Constructing Monthly Residential Locations of Adults Using Merged State Administrative Data

Mark C. Long
University of Washington

Elizabeth Pelletier
University of Washington

Jennifer Romich
University of Washington

ACKNOWLEDGMENT We thank seminar participants at the University of Washington's Center for Studies in Demography and Ecology and the Association for Public Policy Analysis and Management fall conference for helpful comments on this research. We thank the Research and Data Analysis department of the Washington State Department of Social and Health Services, and particularly David Mancuso, Jim Mayfield, and Kevin (Buzz) Campbell, for help in acquiring, merging, and cleaning state administrative data. The authors would like to acknowledge the contributions of the other study investigators, Scott Allard, Heather Hill, Jennifer Otten, Robert Plotnick, and Jake Vigdor, and key staff members Anne Althauser and Anita Rocha. We are grateful for funding from the City of Seattle, the Laura and John Arnold Foundation (now Arnold Ventures), Washington Center for Equitable Growth, and the Economic Self-Sufficiency Policy Research Institute. We benefit from use of the computing resources of the Center for Studies in Demography and Ecology at the University of Washington (Eunice Kennedy Shriver National Institute of Child Health and Human Development research infrastructure grant, R24HD042828). The contents of this article are solely the responsibility of the authors and do not necessarily represent the official views of any funder. Email correspondence to Mark Long <marklong@uw.edu> (corresponding author), Elizabeth Pelletier <epell@uw.edu>, and Jennifer Romich <romich@uw.edu>.

ABSTRACT In any month, administrative data collected by government agencies contain a fraction of the polity's adults, namely those persons who have had interactions with government agencies in that month. For researchers and policymakers who want to evaluate questions that require a spatial location of the whole population of adults at a given time (e.g., job-residence spatial mismatch, impacts of local policies), these fragmentary records are insufficient. Combining administrative data from several agencies in the U.S. state of Washington, we impute residential histories by parameterizing the "decay" in maintenance of an observed address. This process yields an imputed population whose demography and geographic distribution matches well with survey estimates. This work uses drivers' license, voter, social services, and birth records to append address locations to Unemployment Insurance data, a process that could be replicated with administrative records in other U.S. states and countries with sporadic address data from various agencies.

Introduction

Administrative data holds promise as a powerful and cost-effective source for demographic research (Cole et al., 2020; Connelly et al., 2016; Penner & Dodge, 2019). Routine administration of public programs creates records that contain large observation counts (populations rather than samples) over long periods of time with accurate reports on earnings, transfer income, voting, residential addresses, and other factors relevant to studies of human populations. These features of administrative data make it a particularly good source for answering questions about specific geographic areas, small populations, and groups defined by the intersection of demographic or socioeconomic characteristics. For instance, Jan Kabátek and Francisco Perales (2021) use Dutch registry data to show that children of same-sex parented families show higher achievement on many academic performance indicators available in schooling records. In another example, Annamarie Ernsten and colleagues (2018) link administrative records from the National Health Service with longitudinal survey records to examine internal migration in Scotland, discerning different trends between native-born and immigrant populations. Such analyses would not be possible with conventional survey data.

The use of administrative data in published research studies is increasing over time (Chetty, 2012) but as one set of observers notes, "the use of such data for policymaking and research still remains far below its true potential" (Cole et al., 2020). Increasing the use of administrative data from government sources requires overcoming several hurdles on the pathway from administrative records to analytic data, specifically issues pertaining to legality and governance; privacy and ethics; and data processing. In many cases, federal or state laws restrict use of data by third parties or for research purposes, and even in the absence of such laws, researchers may face bureaucratic hurdles or reluctance from agency staff or leadership (Hawn Nelson et al., 2020). Using data from private citizens' interactions with public systems poses privacy concerns and ethical uncertainty about how human subjects standards should apply (Goroff, Polonetsky, & Tene, 2017). Finally, converting records collected for the purposes of program administration into analytic data requires considerable work using approaches different from those developed for cleaning and curating survey data (Cole et al., 2020; Connelly et al. 2016). This third factor is the focus of the current article. Increasing the use of administrative data requires new knowledge about all of these factors, and case studies of administrative data use constitute evidence to build the field (Card et al., 2010; Cole et al., 2020; Penner & Dodge, 2019).

This article makes two contributions to the emerging literature on transforming administrative records into research data. First, we outline considerations that arise in the use of a merged state-level data set that is novel in that it includes a U.S. state's voters and driver's license records. Because these records are available in all 50 states, other researchers may be able to replicate this approach. Second, we describe and test a new method for an address-based population imputation process that yields continuous residential histories from sporadic address observations. This process yields a population and spatial distribution that mirrors Census data well; yet, constructing households based on address co-location over-represents larger households. As a whole, this work advances knowledge and methods for creating census-like data from administrative records.

Administrative Data and Demographic Research

While private businesses, nonprofit organizations, and public agencies all create "organic" or "found" data in the form of administrative records, our focus here is on administrative data from government sources. Studies based on public administrative data contribute important insights on demographic topics such as birth cohort size effects, fertility, education, migration, marriage and divorce, and cause of death (e.g., Agarwal et al., 2021; Monti et al., 2019; Cancian, Chung & Meyer, 2016; Conger, 2015; Figlio, Karbownik, & Salvanes, 2017; Kabátek & Perales, 2021; Ernsten et al., 2018; Gibson-Davis, Ananat, & Gassman-Pines, 2016; Grippo et al., 2020). Nordic countries have several decades of experience in using public registry data for research purposes (Wallgren & Wallgren 2014). In the U.S., researchers commonly use records from Unemployment Insurance (UI) to examine employment and earnings outcomes (e.g. Kornfeld & Bloom, 1999). Records including UI and vital statistics also form the foundation of longstanding federally maintained datasets (National Center for Health Statistics, 2021; U.S. Census Bureau, n.d.).

Several features of administrative records make them better suited than survey records for some research purposes. Administrative data are measured at a higher frequency than other population-level data. For example, the administrative data that we feature in this paper, which come from the U.S. state of Washington, are recorded monthly, and records can be linked to create longitudinal data. In contrast, the U.S. Census, which in similarly comprehensive in scale, is conducted only at ten-year intervals, and public-use records are not linked over time. Administrative data's expansiveness allows researchers the opportunity to study effects of local policies on small subgroups in precise geographic areas (e.g., teenagers in a particular city). In contrast, the Census's American Community Survey (ACS), which has a similar frequency of data collection, only surveys less 0.1% of the population each month. This design produces sample sizes that are too small for such precise micro analysis., and the repeated cross-sectional data cannot answer questions about individual or household changes over time.

However, administrative data have limitations. Administrative data are collected to determine program eligibility and track client participation or compliance within a particular program. These data contain fields necessary for those purposes and their scope includes only the select group of program participants. Many agencies' records contain information on individuals rather than households. Lastly, individuals typically only show up in the data when they have interacted with the agency, limiting the ability to construct a population-level spatial distribution of individuals at specific time points. Administrative records from state unemployment insurance (UI) systems – a valuable source of data for studies of employment and earnings – illustrate these limitations. UI records contain accurate microdata on earnings, employer, and industry, but they lack information on workers' personal characteristics, household composition, and residential location within a state. Merging administrative data across sources can address some of these limitations. For instance, the U.S. Census Bureau amends personal demographic information from survey sources onto employment records in creating the LEHD data (Vilhuber & McKinney, 2014). Researchers who can gain access to the tightly controlled LEHD data can examine questions around earnings, but the data do not contain income from transfer programs nor information about non-workers.

Processing Data to Create Residential and Household Information

Part of processing administrative records into research data involves creating variables of interest to research questions (Wallgren & Wallgren 2014). Our study contributes to a burgeoning literature that attempts to use administrative data from a variety of sources to construct residential histories and indicators of household and family memberships. Administrative address data can locate individuals in physical space, allowing for research on questions about the interplay between environment and outcomes for studies on topics such as neighborhood effects or segregation; longitudinal data can yield residential histories capable of tracking how such factors change over time (Jenkins et al., 2021). Knowing where persons live is also an important precursor to constructing household membership, needed for household-level analyses. Finally, co-residence is an important marker of family membership, relationship through blood or marriage among co-residents, which is in turn a precursor to understanding many dynamics of human life courses. While administrative data can be used to develop household and family membership, researchers need to create and test new methods to do so, and such methods will necessarily vary by the type of administrative records involved (Cuccaro-Alamin et al., 2021).

Some extant examples illustrate how this work can happen. For example, Goldschmidt, Klosterhuber, and Schmieder (2017) use "address and name data from the universe of employment records in Germany" and "develop a new method for identifying married couples in administrative data" (p. 29). Specifically, they identify couples that consist of two individuals living in the same home location, having the same name, where one person is male the other female and the age difference between them being less than 15 years. They note several limitations with this procedure (e.g., adult siblings living together being erroneously labeled as a married couple). They "show supporting evidence that around 89 to 94% of these pairs are indeed married" (p. 29), yet the analysis misses many married couples, identifying "about 17% of all married couples in Germany and about 35% of couples where both spouses are in social security covered jobs or unemployed" (p. 29).

Gath and Bycroft (2018) use linked administrative data to create household and family information that they then compare to New Zealand's census. Their method defines households as individuals who share the same address at a given point in time. Encouragingly, they find that "(w)hen family information was available from admin sources… it matched quite well to census family information" (p. 6). Yet, they caution that,

> "There is not currently sufficient admin data to provide high-quality information on families. Although we combined information from a variety of admin sources to create family nuclei, this methodology resulted in only 60 percent of the census family count" (p. 5).

The Social Wellbeing Agency for New Zealand attempted to improve on this work, but their attempted methods "showed no improvement over the existing address table in the [Integrated Data Infrastructure]" (p. 5, Social Wellbeing Agency, 2020). They identified several key challenges including inconsistent timing of address information in administrative data; point-in-time conflicts between various sources of address information; and the fact that an "address notification" (e.g., "a person's present address at the time of their interaction with the recording organization" (p. 8)) may not indicate an actual address change.

Our work builds on these efforts using data from a U.S. state. We use the same general approach as the New Zealand efforts (Gath and Bycroft, 2018; Social Wellbeing Agency, 2020) in sequencing data and defining households based on co-residence. However, we introduce a new imputation process to deal with the sporadic nature of address information available at the state level. In the sections that follow, we describe our novel merged administrative data, the data processing to add demographic variables, and the residential history imputation process. We then benchmark our resulting data against Census records and discuss the overall strengths and weaknesses of this approach.

Washington Merged Longitudinal Administrative Data

This project uses data compiled from several state agencies under the Washington Merged Longitudinal Administrative Data (WMLAD) effort (Romich et al., 2018). University of Washington researchers developed WMLAD as part of work to understand income and labor market dynamics associated with minimum wage law changes, hence the team chose records that could capture as much of the state's working-age population as possible regardless of whether or not they were currently working.

WMLAD comprises longitudinal and geocoded administrative records from seven Washington state agencies, summarized in Table 1. To our knowledge, this is the first attempt at a state-level merged administrative data source that uses driver's license data and voting records. Two other well-developed state efforts, the Wisconsin Administrative Data Core (Brown et al., 2020) and the California Policy Evaluation and Research Linkage Initiative (California Policy Lab, n.d.), include neither licensing nor voting records, so prior work provided little guidance about our endeavor. Records are linked using a single unique person identifier, allowing researchers to merge data across agency sources and follow individuals over time.

Table 1. Washington Merged Longitudinal Administrative Data component data sources and relevant key information

Record Type	State Agency	Key Information	Time Period	Number of Individuals
Unemployment Insurance (UI)	Employment Security Department (ESD)	Earnings, hours worked, employer's industry	2000-2017	7,699,646 workers
Human services	Department of Social and Health Services (DSHS); Health Care Authority (HCA)	Program participation, race/ethnicity, sex, age, residential address	2010-2017	4,968,258 clients
Birth	Department of Health (DOH)	Race/ethnicity, sex, age, education, residential address	2010-2016	896,558 parents
Voting	Secretary of State (SOS)	Voting history, sex, age, residential address	2006-2016	6,084,439 voters
Licensing	Department of Licensing (DOL)	Age, sex, residential address	2005-2016	8,367,317 licensees
Arrests	Washington State Patrol (WSP)	Arrest characteristics	2000-2018	777,416 people arrested

Table 2 shows the overlap between data sources. Each row represents a population from a key WMLAD data source (i.e., driver's license holders), all defined using records between 2010 and 2016. The columns indicate what share of that population was also present in another key population (i.e., the first row of the fourth column indicates that 60% of driver's license holders worked). The final column indicates what share of individuals who were in a given population were not included in any of the other populations in the table (i.e., 11% of driver's license holders were neither workers nor DSHS clients nor registered voters nor parents of newborns).

Table 2. Overlap Between Washington Merged Longitudinal Administrative Data Populations, 2010-2016

	From Other WMLAD Data, Percentage Who…						
Key WMLAD Populations:	Held a Driver's License	Registered to Vote	Voted	Worked	DSHS Client	Parent of a Newborn	Were Not in Other Listed Sources
Driver's License Holders	100%	64%	50%	60%	43%	10%	11%
Registered Voters	83%	100%	68%	55%	37%	9%	14%
Workers	80%	57%	44%	100%	43%	11%	17%
DSHS clients	66%	44%	28%	50%	100%	11%	28%
Parents of Newborns	87%	58%	41%	72%	58%	100%	9%

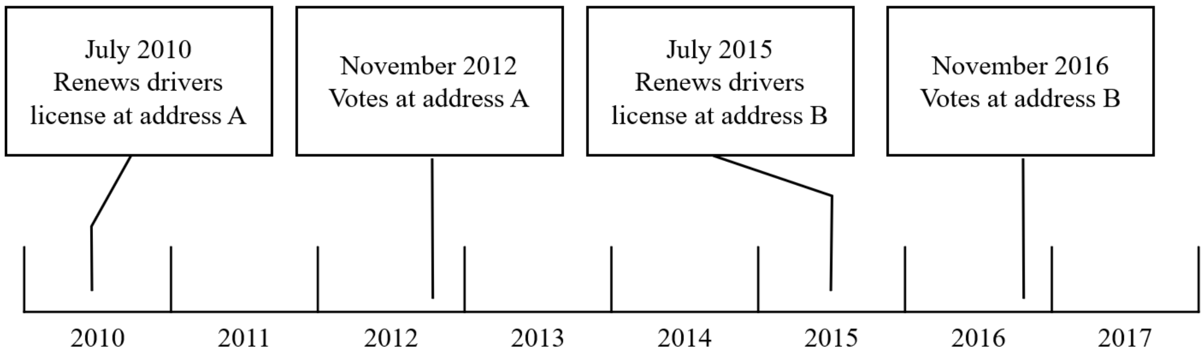
Notes: "Parents of newborns" are parents of children born in Washington between 2010 and 2016.

Although the WMLAD linkage process included quality control measures, the nature of such administrative data linkage is such that no clear standard exists to evaluate match quality in practice (Harron et al., 2017), particularly when linking across several different cases. In such instances, comparing characteristics of the linked data to other known population estimate constitutes an important check. Hence, we benchmark population counts from WMLAD against 2010 Decennial Census and ACS published tables and microdata to assess the overall coverage of our combined data.

WMLAD address records

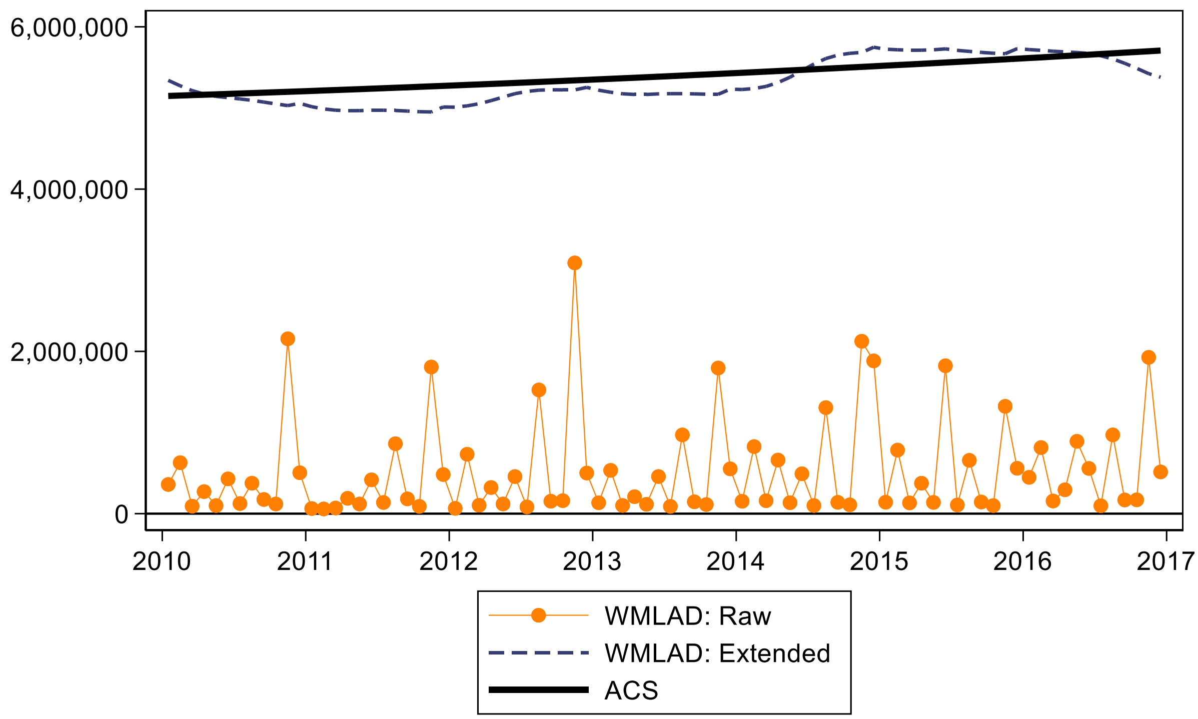
Residential address data are found in human services records from the Department of Social and Health Services (DSHS), birth records from the Department of Health (DOH), voting records from the Secretary of State (SOS), and drivers' license records from the Department of Licensing (DOL) in months when state residents interacted with those agencies. Although some data files contain address observations for every month, these may be "stale" and hence untrustworthy (Jim Mayfield, Washington State DSHS, personal communication). Hence, we rely on address information in months when residents either reported a new address or otherwise interacted with the agency such that we are confident the address information is accurate. Figure 1 illustrates a possible set of address information for one hypothetical person. Within our focal time period, this person is first observed interacting with the DOL by renewing a driver's license at address A. The person then votes at that same address about two years later. The next two observations are at a different address and again consist of a driver's license renewal and voting in a presidential election.

Figure 1. Illustration of address location data for a hypothetical observation



In our data, the availability of address data varies considerably across months based on the varying frequency of state residents' interactions with these agencies. The points on Figure 2 show this variation by month. For example, address data are most thoroughly available in Novembers of congressional and Presidential election years due to voting. On average were 520,444 Washington adults with a valid address reported in our administrative data sources in a given month. Monthly totals vary widely, with a standard deviation of 606,285, ranging from a minimum of 57,786 (February 2011) to a maximum of 3,090,483 (November 2012).

Figure 2. Number of Adults in ACS and WMLAD, before and after extending (imputing) observed addresses to months with missing addresses



The solid line in Figure 2 shows an estimate of the number of adults in Washington during this period and is based on the ACS (Ruggles et al., 2020) that is smoothed by a regression of the person-weighted annual population estimate regressed on year and year-squared. We use this total, as well as totals for demographic subgroups, as a target for the imputation process described next.

Methodology: Overview

The goal of the imputation process is to move from sporadic address point observations as illustrated in Figure 1 to complete residential histories consisting of an address for every month in the data. A simple approach to this task would be to assume a person remains at an address until observed elsewhere. For instance, this would assume that the hypothetical person in Figure 1 remains at address A until they change their address on record to address B at the point of renewing their driver's license. This approach would fail to capture two important considerations. First, there may be lags between residential moves and interactions with state agencies. While some citizens may update their drivers license address, voting registration, and other address data when they move, we believe that others may not. Second, while we cannot observe moves out-of-state or deaths in our data, some proportion of the population will make such moves. Extending the last in-state data observation into the future fails to acknowledge the right-censored nature of our observations. To account for such behaviors, we have developed a new method for imputing residential location that benchmarks the total population and demographic subgroups against ACS adult populations.

For the imputation, we predict the probability of continued residence at an observed residence forward and backward into months during which the individual's address is not observed. We assume that the probability of persistence decays in the months that follows. We use a flexible function to model this decay, with the shape and speed of this decay controlled by four parameters: γ_1 and γ_2 (for forward decay) and γ_3 and γ_4 (for backward decay). We identify the values of the gamma parameters that minimize the sum of squared differences between the size of our imputed WMLAD adult population and the adult population for the state of Washington that is estimated by the ACS. We use "Particle Swarm Optimization" (PSO), described below, as our preferred optimization method to identify optimal values of these gamma parameters. Finally, we improve the fit of our imputed population to the sex, age, and race/ethnicity demographics of the state by scaling the decay functions by a set of beta parameters (e.g., β_{Male}). If these beta parameters take a value less than 1, then it implies that this demographic group persists at their observed addresses longer than the base demographic group. The values of the beta parameters are optimized iteratively after successive runs of the PSO.

Next, we describe the address data preparation before turning to the imputation process. Our imputation includes benchmarking by demographic subgroups defined by age, sex, and ethnorace. Appendix A contains information on how we created the demographic variables.

Data Preparation: Determine an Observed Address for Month m

State administrative voter, social service, birth, and license data contain addresses. When these address data conflict for month m, we prioritize the data source for which we have the highest confidence. We place the highest priority on voting records from SOS. Since voting in Washington occurs exclusively by mail, we assume that if an individual successfully voted in month m they have an accurate address at that time. Second, we add observations of address *changes* from the SOS and then DSHS. We assume that an update to the database is likely to accurately reflect residential location in month m. Next, we add the address of biological parents of newborns born in month m from DOH. Finally, we incorporate biannual snapshots of the DOL driver's license database. Individuals are included in a snapshot if they had an active driver's license on file. Since there are limited incentives to update DOL records following a move, we only use license data in months when individuals have an updated address relative to the last snapshot. The sequence of observed addresses is constructed for all months between January 2010 to December 2016, which are indexed from month $m = 1$ to month $m = 84$.

We use a similar process to establish the best address as of January 2010, the beginning of our focal period and the date before which data availability is more sporadic for most sources.

This pre-January 2010 address data is used to impute the January 2010 data per the method described in the text. For each individual, we identify the most recent month pre-January 2010 in which one of the following types of information was available: a voter who voted in a given month, an update to the voter rolls, an update to the DSHS client records, or an update to the DOL database (included as the first possible month). If none of this information is available, we use either the first available month of voter data or the first available month of DOL data as the pre-January 2010 address. (Birth records are not available prior to 2010). We use the ranking described above to reconcile conflicting information, then use the most recent available data point as the individual's pre-January 2010 address.

Imputation: Extend Observed Residential Address to Missing Months

Functional form

We begin by estimating a probability that an individual with an address in month m persists at this address in month $m + 1$ (and, if so, to month $m + 2$, and so on). We draw a random number uniformly distributed from 0 to 1 and impute continuance between months m and $m + 1$ if the random number is below the estimated probability of persistence.

For an individual who was female, age 18 to 29, and White, we estimate the probability of this person i being at their observed month m address in missing month $m + j$ using the following equation:

$$(1) \quad pr\left(Address_{m+j} = Address_m\right) = \left(1 - \frac{\gamma_1 j^{\gamma_2}}{1+\gamma_1 j^{\gamma_2}}\right)^2$$

For example, the probability that an observed January 2010 address persists to February 2010 would be given by the following equation:

$$(2) \quad pr(Address_2 = Address_1) = \left(1 - \frac{\gamma_1 1^{\gamma_2}}{1+\gamma_1 1^{\gamma_2}}\right)^2 = \left(1 - \frac{\gamma_1}{1+\gamma_1}\right)^2$$

Furthermore, the probability that the person's January 2010 address persisted to March 2010 would be estimated as follows:

$$(3) \quad pr(Address_3 = Address_1) = \left(1 - \frac{\gamma_1 2^{\gamma_2}}{1+\gamma_1 2^{\gamma_2}}\right)^2$$

The functional form that we use for this predicted probability of persistence has desirable features. First, note that if $j = 0$, then the $pr(Address_{m+0} = Address_m) = 1$. That is, for the month with an observed address, where no imputation is needed, the probability of being at this address is 100%, and this forms an anchor from which the probability of persistence smoothly decays. Second, the functional form allows for various shapes of decay, with the γ_1 and γ_2 parameters controlling the speed and shape of decay in the probability of still residing at person i's month m address. For example, Appendix Figure 1 shows examples where the decay can be characterized as a sigmoid function bounded between 0% and 100% (shown with $\gamma_1 = 0.0005$ and $\gamma_2 = 3$) and a second example where the shape can be characterized as reflecting exponential decay from a base of 100% (shown with $\gamma_1 = 0.05$ and $\gamma_2 = 1$).

If the observed address is from a month prior to January 2010 (during which the address data is spottier) and we are imputing the probability of persistence at this address into the period beginning with January 2010, we modify equation 1 by incorporating γ_0 as follows:

$$(4) \quad pr\left(Address_{m+j} = Address_m\right) = \left[\left(1 - \frac{\gamma_1 j^{\gamma_2}}{1+\gamma_1 j^{\gamma_2}}\right)^2\right]^{\gamma_0}$$

γ_0<1 results in discounting of the address information prior to January 2010. γ_0 is set to 1 for imputing persistence of all address data beginning with January 2010.

Finally, our base case, described above, is based on the most populous subgroups: female, age 18 to 29, and White. (Using a different subgroup as the base case would yield functionally equivalent results). For other groups, we allow the shape of the probability of address persistence to be scaled upwards or downwards as follows:

$$(5) \quad pr\left(Address_{m+j} = Address_m\right) = \left[\left(1 - \frac{\gamma_1 j^{\gamma_2}}{1+\gamma_1 j^{\gamma_2}}\right)^2\right]^{\gamma_0 \beta_{1i} \beta_{2i} \beta_{3i}}$$

β_{1i}, β_{2i}, and β_{3i} are parameters that are greater (less) than 1 if we need to decrease (increase) this person's probability of being at their month m address on account of person i's sex, age, and race as follows. β_{1i}, β_{2i}, and β_{3i} are shorthand for the following are expanded expressions:

$$(6) \quad \beta_{1i} = Female_i + Male_i \beta_{Male} + (1 - Female_i - Male_i)$$

$$(7) \quad \beta_{2i} = Age18to29_i + Age30s_i \beta_{Age30s} + Age40s_i \beta_{Age40s} + Age50s_i \beta_{Age50s} + Age60s_i \beta_{Age60s} + Age70plus_i \beta_{Age70plus} + (1 - Age18to29_i - Age30s_i - Age40s_i - Age50s_i - Age60s_i - Age70plus_i)$$

(8) $\beta_{3i} = White_i + Hispanic_i\beta_{Hispanic} + Black_i\beta_{Black} + API_i\beta_{API} + AIAN_i\beta_{AIAN} + Multiracial_i\beta_{Multiracial} + (1 - White_i - Hispanic_i - Black_i - API_i - AIAN_i - Multiracial_i)$

Note that for the base case (i.e., female, age 18 to 29, and White), β_{1i}, β_{2i}, and β_{3i} each equal 1 as does their product. For persons whose sex is missing or "other" or "unknown" we treat their probability of persistence the same as the base case, female. Similarly, for those with missing ethnorace or age indicators, we treat their probability of persistence as the same as the base cases, non-Hispanic White alone and age 18 to 29, respectively. Consequently, the beta values for these groups equal 1, as shown in the final column of Table 3.

Table 3. Demography of the state of Washington for adults in ACS and adults in WMLAD with address information during January 2010 through December 2016, before and after extending observed addresses to months with missing addresses

		WMLAD		
Characteristic	**ACS**	**Raw**	**Extended**	**Beta**
All Adults	5,444,135	520,444	5,315,145	
Male	2,697,433	228,372	2,540,510	0.69
Female	2,746,702	281,485	2,633,185	1
Other/Unknown		28	324	1
Sex is Missing		10,559	141,126	1
18-29	1,175,504	98,273	1,130,295	1
30s	972,486	86,750	927,560	1.40
40s	929,903	76,375	885,615	0.76
50s	962,490	88,824	915,904	0.72
60s	768,413	85,217	731,092	0.79
70 and above	635,339	75,304	607,540	1.21
Age is Missing		9,702	117,138	1
Hispanic	411,057	40,150	381,350	4.59
White Alone, Non-Hispanic	4,033,758	383,010	3,768,198	1
Black Alone, Non-Hispanic	184,543	18,619	171,311	3.03
API Alone, Non-Hispanic	466,739	24,018	397,637	0.07
AIAN Alone, Non-Hispanic	59,262	4,829	54,529	1.09
Other Alone, Non-Hispanic	6,629			1
Multi-racial, Non-Hispanic	282,146	19,066	212,270	0.01
Race/Ethnicity is Missing		30,753	329,851	1

Notes: ACS data come from Ruggles et al. (2010) and show the average counts (implied by the weight "perwt") for the years 2010 to 2016. WMLAD columns show the average counts for the months January 2010 to December 2016. "API" denotes "Asian or Pacific Islander" and "AIAN" denotes "American Indian or Alaskan Native". The "Beta" column shows the final values of the beta parameters that are selected by our optimization.

Some demographic groups are under- or over-represented in our raw data with observed address-months and our extension process adjusts accordingly. For example, in an average month, among those identified as “male” or “female”, 44.8% are male. This share compares to 49.5% male per ACS. Underrepresentation of males could be caused by males being less likely to be present in the state’s data, females updating addresses more frequently, and/or females moving into and out of Washington at a faster rate than males. Our process does not adjudicate between these possible explanations. Rather, to achieve an imputed dataset that matches the state’s demography, per ACS, we incorporate a greater persistence of males at observed addresses relative to females, which will be achieved by β_{Male} being less than 1.

Note that our procedure assumes that the basic shape of the decay in the probability of persistence of a person at their observed address is common across all persons, yet that that common shape is scaled upwards or downwards for demographic subgroups. An alternative approach, which we did not fully explore but which could be attempted in future research, would be to estimate the gamma parameters separately for each subgroup. For example, one could estimate $\gamma_{1,Female}, \gamma_{2,Female}, \cdots, \gamma_{1,Multiracial}, \gamma_{2,Multiracial},$ Of course, such a procedure produces many more parameters to estimate. Early versions of this paper explored similar, more flexible specifications, but we found that the parameters were difficult to estimate and produced poor fitting results.

After running through this forward imputation to December 2016, we repeat the process in the reverse order, estimating the probability that the individual was at their observed month m address in missing month $m-1$ (and, if so, to month $m-2$, and so on) going back to January 2010. If the person was not observed at an address in any month prior to month m, we estimate the probability of person i being at their observed month m address in month $m-j$ as follows:

$$(9) \quad pr\left(Address_{m-j} = Address_m\right) = \left[\left(1 - \frac{\gamma_3 j^{\gamma_4}}{1+\gamma_3 j^{\gamma_4}}\right)^2\right]^{\beta_{1i}\beta_{2i}\beta_{3i}}.$$

If the person was previously observed at an address in month $m-j-J$, we estimate the probability of person i being at their observed month m address in month $m-j$ as follows:

$$(10) \quad pr\left(Address_{m-j} = Address_m\right) = \left[\left(1 - \frac{\gamma_3 j^{\gamma_4}}{1+\gamma_3 j^{\gamma_4}}\right)^2\right]^{\beta_{1i}\beta_{2i}\beta_{3i}} \times \left(1 - \left[\left(1 - \frac{\gamma_1 J^{\gamma_2}}{1+\gamma_1 J^{\gamma_2}}\right)^2\right]^{\gamma_0\beta_{1i}\beta_{2i}\beta_{3i}}\right).$$

The last term in Equation (10) captures the probability that the observed address in month $m-j-J$ has not persisted through to month $m-j$. The γ_3 and γ_4 parameters in Equations (9) and (10) control the speed and shape of decay in the probability of residing at person i’s month m address in prior months. Finally, note that the same beta parameters are used to scale the backward and forward decay in the probability of address persistence.

Parameter estimation

The gamma and beta parameters are estimated iteratively. First, we set all of the beta parameters to 1 and estimate the gamma parameters by PSO (Eberhart and Kennedy, 1995). Particle swarm optimization is one type of algorithm inspired by optimization in nature, such as a flock of birds or a school of fish that swarm locations in order to identify the best location for food. Each bird in the flock, it is assumed, chooses its direction of flight based on the best location of where it has found food and the best location where the flock, as a whole, has found food. Birds in this model share information. After a period, the birds will converge at the same optimal location – each bird's best location will be identical to the flock's best location and each bird will slow their speed as they approach this optimal location. PSO is well suited to our problem as it does not require the computation of derivatives of the loss function, which would be challenging for our problem and which would be required for other optimization methods (e.g., gradient decent).

For our PSO, we use ten independent particles (i.e., "birds") and each particle contains a candidate set of parameters and velocities (i.e., speeds at which each parameter is moving). The directions in which each particle moves its parameters is influenced by the local best set of parameters the particle has found on its own journey and the global best set of parameters that have been found across the ten particles.

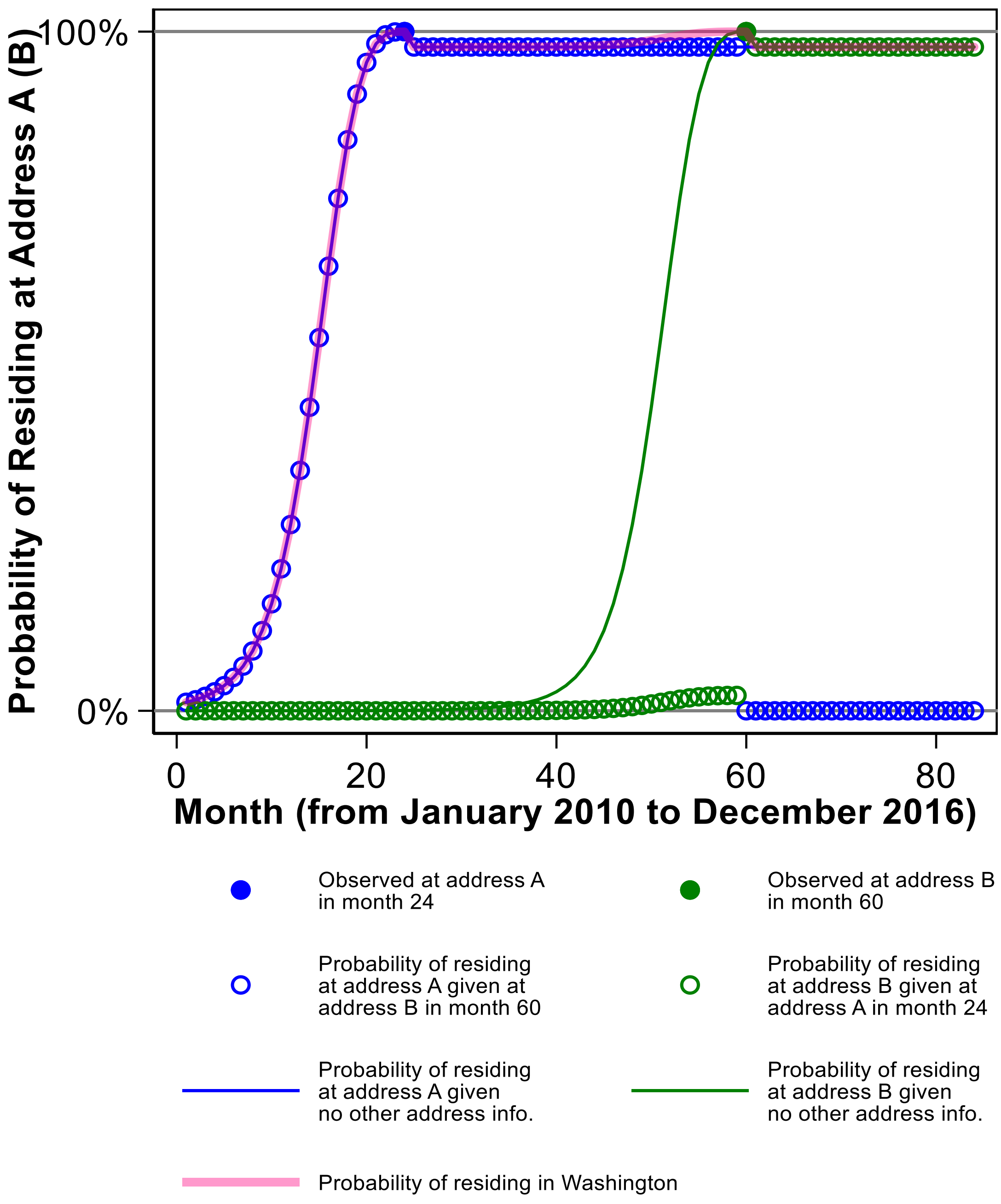
Following Clerc (1999) and Eberhart and Shi (2000), we include a constriction factor, K, to constrain the particle velocities. We set K to 0.73 per Carlisle and Dozier (2001). The initial base values for the gamma parameters, which were informed by earlier optimization explorations, were as follows: $\gamma_0 = 0.41$, $\gamma_1 = 0.0018$, $\gamma_2 = 0.84$, $\gamma_3 = 0.0002$, and $\gamma_4 = 3.14$. For each of the 10 particles, these base values were multiplied by a random number uniformly distributed between 0.8 and 1.2.

The gamma parameters seek to minimize the loss function: $Loss = \sum_{m=1}^{84}(WMLAD.Imputed.Adult.Population_m - ACS.Adult.Population_m)^2$. As the PSO progresses, each particle's search direction and speed are influenced by the set of gamma parameters that it has investigated and which has produced the lowest $Loss$ it has observed and the set of gamma parameters that the flock has identified which has produced the lowest $Loss$.

After a period, which averaged five hours, we paused the PSO and readjusted the beta parameters. The new value of beta was set equal to the old value multiplied by the current imputed number of persons in that demographic subgroup (averaged across all months) divided by the expected number of persons in that subgroup. For example, to adjust β_{Male}, we would multiply it by $\frac{WMLAD.Imputed.Adult.Male.Population_m}{49.5\% \times WMLAD.Imputed.Adult.MaleOrFemale.Population_m}$ (where 49.5% reflects the male share of the population per ACS). If the ratio shown in the prior sentence was greater than one, it would imply that the imputed WMLAD had too many males and thus β_{Male} would be increased. If the ratio was less than one, it would imply that the imputed WMLAD had too few males and thus β_{Male} would be decreased. Following this adjustment, we restart the PSO to continue to optimize the gamma parameters. Note that as this process continued, the fractions by which we would multiply the beta parameters would get closer and closer to 1 as we converged on the best set of beta parameters.

Convergence was defined as each of the beta parameters for sex and age being multiplied by a fraction lying in the interval of (0.99, 1.01) followed by less than a 0.5% reduction in the loss function in the last 10 hours of gamma parameter estimation. Convergence was achieved after the beta parameters had been adjusted twenty times yielding the parameters shown in the final column of Table 3.

Figure 3. Illustration of the predicted probability of residing at address A and B for use in extending addresses to months with missing addresses.



Results

The PSO converged upon the following gamma parameters: $\gamma_0 = 0.34, \gamma_1 = 0.0115, \gamma_2 = 0, \gamma_3 = 0.0002, \gamma_4 = 3.33$. Figure 3 illustrates these parameters for a hypothetical person who is seen residing at address A in month 24 and address B in month 60 and for whom $\beta_{1i}\beta_{2i}\beta_{3i} = 1$. The predicted probability of residing at A prior to month 24 rises sharply as we approach month 24. The probability of remaining at A falls immediately to 0.977 where it stays through month 60 and then is set to 0 thereafter. The PSO converging at $\gamma_2 = 0$ is what generates the constant probability of remaining at address A in the intervening months. The unconditional predicted probability of residing at B in months 25 to 59 is multiplied by (1 - probability of remaining at A during these months) to produce the conditional probability of residing at B during these months.

The thick, pink line in Figure 3 provides the probability of being imputed to live in Washington at either address A or B. This pink line is the same as the predicted probability of residing at A prior to month 24, and thus it rises sharply as we approach month 24. The probability of being at A or B is nearly 100% during the in-between months, having a minimum of 97.7% in month 25, with the residual 2.3% being the probability of residing out-of-state. After month 60, the pink line is the same as the predicted probability of residing at B after month 60, and thus it falls to 97.7% and remains constant thereafter. Thus, the model suggests that such an individual is very likely to have remained in Washington for months 24-84, but this individual had less than a 50% chance of having resided in Washington prior to month 14 given the lack of address information in state administrative data before month 24.

The convergence of γ_2 to zero is a surprise, and it produces the odd result of constant persistence at a fixed probability after an address is observed. However, note that we are estimating persistence over a short window (6 years) and given that the observed address may come at any point during that window, this odd result may not be that strange.

These parameters do well in yielding an adult population that matches the count of Washington's adults per ACS as shown by the dashed line in Figure 2. In an average month, the absolute difference between the ACS population estimate and the WMLAD population with observed or imputed addresses is 3.0%.

As shown in the third column of Table 3, the fit is generally strong for each demographic group, with the average absolute difference between the demographic group's share in ACS and share in imputed WMLAD being just 0.3 percentage points. The fit is particularly strong for the male/female shares and shares by age group, with each deviation being less than 0.4 percentage points.

Our imputed population underestimates the numbers of API and Multiracial persons. As shown in the fourth column of Table 3, the beta parameters for these groups converge at nearly 0, which means that the imputation probabilities are near 100% for each month in which these persons are not observed. Even with these low parameters, there are insufficient numbers in these groups. ACS suggests we should expect 8.6% (5.2%) of the Washington adult population to be API (Multiracial), whereas our imputed WMLAD population contains 7.6% (4.3%) API

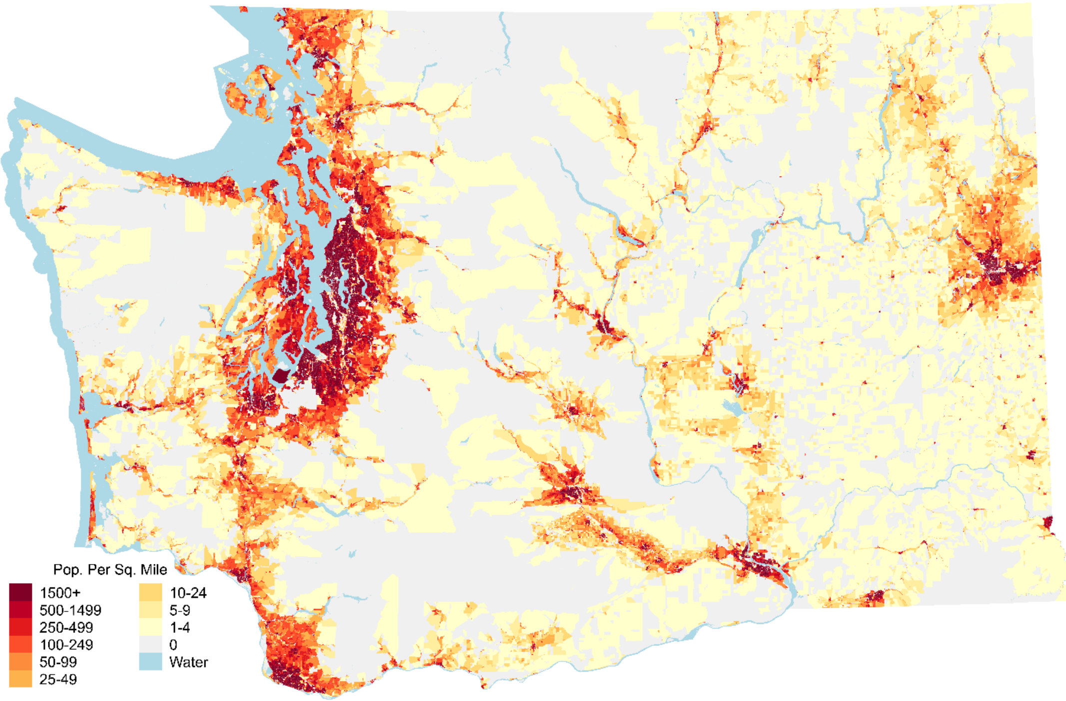
(Multiracial). Also note that the beta parameter for Hispanics is large, 4.59. This large parameter will mean that Hispanics are estimated to remain at their observed residences for shorter durations. Migrant farm workers may be contributing to the size of this parameter. Stromsdorfer (2007) notes that farm employment peaks in June, July, September, and October. Based on WMLAD's observed, non-imputed address data, Hispanics comprise 14.0% of the state's population during these months and 9.6% of the state's population in other months. Appendix Figure 2 graphically illustrates the effects of the beta parameters for an individual who is observed to reside in Washington during month 24 and at no other time. As this figure illustrates, the beta coefficients for sex and age have only modest effects on the predicted decay in the probability of residing at the observed month 24 address. In contrast, the estimated probabilities of persistence vary substantially by race.

Figure 4 shows the population density of each Washington Census block based on adult population estimates from the 2010 Census (Panel A) and the average month in 2010 using imputed WMLAD populations (Panel B). Spatial distributions of population densities are very similar; the block-level correlation between these two data sources is 0.913. Note, however, that this 0.913 correlation is only a modest improvement over the block-level correlation between the raw, non-imputed WMLAD address data and the 2010 Census data, which is 0.890.

We then combined all persons residing simultaneously at a given address into quasi-households. Table 4 compares WMLAD quasi-households with ACS data. Our process yields an imputed population that matches ACS well in terms of households with one adult; both datasets suggest 1 million Washington adults in such households. However, WMLAD contains too few adults imputed to be in households with two adults (2.9 v. 1.8 million) and too many adults imputed to be in households with four or more adults (0.6 v. 1.3 million). This result appears to be an artifact of how the raw addresses were converted (by state administrators) into anonymized address ids prior to sharing the data with us. The ACS definition of a household is analogous with ours in that it does not require any sort of familial or economic ties, but rather considers everyone living in the same "housing unit" to be part of the same household. Therefore, we believe that the discrepancy in household size estimates between WMLAD and the ACS likely emerges from a difference between "addresses" as we are able to observe them and "housing units" as the ACS defines them. Specifically, addresses in the WMLAD data may encompass multiple housing units. A challenge is dealing with apartment building addresses or dwellings like duplexes when the administrative data lack the unit number. Such units would appear as a single "household" and thus generate artificially large households. More conservative assumptions yielded similar results. Solving this challenge would require finer grained address information than what is available in the current data.

Figure 4. Population density for each Census block in the state of Washington based on the 2010 Census (Panel A) and the average population during January through December 2010 in WMLAD with observed addresses extended to months with missing addresses (Panel B)

Panel A: 2010 Census



Panel B: Jan.-Dec.2010 WMLAD (post-imputation)

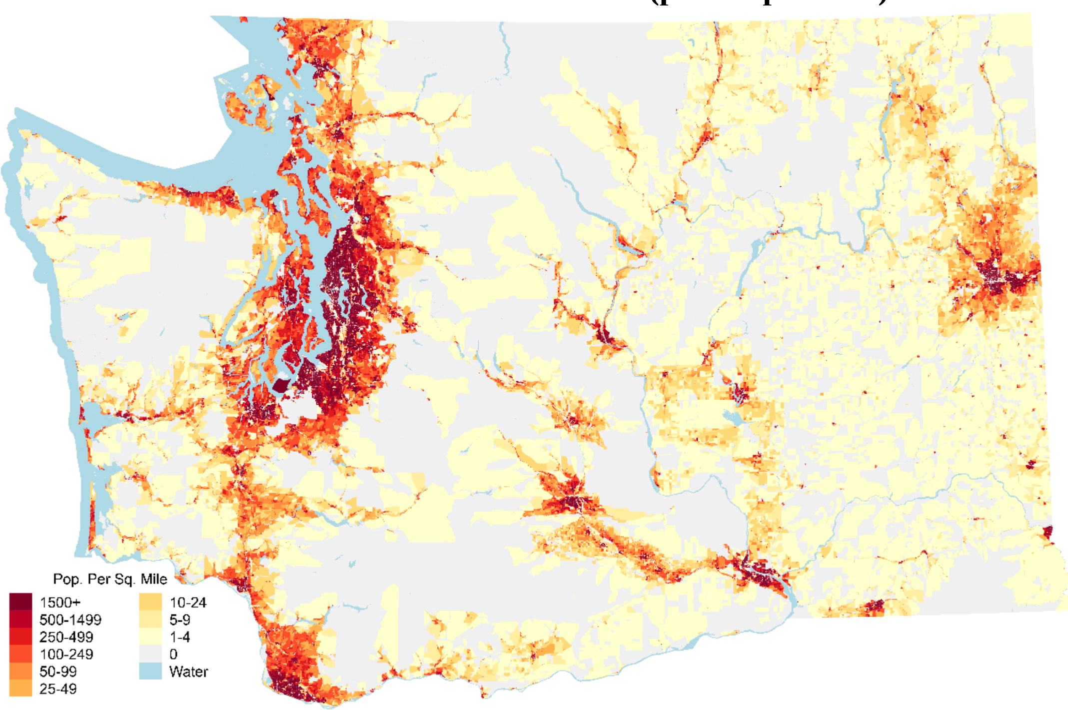


Table 4. Number of adults per household in the state of Washington from ACS and WMLAD

Characteristic	ACS	WMLAD
All Adults	5,444,135	5,315,145
Living in HH With 1 Adult	1,006,008	1,011,228
Living in HH With 2 Adults	2,917,162	1,803,298
Living in HH With 3 Adults	895,175	1,151,820
Living in HH With 4+ Adults	625,790	1,348,799

Notes: ACS data come from Ruggles et al. (2010) and show the average counts (implied by the weight "perwt") for the years 2010 to 2016. WMLAD columns show the average counts for the months January 2010 to December 2016 after extending observed addresses to months with missing addresses.

Discussion and Conclusion

The methods described here add to the value of administrative data for applied demographic research. We demonstrate the viability of using merged government administrative data to create a month-by-month dataset of residential address histories whose total population and geographic and demographic distribution matches well with snapshot data from the U.S. decennial Census and monthly data from the ACS survey and improves on decennial Census data by being higher frequency and improves on ACS data by covering the population rather than a sample. Our data thus permits analysis of issues that require large samples with high frequency. In this section, we discuss how our novel method compares to previous approaches in the literature and then overview how the resulting WMLAD data – or other similar state-level efforts – can be used to examine important demographic questions.

Imputation method

The novel contribution of our analysis is a procedure for parameterizing the decay in the probability that the individual is likely to remain at their most recent address that is known in administrative data and to have already been at that address in months prior to it being known by the administrative data. We illustrate a method for identifying decay parameters as a function of individual characteristics, including age, sex, and ethnorace. We show that using this method results in longitudinal residential histories that generate aggregate populations that match census estimates of the adult population of the state. This method improves on existing methods that have been previously used (e.g., New Zealand's "existing approach uses a person's most recent address as the best estimate for their residence at a point in time" (p. 7, Social Wellbeing Agency, 2020)) as it allows the information about the most recent address to decay both forward and backward in time.

The potential error introduced by the simple approach of extending the most recent address forward in time will vary by data structure and the underlying population dynamics. In the case of the current study's data, our findings suggest that this simple method would provide larger estimates of the state adult population relative to ACS estimates, and that these over-estimates would be slightly larger for some demographic groups, including women and Hispanic/Latino residents.

By sequencing observed addresses and parameterizing decay functions for the length of time a person likely spends at an address, we impute residential histories for approximately all adults in the state and produce quasi-households. All data contain imperfections, however. Given the complexity of WMLAD, its shortcomings and biases must be understood in relationship to specific types of inquiry, the topic we turn to next.

Possible uses

Such merged data with the construction of monthly pseudo-residential address information facilitates demographic and policy research. Here we discuss three types of research for which such a merged, longitudinal administrative dataset is uniquely powerful – spatial, policy impact, and household analyses– and also the limitations of the data that might arise during such applications. We draw specific examples from uses planned as part of our larger team's work studying Seattle's $15 minimum wage (UW Minimum Wage Study, n.d.) but all relate clearly to general questions about social and economic topics falling within the field of applied demography.

First, the residential histories created from merged longitudinal data allow for a finer-grained approach to spatial questions than is available via survey or administrative data from single sources. Important topics in spatial demography include residential segregation, spatial match or mis-match between residential locations and employment, and neighborhood change. One line of questions related to income and policy in Seattle is the extent to which higher-paid workers are displacing lower-wage workers within the city limits—and hence potentially blunting the possible impact of Seattle's $15 wage. Indeed, early analyses using WMLAD show that lower-paid Seattle workers moved more times than higher-paid workers (Foster et al., 2021) over the period that the higher wage took effect. Future work will look at the implications of those moves for commute times.

Second, such data are very useful for examining the impact of city- or county-level policy interventions. While our larger research team has examined the impact of the Seattle wage law on employees of firms inside Seattle using UI data, such data are insufficient to answer more general questions about the impact of this or any policy on residents of the city. Using survey data, like the ACS, for the purposes of evaluating the effects of a geographic-based policy is challenging and likely to produce large standard errors due to small sample sizes in micro areas. In contrast, the imputed WMLAD's large (approximately comprehensive) population size makes it ideal for such policy analyses. Similarly constructed data could form the basis of impact studies of other local policies, such as the city- and county-level eviction bans instituted during the pandemic. However, we should note that uncertainty in the process of imputing continuance at a residential address could generate attenuation bias in such policy analyses. For example, if a person is imputed to remain at an address that is affected by a geographic-based policy, but the person did not, in fact, remain at that address, we might incorrectly conclude that the policy did not affect the person (rather than correctly conclude that the person's outcomes were unaffected because the person no longer resided in that policy jurisdiction).

Third, by linking persons via co-residence, this data permit analysis of households (persons living together) or families (persons related by blood or marriage). This includes important demographic topics such as poverty (a household-level construct); family formation and dissolution; and inter-generational mobility. For instance, one question within the larger minimum wage literature is the extent to which minimum-wage workers are young persons from middle- or high-income households (Newmark & Wascher 2007). With WMLAD, we will be able to identify young workers at or near the minimum wage who live in households with other higher-paid workers—and our longitudinal data will allow us to examine the earnings trajectories of young workers by their parents' earnings levels even after they have left their parents' households. For the subset of the persons in our data who have records in the DSHS client data, we will be able to triangulate our constructed households against household rosters reported to DSHS to determine program eligibility. Public assistance program eligibility rules typically apply at the household or family level, known to program administrators as "assistance units." By comparing households constructed via the current address co-location method to DSHS assistance units, we will be able to better understand potential systematic biases in our records. While neither assistance units nor address co-location necessarily represent the ground truth on household membership much less family membership, we believe this combination of data will yield a set of helpful and fairly accurate working definitions for examining income and poverty at the household and family level.

Conclusion

To realize the promise of administrative data for demographic research, the scholarly community needs to create and share methods for meeting the governance, ethics, and data processing challenges inherent in this endeavor (Cole et al., 2020; Penner & Dodge, 2019). Toward that end, this paper constitutes an important "use case" of how merged records from state-level public agencies can be transformed into helpful longitudinal data. Our novel address imputation method builds and improves on previous efforts, yielding useful evidence for approaching important spatial, economic, and social questions.

References

Agarwal, S., W. Qian, T. F. Sing, & P. L. Tan. 2021. Fortunes of dragons: Cohort size effects on life outcomes, *Population Studies* 75(2): 191-207.

Brown, P. R., K. Thornton, D. Ross, J. A. Smith, & L. Wimer. 2020. *Technical Report on Lessons Learned in the Development of the Institute for Research on Poverty's Wisconsin Administrative Data Core*. University of Wisconsin-Madison. https://www.irp.wisc.edu/wp/wp-content/uploads/2020/08/TechnicalReport_DataCoreLessons2020.pdf

California Policy Lab. n.d. *Life Course Dataset – California Policy Lab*. Available: https://www.capolicylab.org/life-course-dataset/ (accessed: 19 December 2021).

Cancian, M., Y. Chung, & D. R. Meyer. 2016. Fathers' imprisonment and mothers' multiple-partner fertility, *Demography* 53(6): 2045-2074.

Card, D. E., R. Chetty, M. S. Feldstein, & E. Saez. 2010. Expanding Access to Administrative Data for Research in the United States. *American Economic Association, Ten Years and Beyond: Economists Answer NSF's Call for Long-Term Research Agendas*. http://dx.doi.org/10.2139/ssrn.1888586

Carlisle, A., & G. Dozier. 2001. An off-the-shelf PSO, *Proceedings of the 2001 Workshop on Particle Swarm Optimization.*

Chetty, R. 2012. *Time trends in the use of administrative data for empirical research*. Presentation to NBER Summer Institute. http://www.rajchetty.com/chettyfiles/admin_data_trends.pdf

Clerc, M. 1999. The swarm and the queen: towards a deterministic and adaptive particle swarm optimization, *Proceedings of the 1999 Congress on Evolutionary Computation-CEC99:* 1951-1957.

Cole, S., I. Dhaliwal, A. Sautmann, & L. Vilhuber. 2020. *Handbook on Using Administrative Data for Research and Evidence-based Policy*. https://admindatahandbook.mit.edu/book/v1.0-rc5/index.html.

Conger, D. 2015. Foreign-born peers and academic performance, *Demography* 52(2): 569-592.

Connelly, R., C. J. Playford, V. Gayle, & C. Dibben. 2016. The role of administrative data in the big data revolution in social science research, *Social science research* 59: 1-12.

Consumer Financial Protection Bureau. 2014. *Using Publicly Available Information to Proxy for Unidentified Race and Ethnicity: A Methodology and Assessment*. Washington, DC: Consumer Financial Protection Bureau. https://files.consumerfinance.gov/f/201409_cfpb_report_proxy-methodology.pdf

Cuccaro-Alamin, S., A. L. Eastman, R. Foust, J. McCroskey, H. T. Nghiem, E. Putnam-Hornstein. 2021. Strategies for constructing household and family units with linked administrative records, *Children and Youth Services Review* 120.

Eberhart, R., & J. Kennedy. 1995. A new optimizer using particle swarm theory, *Proceedings of the Sixth International Symposium on Micro Machine and Human Science*: 39-43.

Eberhart, R.C., & Y. Shi. 2000. Comparing inertia weights and constriction factors in particle swarm optimization, *2000 Congress on Evolutionary Computing* 1: 84-88.

Elliott, M. N., P. A. Morrison, A. Fremont, D. F. McCaffrey, P. Pantoja, P., & N. Lurie. 2009. Using the Census Bureau's surname list to improve estimates of race/ ethnicity and associated disparities, *Health Services and Outcomes Research Methodology* 9(2): 69–83.

Ernsten, A., D. McCollum, Z. Feng, D. Everington, & Z. Huang. 2018. Using linked administrative and census data for migration research, *Population Studies 72*(3): 357-367.

Figlio, D., K. Karbownik, & K. Salvanes. 2017. The promise of administrative data in education research, *Education Finance and Policy* 12(2): 129-136.

Foster, J. L., D. Glass, C. Salazar, & M. Tamene. 2021. Geography, equity, and the Seattle $15 Minimum Wage Ordinance. Data Science for Social Good 2021 Project Results, University of Washington eScience Institute.

Gath, M., & C. Bycroft. 2018. The Potential for Linked Administrative Data to Provide Household and Family Information. Wellington, New Zealand: Stats NZ.

Gibson-Davis, C. M., E. O. Ananat, & A. Gassman-Pines. 2016. Midpregnancy marriage and divorce: Why the death of shotgun marriage has been greatly exaggerated, *Demography* 53(6): 1693-1715.

Goldschmidt, D., W. Klosterhuber, & J. F. Schmieder. 2017. Identifying couples in administrative data, *Journal for Labour Market Research* 50(2): 29-43.

Goroff, D., J. Polonetsky, & O. Tene, O. 2017. Privacy Protective Research: Facilitating Ethically Responsible Access to Administrative Data, *The ANNALS of the American Academy of Political and Social Science* 675(1): 46-66.

Grippo, F., A. Désesquelles, M. Pappagallo, L. Frova, V. Egidi, & F. Meslé. 2020. Multi-morbidity and frailty at death: A new classification of death records for an ageing world, *Population Studies* 74(3): 437-449.

Harron, K., C. Dibben, J. Boyd, A. Hjern, M. Azimaee, M. L. Barreto, & H. Goldstein. 2017. Challenges in administrative data linkage for research, *Big Data & Society* 4(2): 1-12. doi:10.1177/2053951717745678.

Hawn Nelson, A., D. Jenkins, S. Zanti, M. Katz, T. Burnett, D. Culhane, K. Barghaus, et al. 2020. Introduction to Data Sharing and Integration. *Actionable Intelligence for Social Policy*. University of Pennsylvania.

Jenkins, D., E. Berkowitz, T. Burnett, D. Culhane, A. Hawn Nelson, K. Smith, & S. Zanti. 2021. Expanding Mobility: The Power of Linked Administrative Data for Spatial Analysis. *Actionable Intelligence for Social Policy*. University of Pennsylvania.

Kabátek, J., & F. Perales. 2021. Academic achievement of children in same-and different-sex-parented families: A population-level analysis of linked administrative data from the Netherlands, *Demography* 58(2): 393-418.

Kornfeld, R., & H. S. Bloom. 1999. Measuring program impacts on earnings and employment: Do unemployment insurance wage reports from employers agree with surveys of individuals? *Journal of Labor Economics* 17(1): 168-197.

Monti, A., S. Drefahl, E. Mussino, & J. Härkönen, J. 2020. Over-coverage in population registers leads to bias in demographic estimates, *Population Studies* 74(3): 451-469.

National Center for Health Statistics. 2021. National Vital Statistics System Improvements (NCHS Fact Sheet). National Center for Health Statistics, Office of Planning, Budget, and Legislation. https://www.cdc.gov/nchs/data/factsheets/2020-NVSS-improvement-factsheet-508.pdf

Neumark, D., and W. L. Wascher. 2007. Minimum wages and employment, *Foundations and Trends in Microeconomics* 3(1-2):1-182.

Penner, A.M., & K. A. Dodge. 2019. Using administrative data for social science and policy, *RSF: The Russell Sage Foundation Journal of the Social Sciences* 5(2): 1-18.

Romich, J., M. Long, S. Allard, & A. Althauser. 2018. The Washington State Merged Longitudinal Administrative Database (Abstract). Administrative Data Research Facilities Network Conference Proceedings. *International Journal of Population Data Science 3*(5). doi.org/10.23889/ijpds.v3i5.1066.

Ruggles, S., S. Flood, R. Goeken, J. Grover, E. Meyer, J. Pacas, & M. Sobek. 2020. *IPUMS USA: Version 10.0 [dataset]*. Minneapolis, MN: IPUMS. https://doi.org/10.18128/D010.V10.0.

Social Wellbeing Agency. 2020. *Constructing households from linked administrative data: An attempt to improve address information in the IDI*. Wellington, New Zealand: Social Wellbeing Agency.

Stromsdorfer, E. W. 2007. Agricultural Workforce in Washington State 2006. https://migration.ucdavis.edu/rmn/more.php?id=1255.

U.S. Census Bureau. n.d. *Center for Economic Studies Publications and Reports Page*. Available: https://lehd.ces.census.gov/ (accessed: 18 December 2021).

Vilhuber, L., & K. L. McKinney. 2014. LEHD infrastructure files in the Census RDC – Overview. US Census Bureau Center for Economic Studies Paper No. CES-WP-14-26. https://ideas.repec.org/p/cen/wpaper/14-26.html

Wallgren, B., & A. Wallgren. 2014. *Register-Based Statistics: Statistical Methods for Administrative Data*. New York: John Wiley & Sons.

Appendix A. Demographic data generation and imputation

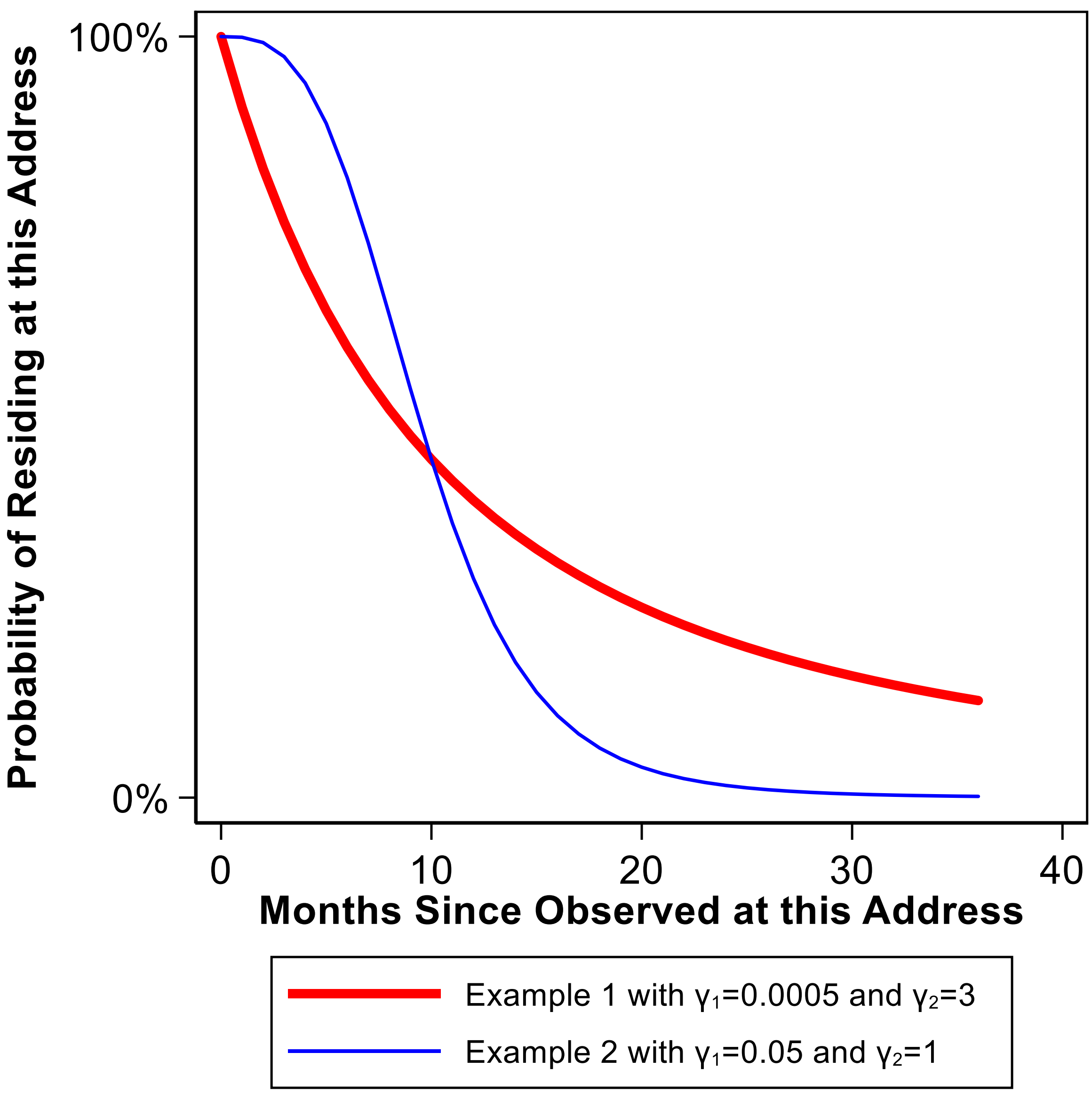
For each individual in the address data, we use administrative records to determine age, sex, and race/ethnicity. Age and sex information is drawn from DOL, SOS, DSHS, and DOH. Residents self-report ethnorace information in DSHS client records and birth parent records. We impute race and ethnicity for an additional subset of the population by combining information on residential location and last name using the Bayesian Improved Surname Geocoding (BISG) method (Elliot et al., 2001; Consumer Financial Protection Bureau, 2014). We standardize each data source to comprise seven ethnorace categories: White alone, non-Hispanic (hereafter referred to as "White"); Black alone, non-Hispanic (hereafter "Black"); Native American/American Indian or Alaska Native alone, non-Hispanic (hereafter "AIAN"); Asian or Pacific Islander alone, non-Hispanic (hereafter "API"); Multiracial or some other race, non-Hispanic (hereafter "Multiracial/Other"); and Hispanic or Latino, any race (hereafter "Hispanic").

The BISG method uses Bayes' formula to estimate the probability an individual belongs to a given ethnoracial group by combining information on the person's surname and their census block of residence. First, the probability that an individual belongs to a given ethnoracial group r given their surname s, $p(r|s)$, is estimated using Census data on the ethnoracial distribution of the population with a given surname. Next, the proportion of people of a given group that live in a given geographic area g, $q(g|r)$, is calculated using block-level data from the 2010 Decennial Census and the individual's most recent known address across WMLAD address data sources. Applying Bayes' theorem, the probability that a given individual belongs to group r is $\Pr(r|g,s) = \frac{p(r|s)*q(g|r)}{\sum_{r \in R} p*q}$.

When we compare ethnorace information in reported WMLAD data (from the DSHS database and the DOH records) to the BISG imputation results, among a subset of individuals for whom we have both reported and imputed ethnorace data, the overall predicted distribution of ethnoracial groups matches the distribution in the reported data quite well. The composition of the population according to the reported data, with the BISG imputed breakdown in parentheses, is as follows: White 68% (65%); Black 5% (5%); AIAN 1% (1%); API 7% (7%); Hispanic 15% (17%); Multiracial/Other 4% (4%). However, the accuracy of the imputation varies widely by group. For example, among persons reported to be White in the DOH/DSHS data, the BISG method generates a predicted probability of being White of 87%, on average. In contrast, among persons reported to be Black, the BISG method predicts their probability of being Black to be only 31%, on average. The BISG method is most accurate for White, API, and Hispanic individuals, and less accurately imputes ethnorace for Black, AIAN, and Multiracial/Other individuals. When we use these imputed ethnorace data to impute residential address histories, fitting our WMLAD population to totals from the ACS, the inaccuracies generated in the BISG imputation process will create attenuation bias. For example, when we impute differences in residential persistence at a given address based on ethnorace (described below), inaccuracies in imputing ethnorace generated by the BISG approach will attenuate the differences in the estimated parameters across ethnorace, relative to the true parameter differences.

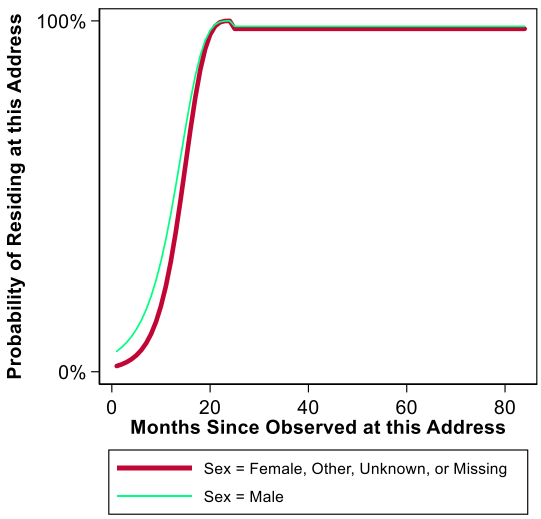
If reported demographic data conflicts across or within data sources, we choose a single value as follows. First, we collapse the records within a data source by person and demographic characteristic. For example, a person with two values of age within a particular dataset would have two rows for this dataset after collapsing. Then, we append the data sources together and identify the modal value. For age, if there is no unique modal response, we take the average if this range is less than or equal to 5. If there is no unique modal response – and, in the case of age, the range is greater than 5 – we then prioritize, in order, information from DOL, SOS, DSHS, or DOH. If there were multiple different values reported in the highest-priority source, we prioritize more recent observations from that source. Imputed ethnorace is used if no reported ethnorace information is available.

Appendix Figure 1. Illustration of the possible decay in the predicted probability of residing at a particular observed address as a function of the gamma parameters.



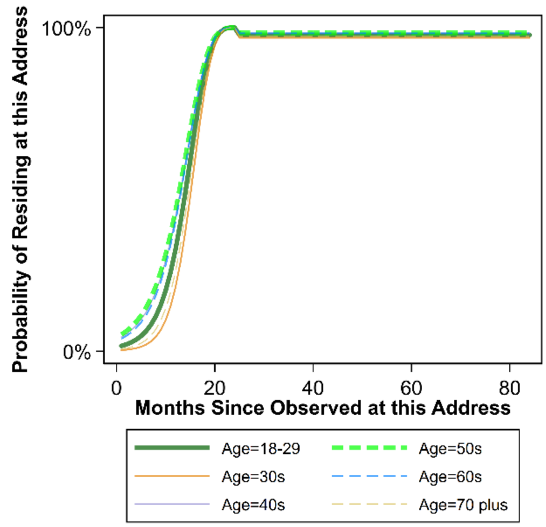
Appendix Figure 2 Illustration of the decay in the predicted probability of residing at an address that is observed in month 24 as a function of the beta parameters.

Panel A: By Sex



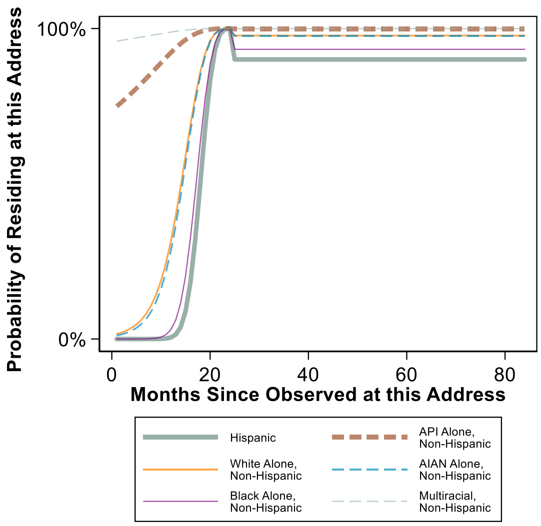
Assuming Age 18 to 29 and White Alone, Non-Hispanic

Panel B: By Age



Assuming Female and White Alone, Non-Hispanic

Panel C: By Race



Assuming Female and Age 18 to 29