CALTECH/MIT
VOTING TECHNOLOGY PROJECT

A multi-disciplinary, collaborative project of
the California Institute of Technology – Pasadena, California 91125 and
the Massachusetts Institute of Technology – Cambridge, Massachusetts 02139

TITLE Interstate Voter Registration Database Matching: The Oregon-Washington 2008 Pilot Project

Name R. Michael Alvarez
University Caltech

Name Jeff Jonas
Organization: IBM Entity Analytics

Name William E. Winkler
Organization Bureau of the Census

Name Rebecca N. Wright
University Rutgers University

Key words:

VTP WORKING PAPER #84
August 10, 2009

Interstate Voter Registration Database Matching: The Oregon-Washington 2008 Pilot Project*

R. Michael Alvarez
Professor of Political Science
California Institute of Technology

Jeff Jonas
IBM Distinguished Engineer
Chief Scientist, IBM Entity Analytics

William E. Winkler
Principal Researcher
Bureau of the Census

Rebecca N. Wright
Professor of Computer Science
Rutgers University

Abstract

Voter registration databases maintain lists of registered voters that are used to determine who is and is not eligible to vote in an election. As such, accurate voter registration databases form a cornerstone of the electoral process. In the United States, each state maintains its own voter registration database. It is not uncommon for a voter to become registered in two states, for example as a result of moving from one state to the other or of living in one state and working in one another.

In this paper, we report on a pilot interstate voter registration database matching project between the two states of Oregon and Washington whose goal was to explore the feasibility of using database matching to identify voters registered in the two states, and to do so with as much openness and transparency as possible. We describe the matching algorithms used, the procedures taken with found matches, and the resulting actions taken on actual voter registrations. We also discuss several directions for improving matching algorithms and procedures.

1 Introduction

The Help America Vote Act (HAVA), passed into law in the United States in 2002, required that states implement a sweeping set of changes to election administration and voting technology. Among the more important of these changes was a requirement that states develop statewide voter registration files. Specifically, Section 303 of HAVA [1] mandated that each state develop "a single, uniform, official, centralized, interactive computerized statewide voter registration list defined, maintained, and administered at the State level that contains the name and registration information of every legally registered voter in the State..."

As of August 2008, information provided by the Pew Center on the States indicates that most states have implemented their statewide database [2].[1] But as states have moved to develop and implement their statewide voter registration databases as required by HAVA, scholars and policymakers have begun to explore how statewide voter databases can be shared across different states, in an effort to improve the accuracy of each state's database [7, 11]. Interstate voter registration database exchange and matching may be of particular interest to election officials in states that have metropolitan areas that span state borders and in situations where there is a great deal of residential mobility between two or more states. As few eligible citizens think about canceling their registration status at their former place of residence when they move, interstate matching in places with high rates of residential mobility across state borders might help identify potential duplicate registration records, and when dealt with appropriately, should result in more accurate voter registration databases in participating states.

While the issue of interstate database exchange has received some attention in the election administration community, we are aware of only one large-scale database exchange project, the "Midwest States Voter Record Exchange" [4].[2] That project, initiated in late 2005, originally involved Iowa, Missouri, Nebraska and Kansas; it has more recently been expanded to include Min-

*We thank Katie Blinn, Dave Franks, Nick Handy, Shane Hamlin, Ericka Haas, Tim Likness, John Lindback, Paul Miller, David Motz, and Randy Newton for their support in carrying out this project. We also thank Herb Lin and our colleagues on the National Academies of Science Committee on State Voter Registration Databases for their comments.

[1] Notable states that have not implemented their statewide databases are California and New York.

[2] There are other interstate matching efforts, for example, involving some exchanges in the southeast and mid-central Atlantic states. The Midwest project is the largest and most established project that we are currently aware of.

nesota and South Dakota. This effort has involved annual database exchanges. An exact character-by-character matching algorithm is applied between the databases using each registrant's first name, middle name, last name and recorded date-of-birth. If a match is found between two states for a particular registrant, that information is passed back to administrators in the two states for resolution using each state's existing procedures. As this point, only limited and highly aggregated information has been provided to researchers and the public about the Midwest effort, and so far no detailed project analysis has been released.

In this paper, we provide a detailed analysis of an interstate voter registration database matching pilot project in Oregon and Washington. This project was initiated in August 2008 with the cooperation of the Secretary of State offices in both states, and involved the matching of databases of active voters from the two states: the database of just over 3.4 million Washington active voters was matched to the database of the just over 2 million Oregon active voters. A subset of counties was selected to move forward in a pilot project with voter contact aimed at attempted resolution to determine whether found matches were in fact duplicate records. In the remainder of the paper, we describe two different matching algorithms that were employed and their outputs. We describe the overall end-to-end process including notification, response rates, and actions taken as a result. We also discuss the implications of these results for other interstate data exchange programs in the future and consider improvements to matching through the use of more sophisticated techniques.

2 Oregon-Washington 2008 Pilot Project

Oregon and Washington make an excellent case for a pilot project such as this one. These two states have similar progressive political cultures and there is a history of collaboration between the election officials in these states. Oregon and Washington are also interesting for an interstate matching project as their statewide voter registration systems were not developed by the same vendor, suggesting that it is not necessary or important that states share similar voter registration systems in order to carry out efficient interstate data matching.[3] Also, they both have mobile populations; data from the 2007 American Community Survey show that only about half of each state's population was born in that state (53.6% in Washington, ranking 40th of all states in percentage of the current native-born state residents who were born there, and 50.3% in Oregon, ranking 41st. See Table R0601 of [3].) Other data from the 2007 American Community Survey paint a similar portrait; in terms of the percentage of people aged one year and over who lived in a different state one year ago, Oregon ranks 21st in the United States (3.4%) while Washington ranks 24th (3.2%) (See Table R0703 of [3]). Finally, the Portland-Vancouver-Beaverton metropolitan statistical area, which straddles the Oregon-Washington border, is one of the largest in the nation, ranking 23rd, with an estimated population of 2,137,565 in 2006.

The primary objective of the Oregon-Washington pilot project was to compare their state voter registration databases, to identify potential duplicate records between the files, and then to develop procedures for resolving those potential duplicates. By developing and carrying out processes and procedures for matching between Oregon and Washington, the hope of pilot project participants was that each state would in the end have a more accurate voter registration database; more accurate databases in both states should lead to more efficient election administration. Because such an interstate matching project had never been performed between these two states, the precise procedures for both stages (identification of potential duplicates and procedures for resolution of potential duplicates) were developed as part of the pilot project. The procedures developed by this pilot project may therefore be of interest to other state election officials considering interstate matching efforts.

The Oregon-Washington project began in early August 2008 with a series of telephone conference calls among the project team (which consists of the authors of this report as well as representatives from Secretary of State offices of both Oregon and Washington as well as some local county officials). On August 13, 2008 the Oregon Secretary of State's Office (OSOS) received the complete voter registration database from the State of Washington. At the time this pilot project was undertaken, the Oregon voter database had 2,053,444 records (approximately 280MB of data), and the Washington voter database had 3,407,596 records (approximately 465MB of data), for a total of 5,461,040 records. Notably, these files were consistent in structure and content with the public voter registration data available as a matter of public record—no additional information was contained in these files. For this reason, any private or public entity could have engaged in the matching aspect of this project.

The OSOS staff proceeded to carry out matching on the Washington and Oregon databases. The two databases proved to be identical in formatting and naming conventions, except for a minor discrepancy between the format of the date-of-birth field.[4] Following the methodol-

[3] Washington's system was based on a Microsoft platform; Oregon uses a system from Saber Government Solutions.

[4] Washington uses a format mm/dd/yyyy for their date-of-birth field, while Oregon uses mm-dd-yyyy; OSOS staff converted the Washington

ogy that has been used by the Midwest interstate matching project, the first matching algorithm used by OSOS matches on the complete first name, complete middle name, complete last name and date-of-birth. Additionally, OSOS used a second matching algorithm that matches on the complete first name, the first character from the middle name field, complete last name, and the date-of-birth.

The matching was done using FileMaker Pro, by OSOS staff, using standard desktop machines. After the date-of-birth fields were made comparable, the OSOS staff merged the two databases together (creating one large data file with approximately 5.4 million records). Within that file, they created two new fields, one that concatenated full first name, full middle name, full last name, and date-of-birth; the second concatenated the full first name, the first character of the middle name field, full last name, and date-of-birth. These two new concatenations "match keys" that were then used to identify matches between the two state files.[5] Merging the two files took approximately 90 minutes on an iMac; the analysis identifying matches based on the second concatenation took approximately 50 minutes.

We define the *match rate* as the number of matched records found divided by the total number of records from both states, expressed as a percentage. This excludes in-state matches, which were not investigated in the pilot project. Using the first matching algorithm, when the full middle name was used, 3,482 matched records between the two states were found, for a match rate of 0.064%.[6] However, when only the middle initial was used in the matching algorithm, there were 8,292 matching records between the two states, a match rate of 0.152%. This larger set of matches includes all of the original 3,482 matches; the use of only the middle initial added 4,810 cases to the set of identified matches. Notably, this subtle matching difference (middle name versus middle initial) more than doubled the number of match candidates found—specifically, it resulted in 2.375 times more match candidates. Human inspection of both matching methods produced a great degree of confidence in that these methods were yielding potential duplicates.

format to the Oregon format.

[5]The matching was not case sensitive. Although no preprocessing was done to remove hyphens, more advanced matching methods routinely are able to deal effectively with hyphenated names. The details of the more advanced matching methods are beyond the scope of the current paper.

[6]This number means that there were 3,482 records in the Oregon file that had exact matches with 3,482 records in the Washington file; the total number of records across both states that were matched is thus twice this number.

3 Analysis of Interstate Matches

Once the matching algorithms yielded lists of potential duplicates, the next step was to determine how to proceed. The election officials involved in the pilot unanimously decided to use the larger set of matched records as the starting point for this pilot project. This involved the 8,292 duplicate records that arose from the matching algorithm using the middle initial, for the second component of the pilot: selection of a pilot subset, voter notification and changes to the voter registration database. Before describing how potential duplicates were handled by the election officials, we first summarize the results of the second matching algorithm.

Our first analysis, presented in Table 1, identifies the counties in each state that had the most matched records. There we give the fifteen counties with the most matched records (rank-ordered by number of matched records); we also give in the second column of Table 1 data on the number of registered voters in each of these same counties to provide a better sense of the extent to which population alone might make counties particularly susceptible to being among the top set of counties with matches.

Not surprisingly, we see that the more populous Washington and Oregon counties had the greatest numbers of matched records. As Table 1 shows, in Washington the two counties with the most matched records were King and Clark, followed by Pierce, Snohomish, and Spokane counties. A similar pattern exists in Oregon: the counties there with the most matches are also among the more populous Oregon Counties, including Multnomah, Washington, Clackamas, Lane and Marion counties. This analysis makes clear that the number of matches that arise due to intrastate matching efforts in any particular county is to some extent simply related to the county's population.

In Table 2, we give the county combinations that we observe in the dataset of matched records, ordered by their frequency of occurrence. In other words, this table presents the county combinations across the two states that generate the highest number of matches. We see that the populous counties again are likely to be at the top of this list: of the 8,292 matches, 991 matches come from Multnomah County in Oregon and King County in Washington; 790 come from a Multnomah-Clark County match; and 398 are due to matches between Washington State's King County and Oregon's Washington County.

Furthermore, many of the counties that rank high in numbers in the matched dataset are also counties at or near the Oregon-Washington border, which also suggests that some of these individuals in the matched dataset might reside in geographic areas where there is substantial cross-state mobility. This is particularly true for a number of the county pairs that we see in Table 2: Multnomah-Clark and Clackamas-Clark are good exam-

Washington County	Matches	Total Voter Registrations	Matching %
King	2774	1108128	0.25
Clark	1765	216508	0.82
Pierce	534	411103	0.13
Snohomish	348	372636	0.09
Spokane	334	258952	0.13
Thurston	329	148527	0.22
Cowlitz	275	55331	0.50
Walla Walla	209	31625	0.66
Kitsap	197	144690	0.14
Klickitat	155	12171	1.27
Yakima	147	97856	0.15
Whatcom	132	115314	0.11
Benton	107	87059	0.12
Pacific	88	13052	0.67
Chelan	82	38650	0.21
State Total	**7476**	**3629851**	**0.21**
Oregon County	**Matches**	**Total Voter Registrations**	**Matching %**
Multnomah	2717	422336	0.64
Washington	1058	266523	0.40
Clackamas	876	220448	0.40
Lane	537	204976	0.26
Marion	380	147849	0.26
Deschutes	374	91681	0.41
Jackson	307	119231	0.26
Umatilla	228	31762	0.72
Benton	174	49895	0.35
Yamhill	140	50048	0.28
Clatsop	133	21503	0.62
Linn	123	61954	0.20
Columbia	119	28521	0.42
Douglas	118	64526	0.18
Polk	117	41479	0.28
State Total	**7401**	**2113668**	**0.35**

Table 1: Counties with Highest Numbers of Matches

Oregon County	Washington County	Frequency
Multnomah	King	991
Multnomah	Clark	790
Washington	King	398
Clackamas	Clark	302
Washington	Clark	244
Clackamas	King	235
Lane	King	234
Deschutes	King	147
Multnomah	Pierce	145
Marion	King	135
Jackson	King	123
Multnomah	Thurston	111
Multnomah	Spokane	106
Umatilla	Walla Walla	91
Multnomah	Cowlitz	82

Table 2: Top Fifteen Combinations of County Matches

ples of this phenomenon. The results in Tables 1 and 2 also demonstrate that, at least between Oregon and Washington, cross-state matches occur more frequently from larger population counties, especially those such as Multnomah that are on the border with another state.

Another question of interest involves the date of registration for the potential matches that were identified between the two state databases. Each state database has the date of most recent registration in that state, and we were able to compare those registration dates in a number of ways. First, we analyzed the difference between the registration dates for each of the 8292 matched records, and found that the median absolute difference in registration dates was 1538 days, or approximately 4.2 years. That implies that many of the matched cases involve records where the difference in registration dates between the states is four years or more. When we examined records where the voter registration date was most recent in Washington (2739 cases), we found that there was a very wide range of Oregon registration dates: in this analysis one record had an Oregon registration date of 1941, though most of the Oregon registration dates were much more recent, including 603 in 2004, 329 in 2006, and 95 in 2008. The same analysis of the 5553 cases where the Oregon registration date was most recent revealed two cases with a Washington registration date of 1944, though again most were much more recent, with 877 in 2004, 518 in 2006, and 148 in 2008.

Of particular interest are the cases where the matched records involve registration dates in both states in 2008, which in our analysis of the 8292 potential matches was a total of 243 cases. Most of these cases where we see

matched records with a 2008 registration date in both states involve records from border counties; for example, 84 of these matched records with 2008 registration dates in both states involve an Oregon address in Multnomah County, while 77 involve a Washington address in Clark County. Visual examination of the two registration dates among all of these 243 these records often turned up registration dates that were relatively close in time: for example, in one case we see an Oregon registration date of April 3, 2008 and a Washington registration date of April 7, 2008, another case had an Oregon registration date of June 19, 2008 and a Washington registration date of June 24, 2008.

4 Resolution of Matches

With the matched files in hand, the election officials from both states involved in this pilot project made an early decision to mitigate potential risk via a partial project rollout—limiting the scope of the study to a subset of the population. Voter registration list processing for duplicates can come with much public scrutiny. Just a few missteps involving even a very small percentage of voters can be raise significant concern, be easily mischaracterized, and even bring an unfortunate and untimely end to such an effort. An additional issue was that the 2008 general election loomed just a few months away and with this a potential for extraordinary scrutiny and concern from media, privacy community, oversight entities and other third parties. It is easy to imagine the initial rollout of a project suffering missteps such as bugs in the implementation of the matching process, or notification letters being mailed with a typographical error in the contact phone number or inadvertently missing pre-paid postage envelopes. Making such a mistake on a statewide scale was deemed worth avoiding. With these risks in mind, it was mutually agreed that a partial rollout would allow the two participating states to ensure all of the project elements worked as planned before proceeding further. With this conservative approach in mind, the joint team decided to perform the end-to-end process (matching, notification, and changes to the voter rolls) on only a subset of the total available statewide data. As a result, the pilot project was able to remedy a process error in the mailing, determine estimated response rates, and deal with media messaging issues that had began to emerge. Using this knowledge, the potential costs and consequences of a full statewide effort can be better estimated in terms of resources and outcomes while at the same time considerably reducing execution risk.

The scope of the pilot project with respect to deeper analysis, notifications and procedural remedies was limited to Clackamas, Multnomah and Washington Counties in Oregon, and Clark County in Washington. These counties are border regions between the two states, in areas where it was believe that there was a relatively high degree of cross-border mobility. Once the scope of the project was limited to these counties, a list of 1,312 matched individuals limited to these counties was produced and a procedure was developed for contacting individuals on the matched list.[7] Each state developed a contact letter. The Oregon contact letter was sent to the 686 individuals from the matched list whose most recent registration date was in Washington; the Washington contact letter was sent to the 626 individuals from the matched list whose most recent registration date was in Oregon. In each letter, the individual was asked if he or she wished to cancel his or her registration in the other state (i.e., the state with the less recent registration). Examples of these letters are provided in the Appendix. When, and only when, an individual returned the form indicating that he or she wished to be removed from the registration list in the particular state, that information was recorded in the respective state election office, and that information was then forwarded to the appropriate county election official for resolution according to their normal handling procedures for removal requests (including checking the signatures). Some of the notifications were returned undeliverable as addressed, and that information was also recorded by the state election office handling the particular matched individual. It is important to stress that no removal action was taken in cases were the signed request to be removed was not returned; no individuals were removed from either state's registration list unless the signed and completed request for removal was received.[8]

Data on resolution is provided in Table 3, which gives the total number of mailings in each state, the number that appear to have been delivered, the number that yielded responses and the response rate of those delivered, the number of cancelled voters, and the number of unresolved responses as of April 16, 2009. Oregon mailed 686 individuals from the matched list; 650 of those mailings appear to have been delivered (95%). Of those Oregon letters that appear to have been delivered, 391 generated a response from the individual, a response rate of 60% of delivered mailings. Of the 391 responses, 379 responses were forwarded to the appropriate county election official

[7] The matching analysis identified 1,336 matches in these counties, as reported in Table 2 above. Before proceeding with contacting these individuals, election officials updated the mailing list by removing names of people who they confirmed had notified them of their move on their new voter registration card, which is why the number of individuals in the mailing was slightly lower than the number originally matched.

[8] Thus, no individuals were removed from the registration list without confirmatory information. This is an issue that we will discuss in more detail in subsequent research. See [17] for a discussion of this issue.

	Oregon	Washington
Total mailed	686	626
Delivered	650	599
Response received	391	362
Response rate of delivered	60%	60%
Cancellations	379**	352
Unresolved responses	12	8**

**Two cancellation requests received by Washington were forwarded to Oregon for processing and are included in the Oregon cancellation number.

Table 3: Mailing Resolution (Data as of April 16, 2009.)

in Clackamas, Multnomah or Washington Counties and resulted in cancelled voter registration records. Twelve of the responses were unresolved as the registered voter did not provide enough information for the removal process to proceed.

The response data from Washington are quite similar. Washington sent 626 letters to individuals, and of those 599 appear to have been delivered (96%). Of those delivered, 362 generated a response from the individual, a 60% response rate of those delivered. 352 of the responses resulted in a cancellation of the individual's registration record in Clark County, Washington. Eight responses did not have enough information for resolution, and two of the responses received by Washington were forwarded to Oregon for processing (and are included in the Oregon cancellation number).

An important issue that arises when interstate voter registration matching is concerned is the possibility that the individuals who have active voter registrations in each state might be obtaining and casting ballots in each state, potentially engaging in double voting. During the course of the pilot project, in the midst of the mailing effort (mid-October 2008), the election officials in each state decided to examine the voter registration dates and last voted dates of those individuals in the matched subsample used for the pilot project mailing effort. This study produced 67 voter records that merited further examination by the respective county election officials. Of these 67 voter records, after further investigation the respective county election officials determined that 12 of these individuals might have voted in both Washington and Oregon in prior elections. However, the potential double voting had occurred sufficiently far in the past that in was not possible to determine whether or not double voting actually occurred because signature envelopes and other documentation were no longer available for further review. Of the twelve under examination, six returned a form requesting that their voter registration record be cancelled in one of the two states. Of the remaining six, one returned a ballot to Oregon for the November 4, 2008 general election but did not submit a ballot in Washington (even though this individual's most current registration date is in Washington). At this point, there is no further investigation going on regarding historical instances of potential double voting; however, election officials in both states plan to revisit the potential double voting issue in the near future, as they plan to run this analysis again to determine if there were any potential instances of double voting in the November 4, 2008 general election, in which case documentation may exist to allow for closer investigation.

5 Discussion

During this pilot a basic matching algorithm was used: an exact character-by-character match on name and date-of-birth, with the only exception being that middle names were truncated to only first initial. This truncation enabled a middle name of *Edward* to match with a middle initial of *E*. The exact matching approach favors the false negative (missing real duplicates due to very minor discrepancies in the data—for example, one has a middle initial and the other record has no middle initial).

However, there are often inconsistencies between records in part because voter registration data is created to a great extent via manual data entry. Manual data entry errors in voter registration have been documented to be as high as 20% in some cases. Specifically, the Brennan Center found that data entry error was nearly 20% when a 15,000-record audit was performed in New York City in 2004 [16]. However, it is not clear that such data error entry rates are necessarily that great in other states, in particular states like Oregon and Washington that are in frequent contact with registered voters, especially in situations where mail is used frequently to provide voters with election materials. Data from the 2004 general election in Oregon, where ballots are delivered by mail to voters, show undeliverable rates ranging from approximately 2 to 8% [13].

Furthermore, individuals who wish to engage in fraud may try to use a degree of variation/inconsistency in their registration data (using the middle name *Bob* instead of *Robert*, wrong middle initial, a transposition error in their date-of-birth or SSN+4 [9], etc.). As a result, the kinds of matching algorithms we used are better suited to finding accidental duplicate registrations rather than intentional fraud. (We note that we are not aware of research indicating that such intentional fraud on statewide voter registration lists has actually been attempted, even though

[9] SSN+4 refers to the last four digits of a Social Security Number. Use of SSN+4 in voter registration is a common practice and often prescribed by law to prevent unintended disclosure of full Social Security Numbers.

potential vulnerabilities like these have been discussed in recent studies such as [6].)

Should this project proceed into future phases, especially should other states consider engaging in interstate matching, use of more sophisticated matching algorithms (e.g. [19, 8, 15]) will increase the detection of proper matches and, if done with some caution, at the same time limit the number of new false positives (found matches that do not in fact represent a duplicate voter). We suggest experimenting with new matching approaches starting with different methods of matching character data, as well as potentially utilizing tertiary data to improve matching results. Because false positives are always a concern, we recommend that a notification process similar to ours be followed. Specifically, no changes should be made to the voter rolls unless contact with the voter was made to confirm the match. Further, if "fuzzier" matching techniques are used, it is advised that one apply at minimum a similar notification process and additionally consider closer human inspection of the most fuzzy matches (e.g., comparison of the signature cards) before even attempting to contact the voter. In future phases, we also hope to obtain some quantitative measures of the false positive rate, perhaps by following up by telephone calls or other communication with a sampling of the potential duplicates.

Some other potential matching improvements to consider include the following:

- **Name Roots.** Name root libraries enable the matching process to recognize that names within name families such as *Bob/Bobby/Robert/Rob* and *Liz/Elizabeth/Beth*, etc. are comparable.
- **Name transliteration.** Name transliteration libraries enable matching algorithms to recognize that pairs such as *Mohammed* and *Mohamed* are comparable.
- **Name classifiers and name order evaluations.** One common data error involves name transposition such as first and last names transposed or first and middle transposed. Errors of this nature can be more common among names from certain cultural groups—for example, Asian first and last names are transposed more frequently in American usage than names from some other cultures. Name classifiers can help recognize, for example, that *Smith* and *Tan* are more often last names than first names. Name order evaluations are used to analyze multi-part names for misplacement, whether the names are traditional Western three-part names, or the many-part names that are commonly seen in non-Western names.
- **Name closeness testing.** Typographical errors in names (e.g., *Donna* versus *Dona*, *Mark* versus *Marek*) can be detected via algorithms designed to evaluate name closeness, such as the classic Soundex technique [18], the more advanced Jaro-Winkler method [14, 21], or even simply edit distance (i.e., the minimum number of substitutions, deletions, and substitutions to get from one string to another), which has been shown to be a reasonable alternative to the Jaro-Winkler method [10, 9].
- **Date-of-birth closeness testing.** One of the more common data errors in dates of birth occurs when the month and day are transposed (e.g., *12/06/64* versus *06/12/64*). Special field-level evaluations designed for date-of-birth challenges help detect a variety of "closeness" issues including transposition and single digit error.
- **Use additional fields such as SSN+4, driver's license fields, phone number, gender, and address.** If and when such information can be shared between states, such additional attributes can potentially be helpful to the matching process. This might not be possible in some states as state law may prohibit sharing of certain information with another state. Even if data is able to be shared, this also requires that records being matched have the same values (e.g., under current policy, voters may use either a driver's license, SSN+4, or in some states full SSN, or if the voter has none of these, the registrar will assign an internally generated unique number). The use of tertiary sources such as the Department of Motor Vehicles (DMV) and Social Security Administration databases can further improve matching accuracy. For example, an error in a date-of-birth can be detected in the voter registration record if it can be matched to DMV on name, address, and driver's license number. Such detection can both improve automated match quality and assist election office staff in manual review and disposition.
- **Use publicly available data (e.g., phone books) and public records (e.g., property ownership).** Using these sources can improve both automated and manual review-based matching processes. For example, *Mark Smith* born *6/21/74* residing at *312 Palmyra Street* might look to be a match with *Mark K. Smith* born *6/21/74* residing at *312 Palmyra Street*, except that one has a middle initial and one does not. However, if the phone book has an entry for *Mark Thomas Smith* residing at *312 Palmyra Street*, this could be interpreted to provide evidence that in fact these are two different entities as this indicates that *Mark Smith* does not have a middle name starting with "*K*". While use of external data sources can reduce the false negatives and reduce the

false positives and should be the subject of further research, such matching techniques are beyond the scope of this paper. More about the value and challenges of using third party data can be found in [11].

- **Automated signature analysis.** To date signature analysis comparing voter registration cards is generally a manual, human process. One method already being used by some banks to compare signatures on financial instruments (e.g., checks) involves automated signature analysis (e.g., [5]). While efficacy of these emerging technologies may or may not meet the expectations necessary for voter registration processes, it is conceivable that eventually such technology will be helpful to the voter registration matching process. Of course, matching on the basis of digitized signatures requires that both state databases have a digitized signature; in some implementations, such as "bottom-up" implementations of statewide voter registration databases, a digitized signature might not be available in the statewide file.

6 Results of Enhanced Matching

As a way to study some of these potential matching improvements, we examined some methods that go beyond an exact character-by-character match on the entire Oregon and Washington data files, which we report on here. We note that these exploratory matching results have not yet been followed up with further review (such as systematic human review and/or any voter contact).

In these experiments, we looked at methods that are more resilient to minor typographical errors. Our strategy uses three enhancements in the literature [12, 20]). First, we bring together pairs of records using a subset of the information (referred to as *blocking*). Second, we allow very minor typographical errors in first name and last name, as well as very slight deviations in the year-of-birth. Third, we apply a relative weighting strategy that assigns slightly different distinguishing power to first name, middle initial, last name, and the components of date-of-birth. Typically, first name and last name have more distinguishing power than other fields.

We performed three matching passes of the two files. The first pass brings together pairs of records that agree exactly on date-of-birth and first character of the surname and computes a matching score based on first name, middle initial, and last name. We use a string comparator function that allows us to account for very minor typographical error (*Smith* versus *Smoth*). In the second pass, we bring together pairs using month-of-birth, day-of-birth, and first three characters of last name and compute a matching score based on first name, middle initial, last name, and year-of-birth. We again use the string comparator for first and last name. We also use a function that allows slight deviation in the year-of-birth. In the third pass, we bring together pairs using the first three characters of first name and the first three characters of last name and compute a score using the remainder of first name, the remainder of last name, middle initial, day-of-birth, month-of-birth, and year-of-birth. We again allow a slight deviation in year-of-birth. Next, we slightly modify the relative score weights of different fields and choose a cutoff score above which we keep pairs that will be flagged for further review. Pairs with a score below the cutoff are considered non-matches. The intent of the relative weights and cutoff score is to keep pairs that look very much like those obtained during exact character-by-character matching but might have minor typographical error or missing data, without keeping too many pairs that are not reasonable duplicate candidates. In this case, we used manual review and expert "eye-balling" to choose weights and a cutoff score, but one could investigate the use of more analytic methods.

Pass	Total Pairs Identified	Pairs Above Cutoff
1	18,367,051	10,175
2	45,899,938	1,713
3	129,418,435	35,391

Table 4: Pairs Found in Three Passes

Table 4 shows the number of pairs and the number of pairs above cutoffs found on each pass. Of approximately 194 million pairs, approximately 47,300 pairs are above the cutoffs and suitable for followup. The latter number includes the approximately 8,300 pairs that agree exactly on all the fields. We note that it is not possible to compare the effectiveness of different algorithms without further knowledge about false positives (how many flagged pairs are in fact not duplicates) and false negatives (how many duplicate records are not flagged), as well as the cost and risk of each kind of error. Procedures such as never removing a voter without the voter's explicit request to do so can reduce the risk of false positives, but other factors such as cost of mailings and risk of voter confusion must also be considered.

7 Conclusion

The Oregon-Washington interstate voter registration database matching project was a fruitful exercise. First, it gave election officials in both states hands-on experience with voter registration database matching with a neighboring state. Second, it gave county election of-

ficials in the four counties included in the pilot project an opportunity to clean portions of their voter registration lists immediately prior to the November 4, 2008 general election. Third, the pilot project allowed for investigation of potential double voting in the past (and found only a very small handful of instances of potential double voting, none of which has been confirmed to be actual double voting at this time), as well as setting the stage for further investigation of double voting in the just-conducted general election. Fourth, because the election officials in both states wished for this pilot project to be conducted in an open and transparent fashion, they allowed the authors—as independent and neutral evaluators—access to the communications between election officials in each state as well as access to data generated by the pilot project.[10] By starting with a relatively small pilot project, much was learned about the two state's voter registration databases, and also about the potential gains from undertaking a more ambitious project. Finally, it was demonstrated that an interstate matching project can be carried out with relative ease despite the fact the voter registration systems were developed by different vendors.

Based on our experiences in this project, we provide several recommendations, some of which are specific to future Oregon-Washington matching efforts, and some of this apply to any attempt at interstate matching. First, regarding the specific of the Oregon-Washington interstate data matching efforts, both states should work to contact and resolve all matches between the two states before the next federal election in 2010. Second, election officials in both states should develop a set of procedures to deal with undeliverable mailings; at the time of this writing, those in the matched set (multiple interstate registrations) whose mail notifications were returned by the postal service have not been followed up. Such procedures might include, for example, another mailing to them, checking their address information against the USPS National Change Of Address (NCOA) database, examination of their voter registration records to determine if there are errors in data entry, or other procedural actions. Third, both states should develop an appropriate mechanism and related procedures to detect and investigate possible double voting, for example by including voting history in the matching process. Fourth, this project should be extended to involve other neighboring states, perhaps Idaho and California, in future matching pilot projects (perhaps focusing initially on border counties, counties that are known to have high intercounty mobility rates, or counties that have large number of matched cases). Fifth, states that in the future engage in matching projects should consider the use of matching methods more advanced than exact matching, including some of the advanced methods we discussed in this paper. Finally, it is imperative that future projects use the same high degree of transparency and open process that characterized this pilot project so that the research community and the public can understand exactly what the interstate voter registration matching process involves, how matches are resolved, and related policy and technical issues.

[10] This openness and transparency was firstly generated by the commitment provided by the election officials in both states to allowing us to participate in their periodic telephone conference calls about the project, by providing us with access to all of the relevant data and other materials they generated during this pilot project, and through their commitment to answer all of our questions about the project. We believe that this was a very successful model for how independent researchers and election officials can collaborate on pilot projects like these.

References

[1] Help America Vote Act of 2002. Sec. 303(a)(1)(A), Public Law 107-252. The complete text of HAVA is available at http://www.eac.gov/election/docs/help-america-vote-act-of-2002.pdf/attachment_download/file.

[2] State-by-state voter registration database status. The Pew Center on the States. Available at http://www.pewcenteronthestates.org/template_page.aspx?id=42364.

[3] American FactFinder. U.S. Census Bureau, American Community Survey, 2007. Ranking tables available at http://www.factfinder.census.gov/servlet/GRTSelectServlet?ds_name=ACS_2007_1YR_G00_.

[4] Midwest voter registration data-sharing project moves forward: Advocates voice concern, December 2007. Available at http://www.mapj.org/?q=node/118.

[5] SQN implements parascript's signature verification technology to prevent check fraud in Brazil, October 2008. Available at http://www.reuters.com/article/pressRelease/idUS85905+20-Oct-2008+PRN20081020.

[6] R. Michael Alvarez. Potential threats to statewide voter registration systems, October 2005. Available at http://vote.nist.gov/threats/papers/statewide_registration.pdf.

[7] R. Michael Alvarez and Thad E. Hall. The next big election challenge: Developing electronic data transaction standards for election administration. IBM Center for the Business of Government, July 2005. Available at http:

//www.votingtechnologyproject.org/media/documents/AlvarezReport.pdf.

[8] Omar Benjelloun, Hector Garcia-Molina, Hideki Kawai, Tait Eliott Larson, David Menestrina, Qi Su, Sutthipong Thavisomboon, and Jennifer Widom. Generic entity resolution in the SERF project. *IEEE Data Engineering Bulletin*, 29(2):13–20, 2006.

[9] William W. Cohen, Pradeep Ravikumar, and Stephen E. Fienberg. A comparison of string distance metrics for name-matching tasks. In *Proceedings of the ACM Workshop on Data Cleaning, Record Linkage and Object Identification*, 2003.

[10] William W. Cohen, Pradeep Ravikumar, and Stephen E. Fienberg. A comparison of string metrics for matching names and addresses. In *Proceedings of IJCAI-03 Workshop on Information Integration on the Web*, pages 73–78, 2003.

[11] Committee on State Voter Registration Databases, National Research Council. *State Voter Registration Databases: Immediate Actions and Future Improvements, Interim Report*. National Academies Press, 2008.

[12] Ivan P. Fellegi and Alan B. Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64:1183–1210, 1969.

[13] Paul Gronke. Ballot integrity and voting by mail: The oregon experience, June 2005. Available at http://earlyvoting.net/resources/carter-baker-report-pr.pdf.

[14] Matthew A. Jaro. Advances in record linking methodology as applied to the 1985 census of tampa florida. *Journal of the American Statistical Society*, 84(406):414–420, 1989.

[15] Jeff Jonas. Entity resolution systems vs. match merge/merge purge/list de-duplication systems, 2007. Available at http://jeffjonas.typepad.com/jeff_jonas/2007/09/entity-resoluti.html.

[16] Justin Levitt and Wendy R. Weiser and Ana Muñoz. Making the list: Database matching and verification processes for voter registration, March 2007. Available at http://brennan.3cdn.net/96ee05284dfb6a6d5d_j4m6b1cjs.pdf.

[17] Michael P. McDonald and Justin Levitt. Seeing double voting: An extension of the birthday problem. *Election Law Journal*, 7(2):111–122, 2008.

[18] Robert C. Russell. US Patents 1,261,167 (April 2, 1918) and 1,435,663 (November 14, 1922). More information at http://en.wikipedia.org/wiki/Soundex.

[19] William E. Winkler. Advanced methods for record linkage. In *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pages 467–472, 1994.

[20] William E. Winkler. Matching and record linkage. In B. G. Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, M. J. Colledge, and P. S. Kott, editors, *Business Survey Methods*. J. Wiley, New York, NY, 1995.

[21] William E. Winkler. The state of record linkage and current research problems. In *Statistical Society of Canada, Proceedings of the Survey Methods Section*, pages 73–80, 1999. Longer version at http://www.census.gov/srd/papers/pdf/rr99-04.pdf.

Appendix A: Examples of Contact Letters

OFFICE OF THE SECRETARY OF STATE

BILL BRADBURY
SECRETARY OF STATE

ELECTIONS DIVISION
JOHN LINDBACK
DIRECTOR
255 CAPITOL ST NE, SUITE 5
SALEM, OREGON 97310
ELECTIONS — (503) 986-1518
FAX — (503) 373-7414

October 1, 2008

Dear Registered Voter:

A routine check of our state voter list and the state of Washington's voter list has indicated that you may be registered to vote in two different states. Your most recent date of registration is in the state of Washington however, the other voter registration address is located in Oregon. That address may be a previous residence for you, or it might be that of another properly registered voter. This may also be a result of a clerical error.

If you are certain that you have never registered to vote in Oregon, you do not need to do anything and may disregard this notice. However, because there is a state law that prohibits intentionally maintaining voter registrations in two locations simultaneously, you may voluntarily cancel any past registration you may have had in Oregon. To do so, please complete the section at the bottom of this letter and return it in the enclosed postage paid envelope.

If you are not sure or would like more information, please feel free to contact our office at (503) 986-1518 or email us at elections.sos@state.or.us.

Sincerely,

David Franks
HAVA Manager

Please cancel any invalid voter registration listed under my name at:

_____ _____ _____
(Previous Residence address) (Previous Residence City) (Zip Code)

_____ _____ _____
(First Name) (Middle Name) (Last Name)

_____ _____
(Signature) (Date)

Note: The signature you submit on this form will be compared to the signature on the your most recent registration card submitted in Oregon. This safeguard is necessary to ensure that this information submitted is from the elector. Thank you.

Washington
Secretary of State
SAM REED

520 Union Avenue
PO Box 40229
Olympia, WA 98504-0229
Tel: 360.902.4180
Fax 360.664.4619
www.secstate.wa.gov

October 8, 2008

Dear Registered Voter,

A routine check of our state voter list and the state of Oregon's voter list has indicated that you may be registered to vote in two different states. Your most recent date of registration appears to be in the state of Oregon. That address may be a previous residence for you, or it might be that of another properly registered voter. It might even have resulted from a clerical error.

If you are certain that you have never registered to vote in Washington, you do not need to do anything and may disregard this notice. If, however, you think you may have an old voter registration record in Washington State, I encourage you to voluntarily cancel that voter registration by completing the bottom section of this letter and returning it in the enclosed, postage-paid envelope.

If you are not sure or would like more information, please contact Dave Motz, Voter Services Manager, by calling (360) 725-5786 or by email (dmotz@secstate.wa.gov).

Sincerely,

Voter Registration Services
Elections Division, Office of the Washington State Secretary of State

---

Please cancel any invalid voter registration listed under my name at:

_____ _____ _____
(Previous Washington residence address) (Previous WA city) (ZIP Code)

_____ _____ _____
(First name) (Middle name) (Last name)

_____ _____
(Signature) (Today's Date)

Note: The signature you submit on this form will be compared to the signature on the registration record in question before it is cancelled. The same safeguards created for voter registration applications bearing voters' signatures will be used when processing this form.