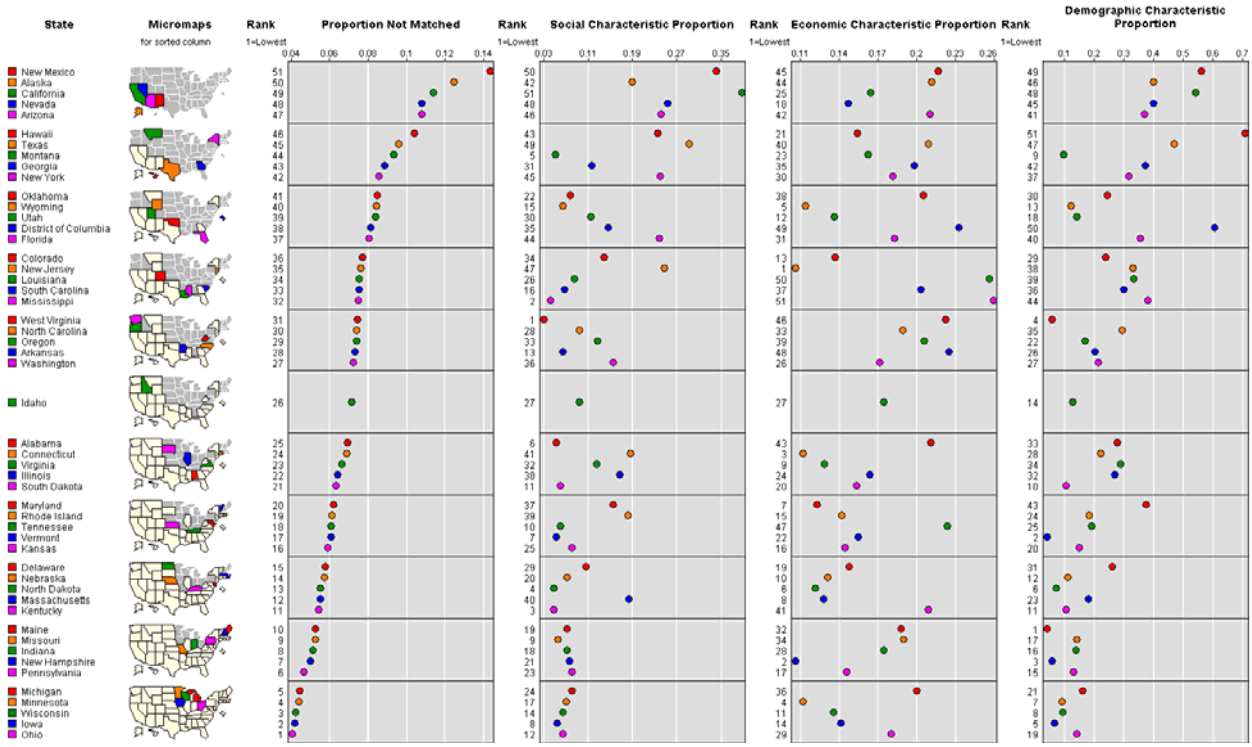
To examine this more closely, we constructed three composite variables to represent some of the characteristics in **Exhibit 12** within the social, economic, and demographic groupings. These characteristics are defined as follows using the ***self-reported*** information in the ACS 2009.

- Social Characteristic – a person who is a non-English speaker at home or a not a U.S. citizen
- Economic Characteristic – a person whose income is below the poverty line or is a food stamp recipient
- Demographic Characteristic – a person that is either non-white or Hispanic

Exhibit 13 is a linked micromap comparison of the ACS 2009 unmatched proportions and the proportion of ACS 2009 records with the social, economic and demographic characteristics of interest by state. There is some correlation between the social characteristic proportion, the demographics characteristic proportion, and the unmatched proportion. Thus, there appears to be an association between the social and demographic characteristics of interest and the unmatched proportion. It is not clear from this plot that the economic characteristic is correlated with any of the other factors. In which case, the unmatched proportion may not be affected by the economic characteristics that we have chosen to consider at the state level.

[15] Unemployment percentages are calculated as a percentage of total civilian labor force.
[16] Educational attainment percentages are calculated as a percentage of persons aged 25 years and above.

Exhibit 13: ACS 2009 Unmatched Proportion and Social, Economic, and Demographic Characteristics by State as Reported in the ACS 2009 Sorted by ACS Unmatched Proportion



One possible reason that records of persons in these social, economic and demographic groups do not match to the PVS reference files is that corresponding records may not exist in the reference file. But, another reason could be that the data quality of the records for people in these groups is low, i.e., there is a high degree of missingness in key blocking and matching variables. We look at how missingness relates to the unmatched records in the next section. An analysis of an association between social, economic and demographic characteristics and missingness is discussed in the **Association between Socio-economic Factors and Missingness in Unmatched Records** section.

1.3.3 Blocking and Matching Variable Missingness Analysis

In this section, using unmatched ACS 2009 records, we examine the quality of different variables used directly in matching the input file with the reference files. Date of Birth (DOB), Geokey (street name, street name prefix and suffix, house number, rural route and box, and ZIP code), and Name (first name, last name, middle name, middle initial, suffix) are used in the search modules, and are highly related to whether or not records in the incoming and reference files match. As with the socioeconomic and

demographic variables, we limited this investigation to three variables: DOB, ZIP code and fake or incomplete names.

Both the GeoSearch and NameSearch modules use DOB to match records,[17] and the DOB can be broken into month, day and year components. A DOB can be partial in that some of the components may be missing, while others are present. Because records with completely missing DOB are likely to have the greatest impact on the match rate, we focused on those records.

For Geokey variables, it is not straightforward to judge the quality of incoming records in terms of a missing percentage. An address does not usually contain all of the information allowed for in the Geokey. For example, a rural route address usually does not have a street name and other street related characteristics, and a city style address does not include rural route information. The original address provided in the ACS 2009 is run through an address standardization algorithm to parse the address and form the Geokey. In situations where the ZIP code is missing, it may be the quality of the address was not good enough to determine a ZIP code. Therefore, as a proxy for Geokey quality, we looked at whether or not the ZIP code was missing.

For the name variables (first name and last name), **Exhibit 9** indicates that records with missing (blank) first or last name have higher unmatched proportions. But the graphic also indicates that names that begin with letters, such as "O," also have high unmatched proportions. This may be due to fake or incomplete names that are used to fill-in a survey response when a respondent wishes to remain anonymous. There is a PVS name-editing step, which attempts to remove fake names (see **Appendix B**). The fake names that are caught in this step are set to blank, and this may cause a record to be removed from the PVS process in the event that both the first and last names are blank. This name editing step may not set all fake names to blank because of various spellings (or misspellings) of the fake names.[18] Additionally, the fake name reference list may not include some that were used in the ACS 2009 file.

Using the unmatched ACS 2009 records, we constructed a list of possible fake names or incomplete names that may have been missed, or where not set to blank, by the PVS name-editing step. This list is provided in **Appendix B**. We have included certain cases where only an initial is used for a name

[17] NORC understands that the Census Bureau is working on a DOB-based search module for PVS. Early experiments with the model indicate that an additional set of incoming records will be matched. We do not know at this time whether this module will overcome some of the issues related to missing DOB information.

[18] The PVS name editing step program (/pvs/pvs/code-template/ver-4/pbde_pvs-name-edit-macro.sas) attempts to remove fake names, which are defined in the datafile /pvs/pvs/code-template/ver-4/pbde_fakenamelist.dat. John and Jane Doe are allowed to stay, but the matching requirements are stricter. But Baby Doe, Boy Doe, Girl Doe, are all set to blank.

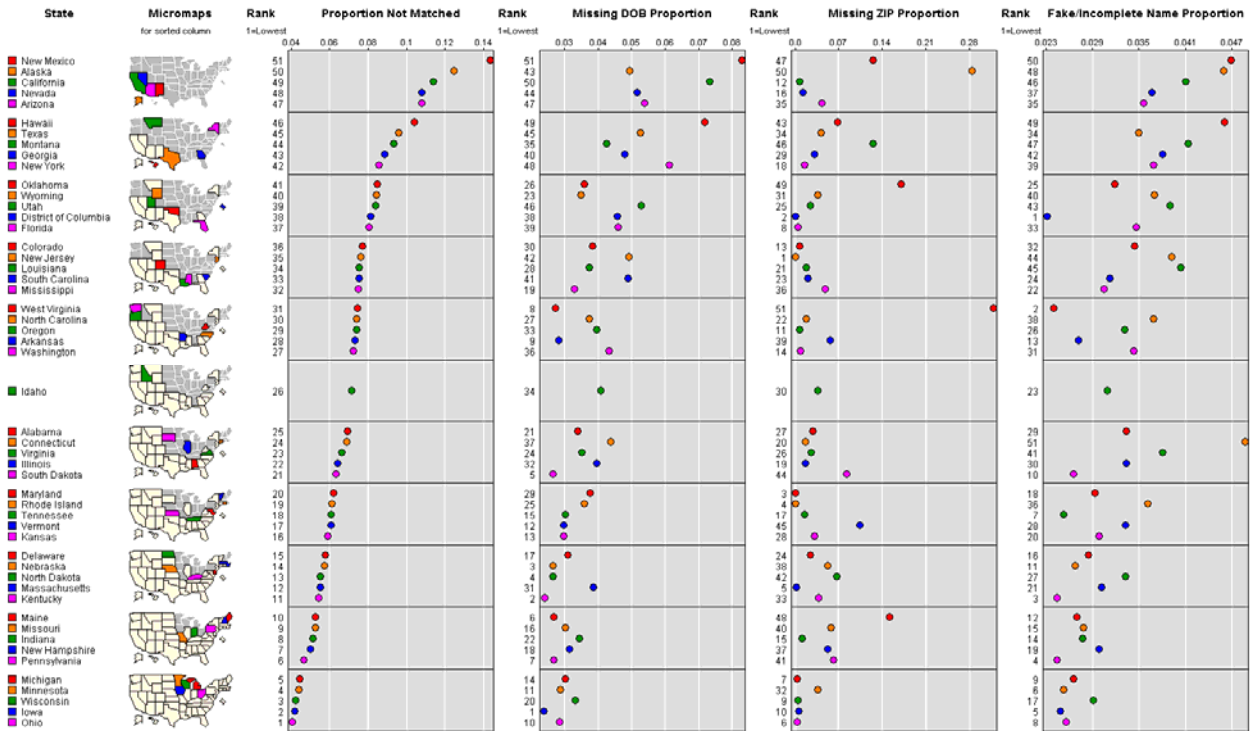
(incomplete name). For first and last name, instead of categorizing the variables as missing and non-missing, the quality of information categorization is 'real' or 'fake/incomplete.' Fake/incomplete includes cases where the first or last name is completely missing.

Overall, for the missingness of matching and blocking variable factors, we consider the following characteristics.

- Missing DOB – a record with completely missing information DOB
- Missing ZIP Code – a record with no ZIP code in the Geokey
- Fake or Incomplete Name – a record that has a fake/incomplete first or last name found in the NORC generated lists in **Appendix B**; this includes records with blank first or last names

Exhibit 14 is a linked micromap comparison of the ACS 2009 unmatched proportions and the proportion of ACS 2009 records with missing DOB, missing ZIP code, or fake/incomplete names. There is some correlation between the missing DOB proportion, the fake/incomplete names proportion, and the unmatched proportion. This is expected, as these are key blocking and matching variables in the PVS process. When the quality of this information is poor, it becomes hard to match records between the incoming and reference files. It is not clear from this plot that missing ZIP code is correlated with any of the other factors. While ZIP code plays a key role in the blocking and matching within GeoSearch, it plays no role in NameSearch. As was noted in the **Cut and Blocking Strategy Effects** section, many of the ACS 2009 records in the "000" geo-cut, which includes all cases with a missing ZIP code, are matched in NameSearch. Therefore, missing ZIP code does not have a substantial impact on the match rate of the full PVS process.

Exhibit 14: ACS 2009 Unmatched Proportion and Missing Characteristic Proportions by State Sorted by ACS Unmatched Proportion



1.4 Reference File Coverage Assessment

As has been mentioned above, one reason that some of the ACS records are unmatched could be that a person represented by a record does not have a corresponding record in the reference files. The current PVS reference file is built from the SSA Numident file, IRS data, and other federal administrative records data. These administrative records may not include a number of people residing in the U.S. For example non-U.S. citizens[19], children and people not in the work-force may not be adequately covered by the administrative records used for the reference file. NORC investigated two issues that are related to the coverage of the reference file.

[19] NORC understands that the Census Bureau has undertaken an effort to enhance the PVS reference files with IRS files that include Individual Taxpayer Identification Numbers (ITIN). For those people who are required to file a tax return but do not have, and may not want an SSN—such as a non-U.S. citizen—the IRS issues the taxpayer an ITIN. This enhancement to the PVS reference file may help to match more non-U.S citizens.

First, we attempted to match the unmatched ACS 2009 records to the Census 2010 DRF unmatched records with the idea that if there is a large overlap then certain records are not getting assigned PIKs from one PVS run to the next. If so, it could be these person records are not in the reference file. Second, we investigated whether there is a strong association between social, economic and demographic variables and missingness of blocking and matching variables. If there is no noticeable association, then it might be that the records of people with characteristics listed in **Exhibit 12** do not have records in the PVS reference files.

1.4.1 Comparison of Unmatched Records between Incoming Files – ACS 2009 vs. Census 2010 DRF

NORC set up a PVS run to compare the unmatched ACS 2009 records to the unmatched 2010 Census records. The parameters of the run were based on the current parameters used to match ACS records to the reference files. The unmatched Census 2010 DRF records were considered the reference file in this exercise. No cutting strategy was used—the complete ACS unmatched file was compared to the complete unmatched Census file. The run was set up in five passes: for GeoSearch, passes 1, 2, and 6 currently used for the ACS matching; and for NameSearch, passes 1 and 2 currently used for the ACS matching. These are the most successful passes in the respective modules.

The purpose of this investigation is an indirect assessment of the reference files. If unmatched records in different incoming files match each other using PVS, when they failed to match to the reference files, this suggests the individuals are missing or at least defined differently in the reference files. For many survey files this idea would not work because each survey file is likely an independent sample of residents. The chance of a sizeable overlap between the two complete files would be small, and smaller still for the sets of unmatched records. But because the census file includes all the individuals enumerated in the U.S., the complete ACS file should be almost fully contained in the census file.[20] In which case, the ACS records that are unmatched due to a lack of coverage in the PVS reference files would match to a subset of census records that are unmatched records due to a lack of coverage in the PVS reference files. Although the collection timeframe of for ACS 2009 is different than that of Census 2010, we felt it close enough that the matching exercise would still provide insight into the reference file coverage issue.

[20] The ACS is a series of monthly samples used to produce annually updated data, whereas the 2010 Census is an enumeration of people at their "usual residence" as of Census Day (April 1, 2010). Thus, there is a possibility that the ACS 2009 and Census 2010 would record different information for a person, especially for large time-gaps between responses, e.g. an ACS response record in January 2009 versus the April 2010 Census response.

Exhibit 15 is a summary of this special PVS matching run. There is an overall match rate of 13 percent, hinting there is some small degree of under-coverage in the PVS reference files.

Exhibit 15: Summary of Matches between Unmatched Census 2010 DRF and ACS 2009 Records

Category	ACS Records	Percent of ACS Records
Total Unmatched ACS Records	328,364	100.0
Matched to the Census 2010 DRF Unmatched Records (including ACS records with more than one Census Record match)	43,223	13.2
Duplicate Matches (ACS records with more than one Census Record match)	6,704	2.0
Non-duplicate Matches (ACS records with only one Census Record match)	36,519	11.1
Non-duplicate Matches matched in GeoSearch pass 1	26,811	8.2
Non-duplicate Matches matched in GeoSearch pass 2	3,165	1.0
Non-duplicate Matches matched in GeoSearch pass 6	3,674	1.1
Non-duplicate Matches matched in NameSearch pass 1	2,296	0.7
Non-duplicate Matches matched in NameSearch pass 2	573	0.2

While overlapping records in input files may point to under-coverage in the reference files, the duplicate matches in the input files could point to quality issues with the records in both files. The 6,704 duplicate matches (see Exhibit 15) were examined more closely. Exhibit 16 provides a frequency distribution of these duplicate matches.

Exhibit 16: Frequency Distribution of Duplicate Matches

Number of Census DRF Record Matches	ACS Records	Percent of Duplicate Matches
2	5,668	84.5
3	782	11.7
4	157	2.3
5	35	0.5
6 to 10	43	0.6
11 to 20	4	0.1
21 to 50	8	0.1
51 to 86	7	0.1

Some records have a very high duplication rate—one ACS 2009 record, for example, matched 86 Census 2010 DRF records. Looking closely at these high-ranking duplicates, they are seen to be in a few institutions where groups of people share a permanent home, for example, mental institutions and state prisons. These are cases where a large number of individuals have that same exact Geokey. There can be additional factors preventing unique matching within these institutions because the name and DOB may be intentionally inaccurate. There are a large number of "John" and "Jane" "Doe" records in these institutions. Some prisons appear to commonly use only last names, while putting what appears to be a one-character code in the first name field instead of a real name. Also, it appears common to only use a decade for DOB instead of a real date (there are many DOBs which are 1/1/1970, 1/1/1980, etc). These duplicates and multiple sets ultimately become unmatched in the PVS because a true match cannot be resolved. This may be a situation without a good resolution. Survey and census records would need to use real names for institutionalized individuals for the records to match to reference file records.

While 13 percent is an overall match rate between ACS 2009 unmatched records and Census 2010 DRF unmatched records, this may provide an incomplete picture of records missing from the reference files. Only unmatched records were compared. A more complete analysis would be to compare the entire ACS input to the entire Census input. We can then look at numbers in the 2-by-2 table: Matches-Each-Other vs. Matches-Reference-Files. A final improvement to this analysis is to extend the comparisons to other surveys besides just the ACS, particularly surveys that took place in 2010.

1.4.2 Association between Socioeconomic/Demographic Factors and Missingness in Unmatched Records

NORC's final investigation for the PVS assessment explores whether there is an association between socioeconomic and demographic characteristics and the missingness of key blocking and matching variables in the unmatched ACS 2009 records. As was demonstrated in the **Blocking and Matching Variable Missingness Analysis** section, the matched percentages are affected by missingness in key variables related to DOB and first and last name. If there is an association between the socioeconomic and demographic characteristics of interest and the missingness of incoming records, then there is no clear-cut argument that correcting a possible under-coverage in the reference files of persons with the characteristics of interest will necessarily increase the match rate for the incoming records associated with these persons. If the likelihood of missingness in records of persons with the characteristics is high, then match rates may not increase as much as might be expected if additional administrative data—for example with Department of Education data—is used to decrease under-coverage in the reference files.

Socioeconomic/demographic characteristics and the missingness characteristics are categorical in nature. To study the association between categorical variables loglinear models are used (Fienberg, 2007). These models provide a more generalized version of an association test between multiple categorical variables when compared to standard chi-squared tests. Loglinear methodology is appropriate when there is no clear distinction between response and explanatory variables. By treating all the variables as response, loglinear models focus on statistical independence and dependence. But, fitting loglinear models is a computationally intensive process. If there are too many categorical variables in the analysis the algorithm may not converge properly. Therefore, we limited the number of factors in the analysis.

For the socioeconomic and demographic factors, we used the same factors used in **Exhibit 13** for the loglinear analysis. Each of the social, economic and demographic variables was defined as a two-category factor: either a record had the characteristic of interest or it did not. Because of item nonreponse in the ACS, some records contain missing information for the social and economic characteristics. We decided to remove these records from consideration, and this reduced the number of unmatched records to 292,071.[21]

With respect to missingness of key matching and blocking factors, **Exhibit 14** indicates that the unmatched proportion is related to missing date of birth (DOB) and fake/incomplete name. Therefore, these two variables were used as two-category factors in the loglinear analysis. Missing ZIP code was not used because it has little-to-no effect on the match rate. Because there appears to be a regional effect for the unmatched proportion of records, we also included a geographic factor in the loglinear analysis. In order to limit the number of geographic categories, we used Census Divisions—nine groups of states as defined in http://www.census.gov/geo/www/us_regdiv.pdf.

We fit a "saturated" loglinear model using six factors: Social, Economic (Econ), Demographic (Demo), Census Division (CensusDiv), Fake/Incomplete Name (FakeName), and Missing DOB (MissDOB). This model includes all main effects and all possible interactions of the variables. Significant interaction terms indicate a dependency between the factors. **Exhibit 17** provides the significant interaction terms that include at least one missingness factor and one socioeconomic/demographic/geographic factor. A

[21] All ACS 2009 records were used to calculate social, economic and demographic characteristics within each state. If an ACS record was missing the information for the characteristic of interest it was not counted as having the characteristic, but it was not removed from the population. In other words, the denominator of the proportion included all state records. The loglinear analysis differs in two ways. First, only unmatched ACS 2009 records are considered. Second, records with missing information related to the social and economic characteristics were removed.

complete list of the main effects and interactions is provided in Appendix C, along with the SAS code used to fit the model.

Exhibit 17: Significant Interaction Terms from the Saturated Loglinear Model of the Factors Social, Econ, Demo, CensusDiv, FakeName, and MissDOB

Interaction	Degrees of Freedom	Chi-Squared Value
CensusDiv×FakeName	8	95.57
Social×FakeName	1	770.29
CensusDiv×Social×FakeName	8	36.64
Econ×FakeName	1	365.03
CensusDiv×Econ×FakeName	8	57.92
CensusDiv×Social×Econ×FakeName	8	40.67
Demo×FakeName	1	198.5
CensusDiv×Demo×FakeName	8	62.04
Social×Demo×FakeName	1	97.62
CensusDiv×Social×Demo×FakeName	8	21.47
Econ×Demo×FakeName	1	29.83
CensusDiv×Econ×Demo×FakeName	8	19.35
CensusDiv×Social×Econ×Demo×FakeName	8	23.22
CensusDiv×MissDOB	8	336.5
Social×MissDOB	1	24.69
CensusDiv×Social×MissDOB	8	58.08
Econ×MissDOB	1	15.38
CensusDiv×Social×Econ×MissDOB	8	26.62
Demo×MissDOB	1	138.22
CensusDiv×Demo×MissDOB	8	26.36
Social×Demo×MissDOB	1	43.15
CensusDiv×Social×Demo×MissDOB	8	34.74
Econ×Demo×MissDOB	1	15.19
CensusDiv×Econ×Demo×MissDOB	8	43.87
CensusDiv×Social×Econ×Demo×MissDOB	8	20
Social×FakeName×MissDOB	1	128.79

Interaction	Degrees of Freedom	Chi-Squared Value
CensusDiv×Social×FakeName×MissDOB	8	48.15
Econ×FakeName×MissDOB	1	8.17
CensusDiv×Social×Econ×FakeName×MissDOB	8	17.17
CensusDiv×Demo×FakeName×MissDOB	8	28.91
Social×Demo×FakeName×MissDOB	1	7.16
CensusDiv×Econ×Demo×FakeName×MissDOB	8	22.37

These results suggest that there are a number of dependencies between the missingness factors and the socioeconomic, demographic and geographic characteristics. Because of the dependency between these factors, we conclude that there is an association between socioeconomic and demographic characteristics and factors that are known to reduce the likelihood of record linkage—missing information in the blocking and matching variables.

Given this association, it will be difficult to increase the PVS match rates without addressing the quality of DOB and name variables in the incoming file. The PVS reference files may under-cover certain population segments, and it would be beneficial to include more administrative records from sources that cover the underrepresented population segments. However, in the case of the populations segments defined by the socioeconomic and demographic factors we have considered, the fact that they are associated with factors that degrade the probability of a match means that these records may still not get matched to the reference files no matter how well the reference files cover the population.

2 Recommendations

In this section of the report we outline possible alternatives and improvements for the PVS. These recommendations are based on the studies and exploratory analyses described in the **Review of the Person Identification Validation System** section of this report, and we rely on information learned from the **Environmental Scan of Record Linkage Methods** (**Appendix A**). We organize the recommendations into three sets: 1) recommended research based on the investigation undertaken in our PVS assessment, 2) recommended research based on best practice concepts voiced by others who have used or reviewed the PVS, as well as the application of record linkage best practice concepts discussed in the **Environmental Scan of Record Linkage Methods**, and 3) a recommendation to consider creating a research and evaluation environment for PVS so that on-going research will not interfere with or jeopardize PVS production runs.

2.1 Extended Assessment Research

Most of the investigations undertaken in NORC's PVS assessment consider only ACS 2009 data for incoming records and the version of PVS software used by surveys. Often only the unmatched records were considered. More extensive investigations are needed that consider other incoming files of different types—survey, census, administrative record, and possibly commercial files—investigated in similar ways to see if similar observations are made. We provide descriptions of what some of this research would entail.

2.1.1 Cut and Blocking Strategies

NORC's investigation of cutting and blocking strategies in the PVS indicate that the strategies are generally effective, but there are some issues that may prevent PIK assignments to a small set of records. These issues were uncovered by rerunning all unmatched records through each module without regard to the module cuts. If these results are consistent for other incoming files then there are some enhancements that could be done to increase the number of matches.

- Incoming records in the ZIP3 "000" geo-cut can be run against all non-"000" reference file geo-cuts, and all unmatched output from GeoSearch can be run against the "000" reference file geo-

cut. If this is done, incoming file records in geo-cut "000" do not need to be run against the reference file "000" geo-cut.

- Consider changes to the name-cuts. NORC found that C and K were often interchanged for the first letter in a name, e.g. Cathy is the first name in the incoming file, but Kathy is the first name in the reference files. A change of the letter groups that define name-cuts so that C and K are together would mitigate this. Analysis of common letter interchanges is needed across all types of files in order to find the best combinations of letter groups for the name cuts. A study of using the first letter from the Soundex or NYSIIS code of first and last names is another possibility, but the effect of doing this with Hispanic names would need to be considered.

2.1.2 Relationship between Social, Economic and Demographic Factors and the Likelihood of a PVS Match

The relationship between social, economic and demographic factors and the likelihood of record matching is important to understand. First, to the extent it exists, social, economic and public policy researchers need to be informed about the issue, understand how to incorporate it into their research, and discuss potential limitations to their analyses. Meyer and Goerge (2010) used a weighting adjustment in their research on the misreporting of Food Stamp Program (FSP) benefits. Two types of research projects may help with this problem.

- Research is needed to help determine the best type of statistical adjustment to account for the fact that certain classes of people from certain parts of the country are less likely to be represented in the set of records assigned to a PIK. The problem may be due to under-coverage of these groups in the reference file, an association of the group's records with poor quality survey records—records lacking good name, address and date of birth information—or both. Other factors may have an influence as well. Regardless of the cause, users of PIK-linked data need information on proper usage of the data in their research efforts.
- A study is needed of the social, economic and demographic characteristics of matched and unmatched incoming file records to see if additional administrative records that include records for these groups can be added to the reference files. Geographic effects should be studied as well because our investigations noticed a regional effect in the unmatched data for both the ACS 2009 and Census 2010.

2.1.3 The Effect of Incoming Record Data Quality on Matching

Blocking and matching variables related to a person's name, address and data of birth are key elements in the PVS processes of linking incoming files to reference files. If this information is missing, or fake information is substituted for the true information in an incoming record, then the probability that the record will be assigned a PIK will be decreased. The magnitude of the decrease needs to be studied to look for ways to mitigate the problem. If the problem cannot be mitigated, then understanding the problem may provide a way to reassess the success of a PVS run. For example, if we know that 5 percent of the incoming records have poor information then a 93 percent match rate is quite good. Specific issues that can be studied include:

- Re-examine pre-processing rules for determining fake names. Currently, a list of fake names is used to blank out first or last names in records that have them. When both name variables are blanked, a record will be removed from the PVS process in the initial edit stage. But misspellings of these names are missed by the initial edit, e.g. "gentelman" is recorded instead of "gentleman." Studying frequency distributions of unmatched record first and last names across many different PVS runs will help to identify additional fake names, along with their misspellings. The list can then be updated for future use. It may also be the case, that different survey programs will use different rules for assigning fake names. In this case, separate lists of the names will be needed for each program.
  - The fake names may be associated with certain socio-economic and demographic groups. Consequently, the handling of this issue is related the problem of lower match rates for these groups.
  - Fake names also seem to frequently be used for institutionalized persons. Research should be done to see if there may be a way to record such people who are essentially permanent residents of the institution in a way that would allow for matching to a reference file record.

2.1.4 Matching Cause and Effect Research

Unmatched records can be a cause for concern when particular social, economic, demographic and geographic groups are over-represented in the unmatched records. As mentioned above, if this is due to a lack of coverage for the groups in the PVS reference file, then one way to mitigate the problem may be to enhance the PVS reference file so it has better coverage of the groups. However, if the reason an incoming record is unmatched is another factor—for example, poor quality of blocking and matching variables—the association observed between the socio-economic groups and the unmatched records may be due to an association between the socio-economic groups and an exogenous factor that is truly the

reason for the lack of a match. Fixing the over-representation problem will not just be a reference file coverage issue. Therefore, understanding the true reasons for PVS matches is important. The investigation NORC conducted into the association between certain social, economic and demographic characteristics should be expanded and repeated using additional incoming files.

- Study the association between social, economic, demographic and geographic characteristics and matched/unmatched status using the complete set of ACS records. NORC's investigation of this issue looked only at unmatched records. More may be learned by considering all records.
- Perform similar studies with other incoming files. Because of the size of the Census 2010 file, a sample from this file might be considered for this analysis.
- Consider extending the exercise of matching the ACS 2009 unmatched records to the Census 2010 records. This exercise could be done for all ACS 2010 and all Census 2010 records. Analyzing records in the four groups of a two-way table—matched each other and matched to the reference files, matched each other but did not matched to the reference files, did not match each other but did match to the reference files, and did not match each other and did not match the reference file—would provide insight about the characteristic of records that match and don't match the reference file. It may also provide more insight into data quality issues that affect how the PVS works.
- Investigate whether there are other causes for unmatched records aside from the poor quality of blocking and matching variables and under-coverage in the reference files.

2.1.5 Reference File Assessments

Under-coverage of U.S. population segments in the reference file is an important issue. As we have discussed, much can be learned about the reference file by analyzing matched and unmatched records. Nonetheless, an assessment of the reference file, and the files that are used to create it, are important as well.

- To the extent possible, compile statistics on the number of person records in the reference file for various socioeconomic, demographic and geographic groups, and compare these numbers to current population estimates. Note that removal of records because the associated person has died is required for this analysis. Adjustments for population migration may also be needed.
- Start keeping track of reference file PIKs that have been linked to incoming records. And when recording a link, also record the related metadata: incoming file type, vintage of the records in the

file, etc. Metadata can be saved to a file that can easily be summarized and linked by PIK to any reference file for later review.[22] Over time this information can be used to learn more about the reference file and the PVS system. Analysis of this information may reveal interesting outcomes such as finding that records with a high match expectation never or seldom get matched. Researching the cause of unexpected outcomes will lead to improvements in the system, or improvements in the way people analyze PVS linked data.

2.2 Best Practices Research

Beyond the recommended research that is an extension of NORC's investigation, there are numerous research activities that should also be considered. We list research areas that were suggested by Census Bureau staff or contractors, along with topics from the environmental scan of record linkage methods that seem applicable to the PVS.

- **The use and improvement of checklists** has been shown to be an effective tool to enhance the quality of any endeavor and improve outcomes (Gawande, 2009). The PVS process has checklists, and NORC recommends that the checklists be periodically reviewed to look for process improvement.
- **Research of the matching parameters** should be considered. Every PVS run includes a parameter input file that contains match weight cutoff values and other factors needed to perform record linkage. NORC understands that the parameters are often set based on past experience with the type of incoming file and some test runs with manual inspection of the linked records. Consideration should be given to creating a database of these parameters from all past projects so that an analysis of the information can be done. It may be possible to create a model that could predict the parameter values based on factors such as the file type, the year of the study, etc. Interactions between the parameters could be studied as well. Overall such an analysis would be beneficial in helping users to better understand the PVS process.
- **The probability of a match for each linkage** is related to the match weight of a record pair, which is produced by PVS. Transforming this weight into an estimate of a probability of a match depends on assumptions of independence across the variables used in the linkage process. Empirical evidence suggests that independence of these variables may not be a realistic

[22] This file is similar to the PIK crosswalk file, except that it will be continually updated over time and it only needs to store records for PIK that are assigned during a PVS run. If a PVS data warehouse is constructed, this would just be a special table in the database.

assumption. The Census Bureau's contractors for creating additional PVS modules, Gunnison Consulting Group and subcontractor Westat (Gunnison/Westat), have recommended considering a logistic regression approach to generate a predictive probability for each record pair. NORC concurs with this assessment. Logistic regression approaches have been used in propensity matching problems in other settings, and may provide a good alternative to models that require independence.

- **Research on estimating record linkage error rates** would also help users of PVS linked data understand the uncertainties inherent in the data. Belin and Rubin (1995) proposed a mixture model approach for estimating the false match rate, and Winkler (2007) provides an alternative mechanism for automatically estimating record linkage false match rates in situations where the subset of the true matches is reasonably well separated from other pairs and there is no training data. But Winkler cautions that this situation may be rare.

  A 2004 PVS evaluation used a truth deck approach to estimating match error rates. CPS 2001 records with verified SSNs were assumed to be true matches. These records were processed through the PVS search modules, and sets of false matches and failed matches were identified to estimate false-match and failed-match rates. This approach is no longer feasible for most survey files because SSNs are no longer collected. While we would like to think that these error rates are representative of the PVS, they may only apply to the data that were processed. We recommend that more of these types of studies be conducted, even on past incoming files. Investigations can be done to see if a model based on a number of factors—the type of incoming file (survey, census, and administrative records), data collection year, search module matched proportions, disagree proportion, name field missingness measures, etc.—can reliably predict the false-match/failed-match rates.

  Another idea on the creation of truth decks comes from Deborah Wagner, Chief of the CARRA Census Applications Group. An administrative record file that contains verified SSNs could be randomly modified by blanking out various matching fields within records. This deck could then be run through the PVS search modules to estimate error rates. NORC thinks that this idea is promising and should be pursued—possibly in conjunction with the modeling approach described above.

- **Research on the use of linked data** is needed so that users of linked data will use appropriate analysis techniques. Scheuren and Winkler (1993) investigated the effect of mismatch errors of regression coefficients and proposed a method of adjusting for the bias. Scheuren and Winkler (1997) advanced the work further with an iterative procedure that modified the regression and

matching results for apparent outliers. Lahiri and Larsen (2005) consider an alternative to the bias correction method of Scheuren and Winkler (1993).

In their work, Scheuren and Winkler (1997) incorporated quantitative information to enhance the linkage process. This is a possibility for PVS, at least for some incoming files. Quantitative information such as income can be incorporated into the reference files from SSA and federal administrative data. When similar information is present in an incoming file, the information could be used to improve the linkages.

- **Changes to the PVS programming environment** could decrease the run time of the PVS process. The PVS is programmed in SAS and data are stored in SAS files. PVS processing has a long, but acceptable, run time for many of the large files that are run in the system. An ACS file of approximately 4 million records may take 2 or 3 days to process. Small incoming files like a set of CPS records take less time to run, and extremely large files like Census 2010 may take weeks to completely process. Parallel processing can reduce the processing time for the records—for the Census 2010 DRF (over 344 million records), parallel runs allowed GeoSearch to complete in two days, while NameSearch took one day. NORC understands that personnel in the CARRA find these run-times acceptable, so there may be no need for an enhancement. But if there comes a time when the processing time is too long, consideration should be given to porting the system to the C programming language. Run times are much faster in C, and some of the programming may be simpler.

2.3 A PVS Research and Evaluation Environment

The PVS is a complex system with many parts. Aside from the SAS programs that implement the processing, the system includes a parameter file that defines the matching rules for each PVS run, and a set of reference files for every year associated with the vintage of PVS processed incoming files. Copies of the incoming files are saved as well.

Valuable information about the PVS process is contained in these files, but it may be hard to analyze due to storage capacity of the computing environment and the need to provide higher priority to PVS production runs. If a separate computing environment was set up for PVS research and evaluation, CARRA could setup a PVS research program that would not share resources with PVS production runs, or other CARRA computing activities. Research and evaluations that are described in previous sections of this report could be conducted in an efficient and ongoing way. The following, is a brief description of

some of the benefits of such an environment. Additionally, we discuss the need for a robust data management system.

- **A PVS research and evaluation environment could be constructed within an administrative record research environment.** The use of administrative records in survey and census research is increasing. For example, Mulry, et al (2006) used administrative records to examine the quality of the estimates of duplicate enumerations in Census 2000. NORC also understands that the Census Bureau is looking for ways to use administrative records in Census 2020. The PVS will likely play a big role in such research, and a PVS research and evaluation environment would fit in well with these plans.
- **A record linkage research database** could be considered for this environment. This would provide researchers easy access to important articles and papers on record linkage methodology.
- **The implementation of systems thinking** would be enhanced by a PVS research and evaluation environment. Research based on a single PVS project has benefits, but a review across many projects will be more beneficial. In talking about better graphics design, Edward Tufte notes that "local optimizing leads to global pessimizing" (Wehr, 2009 and many others). But, this concept is true in many other fields. A research and evaluation environment would provide researchers access to a number of PVS project results so that a more global view can be taken.
- **Shifting focus from exact data linkage to information generation** may also be a benefit of a PVS research and evaluation environment. NORC's PVS assessment notes that the use of fake names in survey files can make some incoming file records unmatchable. There may be other data issues that prevent matching as well. This lack of matchability and the fact that matches made using probabilistic record linkage have a degree of uncertainty needs to be recognized by those who use linked data. We noted that research needs to be done to help data users understand appropriate methods for analyzing the data. Such research is more likely to proceed and more easily carried out if a research and evaluation environment is available.

2.3.1 Data Management

Data management within a research and evaluation environment will become a major problem if a reliable data management system is not part of the environment. A data warehouse data management system should be considered. With such a system, management of large data file would become easier, and storage of the information more efficient. Security of the information may be more manageable, and a permissions system can be set up to control data access to researchers.

References

Census Bureau Documents and Spreadsheets

ACS PVS Results All Years for Groves Briefing.xls

CPS97 PVS Module Test finalstats.xls

Haines, D.E., Wagner, D.A., Resnick, D. M. (2008), Data Integration Division DIDMatch Software Documentation v0.1.doc

DRAFT PVS One Pager 20091015.doc

DRAFT PVS Version 4 Overview.vsd

Person Identification Validation System Match Steps 2010_07_12.doc

Killion, R.A. (2002), "Development of a Production Capability for Social Security Number (SSN) Verification at the Census Bureau," PVS ARR#58 _final_.pdf

PVS Description Groves Briefing 2010_07_15 v3.doc

Final Report: PRED Social Security Number Validation System Research Project (2004), PVS Final Evaluation Report 10242006.doc

PVS Op Spec Sheet (ACS).doc

PVS Survey QC Steps.doc

Wagner, D. A. (2007), PVS Technical Documentation (5_30_07)_updated 9_15_10.doc

PVS Template with MAFMatch No Paths 2010_10_26.pdf

PVS Validation Flag Values.doc

Roemer M. and Stinson M. (2003), PVS-Review_Final.doc

PVS_GEOKEYS_Overview.vsd

Wagner, D.A. (2009), PVS_Overview_Presentation_20090825.doc

StARS 2009 PVS Results.doc

Books, Papers, and Presentations

Belin, T. R., and Rubin, D. B. (1995), "A Method for Calibrating False- Match Rates in Record Linkage," Journal of the American Statistical Association, 90, 694-707.

Carr, D. B., A. R. Olsen, J. P. Courbois, S. M. Pierson, and D. A. Carr. 1998. "Linked Micromap Plots: Named and Described," *Statistical Computing & Graphics Newsletter*, Vol. 9 No. 1 pp. 24-32.

Fienberg, S. (2007), *The Analysis of Cross-Classified Categorical Data, Second Edition*, New York, N.Y.: Springer.

Gawande, A., (2009), *The Checklist Manifesto: How to Get Things Right*, Metropolitan Books

Herzog, T. N., Scheuren, F., and Winkler, W.E., (2007), *Data Quality and Record Linkage Techniques*, New York, N.Y.: Springer.

Linked Micromaps, Version 1.0.2. March 2009. Statistical Research and Applications Branch, National Cancer Institute.

Meyer, B. D. and Goerge, R. M. (2010). "Errors in Survey Reporting and Imputation and their Effects on Estimates of Food Stamp Program Participation," working paper.

Mulry, M. H., Bean, S. L., Bauder, D. M., Wagner, D., Mule, T., Petroni, R. J., (2006), "Evaluation of Estimates of Census Duplication Using Administrative Records Information," *Journal of Official Statistics*, Vol. 22, No. 4, pp. 655–679.

Resnick, D. M. (2010) "Current Records Linkage Research and Practice at the U.S. Census Bureau," *Proceedings of the 2010 Joint Statistical Meetings*.

Scheuren, F., and Winkler, W. E. (1993), "Regression Analysis of Data Files that are Computer Matched," *Survey Methodology*, **19**, 39-58.

Scheuren, F.,and Winkler, W. E. (1997), "Regression analysis of data files that are computer matched, II," *SurveyMethodology*, 23, 157-165, http://www.fcsm.gov/workingpapers/scheuren_part2.pdf.

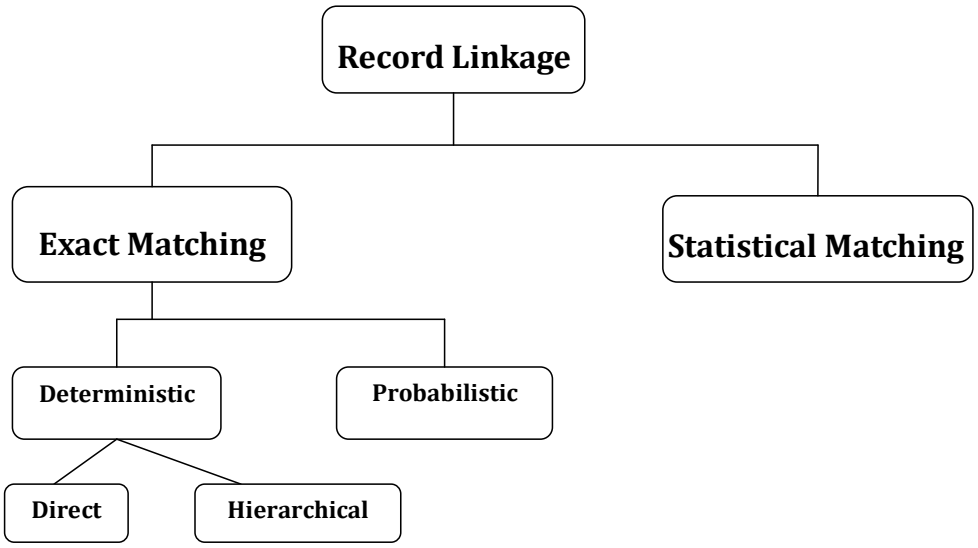
Wehr, J. (2009) "Edward Tufte Course Notes and Reactions," http://wehrintheworld.blogspot.com/2009/03/edward-tufte-course-notes-and-reactions.html

Winkler, W. E. (2007), "Automatically Estimating Record Linkage False Match Rates," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, CD-ROM http://www.census.gov/srd/papers/pdf/rrs2007-05.pdf.

Appendix A:
Environmental Scan of Record Linkage Methods

Record linkage is the task of quickly and accurately identifying records corresponding to the same entity from one or more data sources and combining them. Record linkage is closely related to the terms data cleaning, entity resolution, and the merge/purge problem (Herzog, Schreuren and Winkler, 2007). In the absence of unique identifiers, the basic methods compare name and address information of entity records across data files of interest to determine those sets of records within and across files that are associated with the same entity. An entity might be a business, a person, or some other type of unit that is listed. Record linkage is an important tool for the creation of a large, integrated, coherent database. Such databases are used for: health research, social or economic statistical studies, epidemiological cohort studies, computer science applications, or other research of public interest.

In this environmental scan, we survey the best practices and recent innovations in record linkage methods in government agencies in the US and other countries, as well as activities in academia, and the private sector in the US. We also briefly review the methodologies developed within the computer science arena that are generally referred to as entity resolution (e.g., Brizan and Tansel, 2006). The following chart (Fox and Stratychuk 2010) gives a broad overview of various record linkage techniques. Herzog, Scheuren and Winkler (2007) generally follow this organization of topics for record linkage, and we have adopted it as well for this environmental scan report.



References:

Brizan, D. G. and Tansel, A. U. (2006), "A Survey of Entity Resolution and Record Linkage Methodologies," Communications of the IIMA, Volume 6, Issue 3.

Fox, K. and Stratychuk, L. (2010), Proceedings of the Statistics Canada Symposium 2010, Workshop1: Record Linkage Methods.

Herzog, T. N., Scheuren, F., and Winkler, W.E., (2007), Data Quality and Record Linkage Techniques, New York, N.Y.: Springer.

Statistical Matching

Exact matching is used to integrate datasets with substantial overlap (with regard to observed entities as well as variables). Matching of records belonging to identical entities is the objective. If this is not possible (or not even necessary), e.g., in a situation where two or more surveys are based on different samples with low likelihoods of overlap, statistical matching may be used as an approximation of exact matching. For a comparative discussion of exact and statistical matching see the FCSM Statistical Policy Working Paper 5 (Federal Committee on Statistical Methodology, 1980).

In statistical matching the linkages are based on similar characteristics rather than unique identifying information. Linked records need not correspond to the same unit. In a statistical match each observation in one microdata set (the "base" set) is assigned one or more observations from another microdata set (the "nonbase" set); the assignment is based upon similarity in selected characteristics. Conceptually, statistical matching is closely related to imputation. The method relies on the joint distribution of the variables (i.e., the characteristics forming the basis for matching). This can lead to inaccurate analysis if the joint distribution is incorrectly specified. For example, if we want to match units based on the distribution of household income and household type, and we only link single males, the distribution for single persons will be misspecified (Fox and Stratychuk 2010).

At the Italian National Statistical Institute (ISTAT), statistical matching is considered advantageous over conducting a new survey to obtain information about certain variables of interest. Among the advantages, the use of already available sample surveys and or administrative databases makes it possible to obtain timely, inexpensive results. Furthermore, a reduction of response burden on the survey units is expected. For the statistical matching procedure, most of the data sources come from sample surveys and administrative databases that either lack unit identifiers due to privacy constraints, or that have little overlap of units.

One important field of application, in the context of ISTAT, is related to the integrated analysis of two economic variables: consumer's expenditures and income. Even if many surveys observe these variables jointly, there is not a single source that describes both of them with high quality and high level of detail. More often, each survey focuses alternatively on either consumer's expenditures or income. As far as income is concerned, the Household Balance survey (HB) managed by the Bank of Italy is considered the most detailed and complete. Different sources can be used for expenditures: among the others, the Household Expenditure survey (HE) and the Household Multipurpose survey (HM), both managed by ISTAT. To achieve their objective, ISTAT links these two datasets through various statistical matching methodologies (D'Orazio, Di Zio, and Scanu 2001, 2006). Examples of linking such datasets in ISTAT are the following:

The construction of the Social Accounting Matrices: The Social Accounting Matrix (SAM) is a system of statistical information containing economic and social variables in a matrix formatted data framework. The matrix includes economic indicators such as per capita income and economic growth. In Italy an archive of this information is not available; hence, SAM's are built by means of combining the data sources HB, HE and the National Accounts.

The analysis between income and health expenditures: During the 2000 Annual Report of ISTAT, the problem of evaluating the relation between income and health expenditures arose. Due to the lack of time and of additional funds for an *ad hoc* survey, the only feasible way to reach the scope was to combine information coming from the 1994 HM and the 1995 HB. The objective consisted in an estimate of a parameter representative of the relation between health expenditures and income.

The construction of comprehensive data-sets for flexible statistical analysis: The surveys HE and HB can be integrated so that a complete dataset of units becomes available in order to: (a) analyze family's saving; (b) analyze the decisions for groups of non-durable (or durable) goods; (c) implement microsimulation models for the analysis of public policies; (d) supply a multidimensional analysis of poverty.

More generally, in Europe, there is a growing demand for new indicators and statistical surveillance tools cutting across several domains in socioeconomic areas (Leulescu and Di Meglio 2010). The importance and urgency of this demand is demonstrated by recent European initiatives: the GDP and beyond communication, the Stiglitz-Sen-Fitoussi Commission' Report (September 2009) and Europe 2020 Strategy. One of the key improvements foreseen in the coming years is finding broader ways for the measurement of quality of life on the basis of the Stiglitz Commission report that encompasses several

key dimensions such as, living conditions (including income, consumption and wealth), health, education, personal activities (paid work, unpaid domestic work, commuting, leisure, and housing), political voice and governance, social connections, environmental conditions, personal insecurity, economic insecurity etc. But there is no agreement reached on what are the appropriate outcomes within all these domains and on how they should be combined in an overall index. Furthermore, there is a need to go beyond aggregates and capture heterogeneity in the population: distributional and inequality aspects, sub-national statistics, and vulnerable sub-groups. Europe 2020 sets out an example of such subpopulations disadvantaged in several domains: people at-risk-of-poverty and social exclusion. In order to obtain such an overall index of quality of life, Leulescu and Di Meglio (2010) describe and evaluate the utility of statistical matching methods for the integration of various surveys.

References:

D'Orazio, M., Di Zio, M. and Scanu, M. (2001) Statistical Matching: a tool for integrating data in National Statistical Institutes. In *Proc. of the Joint ETK and NTTS Conference for Official Statistics*, Crete.

D'Orazio, M., Di Zio, M. and Scanu, M. (2006) *Statistical Matching: Theory and Practice*.Wiley Series in Survey Methodology.

Federal Committee on Statistical Methodology (1980), Statistical Policy Working Paper 5: Report on Exact and Statistical Matching Techniques," Washington, DC: Office Federal Statistical Policy and Standards, U.S. Department of Commerce. Available at http://www.fcsm.gov/working-papers/wp5.html

Fox, K. and Stratychuk, L. (2010), Proceedings of the Statistics Canada Symposium 2010, Workshop1: Record Linkage Methods.

Susanne Rässler, S. (2002) *Statistical Matching: A Frequentist Theory, Practical Applications and Alternative Bayesian Approaches.* Springer Lecture Notes in Statistics.

Exact Matching

An exact match is a linkage of records for the same unit from different databases. In some instance however, the linkage procedures may erroneously link units that are not the same. Exact matching uses unique identifying information such as a government program identification number—for example a Social Security Number (SSN) in the US, or a Social Insurance Number in Europe or Canada—date of birth (DOB), name, address etc. Various forms of exact matching are used by statistical agencies worldwide.

Suppose file A (e.g., survey data) has n_a records and file B (e.g., administrative data) has n_b records, then the file A × B contains $n_a \times n_b$ record pairs. Each of the n_b records in file B is a potential match for each of

the n_a records in file A. Thus there are $n_a \times n_b$ record pairs whose match/non-match status is to be determined. The linkage process uses the unique identifiers in files A and B to classify the $n_a \times n_b$ record pairs in the file A × B as either matches or non-matches. In practice, in order to reduce the number of pairs that have to be investigated by the matching procedure, the set of all record pairs is decomposed into (i) blocks containing candidate pairs that agree on certain variables (called blocking variables) which are then further examined to determine match status, and (ii) a residual set of determinate non-matched pairs that do not belong to the same block. For additional details on the use of blocking variables see Gomatam et al. (2002), Jaro (1989), Winkler (1985), and Herzog, Scheuren, and Winkler, 2007). Two major classes of exact linkage strategies are deterministic and probabilistic, which are discussed in more detail below.

In most applications, a combination of available methods seems to work best. A quite common pragmatic approach is to use deterministic linkage, followed by probabilistic linkage (including string comparators, if necessary), then followed by clerical review (Denk and Hackl 2003, Gill 2001). For example, the Person Identification Validation System (PVS) developed by the U.S. Census Bureau is a combination of deterministic and probabilistic record linkage techniques (see section 3.4.1.3 for further details).

Deterministic Methods

Deterministic algorithms for exact matching can range from the simple to complex. They result in direct matching, (the simplest deterministic record linkage method) and hierarchical exact matching.

Direct Matching

Direct matching is a simple method of linking records using a unique identifier or collection of unique identifiers. This is also known as the match-merge method. In direct matching, matches are determined by 'all-or-nothing' comparisons; that is, agreement on all key identifiers. In this kind of matching when comparing two records on first and last name, for example, the records are considered matches only if the names on the two records agree on all characters (Gomatam et al., 2002).

Entity-level identifiers serving as keys generally provide good, but not perfect, match-merge results. Problems may occur due to omissions or errors. Data errors and omissions occur for numerous reasons. Information may be omitted because the respondent is unwilling or unable to supply it. For example, people are often reluctant to supply their Social Security Number (SSN). Even when the respondents supply personal information, it may be recorded incorrectly. Digits may be transposed in Medicaid IDs, SSNs, names may be difficult to spell or the letters ordered incorrectly. Data errors may occur when written information is illegible. If the key is missing for a given record, matching that record is not

possible. Two outcomes are possible for records with incorrect key values: either the incorrect records do not match with any records on the second file or they may match with the wrong record.

Examples of Direct Matching

An example of the simple deterministic algorithm is found in the State of California's Family Outcomes Project (FOP). The FOP has adopted a Common Patient Identifier (CPI) constructed from the following data elements: gender, date of birth, birthplace, first 3 characters of first name, and first 3 characters of last name (Campbell 2009). Another example of a match-merge is the use of a Medicaid ID to combine current Medicaid Eligibility information with Medicaid Claims (Whalen et al. 2001). A match-merge can use a simple, single key, as with the Medicaid example, or use a more complex key made up of several variables.

Hierarchical Exact Matching

Hierarchical exact matching uses multiple passes with different matching criteria to match records from one file to another. Multiple identifiers are used when a highly reliable single unique identifier is not available. The records are linked in a sequence of steps each of which decides the linkage status (either match or non-match) of the record pair by considering *exact* agreement on a particular subset of identifiers. The matching process usually starts with the most stringent criteria and moves towards the least stringent criteria. This in turn ensures that the first step implemented in the hierarchical matching procedure would have the lowest false match (or false positive) rate and the highest false non-match rate. Each successive step would increase the overall false match rate and decrease the false non-match rate. This allows the user to have some indirect control over error rates via a choice of the number of steps to be executed (Gomatam et al. 2002).

At each step, the unique matches are extracted; the duplicates and the remaining unlinked observations in each of the two data sets (the residuals) form the input to the next step in the data linkage process, which continues with a different subset of identifiers. Although each identifying variable is still tested for total agreement (as in direct matching) at each step, by implementing the succession of steps described above, the effect is similar to that of considering agreement among a partial subset of the complete set of unique identifiers. Hierarchical exact matching is also known as stepwise deterministic strategy (SDS; Gomatam et al. 2002).

Preprocessing for Hierarchical Exact Matching

Names and other variables can include variations and errors such that exact string matches may fail when a human reader might recognize them as equivalent (e.g. "Jim" and "James"). Pre-processing names using

a phonetic compression algorithm would help overcome such variations and errors. There are several phonetic compression algorithms; examples include Soundex (Knuth 1998), Metaphone, and the New York State Identification and Intelligence System algorithm (NYSIIS; Lynch and Arends 1977). The NYSIIS algorithm has high discriminating power (Newcombe 1988). Jaro (1989) applied the Soundex encoding to transform a person's name into code that tends to bring together all variants of the same name. The US Census Bureau's Person Identification Validation System (PVS) uses the Soundex of street name and NYSIIS code of the first and last name as blocking variables, albeit in a probabilistic record linkage algorithm.. Gomatam et al. (2002) used NYSIIS codes created for first, middle and last names. Although using phonetic codes, such as the NYSIIS, for names can handle some kinds of partial agreements and phonetic/spelling errors in names, partial agreements caused by transpositions of letters, and other more complicated forms of agreement would be difficult to deal with. To deal with these problems one may need to consider string comparator metrics (Winkler 1990), as discussed later.

In order to make reasonable comparisons of string variables, adequate pre-processing by standardizing and parsing the strings is essential (Denk and Hackl 2003). Appropriate parsing of name and address components is the most crucial part of computerized record linkage. Without it, many true matches would erroneously be designated as nonmatches because common identifying information could not be compared (Winkler 1993). DeGuire (1988) presents an overview of the ideas needed for parsing and standardizing addresses. The basic idea of standardization is to replace the many spelling variations of commonly occurring words with standard spellings such as a fixed set of abbreviations or spellings. Parsing divides a string variable into a set of string components which are then individually compared. For further details on standardization and parsing of name and address, with examples, see Winkler (1993).

Examples of Exact Hierarchical Matching

An example of exact hierarchical matching (or SDS linkage) can be found in Gomatam et al. (2002). They combined Medical records from children born in the years 1989–1992 and treated in Florida's Regional Perinatal Intensive Care Centers (RPICC) with data on their subsequent educational performance recorded by Florida's Department of Education (DOE). The objective was to model the association between school outcome as a function of the medical and family sociodemographic conditions at birth. For the SDS linkage of the two data sets mentioned above, records were matched using a sequence of unique identifiers such as the last name, first name, middle name, date of birth, race and sex. A county code was also present in both data sets – the county of mother's residence at child's birth was available in the RPICC data set, while that of current enrolment was available in the DOE data. NYSIIS codes created for first, middle and last names were also used. Only exact character-by-character matches

were considered for pairs of strings. Gomatam et al. (2002) have considered four passes based on the unique identifiers.

Using records from two hospital (in central Indiana) systems' patient registries (a public inner-city hospital system with a large Medicare/Medicaid population and another private urban hospital system that invested in extensive patient registry clean-up in 1999), Grannis et al. (2002) applied a hierarchical exact matching method to link individuals to the Indiana subset of the Social Security Death Master File (SSDMF) based on a set of unique identifiers. For further details see Grannis et al. (2002).

Probabilistic Record Linkage

Probabilistic record linkage identifies a match between records based on a formal probabilistic model. The advantage of probabilistic record linkage is that it uses all available identifiers to establish a match (e.g., name, sex, date of birth, SSN, race, address, phone number) and does not require identifiers to match exactly. Identifiers that do not match exactly are assigned a "distance" measure to express the degree of difference between files. Each identifier is assigned a weight and the total weighted comparison yields a score, which is used to classify records as linked, not linked, or uncertainly linked according to whether the probability of a match exceeds a certain threshold. Herzog, Scheuren, and Winkler (2007) describe the key principles of probabilistic record linkage. Below we describe probabilistic record linkage and measures of its analytical quality. We conclude the section by providing a summary of comparisons of probabilistic record linkage to deterministic record linkage.

Optimal Linkage Rule

Although Newcombe (1959, 1962) introduced the use of the frequency ratio for record linkage in earlier work, Fellegi and Sunter (1969) are recognized as the first researchers to rigorously present the mathematical model and theoretical foundation for probabilistic record linkage. Their framework groups possible pairs into three sets, referred to as links (L), non-links (N), and possible links (P), based on objective criteria. Each set (L, P, and N) has associated error rates. An optimal linkage rule is defined as one that minimizes the probability of classifying a pair as belonging to set P for fixed error levels in L and N. The decision rules in the Fellegi-Sunter model are optimal in the sense that, given fixed upper bounds on the rate of false matches and false nonmatches, the decision rules minimize the size of the possible (indeterminate) links (P).

Assume that every record pair (a, b) are compared on the basis of k unique identifiers. We can define a vector γ, consisting of 1's and 0's, where 1 (0) indicates that the record pair agrees (disagrees) on

component *j*. Also define m_j and u_j to be the conditional probability that component *j* matches, given that the record pair (*a*, *b*) is a true match and true non-match, respectively. Fellegi and Sunter (1969) show that the optimal rule can be expressed in terms of the composite weight $\sum_{j=1}^{k} w_j$, where the weight $w_j = \log\left(m_j / u_j\right)$, if for a given record pair, component *j* agrees (matches), and $w_j = \log\left(\left(1-m_j\right)/\left(1-u_j\right)\right)$ if the component *j* disagrees. Since $m_j > u_j$ in most cases, unique identifiers that agree make a positive contribution to the composite weight, whereas the identifiers that disagree make a negative contribution.

Blocking

If all possible record pair comparisons between two files were actually carried out, the number of comparisons could be very large even for relatively small files. In practice, as mentioned in the context of hierarchical exact matching, to reduce the number of comparisons, only the record pairs from a subset of the files (based on blocking variables) are examined to determine match status (Gomatam et al. 2002). A good multiple blocking strategy example is provided in Jaro (1989), which describe a probabilistic record linakge of the 1985 census of Tampa, Florida to an independent post-enumeration survey (PES). The current Census PVS also uses multiple blocking strategies in the GeoSearch and NameSearch modules.

Estimation of Weights

Estimating the conditional probabilities m_j and u_j and hence the weights is not a trivial task. Fellegi and Sunter (1969) propose two methods of estimating the weights – one requires the knowledge of various error rates associated with the unique identifiers in the data sets to be linked, whereas the other gives a closed-form solution when there are three matching variables, under the assumption of conditional independence of the components of the γ vector. Jaro (1989) used an EM algorithm (Dempster et al. 1977), after introducing latent variables (true match or non-match status), to estimate weights and tested the method in the context of a linkage between 1985 Tampa, FL census data to an independent post-enumeration survey (PES). Jaro also assumed conditional independence of the components of the γ vector. However, the conditional independence assumption may not hold in practice. For example, if a pair of records agrees on the nine-digit zip code, then it is more likely to simultaneously agree on characteristics such as house number, and street name. This is regardless of whether the record pairs are a match or a non-match. Although such departures from conditional independence can be quite pronounced, the decision rules obtained from the Fellegi-Sunter framework can still be quite accurate (Herzog, Scheuren, and Winkler 2007). Winkler (1988) describes a method for estimating weights using the EM

Algorithm under less restrictive assumptions, where in the weight computation automatically incorporates a Bayesian adjustment based on file characteristics.

As a special case of their general theory of record linkage, Fellegi and Sunter (1969) presented a formal model for matching that uses the relative frequency of strings being compared. For instance, a surname that is relatively rare in pairs of records taken from two files has more distinguishing power than a common one. Most applications of frequency-based matching have used close variants of the basic model but have made different simplifying assumptions that reduce computation and facilitate table building (Winkler 1989b). Winkler (1989b) introduces an extended methodology under weaker assumptions. While the amount of computation is significantly increased (as much as an order of magnitude), the need for expert human intervention is reduced. Most or all of the matching parameters can be automatically computed using file characteristics alone. The methodology does not require calibration data sets on which true match status has been determined. No a priori assumptions about parameters or previously created lookup tables are needed.

Copas and Hilton (1990) measure the evidence that a pair of records relates to the same, rather than different, individuals. Their paper emphasizes statistical models which can be fitted to a file of record pairs known to be correctly matched, and then used to estimate likelihood ratios. A number of models are developed and applied to UK immigration statistics. The combination of likelihood ratios for possibly correlated record fields is discussed.

Many applications of the Fellegi-Sunter model use simplifying assumptions and ad hoc modifications to improve matching efficacy. Because of model misspecification, distinctive approaches developed in one application typically cannot be used in other applications and do not always make use of advances in statistical and computational theory (Winkler 1993). In Winkler's paper, an Expectation-Maximization (EMH) algorithm that constrains the estimates to a convex subregion of the parameter space is given. The EMH algorithm provides probability estimates that yield better decision rules than unconstrained estimates. The algorithm is related to results of Meng and Rubin (1993) on Multi-Cycle Expectation-Conditional Maximization algorithms and makes use of results of Haberman (1977) that hold for large classes of loglinear models.

Decision Problem

After data collection, preprocessing of data, and determination of weights, the next step is the assignment of candidate matched pairs where each pair of records consists of the best potential match for each other from the respective data bases. According to specified rules, a scalar weight is assigned to each candidate pair, thereby ordering the pairs. The final step of the record linkage procedure is viewed as a decision

problem, based on the weight, where three actions are possible for each candidate matched pair: declare the two records matched, declare the records not matched, or send both records to be reviewed more closely. Belin and Rubin (1995) outline a general strategy for the decision problem, that is, for accurately estimating false-match rates for each possible cutoff weight. The strategy uses a model where the distribution of observed weights is viewed as a mixture of weights for true matches and weights for false matches. An EM algorithm for fitting mixtures of transformed-normal distributions is used to find posterior modes; associated posterior variability is due to uncertainty about specific normalizing transformations as well as uncertainty in the parameters of the mixture model, the latter being calculated using the supplemented EM (SEM) algorithm. This mixture-model calibration method is shown to perform well in an applied setting with census data.

Winkler (2006b) provides a mechanism for automatically estimating record linkage false match rates in situations where the subset of the true matches is reasonably well separated from other pairs and there is no training data. His method provides an alternative to the method of Belin and Rubin (1995) and is applicable in more situations. He provides examples demonstrating why the general problem of error rate estimation (both false match and false nonmatch rates) is likely impossible in situations without training data and exceptionally difficult even in the extremely rare situations when training data are available.

In record linkage problems, patterns of agreements on variables are more likely among records pertaining to a single person than among records for different people, the observed patterns for pairs of records can be viewed as arising from a mixture of matches and nonmatches (Larsen and Rubin 2001). Mixture model estimates can be used to partition record pairs into two or more groups that can be labeled as probable matches and probable nonmatches. The marginal information in the database can be used to select mixture models, identify sets of records for clerks to review based on the models and marginal information, incorporate clerically reviewed data, as they become available, into estimates of model parameters, and classifies pairs as matches, nonmatches, or in need of further clerical review.

String Comparator Metrics

Locating matches across a pair of lists not having unique identifiers such as a social security number is often difficult. Typically available identifiers such as first name, last name, and various demographic, economic, or address components may not uniquely identify matches because of typographical variations. When comparing values of these string variables, such as, names or addresses, it usually does not make sense to just discern total agreement and disagreement. Typographical error may lead to many incorrect disagreements.

Several methods for dealing with this problem have been developed: string comparators are mappings from a pair of strings to the interval [0, 1] measuring the degree of compliance of the compared strings (Winkler, 1990). String comparators may be used in combination with other exact matching methods, for instance, as input to probabilistic linkage method. The simplest way of using string comparators for exact matching is to define compliance classes based on the values of the string comparator. In order to make reasonable comparisons of string variables, adequate pre-processing by *standardizing* and *parsing* the strings is essential, as discussed earlier. This holds, in particular, when matching business data, since inconsistencies of name and address information are typically even greater for this kind of data (Winkler, 1999).

In the context of the Fellegi-Sunter Model of Record Linkage, Winkler (1990) describes a string comparator metric that partially accounts for typographical variation in strings such as first name or surname, decision rules that utilize the string comparator, and improvements in empirical matching results.

Jaro (1989) introduced a string comparator that accounts for insertions, deletions, and transpositions. The basic steps of this algorithm include computing the string lengths and finding the number of common characters in the two strings and the number of transpositions. Jaro's definition of "common" is that the agreeing character must be within the half of the length of the shorter string. Jaro's definition of transposition is that the character from one string is out of order with the corresponding common character from the other string. Porter and Winkler (1999) modified the original string comparator introduced by Jaro in the following three ways:

- A weight of 0.3 is assigned to a 'similar' character when counting common characters. Winkler's model of similar characters includes those that may occur due to scanning errors ("1" versus "l") or key punch errors ("V" versus "B").
- More weight is given to agreement at the beginning of a string. This is based on the observation that the fewest typographical errors occur at the beginning of a string and the error rate then increases monotonically with character positions through the string.
- The string comparison value is adjusted if the strings are longer than six characters and more than half the characters beyond the first four agree.

String comparison in record linkage can be difficult because lexicographically "nearby" records look like "matches" when they are in fact not. Because pairs of strings often exhibit typographical variation (e.g., Smith versus Smoth), the record linkage needs effective string comparison functions that deal with

typographical variations. Hernandez and Stolfo (1995) discussed three possible distance functions for measuring typographic errors: edit distance, phonetic distance, and typewriter distance. They developed an Equational Theory involving a declarative rule language to express these comparison models.

Accuracy of Record Linkage Methods

To evaluate the performance of record linkage methods in matching records, the following criteria have been suggested in the literature (Blakley and Salmond 2002, Gomatam et al. 2002). For ease of illustration, let's assume that the record linkage procedure yields the following two-way table:

	Matches	Non-matches
Linked	A	B
Non-linked	C	D

Proportion of true match (*sensitivity* in epidemiological terms) = A/(A+C). This can also be viewed as 1- probability of false positives (matches).

Proportion of true non-match (*specificity* in epidemiological terms) = D/(B+D). This can be viewed as 1- probability of false negatives (non-matches).

Proportion of linked records that are valid (positive predictive value, PPV) = A/(A+B)

A common practice for evaluating the accuracy of any record linkage procedure, is to set up "truth decks" of records for which know the true match/non-match status of all possible record pairs constructed from the decks. In general, such decks are not common, but they can be constructed from some of the records of interest in a linkage study. For example, while linking medical records from the Florida's Regional Perinatal Intensive Care Centers database (49,862 RPICC records) and educational records from the Florida Department of Education (628,860 DOE records), Gomatam et al. (2002) considered only the subset of 1,156 records in the RPICC database, for which unique social security number matches were available in the DOE database, to compare a stepwise deterministic linkage strategy with a probabilistic strategy.

The U.S. Census Bureau, while performing error rate analysis for the PVS process using 2001 CPS data, used a truth deck constructed of records that were verified in the PVS process using reported SSNs for the CPS source records. These records were run through the PVS process without using SSNs and an initial estimate of false match and false nonmatch rates were obtained.

Comparison of Deterministic and Probabilistic Record Linkage

Probabilistic linkage (PL) and deterministic linkage (DL) both combine data using identifying information from two or more files. Like PL, DL also uses multiple criteria to link records, but it's usually based on an exact agreement criteria, although in the literature (Whalen et al. 2001) deterministic linkage method has also been described as one that uses scores (instead of exact match) to establish record links. The difference lies in the manner in which points and thresholds are set. With deterministic linking, agreement points and linkage thresholds are set outside of and known prior to the linking process. This is not the case with probabilistic linking. Agreement points, referred to as "weights", are determined by the data; these are scaled relative to the value of the identifier. The idea of creating data-driven weights makes it a flexible method that adapts to differing conditions and overcomes the main weakness of deterministic linking – the arbitrary and rigid assignment of agreement weights. In general, disagreements are ignored with deterministic linkage, while probabilistic linkage uses both agreements and disagreements for all identifiers. This is a unique characteristic of probabilistic linkage.

Gomatam et al. (2002) compare a stepwise deterministic linkage strategy with a probabilistic strategy for a situation in which the truth is known. The comparison was carried out on a linkage between medical records from the Regional Perinatal Intensive Care Centers database and educational records from the Florida Department of Education. Social security numbers, available in both databases, were used to decide the true status of each record pair after matching. Match rates and error rates for the two strategies are compared and a discussion of their similarities and differences, strengths and weaknesses is presented.

As a general rule, positive predictive values (PPV) of deterministic protocols are slightly higher than those of probabilistic protocols (Gomatam et al. 2002, Grannis et al. 2002). The sensitivity of deterministic protocols is usually lower than those produced by probabilistic protocols. Thus, in situations where sensitivity or overall accuracy is more important than PPV, probabilistic linkage is recommended. For example, in the Census Coverage Measurement program, to evaluate the quality of census and make decisions on adjustments, a high degree of overall accuracy of the linkage process is required. In such cases it would be preferable to use probabilistic methods, with clerical review of possible matches to ensure high accuracy (Gomatam et al. 2002). On the other hand, when we are interested in creating a dataset for analysis by record linkage, then the goal is to maximize the valid links and hence a high PPV is desirable. In such cases deterministic linkage rule can be applied. A PPV close to 1 can be attained by choosing subsets of identifiers so that a match on the least conservative subset (that is used in the last of the multiple passes) is highly likely. Usually a PPV less than 1 results in decreased efficiency of the statistical analyses due to addition of noise resulting from linkage error.

Deterministic strategies have the disadvantage of being ad hoc, that is, they are situation-specific, can sometimes be very time-consuming, and are very practitioner-dependent. They also have the disadvantage of not being able to handle partial agreements easily to the extent that probabilistic strategies can. Advantages of deterministic strategies are that they are easier to interpret, and allow the practitioner to incorporate specific knowledge about the databases. Hybrid approaches involving deterministic strategies in conjunction with probabilistic strategies have been used in practice (see Winkler 1985), and may sometimes enhance the performance of the linkage.

References:

Belin, T. R., and Rubin, D. B. (1995), "A Method for Calibrating False- Match Rates in Record Linkage," *Journal of the American Statistical Association*, 90, 694-707.

Blakely, T. and Salmond, C. (2002) Probabilistic record linkage and a method to calculate the positive predictive value Int. J. Epidemiol. 31: 1246-1252.

Campbell, K.M. (2009) Impact of record-linkage methodology on performance indicators and multivariate relationships, *Journal of Substance Abuse Treatment*, 36, pp. 110-117.

Copas, J. R., and Hilton, F. J. (1990), "Record Linkage: Statistical Models for Matching Computer Records," Journal of the Royal Statistical Society, A, 153, 287-320.

DeGuire, Y. (1988), "Postal Address Analysis," *Survey Methodology*, **14**, pp. 317-325.

Dempster AP, Laird NM, Rubin DB (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B; 39:1–38.

Denk, M. and Hackl, P. (2003). Data Integration and Record Matching: An Austrian Contribution to Research in Official Statistics**,** *Austrian Journal of Statistics*, Volume 32, Number 4, 305-321.

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

Grannis S, Overhage J, McDonald C (2002). Analysis of identifier performance using a deterministic linkage algorithm. *Proceedings of American Medical Informatics Association Symposium*, Philadelphia, PA. Hanley and Belfus.

Gill, L.E. (2001). Methods for automatic record matching and linking in their use in National Statistics. GSS Methodology Series, NSMS25. Office for National Statistics, UK, 2001.

Gomatam, S., Carter, R., Ariet, M., Mitchell, G. (2002), An empirical comparison of record linkage procedures, *Statistics in Medicine*, 21, pp 1485—1496.

Gu, L., Baxter, R. Vickers, D., and Rainsford, C. (2003). "Record linkage: current practice and future directions."

Haberman, S. J., (1977), "Product Models for Frequency Tables involving Indirect Observation," *Annals of Statistics*, **5**, 1124-1147.

Hernandez, M.A and Stolfo, S.J. (1995), "The Merge/Purge Problem for Large Databases." In Proc. of 1995 ACT SIGMOD Conf., pages 127–138.

Jaro, M. A. (1989), "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida," *Journal of the American Statistical Association*, 84, 414-420.

Jaro MA (1995). Probabilistic linkage of large public health data files. Statistics in Medicine; 14:491–498.

Knuth D.E. (1998) The Art of Computer Programming, Volume 3: Sorting and Searching, Second Edition. Addison-Wesley Publishing Company.

Larsen, M. D., and Rubin, D. B. (2001), Iterative Automated Record Linkage Using Mixture Models, *Journal of the American Statistical Association*, 79, 32-41.

Lynch B.T., Arends W.L. (1977) Selection of a surname coding procedure for the SRS record linkage system. Washington, DC: US Department of Agriculture, Sample Survey Research Branch, Research Division.

Meng, X. L., and Rubin, D. B., (1993), "Maximum Likelihood Via the ECM Algorithm: A General Framework," *Biometrika*, **80**, 267-278.

Newcombe, H. B., Kennedy, J. M. Axford, S. J., and James, A. P. (1959), "Automatic Linkage of Vital Records," *Science*, 130, 954-959.

Newcombe, H.B. and Kennedy, J. M. (1962) "Record Linkage: Making Maximum Use of the Discriminating Power of Identifying Information" *Communications of the Association for Computing Machinery*, 5, 563-567.

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

Porter, E. H., and Winkler, W. E. (1999), "Approximate String Comparison and its Effect in an Advanced Record Linkage System," in Alvey and Jamerson (ed.) Record Linkage Techniques - 1997, 190-199, National Research Council, Washington, D.C: National Academy Press.

Whalen D, Pepitone A, Graver L, Busch J.D. (2001). Linking Client Records from Substance Abuse, Mental Health and Medicaid State Agencies. SAMHSA Publication No. SMA-01-3500. Rockville, MD: Center for Substance Abuse Treatment and Center for Mental Health Services, Substance Abuse and Mental Health Services Administration.

Winkler, W.E. (1985). Exact matching lists of businesses: blocking, subfield identification, and information theory. *Proceedings of the Section on Survey Research Methods, American Statistical Association*; 438–443.

Winkler, W. E., (1988), "Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage," *American Statistical Association*, *Proceedings of the Section on Survey Research Methods*, 667-671.

Winkler, W. E. (1989a), "Methods for Adjusting for Lack of Independence in an Application of the Fellegi-Sunter Model of Record Linkage," *Survey Methodology*, 15, 101-117.

Winkler, W. E. (1989b), "Frequency-based Matching in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Section on Survey Research Methods*, *American Statistical Association*, 778-783.

Winkler, W. E. (1990), ``String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage", *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 472-477.

Winkler, W. E. (1993), "Improved Decision Rules in the Fellegi-Sunter Model of Record Linkage," P*roceedings of the Section on Survey Research Methods*, *American Statistical Association*, 274-279, also http://www.census.gov/srd/papers/pdf/rr93-12.pdf

Winkler., W.E. (1999). The State of Record Linkage and Current Research Problems, RR99-04, U.S. Bureau of the Census, 1999. See http://www.census.gov/srd/www/byyear.html.

Winkler, W. E. (2004), "Approximate String Comparator Search Strategies for Very Large Administrative Lists," *Proceedings of the Section on Survey Research Methods*, *American Statistical Association*, CD-ROM. Also available at http://www.census.gov/srd/papers/pdf/rrs2005-02.pdf

Winkler, W. E. (2006a), "Overview of Record Linkage and Current Research Directions," U.S. Bureau of the Census, Statistical Research Division Report. Available at http://www.census.gov/srd/papers/pdf/rrs2006-02.pdf

Winkler, W. E. (2006b), "Automatically Estimating Record Linkage False Match Rates," *Proceedings of the Section on Survey Research Methods*, *American Statistical Association*, CD-ROM, also at http://www.census.gov/srd/papers/pdf/rrs2007-05.pdf

Analysis with Linked data

If, for a particular record linkage method, the proportion of valid links (PPV) is not 1, statistical analyses based on linked data can be adversely affected. Neter *et al*. (1965) studied the effect of mismatch errors in finite population sampling. They observed that a relatively small mismatch error could lead to a substantial bias in estimating the relationship between response errors and true values. Scheuren and Winkler (1993) investigated the effect of mismatch errors on the bias of ordinary least squares estimators of regression coefficients in a standard regression model and proposed a method of adjusting for the bias. Scheuren and Winkler (1997) advanced the work further with an iterative procedure that modified the regression and matching results for apparent outliers. Lahiri and Larsen (2005) consider an alternative to the bias correction method of Scheuren and Winkler (1993). For known linkage probabilities, Scheuren and Winkler (1993) obtained their estimator of regression coefficient by adjusting the bias of the ordinary least square estimator for the regression model with mismatch errors, whereas the Lahiri and Larsen (2005) proposed method provides an unbiased estimator directly for a transformed regression model. In the context of data obtained after probability linkage of administrative registers, Chambers et al. (2009) describe some approaches to eliminating this bias when parametric inference is based on solution of an

estimating equation, with an emphasis on linear and logistic regression analysis. In the context of epidemiological cohort studies, Blakley and Salmond (2002) study the effect of various mismatch errors on subsequent analyses of the association of exposure variables with the outcome.

References:

Blakely T, Salmond C. (2002) Probabilistic record linkage and a method to calculate the positive predictive value. International Journal of Epidemiology; 31(6):1246–52.

Chambers et al. (2009), Inference Based on Estimating Equations and Probability-Linked Data, Working Paper, the University of Wollongong. Available at http://ro.uow.edu.au/cgi/viewcontent.cgi?article=1037&context=cssmwp

Lahiri, P. A., and Larsen, M. D. (2005) "Regression Analysis with Linked Data," *Journal of the American Statistical Association*, 100, 222-230.

Neter, J., Maynes, E. S., and Ramanathan, R. (1965), "The effect of mismatching on the measurement of response errors," *Journal of the American Statistical Association*, 60, 1005-1027.

Scheuren, F.,and Winkler, W. E. (1993), "Regression analysis of data files that are computer matched," *Survey Methodology*, 19, 39-58, also at http://www.fcsm.gov/working-papers/scheuren_part1.pdf.

Scheuren, F.,and Winkler, W. E. (1997), "Regression analysis of data files that are computer matched, II," *Survey Methodology*, 23, 157-165, http://www.fcsm.gov/working-papers/scheuren_part2.pdf.

Record Linkage Practice

We begin this section by presenting record linkage practices in U.S. government agencies (e.g., National Center for Health Statistics, Substance Abuse and Mental Health Services Administration, and the Census Bureau) followed by practices in foreign governments (Canada and Europe), academia, and private/marketing industry. Examples of record linkage practice associated with each group are presented for practical illustration and their potential relevance to the U.S. Census Bureau.

Record Linkage Practice in US Government Agencies

Data Linkage Activities at the National Center for Health Statistics (NCHS)

The National Center for Health Statistics (NCHS) has developed a record linkage program designed to maximize the scientific value of population-based health surveys. Linked data files enable researchers to examine factors that influence disability, chronic disease, health care utilization, morbidity, and mortality. NCHS is currently linking various NCHS surveys with air monitoring data from the Environmental Protection Agency (EPA), death certificate records from the National Death Index (NDI), Medicare enrollment and claims data from the Centers for Medicare and Medicaid Services (CMS), and Retirement,

Survivor, and Disability Insurance (RSDI) and Supplemental Security Income (SSI) benefit data from the Social Security Administration (SSA).

NCHS Data Linked to Air Quality Data

NCHS collects data that describe the general and specific health of the United States population. NCHS has linked air monitoring data, available from the Environmental Protection Agency (EPA) internet site, to the National Health Interview Survey (NHIS), the National Health and Nutrition Examination Survey (NHANES) (Kravets and Parker 2008), and theNational Hospital Discharge Survey (NHDS). These linked SAS data files are available in the NCHS Research Data Center (RDC) but, because they contain geography-specific information, are not available for public-use.

NCHS Data Linked to Mortality Files

NCHS is currently linking various NCHS surveys with death certificate records from the NDI. Linkage of the NCHS survey participants with the NDI provides the opportunity to conduct a vast array of outcome studies designed to investigate the association of a wide variety of health factors with mortality.

NCHS has updated the mortality linkage of the NHIS for years 1986-2004 to death certificate data found in the NDI. The updated NHIS Linked Mortality Files provide mortality follow-up data from the date of NHIS interview through December 31, 2006. Mortality ascertainment is based primarily upon the results from a ***probabilistic match*** between NHIS and NDI death certificate records. There are two versions of the NHIS Linked Mortality Files: public-use files that include a limited set of mortality variables for adult NHIS participants and restricted-use files that include more detailed mortality information and mortality follow-up for children. Each NHIS survey year (1986-2004) is available on a separate date file.

NCHS has conducted a mortality linkage of the NHANES to death certificate data found in the NDI. The NHANES Linked Mortality Files include the years 1999-2004 and provide mortality follow-up data from the date of survey participation through December 31, 2006. Mortality ascertainment is based upon the results from a ***probabilistic match*** between NHANES and NDI death certificate records.

NCHS has also conducted mortality linkages of the 1985, 1995, 1997, and 2004 National Nursing Home Surveys (NNHS) to death certificate data found in the NDI. The NNHS Linked Mortality Files provide mortality follow-up data from the date of NNHS interview through December 31, 2006. Mortality ascertainment is based primarily upon the results from a ***probabilistic match*** between the NNHS and NDI death certificate records.

Other NCHS surveys such as, the Second Longitudinal Study of Aging (LSOA II), are also linked to NDI data.

NCHS Data Linked to CMS Medicare Enrollment and Claims Files

Medicare enrollment and claims data are available for those respondents to NCHS surveys who agreed to provide personal identification data to NCHS and for whom NCHS was able to match with Medicare administrative records. CMS provided NCHS with Medicare benefit claims data for 1991 through 2007 for all successfully matched NCHS survey participants. Many surveys (NHIS, NHANES, NNHS) conducted by NCHS are linked to CMS Medicare data.

The process of linking each NCHS survey with Medicare data began by matching individual survey respondents with Medicare's Enrollment Database (EDB). EDB is a master enrollment file of all people ever entitled to Medicare. CMS identified potential matches between NCHS survey participants and records in the EDB. CMS based potential matches on whether NCHS records matched EDB records on (1) Health Insurance Claim number, (2) SSN, or (3) name and date of birth. For these potential matches, NCHS employed a ***deterministic matching algorithm*** to determine which matches were correct. For further details see National Center for Health Statistics, Office of Analysis and Epidemiology (2010).

The linked NCHS-CMS Medicare files are restricted-use files that can be accessed through the NCHS RDC. NCHS has created Feasibility Study Data files to assist researchers who are considering submitting an RDC application to analyze the linked NCHS- CMS Medicare data.

NCHS Data Linked to Social Security Benefit History Data

Several NCHS health surveys are linked to ***five SSA Administrative Data Files***: the Master Beneficiary Record (MBR) file, the Supplemental Security Record file (SSR), the Payment History Update System (PHUS) file, the 831 Disability Master File (831) and a special extract of summarized quarters of coverage (QOC) from the Master Earnings File. An overview of the files can be found in the description of the linkage at http://www.cdc.gov/nchs/data/datalinkage/description_of_nchs_ssa_2009.pdf.

Application of Record Linkage to Estimate Multiple Program Participation for the Food Assistance & Nutrition Research Program

Administrative data from USDA's food assistance and nutrition programs (FANPs) provide statistics on the number and characteristics of program participants. However, policymakers and researchers often want more information than these administrative data provide, particularly the characteristics of families who choose to participate in some, but not all, programs for which they are eligible.

This study investigates the feasibility of linking administrative data across multiple FANPs to provide statistics on multiple-program participation. The first phase of the study is based on the Survey of Food Assistance Information Systems, taken in 26 States from directors of the Food Stamp Program (FSP), the Special Supplemental Nutrition Program for Women, Infants, and Children (WIC), and Child Nutrition Programs. The survey collected information about the characteristics and content of FANP information systems. Phase 2 of the study collected and linked 2000-02 administrative data on clients of the FSP and the Special Supplemental Program for WIC in Florida, Iowa, and Kentucky. Records from the FSP and WIC programs were matched using probabilistic record linkage software provided by the U.S. Census Bureau. Match results were used as estimates of multiple program participation within each State.

Two measures of multiple program participation are of interest in characterizing the experiences of program participants: contemporaneous participation and exposure. Contemporaneous participation is participation in multiple programs at a point in time; exposure is participation in multiple programs during an extended period, but not necessarily at the same time. For further details see (Cole 2003).

FSP and WIC do not share a common information system and do not share a common person ID. FSP and WIC records cannot be reliably linked via a merge on SSN because the SSN may not be equally reliable in the two files: FSP validates SSNs but WIC does not (according to the Phase 1 survey). Because SSNs are not validated by both programs, there is potential for false positive and false negative results from a match on SSN. For this reason, probabilistic matching was the primary approach used for this study, with deterministic matching conducted for sensitivity analyses.

Electronic Record Linkage to Identify Deaths among Persons with AIDS in the District of Columbia, 2000-2005

In 2007, to identify all deaths that occurred during 2000-2005 among persons with AIDS who resided or received their diagnosis in the District of Columbia (DC), the HIV/AIDS Administration of the DC Department of Health, with assistance from CDC, performed an electronic record linkage. The results indicated that electronic record linkage for death ascertainment is necessary to more accurately estimate the prevalence of persons living with HIV/AIDS (CDC 2008).

The DC HIV/AIDS Reporting System (HARS) is a confidential, name-based reporting system developed by CDC to manage HIV/AIDS surveillance data. HARS contains vital status information but does not contain information on cause of death. To perform the electronic record linkage, Link Plus, a free program developed at CDC, was used to link AIDS patients in the HARS data file to records in two other computer data files: 1) the DC Vital Records Division's electronic death certificate file (eDCF) and 2) the Social Security Administration's Death Master File (SSDMF). The eDCF includes all deaths that occur in

DC, regardless of state of residence, and some deaths of DC residents that occur in Maryland or Virginia. The SSDMF contains information on all deaths reported to the Social Security Administration, regardless of state of residence or where the death occurred. The eDCF has information on causes of death, but the SSDMF does not.

The variables used for record linkage were name, date of birth, Social Security number, and sex. Three linkages were performed. Linkage 1 and linkage 2 matched the HARS file to eDCF and SSDMF records, respectively, to identify deaths among persons listed in HARS with reported AIDS. HARS cases that were successfully linked to eDCF or SSDMF records were categorized by whether the death had been previously reported to HARS. To identify potential new AIDS cases never previously reported to HARS, linkage 3 identified those death certificates within eDCF that indicated HIV infection as a cause of death but had not been linked to HARS via linkage 1.

Record Linkage for the Surveillance Program to Monitor the Occurrence of Birth Defects in the Metropolitan Atlanta Area

As part of the surveillance program to monitor the occurrence of birth defects in the metropolitan Atlanta area, Jurczyk et al. (2008) developed a record linkage software tool that provides latitude in the choice of linkage parameters, allows for efficient and accurate linkages, and enables objective assessments of the quality of the linked data. They developed and implemented a Java-based fine-grained probabilistic record integration and linkage tool (FRIL) that incorporates a rich collection of record distance metrics, search methods, and analysis tools. Along its workflow, FRIL provides a rich set of user-tunable parameters augmented with graphic visualization tools to assist users in understanding the effects of parameter choices. The authors used this software tool to link data from vital records (n = 1.25 million) with birth defects surveillance records (n = 12,700) from the metropolitan Atlanta Congenital Defects Program (MACDP) for the birth years 1967–2006.

Probabilistic Record Linkage in CODES Program to Improve Highway Safety Applications at the State Level

Evolving from a need to quantify and report on the benefits of safety equipment and legislation in terms of mortality, morbidity, injury severity, and health care costs at State and national levels, the Crash Outcome Data Evaluation System (CODES) has built proactive partnerships between traffic safety and public health agencies, which own the State data, and the National Highway Traffic Safety Administration (NHTSA), which provides access to the software and training resources that make the linkage feasible. CODES uniquely uses probabilistic methodology to link crash records to injury outcome records collected at the scene and en route by emergency medical services, by hospital personnel after arrival at the emergency department or admission as an inpatient and/or, at the time of death, on the death

certificate. Analyses of linked data help inform State traffic safety professionals and coalitions to determine and implement data-driven traffic safety priorities.

CODES record linkage is conducted using CODES2000, commercially available software that implements an extension of Fellegi and Sunter's statistical theory of record linkage (Fellegi and Sunter, 1969; McGlincy, 2004, 2006). CODES2000 determines the posterior odds for a true link by applying Bayes' rule for odds (Gelman et al., 2004, pg. 9). Parameters of the linkage model are determined using Markov Chain Monte Carlo data augmentation (Schafer, 1997, pg. 72). For further details on CODES linkage methods see NHTSA (2010).

Missing values and reporting errors in the data collection processes may lead to low probabilities being assigned to many true matches. To be able to include these low-probability matches in outcome studies, CODES2000 completes five linkage imputations; that is, missing links are determined five times resulting in five complete datasets. Note that multiple imputations does not attempt to identify each missing link but instead constructs samples representative of the distribution of low to high probability links. As a result, analyses yield valid statistical inferences that reflect the uncertainty associated with having low-probability true links. Standard statistical analyses are performed on each of the five datasets and then combined to produce final results using procedures in SAS.

Integrating the New York Citywide Immunization Registry and the Childhood Blood Lead Registry

In February of 2004, the New York City Department of Health and Mental Hygiene completed the integration of its childhood immunization registry (CIR) and blood lead test registry (LQ) databases, each containing over 2 million children. A modular approach was used to build a separate integrated system, called Master Child Index (MCI), to include all children in both the immunization and lead test registries. The principal challenge of this integration was to properly align records so that a child represented in one database is matched with the same child in the other database. To accomplish this task as well as to identify internal duplicate records within each database, an artificial intelligence (AI) record linkage system was created.

Before the integration, both CIR and LQ were using custom-designed software to automatically match and merge incoming records. The LQ matching system employed a specific set of criteria, known as rules, for making matching decisions. Combined with human review by 5 full-time equivalent staff, LQ's record duplicate rate was kept at an acceptable level of an estimated 10%.

The CIR was also using a rules-based approach to automatically match incoming records. This system alone proved ineffective, resulting in a 50% duplication rate. The CIR added an automated clean-up

process using artificial intelligence (AI) matching software developed in collaboration with a consultant (Borthwick, Papadouka, Walker 2000). Immediately after its initial run, this matching system reduced the duplication rate from 50% to 15%–20%. The advantage of using AI software, instead of rules-based software, is that it can make sense of conflicting information (e.g., same first name, different spellings of last name, slightly different dates of birth [DOBs]).

In a rules-based system, each rule results in a definitive decision. A rule holding that two records with different DOBs do not belong to the same child will result in separating the records, regardless of any other similarities. In contrast, an AI system could merge these records if there are similarities that outweigh the differences in DOBs. For further discussion on the differences between AI system and rules-based system see Papadouka et al. (2004).

Substance Abuse and Mental Health Services Administration Integrated Database (IDB) Project

The Integrated Database Project is a contract jointly funded by the Substance Abuse and Mental Health Services Administration's (SAMHSA's) Center for Substance Abuse Treatment (CSAT) and Center for Mental Health Services (CHMS). The goal of the project is providing technical assistance to states for integrating and using mental health (MH), substance abuse (SA), and Medicaid data.

Whalen et al. (2001) describes the concepts behind record linking and the specific application of record linking in building databases integrating information about mental health (MH) and alcohol/drug (AOD) services. The MEDSTAT Group (a SAMHSA contractor) constructed these databases. Each Integrated Database (IDB) includes comprehensive information for MH and AOD services from the MH and AOD State Agencies, as well as Medicaid Agencies for three States: Delaware, Oklahoma, and Washington. A variety of methods have been employed to link records from different data sources and these methods vary in terms of complexity, efficiency, and accuracy. Simple matching and deterministic methods are useful for certain applications, and while these methods are relatively simple to implement, they can also produce inaccurate results. By contrast, probabilistic linking methods are relatively complex, but tend to produce more accurate results.

Linking routines are available on the SAMHSA Web site at http://www.csat.samhsa.gov/IDBSE/idb/modules/linking/recordlink.aspx. Data-linking protocols for the IDB project are written in SAS code. SAS routines are included for data-deduplication and linking algorithms.

In constructing integrated MH-SA-Medicaid databases, the following numbers exhibit the percentage of links found in employing three different methods: Probabilistic linking: 80-86%, Match merge: 51-72%,

Deterministic links: 59-76%. Probabilistic linking consistently found more links than other methods. The more sophisticated the deterministic linking, the better the results. Deterministic linking that uses blocking were the most effective of the deterministic linking methods.

Analysis of IDB data

Under confidentiality agreements with the States, data from this CSAT/CMHS Integrated Data Base (IDB) Project has been analyzed and the report can be found in Coffey et al., 2001. They presents findings from analyses of a subset of IDB records - persons with a primary mental or substance abuse disorder who are under age 65. Information about three groups of clients is presented: clients with mental disorders only (MH-only clients), clients with substance abuse disorders only (SA-only clients), and clients with dual MH+SA disorders (MH and SA clients). The study answers questions about the treatment services received by these populations under three different State auspices - the State MH and/or SA agency, Medicaid, or multiple auspices.

Integrating State Administrative Records to Manage Substance Abuse Treatment System Performance

State agencies collect a variety of information on individuals they serve or encounter, and they maintain official records as a routine part of their operations. Heil et al. (2007) describes the utility and practice of integrating the information available in State agency data sets with information on clients of AOD services.

Developing and using integrated data afford a State a readily accessible data repository for answering questions about clients (e.g., demographics, family and social arrangements, substance use), services (e.g., modalities, length of stay, funding source), and outcomes (e.g., treatment completion, employment, arrests). Data-integration efforts by a State can range from one-time linkage of selected data sets to address a particular question of interest to developing a more comprehensive integrated-data system that regularly links AOD data with one or more other agency data sets, stores the collected data, and uses such data to support reporting requirements and systemic decision making. Data-integration strategies can also enhance State efforts to identify unique clients served by the AOD system, because admission data at the encounter level can be matched to itself to detect clients who may have had multiple encounters, have more than one identity within the client-data system, or both.

Software

The record linakge software used in several U.S. federal agencies' projects have already been mentioned in the context of the project. In this section we describe only two important software packages with considerable details:

LinkPlus

Link Plus is public domain probabilistic record linkage software developed at CDC's Division of Cancer Prevention and Control in support of CDC's National Program of Cancer Registries (NPCR). Link Plus is a record linkage tool for cancer registries. It is an easy-to-use, standalone application for Microsoft Windows that can run in two modes—

- To detect duplicates in a cancer registry database.
- To link a cancer registry file with external files.

Although originally designed to be used by cancer registries, the program can be used with any type of data in fixed width or delimited format. Used extensively across a diversity of research disciplines, Link Plus is rapidly becoming an essential linkage tool for researchers and organizations that maintain public health data. It computes probabilistic record linkage scores based on the theoretical framework developed by Fellegi and Sunter (1969). For detailed features of the software see National Program of Cancer Registries (2007).

Link King

Heil et al. (2007) also describes the usefulness of the Link King software in various data linkage context. Link King conducts an elaborate deterministic evaluation and makes a deterministic decision regarding the appropriateness of a link independent of the probabilistic decision. Link King is a public domain record linkage and deduplication program developed by Washington State's Division of Alcohol and Substance Abuse (DASA). Portions of The Link King protocol were adapted from algorithms developed by MEDSTAT for the SAMHSA IDB project. The URL for the Link King site is http://the-link-king.com. Link King requires a SAS license but no SAS programming experience. Features include a data importing and formatting wizard, artificial intelligence to determine appropriate linking protocols, an interface for manual review of "uncertain" record pair matches, and an ability to generate random samples of record matches to allow for validation of matched pairs.

An extensive list of currently available record-linkage and deduplication software can be found at a comprehensive website sponsored by the Australian National University Data Mining Group (http://datamining.anu.edu.au/projects/linkage-links.html). An independent review of relatively low-cost commercially available client data-matching software (Jones & Sujansky, 2004) from the California Health Care Foundation (CHCF) is available at http://www.chcf.org/documents/ihealth/PatientDataMatchingBuyersGuide.pdf.

Census Bureau's Person Identification Validation System (PVS)

The Census Bureau uses the Person Identification Validation System (PVS)to link numerous survey, census and administrative record files. It uses a combination of deterministic and probabilistic record linkage techniques that are calibrated by analyst review and judgment of test pairs to search for persons in the input file against a set of reference files based on the Social Security Administration's Numident file, one of which is enhanced with addresses, to assign a unique person identifier, the Protected Identification Key (PIK). The current PVS process consists of an initial edit process and any or all of the three modules defined below in sequential order. That is, if both GeoSearch and NameSearch are used, GeoSearch processes the input file first, and then unmatched records are processed by NameSearch.

Initial Edit Process

Perform name and address edits. If an input referent record has a blank first name and last name, then that record is not included in the match process.

Module 1: Verification

For input records containing reported SSNs, the Verification module matches the SSN to the Census Numident and if found, compares person characteristics (name, date of birth, and sex) to those on the Numident. If the data match sufficiently, the record is assigned a protected identity key (PIK). Note that this module is a deterministic protocol but does not rely completely on total agreement of person characteristics; partial agreement (with weights above a certain threshold) is also allowed. The Verification module is not used if an input file record does not contain reported SSNs. For example, after the 2005 Current Population Survey (CPS), SSN is no longer captured in the CPS. Source records linked during verification do not proceed to subsequent modules. This module does not use address components as a matching variables.

Module 2: GeoSearch

For source person records failing the Verification module or without reported SSN, the GeoSearch module compares person-level characteristics to the Census Numident—including alternate names and alternate dates of birth—augmented with addresses to attempt to determine the PIK. The GeoSearch module is not used if input file does not contain address.

The match strategy used in this module relies on blocking factors and match variables. This module uses a hierarchical process with multiple passes through the unmatched records. Each pass uses different sets of blocking factors to sort the files and compares only those records that agree on blocking factor (sort key). Each match field has a comparison type and weight associated with the field. Any matched pair with composite score—which is the sum of scores from all match fields—over a user-defined threshold value

is considered to be linked and is assigned a PIK from the reference file. Source records without PIK in one pass proceed to the next pass that typically has less restrictive blocking features.

Module 3: NameSearch

Input file records failing the Verification and GeoSearch modules, or those without full SSN and address information, have person characteristics compared to the Census Numident—including alternate names and alternate DOB—through the NameSearch module. This is also a hierarchical process having multiple passes. Source records failing to receive a PIK in one pass proceed to next pass where the records are blocked by some criteria related to first/last name or DOB. Records that agree on blocking variables are matched on name, DOB, sex, and SSN if it is present in full or partial form. Any matched pair with a composite score (sum of the scores from all match fields) over a user-defined threshold value is considered to be linked and is assigned SSN or PIK from the reference file.

References

Borthwick A, Papadouka V, Walker D. (2000). Principles and results of the NY Citywide Immunization Registry's MEDD De-Duplication Project. Paper presented at: The 34th National Immunization Conference; Washington, DC; July 7, 2000.

Centers for Disease Control and Prevention (2008) "Electronic Record Linkage to Identify Deaths among Persons with AIDS --- District of Columbia, 2000—2005", *Morbidity and Mortality Weekly Report*, June 13, 2008 / Vol. 57 / No. 23.
Available at http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5723a4.htm

Cole, N. (2003), "Feasibility and Accuracy of Record Linkage to Estimate Multiple Program Participation". Available at http://www.ers.usda.gov/Publications/EFAN03008/

Coffey, R.M., Graver, L., Schroeder, D., Busch, J.D., Dilonardo, J., Chalk, M., & Buck, J.A. *Mental Health and Substance Abuse Treatment: Results from a Study Integrating Data from State MH, SA, and Medicaid Agencies.* SAMHSA Publication No. SMA-01-3528. Rockville, MD: Center for Substance Abuse Treatment and Center for Mental Health Services, Substance Abuse and Mental Health Services Administration.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman & Hall/CRC.

Heil, S. K. R., Leeper, T. E., Nalty, D., & Campbell, K. (2007). *Integrating State administrative records to manage substance abuse treatment system performance,* Technical Assistance Publication (TAP) Series 29. Rockville, MD: Center for Substance Abuse Treatment, Substance Abuse and Mental Health Services Administration. Available at http://kap.samhsa.gov/products/manuals/pdfs/TAP29_06-07.pdf

Jones, L., & Sujansky, W. (2004). *Patient data matching software: A buyer's guide for the budget conscious.* Oakland, CA: California Health Care Foundation.

Jurczyk, P., Lu, J.J., Xiong, L., Cragan, J.D., and Correa, A. (2008), Fine-grained record integration and linkage tool, *Birth Defects Research Part A: Clinical and Molecular Teratology*, 82, 11, 822-829.

Kravets N, Parker JD. (2008) Linkage of the Third National Health and Nutrition Examination Survey to air quality data. National Center for Health Statistics. Vital Health Stat 2(149). Available at http://www.cdc.gov/nchs/data/series/sr_02/sr02_149.pdf

McGlincy, M. A. (2004). Bayesian Record Linkage Methodology for Multiple Imputation for Missing Links. *Proceedings of the Section on Survey Research Methods*, *American Statistical Association*, 4001-4008.

McGlincy, M. A. (2006). Using Test Databases to Evaluate Record Linkage Models and Train Linkage Practitioners. *Proceedings of the Section on Survey Research Methods*, *American Statistical Association*, 3404-3410.

National Program of Cancer Registries (2007). *Link Plus*, Atlanta, GA: US Department of Health and Human Services, CDC. Install and upgrade from http://www.cdc.gov/cancer/npcr/tools/registryplus/lp.htm.

National Center for Health Statistics Statistics, Office of Analysis and Epidemiology (2010). Linkages between Survey Data from the National Center for Health Statistics and Medicare Program Data from the Centers for Medicare and Medicaid Services. Available at http://www.cdc.gov/nchs/data/datalinkage/cms_medicare_methods_report_final.pdf

NHTSA (2010). The Crash Outcome Data Evaluation System (CODES) And Applications to Improve Traffic Safety Decision Making. Available at http://www-nrd.nhtsa.dot.gov/Pubs/811181.pdf

Papadouka, V. and others (2004) Integrating the New York citywide immunization registry and the childhood blood lead registry, *Journal of Public Health Management and Practice*, **10**, p S72-S80.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall/CRC.

Thoburn KK, Gu D, Rawson T. (2007). Fundamentals of linking public health datasets. Link Plus: probabilistic record linkage software. Probabilistic Linkage Webinar 2, March 30. Available at: http://www.nri-inc.org/projects/OSA/LinkPlusOverviewMarch2007.pdf

Whalen D, Pepitone A, Graver L, Busch J.D. (2001). Linking Client Records from Substance Abuse, Mental Health and Medicaid State Agencies. SAMHSA Publication No. SMA-01-3500. Rockville, MD: Center for Substance Abuse Treatment and Center for Mental Health Services, Substance Abuse and Mental Health Services Administration.

Record Linkage Practices in Foreign Government Agencies

Canada

Background

The idea of computerized record linkage emerged in Canada with the vision of using existing administrative and health records to answer research questions relating to genetics, occupational and environmental health, and medical research (Fair 2004). Newcombe and his associates (1959) required quantitative data regarding the effects of radiation in human populations. They envisioned the possibility of using computerized record linkage of vital statistics and health surveillance records to help answer this

question. In the absence of unique identifiers on the files, Newcombe and some other researchers developed ad hoc computer programs to carry out the linkages of vital records into individual and family groupings (Newcombe 1988). The mathematical theory of record linkage work of Fellegi and Sunter (1969) at Statistics Canada was not motivated by health research issues. Rather, it was explicitly oriented to the problem of merging the information content of large administrative files in order to create a statistically useful source of new information (Fair 2004).

Unlike many countries, most statistical activity in Canada is carried out within a single national agency, Statistics Canada. Statistics Canada will carry out linkages of different records pertaining to the same individual only for statistical purposes and only when the results of the linkage would yield a potential public good which clearly outweighs the potential invasion of the privacy rights of individuals included in the linkage. This activity is conducted in accordance with the Agency's Policy on Record Linkage which has been in place since 1986. All record linkage proposals must satisfy a prescribed review and approval process which involves the submission of documented proposals to an internal expert committee. A description of all linkages approved since 2000 are available on the website http://statcan.gc.ca/record-enregistrement/summ-somm-eng.htm. Some selected examples are given below.

Economic Impacts Analysis of Public Research and Development Performers and Programs in Canada

The purpose of the project is to measure the socio-economic impact of National Research Council (NRC) science and technology (S&T) programs on participating firms. Changes over time in employment, research and development expenditures, export patterns, and other performance indicators of NRC program participants will be compared to those of a sample of similar firms who were not program participants, in order to evaluate the impact of NRC S&T programs.

The NRC S&T client list will include firms that participated in these programs from 2001 to 2006 inclusively. Statistics Canada will link, at the enterprise level, the Business Register (BR), the Longitudinal Employment Analysis Program (LEAP), the Research and Development in Canadian Industry (RDCI) Survey, the Exporter Register, and the General Index of Financial Information (GIFI) (tax) databases for reference years 2000 to 2006, to the S&T program participant file provided by NRC. The names and addresses of enterprises that were NRC S&T program participants will be used as key identifiers. A cohort of non-participating firms with similar characteristics to the NRC client firms will be selected for comparative analysis from the linked Statistics Canada files.

Canadian Forces Cancer and Mortality Study, 1972 to 2009

The purpose of the project is to measure cancer and mortality risks of current and former Canadian Forces (CF) members related to their occupational exposures. This information will assist the Department of National Defence (DND) and Veterans Affairs Canada (VAC) to develop and enhance health promotion and health protection policies and programs for serving personnel. DND and VAC do not have access to complete information on mortality and cancer outcomes of serving and retired CF personnel. Statistics Canada will undertake the Canadian Forces Cancer and Mortality Study to address these health information gaps.

DND will provide Statistics Canada with a list of approximately 312,500 personnel who enrolled on or after January 1, 1972 and have served or are still serving with the Canadian Forces at any point in the period from January 1, 1972 to December 31, 2009. The records of this CF cohort will be linked to the following files maintained by Statistics Canada: the 1984 to 2010 Tax Summary Files; the 1972 to 2007 Canadian Mortality Database (CMDB); and the 1969 to 2009 Canadian Cancer Database. Linkage to the Tax Summary Files will assist in the record linkage, the manual resolution of doubtful links, and to verify the total number in the cohort who are found alive at the end of the study period and not lost to follow up: these files contain no income data.

A random Statistics Canada-generated unique identifier will be attached to each record in the CF cohort, as well as to each record in the output file generated by the mortality linkage, and the output file generated from the cancer linkage. In addition, Statistics Canada will attach the unique identifier to each record in a DND cohort work history file and a VAC client administrative database file. This will enable linkage of the output files with the DND and VAC files by Statistics Canada, DND or VAC.

2011 Census of Agriculture to 2011 National Household Survey linkage

Linkage of the 2011 Census of Agriculture to the 2011 National Household Survey will provide a great depth of socio-economic information on farm operators, their families and their households, without increasing respondent burden. The Census of Agriculture was linked to the Census of Population for the Census years 1971, 1981, 1986, 1991, 1996, 2001 and 2006, to produce a database of socio-economic characteristics of farm operators and their families and households. The Censuses of Agriculture were linked to both the short form and the long form of the Censuses of Population.

The information previously collected by the long-form census questionnaire will be collected as part of the new voluntary 2011 National Household Survey, to be conducted shortly after the May 2011 Census of Population. Linkage of the 2011 Census of Agriculture and the 2011 National Household Survey will produce a database of socio-economic information on farm operators and their families.

Study of Doctoral Graduates: Linkage of the National Graduates Survey and the Survey of Earned Doctorates

This study is conducted to analyze the labor market outcomes of doctoral graduates from three perspectives: 1) in relation to their plans at the time of graduation; 2) in relation to whether they were pursuing a post-doctoral fellowship; and 3) in relation to mobility within the two-year period following graduation. Linking the Survey of Earned Doctorates and the National Graduates Survey introduces a longitudinal dimension to studying the pathways of doctoral graduates between the time of graduation and two years later. This linkage will differentiate doctoral graduates who were pursuing post-doctoral training and those who were not, and enable the study of differences in labor market outcomes between these two groups.

Records of doctoral graduates from the 2007 National Graduates Survey (Class of 2005) are linked to the 2004-2005 and 2005-2006 Survey of Earned Doctorates master files. The files are linked deterministically using the name of the institution where the doctorate was obtained and the graduate's first name, last name and date of birth.

Linkage to T1 Income Tax Files for Purposes of the 2006 Census Income Question

For the first time in a census, respondents will have the option of giving Statistics Canada the permission to use income information available in their 2005 income tax files (T1) as a way of answering the 2006 Census income question. The income question on the long-form questionnaire requires detailed reporting on eleven different sources of income, total income, as well as income taxes paid. Accurate reporting often requires that respondents consult their own personal records. This option, which would allow Statistics Canada to access respondent's income tax files, is offered to reduce the respondent's response burden and the time required to fill out the census long-form questionnaire. Granting this permission may also contribute to improving the quality of census data, which are widely used by all sectors of Canadian society, namely allowing a measure of after-tax income.

For only those respondents who gave permission, information corresponding to the 2006 Census income question on the long-form questionnaire will be retrieved from their 2005 personal income tax files (T1). In order to allow for an evaluation of the quality of data, the linking keys with personal identifiers will be maintained until June 2010, after which they will be destroyed. Respondents who prefer not to grant permission will be required to fill in the information on the 2006 Census long-form questionnaire.

Socio-Economic Status and Perinatal Health: Linkage of the Nova Scotia Perinatal Database to the T1 Family File

There are two studies to be derived from this linkage. The first is to examine further the relationship between socioeconomic status and the receipt of health services. There have been changes in maternal characteristics, in obstetric practices, in technologies and in health care funding over the past 10 years, all of which have had consequences for perinatal health in Canada. The study will review health care delivery and infrastructure programs in light of these new emerging issues and their connection with longer-term temporal trends among lower and higher socioeconomic status women.

The second study will further examine the relationship between socioeconomic status and perinatal outcomes: the expanded years of data will allow researchers to assess relationships for less frequent outcomes. Results of these two studies may lead to a review of health care delivery and infrastructure, as relevant programs can be effectively targeted and modifications introduced where appropriate.

This study involves linkage of the Nova Scotia Perinatal Database (NSPD) to the T1 Family File (T1FF), for the years 1988 to 1995. This one-time linkage adds data for the years 1996 through 2003. Thus, the study period will range from 1988 to 2003.

Understanding the Early Years - National Longitudinal Survey of Children and Youth Community Component Linkage

This linkage will help to increase knowledge about the development of children in the early years, monitor progress in improving outcomes for young children and encourage community action to improve children's readiness to learn when entering school. The results could change certain factors in our communities which would put children in a better position to learn and succeed at school.

Human Resources Development Canada (HRDC) has contracted with Statistics Canada to carry out the Understanding the Early Years (UEY) interviews using the same instruments developed for the National Longitudinal Survey of Children and Youth (NLSCY). The project also includes the collection of information using the Early Development Instrument (EDI) developed by McMaster University under contract to HRDC. The information collected by McMaster University from teachers using the EDI will be added to the information collected by Statistics Canada for the UEY project.

Linkage between the Survey of Labour and Income Dynamics and Federal Child Tax Benefit Programs

The Survey of Labour and Income Dynamics (SLID) is designed to measure changes in the economic well-being of individuals and uncover the factors that influence those changes. The SLID is the main source of information on individual and family income.

SLID respondents have the option of giving Statistics Canada permission to access their income tax returns and extract the data required by the income questions in the SLID. It reduces the response burden and improves data quality. Personal income data from administrative files are generally of higher quality than data obtained directly from survey respondents. To date, Statistics Canada has been using T1 Income Tax and Benefit Returns. This linkage will provide access to new information: Child Tax Benefit (CTB) data held by the Canada Revenue Agency.

SLID respondents take part in the survey for up to six years. In the first year, the data supplied by respondents who have agreed to the linkage are statistically matched with the T1 file using six key variables: last name; first name; postal code of residence; date of birth; marital status; and spouse's first name. When there is a match, the respondent's income data and social insurance number (SIN) are saved. In subsequent years, the SIN is used for matching with the T1 file. Linkage with the CTB file is based on the SIN. Since the Canada Revenue Agency produces two CTB files per year, linkage will be performed twice a year for every year the respondents participate. The data extracted from the tax files will be combined with the other information provided by respondents in the survey, and then stored in the SLID database.

Post-Censal Survey - Aboriginal Peoples Survey (APS)

The Aboriginal Peoples Survey is mandated under the federal government's Aboriginal action plan "Gathering Strength". It will provide a profile of the lifestyles and living conditions of the Aboriginal populations, for both adults and children resident on and off-reserves. Special components will provide further insight into the situation of the Métis and the Inuit. APS will offer comprehensive information on subjects such as: employment, education, language, tradition, technology, health, social issues and housing that will be used by Aboriginal peoples and by governments at all levels for the development of policies and programs designed for Aboriginal peoples.

The APS was conducted in the fall of 2001 and the spring of 2002. Four questions on the 2001 Census of Population served to identify the target population and to draw the sample for the survey. There are two different linkages. The first involves linking the survey respondent's own census data to the APS master file. This activity adds information on the respondent's socio economic characteristics to the APS eliminating the need to collect this information. The second type of linkage involves deriving variables from the Census data pertaining to the respondent's family or household members. This linkage activity will select data from the census records of family members, derive a "family" level variable and place this information on the respondent's record on the APS file.

Software

Large-scale record linkage using probabilistic matching techniques is done at Statistics Canada using the Generalized Record Linkage System (GRLS). The current version of GRLS (version 4) runs in a client-server environment with ORACLE and a C compiler. The software will also run on a PC or workstation that supports the UNIX operating system (Fair 2004). The GRLS is particularly suited to applications where there are no unique identifiers available to carry out the linkage.

The New York State Intelligence and Information System (NYSIIS) and Russell Soundex code routines are available in the GRLS. However, in most of the Statistics Canada project the phonetic coding of name and address, postal code conversions are done outside the record linkage system, and preferably using SAS at the data handling and pre-processing phase. Based on statistical decision theory, GRLS breaks the linkage operation into three major phases: 1) A searching phase, 2) A decision phase, and 3) A grouping phase. For further details see Fair (2004).

References (For references not in this list check the previous references lists)

Fair M. (2004) Generalized Record Linkage System – Statistics Canada's record linkage software. *Austrian Journal of Statistics*; 33(1&2): 37-53.

European Countries

United Kingdom

Background

The first reported application of record linkage in the United Kingdom (UK) was in the area of health studies, where it was used to link patient records from hospitals and death certificates in order to study morbidity and mortality - the Oxford Record Linkage Study (ORLS) and numerous occupational health studies are typical examples (Gill 2001). More recently linkage has been making inroads in official statistics in the UK - the Office for National Statistics (ONS) Longitudinal Study (ONS, 1995) is a typical example, and its role is expected to increase. The ONS Methodology Directorate now has a small team of staff dedicated to working on record linkage methodology.

The Oxford Record Linkage Study

One of the pioneering practical studies of record matching in the UK health field was undertaken by the Oxford Record Linkage Study (ORLS).The initial data were limited to brief extracts of each birth, hospital inpatient discharge, and death for a population of about 350,000, and it was hoped to link these data together using the National Health Service (NHS) number. In practice only a minority of the records contained the NHS number and the decision was made to adopt and use the probabilistic linkage method.

The study was extended to include the entire Oxford health region by 1985 with a population of 2.5 million. The applications using the ORLS file include the statistical analysis of person-based longitudinal files, and tables of hospital morbidity rates for a range of conditions. Later developments include: studies of the association between diseases; outcomes; and studies of the health services. For further details see Gill (2001).

The ONS Longitudinal Study

The ONS Longitudinal Study started in 1974 with a one per cent sample drawn from the resident population of England and Wales enumerated at the 1971 census and containing Census and vital events. Subsequent samples have been drawn and linked from the 1981 and 1991 censuses. Linkage of the data became possible with the recording of date of birth rather than age in the decennial census and in birth and death registration. The matching of the file is undertaken by the National Health Service Central Register (NHSCR) using data from subsequent censuses, the national cancer registry, and death certificates. The longitudinal study contains selected records arranged in personal cumulative files. The uses of the longitudinal study include the analysis of occupational mortality, and to provide better information on fertility and birth spacing. Further uses include the analysis of migration and other socio-demographic studies (Gill 2001).

Use of School Census data to Improve Population and Migration Statistics

Good quality population and migration statistics are essential for providing the evidence base for managing the UK economy, planning, and allocating resources. Improving the quality and range of these statistics is a priority for the ONS. The initial assessment is that the School Census is a good quality data source that offers potential for improving migration and population statistics (ONS 2009). Its main strength is that it has very good coverage for a defined subset of the population (i.e. children 5-15 within England). It also collects a broad range of demographic information, including variables such as language and ethnicity which are not available from other administrative data sources.

The School Census collects data on schools and pupils in England and is administered by the Department for Children, Schools, and Families (DCSF). The School Census is now carried out three times each year. School Censuses are also carried out in Wales, Scotland, and Northern Ireland. There are variations in frequency, timing, content, and coverage among the constituent countries of the UK. The pupil level data available from the School Census give individual level information on a range of variables (unique pupil number, name, date of birth, sex, first language, ethnic group, address, school identifier, etc.). Some variables, such as unique pupil number, name, and date of birth, can be used for linking information

within the School Census across time, and also linking with other data sources. Research is now underway on the potential uses of pupil level data and data linkage.

Scotland Community Health Index and NHSCR

Record linkage methods are to link the Community Health Index and the National Health Service Central Register (NHSCR) in Scotland to provide a basis for a national patient index. The linkage used a combination of deterministic and probability matching techniques. A best-link principle was used by which each Community Health Index record was allowed to link only to the NHSCR record with which it achieved the highest match weight. This strategy, applied in the context of two files which each covered virtually the entire population of Scotland, increased the accuracy of linkage approximately a thousand-fold compared with the likely results of a less structured probability matching approach. By this means, 98.8% of linkable records were linked automatically with a sufficient degree of confidence for administrative purposes. For further details about the study see Kendrick et al. (1998).

Software

Gill (1999) describes the major features of the Oxford record linkage system (OX-LINK), with its use of the Oxford name compression algorithm (ONCA), the calculation of the names weights, the use of orthogonal matrices to determine the threshold acceptance weights, and the use of combinational and heuristic algebraic algorithms to select the potential links between pairs of records.

The system was developed using the collection of linkable abstracts that comprise the ORLS, which includes 10 million records for 5 million people and spans the period from 1963 to date. The linked dataset is used for the preparation of health services statistics, and for epidemiological and health services research. The policy of the Oxford unit is to comprehensively link all the records rather than prepare links on an ad-hoc basis.

The OX-LINK system has been further developed and refined for internally cross matching the whole of the NHSCR against itself (57.9 million records), and to detect and remove duplicate pairs, as a first step towards the issue of a new NHS number to everyone in England and Wales. A recent development is the matching of general practice (primary care) records with hospital and vital records to prepare a file for analyzing referral, prescribing and outcome measures.

References

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

Gill, L.E. (2001). Methods for automatic record matching and linking in their use in National Statistics. GSS Methodology Series, NSMS25. Office for National Statistics, UK.

Kendrick S W, Douglas M M, Gardner D and Hucker D (1998). The best-link principle in the probability matching of population datasets: the Scottish experience in linking the Community Health Index to the National Health Service Central Register, *Methods for Information in Medicine*, 37, 64-68.

Office for National Statistics (1995). Longitudinal Study 1971-1991: History, Organisation and Quality of Data. TSO (London: 1995).

Office for National Statistics (2009). Use of School Census data to Improve Population and Migration Statistics, Research Paper.

Netherlands

Background

Statistics Netherlands is entitled by law to use and link micro data from registers and various other data sources, but only for statistical purposes and with stringent obligations to ensure data confidentiality. Making maximum use of existing data sources reduces the need for primary data collection by means of surveys. Moreover, it is expected that linking the available sources will create an important asset, as it enables integrated analysis of (professional) health data and socio-economic data. In addition it is possible to do follow-up studies of population groups in time, and thus produce longitudinal statistics.

The Netherlands population register (PR) is used as the backbone of the linkages with other person-related data sources. The PR contains demographic and household information on all residents of the Netherlands, and has been available electronically since 1995. Statistics Netherlands receives micro data (at the person level) from the PR for a set of demographic variables, such as birth, death, registered address, country of birth, and familial relations. For the purpose of linkage, Statistics Netherlands includes these data with the individual changes in a longitudinal database.

Virtual Census 2001

The last traditional population census in the Netherlands happened in 1971. Since then the willingness of citizens to participate in a census declined because of privacy considerations. Census information is still necessary for policy and research purposes. For the 1981 and 1991 Census Rounds, demographic data were drawn from the Population Register. Data on socio-economic characteristics, such as on labor and education, were provided by the Labor Force Survey. These sources, however, were used separately, which means that no special attention was paid to coherence of the information at the micro-level.

For the Census 2001 Program, Statistics Netherlands launched a new approach, which is unique in Europe, known as the Virtual Census (Linder 2004). The advantages are its low response burden on the population and considerably lower costs. The Virtual Census 2001 uses the Social Statistical Database

(SSD) as its source. The SSD contains a huge amount of data on demographic and socioeconomic issues. It is constructed by micro-linking several administrative registers and household sample surveys. A micro-integration process ensures coherence, consistency and completeness of the SSD data. The data sources for the SSD include the PR, the Employee Insurance Schemes Registration System (data on employees and unemployment insurance), the Survey on Employment and Earnings, the FiBase-register (data on labor and social security income that is subject to advance tax payments, life insurances and pensions from former activities), the Social Assistance Benefits Administration, and the Labor Force Survey. For further details on these sources see (Linder 2004).

Approximately 2%-3% of the LFS records could not be linked to the PR. All together this is a good result, but selectivity in the micro-linkage process is not to be ruled out. Analysis in the past has indicated that young people in the 15-24 age group show a lower linkage rate in household sample surveys than other age groups. The reason for this is that they move more frequently, therefore they are often registered at the wrong address. The linking rate for persons living in the four large cities of Amsterdam, Rotterdam, The Hague, and Utrecht is lower than for persons living elsewhere. Ethnic minorities also have a lower linkage probability, among other things because their date of birth is often less well registered (Arts et al., 2000).

Record Linkage of Hospital Discharge Register with Population Register

The development of a population-based health statistics dataset is a part of a strategic research project at Statistics Netherlands. The aim is to build a system of coherent information on use of medical services and health status by linking the available national data sources, i.e., medical registers, registers with socio-economic data, and survey data. With regard to the health data sources, Statistics Netherlands has kept the causes of death register since 1901 and has conducted the national health interview survey since 1981. Both these sources are linked to the PR from 1995 onwards. However, with a view to building a more comprehensive health statistics database, Statistics Netherlands has started to explore external data sources that may be used for this purpose. Bruin et al. (2004) explored the national hospital discharge register (HDR) as the first external register because of the economic importance of this aspect of health care and because the register has a high coverage rate. The HDR contains data on hospital admissions, covering all general and university hospitals and most specialized hospitals. Information is collected on both in-patients and day patients. The information concerns administrative patient data, admission and discharge data, diagnoses, surgical procedures, and the medical specialties concerned.

1995–2001 data were selected from the HDR. These files were first adjusted to a uniform population of hospitals and admissions. Furthermore, the records of patients admitted to Dutch hospitals but not

resident in the Netherlands were excluded from the HDR data, as Statistics Netherlands' population statistics are based on the resident population, registered in the Dutch population register. For the purpose of record linkage, Statistics Netherlands has created a central linkage file of persons (CLFP), a longitudinal file containing persons registered in the PR. The CLFP starts on January 1, 1995 and is updated to the year before the current calendar year. For details on the linkage method, matching variables see Bruin et al. (2004). Overall, 87.6% of all the HDR records were uniquely linked to a person record in the PR.

Software

Statistics Netherlands has developed a software for Statistical Disclosure Control, called ARGUS. No reference has been identified on use of particular software in linking records.

References

Arts, C.H., Bakker, B.F.M., and van Lith. F.J. (2000). Linking Administrative Registers and Household Surveys'. *Netherlands Official Statistics*, Vol. 15 (Summer 2000): Special Issue, *Integrating Administrative Registers and Household Surveys*, ed. P.G. Al and B.F.M. Bakker: pages 16-22.

De Bruin, A., Kardaun, J., Gast, F., de Bruin, E., van Sijl, M., and Verweij, G. (2004). Record linkage of hospital discharge register with population register: experiences at Statistics Netherlands, *Statistical Journal of the United Nations Economic Commission for Europe*, 21, 23-32.

Linder, Frank (2004). The Dutch Virtual Census 2001: A new approach by combining Administrative Registers and Household Sample Surveys, *Austrian Journal of Statistics*, volume 33, pp. 69-88.

Italy

The Italian National Statistical Institute is mostly engaged in statistical matching as described in Section 2. However, in the context of linking epidemiological registries, Maso et al. (2001) describe the program software for automated linkage in Italy (SALI), aimed at matching individual records from medium-sized registries (in the order of 100,000 records), where the desired outcome is to miss as few links as possible and, because of low link-likelihood (>1%), a manual revision of matched pairs is feasible.

References

L. Dal Maso, C. Braga, and S. Franceschi (2001). Methodology Used for Software for Automated Linkage in Italy (SALI). Journal of Biomedical Informatics, 34:387–395.

Australia

The Australian Bureau of Statistics (ABS) applies data linking methods and quality measures to Australian census data. They use blocking and linking variables and conduct quality assurance by using meaures of linkage quality. The ABS seems to be extensively using the methods developed by Winkler, whose work is based on classic approaches by Fellegi and Sunter. Bishop and Khoo (2006) describe

recent developments in data linking at the ABS, review the data linking methodology and quality measures they have considered, and present results using the Australian Census Dress Rehearsal data. A goal was to develop a Statistical Longitudinal Census Data Set (SLCD) by choosing a 5% sample of people from the 2006 Australian population census to be linked probabilistically with subsequent censuses. ABS plans to enhance SLCD further by probabilistically linking it with births, deaths, immigration settlements or disease registers.

References

Glenys Bishop and Jonathon Khoo (2006). Methodology of Evaluating the Quality of Probabilistic Linking. Proceedings of Statistics Canada Symposium 2006. Methodological Issues in Measuring Population Health.

University Programs in Data Integration

Stanford Entity Resolution Framework

In this section, we briefly describe different approaches for entity resolution (ER), particularly in the context of the Stanford Entity Resolution Framework (SERF) project. The review paper by Brizan and Tansel (2006) can be considered as an excellent starting point for various entity resolution methods and its relationship with record linkage (RL) techniques. Entity Resolution (ER) (also referred to as deduplication) is the process of identifying and merging records judged to represent the same real-world entity. Although entity resolution and record linkage are conceptually similar, there is a basic implementation difference between the two approaches. Record linkage techniques are meant to link records from multiple databases, where as the ER methods are useful for identifying non-identical duplicates in a database and merging the duplicates into a single record.

The basic ER problem may arise in many applications. For example, consider a comparative shopping website, aggregating product catalogs from multiple merchants. Identifying records that match, i.e., records that represent the same product, is challenging because there are no unique identifiers across merchant catalogs. A given product may appear in different ways in each catalog, and there is a fair amount of guesswork in determining which records match. Deciding if records match is often computationally expensive. Merging records that match is often also application dependent. Most existing work on ER focuses on developing techniques to achieve the best quality for ER, measured in terms of precision and recall, on some class of data or applications. But the goal of the SERF project is to develop a generic infrastructure for ER. In their generic approach, the methods are dependent on the black-box

functions, and their focus is rather on the framework and algorithms in which these black-boxes are used (Benjelloun et al. 2006 and some other papers from the website http://infolab.stanford.edu/serf/).

The generic ER model of SERF (Benjelloun et al. 2006) is based on two black-box functions provided as input to the ER computation: match and merge. A match function M is a function that takes two records as input and returns a Boolean value. Function M returns true if the input records represent the same entity, and false otherwise. Such a match function reflects the restrictions SERF researchers are making that: (i) matching decisions can be made "locally", based on the two records being compared; and (ii) that such decisions are Boolean, and not associated with any kind of numeric confidence. The match function is based on a particular important attribute being equal, or all attributes of the records being highly similar. A merge function μ is a function that takes in two records and returns a single record. Function μ is only defined for pairs of matching records, i.e., records known to represent the same entity. Its output is a "consolidated" record representing that entity. Another important concept defining generic ER is domination. Intuitively, if two records r1 and r2 are about the same entity but r1 holds more information than r2, then r2 is useless for representing this entity and r1 dominates r2. The generic ER methods include the domination rule in the match and merge functions.

There are two main characteristics of SERF entity resolution approach: (a) in general, they do not assume any knowledge about which records may match, so all pairs of records need to be compared using the match function; and (b) merged records may lead to discover new matches, therefore a "feed-back loop" must compare them against the rest of the data set. Benjelloun et al. (2006) define entity resolution in a more formal way as follows: given a set of input records R, an ER of R, denoted ER(R) is a set of records such that:

- Any record in ER(R) is derived (through merges) from records in R;
- Any record that can be derived from R is either in ER(R), or is dominated by a record in ER(R);
- No two records in ER(R) match, and no record in ER(R) is dominated by any other.

They also introduce four simple and practical conditions on the match and merge functions, which guarantee that ER is "consistent", i.e., that it exists, is unique and finite. For a detailed definition of the properties see Benjelloun et al. (2006).

Software

The generic ER methods developed by SERF can be implemented by the SERF software and can be downloaded from http://infolab.stanford.edu/serf/. This package provides an implementation of the R-

Swoosh algorithm (Benjelloun et al. 2009). R-Swoosh is the most efficient ER algorithm, out of the 3 compared by SERF researchers, that satisfy four properties mentioned earlier. The algorithm takes as input a dataset of records (in XML) and a "MatcherMerger" class that implements functions to match and merge pairs of records, and returns a dataset of resolved records. A sample dataset of product records, along with a simple MatcherMerger implementation are provided as an example. Products are matched based on the similarity of their titles and prices.

References

Birzan, D.G., Tansel, A.U. (2006) 'A Survey of Entity Resolution and Record Linkage Methodologies', *Communications of the IIMA*, Vol. 6, No. 3, pp. 41-50.

Omar Benjelloun, Hector Garcia-Molina, Hideki Kawai, Tait Eliott Larson, David Menestrina, Qi Su, Sutthipong Thavisomboon, Jennifer Widom (2006). Generic Entity Resolution in the SERF Project. *IEEE Data Engineering Bulletin*, June 2006.

Omar Benjelloun, Hector Garcia-Molina, David Menestrina, Qi Su, Steven Euijong Whang, Jennifer Widom (2009). Swoosh: A Generic Approach to Entity Resolution, *The VLDB Journal,* January 2009.

University of Arkansas Entity Resolution Program

The Center for Advanced Research in Entity Resolution and Information Quality (ERIQ) was established to advance research and best practices in the areas of entity resolution and information quality. ERIQ is located at Donaghey College of Engineering and Information Technology (EIT) at the University of Arkansas at Little Rock (UALR). The Center was established in 2006 to support research and advanced studies by the faculty, students, and research partners of the UALR Information Quality Graduate Program. In 2009, the Center was designated as a statewide research center by the Arkansas Department of Higher Education. Infoglide Software Corporation of Austin, Texas (www.infoglide.com), and the UALR ERIQ Laboratory signed a collaboration agreement providing ERIQ access to Infoglide's Identity Resolution Engine (IRE) software for evaluation and research.

University of Maryland

The University of Maryland has a research institute named the University of Maryland Institute for Advanced Computer Studies (UMIACS). The mission of UMIACS is to foster and enhance interdisciplinary research and education in computing across the College Park campus. The UMIACS faculty conducts research programs covering a broad range of areas, addressing both fundamental core computer science issues and fundamental problems at the interface between

computer science and other disciplines. While the program is not an entity resolution program per se, research on entity resolution has been conducted (Bhattacharya and Getoor, 2006).

Bhattacharya, I. and Getoor, L. (2006) "A Latent Dirichlet Model for Unsupervised EntityResolution," *Proceedings of the Sixth SIAM International Conference on Data Mining*.

Record Linkage Practice in Private Industry

There are many industries in private industry that use record linkage or entity resolutions. Axiom is a well know professional services firm that supplies marketing research information to clients. Based in Little Rock, AR, the company has developed relationships with the Center for Advanced Research in Entity Resolution and Information Quality at the University of Arkansas Little Rock. In the marketing research area the company provides clients with detailed demographic, statistical, and trend analysis, some of which is based on records linked across database.

In the field customer relations management (CRM) record linkage means obtaining related data elements of value from account record. Examples of linked records include contacts, credit reports, or other global sites within a corporate family tree. Providing the linked attributes around account records enables marketing to define and report progress in target markets and sales to better qualify prospects and determine decision-makers.

Little is known about the record linkage or entity resolutions methods that are used in practice. At the 1997 FCSM conference on record linkage, Ivan Fellegi expressed his concerns about record linkage practices in the private sector. "Not that I know much about it—and I suspect the same applies to most of you. But this is precisely the sign of a potentially very serious problem: the unrecognized and undiscussed threat of privately held data banks and large scale record linkage."

Appendix B: List of Fake and Incomplete Names

The following lists of fake and incomplete names were extracted from the PVS unmatched records of the ACS 2009 incoming file. Frequency distributions of first and last names in the unmatched records were constructed. The fake and incomplete names were those names among names with frequency counts of 50 or higher that were: blank, one- or two-characters long, or names that are usually considered fake, for example, anonymous. The designation [blank] means that the name field was empty.

List of First Names Considered Fake or Incomplete

[blank]	GIRL	MOM
A	GOH	MOTHER
ADULT	GRANDCHILD	MR
ADULT MALE	GRANDDAUGHTER	MRS
B	GRANDSON	MS
BABY	H	N
BOY	HIJA	NEPHEW
BROTHER	HIJO	NINO
C	HOUSE	O
CHILD	HUSBAND	OLDEST
CHILD F	INMATE	ONE
COH	J	P
D	K	PERSON
DAD	KID	R
DAU	L	RESIDENT
DAUGHTER	LADY	RESPONDENT
DAUGHTER OF	LADY IN THE	S
DOH	LADY OF	SENOR
E	LADY OF HOUSE	SENORA
F	LADY OF THE	SISTER
FATHER	LOH	SOH
FEMALE	M	SON
FEMALE CHILD	MALE	SON OF
FRIEND	MALE CHILD	T
G	MAN	V
GENT	MAN IN THE	W
GENTELMAN	MAN OF	WIFE
GENTLE	MAN OF THE	WOMAN
GENTLEMAN	MINOR	YOUNGEST
GENTLEMAN OF	MISS	
GENTLEMEN	MOH	

List of Last Names Considered Fake or Incomplete

[blank]	HH	OF THE HOUSE
A	HHM	ONE
ADULT	HOME	OWNER
ANON	HOUSE	P
ANONYMOUS	HOUSEHOLD	PARENT
APELLIDO	HOUSEHOLDER	PERSON
B	HUSBAND	R
BOY	J	REF
C	K	REFUSE
CASA	L	RESIDENT
CHILD	LADY	RESP
COH	LADY OF HOUSE	RESPONDANT
D	LADY OF THE HOUSE	RESPONDENT
DAUGHTER	LAST NAME	S
DE CASA	LOH	SOH
DE LA CASA	M	SON
DECLINED	MALE	T
DOE	MAN	THE HOUSE
DOH	MAN OF THE HOUSE	THREE
DONT KNOW	MOH	TWO
E	N	UNK
F	NA	UNKNOWN
FEMALE	NO	W
FOUR	NO LAST NAME	WIFE
FRIEND	NO NAME	X
G	NONE	XXX
GIRL	O	Y
GOH	OCCUPANT	YOUNGER
H	OF HOUSE	
H AGE	OF THE HOME	

The following fake/incomplete names are in the PVS lookup reference table located at **prbu01:/pvs/pvs/code-template/ver-4/pbde_fakenamelist.dat**. The survey version PVS name-editing step attempts to remove these fake names from the incoming file records by setting them to blank. Records with both first and last names blank are not assigned a PIK. Thus, setting a record with one of these names to blank may result in the removal of the record from PVS processing. The fake name list NORC used was created from unmatched records processed by PVS. Therefore, even though some of the names in NORC's list appear below, the names were not set to blank by the PVS name-editing step.

(CONFIDENTIAL)
(NO MIDDLE NAME)
A RELUCTANT
ADULT
ADULT F
ADULT FEMAL
ADULT FEMALE
ADULT M
ADULT MALE
ADULT ONE
ADULT THREE
ADULT TWO
ADULTFEMALE
ADULTMALE
ADULTONE
ADULTTWO
ADYULT
ANON
ANONYMOUS
ANONYMOUS LADY
ANOTHER
ANOTHER BABY
AS ABOVE
AT THIS ADDR
AT THIS ADDRE
AU PAIR
AUNT
BABY
BABY BOY
BABY DAUGHTER
BABY FEMALE
BABY FEMALE DAUGHTER
BABY DOE
BABY GIRL
BABY MALE
BABY MALE SON
BABY OF THE
BABY SON
BABYBOY
BABYGIRL
BOARDER MALE
BOY
BOY !
BOY CHILD
BOY DOE
BOY FOUR
BOY FRIEND
BOY GRANDCHIL
BOY GRANDCHILD
BOY GREAT
BOY OF
BOY OLDER
BOY ONE
BOY TWO
BOY YOUNGER
BOYCHILD
BOYFRIEND
BROTHER
BROTHERINLAW
CH1LD
CHIL
CHILD
CHILD 11
CHILD A
CHILD AGE
CHILD B
CHILD BOY
CHILD C
CHILD D
CHILD DOE
CHILD F
CHILD F OF D
CHILD FEM
CHILD FEMALE
CHILD FIVE
CHILD FOUR
CHILD GIRL
CHILD HOUSE
CHILD I
CHILD IN
CHILD M
CHILD M OF D
CHILD MALE
CHILD NO
CHILD OF
CHILD OF HOUSE
CHILD OF HOUSEHOLD
CHILD OF THE
CHILD OF THE H
CHILD OF THE HO
CHILD OF THE HOU
CHILD OF THE HOUS
CHILD OF THE HOUSE
CHILD ONE
CHILD REFUSED
CHILD SON
CHILD THREE
CHILD TWO
CHILD YOUNGER
CHILD1
CHILDREN
CONFIDENTIAL
CONFIDENTIAL FOSTER
CONFIDENTIAL WILL NO
D WOMAN
DAD
DAUGH
DAUGHETER
DAUGHT
DAUGHTER
DAUGHTER A
DAUGHTER B
DAUGHTER IN
DAUGHTER NAME
DAUGHTER OF
DAUGHTER OF D
DAUGHTER OF T
DAUGHTER OF THE H
DAUGHTER OF THE HO
DAUGHTER OF THE HOU
DAUGHTER OF THE HOUS
DAUGHTER OLD
DAUGHTER ONE
DAUGHTER THREE
DAUGHTER TWIN
DAUGHTER TWO
DAUGHTER YOUNGER P
DAUGHTER YOUNGER PER
DAUGHTER1
DAUGHTERII
DAUGHTERINLAW
DAUGHTERONE
DAUGHTERS
DAUGHTERTHRE
DAUGHTERTWO
DAUGHTETER
DAUGHTR
DAUGNTER
DAUGTHER
DAUGYHTER
DAUHTER OF
DDAUGHTER
DECEASED
DECEASEDWIFE
DECLINE
DECLINE TO STATE
DECLINED
DOE BOYFRIEND
DOES NOT EXIST
DON T KNOW
DONT KNOW
DOUGHTER
ELDERDAUGHTER
ELDEST BOY
ELDEST GIRL
F CHILD
F CHILD RESID
F HEAD OF H
F HEAD OF HOUSE
F OCCUPANT
FATHER
FATHER IN
FATHER IN LAW
FATHER OF
FATHER OF CHILDREN
FATHER OF THE
FATHERINLAW
FEAMLE CHILD
FEAMLE SHILD
FEMAALE
FEMAIL
FEMAL CHILD
FEMALE
FEMALE A
FEMALE ADULT
FEMALE AGE
FEMALE B
FEMALE CHIL
FEMALE CHILD
FEMALE CHILD A
FEMALE D
FEMALE DAUGHTER
FEMALE FRIEND
FEMALE GRANDC
FEMALE GRANDDAUGHTER
FEMALE H OF HH
FEMALE HEAD
FEMALE HEAD OF HOU
FEMALE HEAD OF HOUS
FEMALE HEAD OF HOUSE
FEMALE HH
FEMALE HHM
FEMALE HOH
FEMALE HOUSEHOLD M
FEMALE HOUSEHOLD MEM
FEMALE I
FEMALE II
FEMALE III
FEMALE M
FEMALE NUM
FEMALE OCCUPANT
FEMALE OF
FEMALE OF HOUSE
FEMALE OF HOUSEHOL
FEMALE OF HOUSEHOLD
FEMALE OF THE
FEMALE ONE
FEMALE PARENT
FEMALE PERSON
FEMALE RES
FEMALE RESI
FEMALE RESIDE
FEMALE RESIDENT
FEMALE SISTER
FEMALE TEEN
FEMALE TWO
FEMALE YR
FEMALE YRS
FEMALECHILD
FIRST
FIRST CHILD
FIRST DAUGHTE
FIRST FEMALE
FIRST MALE
FIRST SON
FOSTER CHILD
FOSTER CHILD VARIOUS
FOSTER CHILD-VARIOUS
FOSTER DAUGHTER
FOSTER PERSON
FOURTH CHILD
FOURTH SON
FRIEND
FRIEND OF
FRIENDCHILD
FRIENDS SON
FROM SNLAW
FRONT OF BOOK
FRONT PAGE
G CHILD
G O H
GENT
GENT OF HOUSE
GENT OF THE
GENTLEMAN
GENTLEMAN OF
GENTLEMAN OF HOUSE
GENTLEMAN OF HOUSEHO
GENTLEMAN OF HOUSEHO
GENTLEMAN OF THE
GENTLEMAN OF THE H
GENTLEMAN OF THE HO
GENTLEMAN OF THE HOU
GENTELMEN
GENTLEMEN
GENTLEMEN OF
GENTLEMEN OF THE H
GENTLEMEN OF THE Ho
GENTLEMEN OF THE HOU
GENTLEMEN OF THE HOUS
GENTLEMEN OF THE HOUSEHO
GENTLMAN
GIRL
GIRL1
GIRL A
GIRL B
GIRL CHILD
GIRL CHILD OF
GIRL DOE
GIRL GRANDCHI
GIRL III
GIRL OF
GIRL ONE
GIRL YOUNGEST
GIRL YRS
GIRLCHILD
GIRLFRIEND
GIRLTODDLER
GOD DAUGHTER
GOD SON
GOH
GR DAUGHTER
GRANCHILD
GRAND DAUGHTER
GRANDAUGHTER
GRANDCHILD
GRANDCHILD II
GRANDCHILD1
GRANDDAUGHTE
GRANDDAUGHTER
GRAND FATHER
GRANDFATHER
GRANDMA
GRAND MOTHER
GRANDMOTHER
GRANDPARENT
GRANDSON
HEAD FEMALE OF THE
HEAD MALE OF THE
HEAD OF
HEAD OF HOUSE
HEAD OF HOUSE HOLD
HEAD OF HOUSEHOL
HEAD OF HOUSEHOLD
HEAD OF HS
HEAD OO HOUSEHOLD
HEADHOUSEHOLD
HEADOFHOUSE
HH F
HH FEMALE
HH M
HH MALE
HH MEMBER
HOME OWNER
HOMEMAKER
HOMEOWNER
HOMEOWNER NUMBER O
HOMEOWNER NUMBER ONE
HOMEOWNER NUMBER T
HOMEOWNER NUMBER TWO
HOUSBAND
HOUSE
HOUSE HOLD
HOUSE MALE
HOUSEHOLD
HOUSEHOLD HEAD
HOUSEHOLD MEM
HOUSEHOLD MEMBER
HOUSEHOLDER
HOUSEHOLDMEMBER
HSEWIFE
HUBAND
HUSB
HUSBAMND
HUSBAND
HUSBAND OF
HUSBAND OF TH
HUSBAND OF THE
I FEMALE
I MALE
IN LAW
INFANT
KID ONE
KID THREE
KID TWO
LADAY OF HOUSE
LADY
LADY A
LADY AS OF TH
LADY HOUSE
LADY O HOUSE
LADY OF
LADY OF HH
LADY OF HOME
LADY OF HOUSE
LADY OF HOUSEHOLD
LADY OF HOUSEHOLD
LADY OF HS
LADY OF HSE
LADY OF THE

LADY OF THE H
LADY OF THE HO
LADY OF THE HOME
LADY OF THE HOU
LADY OF THE HOUS
LADY OF THE HOUSE
LADY OF THE HOUSEH
LADY OF THE HOUSEHOL
LADY OF THGE
LADY OIF THE
LADY OPF THE
LADY POF THE
LADYA
LADYOFTHEHOUSE
LDY OF THE
LITTLE BOY
LITTLE GIRL
LIVING HERE FEB 2001
LIVING HERE FEB OOL
LOCAL
LOH
LOOK ON FRONT
M CHILD
M HEAD OF H
M HEAD OF HOUSE
M OCCCUPANT
M OCCUPANT
MALE
MALE A
MALE ADULT
MALE AGE
MALE B
MALE CHIL
MALE CHILD
MALE DECLINED
MALE FRIEND
MALE H OF HH
MALE HEAD
MALE HEAD OF
MALE HEAD OF HOUS
MALE HEAD OF HOUSE
MALE HEAD OF HOUSEHO
MALE HEAD OR
MALE HH
MALE HHM
MALE HOH
MALE HOUSEHOLD MEM
MALE HOUSEHOLD MEMBE
MALE I
MALE II
MALE III
MALE IN
MALE NO
MALE NUM
MALE OF
MALE OF HOUSE
MALE OF HOUSEHOLD
MALE OF THE
MALE ONE
MALE PARENT
MALE PERSON
MALE REF
MALE REFUSED
MALE RENTER
MALE RESIDE
MALE RESIDENT
MALE RESP
MALE SON
MALE TEEN
MALE TWO
MALECHILD
MAN
MAN O
MAN OF
MAN OF HOIUSE
MAN OF HOUSE
MAN OF HOUSEHOLD
MAN OF HOUSEM
MAN OF HS
MAN OF THE
MAN OF THE HO
MAN OF THE HOME
MAN OF THE HOU
MAN OF THE HOUS
MAN OF THE HOUSE
MAN OF THE HOUSEHO
MAN OF THE HOUSEHOLD
MAN OF THE HS
MAN ONE
MAN OR THE
MANA OF THE
MARRIED
MASTER OF HOUSE
MATERNAL
ME I LIVE ALONE
ME SEE FRONT
ME-I LIVE ALONE
ME-SEE FRONT
MIDDLE
MIDDLE BOY
MIDDLE BROTHER
MIDDLE CHILD
MIDDLE DAUGHTER
MIDDLE GIRL
MIDDLE SON
MINOR CHILD
MISS
MOM
MOTHER
MOTHER IN
MOTHER IN LAW
MOTHER OF
MOTHER OF CHILDREN
MOTHER OF LN 01
MOTHER OLDER PERSON
MOTHERINLAW
MR
MR MALE
MR REFUSAL
MR REFUSED
MR RESIDENT
MR RESP
MRREFUSED
MRS
MRS REFUSAL
MRS REFUSED
MRS RESIDENT
MRS RESP
MRS.
MS
MS HOMEOWNER
MS NAME
MS RESPONDANT
MY CHILD
MY DAUGHTER
MY FATHER
MY MOTHER
MY SON
MY WIFE
MYSELF
NA
NAME FEMALE
NAME MALE
ND BOY
ND CHILD
ND DAUGHTER
ND FEMALE
ND GIRL
ND MALE
ND MAN OF
ND SON
ND SON OF
ND WOMAN
NEPHEW
NEW BABY
NICE LADY
NIECE
NO DAUGHTER
NO LAST NAME
NO MIDDLE NAME
NO NAME
NO NAME TWO
NO NAMES PLEASE
NO NEED
NO ONE
NO ONE ELSE
NON OF YOUR
NONAME
NONE
NONE OF YOUR BUSINES
NOT REQUIRED
NUMBER
NUMBER ONE
NUMBER TWO
OCCUPANT
OF HOUSE
OF HOUSE M
OF HOUSEHOLD
OF RESIDENT
OF THE HH
OF THE HOUS
OF THE HOUSE
OF THE HS
OLD DAUGHTER
OLDER BOY
OLDER CHILD
OLDER DAUGHT
OLDER DAUGHTER
OLDER FEMALE
OLDER GIRL
OLDER KID
OLDER MALE
OLDER SON
OLDEST
OLDEST BOY
OLDEST CHILD
OLDEST DAU OF
OLDEST DAUGHT
OLDEST DAUGHTER
OLDEST GIRL
OLDEST SON
ON FRONT PAGE
ONE
OTHER MALE
PARENT
PERSON
PERSON A
PERSON B
PERSON D
PERSON E
PERSON FOUR
PERSON I
PERSON OF THE HOUS
PERSON OF THE HOUSE
PERSON ONE
PERSON THREE
PERSON TWO
PERSONAL INFORMATION
R HOUSE
RD BOY
RD DAUGHTER
RD GIRL
RD SON
RDAUGHTER
REF
REF DAUGHTER
REF MRS
REFUSAL
REFUSD
REFUSE
REFUSED
REFUSED MALE
REFUSED NAME
REFUSED SON
REFUSEDNAME
RELATIVE
RELUCTANT
REPONDENT
RESIDENCE
RESIDENT
RESIDENT A
RESIDENT B
RESIDENT C
RESIDENT D
RESIDENT I
RESIDENT II
RESIDENT NO
RESIDENT NO ONE
RESIDENT NO THREE
RESIDENT NO TWO
RESIDENT NUMBER ON
RESIDENT NUMBER ONE
RESIDENT NUMBER TW
RESIDENT NUMBER TWO
RESIDENT OF
RESIDENT OF N
RESIDENT ONE
RESIDENT OWNER
RESIDENT SEE P
RESIDENT TWO
RESIDENT-OWNER
RESIDENTS
RESPONDANT
RESPONDANT ONE
RESPONDENT
REUSED
ROBIN IS MIDDLE
ROOMATE
ROOM MATE
ROOMMATE
SAME AS FRONT
SAME AS ON FRONT
SAME AS ON PG
SAME AS PAGE
SAME AS PG
SECOND
SECOND BOY
SECOND CHILD
SECOND DAUGHTER
SECOND FEMALE
SECOND GIRL
SECOND MALE
SECOND OLDEST
SECOND RESIDE
SECOND RESIDENT
SECOND SON
SEE FRONT
SEE FRONT PAGE
SEE PAGE ONE
SEE SHEET
SISTER
SISTER OF THE HOUS
SISTER OF THE HOUSE
SISTER IN LAW
SISTERINLAW
SON
SON A
SON B
SON C
SON CHILD
SON I LAW
SON IN LAW
SON INLAW
SON N
SON OF
SON OF D
SON OF D LADY
SON OF HOUSE
SON OF MRS
SON OF THE
SON OF THE HO
SON ONE
SON TWO
SON1
SONINLAW
SPOUSE
SPOUSE OF
ST CHILD
ST DAUGH OF
ST DAUGHTER
ST FEMALE
ST MALE
STDAUGHTER
STEP DAUGHTER
STEP FATHER
STEP MOTHER
STEP SON
STEPDAUGHTER
STEPSON
TEENAGER
TH GIRL
THE
THE GENTLMAN OF
THE HOUSE
THE HOUSEHHOLD
THE HOUSESITTER
THE HUSBAND
THE LADY OF HOUS
THE WIFE
THIRD
THIRD CHILD
THIRD FEMALE
THIRD MALE
THIRD OLDEST
THIRD SON
THREE
TODDLER
TOO PERSONAL
TWIN GIRL
TWO
UNBORN CHILD
UNCLE
UNCOMFORTABLE
UNKNOWN
UNKNOWNALSO
UNNAMED
WHITE MALE
WHITFEMALE
WIFE
WIFE OF
WIFE OF THE
WIFE OF THE H
WIFE REFUSE
WIFE TO
WILL NOT GIVE N
WOMAN
WOMAN FRIEND
WOMAN I
WOMAN O
WOMAN OF
WOMAN OF HOME
WOMAN OF HOUSE
WOMAN OF HOUSEHOLD
WOMAN OF THE
WOMAN OF THE H
WOMAN OF THE HO
WOMAN OF THE HOU
WOMAN OF THE HOUS
WOMAN OF THE HOUSE
WOMEN
WOMEN OF
WOMEN OF THE
WOMEN OF THE HOUSE
WON'T
YEAR OLD
YO BOY
YO GIRL
YONG DAUGHTER
YOU DONT NEED
YOUNG BOY
YOUNG CHILD
YOUNG DAUGHTER
YOUNG GIRL
YOUNG SON
YOUNG TODDLER BRO
YOUNGER
YOUNGER BOY
YOUNGER DAUGHTER
YOUNGER GIRL
YOUNGER KID
YOUNGER MALE
YOUNGER SON
YOUNGERFEMALE
YOUNGEST
YOUNGEST BOY
YOUNGEST DAU O
YOUNGEST DAUG
YOUNGEST GIRL
YR BOY
YR GIRL
YR OLD

Appendix C: Loglinear Model SAS Code and Output

For the **Association between Socioeconomic/Demographic Factors and Missingness in Unmatched Records** analysis, a saturated loglinear model was fit using the unmatched ACS 2009 records and following factors.

- Social Characteristic (SocialCh) – two categories: 1) a person who is a self-reported non-English speaker at home or a self-reported non U.S. citizen, or 2) all others
- Economic Characteristic (EconCh) – two categories: 1) a person whose self-reported income is below the poverty line or is a self-reported food stamp recipient, or 2) all others
- Demographic Characteristic (DemoCh) – two categories: 1) a person that is either non-white or Hispanic, or 2) all others
- Census Division (CenRegion) – nine state divisions as defined by the U.S. Census Bureau
- Missing DOB (DBD_ALL) – two categories: 1) a record with completely missing Date of Birth (DOB) information, 2) records with full or partial DOB
- Fake or Incomplete Name (FakeName) – two categories: 1) a record that has a fake/incomplete first or last name found in the NORC generated lists in **Appendix B**; this includes records with blank first or last names, or 2) all other records

Because of item nonresponse, some ACS records did not have information for the social and economic variables used to define SocialCh and EconCh. These records were removed, and the loglinear fit used 292,071 unmatched ACS 2009 records. Below is the SAS program used to fit the loglinear model followed by the SAS output for the maximum likelihood analysis of the main effects and interaction terms.

```
options nocenter;
libname santa "/home/prama001/";

/* Remove records with missing Economic Characteristic */
data tmp;
set santa.llmodel_Ed(where=(EconCh^='EconMiss'));

/* Remove records with missing Social Characteristic */
data tmp2;
set tmp(where=(SocialCh^='SocialMiss'));

tables CenRegion*SocialCh*EconCh*DemoCh*FakeName*DBD_ALL/ out=Combos noprint;
```

```
proc catmod data=Combos;
weight count;
model CenRegion*SocialCh*EconCh*DemoCh*FakeName*DBD_ALL=_response_/ noprofile
noresponse noiter noparm;
loglin CenRegion|SocialCh|EconCh|DemoCh|FakeName|DBD_ALL;
run;
```

The CATMOD Procedure

Data Summary

Response	Cen*Soc*Eco*Dem*Fak*DBD_	Response Levels	288
Weight Variable	COUNT	Populations	1
Data Set	COMBOS	Total Frequency	292071
Frequency Missing	0	Observations	288

Maximum Likelihood Analysis

Maximum likelihood computations converged.

Maximum Likelihood Analysis of Variance

Source	DF	Chi-Square	Pr > ChiSq
CenRegion	8	11214.36	<.0001
SocialCh	1	9646.87	<.0001
CenRegion*SocialCh	8	2208.63	<.0001
EconCh	1	4943.19	<.0001
CenRegion*EconCh	8	161.25	<.0001
SocialCh*EconCh	1	0.26	0.6098
CenRegio*SocialCh*EconCh	8	33.62	<.0001
DemoCh	1	1741.42	<.0001
CenRegion*DemoCh	8	2408.92	<.0001
SocialCh*DemoCh	1	10851.34	<.0001
CenRegio*SocialCh*DemoCh	8	443.42	<.0001
EconCh*DemoCh	1	997.78	<.0001
CenRegion*EconCh*DemoCh	8	39.33	<.0001
SocialCh*EconCh*DemoCh	1	84.28	<.0001
CenRe*Socia*EconC*DemoCh	8	73.52	<.0001
FakeName	1	14713.36	<.0001
CenRegion*FakeName	8	95.57	<.0001
SocialCh*FakeName	1	770.29	<.0001
CenRegi*SocialC*FakeName	8	36.64	<.0001
EconCh*FakeName	1	365.03	<.0001
CenRegio*EconCh*FakeName	8	57.92	<.0001
SocialCh*EconCh*FakeName	1	0.25	0.6158
CenRe*Socia*EconC*FakeNa	8	40.67	<.0001
DemoCh*FakeName	1	198.50	<.0001
CenRegio*DemoCh*FakeName	8	62.04	<.0001
SocialCh*DemoCh*FakeName	1	97.62	<.0001
CenRe*Socia*DemoC*FakeNa	8	21.47	0.0060
EconCh*DemoCh*FakeName	1	29.83	<.0001
CenRe*EconC*DemoC*FakeNa	8	19.35	0.0131
Socia*EconC*DemoC*FakeNa	1	0.10	0.7520

CenR*Soci*Econ*Demo*Fake	8	23.22	0.0031
DBD_ALL	1	4483.91	<.0001
CenRegion*DBD_ALL	8	336.50	<.0001
SocialCh*DBD_ALL	1	24.69	<.0001
CenRegi*SocialCh*DBD_ALL	8	58.08	<.0001
EconCh*DBD_ALL	1	15.38	<.0001
CenRegion*EconCh*DBD_ALL	8	15.01	0.0589
SocialCh*EconCh*DBD_ALL	1	0.12	0.7251
CenRe*Socia*EconC*DBD_AL	8	26.62	0.0008
DemoCh*DBD_ALL	1	138.22	<.0001
CenRegion*DemoCh*DBD_ALL	8	26.36	0.0009
SocialCh*DemoCh*DBD_ALL	1	43.15	<.0001
CenRe*Socia*DemoC*DBD_AL	8	34.74	<.0001
EconCh*DemoCh*DBD_ALL	1	15.19	<.0001
CenRe*EconC*DemoC*DBD_AL	8	43.87	<.0001
Socia*EconC*DemoC*DBD_AL	1	1.84	0.1745
CenR*Soci*Econ*Demo*DBD_	8	20.00	0.0103
FakeName*DBD_ALL	1	3337.20	<.0001
CenRegi*FakeName*DBD_ALL	8	118.90	<.0001
SocialC*FakeName*DBD_ALL	1	128.79	<.0001
CenRe*Socia*FakeN*DBD_AL	8	48.15	<.0001
EconCh*FakeName*DBD_ALL	1	8.17	0.0043
CenRe*EconC*FakeN*DBD_AL	8	8.69	0.3694
Socia*EconC*FakeN*DBD_AL	1	3.25	0.0714
CenR*Soci*Econ*Fake*DBD_	8	17.17	0.0284
DemoCh*FakeName*DBD_ALL	1	1.12	0.2896
CenRe*DemoC*FakeN*DBD_AL	8	28.91	0.0003
Socia*DemoC*FakeN*DBD_AL	1	7.16	0.0075
CenR*Soci*Demo*Fake*DBD_	8	7.41	0.4927
EconC*DemoC*FakeN*DBD_AL	1	1.33	0.2488
CenR*Econ*Demo*Fake*DBD_	8	22.37	0.0043
Soci*Econ*Demo*Fake*DBD_	1	1.01	0.3149
Cen*Soc*Eco*Dem*Fak*DBD_	8	14.83	0.0625
Likelihood Ratio	0	.	.

Appendix D: Glossary

Assignment of a PIK	The process of loading data from a reference file to an incoming file. This is a function of the PVS process where the PIK from the Numident reference file is assigned to the incoming record when an incoming recorded is validated (see **Validation** and **Verification**).
Census NUMIDENT File	A version of the NUMIDENT file that consolidates the Social Security transaction records into one record per SSN. Alternate name and date of birth data are retained in separate files. The Census NUMIDENT is recreated each year, to reflect Social Security transaction records through March of each year.
False Match	A record that is incorrectly matched. For the PVS process, an incorrect PIK has been assigned to an incoming record.
Failed-Match	A record that is not linked which should have been linked. For this process, no PIK has been assigned to an incoming record that should have received a PIK. Also referred to as a **False-Nonmatch**.
Geokey	The Geokey describes the address variable on the incoming and reference files used for linkage.
GeoSearch Reference File	A version of the NUMIDENT file that contains address data linked to SSN records from the Census Numident file. The address data is extracted from various administrative source files. Records are created for each SSN listing all possible combinations of address, Census NUMIDENT name and date of birth data, alternate name and date of birth data. GeoSearch Reference Files are created for various time frames to reflect the needs of the survey/source files requiring validation. The GeoSearch module tries to match incoming records to this file.
Incoming File or Record	A file of survey respondents or administrative records requiring SSN validation.
Link	See **Match.**
Match	The output from the record linkage software system. Two records are considered linked/matched when the scores computed by the software exceeds the thresholds set by the user. When a match is made in PVS, a PIK is assigned to the incoming record.
Match Rate	The percentage of incoming records assigned a PIK.

NameSearch Reference File	A version of the NUMIDENT file which contains records for each SSN listing all possible combinations of Census NUMIDENT name and date of birth data, and alternate name and date of birth data. Name Reference Files are recreated for each new version of the Census NUMIDENT. The NameSearch module tries to match incoming records to this file.
PIK	Protected Identity Key
PVS	The Personal Identification Validation System
Reference File	A file containing the data used for assignment to a source file. The PVS system uses 3 types of reference files, the Census NUMIDENT, the Geokey Reference File, and the Name Reference File.
Validation	The complete PVS process applied to an incoming file. The process may contain any or all of the PVS phases, Verification, GeoSearch, and NameSearch. A full PVS would contain a verification phase, a GeoSearch phase, and a NameSearch phase in this order. Incoming files with no SSN would proceed directly to the search phases. The final SSN linked to the incoming record is considered the validated SSN.
Verification	The PVS process that verifies the accuracy of an SSN on an incoming record, using the Census NUMIDENT File as the reference file. The alternate names and dates of births are also available for this process. While verification can be accomplished using less fields, depending on data availability, full verification requires the following data fields: SSN, Name (First, Last and Middle initial), Date of Birth (DOB), Gender