Westat®

Improving lives
through research™

NCI/SEER Residential History Project

Technical Report

Authors

David G. Stinchcomb
Allison Roeser

May 18, 2016

Prepared for:
National Cancer Institute
Rockville, Maryland 20850

Contract # HHSN261201500371P

Prepared by:
Westat
An Employee-Owned Research Corporation®
1600 Research Boulevard
Rockville, Maryland 20850-3129
(301) 251-1500

Contents

Recommended citation:

Stinchcomb DG, Roeser, A. (2016). *NCI/SEER Residential History Project Technical Report*. Rockville, MD: Westat, Inc.

Executive Summary

Cancer research studies often need data on where people have lived throughout their lifetimes to assess prior risk exposures, both socioeconomic and chemical. Residential information in cancer registry data is generally limited to the address for a person at the time of diagnosis. A key problem facing researchers who wish to account for residential mobility in their analyses is the cost and difficulty of obtaining residential histories.. Recent studies have shown that commercial vendors are viable sources for information about prior residential locations. For this study, we identified three commercial vendors that could provide previous address data. To assess the accuracy of the commercially provided data, a set of self-reported residential histories was collected from volunteer participants at the National Cancer Institute and the National Institute of Environmental Health Sciences. We compared the accuracy and completeness of the residential histories derived from the vendor data with the self-reported residential histories.

For this study, the data from LexisNexis® was the most complete and accurate. The commercial data start around 1980 – there is very little data available before then. Data is available for deceased individuals. Only U.S. addresses are reported. The data that commercial vendors provide consist of a set of addresses associated with each individual rather than an actual residential history for the individual. We needed to develop an algorithm to construct residential histories from the vendor data. The derived residential histories were reasonably accurate and complete. We concluded that reasonable residential histories can be derived from vendor data and the derived histories yield significant accuracy improvements compared to assuming the person always lived at their current residence. This study demonstrates how a wide range of cancer research studies that need data on where people have lived prior to diagnosis can be conducted using existing data in cancer registries linked with commercial residential data.

1. Introduction

There is a growing recognition that residential mobility is an important factor in epidemiologic studies of cancer for evaluation of clusters, for reconstructing exposures, and as a source of exposure misclassification (Jacquez 2011, Pronk et al 2013, Boscoe 2011, Meliker et al 2010). Residential information in cancer registry data is generally limited to the address for a person at the time of diagnosis. A key problem facing epidemiologists who wish to account for residential mobility in their analyses is the cost and difficulty of obtaining residential histories. Recently, researchers have explored the use of commercially available sources to obtain residential histories. A study by Jacquez, et al (Jacquez 2011) evaluated the accuracy of residential histories from LexisNexis and found results that provided a level of accuracy to indicate that routine use of residential histories from commercial vendors is feasible. More recently, Wheeler and Wang also evaluated the use of LexisNexis data and concluded that these commercial data sources can be useful for reconstructing residential histories (Wheeler & Wang, 2015). Another less explored avenue for obtaining residential histories could be through use of publically available "big data" sources. Aligned with federal interest in using big data and in response to a request from NCI/SEER, Westat undertook a two part study to explore publically available big data sources of residential histories.

Part 1 of the study, the identification and evaluation of vendors providing residential history data, is described in a previous report (Westat, 2014). A brief summary is provided below. The main purpose of this report is to describe the study objectives, implementation and results of Part 2 of the study: testing the accuracy and completeness of the vendors able to provide residential history data.

1.1 Summary of Part 1

Westat performed a scan of publically available sources of big data to determine if residential history data could be obtained. Over 100 potential sources were reviewed for available residential history and mobility information, identified through a number of sources:

- Big data sources and initiatives were reviewed from the White House Fact Sheet "Big Data across the Federal Government"
- Social media data sources (e.g., Twitter, Foursquare, FullContact) were identified through consultation with the Westat social media group
- Sources, companies, and open-access data providers were extracted from an online search of the top 20 info-graphics returned using the search phrase "big data landscape"
- Additional online searches identified a number of sources providing "data-as-a-service" using search terms such as residential history data, GIS data, or other data types and formats associated with "big data."

Findings indicated that the most reliable information on residential mobility and residential history still appears to be developed and maintained by private "data-as-a-service" providers such as business intelligence services, marketing companies, or credit reporting agencies. Part 1 of the study yielded a set of possible vendors from which to obtain residential history data. Part 2 of the study then examined the accuracy and completeness of these vendors.

1.2 Objectives for Part 2

Westat assessed the feasibility of the top-ranked potential data sources from Part 1 by following-up with source representatives to establish logistics for a scalable method for obtaining residential histories. Only three commercial vendors were able to deliver the needed data. We refer to the vendors by number in this report so that the results can more readily be shared in situations where identifying the specific vendors may not be appropriate.

The objective of Part 2 was to evaluate the quality, completeness, and accuracy of the residential address data provided by these three sources. In addition to evaluating the three data sources individually, we sought to understand if more accurate or complete residential histories could be obtained by combining data from more than one vendor. We also wanted to know if data were available from the vendors on deceased individuals for research studies of highly fatal cancers or if the vendors removed data from their databases once an individual has died. Another objective was to characterize data completeness and accuracy by time period: is the data for more recent decades better than older data. Finally, we also sought to characterize the accuracy of the data both by distance (how far is the vendor data from the true location) and by changes in the geographic areas that are used to describe social determinants of health (is the true location in a different census tract or ZIP code).

2. Methods and Procedures

The study design included receiving residential histories from NCI and NIEHS employees. Westat then obtained vendor residential history data on the same set of individuals to assess accuracy and completeness.

2.1 Survey Design and Implementation

To test the accuracy of the available residential histories, Westat worked with colleagues from the National Cancer Institute (NCI) and the National Institute of Environmental Health Sciences (NIEHS) to recruit a convenience sample of employees and to administer a survey to obtain actual residential histories to use to compare with vendor-supplied data. A total of 66 residential histories were obtained including 10 from deceased relatives (for studies of cancers with poor survival rates).

The survey was administered over the web with open text fields. There was also a "Google Map" option where the respondent was able to "point and click" to receive locational coordinates (latitude and longitude) of their address or approximate location when a specific street address could not be recalled. In instances where participants lived outside of the U.S. or simply could not recall their street address, they could specify the locational coordinates from the Google Map interface. Respondents were asked to rate the accuracy of their address entries (being 100% certain of the location, within 1-2 blocks, etc.)

We asked respondents to include their full life-time residential history. However, respondents had the option to refuse providing any type of residential history data during any time period. To reduce the burden on the respondents, we instructed them to leave out addresses where they lived for less than six months (for example, summer addresses during college years). As a result, there reported residential histories sometimes included gaps in time.

2.2 Human Subjects Protection

Although this was a data quality study rather than a research study, all correspondence with participants as well as the design for the web survey was submitted through Westat IRB for approval. An electronic invitation for employees to participate was created, including a username and temporary password. Once the participant logged onto the site using their username and temporary password, they were asked to provide consent (if participating). After consent was provided, in order to secure the maximum level of protection, the participant was asked to provide a telephone number for two factor phone authentication before beginning the survey.

Because we asked participants for personal identifying information such as their social security number and date of birth, all data collected was treated as confidential data and secure data handling procedures were used. All data entered into the survey at the respondent's computer was encrypted by the internet browser before being transmitted to a secure central server using Secure Socket Layer

(SSL) technology. Response data were secured on the server using industry standard security controls, including firewalls and encryption. Electronic data was stored on a secure network server in a project folder with access restricted to specific project personnel. All personnel with access to the data received training in human subject protection and signed a data-use and confidentiality protection agreement. All data transmission with vendors and with study participants for the reconciliation process (see Section 2.5) used encrypted data and secure transmission protocols.

2.3 Vendor Data Requests

Westat submitted identifiers for these 66 individual to each of the three identified vendors and received residential address information back from each of them. The identifying information was selected based on individual identifiers that are commonly available in cancer registries which includes the individual's SSN. Table 1 summarizes the identifying information provided to each vendor for matching.

Table 1 – Matching information provided to vendors

Vendor	Information Provided
Vendor 1	First name, last name, date-of-birth, SSN, current street address, city, state, ZIP code, country
Vendor 2*	First name, last name, date-of-birth, current street address, city, state, ZIP code, country
Vendor 3	First name, last name, date-of-birth, SSN, current street address, city, state, ZIP code, country

* Vendor 2 does not accept social security numbers.

Vendor 1 is LexisNexis, the same data source as reported in the Jacques et al. 2011 and Wheeler and Wang 2015 papers. Where feasible, we compare our results to the results reported in these papers.

All vendors provided a series of addresses associated with matched individuals. However, there was significant variation in the format, completeness, and utility of the data returned by the vendors. Table 2 summarizes the data returned by each of the vendors.

Table 2 – Information returned by vendors

Vendor	Address Information	Other Information
Vendor 1	Current and all known previous addresses with from and to dates for each (month and year).	Other names, phone numbers, date of death if deceased
Vendor 2	Current and up to 5 previous addresses with a single effective date (start date) for each.	
Vendor 3	Current and previous addresses with from and to dates for each (day, month, and year)	Date of death if deceased

The data provided by the vendors consisted of a series of addresses associated with each individual rather than a residential history *per se*. Addresses usually included either a start and end date or an effective date but many of the dates were overlapping and some were missing altogether. This made it challenging to construct actual residential histories based on the vendor-provided residential address information – see Section 2.7.

2.4 Initial Address Matching

Name and social security number were used to identify the correct person in the vendor databases; in cases with incomplete social security numbers, names and dates of birth were used. Vendor person-match rates were calculated for each vendor with separate accounting for deceased individuals. Additional statistics were calculated to compare the number of addresses, the years at each address, and the distribution of address years for the survey-reported address information and the vendor data. To compare the availability of data across time, the number of addresses reported per month was calculated for each data source and plotted.

Individual addresses were then matched between the survey-reported address information and the vendor data. We used probabilistic matching methods to identify matches so that minor variations in the address information would not prevent a match. Matches on the full address, the street name only, and the city only were reported separately. Addresses were geocoded prior to matching so that distances could be used in the matching process. This helped identify matches when there were alternative names for the same street. ZIP codes were used to identify city-level matches where different city names were used for the same ZIP code (for example, Gaithersburg, MD and Montgomery Village, MD are alternative city names for the same ZIP code). Address match rates were calculated in two directions: the number of survey addresses found in the vendor data and the number of vendor addresses found in the survey results.

2.5 Residential History Visualization

To better understand the relationships between vendor address time-frames and survey-reported address time-frames, temporal bar charts were developed depicting each person's survey-reported residential history over time with matched and unmatched vendor addresses. To avoid including any specific identifying information, survey-reported addresses are referred to simply as "Address 1", "Address 2", etc. These bar charts helped to demonstrate the complexity of the data received from the vendors and provided a starting point for the development of an algorithm to derive actual residential histories from the collection of vendor residential addresses (see Section 2.8).

2.6 Reconciliation Process

The matched addresses were organized into one spreadsheet for each participant for their review. Address information was organized by self- reported address with any matching addresses from each of the three vendors beside it. Various types of discrepancies where highlighted including vendor addresses that did not appear on the survey-reported list, vendor addresses that had move-in dates before the survey-reported address, or vendor addresses with street numbers or spellings that differed from the self-report. Respondents were given an opportunity to correct or explain the inconsistencies between the survey-reported and vendor results. Afterwards, the results were compiled and any changes to the survey-reported residential histories were used to update the original data.

2.7 Time-frame Comparisons

For matched addresses, we calculated several time-frame metrics that were reported in previous studies of commercial residential history data. Jacquez et al. report the mean number of survey-reported years at addresses that were matched by vendor addresses (Jacquez et al. 2011, Metric 5). Wheeler and Wang report the difference between the vendor reported and the survey-reported time spent at each address (Wheeler and Wang 2015, Metric 6). We calculated these same metrics for comparison purposes. We calculated two additional metrics: the difference in years between the starting date of the survey-reported address and the matched vendor address, and a similar metric for the ending date. Since we know the vendor ending dates are not reliable due to the forwarding of mail after a move, we wanted a pair of time-reliability measures that treated the starting and ending dates separately. We use the results of the starting date measure as a weighting variable in the algorithm described in the next section.

2.8 Deriving Residential Histories from Vendor Data

The address information returned by the data vendors consisted of individual addresses with associated dates. Many addresses appeared more than once and had overlapping and conflicting dates. There were also many gaps in time when there was no address associated with an individual. To convert the data received from the vendors into a residential history, an algorithm was developed.

The algorithm was based on first combining all of the duplicate addresses including those with slightly different forms of a street address, deciding on the most likely time-frame for that address, and then constructing a complete sequential residential based on the known information. We have been told that the ending dates in the vendor data files tend to extend past the time when people move to new addresses because mail is often forwarded for some time after a move. The algorithm takes this into account, preferring the starting date of a subsequent address over the ending date of the previous address.

The basic steps in the algorithm are:

1. Match addresses within each vendor to identify duplicates and synonyms.
2. Match addresses across vendors (Vendor 1 with Vendor 2, Vendor 1 with Vendor 3, and Vendor 2 with Vendor 3).

3. Combine matched addresses from all of the vendors.
4. Decide on a time-frame for each address. This is done by combining time frames using a histogram-based analysis of the time frames for all combined addresses. Variable thresholds were included to control time frame selection (for example: 0% to 100% (earliest date to latest date), 10% to 90%, 25% to 75%, etc.).
5. Weed out short duration addresses. Note that any gaps in time will be filled in the next step.
6. Build residential history working backwards from the most recent address. Use current start date as end date for the previous address.

The method used in the last step is analogous to the "last observation carried forward" (LOCF) method used with missing data.

Separate residential histories were derived for each vendor independently as well as for all combinations of pairs of vendors and all three vendors combined. Hence, a total of seven different residential histories were derived from the vendor-supplied data.

A number of tuning parameters were included in the algorithm. By varying these parameters and comparing the accuracy of the resultant residential histories to survey-reported histories, parameter settings that provide the most accurate and complete histories can be established. We explored variations in the tuning parameter values and calculated the how the setting impact both accuracy and completeness of the derived histories.

2.9 Assessing the Quality of Derived Residential Histories

To compare the accuracy and completeness of the derived residential histories with the survey-reported histories, we compared the geographic location of individuals at all points of time. We measured completeness by calculating the proportion of time with survey-reported locations that we also had locations from vendor data. We refer to this measure as the coverage. We limit the measure to times when survey-reported locations are in the U.S.

Two measures of accuracy were used, one based on distance and one based on changes to geographic areas. For the distance calculation, we report descriptive statistics for the time-weighted distance between survey-reported and vendor-reported locations as well as the proportion of time that the distance is zero, less than 1 kilometer, less than 5 kilometers, and less than 10 kilometers. The measures based on geographic areas are the proportion of time that there is a difference in the census tract, ZIP code, or county of the survey-reported and vendor-reported locations.

To assess how accuracy and completeness vary over time, we compared residential histories for seven different time periods: the full reported life history[*]; three time periods ending in the current

[*] Because some survey respondents started their survey-reported histories significantly after their date-of-birth, we measure the full life span from the start date of the first survey-reported address rather than from the date-of-birth.

year: 30 years from 1986 to 2015, 20 years from 1996 to 2015, and 10 years from 2006 to 2015; two time periods ending in 2005: 20 years from 1986 to 2005 and 10 years from 1996 to 2005; and the additional individual decade from 1986 to 1995. Current geographic boundary files were used for all time periods.

Hence, a complete set of comparisons involved comparing the survey-reported histories with seven different vendor combinations and seven different time periods for a total of 49 comparisons. These 49 comparisons were repeated for each set of algorithm tuning parameters that was tested.

One final type of residential history was developed for comparison purposes. For this history, we assumed that each individual lived at their current address for their entire lives. This type of history is the equivalent of the approach used in many health research studies when no information is available about residential histories. For cancer research, the address at the time of diagnosis is used rather than the current address, but the same assumption is made: that the individual does not change residential locations for the duration of the study period. We used this derived residential history for the final set of comparisons of the tuned algorithm.

3. Results

A variety of results were obtained from Part 2 of the NCI/SEER Residential History Study. The section presents summary statistics describing the data received from the three commercial vendors, how well the addresses in these data match with the survey-reported addresses, and visualizations showing the relationship of the vendor data to the survey-reported data over time, We then describe the results of the reconciliation process and comparisons of time-frames of the addresses after reconciliation. Finally, we present the results of changing the parameters of the algorithm to derive residential histories from the vendor-supplied address information and the final comparisons to assess the accuracy and completeness of the resulting residential histories.

3.1 Data Received from Vendors

We submitted identifying information for the 66 participants in our survey to each of the three data vendors. Tables 3, 4, 5, and 6 summarize the characteristics of the data we received in response. In Table 3, we describe the ability of the vendors to match the participants in their databases. Table 4 gives the total number of addresses, addresses per person, and rates of missing date information. Table 5 contains information on the geocoding rates. Table 6 provides statistics on the length of time people spent at each address.

Table 3. Person match rates

	Total number of individuals		Living individuals		Deceased individuals	
	N	Pct.	N	Pct.	N	Pct.
Survey	66	100%	56	100%	10	100%
Vendor 1	64	97%	56	100%	8	80%
Vendor 2	52	79%	48	86%	4	40%
Vendor 3	57	86%	50	89%	7	70%

All of the vendors were able to match a reasonable number of the survey participants. Vendor 1 matched all but two; Vendors 2 and 3 matched from 79% to 86% of the survey participants. All vendors had data on deceased individuals with Vendor 1 able to provide data on 8 of the 10 deceased individuals in the survey and Vendor 3 able to provide data on 7 of 10 deceased individuals. There was no evidence of false positive matches at the person level: we were able to identify at least one common address between the survey and the vendor for each matched individual.

Table 4. Address counts and date completeness

	Number of address				Addresses per person		Addresses with from-to dates		
	Total	US	Non-US	APO*	Avg.	Max.	N	Pct.	
Survey	587	549	38	0	8.89	24	587	100%	🟦
Vendor 1	636	631	0	5	9.94	15	621	98%	🟩
Vendor 2	156	154	0	2	3.00	6	153	98%	🟩
Vendor 3	596	589	0	7	10.46	26	584	98%	🟩

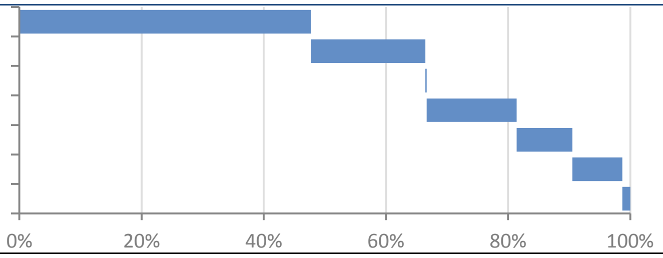
* APO addresses are overseas military mailing addresses.

The vendor results did not include any addresses outside the U.S. but did include some APO (overseas military) addresses. In the survey results, there were a little less than nine addresses reported per person with one person reporting 24 resident addresses. Vendor 1 and Vendor 3 data had slightly more addresses per person. Vendor 2 had substantially fewer addresses per person. There were addresses with missing date information for all of the vendors. However, Vendors 1, 2, and 3 data had dates for over 95% of their addresses.

Table 5. Geocoding rates for U.S. addresses

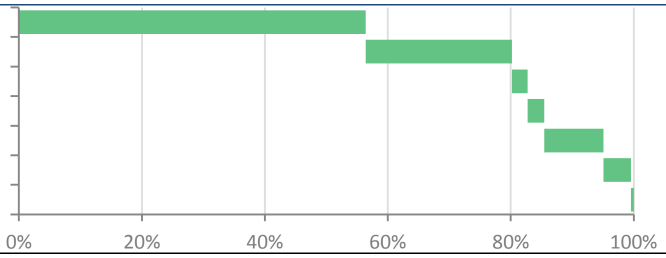
Survey:

Geocode level	N	Pct.	Cumulative N	Cumulative Pct.
1: Point location	262	47.7%	262	47.7%
2: Street address	103	18.8%	365	66.5%
4: 9 digit ZIP	1	0.2%	366	66.7%
5: Street name	81	14.8%	447	81.4%
6: 5 digit ZIP	50	9.1%	497	90.5%
7: Admin place	45	8.2%	542	98.7%
9: Unable to match	7	1.3%	549	100.0%



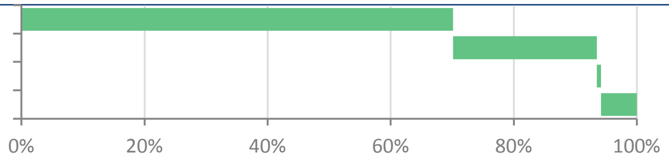
Vendor 1:

Geocode level	N	Pct.	Cumulative N	Cumulative Pct.
1: Point location	356	56.4%	356	56.4%
2: Street address	150	23.8%	506	80.2%
4: 9 digit ZIP	16	2.5%	522	82.7%
5: Street name	17	2.7%	539	85.4%
6: 5 digit ZIP	61	9.7%	600	95.1%
7: Admin place	28	4.4%	628	99.5%
9: Unable to match	3	0.5%	631	100.0%

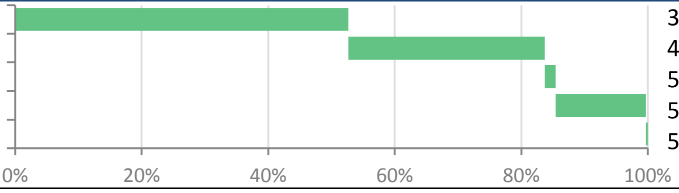


Vendor 2:

Geocode level	N	Pct.	Cumulative N	Cumulative Pct.
1: Point location	108	70.1%	108	70.1%
2: Street address	36	23.4%	144	93.5%
5: Street name	1	0.6%	145	94.2%
6: 5 digit ZIP	9	5.8%	154	100.0%



Vendor 3:			Cumulative	
Geocode level	N	Pct.	N	Pct.
1: Point location	310	52.6%	310	52.6%
2: Street address	183	31.1%	493	83.7%
5: Street name	10	1.7%	503	85.4%
6: 5 digit ZIP	84	14.3%	587	99.7%
7: Admin place	2	0.3%	589	100.0%



Geocoding rates were generally good, particularly for the vendor-supplied addresses. Addresses were geocoded to the street address level or better for 66% of the survey addresses, 80% of Vendor 1 addresses, 93% of Vendor 2 addresses and 84% of Vendor 3 addresses. Addresses were geocoded to the street name level or better (including 9-digit ZIP codes) for 81% of the survey addresses, 85% of Vendor 1 addresses, 94% of Vendor 2 addresses and 85% of Vendor 3 addresses.

Table 6. Number of years at each address

	Years at address		
	Average	Min.	Max.
Survey	4.7	-0.2 *	44.2
Vendor 1	5.3	0.1	55.3
Vendor 2	6.9	0.0	32.1
Vendor 3	2.1	0.0	24.7

* For one survey-reported address, the end date was before the start date.

The survey data reported that people stayed at a residence location for an average of 4.7 years. The time at each address was slightly higher for Vendor 1 and more substantially higher for Vendor 2. The time at each address was substantially lower for the Vendor 3 data.

To evaluate the availability of vendor data over time, we calculated the number of addresses that were reported in the survey responses for each year. Since the survey includes deceased individuals, the data starts in the early 1900s with the bulk of the data from the 1950s through the present. We then calculated the same measure for each of the three vendors. The results are shown in Figure 1.

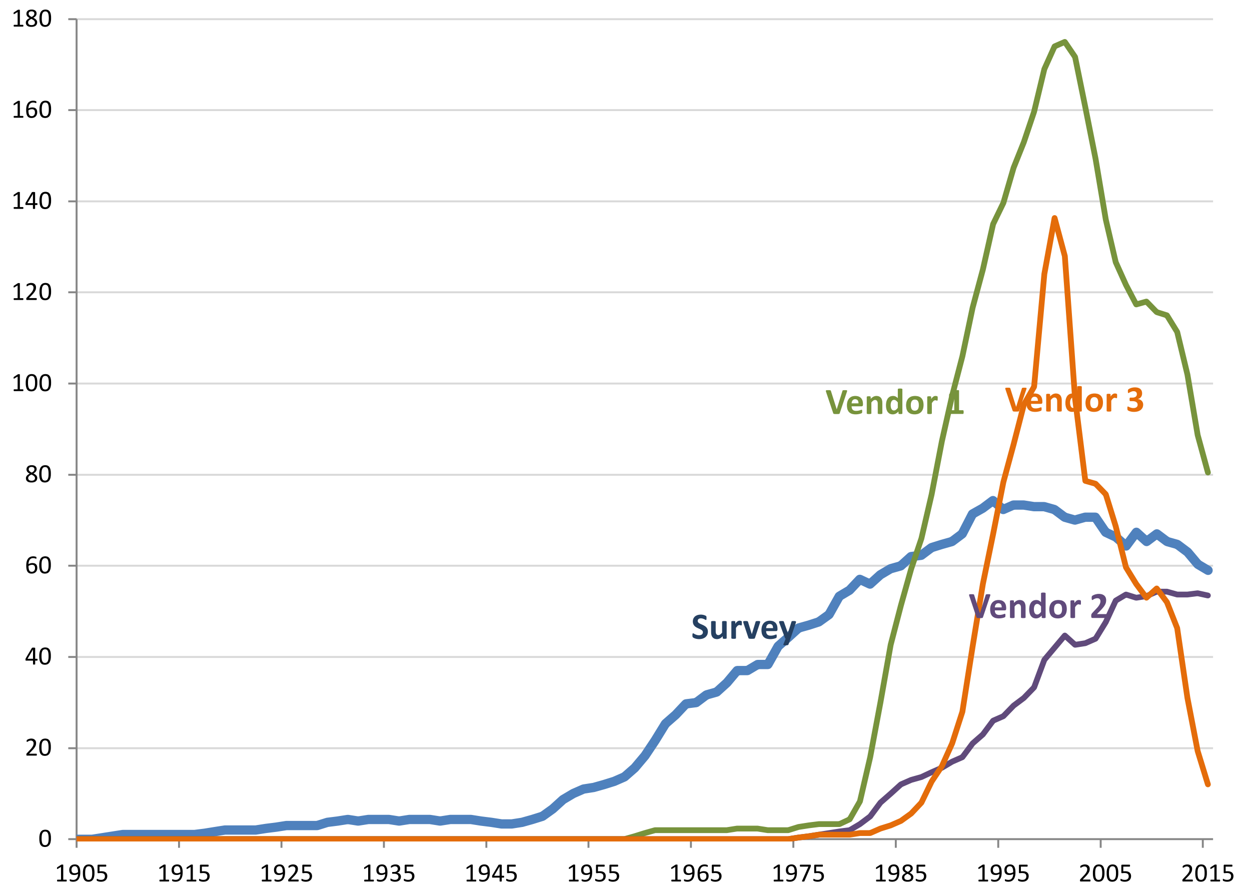


Figure 1. Number of addresses per year for survey and vendor data

Since there are 66 individuals in the survey results, the number of addresses per year should not be much higher than 66 (people will be at more than one address per year if they moved during that year). The survey data peak at about 74 addresses per year in 1994. It drops below 66 in the more recent years because deceased individuals were no longer reporting addresses. The Vendor data does not really start until the 1980s. The number of addresses per year for Vendor 1 increases quickly in the 1980s with substantial data available by 1985. The data from Vendor 1 includes many duplicate addresses with slightly different street address information, for example “123 Main Street” versus “123 Main St.” We speculate that the drop in the number of addresses per year from Vendor 1 around year 2000 is due to an effort by the vendor to reduce this duplication in their database. The data from Vendor 2 increases more gradually starting in the 1980s. It approaches the number of matched people (52 of the 66) around 2005. Vendor 3 begins ramping up in the late 1980s and includes multiple addresses per person in the 1990s and 2000s. In the more recent years, Vendor 3 data drops below the expected number of addresses based on the survey responses.

3.2 Address Match Results

To get an initial assessment of the agreement between survey addresses and vendor addresses, we used probabilistic matching methods to identify common addresses. Tables 8 and 9 describe the results for detailed matches (full street address, city and state), street level matches (street name but not the street number), and city level matches. Table 7 reports the number of vendor addresses found in the survey results and Table 8 reports the number of survey addresses found in the vendor data. Both tables include data from prior studies using data from Vendor 1.

Table 7. Vendor addresses that match a survey response address

	Number of addresses	Detailed match			Street match			City match *		
		N	Pct.		N	Pct.		N	Pct.	
Jacquez** Vendor 1	2,388	1,259	53%	🟩	1,475	62%	🟩	1,701	71%	🟩
Vendor 1	636	280	44%	🟩	352	55%	🟩	497	78%	🟩
Vendor 2	156	116	74%	🟩	123	79%	🟩	140	90%	🟩
Vendor 3	596	281	47%	🟩	329	55%	🟩	458	77%	🟩

* Includes 5-digit ZIP code matches

** Jacquez et al. 2011

Our match rates for Vendor 1 agree fairly well with those from the prior study. Vendor 2 has higher match rates indicating that they have fewer addresses that are not part of the survey-reported address histories than Vendor 1. The match rates for the Vendor 3 data are similar to those of Vendor 1.

Table 8. Survey response addresses that match a vendor address

	Number of addresses	Detailed match			Street match			City match *		
		N	Pct.		N	Pct.		N	Pct.	
Wheeler** Vendor 1	10,327	8,871	86%	🟩	8,919	86%	🟩	9,100	88%	🟩
Vendor 1	583	261	45%	🟩	307	53%	🟩	389	67%	🟩
Vendor 2	485	111	23%	🟩	118	24%	🟩	170	35%	🟩
Vendor 3	505	202	40%	🟩	228	45%	🟩	293	58%	🟩

* Includes 5-digit ZIP code matches

** Wheeler and Wang 2015. Addresses limited to study period: 1995 to 2013.

Our match rates for Vendor 1 are lower than those from the prior study. The prior study only had addresses during the study period rather than the full life span. These results indicate that Vendor 1 has about half of the survey-reported addresses (45% at the detailed match level and 53% at the street level). Vendor 2 has substantially fewer: between a quarter and a third of survey addresses. The match rates for the Vendor 3 data are similar to but slightly less than the Vendor 1 match rates. Note that the total number of survey addresses is different for each vendor because these calculations use only addresses from matched individuals (64 for Vendor 1, 52 for Vendor 2 and 57 for Vendor 3 data).

3.3 Visualization of Reported Histories

To better understand the relationships between the survey-reported addresses and the vendor addresses across both space and time, we displayed the data on temporal bar charts. Each survey-reported address is presented in sequence from top to bottom with the bar showing when they lived at that address. If vendor addresses match a survey-reported address, it is shown with the survey-reported address. Unmatched vendor addresses are shown where they would appear chronologically in the history. Vendor addresses with missing date information is shown with small bars on the right side beyond the 2015 end date of the study. An example is shown in Figure 2.

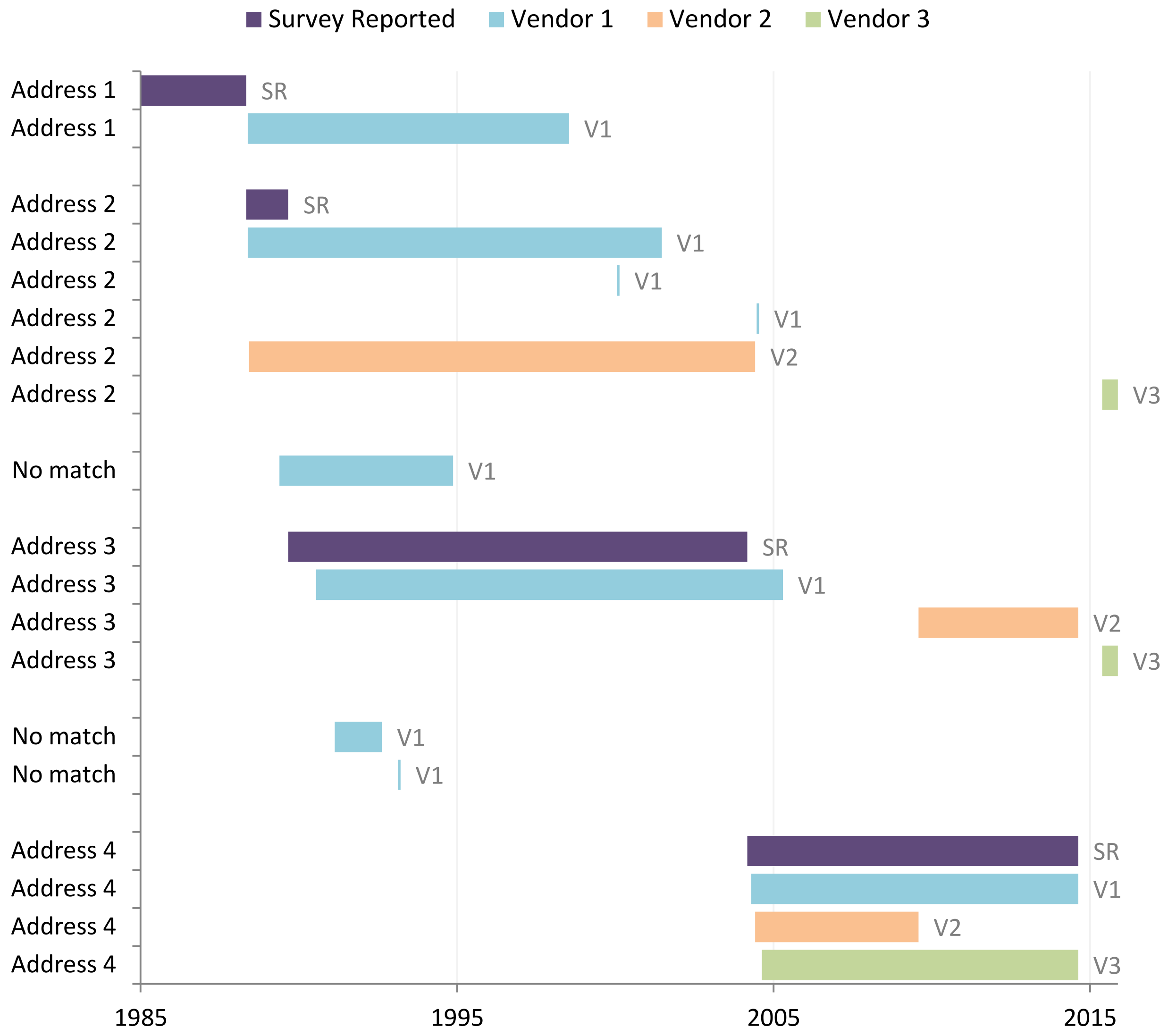


Figure 2. Example of a temporal bar chart for one study participant

The darker bars (labeled "SR") show the survey-reported addresses in sequence moving from Address 1 to Address 4 over the course of 30 years. Vendor 1's data includes Address 1 but has a start date about the same time as the person moved from Address 1 to Address 2. Vendor 1 has three address records matching Address 2: one record that overlaps the survey-reported time frame but extends much further in time and two records with very short durations both well after the person left that address. Vendor 1 had matching records for both Addresses 3 and 4 that align fairly well in time. Vendor 1 also has three records for two addresses that do not appear in the survey-reported history. Vendor 2 matches three of the four survey-reported addresses but has them in the wrong order, moving from Address 2 to Address 4 and then to Address 3. Vendor 3 (original data) has accurate information about Address 4 (the current address) and knows about Addresses 2 and 3 but does not have a time-frame associated with these addresses.

3.4 Reconciliation Results

We received 52 responses from the reconciliation requests (79%). These responses included answers to 335 specific questions about observed discrepancies between the survey-reported information and the data provided by vendors. A dropdown list of possible responses was provided for each question with an "other (please specify)" option available if needed. Table 9 provides a summary of the responses to these questions.

Table 9. Summary of responses to discrepancy questions

Question Type	Response	Count	Percent
Address details	Yes, vendor data is correct	58	77%
	No, vendor data is not correct	17	23%
Unreported address	Yes, I just forgot this one	39	23%
	Yes, but not part of my residential history	87	52%
	Yes, I lived there less than 6 months	15	9%
	Yes, but this is not a residential address	24	14%
	Yes, other	38	23%
	Yes, this was temporary/part time	10	6%
	No, I don't recognize this address	41	24%
Address time-frames	Yes, vendor data is correct	21	27%
	No, vendor data is not correct	56	73%

For questions about specific details of an address, usually the street number of the street name, the respondents indicated that the vendor information was correct 77% of the time. For unreported addresses (addresses that appeared in the vendor data but not in the survey-reported data), respondents indicated that about half the time (52%) the address was associated with them but was not a residence. About a quarter (23%) were addresses that the respondent forgot to include. The remaining quarter (27%) were addresses that the person did not recognize.

Respondents indicated that some of the unreported addresses were work addresses. This could be caused by using a work address for household type activities such as magazine subscriptions or utility bills. For constructing residential histories from vendor data, it might be useful to be able to weed-out these non-residential addresses. It would be possible to use the US Postal Service's "Residential Delivery Indicator" product (https://www.usps.com/nationalpremieraccounts/rdi.htm, available from a variety of vendors) to identify business and residential addresses,

Based on the reconciliation responses, we were able to update the survey-reported residential histories. Table 10 gives a summary of the types of updates.

Table 10. Changes to survey-reported histories based on reconciliation responses

Type of Change	Count
Address addition	39
Address deletion	1
Change of address details:	108
Street Number/Name	66
ZIPcode	2
From Month/Year	24
To Month/Year	16

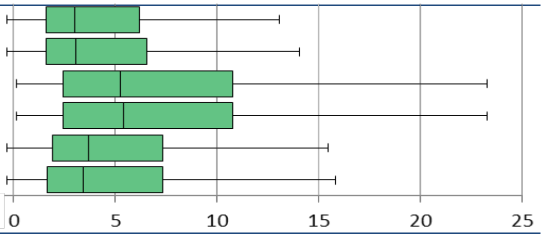
We were able to add 39 address records, delete one address record (this was an accidental duplicate entry), and made 108 changes to reported address details. This updated set of residential histories was used for the assessment of the residential histories derived from the vendor data.

3.5 Time-frame Comparison Results

Using the updated survey-reported residential histories after reconciliation, we calculated several measures to quantify differences in time-frames for matched addresses. In Table 11, we report statistics for the number of years at addresses that were matched by vendor addresses (analogous to Metric 5 in Jacquez et al. 2011). Table 12 gives statistics for the difference between the vendor reported and survey-reported time spent at each address (analogous to Metric 6 in Wheeler and Wang 2015). Positive values indicate longer vendor time intervals; negative values indicate longer survey-reported time intervals. In Table 13 we provide statistics for the difference in years between the starting dates of the survey-reported address and the matched vendor address. Table 14 provides statistics for a similar measure based on ending dates. For both metrics, the absolute value of the difference is used to reflect the overall accuracy of the vendor time-frame data. Box plots in these tables show the median, 25th and 75th percentiles and 1.5 IQR values bounded by the minimum and maximum.

Table 11. Years at matched addresses

Source	Match level	Pct. Of Survey Years	N	Mean	Median
Jacquez* Vendor 1	Detailed	62.6	1,259	10.7	NA
	Street	71.5	1,475	12.2	NA
Vendor 1	Detailed	73.8	367	5.6	3.0
	Street	83.1	406	5.7	3.0
Vendor 2	Detailed	36.0	129	7.7	5.2
	Street	37.5	133	7.8	5.4
Vendor 3	Detailed	77.9	348	6.2	3.7
	Street	81.1	370	6.1	3.4

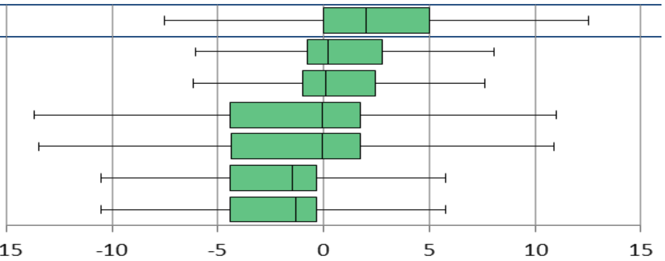


* Jacquez et al. 2011, three most recent addresses.

Comparing the percentage of survey years accounted for by matched addresses, our results are similar to those reported in Jacquez et al, 2011. Vendor1 and Vendor 3 data account for the largest percent of survey years. Comparing the number of years at matched addresses, our results are lower than those reported in Jacquez et al, 2011. Vendor 1 has smaller mean years at address than Vendors 2 and 3 which indicates that addresses are more likely to be in Vendor 2 and 3's databases if the person lived there for a longer period of time. Vendor 1 is better able to pick up short duration addresses. Jacquez et al. intended this metric to indicate what portion of the survey-reported history is covered by matched vendor data. However this metric assumes that the vendor has the same time-frame (both start date and duration) for the matched address which is clearly not a good assumption based on the temporal bar charts in Section 3.3. In addition, this measure inflates coverage because of duplicate addresses in the vendor data.

Table 12. Difference in duration at matched addresses

Source	Match level	N	Mean	Median
Wheeler* Vendor 1	Detailed	8,871	2.9	2.0
Vendor 1	Detailed	359	1.6	0.2
	Street	397	1.0	0.1
Vendor 2	Detailed	127	-0.5	0.0
	Street	130	-0.5	0.0
Vendor 3	Detailed	343	-3.1	-1.5
	Street	365	-3.2	-1.3

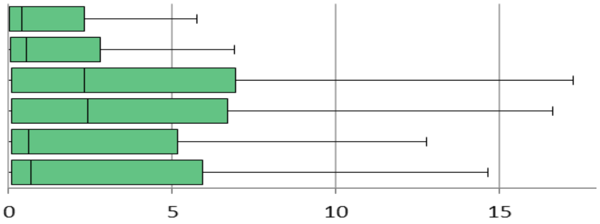


* Wheeler and Wang 2015. Addresses limited to study period: 1995 to 2013.

Our results for Vendor 1 with detailed matches are slightly lower than those reported in Wheeler and Wang 2015 but in the same range. The mean and median values for Vendor 1 are both positive indicating longer time intervals than in the survey-reported data. In contrast, the mean and median values for Vendors 2 and 3 are zero or negative indicating shorter time intervals than in the survey-reported data. Note that Vendor 3 has negative values even for the 75th percentile indicating that Vendor 3 time intervals are frequently shorter than survey-reported intervals.

Table 13. Differences in the starting date of matched addresses (in years)

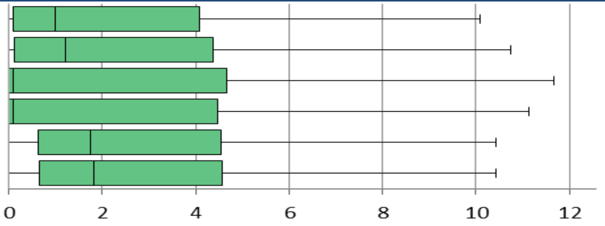
Source	Match level	N	Mean	Median
Vendor 1	Detailed	359	2.6	0.4
	Street	397	2.7	0.5
Vendor 2	Detailed	127	4.9	2.3
	Street	130	4.8	2.4
Vendor 3	Detailed	343	3.7	0.6
	Street	365	3.8	0.7



The median values indicate that Vendor 1's starting dates are the most accurate with about a half a year difference. Vendor 3 data has similar median values but with a larger mean and a wider distribution of values. Vendor 2 has a median value of 2.3-2.4 years. The algorithm used to derive residential histories from the vendor data relies primarily on the starting dates due to known inaccuracies in reporting address ending dates. For tuning the algorithm, we used the median values for the detailed matches from this table (0.4, 2.3, and 0.6) to assign weights to the time-frame data for the different vendors.

Table 14. Differences in the ending date of matched addresses (in years)

Source	Match level	N	Mean	Median
Vendor 1	Detailed	359	3.2	1.0
	Street	397	3.4	1.2
Vendor 2	Detailed	127	3.1	0.1
	Street	130	3.1	0.1
Vendor 3	Detailed	343	3.4	1.7
	Street	365	3.5	1.8



Vendor 2 has the most accurate ending dates with a median value of 0.1 years. Vendor 1 has a median value of 1.0 to 1.2 years. The Vendor 3 data has a median value of 1.7 to 1.8 years. However, the 75th percentiles for all three vendors have ending dates 4 to 5 years after the survey-reported data. As mentioned above, the algorithm used to derive residential histories from the vendor data relies primarily on the starting dates due to known inaccuracies in reporting address ending dates.

3.6 Derived History Algorithm Tuning Results

We experimented with the settings of several tuning parameters in the algorithm used to derive residential histories from vendor-supplied address information. The following figures display the results of these experiments for each of the 49 combinations of vendor input files and time spans. For each, we show the difference in three metrics: the percent of time period coverage, the percent of covered time with distances less than one kilometer, and the percent of covered time with locations in the same census tract. Values are color-coded to facilitate interpretation of the results: large negative values are red indicating poorer performance and large positive values are green indicating better performance.

Differences are reported as the percent change of a test run compared with a base run ("Test:Base"). The tuning runs are labeled in these figures with an abbreviation as described in Table 15.

Table 15.Tuning run abbreviations and tuning parameters

Run abbr.	Minimum duration (days)	Trimming Limits (%)		Trim Groups Only	Time-frame weights		
		Lower	Upper		Vendor 1	Vendor 2	Vendor 3
Base	0	0	100	True	1.0	1.0	1.0
D1mo	32	0	100	True	1.0	1.0	1.0
D6mo	186	0	100	True	1.0	1.0	1.0
T1090	32	10	90	True	1.0	1.0	1.0
T2575	32	25	75	True	1.0	1.0	1.0
T1075	32	10	75	True	1.0	1.0	1.0
T09T	32	0	90	True	1.0	1.0	1.0
T19F	32	10	90	False	1.0	1.0	1.0
T09F	32	0	90	False	1.0	1.0	1.0
W1090	32	10	90	True	2.39	0.43	1.59

Figure 3 shows the results of experiments with the minimum time interval threshold.

		% change D1mo:Base			% change D6mo:Base		
Vendors	Time-span	Coverage	Within 1 km	Same tract	Coverage	Within 1 km	Same tract
V1	Full	-1.02	3.95	4.36	-0.97	4.01	4.47
V2	Full	0.00	0.00	0.00	-0.23	0.07	0.11
V3	Full	-6.52	16.82	16.65	-8.15	16.61	16.47
V1V2	Full	-1.02	4.11	4.54	-1.04	4.16	4.65
V1V3	Full	-1.05	5.50	5.07	-1.23	5.50	5.07
V2V3	Full	-0.49	6.12	5.47	-1.20	6.46	5.82
V1V2V3	Full	-1.01	7.80	7.31	-1.18	7.91	7.43
V1	1986_1995	-1.82	8.08	10.66	-1.82	7.48	10.24
V2	1986_1995	0.00	0.00	0.00	-0.48	0.03	0.03
V3	1986_1995	-1.54	5.05	5.11	-2.58	4.78	4.81
V1V2	1986_1995	-1.80	8.74	11.65	-1.80	8.07	11.18
V1V3	1986_1995	-1.81	6.86	9.77	-1.81	6.46	9.32
V2V3	1986_1995	-0.24	1.27	1.29	-0.84	0.65	0.65
V1V2V3	1986_1995	-1.79	10.11	13.61	-1.79	9.68	13.13
V1	1986_2005	-0.91	6.22	7.03	-0.81	6.37	7.31
V2	1986_2005	0.00	0.00	0.00	-0.25	0.08	0.17
V3	1986_2005	-2.07	14.54	14.41	-2.78	14.48	14.36
V1V2	1986_2005	-0.91	6.65	7.50	-0.92	6.78	7.77
V1V3	1986_2005	-0.91	8.33	8.14	-1.18	7.95	7.79
V2V3	1986_2005	-0.12	5.32	4.22	-0.67	5.94	4.83
V1V2V3	1986_2005	-0.90	11.04	10.79	-1.17	10.86	10.63
V1	1986_2015	-0.59	3.96	4.39	-0.53	4.02	4.51
V2	1986_2015	0.00	0.00	0.00	-0.24	0.08	0.12
V3	1986_2015	-6.60	17.06	16.88	-8.25	16.84	16.70
V1V2	1986_2015	-0.58	4.16	4.61	-0.61	4.22	4.73
V1V3	1986_2015	-0.62	5.72	5.25	-0.81	5.73	5.27
V2V3	1986_2015	-0.50	6.27	5.58	-1.24	6.62	5.94
V1V2V3	1986_2015	-0.58	8.05	7.52	-0.77	8.19	7.66
V1	1996_2005	-0.27	5.06	5.02	-0.11	5.61	5.63
V2	1996_2005	0.00	0.00	0.00	-0.15	0.09	0.20
V3	1996_2005	-2.29	18.45	18.11	-2.87	18.47	18.13
V1V2	1996_2005	-0.27	5.38	5.30	-0.30	5.94	5.93
V1V3	1996_2005	-0.27	9.01	7.15	-0.74	8.68	6.90
V2V3	1996_2005	-0.07	7.10	5.39	-0.60	8.26	6.51
V1V2V3	1996_2005	-0.27	11.38	9.26	-0.74	11.39	9.31
V1	1996_2015	-0.14	2.80	2.79	-0.07	3.02	3.04
V2	1996_2015	0.00	0.00	0.00	-0.19	0.06	0.10
V3	1996_2015	-7.74	19.88	19.58	-9.53	19.71	19.48
V1V2	1996_2015	-0.14	2.92	2.91	-0.17	3.15	3.16
V1V3	1996_2015	-0.19	5.29	4.08	-0.45	5.43	4.23
V2V3	1996_2015	-0.56	7.27	6.40	-1.33	7.83	6.96
V1V2V3	1996_2015	-0.14	7.39	6.04	-0.39	7.69	6.33
V1	2006_2015	0.00	0.91	0.91	-0.03	0.88	0.87
V2	2006_2015	0.00	0.00	0.00	-0.22	0.07	0.07
V3	2006_2015	-14.73	23.07	22.91	-18.08	23.10	23.10
V1V2	2006_2015	0.00	0.91	0.91	-0.03	0.88	0.86
V1V3	2006_2015	-0.11	2.21	1.45	-0.14	2.69	1.90
V2V3	2006_2015	-1.08	7.54	7.37	-2.09	7.63	7.49
V1V2V3	2006_2015	0.00	4.08	3.26	-0.03	4.58	3.72

Legend: <= -5% (loss) > -5% and < -0.5% > 0.5% and < 5% >= 5% (gain)

Figure 3. Changing the minimum time interval threshold

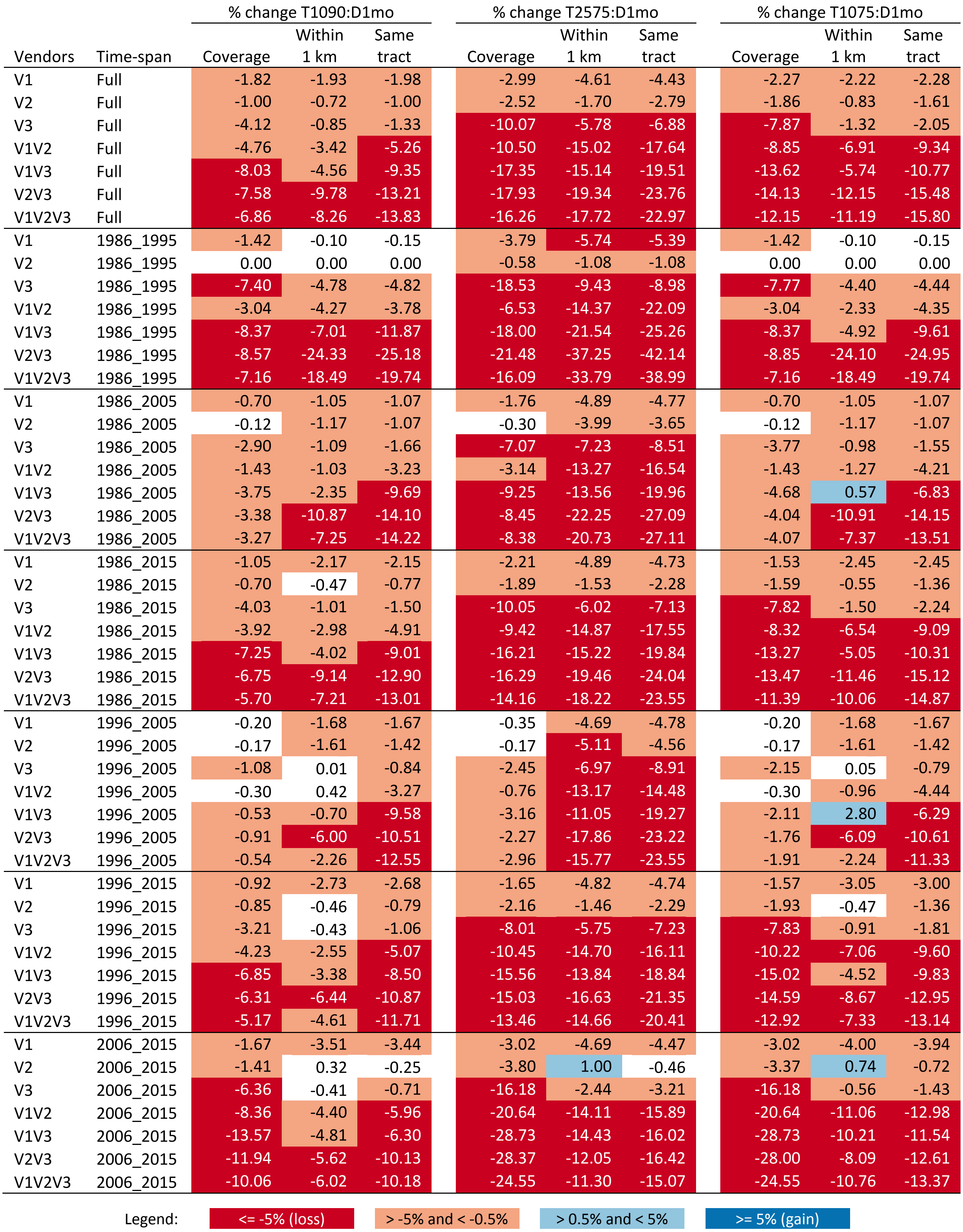
The base configuration of the algorithm uses a value of zero – it does not eliminate any short duration addresses. The test configuration for the first set of results labeled "D1mo:Base" uses an address duration minimum of 32 days – addresses with durations of a month or less are removed. The test configuration for the second set of results labeled "D6mo:Base" uses an address duration minimum of 186 days – addresses with durations of six months or less are removed. This corresponds to the instructions given to survey respondents not to include places where they lived for less than six months.

As can be seen in Figure 3, there is some loss of coverage using minimum time interval thresholds with a more substantial improvement in accuracy. The improvements are largest for the earlier time periods. There is little additional improvement gained by using a six month threshold rather than a one month threshold.

As shown in Figure 3, there is a both a larger loss of coverage and a larger increase in accuracy for residential histories generated from just the Vendor 3 data and just for the time periods that include the most recent decade. This is because the vendor 3 results have a large number of addresses with the same starting and ending dates. Coverage was calculated assuming that they stayed at these addresses for a month. These addresses were not used for the 1 month and 6 month tuning runs. The fact that this trimming resulted in an increase in accuracy indicates that these addresses most often not related to the survey-reported residential histories.

We used the one month threshold setting as a base for subsequent algorithm tuning experiments.

Figure 4 shows the results of experiments with different values for trimming the time interval range for combined addresses.



		% change T1090:D1mo			% change T2575:D1mo			% change T1075:D1mo		
Vendors	Time-span	Coverage	Within 1 km	Same tract	Coverage	Within 1 km	Same tract	Coverage	Within 1 km	Same tract
V1	Full	-1.82	-1.93	-1.98	-2.99	-4.61	-4.43	-2.27	-2.22	-2.28
V2	Full	-1.00	-0.72	-1.00	-2.52	-1.70	-2.79	-1.86	-0.83	-1.61
V3	Full	-4.12	-0.85	-1.33	-10.07	-5.78	-6.88	-7.87	-1.32	-2.05
V1V2	Full	-4.76	-3.42	-5.26	-10.50	-15.02	-17.64	-8.85	-6.91	-9.34
V1V3	Full	-8.03	-4.56	-9.35	-17.35	-15.14	-19.51	-13.62	-5.74	-10.77
V2V3	Full	-7.58	-9.78	-13.21	-17.93	-19.34	-23.76	-14.13	-12.15	-15.48
V1V2V3	Full	-6.86	-8.26	-13.83	-16.26	-17.72	-22.97	-12.15	-11.19	-15.80
V1	1986_1995	-1.42	-0.10	-0.15	-3.79	-5.74	-5.39	-1.42	-0.10	-0.15
V2	1986_1995	0.00	0.00	0.00	-0.58	-1.08	-1.08	0.00	0.00	0.00
V3	1986_1995	-7.40	-4.78	-4.82	-18.53	-9.43	-8.98	-7.77	-4.40	-4.44
V1V2	1986_1995	-3.04	-4.27	-3.78	-6.53	-14.37	-22.09	-3.04	-2.33	-4.35
V1V3	1986_1995	-8.37	-7.01	-11.87	-18.00	-21.54	-25.26	-8.37	-4.92	-9.61
V2V3	1986_1995	-8.57	-24.33	-25.18	-21.48	-37.25	-42.14	-8.85	-24.10	-24.95
V1V2V3	1986_1995	-7.16	-18.49	-19.74	-16.09	-33.79	-38.99	-7.16	-18.49	-19.74
V1	1986_2005	-0.70	-1.05	-1.07	-1.76	-4.89	-4.77	-0.70	-1.05	-1.07
V2	1986_2005	-0.12	-1.17	-1.07	-0.30	-3.99	-3.65	-0.12	-1.17	-1.07
V3	1986_2005	-2.90	-1.09	-1.66	-7.07	-7.23	-8.51	-3.77	-0.98	-1.55
V1V2	1986_2005	-1.43	-1.03	-3.23	-3.14	-13.27	-16.54	-1.43	-1.27	-4.21
V1V3	1986_2005	-3.75	-2.35	-9.69	-9.25	-13.56	-19.96	-4.68	0.57	-6.83
V2V3	1986_2005	-3.38	-10.87	-14.10	-8.45	-22.25	-27.09	-4.04	-10.91	-14.15
V1V2V3	1986_2005	-3.27	-7.25	-14.22	-8.38	-20.73	-27.11	-4.07	-7.37	-13.51
V1	1986_2015	-1.05	-2.17	-2.15	-2.21	-4.89	-4.73	-1.53	-2.45	-2.45
V2	1986_2015	-0.70	-0.47	-0.77	-1.89	-1.53	-2.28	-1.59	-0.55	-1.36
V3	1986_2015	-4.03	-1.01	-1.50	-10.05	-6.02	-7.13	-7.82	-1.50	-2.24
V1V2	1986_2015	-3.92	-2.98	-4.91	-9.42	-14.87	-17.55	-8.32	-6.54	-9.09
V1V3	1986_2015	-7.25	-4.02	-9.01	-16.21	-15.22	-19.84	-13.27	-5.05	-10.31
V2V3	1986_2015	-6.75	-9.14	-12.90	-16.29	-19.46	-24.04	-13.47	-11.46	-15.12
V1V2V3	1986_2015	-5.70	-7.21	-13.01	-14.16	-18.22	-23.55	-11.39	-10.06	-14.87
V1	1996_2005	-0.20	-1.68	-1.67	-0.35	-4.69	-4.78	-0.20	-1.68	-1.67
V2	1996_2005	-0.17	-1.61	-1.42	-0.17	-5.11	-4.56	-0.17	-1.61	-1.42
V3	1996_2005	-1.08	0.01	-0.84	-2.45	-6.97	-8.91	-2.15	0.05	-0.79
V1V2	1996_2005	-0.30	0.42	-3.27	-0.76	-13.17	-14.48	-0.30	-0.96	-4.44
V1V3	1996_2005	-0.53	-0.70	-9.58	-3.16	-11.05	-19.27	-2.11	2.80	-6.29
V2V3	1996_2005	-0.91	-6.00	-10.51	-2.27	-17.86	-23.22	-1.76	-6.09	-10.61
V1V2V3	1996_2005	-0.54	-2.26	-12.55	-2.96	-15.77	-23.55	-1.91	-2.24	-11.33
V1	1996_2015	-0.92	-2.73	-2.68	-1.65	-4.82	-4.74	-1.57	-3.05	-3.00
V2	1996_2015	-0.85	-0.46	-0.79	-2.16	-1.46	-2.29	-1.93	-0.47	-1.36
V3	1996_2015	-3.21	-0.43	-1.06	-8.01	-5.75	-7.23	-7.83	-0.91	-1.81
V1V2	1996_2015	-4.23	-2.55	-5.07	-10.45	-14.70	-16.11	-10.22	-7.06	-9.60
V1V3	1996_2015	-6.85	-3.38	-8.50	-15.56	-13.84	-18.84	-15.02	-4.52	-9.83
V2V3	1996_2015	-6.31	-6.44	-10.87	-15.03	-16.63	-21.35	-14.59	-8.67	-12.95
V1V2V3	1996_2015	-5.17	-4.61	-11.71	-13.46	-14.66	-20.41	-12.92	-7.33	-13.14
V1	2006_2015	-1.67	-3.51	-3.44	-3.02	-4.69	-4.47	-3.02	-4.00	-3.94
V2	2006_2015	-1.41	0.32	-0.25	-3.80	1.00	-0.46	-3.37	0.74	-0.72
V3	2006_2015	-6.36	-0.41	-0.71	-16.18	-2.44	-3.21	-16.18	-0.56	-1.43
V1V2	2006_2015	-8.36	-4.40	-5.96	-20.64	-14.11	-15.89	-20.64	-11.06	-12.98
V1V3	2006_2015	-13.57	-4.81	-6.30	-28.73	-14.43	-16.02	-28.73	-10.21	-11.54
V2V3	2006_2015	-11.94	-5.62	-10.13	-28.37	-12.05	-16.42	-28.00	-8.09	-12.61
V1V2V3	2006_2015	-10.06	-6.02	-10.18	-24.55	-11.30	-15.07	-24.55	-10.76	-13.37

Legend: <= -5% (loss) | > -5% and < -0.5% | > 0.5% and < 5% | >= 5% (gain)

Figure 4. Using different values for trimming the time interval range

The base configuration of the algorithm uses values of 0% and 100% - the time interval range is not trimmed. We experimented with ranges of 10% to 90% (the first set of results labeled "T1090:D1mo"), 25% to 75% (the second set of results labeled "T2575:D1mo"), and 10% to 75% (the third set of results labeled "T1075:D1mo").

The results were negative in all cases indicating that trimming reduces both the coverage and the accuracy of the algorithm.

Figure 5 shows the results of experiments with an option to trim only groups of addresses.

		% change T09T:D1mo			% change T19F:T1090			% change T09F:T09T		
			Within	Same		Within	Same		Within	Same
Vendors	Time-span	Coverage	1 km	tract	Coverage	1 km	tract	Coverage	1 km	tract
V1	Full	-0.56	-2.37	-2.55	-6.04	-1.73	-3.10	-3.03	0.86	-0.74
V2	Full	-0.56	-0.06	-0.39	-7.56	-5.52	-5.10	-3.87	-2.30	-2.02
V3	Full	-2.56	0.06	-0.09	-3.59	-3.55	-3.38	-1.85	-0.47	-0.43
V1V2	Full	-2.78	-3.23	-4.77	-2.83	-1.00	-0.05	-0.52	0.61	1.45
V1V3	Full	-4.50	-2.23	-5.76	-1.05	-0.18	-0.71	-0.16	0.05	0.05
V2V3	Full	-4.72	-6.38	-8.38	-1.24	0.43	0.31	-0.36	0.67	0.67
V1V2V3	Full	-3.34	-2.40	-6.66	-1.09	0.82	-1.24	-0.25	1.18	-0.39
V1	1986_1995	0.00	-0.72	-1.10	-6.01	-3.75	-5.58	0.00	3.12	1.11
V2	1986_1995	0.00	0.00	0.00	-6.63	-13.58	-13.67	0.00	0.00	0.00
V3	1986_1995	0.00	0.00	0.00	-6.21	-6.01	-6.21	0.00	0.00	0.00
V1V2	1986_1995	0.00	-2.67	-1.16	-4.67	-1.14	-2.53	0.00	2.74	1.18
V1V3	1986_1995	0.00	0.71	-2.65	-0.50	0.84	-1.51	0.00	0.00	0.00
V2V3	1986_1995	0.00	-9.01	-8.17	-2.52	0.44	0.37	0.00	0.00	0.00
V1V2V3	1986_1995	0.00	-1.47	-0.65	-0.81	6.41	0.43	0.00	5.31	1.18
V1	1986_2005	0.00	-1.80	-1.95	-2.35	-0.72	-3.28	0.00	3.52	0.65
V2	1986_2005	0.00	0.00	0.00	-4.78	-7.45	-7.14	-0.20	-0.30	-0.26
V3	1986_2005	-0.77	0.53	0.53	-2.67	-4.91	-4.83	-0.71	-0.28	-0.27
V1V2	1986_2005	0.00	-0.49	-2.21	-1.79	-3.62	-1.90	0.00	-0.99	0.73
V1V3	1986_2005	0.00	1.42	-3.37	-0.36	-0.33	-1.12	0.00	0.00	0.00
V2V3	1986_2005	-0.30	-4.76	-5.77	-1.37	-0.57	-0.79	0.00	0.00	0.00
V1V2V3	1986_2005	0.00	1.60	-2.92	-0.50	0.29	-1.77	0.00	0.77	-0.68
V1	1986_2015	-0.60	-2.57	-2.67	-4.79	-1.07	-2.54	-3.27	1.07	-0.71
V2	1986_2015	-0.58	-0.05	-0.38	-6.77	-6.42	-6.00	-4.01	-2.25	-1.96
V3	1986_2015	-2.59	0.06	-0.09	-3.23	-3.64	-3.47	-1.87	-0.47	-0.43
V1V2	1986_2015	-3.00	-3.14	-4.77	-1.75	-0.81	0.21	-0.56	0.44	1.32
V1V3	1986_2015	-4.84	-2.22	-5.92	-0.41	-0.05	-0.51	-0.17	0.06	0.06
V2V3	1986_2015	-4.84	-5.04	-7.37	-1.25	0.44	0.32	-0.37	0.68	0.68
V1V2V3	1986_2015	-3.60	-2.23	-6.71	-0.61	0.93	-1.09	-0.27	1.03	-0.64
V1	1996_2005	0.00	-2.41	-2.39	0.17	0.42	-2.71	0.00	3.75	0.41
V2	1996_2005	0.00	0.00	0.00	-3.92	-5.27	-5.14	-0.29	-0.40	-0.32
V3	1996_2005	-1.08	0.76	0.76	-1.33	-4.75	-4.62	-0.99	-0.36	-0.33
V1V2	1996_2005	0.00	0.69	-2.74	0.17	-5.31	-2.15	0.00	-2.95	0.51
V1V3	1996_2005	0.00	1.81	-3.72	-0.27	-0.90	-0.97	0.00	0.00	0.00
V2V3	1996_2005	-0.45	-2.97	-4.80	-0.87	-1.02	-1.28	0.00	0.00	0.00
V1V2V3	1996_2005	0.00	3.23	-3.99	-0.30	-2.32	-2.74	0.00	-1.54	-1.59
V1	1996_2015	-0.81	-3.00	-3.00	-4.36	-0.49	-1.92	-4.44	0.81	-0.82
V2	1996_2015	-0.70	0.00	-0.36	-6.80	-5.56	-5.12	-4.84	-2.15	-1.79
V3	1996_2015	-3.21	0.18	0.02	-2.55	-3.35	-3.16	-2.34	-0.50	-0.42
V1V2	1996_2015	-4.08	-2.93	-5.26	-0.69	-1.08	0.47	-0.78	-0.10	1.43
V1V3	1996_2015	-6.58	-2.50	-6.11	-0.38	-0.28	-0.30	-0.23	0.09	0.09
V2V3	1996_2015	-6.02	-3.97	-6.92	-0.95	0.33	0.20	-0.47	0.84	0.84
V1V2V3	1996_2015	-4.89	-2.02	-7.67	-0.54	-0.24	-1.42	-0.37	-0.08	-1.07
V1	2006_2015	-1.67	-3.36	-3.38	-9.18	-0.49	-0.32	-9.18	-1.12	-1.18
V2	2006_2015	-1.29	0.28	-0.30	-9.21	-4.51	-4.05	-8.65	-1.10	-0.87
V3	2006_2015	-6.36	0.02	-0.30	-4.43	-1.17	-0.89	-4.43	-0.32	-0.17
V1V2	2006_2015	-8.36	-5.34	-6.82	-1.66	3.22	3.19	-1.66	2.86	2.48
V1V3	2006_2015	-13.57	-5.53	-7.44	-0.52	0.40	0.40	-0.52	0.23	0.22
V2V3	2006_2015	-11.84	-3.62	-7.78	-1.04	1.61	1.62	-1.02	1.76	1.77
V1V2V3	2006_2015	-10.06	-6.22	-10.61	-0.81	1.91	-0.08	-0.81	1.55	-0.46

Legend: <= -5% (loss) | > -5% and < -0.5% | > 0.5% and < 5% | >= 5% (gain)

Figure 5. Changing the option to trim only groups of addresses

This option to trim only groups of addresses makes a difference only when trimming is active which, as seen in Figure 4, has a negative impact on the performance of the algorithm. The base configuration of the algorithm and the previously described experiments all use this option: they only apply time trimming to groups of addresses and they do not apply time trimming to addresses that are not matched with other addresses. To isolate in impact of changing this option, we first ran a test using minimal trimming of just the ending dates (0% to 90%) with the Trim-Groups-Only option set to True (the first set of results labeled "T09T:D1mo"). As with the results shown in Figure 4, these results show a general loss of performance with this level of trimming. The next two sets of results show the impact of changing the Trim-Groups-Only option from True to False. The second set of results (labeled "T19F:T1090"), uses a trimming range of 10% to 90% and the third set of results (labeled "T09F:T09T") uses a trimming range of 0% to 90%.

These result show that changing the Trim-Groups-Only option from True to False – allowing trimming of time periods for individual addresses as well as for groups of addresses – results in poorer coverage in most cases. There is some evidence for improved accuracy particularly with multiple vendors and minimum trimming. However, these gains in accuracy will at best offset the losses in accuracy due to trimming at the cost of additional loss in coverage.

Figure 6 shows the results of experiments with using vendor weights for time selection.

		% change W1090:T1090			% change W1090:D1mo		
			Within	Same		Within	Same
Vendors	Time-span	Coverage	1 km	tract	Coverage	1 km	tract
V1	Full	0.00	0.00	0.00	-1.82	-1.93	-1.98
V2	Full	0.01	-0.01	0.01	-0.99	-0.73	-0.99
V3	Full	0.01	-0.01	-0.01	-4.12	-0.86	-1.34
V1V2	Full	-0.47	4.06	3.00	-5.20	0.50	-2.4
V1V3	Full	0.37	-0.02	-0.03	-7.69	-4.58	-9.37
V2V3	Full	-2.36	3.27	-1.01	-9.76	-6.83	-14.09
V1V2V3	Full	-0.37	6.68	8.85	-7.21	-2.13	-6.21
V1	1986_1995	0.00	0.00	0.00	-1.42	-0.10	-0.15
V2	1986_1995	0.00	0.00	0.00	0.00	0.00	0.00
V3	1986_1995	0.00	0.00	0.00	-7.40	-4.78	-4.82
V1V2	1986_1995	-0.40	9.97	4.26	-3.43	5.27	0.32
V1V3	1986_1995	0.07	-0.17	-0.02	-8.31	-7.17	-11.88
V2V3	1986_1995	0.50	-0.31	-0.31	-8.11	-24.57	-25.41
V1V2V3	1986_1995	-0.58	14.15	12.95	-7.69	-6.95	-9.34
V1	1986_2005	0.00	0.00	0.00	-0.70	-1.05	-1.07
V2	1986_2005	0.00	0.00	0.00	-0.12	-1.17	-1.07
V3	1986_2005	0.00	0.00	0.00	-2.90	-1.09	-1.66
V1V2	1986_2005	-0.23	4.78	3.10	-1.66	3.70	-0.23
V1V3	1986_2005	0.03	-0.11	-0.06	-3.72	-2.46	-9.74
V2V3	1986_2005	0.02	2.21	-0.76	-3.35	-8.91	-14.76
V1V2V3	1986_2005	-0.22	6.66	9.03	-3.48	-1.07	-6.48
V1	1986_2015	0.00	0.00	0.00	-1.05	-2.17	-2.15
V2	1986_2015	0.01	-0.01	0.01	-0.69	-0.48	-0.77
V3	1986_2015	0.01	-0.01	-0.01	-4.02	-1.02	-1.51
V1V2	1986_2015	-0.36	4.15	3.04	-4.26	1.05	-2.02
V1V3	1986_2015	0.36	0.00	0.00	-6.92	-4.01	-9.01
V2V3	1986_2015	-2.33	3.27	-1.04	-8.92	-6.16	-13.80
V1V2V3	1986_2015	-0.27	6.07	8.26	-5.95	-1.58	-5.82
V1	1996_2005	0.00	0.00	0.00	-0.20	-1.68	-1.67
V2	1996_2005	0.00	0.00	0.00	-0.17	-1.61	-1.42
V3	1996_2005	0.00	0.00	0.00	-1.08	0.01	-0.84
V1V2	1996_2005	-0.12	2.14	2.50	-0.42	2.57	-0.84
V1V3	1996_2005	0.00	-0.07	-0.07	-0.53	-0.77	-9.65
V2V3	1996_2005	-0.18	3.05	-0.83	-1.09	-3.12	-11.26
V1V2V3	1996_2005	0.02	3.48	7.35	-0.53	1.13	-6.12
V1	1996_2015	0.00	0.00	0.00	-0.92	-2.73	-2.68
V2	1996_2015	0.01	-0.01	0.01	-0.84	-0.47	-0.78
V3	1996_2015	0.01	-0.01	-0.01	-3.20	-0.44	-1.07
V1V2	1996_2015	-0.35	2.71	2.75	-4.56	0.08	-2.46
V1V3	1996_2015	0.46	0.01	-0.03	-6.43	-3.37	-8.53
V2V3	1996_2015	-3.00	4.09	-0.88	-9.12	-2.62	-11.66
V1V2V3	1996_2015	-0.16	4.34	7.27	-5.33	-0.47	-5.28
V1	2006_2015	0.00	0.00	0.00	-1.67	-3.51	-3.44
V2	2006_2015	0.02	-0.02	0.00	-1.39	0.30	-0.24
V3	2006_2015	0.03	-0.03	-0.03	-6.33	-0.44	-0.74
V1V2	2006_2015	-0.61	3.30	3.03	-8.91	-1.25	-3.12
V1V3	2006_2015	1.02	0.02	-0.09	-12.69	-4.79	-6.38
V2V3	2006_2015	-6.32	5.88	-0.26	-17.50	-0.08	-10.37
V1V2V3	2006_2015	-0.37	5.24	7.23	-10.39	-1.09	-3.69

Legend: <= -5% (loss) | > -5% and < -0.5% | > 0.5% and < 5% | >= 5% (gain)

Figure 6. Using vendor weights for time selection

Weighting only makes a difference if trimming is active. For these experiments, we used weights for each vendor based on the inverse of the median starting date accuracy measures shown in Table 13 above (0.4 for Vendor 1, 2.3 for Vendor 2, and 0.6 for Vendor 3 data). The first set of results labeled "W1090:T1090" uses these vendor weights with trimming values of 10% to 90%. The second set of results labeled "W1090:D1mo" compares running the algorithm with vendor weights and 10% to 90% trimming with not doing either.

The first set of results show an improvement in accuracy with very little loss of coverage by using the vendor weights for the residential histories generated using data from more than one vendor. However, the net impact of using both the vendor weights and trimming shown in the second set of results indicated that the negative impact of the trimming of time frames remains even with the use of vendor weights in most cases.

Overall, the results of these algorithm tuning experiments indicate that, for the combinations tested, the optimal setting of parameter values is to use the one month minimum time interval but none of the other options.

3.7 Assessment of Derived Residential Histories

This section provides the results of a final set of comparisons using the tuned algorithm – for these comparisons we used just the one month minimum time interval option and none of the others. We compared the completeness and accuracy of the derived residential histories for each of the 49 combinations of vendors and time period with the survey-reported residential histories after the updates made as a result of the reconciliation. In addition we included the artificial residential history developed by assuming each individ ual lived at their current residence for their entire lives. In the following figures, this scenario is identified as the "Cur Res" residential history.

Each of the following figures includes a completeness measure: the proportion of survey-reported time with location information that we also have vendor location information. We refer to this measure as the "percent time period coverage". In Figure 7, we report descriptive statistics for the time-weighted distance between survey-reported and vendor-reported locations.

Full life span:

Vendors	Percent time	Time-weighted distance error (km)		
used	period coverage	Mean	Median	75th percentile
V1	58.7	109.9	0.0	5.7
V2	35.4	213.9	0.0	9.4
V3	35.5	63.5	0.0	0.6
V1,V2	58.9	140.0	0.0	8.8
V1,V3	58.8	124.5	0.0	6.3
V2,V3	46.9	110.6	0.0	3.1
V1,V2,V3	59.1	139.9	0.0	7.0
Cur Res	NA	696.1	50.1	955.1

1986 to 2015:

Vendors	Percent time	Time-weighted distance error (km)		
used	period coverage	Mean	Median	75th percentile
V1	89.7	109.4	0.0	4.4
V2	56.3	218.3	0.0	9.4
V3	57.6	64.2	0.0	0.6
V1,V2	90.0	141.8	0.0	6.9
V1,V3	89.9	124.6	0.0	5.3
V2,V3	75.2	111.9	0.0	2.9
V1,V2,V3	90.3	141.5	0.0	6.0
Cur Res	NA	477.5	8.2	374.2

1996 to 2015:

Vendors	Percent time	Time-weighted distance error (km)		
used	period coverage	Mean	Median	75th percentile
V1	96.8	81.1	0.0	0.7
V2	68.2	227.6	0.0	5.3
V3	68.0	65.6	0.0	0.1
V1,V2	96.9	116.5	0.0	2.8
V1,V3	96.9	100.4	0.0	1.4
V2,V3	88.6	111.6	0.0	1.4
V1,V2,V3	97.2	116.4	0.0	1.5
Cur Res	NA	343.9	0.1	35.1

2006 to 2015:

Vendors	Percent time	Time-weighted distance error (km)		
used	period coverage	Mean	Median	75th percentile
V1	98.9	52.8	0.0	0.0
V2	78.2	93.4	0.0	0.0
V3	57.6	58.7	0.0	0.0
V1,V2	98.9	133.4	0.0	0.6
V1,V3	98.4	67.1	0.0	0.0
V2,V3	90.7	63.2	0.0	0.1
V1,V2,V3	98.9	133.4	0.0	0.1
Cur Res	NA	128.7	0.0	4.1

Figure 7. Comparison results – time-weighted distance error

1986 to 2005:

Vendors	Percent time	Time-weighted distance error (km)		
used	period coverage	Mean	Median	75th percentile
V1	85.2	141.1	0.0	8.8
V2	45.6	322.0	1.4	33.0
V3	57.6	66.9	0.0	1.9
V1,V2	85.6	146.5	0.0	11.6
V1,V3	85.8	156.5	0.0	8.8
V2,V3	67.7	143.6	0.0	6.9
V1,V2,V3	86.1	146.0	0.0	8.8
Cur Res	NA	646.4	18.9	625.4

1996 to 2005:

Vendors	Percent time	Time-weighted distance error (km)		
used	period coverage	Mean	Median	75th percentile
V1	94.9	108.1	0.0	5.6
V2	59.1	390.2	1.4	23.4
V3	77.5	70.3	0.0	1.4
V1,V2	95.1	100.5	0.0	7.0
V1,V3	95.6	131.8	0.0	5.3
V2,V3	86.7	157.9	0.0	5.3
V1,V2,V3	95.6	100.3	0.0	5.3
Cur Res	NA	540.9	8.8	432.2

1986 to 1995:

Vendors	Percent time	Time-weighted distance error (km)		
used	period coverage	Mean	Median	75th percentile
V1	74.3	188.5	0.1	51.3
V2	30.5	173.5	2.9	55.3
V3	35.2	58.7	0.0	4.9
V1,V2	75.0	212.1	1.3	69.7
V1,V3	74.8	192.0	0.4	51.3
V2,V3	46.2	113.4	0.0	8.8
V1,V2,V3	75.5	211.0	1.1	64.5
Cur Res	NA	765.0	224.7	947.0

Figure 7 (continued). Comparison results – time-weighted distance error

These results show that coverage statistics are highest for histories derived from the Vendor 1 data (59% for the full life span). Coverage does not improve by combining data from other vendors with the Vendor 1 data. Coverage does not improve when data from the other two vendors are added. Coverage for the original Vendor 3 data is the lowest at 14% for the full life span but coverage for the revised is similar to that of Vendor 2. Coverage for Vendors 2 and 3 together is generally better than each of them independently.

The time-weighted distance measure is highly skewed and the median is almost always zero for the vendor based histories. The mean distance error is lowest for Vendor 3 but the coverage is not as good and Vendor 1. The mean distance error for Vendor 1 by itself is generally lower than it is when data from the other two vendors are added. The distance error based on the current residence is generally much larger, particularly for the older time periods.

In Figure 8, we report the proportion of time that the distance between locations in the vendor and survey-reported histories is zero, less than 100 meters, less than 500 meters, less than 1 kilometer, less than 5 kilometers, and less than 10 kilometers. Coverage results are the same as in Figure 7 – they are repeated here for easy reference.

Full life span:

Vendors used	Percent time period coverage	Pct covered time distance=0	Pct covered time within 100 m	Pct covered time within 500 m	Pct covered time within 1 km	Pct covered time within 5 km	Pct covered time within 10 km
V1	58.7	56.0	64.6	66.4	68.4	74.0	81.3
V2	35.4	49.1	54.4	55.7	56.7	70.0	76.3
V3	35.5	31.6	73.5	74.5	77.0	84.0	90.3
V1,V2	58.9	50.0	58.4	60.8	64.3	71.0	78.5
V1,V3	58.8	25.2	63.1	65.0	65.9	73.1	80.3
V2,V3	46.9	26.2	65.0	66.6	68.1	78.1	84.0
V1,V2,V3	59.1	28.4	60.3	62.9	64.2	72.4	80.1
Cur Res	NA	23.9	24.1	25.4	27.6	32.5	37.4

1986 to 2015:

Vendors used	Percent time period coverage	Pct covered time distance=0	Pct covered time within 100 m	Pct covered time within 500 m	Pct covered time within 1 km	Pct covered time within 5 km	Pct covered time within 10 km
V1	89.7	57.5	66.1	67.7	69.9	75.8	82.0
V2	56.3	50.1	55.5	56.9	57.9	70.3	76.4
V3	57.6	32.1	73.5	74.4	77.0	84.1	90.3
V1,V2	90.0	51.1	59.5	61.8	65.5	72.6	79.0
V1,V3	89.9	25.2	64.7	66.6	67.6	74.9	81.1
V2,V3	75.2	26.3	65.4	67.0	68.3	78.5	84.2
V1,V2,V3	90.3	28.6	61.7	64.4	65.7	74.2	80.8
Cur Res	NA	36.9	37.2	39.2	40.3	47.6	53.2

1996 to 2015:

Vendors used	Percent time period coverage	Pct covered time distance=0	Pct covered time within 100 m	Pct covered time within 500 m	Pct covered time within 1 km	Pct covered time within 5 km	Pct covered time within 10 km
V1	96.8	63.2	72.5	73.0	75.4	80.6	86.2
V2	68.2	54.5	60.1	61.6	62.6	74.0	79.6
V3	68.0	33.1	76.0	76.9	79.5	86.2	91.1
V1,V2	96.9	56.7	65.5	67.0	71.4	77.6	83.3
V1,V3	96.9	27.2	71.5	72.2	73.1	79.9	85.9
V2,V3	88.6	27.4	68.6	70.4	71.4	81.1	85.9
V1,V2,V3	97.2	32.9	68.7	70.4	71.5	79.1	85.3
Cur Res	NA	49.0	49.4	51.4	52.0	59.5	64.8

2006 to 2015:

Vendors used	Percent time period coverage	Pct covered time distance=0	Pct covered time within 100 m	Pct covered time within 500 m	Pct covered time within 1 km	Pct covered time within 5 km	Pct covered time within 10 km
V1	98.9	72.6	79.8	80.2	83.2	87.8	90.1
V2	78.2	69.6	75.5	77.4	79.2	87.5	89.3
V3	57.6	34.6	85.2	85.8	88.6	93.0	95.0
V1,V2	98.9	66.1	72.9	74.8	79.5	84.9	86.7
V1,V3	98.4	26.1	78.7	79.1	80.2	85.4	88.8
V2,V3	90.7	27.6	77.1	79.1	79.8	88.0	89.2
V1,V2,V3	98.9	34.6	75.2	77.1	78.3	83.9	86.7
Cur Res	NA	68.1	68.8	70.5	71.2	76.2	78.2

Figure 8. Comparison results – within distance thresholds

1986 to 2005:

Vendors used	Percent time period coverage	Pct covered time distance=0	Pct covered time within 100 m	Pct covered time within 500 m	Pct covered time within 1 km	Pct covered time within 5 km	Pct covered time within 10 km
V1	85.2	48.9	58.4	60.7	62.5	69.0	77.4
V2	45.6	33.9	38.9	39.8	40.2	56.0	65.6
V3	57.6	30.8	67.8	69.0	71.3	79.8	88.1
V1,V2	85.6	42.7	51.9	54.6	57.7	65.7	74.7
V1,V3	85.8	24.7	56.9	59.6	60.5	69.1	76.8
V2,V3	67.7	25.4	57.7	59.2	60.9	72.3	81.0
V1,V2,V3	86.1	25.3	54.1	57.3	58.7	68.8	77.5
Cur Res	NA	21.8	21.8	24.1	25.3	33.7	41.1

1996 to 2005:

Vendors used	Percent time period coverage	Pct covered time distance=0	Pct covered time within 100 m	Pct covered time within 500 m	Pct covered time within 1 km	Pct covered time within 5 km	Pct covered time within 10 km
V1	94.9	54.2	65.4	66.0	67.9	73.8	82.4
V2	59.1	36.3	41.3	42.4	42.4	57.7	67.8
V3	77.5	32.0	69.8	70.9	73.3	81.5	88.5
V1,V2	95.1	47.8	58.3	59.6	63.7	70.7	80.2
V1,V3	95.6	28.1	64.8	65.8	66.5	74.7	83.1
V2,V3	86.7	27.3	60.4	62.1	63.4	74.6	82.8
V1,V2,V3	95.6	31.3	62.5	64.1	65.2	74.7	84.0
Cur Res	NA	31.6	31.6	34.0	34.4	44.3	52.4

1986 to 1995:

Vendors used	Percent time period coverage	Pct covered time distance=0	Pct covered time within 100 m	Pct covered time within 500 m	Pct covered time within 1 km	Pct covered time within 5 km	Pct covered time within 10 km
V1	74.3	41.3	48.4	53.0	54.7	62.2	70.3
V2	30.5	28.7	33.6	34.1	35.2	52.1	61.0
V3	35.2	27.9	62.7	64.0	66.5	75.7	87.1
V1,V2	75.0	35.5	42.8	47.4	49.2	58.6	67.0
V1,V3	74.8	19.8	45.5	50.8	52.1	61.0	67.8
V2,V3	46.2	21.5	52.1	53.1	55.5	67.5	77.3
V1,V2,V3	75.5	16.8	42.3	47.6	49.5	60.3	68.3
Cur Res	NA	10.8	10.8	13.0	15.1	21.9	28.3

Figure 8 (continued). Comparison results – within distance thresholds

Care should be taken in interpreting these results. Except for distance error of zero, Vendor 3 data generally has the highest percent of time the distance thresholds. However, the coverage for the Vendor 3 data is lower than that of Vendor 1.

Of the vendor combinations with the best coverage, Vendor 1 used independently has the highest values for the accuracy measures. Although the accuracy is not as high as might be desired, it is usually substantially better than the accuracy using the current residence assumption, particularly for older time periods. For newer time periods, the vendor based accuracy improves but there is less of a difference between it and the current residence assumption.

In Figure 9, we report measures based on geographic areas: the proportion of time that there is a difference in the census tracts, ZIP codes, or counties of the survey-reported and vendor-reported locations.

Full life span:

Vendors used	Percent time period coverage	Percent of covered time in same tract	Percent of covered time in same ZIP	Percent of covered time in same county
V1	58.7	67.6	69.3	81.7
V2	35.4	59.5	64.6	79.8
V3	35.5	78.5	80.2	90.7
V1,V2	58.9	63.7	66.0	80.0
V1,V3	58.8	66.0	69.6	80.9
V2,V3	46.9	69.5	75.0	85.5
V1,V2,V3	59.1	64.4	68.6	81.2
Cur Res	NA	25.7	32.4	41.9

1986 to 2015:

Vendors used	Percent time period coverage	Percent of covered time in same tract	Percent of covered time in same ZIP	Percent of covered time in same county
V1	89.7	69.2	70.6	82.4
V2	56.3	60.8	65.9	80.0
V3	57.6	78.5	80.2	90.6
V1,V2	90.0	65.0	67.2	80.4
V1,V3	89.9	67.7	71.0	81.6
V2,V3	75.2	70.1	75.7	85.6
V1,V2,V3	90.3	66.0	70.0	81.8
Cur Res	NA	39.9	46.7	58.1

1996 to 2015:

Vendors used	Percent time period coverage	Percent of covered time in same tract	Percent of covered time in same ZIP	Percent of covered time in same county
V1	96.8	75.4	76.4	86.6
V2	68.2	66.1	71.0	83.5
V3	68.0	81.6	82.6	90.8
V1,V2	96.9	71.7	73.1	85.0
V1,V3	96.9	74.7	77.6	86.1
V2,V3	88.6	73.9	79.1	87.1
V1,V2,V3	97.2	73.3	76.5	86.5
Cur Res	NA	52.6	57.9	69.0

2006 to 2015:

Vendors used	Percent time period coverage	Percent of covered time in same tract	Percent of covered time in same ZIP	Percent of covered time in same county
V1	98.9	82.7	83.1	91.1
V2	78.2	81.5	84.7	90.8
V3	57.6	91.7	90.4	93.5
V1,V2	98.9	79.1	80.2	88.0
V1,V3	98.4	81.1	83.5	89.8
V2,V3	90.7	81.7	87.2	89.7
V1,V2,V3	98.9	79.1	82.0	88.0
Cur Res	NA	71.2	74.6	82.8

Figure 9. Comparison results – in the same geographic areas

1986 to 2005:

Vendors used	Percent time period coverage	Percent of covered time in same tract	Percent of covered time in same ZIP	Percent of covered time in same county
V1	85.2	61.6	63.6	77.5
V2	45.6	43.6	50.3	71.0
V3	57.6	72.1	75.3	89.1
V1,V2	85.6	57.1	59.9	76.2
V1,V3	85.8	60.3	64.1	77.0
V2,V3	67.7	62.5	68.2	82.9
V1,V2,V3	86.1	58.8	63.3	78.3
Cur Res	NA	24.7	33.2	46.2

1996 to 2005:

Vendors used	Percent time period coverage	Percent of covered time in same tract	Percent of covered time in same ZIP	Percent of covered time in same county
V1	94.9	68.4	70.0	82.3
V2	59.1	47.5	54.5	74.7
V3	77.5	74.8	77.2	88.9
V1,V2	95.1	64.6	66.3	82.2
V1,V3	95.6	68.7	72.1	82.6
V2,V3	86.7	66.3	71.5	84.6
V1,V2,V3	95.6	67.8	71.2	85.0
Cur Res	NA	35.6	42.7	56.3

1986 to 1995:

Vendors used	Percent time period coverage	Percent of covered time in same tract	Percent of covered time in same ZIP	Percent of covered time in same county
V1	74.3	51.7	54.4	70.6
V2	30.5	35.3	41.3	63.0
V3	35.2	65.5	70.5	89.7
V1,V2	75.0	46.3	50.9	67.7
V1,V3	74.8	48.2	52.6	68.9
V2,V3	46.2	54.4	61.5	79.5
V1,V2,V3	75.5	45.8	52.0	68.7
Cur Res	NA	12.5	22.5	34.8

Figure 9 (continued). Comparison results – in the same geographic areas

These results are very similar to those shown in Figure 8. Of the vendor combinations with the best coverage, Vendor 1 used independently has the highest values for the accuracy measures. The accuracy improved for larger geographic areas as is expected but the differential improvement over the current residence assumption is smaller (although still substantial for the older time periods).

To better assess the tradeoff between coverage and accuracy, we calculated the proportion of the total time period with accurate data. This value is simply the product of the coverage proportion and the accuracy.

In Figure 10, we report the proportion of the total time period that the distance between locations in the vendor and survey-reported histories is zero, less than 100 meters, less than 500 meters, less than 1 kilometer, less than 5 kilometers, and less than 10 kilometers.

Full life span:

Vendors used	Percent of time distance=0	Percent of time within 100 m	Percent of time within 500 m	Percent of time within 1 km	Percent of time within 5 km	Percent of time within 10 km
V1	32.8	37.9	39.0	40.2	43.4	47.7
V2	17.4	19.3	19.7	20.1	24.8	27.0
V3	11.2	26.1	26.4	27.3	29.8	32.0
V1,V2	29.5	34.4	35.9	37.9	41.9	46.3
V1,V3	14.8	37.1	38.2	38.8	43.0	47.3
V2,V3	12.3	30.5	31.3	31.9	36.6	39.4
V1,V2,V3	16.8	35.7	37.2	38.0	42.8	47.4
Cur Res	23.9	24.1	25.4	27.6	32.5	37.4

1986 to 2015:

Vendors used	Percent of time distance=0	Percent of time within 100 m	Percent of time within 500 m	Percent of time within 1 km	Percent of time within 5 km	Percent of time within 10 km
V1	51.5	59.3	60.7	62.7	67.9	73.5
V2	28.2	31.2	32.0	32.6	39.5	43.0
V3	18.5	42.3	42.9	44.3	48.4	52.0
V1,V2	46.0	53.5	55.6	59.0	65.3	71.1
V1,V3	22.7	58.1	59.8	60.7	67.3	72.9
V2,V3	19.8	49.1	50.4	51.4	59.0	63.3
V1,V2,V3	25.9	55.7	58.1	59.3	67.0	72.9
Cur Res	36.9	37.2	39.2	40.3	47.6	53.2

1996 to 2015:

Vendors used	Percent of time distance=0	Percent of time within 100 m	Percent of time within 500 m	Percent of time within 1 km	Percent of time within 5 km	Percent of time within 10 km
V1	61.2	70.1	70.6	72.9	78.0	83.4
V2	37.2	41.0	42.0	42.7	50.5	54.3
V3	22.5	51.7	52.3	54.0	58.6	61.9
V1,V2	54.9	63.4	65.0	69.2	75.2	80.8
V1,V3	26.3	69.3	70.0	70.9	77.4	83.2
V2,V3	24.3	60.8	62.4	63.3	71.9	76.1
V1,V2,V3	32.0	66.7	68.4	69.5	76.9	82.9
Cur Res	49.0	49.4	51.4	52.0	59.5	64.8

2006 to 2015:

Vendors used	Percent of time distance=0	Percent of time within 100 m	Percent of time within 500 m	Percent of time within 1 km	Percent of time within 5 km	Percent of time within 10 km
V1	71.8	79.0	79.3	82.2	86.9	89.2
V2	54.4	59.1	60.6	62.0	68.4	69.9
V3	19.9	49.1	49.4	51.1	53.6	54.7
V1,V2	65.3	72.1	74.0	78.6	84.0	85.7
V1,V3	25.7	77.4	77.8	78.9	84.0	87.3
V2,V3	25.0	69.9	71.7	72.4	79.8	80.9
V1,V2,V3	34.3	74.4	76.3	77.4	82.9	85.7
Cur Res	68.1	68.8	70.5	71.2	76.2	78.2

Figure 10. Proportion of total time – within distance thresholds

1986 to 2005:

Vendors used	Percent of time distance=0	Percent of time within 100 m	Percent of time within 500 m	Percent of time within 1 km	Percent of time within 5 km	Percent of time within 10 km
V1	41.7	49.8	51.7	53.3	58.8	66.0
V2	15.5	17.8	18.1	18.3	25.5	29.9
V3	17.7	39.0	39.7	41.1	46.0	50.7
V1,V2	36.6	44.5	46.7	49.4	56.3	64.0
V1,V3	21.2	48.8	51.1	51.9	59.2	65.9
V2,V3	17.2	39.1	40.1	41.2	48.9	54.8
V1,V2,V3	21.8	46.6	49.3	50.6	59.2	66.8
Cur Res	21.8	21.8	24.1	25.3	33.7	41.1

1996 to 2005:

Vendors used	Percent of time distance=0	Percent of time within 100 m	Percent of time within 500 m	Percent of time within 1 km	Percent of time within 5 km	Percent of time within 10 km
V1	51.4	62.1	62.7	64.4	70.0	78.2
V2	21.4	24.4	25.0	25.1	34.1	40.0
V3	24.8	54.1	55.0	56.8	63.2	68.6
V1,V2	45.4	55.5	56.7	60.5	67.2	76.2
V1,V3	26.9	61.9	62.9	63.5	71.4	79.4
V2,V3	23.7	52.4	53.9	55.0	64.7	71.8
V1,V2,V3	29.9	59.7	61.2	62.3	71.4	80.2
Cur Res	31.6	31.6	34.0	34.4	44.3	52.4

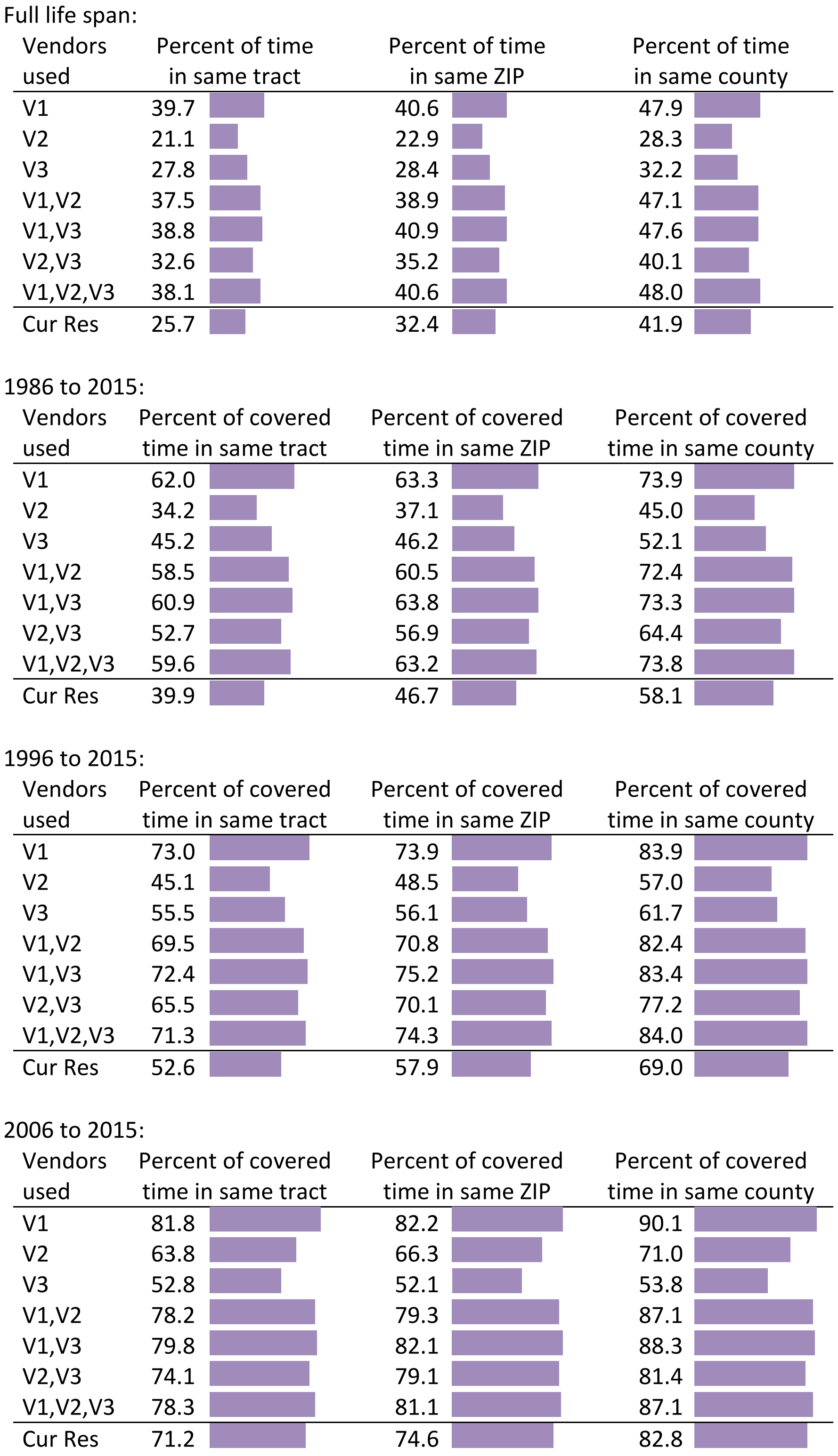
1986 to 1995:

Vendors used	Percent of time distance=0	Percent of time within 100 m	Percent of time within 500 m	Percent of time within 1 km	Percent of time within 5 km	Percent of time within 10 km
V1	30.7	36.0	39.4	40.7	46.2	52.2
V2	8.7	10.3	10.4	10.7	15.9	18.6
V3	9.8	22.1	22.5	23.4	26.6	30.6
V1,V2	26.7	32.1	35.5	36.9	44.0	50.3
V1,V3	14.8	34.0	38.0	38.9	45.6	50.7
V2,V3	9.9	24.1	24.6	25.7	31.2	35.8
V1,V2,V3	12.7	31.9	35.9	37.4	45.6	51.6
Cur Res	10.8	10.8	13.0	15.1	21.9	28.3

Figure 10 (continued). Proportion of total time – within distance thresholds

These results show that the residential histories generated from Vendor 1 by itself generally have the best combination of coverage and accuracy and are better than the current residence assumption in all cases. There are some cases where all three vendors combined are marginally more accurate than Vendor 1 alone but the differences are small. For example, in the time period from 1996 to 2005, the percent of time that the locations are within 10 kilometers is 78.2% for Vendor 1 alone and 80.2% for Vendors 1, 2, and 3 combined. For the smaller distance thresholds, Vendor 1 alone is always better than all three vendors combined.

In Figure 11, we report the proportion of the total time period that there is a difference in the census tracts, ZIP codes, or counties of the survey-reported and vendor-reported locations.



Full life span:

Vendors used	Percent of time in same tract	Percent of time in same ZIP	Percent of time in same county
V1	39.7	40.6	47.9
V2	21.1	22.9	28.3
V3	27.8	28.4	32.2
V1,V2	37.5	38.9	47.1
V1,V3	38.8	40.9	47.6
V2,V3	32.6	35.2	40.1
V1,V2,V3	38.1	40.6	48.0
Cur Res	25.7	32.4	41.9

1986 to 2015:

Vendors used	Percent of covered time in same tract	Percent of covered time in same ZIP	Percent of covered time in same county
V1	62.0	63.3	73.9
V2	34.2	37.1	45.0
V3	45.2	46.2	52.1
V1,V2	58.5	60.5	72.4
V1,V3	60.9	63.8	73.3
V2,V3	52.7	56.9	64.4
V1,V2,V3	59.6	63.2	73.8
Cur Res	39.9	46.7	58.1

1996 to 2015:

Vendors used	Percent of covered time in same tract	Percent of covered time in same ZIP	Percent of covered time in same county
V1	73.0	73.9	83.9
V2	45.1	48.5	57.0
V3	55.5	56.1	61.7
V1,V2	69.5	70.8	82.4
V1,V3	72.4	75.2	83.4
V2,V3	65.5	70.1	77.2
V1,V2,V3	71.3	74.3	84.0
Cur Res	52.6	57.9	69.0

2006 to 2015:

Vendors used	Percent of covered time in same tract	Percent of covered time in same ZIP	Percent of covered time in same county
V1	81.8	82.2	90.1
V2	63.8	66.3	71.0
V3	52.8	52.1	53.8
V1,V2	78.2	79.3	87.1
V1,V3	79.8	82.1	88.3
V2,V3	74.1	79.1	81.4
V1,V2,V3	78.3	81.1	87.1
Cur Res	71.2	74.6	82.8

Figure 11. Proportion of total time – in the same geographic areas

1986 to 2005:

Vendors used	Percent of covered time in same tract	Percent of covered time in same ZIP	Percent of covered time in same county
V1	52.4	54.1	66.0
V2	19.9	23.0	32.4
V3	41.5	43.3	51.3
V1,V2	48.9	51.3	65.2
V1,V3	51.7	55.0	66.0
V2,V3	42.3	46.2	56.1
V1,V2,V3	50.6	54.5	67.4
Cur Res	24.7	33.2	46.2

1996 to 2005:

Vendors used	Percent of covered time in same tract	Percent of covered time in same ZIP	Percent of covered time in same county
V1	64.9	66.4	78.1
V2	28.0	32.2	44.2
V3	58.0	59.8	68.9
V1,V2	61.4	63.0	78.1
V1,V3	65.6	68.9	79.0
V2,V3	57.5	62.0	73.3
V1,V2,V3	64.8	68.1	81.2
Cur Res	35.6	42.7	56.3

1986 to 1995:

Vendors used	Percent of covered time in same tract	Percent of covered time in same ZIP	Percent of covered time in same county
V1	38.4	40.4	52.5
V2	10.8	12.6	19.2
V3	23.0	24.8	31.6
V1,V2	34.8	38.2	50.8
V1,V3	36.1	39.3	51.5
V2,V3	25.1	28.4	36.7
V1,V2,V3	34.6	39.3	51.9
Cur Res	12.5	22.5	34.8

Figure 11 (continued). Proportion of total time – in the same geographic areas

These results are similar to those in Figure 10. They show that the residential histories generated from Vendor 1 by itself generally have the best combination of coverage and accuracy and are better than the current residence assumption in all cases. There are some cases where all three vendors combined are marginally more accurate than Vendor 1 alone but the differences are relatively small.

The overall conclusion from the figures presented in this section is that the best combination of completeness and accuracy is obtained using Vendor 1 data by itself. Although the completeness and accuracy are not as high as might be desired, the derived residential histories are more accurate than using a current residence assumption, particularly for the older time periods.

4. Conclusions and Discussion

Based on this pilot study, we can make a number of conclusions about the availability of residential history data from commercial vendors and the accuracy and completeness of these data. The data that commercial vendors provide consist of a set of addresses that are associated with each individual rather than a residential history *per se* for the individual (i.e., the person lived at location A from time 1 to time 2, location B from time 2 to time 3, etc.) The data includes many addresses not part of survey-reported residential histories – some are work addresses, some are addresses of family members or others. Also, the time frames in the vendor data are frequently missing or incorrect and vendor data often include multiple addresses for a single point in time. For this study, we developed an algorithm to convert the vendor-supplied set of residential addresses into a logical residential history.

In general, the commercial data start around 1980 – there is very little data available before then. All three vendors had data on deceased relatives with Vendors 1 and 3 having more complete data. Vendors reported only U.S. addresses. These did include some military APO addresses that could be used to identify postings at overseas military bases.

All three vendors were able to accurately match the individuals in the study – there was no evidence of false positive matches. Comparing the addresses appearing in the vendor data with the addresses in the survey-reported data, all vendors had reasonable address-match rates and these rates were similar to those reported in previous studies. The process of reconciling differences between vendor-reported addresses and survey-reported addresses resulted in a substantial number of corrections to the survey-reported histories. For studies that have survey-reported residential histories, the commercial data could be a source of valuable additional information.

After using our algorithm to convert the vendor-supplied sets of residential addresses into plausible residential histories, we compared the accuracy and completeness of the derived vendor histories with the survey-reported histories. We conclude that reasonable residential histories can be derived from vendor data. The derived histories yield significant accuracy improvements compared to assuming the person always lived at their current residence although the vendor-derived histories are not as accurate as survey-reported histories. The residential histories derived from Vendor 1 data offer the best combination of completeness and accuracy. Combining data from Vendors 2 and/or 3 with Vendor 1 does not seem to improve results. Vendor derived histories are more accurate for more recent time periods but there is also less of an improvement over the current-residence assumption for these periods.

This pilot study is limited in terms of sample size: a large sample might yield some different conclusions. However, the conclusions about vendor data availability are not likely to change. The study subjects consisted of volunteers from NCI and NIEHS and were not necessarily representative of the general population or of the population of people diagnosed with cancer. In particular, volunteers were health research scientists and professionals: college educated, most with advanced degrees and generally middle class or upper-middle class. There was a limited range of ages: none of the volunteers were very young or very old. The volunteers were living in the Washington DC area or North Carolina at the time of the study so a majority of the residential addresses from the Eastern states (although there were addresses from all parts of the U.S. and many foreign addresses). There were a relatively high percentage of foreign-born individuals.

Because the goal of this study was to better understand the availability and quality of commercial data rather than to draw statistical inferences that could be generalized to the U.S. population as a whole, the non-representativeness of the sample may not be a major limitation. Given that much of the commercial data are derived from various financial transactions (credit applications, home purchases, etc.), the major drawback of our sample is that we do not really know much about the availability of data for the very poor – people without credit cards and bank accounts. We also don't know much about the availability of data for children, teens, and young adults. Finally, residential addresses for older people where someone else has power-of-attorney may reflect the residence of the caregiver rather than the address of the individual.

The algorithm we developed to convert the vendor-supplied sets of residential addresses into plausible residential histories may not be optimal. The algorithm assigns the most likely time frame to each address. An alternative approach to determine the most likely address for each time period was not explored in this study. When combining matched addresses, weights based on measured address accuracy for each of the vendors could be applied. It might be possible to weed out business addresses using a USPS Residential Delivery Indicator service. Finally, a heuristic or adaptive algorithm could make better use of the additional information provided by multiple vendors.

One planned enhancement to the algorithm is to use known previous addresses as additional input. When using the algorithm to generate residential histories for cancer patients, cancer registries will usually have an address for the patient at the time of their cancer diagnosis as well as a current address. In some cases, registries may also have a place of birth (either a state in the U.S. or a foreign country). This additional address information could be used by the algorithm to improve the accuracy of the derived residential history.

In summary, this study shows that commercial residential address data can be used to develop plausible individual residential histories. At this time, Vendor 1 seems to have the most accurate and complete residential address data and combining data from multiple vendors does not seem to help. The commercial data start in the 1980s, contain only U.S. addresses, and include data for deceased individuals. Commercial residential address data may also have value to improve the quality of survey-reported residential histories.

References

Boscoe FP. The Use of Residential History in Environmental Health Studies. In *Geospatial Analysis of Environmental Health*. JA Maantay and SL McLafferty, eds. Berlin: Springer Verlag, 2011. 93-110.

Jacquez, GM, et al. Accuracy of Commercially Available Residential Histories for Epidemiologic Studies. *Am. J. Epidemiol.* (2011) 173 (2): 236-243.

Meliker JR, Slotnick MJ, AvRuskin GA, Schottenfeld D, Jacquez GM, Wilson ML, Goovaerts P, Franzblau A, Nriagu JO. Lifetime exposure to arsenic in drinking water and bladder cancer: a population-based case-control study in Michigan, USA. *Cancer Causes Control.* 2010 May;21(5):745-57.

Pronk A, Nuckols JR, De Roos AJ, Airola M, Colt JS, Cerhan JR, Morton L, Cozen W, Severson R, Blair A, Cleverly D, Ward MH. Residential proximity to industrial combustion facilities and risk of non-Hodgkin lymphoma: a case-control study. *Environ Health.* 2013 Feb 22;12:20

Westat, 2014. *NCI Residential History Project, Part 1 Data Sources.* A technical report provided to NCI on April 16, 2014.

Wheeler, D and Wang, A. Assessment of Residential History Generation Using a Public-Record Database. *Int. J. Environ. Res. Public Health* **2015**, *12*, 11670-11682.