Using SSNs to Identify Immigration, Selection, and Censoring in Administrative Data*

John Klopfer and Corbin Miller

October 22, 2024

Abstract

Most administrative data do not identify the foreign-born population or the time and place of immigration, but contain Social Security Numbers (SSNs). We present a new crosswalk identifying issue months and states for every SSN through 2011, and a new foreign-birth probability table, to accurately and precisely impute immigration months and states, and foreign birth, for the U.S. population. Our imputations may be used to correct for related selection and censoring in administrative data. Both can be linked to government and commercial administrative data without SSA approval.

1 Introduction

Administrative data allow researchers to measure population economic outcomes, but often need to be supplemented with demographics from other sources. Major contributions rely on linked childhood residence (*e.g.* Chetty, Hendren, and Katz, 2016 and Isen, Rossin-Slater, and Walker, 2017); parent-child relationships (*e.g.* Chetty *et al.*, 2014 and Barr, Eggleston, and Smith, 2022); and race (*e.g.* Chetty *et al.*, 2020 and Greenberg *et al.*, 2022).

Most administrative data do not report immigration or foreign birth. Social Security Administration (SSA) records are necessary to identify the foreign-born resident population (Abowd, McKinney, and Zhao, 2018; Abowd et al., 2020). Since foreign-born residents and non-residents hold a fifth of Social Security Numbers (SSNs), immigration and emigration are also major sources of selection and censoring, and therefore bias (Heckman, 1976 and

*Klopfer: Faculty of Business and Economics, The University of Hong Kong; Miller: Office of Research, Applied Analytics, and Statistics, Internal Revenue Service. Manuscript copyright: John Klopfer; not in the public domain (see 17 USC §101, 105, and 201). Code and data copyright: John Klopfer. All rights reserved. Feedback is appreciated. Correspondence: johnklopfer@gmail.com

We thank Casey Breen, Joshua Goldstein, Alan Kwan, Gedeon Lim, Ithai Lurie, Paul Organ, Uta Schoenberg, David Splinter, Heidi Williams, and Constantine Yannelis for help and advice; and Siyao (Julie) Lu for excellent research assistance. Klopfer thanks the Hoover Institution for its hospitality, and the University of Hong Kong and the Research Grants Council of Hong Kong (GRF 17500922) for their support. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors and do not reflect the views or the official positions of the IRS or any of our colleagues.

1979) in estimates from administrative populations.[1] Thus, even when immigration is not of direct interest to the researcher, it is not avoidable. Most methods used to address bias require linked SSA data.[2] Yet, SSA rarely approves linkage.[3] What should researchers do, when they cannot link SSA records to identify immigration or foreign birth?

In this paper, we provide a crosswalk and probability table that researchers may use to impute issue months and foreign birth for any SSN, expanding access to linked population records for immigration and administrative data researchers. Our crosswalk resolves back-dating that we identify in other crosswalks, by up to five years for millions of SSNs. Our probability table, compared to imputations of foreign birth based on crosswalks alone, eliminates bias and halves root mean squared error (RMSE). Both recover tens of millions of missing values in the original SSA records. Neither requires SSA approval to link to other government and commercial administrative data.

Our crosswalk reports issue months and states for SSNs based on their first five digits. We obtain issue months from SSN account records for deceased individuals. Over half of issue months are within a month of the crosswalk median, and more than nine-tenths are within three months. Our crosswalk can be used to impute immigration months: issue months track reported Hungarian, Russian, and Salvadoran immigration from the 1950s to the 2000s.[4] Our crosswalk can also be used to exclude not-yet-issued SSNs from administrative populations, to link and clean data based on known relationships between the SSN issue date and other dates of interest, and to impute the issue age from birth months.

Our probability table reports the likelihood of foreign birth based on the first five digits of the SSN, sex, and birth month. We estimate probabilities using SSN account records linked to our crosswalk, in cells defined by sex and the SSN issue state, year, and age. Because we rely on records for deceased individuals, and deaths are reported for the resident population, probabilities represent the resident population. Our probability table can be used to impute foreign birth using a variety of standard methods. Our probability table can also be used to exclude foreign-born SSNs from administrative populations or to correct for selection and censoring, and to link and clean data based on foreign birth.

We are the first to provide an issue-month crosswalk, and to provide a crosswalk based on SSN account records. Compared to issue-year crosswalks pioneered by Block, Matanoski, and Seltser (1983) and based on SSA documentation, our crosswalk also eliminates back-dating, overcoverage, and undercoverage. By improving accuracy and precision, we advance efforts to impute immigration years from SSN issue years (epidemiologists Mack *et al.*, 1985; Shimizu *et al.*, 1987 and 1991; and Lauderdale *et al.*, 1997) and to impute childhood exposures

[1] Immigrants respond to incentives (Borjas, 1999; Agersnap, Jensen, and Kleven, 2020; Kleven *et al.*, 2020) and return migrate (Lubotsky, 2007; Rho and Sanders, 2021; Akee and Jones, 2024). The bias is like that from internal migration (Foote and Stange, 2022) but cannot be addressed just by using national data.

[2] Immigration and foreign birth inform parametric models (Heckman, 1976 and 1979; Hausman and Wise, 1979), bounds (Manski, 1990; Horowitz and Manski, 2000; Lee, 2009), imputations (Little and Rubin, 1987), or simply excluding the foreign-born population.

[3] Except with internal access at SSA and Census, we could not find immigration research using population outcomes linked to SSA records. Research in Denmark (Foged and Peri, 2016), Germany (Dustmann, Schoenberg, and Stuhler, 2017), and Canada (Dostie *et al.*, 2023) shows the value of broader access.

[4] Rho and Sanders (2021) compare issue years in SSN application records with self-reported arrival years in the 2000 census long form, and argue that SSN issue years better represent immigration. Mehta *et al.* (2016) and Engelman *et al.* (2017) proxy immigration with issue years, at SSA, without validation.

from SSN issue years and states (epidemiologists Mack and Floderus, 1991 and Lauderdale, Thisted, and Goldberg, 1998 who study environmental exposures; and sociologists Jin *et al.*, 2010 and economists Yonker, 2017; Pool, Stoffman, and Yonker, 2012; Bai *et al.*, 2019; and Hu, 2022 who study social ties, home bias, and policy exposures).

We are the first to provide a foreign-birth probability table based on SSN account records, and to impute foreign birth by sex and the SSN issue state, year, and age. Compared to imputations from SSN issue-age cutoffs alone, pioneered by Lauderdale *et al.* (1997), our probability table reduces mis-classification of children as native-born and adults as foreign-born; it reduces the RMSE of imputation by 22, 13, and 50 percent for the population aged 0-25, 25-55, and 55-85 in 2007, and more for earlier cohorts. The disparity would be larger if using issue-year crosswalks. By improving fit, we advance efforts to impute foreign birth from SSN issue years (epidemiologists Raz *et al.*, 2008; Gomez *et al.*, 2010; Keegan *et al.*, 2010 and others who study cancer risk; and economists Yonker, 2017; Bernstein *et al.*, 2022; and Amornsiripanitch *et al.*, 2023 who study CEOs, inventors, and founders).

Our crosswalk and probability table are open-source, and can be refined according to the needs of specific research projects.

Together, our crosswalk and probability table advance research in administrative data that are widely used without linkage to SSA records. Sources include the IRS Data Master (DM1) file, tax returns, and information returns; state unemployment insurance (UI) data;[5] the National Student Clearinghouse (NSC) and more than half of Statewide Longitudinal Data Systems (SLDS); data from the Centers for Medicare and Medicaid Services (CMS) and the Veterans Administration (VA); the National Cancer Institute (NCI) Surveillance, Epidemiology, and End Results (SEER) registries and the Centers for Disease Control (CDC) National Program of Cancer Registries (NPCR); health care providers and insurers; employers and payroll and benefits processors;[6] banks, brokerages, and consumer credit reporting agencies (CRAs);[7] and information brokers including Infutor and LexisNexis.

For the IRS in particular, a lack of data sharing between SSA and the IRS, Treasury, Joint Committee on Taxation (JCT), and Congressional Budget Office (CBO) hinders research and policy-making on immigration and the tax base (see Blau and Mackie, 2017; GAO, 2022; Clemens, 2022; Elmendorf and Williams, 2023 and 2024; and CBO, 2024). Our imputations enable tax policy and tax administration work using IRS data not shared externally.

Our imputations address two major gaps in the literature on intergenerational mobility: by allowing researchers to identify immigrant parents and grandparents, linked to children through dependent claiming in the IRS data, we make it possible to follow second- and third-generation immigrants (Massey, 2010; Grusky and Snipp, 2013; Blau and Mackie, 2017) and to distinguish racial and ethnic groups[8] by generation (Edmonston, 1996).

Our imputations also address gaps in the literature on employment-based immigration:

[5]Hotchkiss, Quispe-Agnoli, and Rios-Avila (2015) identify undocumented workers in state UI data based on invalid SSNs. Our probability table could identify native- and foreign-born documented workers.

[6]See employer (Hall and Krueger, 2018), payroll (Cajner *et al.*, 2020; Autor *et al.*, 2022; Chetty *et al.*, 2024), and benefit (Madrian and Shea, 2001; Choi *et al.*, 2002; Thaler and Benartzi, 2004) data.

[7]Gibbs *et al.* (forthcoming) review the vast literature using CRA data; also see card issuer (Agarwal, Liu, and Souleles, 2007; Mian, Rao, and Sufi, 2013) and bank (Cox *et al.*, 2020; Ganong *et al.*, 2024) data.

[8]Our probability table complements ongoing efforts to impute race and ethnicity into administrative data at IRS, CMS, and elsewhere (*e.g.*, Haas *et al.*, 2019; Imai, Olivella, and Rosenman, 2022; Fisher, 2023).

U.S. Citizenship and Immigration Services (USCIS) H-1B petitions can be linked by firm to commercial (Dimmock, Huang, and Weisbenner, 2021; Mayda *et al.*, 2023), IRS (Doran, Gelber, and Isen, 2022), and BEA (Glennon, 2024) data, but do not identify which employees received a visa.[9] Our imputations of foreign birth and issue month may identify recipients.

2 Related Literature

We contribute to a broader immigration literature that has been built almost entirely on restricted data at SSA and Census. Though influential, the literature is shaped by the access and data limitations we address with this paper.

SSA keeps the most complete record of the foreign-born resident population, linked to a valuable but narrow range of longitudinal earnings and benefit records. Research in population records has been limited to studies on immigrant disability (Engelman *et al.*, 2017) and mortality (Mehta *et al.*, 2016 and others) furthering SSA's actuarial work.[10] Research on immigrant earnings, benefits, and return migration has been limited to 1-percent samples (Lubotsky, 2007; Kopczuk, Saez, and Song, 2007; Schwabish, 2011; Vega, 2015). The challenges at SSA are access, linkages, and capacity; and missing data (Levine, Hill, and Warren, 1985) that may be addressed by our crosswalk and probability table.

Census internal research has addressed some of SSA's challenges. Abowd, McKinney, and Zhao (2018), Abowd *et al.* (2020), Rho and Sanders (2021), Azoulay *et al.* (2022), and Miller, Wherry, and Aldana (2023) use full access to SSA data on immigrants to exclude future immigrants from panel data, characterize the foreign-born population, correctly time immigrant arrivals, and study small populations including company founders and mothers of newborns. We could not find external research given the same access.

Census external research faces greater limitations. Those using LEHD demographics (*e.g.* Kerr, Kerr, and Lincoln, 2015) may access the place of birth, with SSA approval, but not the SSN issue date; because LEHD UI earnings data do not cover all states, do not cover all employers, and do not cover non-employees, they present special challenges when mobile immigrants are the focus of research. Those using census long form or ACS demographics (*e.g.* Kerr and Kerr, 2017; Chetty *et al.*, 2020; Abramitzky *et al.*, 2021) are restricted to 1-in-6 or 1-in-100 cross sections of the population, which do not fully cover employees of each firm, residents of each neighborhood, or members of each family.

[9] Clemens (2013) offers a rare example where H-1B visas can be tied to individual workers, using records from a multinational employer. Individual-level USCIS data are not linked to other individual-level administrative data, with the exception of a now-discontinued project at Census under a rescinded executive order (Abowd et al, 2020; Brown and Murray-Close, 2023, footnote 13). USCIS does release anonymized data: researchers may link H-2B visas by firm (Ameudo-Dorantes *et al.*, 2024; Clemens and Lewis, 2024), and other data including green card petitions by geography (Kaushal, 2005; Saiz, 2007; Cascio and Lewis, 2019).

[10] SSA has also released certain records for deceased individuals, enabling external mortality research. For example, Breen and Goldstein (2022).

3 Issue-Month Crosswalk

Our crosswalk reports issue months and states for SSNs based on their first five digits. It helps to understand how nine-digit SSNs are structured: the format is AAA-GG-SSSS, where AAA is the "area number," GG is the "group number," and SSSS is the "serial number".

We rely on the fact that until mid-2011, SSNs were issued in blocks defined by the area and group numbers, which we call "five-digit SSNs." For each five-digit SSN, we summarize the issue months for its serial numbers: minimum, median, and maximum.

3.1 Building the Crosswalk

We obtain issue months from account records in the SSA Numident file, released to the National Archives and Records Administration (NARA, 2013), for account holders deceased by 2007. Issue months are stored in the "cycle date" of the original application entry (Panis *et al.*, 2000). We account for missing entries and edited cycle dates.

From 2004 onwards, when few account records are available, we rely on the SSA "high group lists" of five-digit SSNs issued each month. Issue months start from the month in which each five-digit SSN is first reported, and end in the month in which the next five-digit SSN is reported. We account for gaps in the reporting, and for delayed reporting.

Our crosswalk differs from crosswalks based on SSA documentation. The SSN account records show that the sequence of issuance for five-digit SSNs reported in the documentation is not the true sequence. We identify and correct for issuance back-dated in other crosswalks by 5-8 years, from 1977-80 to 1972-73,[11] for roughly 2.0 million SSNs.

The deviation from the reported sequence relates to the transition from District Office to Central Office issuance in 1973. In the transition, SSA District Offices left unissued serials within issued five-digit SSNs, and unissued area numbers within issued group numbers (SSA Claims Manual, 1981, Sec. 2891). The SSN application records show that the Central Office later backfilled these serials and area numbers in 1977-80.

The deviation from the reported sequence is not trivial. We find backfilled serial numbers in 111 five-digit SSNs in total, back-dating roughly 0.2 million SSNs, one-half of which were issued in only six states. Our median correctly dates the majority of these serial numbers. We find backfilled area numbers in 183 five-digit SSNs in total, back-dating 1.8 million SSNs, one-third of which were issued in only six states.

Appendix A.1 presents additional details of how we built the crosswalk.

3.2 Assessing the Crosswalk

To assess our crosswalk, we compare it to the SSN account records, and to issue-year crosswalks used in the literature.

Our crosswalk is accurate: its range covers 99.81 percent of recorded issue months in the account records. Our crosswalk is precise: Figure 1 shows that more than half of issue months

[11] Back-dating from 1977-80 to 1972-73 affects years of immigration imputed from other crosswalks, making them appear earlier than they are; and affects foreign birth imputed from other crosswalks, making immigrants issued SSNs in their 20s appear to be natives issued SSNs in their teens. Back-dating is concentrated by state, issue year, and birth cohort, giving these outliers leverage.

are within a month of the median, and more than nine-tenths are within three months. After 1972, more than nine-tenths of issue months are within a month of the median.

Issue-year crosswalks are necessarily less precise than our issue-month crosswalk, but surprisingly, they are also less accurate. We consider the accuracy of the Block, Matanoski, and Seltser (1983; BMS) crosswalk, and the SSN-Verify crosswalk used by Amornsiripanitch *et al.* (2023) and Bernstein *et al.* (2022): only 96.89 percent of issue months are within the range given by the BMS crosswalk, and only 91.33 percent of issue months are within the range given by the SSN-Verify crosswalk. Figure 1 shows that the BMS and SSN-Verify crosswalks are less accurate throughout the entire period, but particularly in 1936-49 when SSA did not publish high group lists, in 1977-80 due to back-dating, in 1989-92 due to missing high group lists, and from 2002 onward for reasons we could not determine.

3.3 Using the Crosswalk

Our crosswalk provides issue months for all of the 466 million SSNs that were issued before SSN randomization in 2011.

3.3.1 Issue Months

The basic use of our crosswalk is to impute issue months where they are not available. Our crosswalk makes issue months available without SSA approval. Our crosswalk also helps to:

Recover Issue Months at SSA.—We estimate that 86 million SSNs have missing or invalid issue dates at SSA due to lost records, data entry error, and editing. Our crosswalk can be used to recover missing issue months in the SSA Numident.

Define Populations.—SSNs are often used to define analysis populations. Given an observation date, the best practice is to include only SSNs issued by that date (Abowd, McKinney, and Zhao, 2018). Our crosswalk makes it possible to correctly exclude SSNs from the population, by month, quarter, or year.

Consider an analyst who wants to calculate mean income for tax year 1996 using Form 1040, but to define the population has only the list of 466 million SSNs issued by 2011: because only 386 million SSNs had been issued by 1996, 80 million SSNs have no filed Form 1040 because they did not exist in 1996. Coding missing incomes as zeroes would bias the mean downward by a factor of 0.83. Since only 120 million SSNs were used to file Form 1040 in 1996, indiscriminately excluding all non-filers could bias the mean upward even more.

Consider an analyst who wants to calculate the fraction of children claimed as dependents by Form 1040 filers in 1996: more than 3.5 percent of SSNs issued by 2011, with age under 19 in 1996, were issued after 1996; SSNs that did not exist in 1996 may account for more than half, and maybe all, of the unclaimed child SSNs reported in that year and subsequent years by Chetty *et al.* (2014) and Chetty and Hendren (2018).

Link Data.—SSNs may also be used to define linkage populations, either narrowing or confirming matches. SSNs issued in a given state and month indicate individuals likely to be found in that state and month. Records related to birth, ITINs, visas, immigration, and work authorization are created before SSN issuance; and records related to public services and benefits, licensing, financial accounts, and death are created after SSN issuance.

Clean Data.—Issue months from our crosswalk can be used to identify incorrect dates in the Numident and other data sources: death dates occurring before the minimum issue month and birth dates occurring after the maximum issue month are due to data entry error. In the Numident, cycle dates of original entries should fall between the minimum and maximum issue month unless the entries are edited.

Calculate Issue Ages.—The issue month can be used with the birth month to define the issue age. The issue age predicts foreign birth, as discussed below, and return migration. From 1938-62 (and often 1972) the issue age was the age at first covered employment.

3.3.2 Immigration Months

When the place of birth is known, SSN issue months from our crosswalk may be used to infer immigration months.

The following case studies from 1956 to 2001 show that issue months tightly track work-authorized immigration months, as does our crosswalk. Issue months do not equate to entry months, whether lawful or unlawful.

2001—El Salvador earthquakes and TPS designation On January 13 and February 13, 2001, two major earthquakes displaced 17 percent of El Salvador's population, leaving many in desperate need of remittances. Attorney General Ashcroft (2001) designated Salvadorans for Temporary Protected Status (TPS), with an 18-month registration period from March 9, 2001 through September 9, 2002: Salvadorans resident in the United States lawfully or not from February 13, 2001 became eligible to apply for TPS and for an Employment Authorization Document (EAD). The Immigration and Naturalization Service (INS) was required to issue the EAD within 90 days (Reno, 1994, p. 62303; GAO, 2001, p. 29, 99). The EAD is needed to apply for a SSN (SSA, 2023), resulting in a lag between the start of the TPS registration period and the first SSN issuance.

In May 2001, two months after Salvadorans became eligible to apply for employment authorization, the NARA Numident shows SSN issuance to Salvadorans rose six-fold from the level seen from December 2000 through February 2001. By July it was ten-fold. Issuance did not return to baseline levels until September 2002, the end of the registration period.

Figure 2 shows that known SSN issue months and median issue months from our crosswalk, for the 2001 El Salvador TPS designation, have a nearly identical time pattern.

1956—Hungarian Uprising and the Hungarian Escapee Program From October 23 to November 11, 1956, Hungarians overthrew their Soviet puppet government, withdrew from the Warsaw Pact, and suffered a crushing defeat by the Red Army. 194,000 Hungarians fled to Austria and Yugoslavia before the USSR sealed the frontier in February 1957. With nowhere to house the Hungarians, and with the host countries' neutrality in question, friendly governments "moved refugees abroad as quickly as possible, generally without proper 'screening'," (Ferenc, 1999; Casardi, 1957, p. 3-4, 8, 12-13).

On November 8, President Eisenhower "directed that extraordinary measures be taken," INS officers detailed to Vienna issued the first visas on November 19, and on November 21 the first planeload of Hungarians arrived at Camp Kilmer, NJ. The first 5,000 arrived by December 1. Operation Safe Haven, an airlift and sealift, brought another 15,000 by January

2, 1957 and 12,373 more by February 28. By May 1, Camp Kilmer had welcomed 32,000 Hungarian arrivals; most stayed less than a month, and "Once they passed their inspections, eligible adults received. . . a social security card." On December 31, with 38,000 admitted, the program ended (H.R. Journal, No. 85-1, p. 163; Swing, 1958, p. 43; NY Times, 1957, p. 5; Markowitz, 1973, p. 56; Eisenhower, 1958, p. 92-93; USCIS, 2022).

The NARA Numident shows SSN issuance to Hungarian-born individuals rose immediately in December 1956 and reached a peak in January 1957; the second-highest month of issuance was February, and the fourth and fifth were March and April. The magnitudes are consistent with known arrivals in December of three times those in late November and likely two times those in January, if SSN issuance is lagged by a few weeks.

Figure 2 shows that known SSN issue months and median issue months from our crosswalk are similar but not identical: the peak is one month late, the second-largest is two months late, and the onset is one month early. The median assigned to some five-digit SSNs falls in September or October 1956, creating a weak pre-trend before the onset of operations in November. Yet, even in 1956, SSN issuance can reliably be dated to within 1-2 months.

Supplementary Cases Appendix A.2.1 presents cases of immigration from El Salvador in 1991 and from Russia in 1977. SSNs track immigration similarly in 1977, 1991, and 2001.

Limitations Appendix A.2.2 discusses limitations to the use of issue months as immigration months before 1972, for those not working in the covered sector. We discuss solutions.

4 Foreign-Birth Probability Table

Our probability table presents the probability of foreign birth conditional on sex and issue state, year, and age from our crosswalk.

4.1 Building the Probability Table

To calculate the probability of foreign birth, we collect key data for each SSN in the NARA Numident: foreign birth, sex, and birth month are edited and imputed as necessary, with limited impact on the results. Appendix A.3 offers a detailed description.

To calculate the probability, we also assume that observations are missing at random (MAR; see Little and Rubin, 1990) conditional on sex and issue state, year, and age. That is, SSN records included in the NARA Numident should be drawn at random from, and therefore represent, the population of interest for imputation.

The most important threat to the MAR assumption is that SSA Numident records are included in the NARA Numident because the account holders are known to be deceased.

We address this threat by implicitly conditioning on mortality risk. Alongside sex and state, the core risk factors for recorded mortality are the issue year, which determines the years in which mortality could be recorded; and the issue age, which determines the annual risk of mortality conditional on survival. The issue year and age also imply the birth year, which captures cohort effects on mortality. Together, these risk factors account for the key

differences in mortality risk by foreign birth, since immigrants enter in different years, at different ages, from different birth cohorts.

The NARA Numident represents the resident population of SSN holders, because it represents the recorded-deceased, and deaths are recorded primarily for residents. The resident population is the relevant population for most imputations.

4.2 Assessing the Probability Table

To assess our probability table, we compare it to the SSN account records, and to issue-age cutoff imputations used in the literature. We start by examining the pattern of foreign birth, by issue year and age, shown in the probability table.

Figure 3 shows that native- and foreign-born SSN account holders are well-differentiated in many issue years and ages. This is due in part to events that include reduced immigration from 1921-46 due to quotas, the Great Depression, and World War II; enumeration of welfare beneficiaries in 1973-74; expanded enumeration of immigrants in 1980-89; and the rollout of enumeration at birth in 1989-97. The probability of foreign birth is affected by current enumeration, but also by past enumeration exhausting native- or foreign-born eligibles.

Figure 4 shows, however, that many SSNs are not issued in the years and ages where account holders are well-differentiated. SSNs are issued to native- and foreign-born applicants at all ages in all years, with overlapping peaks. From 1936, workers obtain SSNs as coverage expands (Nelson, 1985) and as they enter covered jobs; from 1962, non-workers obtain SSNs for federal taxes and benefits; and from 1987-1991 and onwards taxpayers obtain SSNs for children, at increasingly young ages, to claim tax benefits.

These facts matter for foreign-birth imputation. Figure 5 compares two alternatives, by their mean squared error (MSE) relative to SSN account records, by issue year and age. The first alternative, imputation by issue-age cutoff alone, has been used widely in epidemiology, demography, and economics and finance, starting with Lauderdale *et al.* (1997). The second alternative, imputation by our probability table, better fits the SSN account records.

Figure 5, Panel A shows imputation by issue-age cutoff alone. Imputing native birth to SSNs with an issue age below 20, and foreign birth otherwise, yields high rates of misclassification: we find that those aged 25-55 and 55-85 in 2007 were classified as not foreign in 47 and 17 percent of cases where they were foreign (false negatives), and classified as foreign in 0.5 and 10 percent of cases where they were not foreign (false positives). Those aged 25-55 and 55-85 in 2007 were not foreign in 11 and 61 percent of cases classified as foreign (false discoveries), and were foreign in 4 and 1 percent of cases classified as not foreign (false omissions).[12] It is not possible to choose a cutoff age that eliminates misclassification.

Figure 5, Panel B shows imputation by our probability table. The side-by-side comparison

[12]We are the first to estimate rates of misclassification in SSN account records. Lauderdale *et al.* (1997) and Lauderdale, Thisted, and Goldberg (1998) noted the potential for misclassification; later research in epidemiology showed that the false positive and false negative rates were roughly 20 percent in birth cohorts from 1925-1962 observed in 1988-2009 (Raz *et al.*, 2008; Gomez *et al.*, 2010; Keegan *et al.*, 2010; Khan *et al.*, 2013; and Gomez *et al.*, 2015). The economists and demographers who have used the issue-age approach have not generally provided validation, due to their reliance on data sources that do not report foreign birth. An exception, Bernstein *et al.* (2022) find in the ACS that a cutoff at age 20 would mis-classify 39 percent of immigrants as native (false negatives). However, the ACS does not speak to other forms of misclassification.

makes it clear that our probability table reduces misclassification, measured by the MSE, in all issue years and ages. In absolute terms, the MSE is roughly 0.03 (root mean squared error, RMSE, is roughly 0.10) in most years and ages, including all peak years and ages; the MSE is better than both random chance and the issue-age cutoff imputation. In relative terms, our probability table reduces the RMSE of imputation by 22, 13, and 50 percent for the population aged 0-25, 25-55, and 55-85 in 2007, and more for earlier cohorts.

Our probability table outperforms issue-age cutoffs for two reasons. First, it accounts for uncertainty. Second, it improves differentiation, by accounting for variation in the probability of foreign birth jointly by sex and the issue state, year, and age. Due to changing policies and patterns of immigration, these four variables are highly informative. Figures A7, A8, A11, and A12 show that there is strong differentiation by sex and issue state, and also excellent classification within each sex and issue state.

4.3 Using the Probability Table

Our probability table provides the probability of foreign birth conditional on sex, state, year, and age for all of the 447 million SSNs that were issued by the NARA Numident cutoff of 2007. Coverage from 2008-2011 is discussed in Appendix A.4.

4.3.1 Foreign Birth

Our probability table can be used as an input to any form of probabilistic imputation, including those that require probability weights.

We recommend a binary version of the fractional imputation described in Kim and Fuller (2004) and Kim and Rao (2009, p. 924): two copies of the dataset are created, one labeled as foreign-born and the other labeled as native-born, with new weights equaling the original weight multiplied by the probability of foreign or native birth.

Fractional imputation has a number of advantages: it is easy to implement, it distorts parameter estimates less than deterministic or mean imputation (Brick and Kalton, 1996), it eliminates imputation variance (Kim and Fuller, 2004; Kim and Rao, 2009), and it permits conventional inference (Fay, 1993 and 1996).

Even with fractional imputation, the association between the imputed variable (foreign birth) and variables other than the ones used to calculate the probability weights (sex, state, year, and age) will be attenuated (Brick and Kalton, 1996). For example, the correlation between imputed foreign birth and income may be weaker than the true correlation.

However, our probability table improves differentiation and reduces the scope for attenuation bias due to foreign-birth imputation.

4.3.2 Related Uses

Our probability table has related uses for administrative data research:

Recovering Foreign Birth at SSA.—Of 302 million SSNs issued by 1985, the National Academies reported that one-fifth had no recorded place of birth (Levine, Hill, and Warren, 1985, p. 94). Our crosswalk can be used to recover missing foreign birth in the SSA Numident, building on methods described in Appendix A.3.

Defining Populations.—Populations in administrative data may be defined from the universe of SSNs, or from the subset of SSNs used in transactions observed in the data. When attempting to measure the resident population for distributional analysis (Feenberg, Mitrusi, and Poterba, 1997; Cilke, 1998; Piketty, Saez, and Zucman, 2016; Auten and Splinter, 2024) there is a tradeoff between the universe and transaction definitions: using the universe includes SSNs that are inactive because the SSN holder is not resident, and using the transactions excludes SSNs that are inactive even though the SSN holder is resident.

These standard definitions expose estimates to endogenous censoring or truncation. The foreign-born are more likely than natives to become non-residents, as time passes: Akee and Jones (2024) find that by 10 years after arrival, 40 percent of immigrants return migrate and disappear from administrative tax data; return migration depends on earnings histories.

Our probability table and crosswalk can be used to restrict the universe of SSNs to likely residents, or to identify which SSNs are missing from transactions because they are likely non-residents, so their outcomes can be modeled, bounded, or imputed.

Consider an analyst at the IRS who wants to calculate statistics for the native population. Chetty *et al.* (2014) study intergenerational mobility, focusing on citizen children claimed on tax forms by their parents. Claims ensure parents' income is observed, but citizenship does not ensure children's income is observed: naturalized and native-born children of foreign-born parents often return migrate, and therefore have zero observed income even when they are upwardly mobile. This biases mobility down, and may explain the null effect of foreign birth on mobility (*ibid.*, Table 8). Our probability table may be used to address bias from return migration by excluding children of foreign-born parents.

Consider an analyst at the IRS who wants to understand the "tax gap," that is, unpaid tax liabilities. Part of the tax gap is due to non-filers, some of whom may not have a liability. When the IRS attempted to find non-filers from a sample of the universe of living, non-filing SSNs, 43 percent could not be located (Erard and Ho, 2001). Official estimates of the non-filer tax gap similarly start from the universe of SSNs and impute demographics and income assuming residence (Mortenson *et al.*, 2009; Erard, Payne, and Plumley, 2012; Langetieg, Payne, and Plumley, 2017; Krause *et al.*, 2023), which may overstate tax liabilities. Our probability table may be used to improve these imputations and tax gap estimates.

5 References

Abowd, John, Kevin McKinney, and Nellie Zhao. 2018. "Earnings Inequality and Mobility Trends in the United States: Nationally Representative Estimates from Longitudinally Linked Employer-Employee Data," *Journal of Labor Economics* 36:S1, pp. S183-S300

Abowd, John, William Bell, J. David Brown, Michael Hawes, Misty Heggeness, Andrew Keller, Vincent Mule, Joseph Schafer, Matthew Spence, Lawerence Warren, and Moises Yi. 2020. "Determination of the 2020 U.S. Citizen Voting Age Population (CVAP) Using Administrative Records and Statistical Methodology," U.S. Census Bureau, Center for Economic Studies Technical Report 20-33

Abramitzky, Ran, Leah Boustan, Elisa Jacome, and Santiago Perez. 2021. "Intergenerational Mobility of Immigrants in the United States over Two Centuries," *American Economic Review* 111:2, pp. 580-608

Agarwal, Sumit, Chunlin Liu, and Nicholas Souleles. 2007. "The Reaction of Consumer Spending and Debt to Tax Rebates—Evidence from Consumer Credit Data," *Journal of Political Economy* 115:6, pp. 986-1019

Agersnap, Ole, Amalie Jensen, and Henrik Kleven. 2020. "The Welfare Magnet Hypothesis: Evidence from an Immigrant Welfare Scheme in Denmark," *American Economic Review: Insights* 2:4, pp. 527-542

Akee, Randall and Maggie Jones. 2024. "Return Migration Decisions and Declining Earnings: Immigrants in Linked Survey and Administrative Data," *Journal of Human Resources* 59:5, p. 722

Ameudo-Dorantes, Catalina, Esther Arenas-Arroyo, Parag Mahajan, and Bernard Schmidpeter. 2024. "Low-Wage Jobs, Foreign-Born Workers, and Firm Performance," U.S. Census Bureau, Center for Economic Studies Working Paper 24-05

Amornsiripanitch, Natee, Paul Gompers, George Hu, and Kaushik Vasudevan. 2023. "Getting schooled: Universities and VC-backed immigrant entrepreneurs," *Research Policy* 52:7

Ashcroft, John. 2001. "Designation of El Salvador Under Temporary Protected Status Program," *Federal Register* Vol. 66, No. 47, p. 14214-14216

Auten, Gerald and David Splinter. 2024. "Income Inequality in the United States: Using Tax Data to Measure Long-Term Trends," *Journal of Political Economy* 132:7, pp. 2179-2227

Autor, David, David Cho, Leland Crane, Mita Goldar, Byron Lutz, Joshua Montes, William Peterman, David Ratner, Daniel Villar, and Ahu Yildirmaz. 2022. "An evaluation of the Paycheck Protection Program using administrative payroll microdata," *Journal of Public Economics* 211, pp. 1-9

Azoulay, Pierre, Benjamin Jones, J. Daniel Kim, and Javier Miranda. 2022. "Immigration and Entrepreneurship in the United States," *American Economic Review: Insights* 4:1, pp. 71-88

Bai, John, Linlin Ma, Kevin Mullally, and David Solomon. 2019. "What a difference a (birth) month makes: The relative age effect and fund manager performance," *Journal of Financial Economics* 132:1, pp. 200-221

Barr, Andrew, Jonathan Eggleston, and Alexander Smith. 2022. "Investing in Infants: The Lasting Effects of Cash Transfers to New Families," *Quarterly Journal of Economics*,

pp. 1-46

Bernstein, Shai, Rebecca Diamond, Abhisit Jiranaphawiboon, Timothy McQuade, and Beatriz Pousada. 2022. "The Contribution of High-Skilled Immigrants to Innovation in the United States," *NBER Working Paper* No. 30797

Blau, Francine and Christopher Mackie. 2017. *The Economic and Fiscal Consequences of Immigration*. Washington, DC: National Academies Press

Block, Gladys, Genevieve Matanoski, and Raymond Seltser. 1983. "A Method for Estimating Year of Birth Using Social Security Number," *American Journal of Epidemiology* 118:3, pp. 377-395

Borjas, George. 1999. "Immigration and welfare magnets," *Journal of Labor Economics* 17:4, pp. 607–637

Breen, Casey and Joshua Goldstein. 2022. "Berkeley Unified Numident Mortality Database: Public administrative records for individual-level mortality research," *Demographic Research* 47:5, pp. 111-142

Brick, J.M. and Graham Kalton. 1996. "Handling missing data in survey research," *Statistical Methods in Medical Research* 5: pp. 215-238 Brown and Murray-Close, 2023

Cajner, Tomaz, Leland Crane, Ryan Decker, John Grigsby, Adrian Hamins-Puertolas, Erik Hurst, Christopher Kurz, and Ahu Yildirmaz. 2020. "The U.S. Labor Market during the Beginning of the Pandemic Recession," *Brookings Papers on Economic Activity* 2020:2, pp. 3-33

Casardi, Alberico. 1957. "Report on Hungarian Refugees," NATO Restricted Document C-M(57)65, p. 1-20. Accessed online: `https://www.nato.int/nato_static/assets/pdf/pdf_archives_hungarian_revolution/20130904_C-M_57_65-ENG.PDF`

Cascio, Elizabeth and Ethan Lewis. 2019. "Distributing the Green (Cards): Permanent residency and personal income taxes after the Immigration Reform and Control Act of 1986," *Journal of Public Economics* 172, pp. 135-150

Chetty, Raj, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez. 2014. "Where is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States," *Quarterly Journal of Economics* 129:4, pp. 1553–1623

Chetty, Raj, Nathaniel Hendren, and Larry Katz. 2016. "The Effects of Exposure to Better Neighborhoods on Children: New Evidence from the Moving to Opportunity Experiment," 106:4, pp. 855-902

Chetty, Raj and Nathaniel Hendren. 2018. "The Impacts of Neighborhoods on Intergenerational Mobility I: Childhood Exposure Effects," *Quarterly Journal of Economics* 133:3, pp. 1107-1162

Chetty, Raj, Nathaniel Hendren, Maggie Jones, and Sonya Porter. 2020. "Race and Economic Opportunity in the United States: An Intergenerational Perspective," *Quarterly Journal of Economics* 135:2, pp. 711-783

Chetty, Raj, John Friedman, Michael Stepner, and Opportunity Insights Team. 2024. "The Economic Impacts of COVID-19: Evidence from a New Public Database Built Using Private Sector Data," *Quarterly Journal of Economics* 139:2, pp. 829-889

Choi, James, David Laibson, Brigitte Madrian, and Andrew Metrick. 2002. "Defined Contribution Pensions: Plan Rules, Participant Choices, and the Path of Least Resistance," *Tax Policy and the Economy* 16, pp. 67-113

Cilke, James. 1998. "A Profile of Non-Filers," U.S. Treasury Office of Tax Analysis Working Paper 78

Clemens, Michael. 2013. "Why Do Programmers Earn More in Houston Than Hyderabad? Evidence from Randomized Processing of US Visas," *American Economic Review* 103:3, pp. 198-202

Clemens, Michael. 2022. "The Fiscal Effect of Immigration: Reducing Bias in Influential Estimates," IZA Discussion Paper No. 15592

Clemens, Michael and Ethan Lewis. 2024. "The Effect of Low-Skill Immigration Restrictions on US Firms and Workers: Evidence from a Randomized Lottery," NBER Working Paper No. 30589

Congressional Budget Office. 2024. "Effects of the Immigration Surge on the Federal Budget and the Economy," Publication No. 60569

Cox, Natalie, Peter Ganong, Pascal Noel, Joseph Vavra, Arlene Wong, Diana Farrell, and Fiona Greig. 2020. "Initial impacts of the pandemic on consumer behavior: Evidence from linked income, spending, and savings data," *Brookings Papers on Economic Activity* 2020:2, pp. 35-82

Dimmock, Stephen, Jiekun Huang, and Scott Weisbenner. 2021. "Give Me Your Tired, Your Poor, Your High-Skilled Labor: H-1B Lottery Outcomes and Entrepreneurial Success," *Management Science* 68:9, pp. 6355-7064

Doran, Kirk, Alexander Gelber, and Adam Isen. 2022. "The Effects of High-Skilled Immigration Policy on Firms: Evidence from Visa Lotteries," *Journal of Political Economy* 130:10, pp. 2501-2533

Dostie, Benoit, Jiang Li, David Card, and Daniel Parent. 2023. "Employer policies and the immigrant–native earnings gap," *Journal of Econometrics* 233:3, pp. 544-567

Dustmann, Christian, Uta Schoenberg, and Jan Stuhler. 2017. "Labor Supply Shocks, Native Wages, and the Adjustment of Local Employment," *Quarterly Journal of Economics* 132:1, pp. 435-483

Edmonston, Barry. 1996. *Statistics on U.S. Immigration: An Assessment of Data Needs for Future Research.* Washington, DC: The National Academies Press

Eisenhower, Dwight. 1958. "U.S. To End Emergency Program for Hungarian Refugees," *Department of State Bulletin* Vol. 38, Iss. 969, Jan. 20, p. 92-93

Elmendorf, Doug and Heidi Williams. 2023. "A serious case for dynamic scoring," *Slow Boring.* Accessed online (28 September 2024): `https://www.slowboring.com/p/a-serious-case-for-dynamic-scoring`

Elmendorf and Williams, 2024. "How Does Accounting for Population Change Affect Estimates of the Effect of Immigration Policies on the Federal Budget?" *Penn Wharton Budget Model.* Accessed online (28 September 2024): `https://budgetmodel.wharton.upenn.edu/issues/2024/1/18/population-change-effect-immigration-policies-on-federal-budget`

Engelman, Michal, Bert Kestenbaum, Megan Zuelsdorff, Neil Mehta, Diane Lauderdale. 2017. "Work Disability Among Native-born and Foreign-born Americans: On Origins, Health, and Social Safety Nets," *Demography* 54: 2273-2300

Erard, Brian and Chih-Chin Ho. 2001. "Searching for ghosts: who are the nonfilers and how much tax do they owe?" *Journal of Public Economics* 81:1, pp. 25-50

Erard, Brian, Mark Payne, and Alan Plumley. 2012. "Advances in Nonfiling Measures," *IRS Research Bulletin*, Publication No. 1500, pp. 79-89

Fay, Robert. 1993. "Valid Inferences from Imputed Survey Data," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 227-232

Fay, Robert. 1996. "Alternative Paradigms for the Analysis of Imputed Survey Data," *Journal of the American Statistical Association* 91:434, pp. 490-498

Feenberg, Daniel, Andrew Mitrusi, and James Poterba. 1997. "Distributional Effects of Adopting a National Retail Sales Tax," *Tax Policy and the Economy* 11, pp. 49-90

Ferenc, Cseresnyés. 1999. "The '56 Exodus to Austria," *Hungarian Quarterly*, Vol. XL, No. 154, Summer. Accessed online: `https://web.archive.org/web/20041127172402/http://www.hungarianquarterly.com/no154/086.html`

Fisher, Robin. 2023. "Estimation of Race and Ethnicity by Re- Weighting Tax Data," Office of Tax Analysis Technical Paper No. 11

Foged, Mette and Giovanni Peri, 2016. "Immigrants' Effect on Native Workers: New Analysis on Longitudinal Data," *American Economic Journal: Applied Economics* 8:2, pp. 1-34

Foote, Andrew and Kevin Stange. 2022. "Attrition from Administrative Data: Problems and Solutions with an Application to Postsecondary Education," NBER Working Paper No. 30232

Ganong, Peter, Fiona Greig, Pascal Noel, Daniel Sullivan, and Joseph Vavra. 2024. "Spending and Job-Finding Impacts of Expanded Unemployment Benefits: Evidence from Administrative Micro Data," *American Economic Review* 114:9, pp. 2898-2939

General Accounting Office. 2001, "Immigration Benefits: Several Factors Impede Timeliness of Application Processing," *Report to Congressional Requesters* GAO-01-488, May 4

Gibbs, Christa, Benedict Guttman-Kenney, Donghoon Lee, Scott Nelson, Wilbert van der Klaauw, and Jialan Wang. Forthcoming. "Consumer Credit Reporting Data," *Journal of Economic Literature*

Glennon, Britta. 2024. "How Do Restrictions on High-Skilled Immigration Affect Offshoring? Evidence from the H-1B Program," *Management Science* 70:2, pp. 907-930

Gomez, Scarlett, Thu Quach, Pamela Horn-Ross, Jane Pham, Myles Cockburn, Ellen Chang, Theresa Keegan, Sally Glaser, and Christina Clarke. 2010. "Hidden Breast Cancer Disparities in Asian Women: Disaggregating Incidence Rates by Ethnicity and Migrant Status," *American Journal of Public Health* 100:S1, pp. S125-S131

Gomez, Nicole, Sylvia Guendelman, Kim Harley, and Scarlett Gomez. 2015. "Nativity and Neighborhood Characteristics and Cervical Cancer Stage at Diagnosis and Survival Outcomes Among Hispanic Women in California," *American Journal of Public Health* 105:3, p. 538-545

Government Accountability Office. 2022. "Tax Equity: Lack of Data Limits Ability to Analyze Effects of Tax Policies on Households by Demographic Characteristics," Report No. GAO-22-104553

Greenberg, Kyle, Matthew Gudgeon, Adam Isen, Corbin Miller, and Richard Patterson. 2022. "Army Service in the All-Volunteer Era," *Quarterly Journal of Economics* 137:4, pp. 2363-2418

Grusky, David and C. Matthew Snipp. 2013. "Developing New National Data on Social Mobility," Washington, DC: National Academies Press

Haas, Ann, Marc Elliott, Jacob Dembosky, John Adams, Shondelle Wilson-Frederick, Joshua Mallett, Sarah Gaillot, Samuel Haffer, and Amelia Haviland. 2019. "Imputation of race/ethnicity to enable measurement of HEDIS performance by race/ethnicity," *Health Services Research* 54:1, pp. 13-23

Hall, Jonathan and Alan Krueger. 2018. "An Analysis of the Labor Market for Uber's Driver-Partners in the United States," *Industrial and Labor Relations Review* 71:3, pp. 705-732

Hausman, Jerry and David Wise. 1979. "Attrition Bias in Experimental and Panel Data: The Gary Income Maintenance Experiment," *Econometrica* 47:2, pp. 455-473

Heckman, James. 1976. "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement* 5:4, pp. 475-492

Heckman, James. 1979. "Sample Selection Bias as a Specification Error," *Econometrica* 47:1, pp. 153-161

Horowitz, Joel and Charles Manski. 2000. "Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data," *Journal of the American Statistical Association* 95:449, pp. 77-84

Hotchkiss, Julie, Myriam Quispe-Agnoli, and Fernando Rios-Avila. 2015. "The Wage Impact of Undocumented Workers: Evidence from Administrative Data," *Southern Economic Journal* 81:4, pp. 874-906

H.R. Journal, No. 85-1. 1957.

Hu, Yaqin. 2022. "Local CEOs, career concerns and voluntary disclosure," *Journal of Business Finance & Accounting* 50:3-4, pp. 565-597

Imai, Kosuke, Santiago Olivella, and Evan Rosenman. 2022. "Addressing census data problems in race imputation via fully Bayesian Improved Surname Geocoding and name supplements," *Science Advances* 8:49, pp. 1-10

Isen, Adam, Maya Rossin-Slater, and W. Reed Walker, 2017. "Every Breath You Take—Every Dollar You'll Make: The Long-Term Consequences of the Clean Air Act of 1970," *Journal of Political Economy* 125:3, pp. 848-902

Jin, Lei, Felix Elwert, Jeremy Freese, and Nicholas Christakis. 2010. "Preliminary Evidence Regarding the Hypothesis That the Sex Ratio at Sexual Maturity May Affect Longevity in Men," *Demography* 47-3, pp. 579-586

Kaushal, Neeraj. 2005. "New Immigrants' Location Choices: Magnets without Welfare," *Journal of Labor Economics* 23:1, pp. 59-80

Keegan, Theresa, Thu Quach, Sarah Shema, Sally Glaser, and Scarlett Gomez. 2010. "The influence of nativity and neighborhoods on breast cancer state at diagnosis and survival among California Hispanic women," *BMC Cancer* 10:603, pp. 1-11

Kerr, Sari Pekkala, William Kerr, and William Lincoln. 2015. "Skilled Immigration and the Employment Structures of US Firms," *Journal of Labor Economics* 33:S1, pp. S147–S186

Kerr, Sari Pekkala and William Kerr. 2017. "Immigrant Entrepreneurship," *Measuring Entrepreneurial Businesses: Current Knowledge and Challenges.* Chicago, IL: University of Chicago Press

Khan, Fatima, Julie Ruterbusch, Scarlett Gomez, and Kendra Schwartz. 2013. "Differences in the cancer burden among foreign-born and US-born Arab Americans living in metropolitan Detroit," *Cancer Causes Control* 24:11, pp. 1955-1961

Kim, Jae Kwang and Wayne Fuller. 2004. "Fractional hot deck imputation," *Biometrika* 91:3, pp. 559-578

Kim, Jae Kwang and J.N.K. Rao. 2009. "A unified approach to linearization variance estimation from survey data after imputation for item nonresponse," *Biometrika* 96:4, pp. 917-932

Kleven, Henrik, Camille Landais, Mathilde Muñoz, and Stefanie Stantcheva. 2020. "Taxation and Migration: Evidence and Policy Implications," *Journal of Economic Perspectives* 34:2, pp. 119-142

Kopczuk, Wojciech, Emmanuel Saez, and Jae Song. 2007. "Uncovering the American Dream: Inequality and Mobility in Social Security Earnings Data since 1937," NBER Working Paper No. 13345

Krause, Melanie, Barry Johnson, Peter Rose, Dan Rosenbaum, Lamont Ayers, Theodore Black, Jonathan Fieman, Andrew Johns, Valentina Kachanovskaya, Patrick Langetieg, Kara Liebel, Mark Payne, Alan Plumley, Thomas Schuh, and Eric Spitzer. 2023. "Federal Tax Compliance Research: Tax Gap Projections for Tax Years 2020 and 2021," IRS Publication No. 5869 (Rev. 10-2023)

Langetieg, Patrick, Mark Payne, and Alan Plumley. 2017. "Counting Elusive Nonfilers Using IRS Rather Than Census Data," *IRS Research Bulletin*, Publication No. 1500, pp. 197-222

Lauderdale, Diane, Steven Jacobsen, Sylvia Furner, Paul Levy, Jacob Brody, and Jack Goldberg. 1997. "Hip Fracture Incidence among Elderly Asian-American Populations," *American Journal of Epidemiology* 146:6, p. 502-509

Lauderdale, Diane, Ronald Thisted, and Jack Goldberg. 1998. "Is Geographic Variation in Hip Fracture Rates Related to Current or Former Region of Residence?" *Epidemiology* 9:5, p. 574-577

Lee, David. 2009. "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *Review of Economic Studies* 76:3, pp. 1071-1102

Levine, Daniel, Kenneth Hill, and Robert Warren (Eds.). 1985. *Immigration Statistics: A Story of Neglect*. National Academy Press: Washington, D.C.

Little, Roderick and Donald Rubin. 1987. *Statistical Analysis with Missing Data*, Hoboken, NJ: Wiley

Lubotsky, Darren. 2007. "Chutes or Ladders? A Longitudinal Analysis of Immigrant Earnings," *Journal of Political Economy* 115:5, pp. 820-867

Mack, Thomas, Ann Walker, Wendy Mack, and Leslie Bernstein. 1985. "Cancer in Hispanics in Los Angeles County," *National Cancer Institute Monograph* 69, pp. 99-104

Mack, Thomas and Birgitta Floderus. 1991. "Malignant melanoma risk by nativity, place of residence at diagnosis, and age at migration," *Cancer Causes and Control* 2, pp. 401-411

Madrian, Brigitte and Dennis Shea. 2001. "The Power of Suggestion: Inertia in 401(k) Participation and Savings Behavior," *Quarterly Journal of Economics* 116:4, pp. 1149-1187

Manski, Charles. 1990. "Nonparametric Bounds on Treatment Effects," *American Economic Review* 80:2, pp. 319-323

Markowitz, Arthur. 1973. "Humanitarianism versus Restrictionism: The United States and the Hungarian Refugees," *International Migration Review* Vol. 7, No. 1, p. 46-59

Massey, Douglas. 2010. "Immigration Statistics for the 21st Century," *American Academy of Political and Social Sciences* 631:1, pp. 124–140

Mayda, Anna Maria, Francesc Ortega, Giovanni Peri, Kevin Shih, and Chad Sparber. 2023. "Coping with H-1B Shortages: Firm Performance and Mitigation Strategies," *Industrial and Labor Relations Review* 76:5, pp. 919-943

Mehta, Neil, Irma Elo, Michal Engelman, Diane Lauderdale, and Bert Kestenbaum. 2016. "Life Expectancy Among U.S.-born and Foreign-born Older Adults in the United States: Estimates From Linked Social Security and Medicare Data," *Demography* 53, pp. 1109-1134

Mian, Atif, Kamalesh Rao, and Amir Sufi. 2013. "Household Balance Sheets, Consumption, and the Economic Slump," *Quarterly Journal of Economics* 128:4, pp. 1687-1726

Miller, Sarah, Laura Wherry, and Gloria Aldana. 2023. "Covering Undocumented Immigrants: The Effects of A Large-Scale Prenatal Care Intervention." Accessed online (28 September 2024): `https://websites.umich.edu/~mille/MillerWherryUndocumented.pdf`

Mortenson, Jacob, James Cilke, Michael Udell, and Jonathon Zytnick. 2009. "Attaching the Left Tail: A New Profile of Income for Persons Who Do Not Appear on Federal Income Tax Returns," *Proceedings of the National Tax Association*, pp. 88-102

National Archives and Records Administration. 2013. *NUMIDENT Public Use Files: Application (SS-5), Claims, Death, and Documentation Package, Layouts, and Code Lists*, Electronic Records Reference Branch (RRER)

Nelson, William. 1985. "Employment Covered by the Social Security Program, 1935-84," *Social Security Bulletin* Vol. 48, No. 4, p. 33-39

New York Times. 1957. "Attorney Resigns Post as Refugee Coordinator," Mar. 1, p. 5

Panis, Constantijn, Roald Euller, Cynthia Grant, Melissa Bradley, Christine Peterson, Randall Hirscher, and Paul Steinberg. 2000. *SSA Program Data User's Manual*. Baltimore, MD: Social Security Administration Office of Research, Evaluation, and Statistics

Piketty, Thomas, Emmanuel Saez, and Gabriel Zucman. 2016. "Distributional National Accounts: Methods and Estimates for the United States," *Quarterly Journal of Economics* 133:2, pp. 553-609

Pool, Veronika, Noah Stoffman, and Scott Yonker. 2012. "No Place Like Home: Familiarity in Mutual Fund Manager Portfolio Choice," *The Review of Financial Studies* 25:8, pp. 2563-2599

Raz, Dan, Scarlett Gomez, Ellen Chang, Jae Kim, Theresa Keegan, Jane Pham, Jasleen Kukreja, Robert Hiatt, and David Jablons. 2008. "Epidemiology of Non-small Cell Lung Cancer in Asian Americans: Incidence Patterns Among Six Subgroups by Nativity," *Journal of Thoracic Oncology* 3:12, pp. 1391-1397

Reno, Janet. 1994. "Rules and Procedures for Adjudication of Applications for Asylum or Withholding of Deportation and for Employment Authorization," *Federal Register* Vol. 59, No. 232, p. 62284-62303

Rho, Deborah and Seth Sanders. 2021. "Immigrant Earnings Assimilation in the United States: A Panel Analysis," *Journal of Labor Economics* 39:1, pp. 37-78

Saiz, Albert. 2007. "Immigration and housing rents in American cities," *Journal of Urban Economics* 61, pp. 345-371

Schwabish, Jonathan. 2011. "Identifying Rates of Emigration in the United States Using Administrative Earnings Records," *International Journal of Population Research*, pp. 1-17

Shimizu, Hiroyuki, Thomas Mack, Ronald Ross, and Brian Henderson. 1987. "Cancer of the Gastrointestinal Tract Among Japanese and White Immigrants in Los Angeles County," *Journal of the National Cancer Institute* 78:2, pp. 223-228

Shimizu, Hiroyuki, Ronald Ross, Leslie Bernstein, Ryuichi Yatani, Brian Henderson, and Thomas Mack. 1991. "Cancers of the prostate and breast among Japanese and white immigrants in Los Angeles County," *British Journal of Cancer* 63:6, pp. 963-966

Social Security Administration. 1981. "Exhibit—SSN's Issued as of November 2, 1977," *Claims Manual* Part 2, Oct. 1. Sec. 2891

Social Security Administration. 2023. "Social Security Numbers for Noncitizens," Publication No. 05-10096, Apr. Accessed online: `https://www.ssa.gov/pubs/EN-05-10096.pdf`

SSN-Verify. 2024. "Verify Social Security Number Issuance," accessed online (March 7, 2024): `https://www.ssn-verify.com`

Swing, Joseph. 1958. "Hungarian Escapee Program," *I&N Reporter* Vol. 6, No. 4, Apr., p. 43-45, 48

Thaler, Richard and Shlomo Benartzi. 2004. "Save More Tomorrow™: Using Behavioral Economics to Increase Employee Saving," *Journal of Political Economy* 112:S1, pp. S164-S187

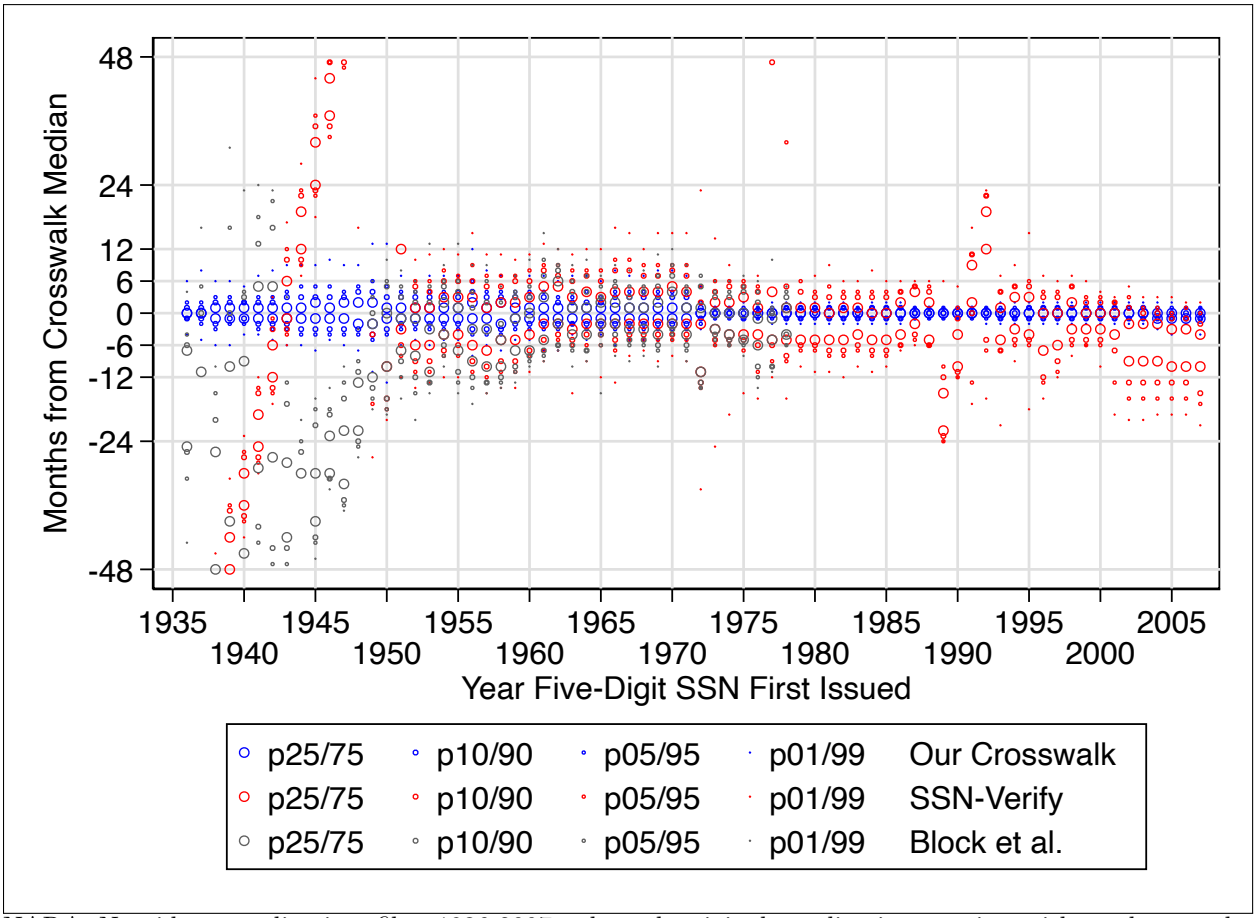
U.S. Citizenship and Immigration Services. 2022. "Operation Safe Haven: The Hungarian Refugee Crisis of 1956," accessed online: `https://www.uscis.gov/about-us/our-history/stories-from-the-archives/operation-safe-haven-the-hungarian-refugee-crisis-of-1956`

Vega, Alma. 2015. "The Impact of Social Security on Return Migration among Latin American Elderly in the U.S.," *Population Research and Policy Review* 34:3, pp. 307-330

Yonker, Scott. 2017. "Geography and the Market for CEOs," *Management Science* 63:3, pp. 587-900

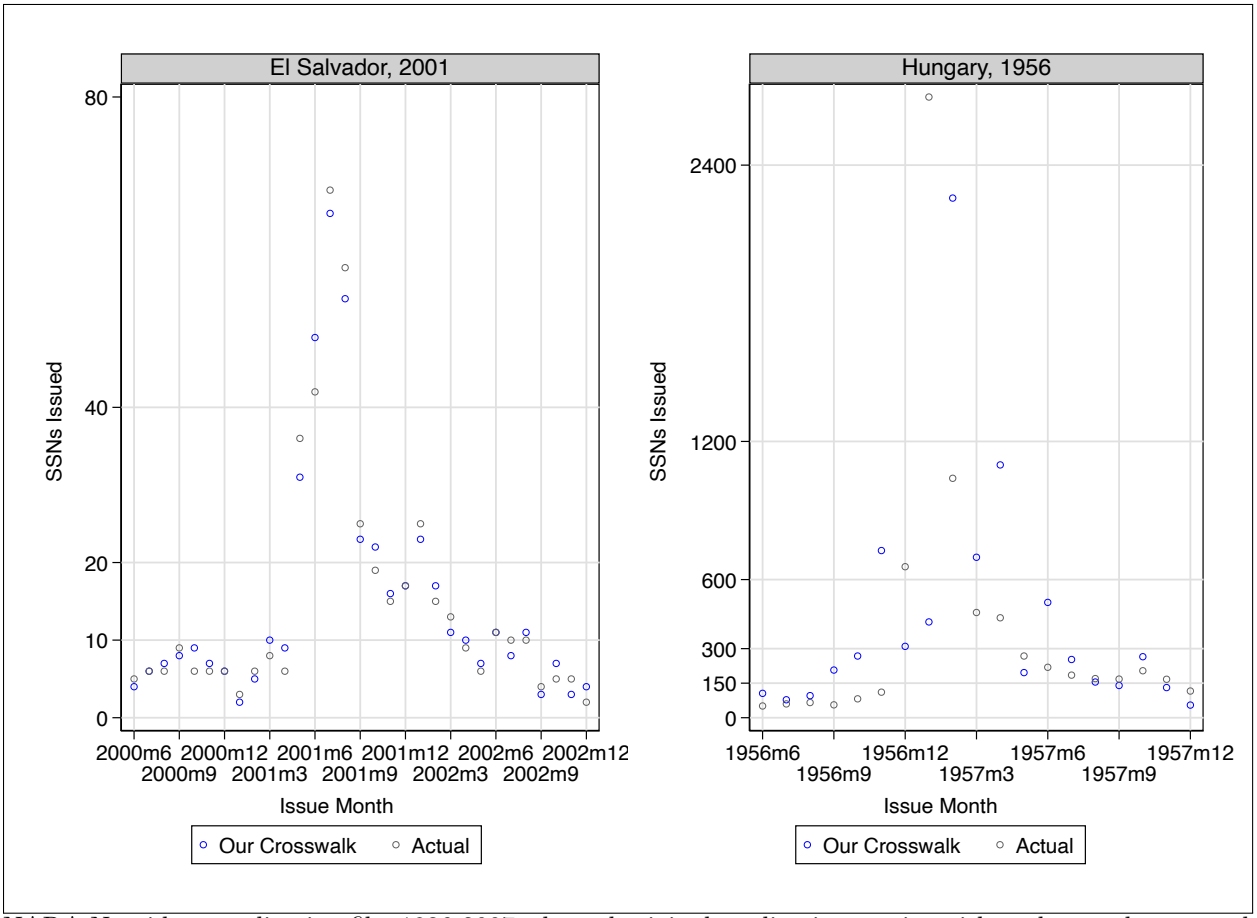
6 Figures

Figure 1: Distribution of Differences between Imputed and Actual Issue Months, Using Our Crosswalk and Issue-Year Crosswalks



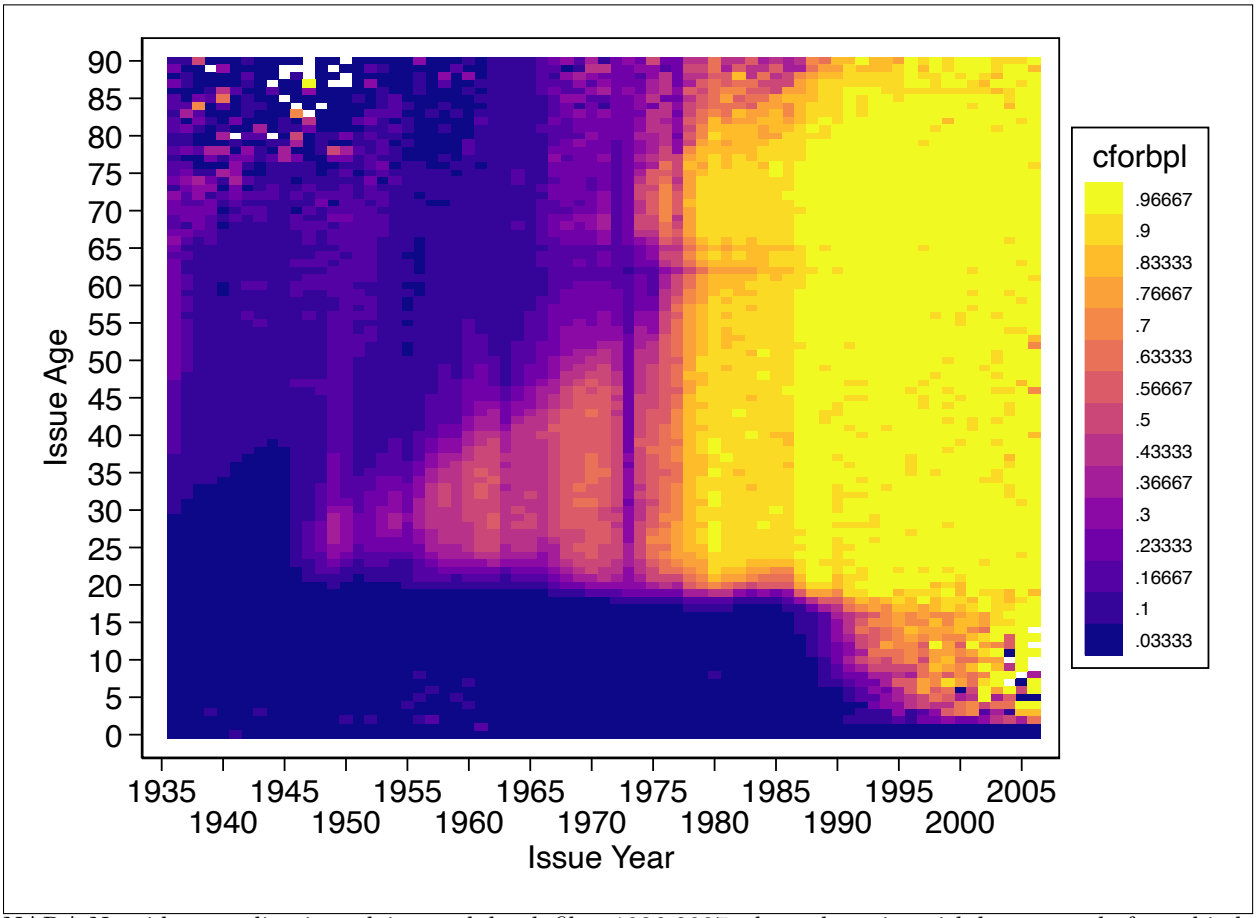
NARA Numident application file, 1936-2007, cleaned original application entries with cycle months (N=35,580,763), merged to our crosswalk and the annual SSN-Verify and Block et al. crosswalks by five-digit SSN. We calculate the median month for the issue-year crosswalks as the rounded average of the January of the first year and December of the last year they report. We calculate the difference in months between the cycle month and the crosswalk median month, for each crosswalk. We group entries by the five-digit SSN's year of first issuance (according to our crosswalk) and calculate the distribution for each crosswalk. The plotted points are percentiles of the differences in months, within the year first issued; we do not plot percentiles outside of 48 months from the median. Overall, this figure shows that both issue-year crosswalks perform very poorly before 1950, and that they also perform poorly from 1950 onward: neither issue-year crosswalk is correctly and consistently centered on the true median. For many SSNs, and in some years most SSNs, using the issue-year crosswalks to impute issue months would deliver an imputed month more than a year from the true month. Our crosswalk (see also Appendix Figure A1) almost always delivers an imputed issue month within 1-4 months of the true month, for any year from 1936 through 2011.

Figure 2: Immigration Months: El Salvador, 2001 and Hungary, 1956



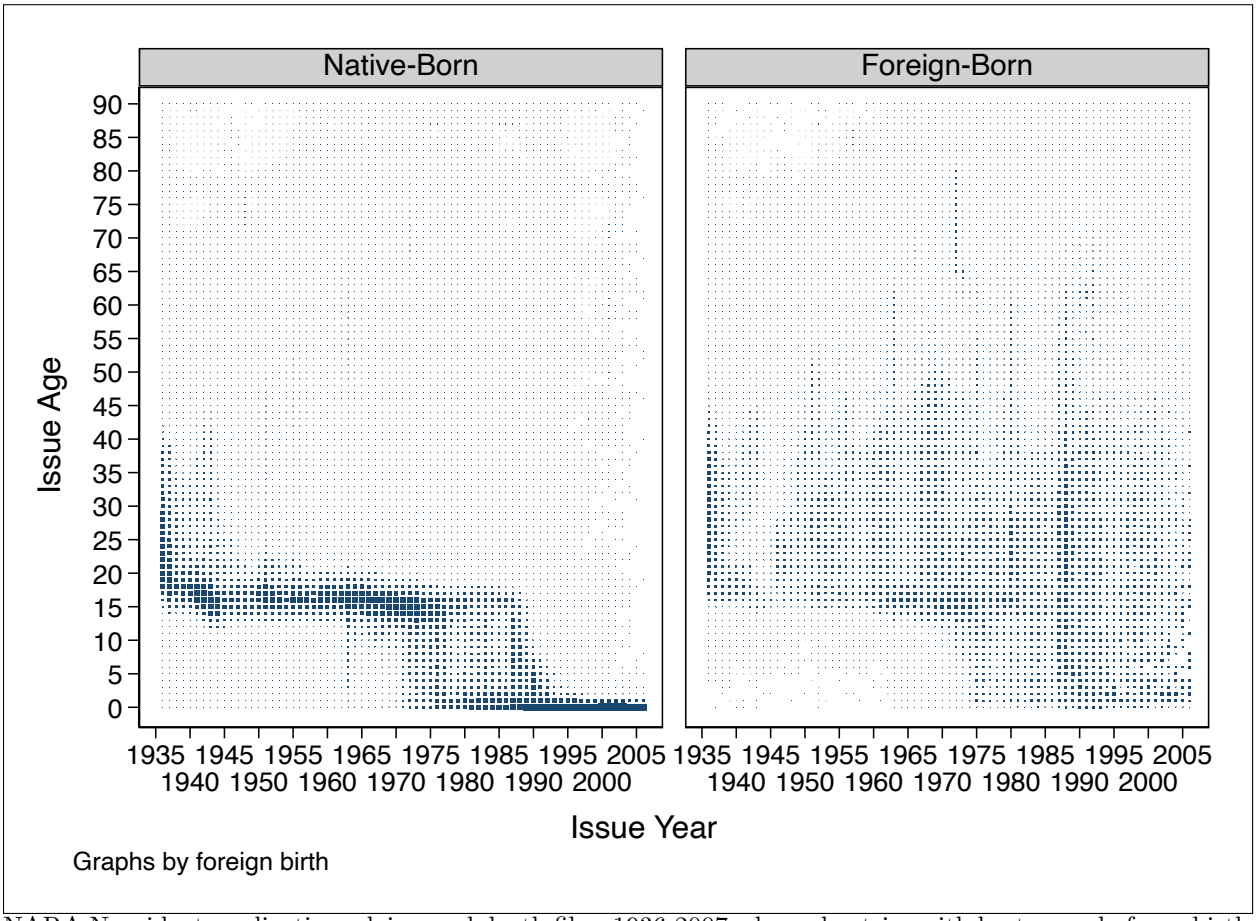
NARA Numident application file, 1936-2007, cleaned original application entries with cycle months; merged to best record of birth month, sex, and birth country code; and merged to our crosswalk for median cycle month. First panel restricted to number-holders born in El Salvador (N=17,230), issued SSNs from June 2000 through December 2002 according to NARA Numident (N=462) or our crosswalk (N=462). This panel shows that SSN issuance to Salvadorans closely tracks their known dates of eligibility under the 2001 Temporary Protected Status (TPS) designation, and that issue months imputed from our crosswalk reproduce the same time pattern. Second panel restricted to number-holders born in Hungary (N=63,012), issued SSNs from June 1956 through December 1957 according to NARA Numident (N=7,206) or our crosswalk (N=7,140). This panel shows that SSN issuance to Hungarian escapees closely tracks their known arrivals in the United States by airlift and sealift and their subsequent processing, and that issue months imputed from our crosswalk reproduce the same time pattern. Absolute numbers of SSNs reported in this figure are of individuals we observe in the NARA Numident, which includes only SSN account holders who were deceased by the end of 2007. The number we observe is therefore much smaller than the total number of SSNs issued: for example, USCIS reported about 100,000 El Salvador TPS beneficiaries by the end of 2002 (GAO, 2020, p. 50), or about two-hundred times the number we observe. See Appendix Figure A2 and A5 for additional notes.

Figure 3: Probability of Foreign Birth by Issue Year and Age



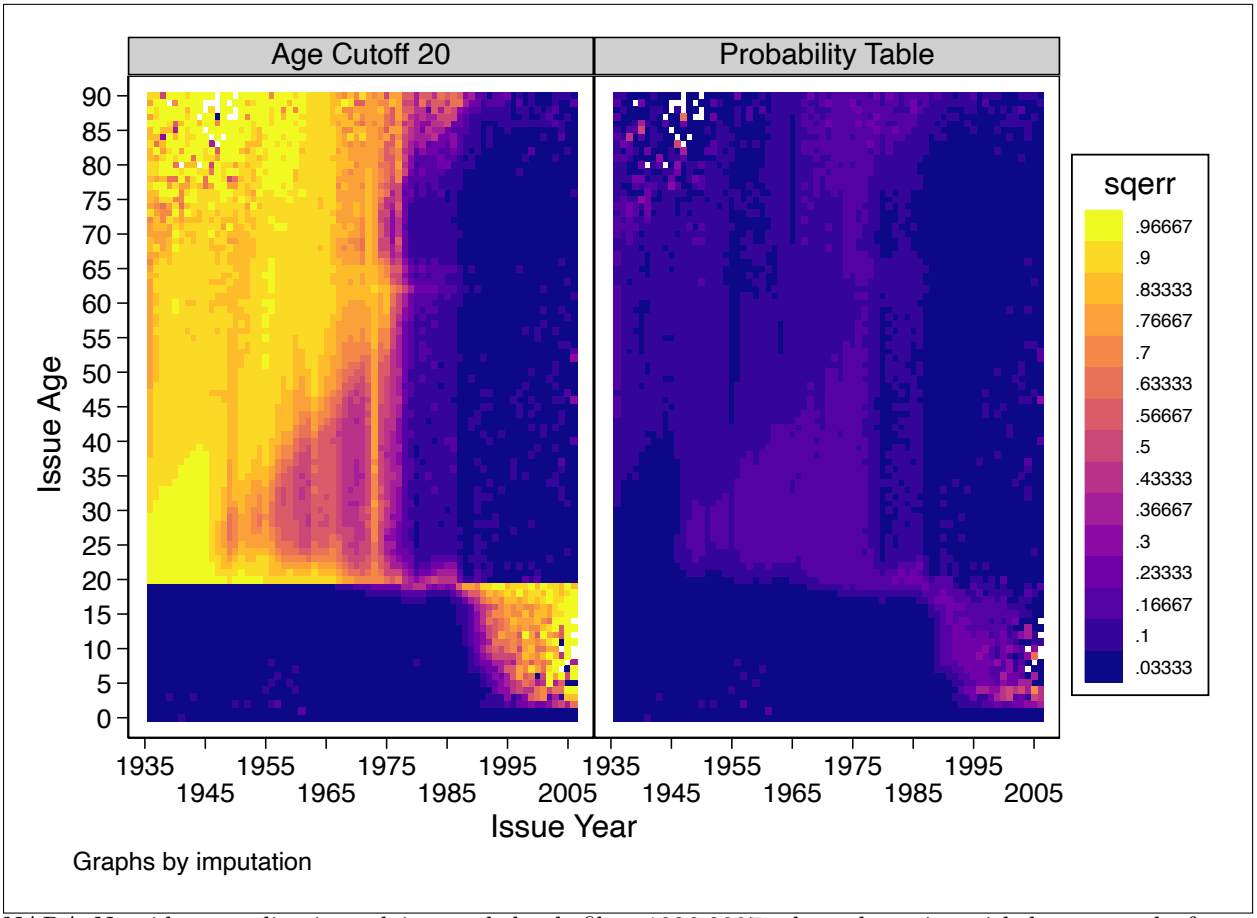
NARA Numident application, claim, and death files, 1936-2007, cleaned entries with best record of sex, birth date, and foreign birth; imputed where missing; merged to our crosswalk by five-digit SSN; restricted to SSN account records with known death date (N=49,379,843). The plot shows the probability of foreign birth by issue year, and the issue age based on the birth month and crosswalk issue month. The probability of foreign birth is affected by current enumeration of the foreign-born relative to the native-born, but also by past enumeration exhausting native- or foreign-born eligibles. The probability of foreign birth for any issue age rarely exceeds 90 percent until 1987, 80 percent until 1980, or 60 percent in earlier issue years. Issuance in childhood is mostly but not only to natives, except in the 1990s. Issuance at age 0 is mostly but not only to natives in any year. The probability of foreign birth markedly declines in entry cohorts 1921-45, in issue years 1951, 1955, 1962-63, 1965, and 1972-73, and at issue ages 0, 62, and 65, as few foreign-born or many native-born individuals are enumerated due to various policy changes discussed in the notes to Figure 4. It rebounds afterwards as fewer native-born individuals remain to be enumerated.

Figure 4: Frequency of Issuance by Year, Age, and Native or Foreign Birth



NARA Numident application, claim, and death files, 1936-2007, cleaned entries with best record of sex, birth date, and foreign birth; imputed where missing; merged to our crosswalk by five-digit SSN; restricted to SSN account records with known death date (N=49,379,843). The plot shows the relative frequency of issuance by issue year and issue age based on the birth month and crosswalk issue month, within the native-born or foreign-born resident population. Observations are weighted by inverse probability of inclusion in the NARA Numident, and then normalized so that each panel has the same total weight, for legibility. SSNs are issued to native- and foreign-born applicants at all ages in all years. For both, the heaviest enrollment is at all ages in 1936-37 as SSNs are launched, at first employment from age 14-18 through 1988, and at age 0-1 from 1973 onward as SSNs were required for federal welfare and tax benefits with notable changes in 1973-74 and 1987-91. For both, enumeration is also driven by covered sector expansions in 1951, 1955-57, and 1966-67 (Nelson, 1985); tax identification requirements in 1962-64 (Hoban, 1962; SSA, 1965, p. 2-3) and 1983; Medicare enrollment in 1965-66; enumeration of secondary beneficiaries in 1972 (SSA, 1973, p. 50; Telling, 1972); and enumeration of SSI and food stamps beneficiaries in 1975 and 1977. Native- and foreign-born applicants are also distinguished by a variety of policies. Many natives are enumerated at birth from 1989 onwards, though not all: some parents (up to 7 percent: see GAO, 2005; SSA, 2021) defer the application to later in childhood or adulthood. Many foreign-born children are enumerated at age 1 or later, though not all: some foreign-born children, half of them born to citizens, obtain an SSN at age 0. We also observe reduced immigration from 1921-46 due to quotas, the Great Depression, and World War II; refugee crises from 1980; and enumeration of undocumented immigrants from 1987-89.

Figure 5: Mean Squared Error of Foreign Birth Imputations by Issue Year and Age



NARA Numident application, claim, and death files, 1936-2007, cleaned entries with best record of sex, birth date, and foreign birth; imputed where missing; merged to our crosswalk by five-digit SSN; restricted to SSN account records with known death date (N=49,379,843). The plot shows the MSE of actual values of foreign birth minus imputed values of foreign birth by issue year and issue age. Panel A shows the MSE for an imputation that assigns native birth to all SSNs at issue age below 20, and foreign birth to all SSNs at issue age at or above 20. This issue-age cutoff imputation yields high rates of misclassification, often from 90-100 percent, because natives frequently obtain SSNs at issue ages above 20; because the foreign-born frequently obtain SSNs at issue ages below 20; because it is completely insensitive to major policy changes; and because it assumes a binary classification where none exists. Panel B shows the MSE for an imputation from our probability table. The side-by-side comparison makes it clear that our probability table reduces misclassification, measured by the MSE, in all issue years and ages. In absolute terms, the MSE is on the order of 0.03 in most years and ages, including all peak years and ages; the MSE is generally much better than random chance, which itself is often better than the issue-age cutoff imputation.

A Online Appendix

A.1 Building the Crosswalk

This appendix presents additional details of how we built the crosswalk, including evidence of SSA enumeration processes.

A.1.1 NARA Numident

We start by obtaining the NARA Numident records of SSN applications (from SSA Form SS-5 and others), which cover two-thirds of all number-holders deceased or presumed deceased by December 2007.[13] Why, if we have these records, do we need a crosswalk? First, the NARA Numident records cover only a fraction of SSNs. Second, the records do not always include issue dates.[14] Third, not all apparent issue dates are correct.

We code all dates by month and year, leaving out the day which is often missing.

To identify the issue dates for each five-digit SSN, we first identify the issue dates for SSNs observed in the NARA Numident.

The issue date for each SSN is not explicitly recorded. Numident *records* for each SSN consist of *entries* from original applications and from updates initiated by SSA or by the applicant. Each entry in the Numident is given a cycle date.[15] The original entry should be the first entry in a record, when the entries are ordered by cycle date.

The issue date is the cycle date of the original entry, when the original entry and correct cycle date can be identified. Because the Numident was converted from paper, punch cards, and microfiche to a computerized system, lost entries, data entry error, and editing mean the original entry may be missing, mis-ordered, or mis-dated. In these cases, the first entry may not be the original entry and the cycle date of the first entry will not be the issue date. We identify issue dates by elimination.

To identify original entries and their cycle dates, we discard entries based on a series of record-, person-, entry-, and order-based logical checks.

Record-based checks. We discard records where we cannot identify the original entry by ordering the entries, either because an entry is un-dated or because an entry is mis-dated

[13] Researchers may obtain the NARA Numident records for a nominal fee by following the instructions at `https://www.archives.gov/research/order/electronic-records`. The catalog page for Form SS-5 files is at `https://catalog.archives.gov/id/23845613`. Technical documentation is available at `https://aad.archives.gov/aad/popup-tech-info.jsp?f=5138`. Interested readers may search a subset of the entries at `https://aad.archives.gov/aad/series-description.jsp?s=5057&cat=all&bc=sl`

[14] First, application forms "were occasionally lost in transit, fire, or flood before [the] Central Office could establish the records" (SSA, 2010): the records were then established with a new form at a later date. Second, application forms were often substituted with later claim forms (OA-C790 or OA-D840) when the Numident was built (Panis *et al.*, 2000, p. 407); most of the 25 million SSNs in the NARA Numident that do not have an original application entry appear instead in the NARA Numident claims file, without the issue date. The records most affected by substitution of claim forms for application forms are the records that had a death, disability, or retirement claim before 1972. Most retirement claims were made upon reaching age 62 or 65, meaning birth cohorts up to 1910 were the most affected. SSNs issued after 1972 are generally not affected because the original application entries were posted directly to the Numident.

[15] For entries under the original paper-based system, the cycle date is the date given on the form; while for entries under the modern computerized system, the cycle date is the date on which the entry was posted to the database. See Panis *et al.* (2000, p. 411), "*SSA Program Data User's Manual.*"

based on information about how it was added to the database. We also remove records where the first entry is neither from an original application nor from an emergency application.

We keep the first entry from each remaining record.

Person-based checks. We discard entries where personal data contradict cycle dates. We discard entries with a birth date after the cycle date, entries with no birth date, and entries where the cycle date implies an age over 110 years. We merge our extract to the Berkeley Unified Numident Mortality Database (BUNMD; Breen and Goldstein, 2022), derived from the NARA Numident death files, to obtain the death date associated with each SSN; we discard entries with a birth date after the death date, or with a cycle date after the death date, excepting cases attributable to infant mortality.

Entry-based checks. We discard entries where record-keeping rules and indicators suggest the cycle date is not the original issue date. First, we discard entries dated before SSNs were issued or after NARA received Numident records. Second, we discard entries for which no original card was issued: these are entries with no card, a duplicate card, or a correction. Third, we discard entries where the cycle date was overwritten. SSA documentation claims that entries are not overwritten and individual entry fields are never deleted "except in rare circumstances" (Panis *et al.*, 2000, p. 407), but we find that the SSA overwrites an entry's cycle date and source whenever another field is edited. We discard entries where the overwritten cycle date does not match the reference number, or where the overwritten source does not match the five-digit SSN: for example, we discard entries created on a computer, for five-digit SSNs issued by paper processing. Fourth, we discard entries with obvious data entry error, based on the source of the entry: for example, we discard entries dated after the end of paper processing, but marked as having been created on paper.

Order-based checks. Because editing and data entry errors do not always lead to contradictions within entries, we use other methods to determine whether entries contain cycle dates that are not issue dates. One way is to check that cycle dates follow the sequence of four-digit serial numbers within five-digit area-group numbers: prior to centralized issuance, SSA usually followed the standard order from 0001-9999; and during centralized issuance, SSA followed a non-standard order substituting a number from the sequences 2000-2999 and 7000-7998 for every fifth number of the standard order, and ending in 7999 (Kestenbaum, 2012; Barron and Bamberger, 1982). Thus, we discard entries where the cycle date is more than two months from the cycle date of the next three entries before and after it in the serial order. Since the adjacent entries change as entries are discarded, we iterate this process until no further mis-ordered entries are identified.

After discarding records and entries, we interpret the remaining entries to be original entries, and their cycle dates to be true issue dates.

Finally, we collapse the cleaned extract of cycle dates by the five-digit SSN area-group number. For each area-group number, we record the number of observations, the first and last cycle dates by serial order, the minimum, the 1st, 2nd, 5th, 10th, 25th, 50th, 75th, 90th, 95th, 98th, and 99th percentiles, and the maximum.

A.1.2 High Group Lists

We next obtain the SSA's summary records of SSN issuance from the "high group lists" issued through 2011. The high group lists report the highest group number issued for each

area number, according to the fixed order of group number assignment, and were released monthly. Recall that a five-digit SSN is a three-digit area number followed by a two-digit group number. We compiled the high group lists from SSA's website, and from archived copies of SSA's website, for the months from January 1999 through June 2011.[16]

We use the high group lists to identify the issue dates for five-digit SSNs issued in the mid-2000s through 2011. We use the overlap between the NARA Numident and the high group lists to check for agreement between sources. We rely on the high group lists in this period because, from the mid-2000s, NARA Numident coverage is poor, and from 2008 onwards, NARA Numident data are not available at all.

For each five-digit SSN, we identify a minimum and a maximum issue month, and interpolate a median issue month.

The minimum issue month is the first month the five-digit SSN is observed in the high group lists, if it is marked as new that month. From September 2003 through June 2011, no gap in the high group lists is longer than one month, so we observe or infer[17] the exact minimum from September 2003 onward.

The maximum *reporting* month is the last month the group number is observed in the high group lists, if we also observe the next month of data. From November 2003 through June 2011, the only missing maximum observations are in August 2009, because that month is followed by a missing high group list in September 2009.

We edit the maximum reporting month to better track the maximum issue month. The maximum reporting month for a five-digit SSN in the high group lists is not necessarily an issue month. High group lists report the highest group issued yet for each area number, which may not be issued in the reporting month. In fact, the maximum issue month may be months or years earlier than the maximum reporting month. We edit two cases:

Months of final group numbers. For all observations of group number 99, we replace the maximum as missing, since group number 99 is not replaced by a higher number when it is exhausted. We drop area numbers where group number 99 was reached before January 1999, the earliest month we observe; and we also drop all area numbers issued by the Railroad Retirement Board (RR), which were not issued after 1963.

Months of temporarily inactive area numbers. SSA issued area numbers for each state sequentially within group numbers, so until the next month an area number was reached, the high group list reports the previous group number issued in that area number. Because issuance did not move to the subsequent five-digit SSN until the previous five-digit SSN was exhausted, we replace the maximum month for a given five-digit SSN with the subsequent minimum, whenever the maximum exceeds the subsequent minimum. If the minimum is missing for the next five-digit SSN in sequence, because it is not reported in the high group list, we replace the maximum as missing.

We interpolate the median issue month as the average of the minimum and the maxi-

[16]High group lists for November 2003 through June 2011 (with the exception of September 2009) are available online at `https://www.ssa.gov/employer/ssnvhighgroup.htm`, and high group lists for January 1999 through September 2003 are archived (with gaps) by the Internet Archive at `https://web.archive.org/web/19990203134611/http:/www.ssa.gov/foia/highgroup.htm`.

[17]To bridge single-month gaps, we identify the minimum as the missing month if the five-digit SSN is not seen in the list before the missing month, but is seen in the list after the missing month and is not reported as new in the list after.

mum, rounded up. The average is reasonable because from 1999 onward, the minimum and maximum are often identical and are always close together.

Finally, we collapse the cleaned high group list extract by the five-digit SSN. For each five-digit SSN, we record the minimum, median, and maximum.

A.1.3 Combining the NARA Numident and the High Group Lists

The crosswalk file covers 99,000 possible five-digit SSNs, generated from the range of area numbers (000-999) and the range of group numbers (01-99).

To build and validate the crosswalk, we add information about the state, territory, entity, or program to which each five-digit SSN was assigned; the issue sequence; the issue process—district office, central office, or hybrid; and whether the five-digit SSN was issued before the randomization of SSNs (SSA Claims Manual, 1981, Sec. 2891; Barron and Bamberger, 1982; SSA, 2005; SSA, 2024a and 2024b). Then, we add our collapsed NARA Numident and high group list extracts onto the combined file.

The crosswalk identifies a single minimum, median, and maximum issue month for each five-digit SSN. For five-digit SSNs issued in 2003 and earlier, the issue months come from the NARA Numident; for five-digit SSNs issued in 2004 and later, the issue months come from the high group lists.[18]

We adjust months from the high group lists for delayed reporting. The NARA Numident reports actual issue dates while the high group lists do not report new issuance until the start of the next calendar month: subtracting months reported in the high group lists from months reported in the NARA Numident, 98 percent of differences are between -2 and 0 months. Thus, in the crosswalk, we subtract one month from each high group report.

We adjust months from the NARA Numident for outliers. The minimum and maximum are sometimes several years away from the adjacent percentiles, within and across five-digit SSNs, because we do not detect all edits and data entry problems in the raw NARA Numident entries. We trim the NARA Numident minimum and maximum to handle outliers according to the rate of SSN issuance within each five-digit SSN.[19]

Finally, we replace a small number of missing maximum and median values from the high group lists due to our edits of group number 99. In each case, the missing values are for the final issue month in a range of area numbers issued in a state, before the state switched to another range of area numbers. We replace the missing maximum with the first issue month for the next range, and interpolate the missing median.

[18]Switching sources between 2003 and 2004 balances the strengths and weaknesses of the NARA Numident and the high group lists. The mean number of observations in the cleaned NARA Numident extract's five-digit SSNs issued in 2003 and earlier is 837 (median: 487, 1st percentile: 16), while the mean number of observations in five-digit SSNs issued in 2004 and later is 13 (median: 12, 1st percentile: 1). The high group lists are missing only one month from 2004 onward, and many for 2003 and earlier.

[19]If the minimum is an outlier, we replace the minimum with the 1st percentile, or if the first cycle date by serial order is larger than the 1st percentile, then we replace the minimum with the smaller of the first cycle date or the 2nd percentile. We define an outlier minimum as one that is farther from the 1st percentile, or farther from the smaller of the first cycle date or the 2nd percentile, than the 10th percentile is from the 50th, or 5 months, whichever is larger. We trim the maximum similarly. We carry out a handful of manual edits on outliers, all of which are documented in the code used to produce the crosswalk.

A.1.4 Special Cases for Crosswalk Users

We document two special cases that are not discussed in SSA publications, in other crosswalks, or in the literature that depends on other crosswalks.

Both special cases involve the reuse of old five-digit SSNs for new issuance.

Special case #1: Backfill of unreached SSNs following the transition from District Office issuance to Central Office issuance. Based on our analysis, by January 1972, SSA shipped the last five-digit blocks of pre-printed Social Security cards and forms to District Offices to issue, leaving unreached area numbers within each group number. The District Offices did not start pre-printed issuance from a new five-digit SSN beyond November 1972. For District Offices that ran out of pre-printed cards and forms before the intended transition date, the Central Office printed and shipped emergency cards and forms from unreached area numbers within the final group number (SSA Claims Manual, 1981, Sec. 2891), which the District Offices started to use between October 1972 and January 1973. The District Offices did not continue issuance beyond May 1973 (SSA, 1974, p. 50), regardless of whether unissued serial numbers remained within a five-digit SSN.

Meanwhile, as group numbers were fully exhausted or as District Offices reached either the planned end of their pre-printed cards and forms, or the final deadline, the Central Office began its own centralized issuance from the next group number.

This transition left a large number of SSNs unissued, in unreached five-digit SSNs within final group numbers and in unreached serials within final five-digit SSNs.

We document here that, in May 1977 or in some cases later, with centralized issuance fully established, the Central Office interrupted the standard issue sequence to go back to unreached serials and five-digit SSNs. The Central Office completed issuance of the unreached SSNs before returning to the standard sequence. Due to the variable number of unreached SSNs by state, and the variable rate of issuance across states, this period of backfill took as little as a month, and as much as three years: backfill of unused serials appears to have been completed by September 1978 or earlier, backfill of unused area numbers within reached group numbers appears to have been completed by November 1979 in all but one case, and backfill of unused group numbers appears to have taken until August 1980.

Five-digit SSNs where SSA backfilled unreached area and group numbers are unremarkable in our crosswalk, except they do not follow the standard issue sequence. These five-digit SSNs can be identified from the way they break sequence, and from indicators we include in our crosswalk. These five-digit SSNs are common, and because other crosswalks rely on the standard issue sequence to impute minimum and maximum SSN issue years, they lead to extreme outliers of the issue year, the issue age, and related variables derived from those crosswalks and relied on in the literature.

Five-digit SSNs where SSA backfilled unreached serial numbers have unusually large ranges between the minimum and maximum issue date in our crosswalk, typically 5-8 years. These five-digit SSNs are few, but attempting to use the minimum or maximum to classify the issue month, the issue age, and related variables will result in some extreme outliers. These five-digit SSNs can be identified either from the indicators we provide, or from the mixture of pre-conversion (A, C, M) and post-conversion (F) system indicators. In some cases, the backfilled serial numbers make up a very large proportion or a very small proportion of the total, suggesting that a narrower range of issue months could be imputed, or the median

relied upon, without leading to too much misclassification. We leave this imputation to individual users, and report the full range of percentiles of issue dates found in our NARA Numident extract to facilitate alternative classifications.

Special case #2: SSN randomization. In June 2011, SSA replaced the standard sequence-based assignment of SSNs with fully-randomized assignment. The change in procedure was motivated in part by a need to allocate numbers across states in such a way that all of the available numbers could be used, and in part by concerns about the increasing guessability of SSNs (Acquisti and Gross, 2009).

Because sequence-based number assignment ended on June 24, 2011 and was replaced with randomized number assignment on June 25, 2011, we have a situation analogous to the transition from District Office issuance to Central Office issuance: certain serial numbers within actively-issued five-digit SSNs were not reached.

Serial numbers that were not reached could in principle be backfilled under randomized number assignment, in which case the true issue date would be potentially years after the issue date reported in our crosswalk.

At most 55 five-digit SSNs have serial numbers that were not reached. These five-digit SSNs, one for each state, DC, Puerto Rico, Virgin Islands, Outlying Territories, and Enumeration at Entry, were actively issued as of June 24, 2011. We flag these five-digit SSNs in the crosswalk. Further, with fewer than 1.3 percent of remaining SSNs issued each year, it would take more than 77 years to fully backfill the remaining serials in these 55 five-digit SSNs. If half of the serials had already been reached as of June 24, 2011, fewer than 0.7 percent of all the serials in these five-digit SSNs will be backfilled under randomization each year, representing fewer than 0.0008 percent of all serials covered by our crosswalk.

For all other five-digit SSNs, either all of their serial numbers were exhausted by June 24, 2011 and none are available to be randomized,[20] or none of their serial numbers were issued as of June 24, 2011 and we have flagged that fact in the crosswalk.

A.2 Using the Crosswalk

This appendix presents additional details on usage of our crosswalk.

A.2.1 Immigration Months—Supplementary Cases

We presented case studies of immigration in 2001 and 1956 in the NARA Numident. These additional case studies show that the relationship between immigration and SSN issuance in the 2001 and 1956 cases is not unique, and also appears in 1991 and 1977.

1991—El Salvador civil war and TPS designation By 1989, El Salvador's decade-long civil war displaced more than half a million undocumented Salvadorans to the United States (Maydans, 1990, p. 30). Amid brutal paramilitary retaliation against civil society and the

[20]The only serial numbers that were never assigned are numbers that were applied for on paper, followed by a loss of the paper form before it could be entered into the SSA database, either individually or as an entire pre-printed block of numbers (for example during shipment back to the Central Office). SSA procedure required that such numbers be marked void and not re-used, so that an applicant could later re-establish the account through alternative procedures (SSA, 2010).

church, Congress found support to legislate a solution (136 Cong. Rec. 36837-43 and 36831, 1990). The Immigration Act of 1990 (Pub. L. 101-649) created TPS, and granted TPS to Salvadorans: from the law's enactment on November 29, 1990, Salvadorans physically present in the United States became eligible to apply for TPS and an EAD either before or during a 6-month registration period ending June 30, 1991, later extended to October 31, 1991 (Pub. L. 102-65). The judicial settlement in American Baptist Churches vs. Thornburgh (ABC; 1990) drove registration that summer by offering additional protection from deportation to Salvadorans who applied for TPS or ABC benefits by June 30, 1991, with notice of these benefits mailed to eligible Salvadorans by March 31, 1991 (Rubin, 1991). In addition, the INS reduced fees on May 15 under Congressional pressure (H.R. Judiciary Comm., No. 15, 1991, p. 1-9 and 74). At that time, the INS was required to issue the EAD within 60 days (Nelson, 1987, p. 16228), resulting in a one- or two-month lag to SSN issuance.

In January 1991, one month after the law's enactment and the first month of the registration period, the NARA Numident shows SSN issuance rose 1.5-fold, and in February issuance rose 2.5-fold. By May and June, two months after the ABC letter, SSN issuance was four-fold. By July and August, two months after fees were reduced and shortly after the ABC deadline and the original TPS deadline, SSN issuance was six-fold. Consistent with the extension, issuance continued to December 1991, one month after the end of registration, and declined to baseline by February 1992.

Figure A3 shows that known SSN issue months and median issue months from our crosswalk, for the 1991 TPS designation, are again nearly identical.

1977—Soviet-Jewish "dropouts" In the early 1970s, the USSR granted exit visas to Soviet Jews for onward travel via Vienna to Israel. After Austrian authorities intervened in 1973 to let Soviet Jews choose their own destination, a growing number of "dropouts" sought to settle in the United States instead of Israel. By 1976 half of the Soviet Jews diverted to the INS processing center in Rome, seeking entry to the United States (Lazin, 2009, p. 19). We identify the following events in the NARA Numident and our crosswalk:

1. With the USSR increasing exit quotas in 1977, about 700 "dropouts" sought entry to the United States each month. These arrivals exhausted US conditional entry quotas under Sec. 203(a)(7) of the Immigration and Nationality Act, and a backlog developed in Rome. On January 13, 1977, Attorney General Levi paroled 4,000 refugees into the United States under Sec. 212(d)(5)[21]to clear the backlog, manage cost, and maintain Italian and Soviet cooperation. By the end of January one-third of the 4,000 were paroled or issued visas, and some had entered the United States (S. Judiciary Comm., No. 96-1, 1979, p. 105 and 197; Staats, 1977, p. 36, 46, and 78; Kissinger, 1976; Carlin, 1989, p. 184).

2. As the signatories to the Helsinki Final Act met in Belgrade from October 4, 1977 through March 9, 1978 to review East-West cooperation in Europe, the USSR increased emigration quotas to defend its rights record: from 1,150 per month in 1976, it issued 1,800 exit visas per month during the meetings and 2,000 in March (Snyder, 2011, p. 106; CSCE, 1978, p. 5-6 and 8). To address a growing backlog in Rome, Attorney General Bell paroled 5,000 Soviet refugees on December 14, 1977, with the quota fully used by April 1978 (S. Judiciary Comm., No. 96-1, 1979, p. 197 and 206; INS, 1978, p. 16).

[21]For a detailed discussion of the parole authority, see Bruno (2020).

3. The USSR sought better trade relations with the US, which the Jackson-Vanik amendment to the Trade Act of 1974 linked to free emigration. From 1978 through 1979, the USSR increased exit quotas to record levels, persuading Representative Vanik and President Carter to put trade normalization on the table for Carter's summit meeting with Chairman Brezhnev in June 1979 (Rosenburg, 2015, p. 424; Kaiser, 1979; Shulman, 1979). Most exiting Soviet Jews, 1,400 per month, sought entry to the United States, contributing to a backlog of 5,000 refugees in Rome. The Attorney General authorized parole for 12,000 Soviet refugees from June 14, 1978 through April 30, 1979 (Nash, 1983, p. 394-397). Processing began slowly, until on October 18, 1978 Congress authorized funds to address the growing strain on the INS and resettlement agencies (Pub. L. 95-481; Martin, 1982, p. 95).

4. The USSR continued to outpace expectations: by March 14, 1979, the Rome backlog ballooned to more than 10,000 Soviet refugees, with 3,400 arriving monthly. Under pressure from Congress, the Attorney General authorized parole for 25,000 from April 30 through September 30, 1979, with the INS and the State Department hiring staff to clear the backlog by the end of September. By September, Congress expected to pass the Refugee Act of 1979 to eliminate the need for parole (S. Judiciary Comm., No. 96-1, 1979, p. 23 and 185; Nash, 1983, p. 394-397 and 415-416). When the Act stalled, the Department of State requested an extension and an additional monthly quota of 3,000 parole entries for Soviet refugees until the passage of the Act, which Attorney General Civiletti granted on October 23, 1979 (Nash, 1983, p. 415-416). From October 1979 through March 1980, the INS admitted the monthly quota of 3,000 (S. Judiciary Comm., No. 96-55, 1980, p. 66).

5. In 1980, the USSR began to tighten emigration: their December 24, 1979 invasion of Afghanistan ended the negotiations over freer trade and arms limitation that motivated their loosening thus far. From January 1980, exit permits required evidence of "first-degree kinship" with sponsors in Israel (Rutland, 2017). From a total of 4,867 refugees arriving in Rome in October 1979, numbers fell to 2,563 in March 1980. Upon the passage of the Refugee Act of 1980, effective April 1, the Administration set an optimistic new monthly quota of 2,500; but monthly arrivals continued to fall, reaching 680 in July (S. Judiciary Comm., No. 96-55, 1980, p. 66, 71-72 and No. 96-79, 1980, p. 68).

Figure A4 shows that the NARA Numident closely tracks these events, and that known SSN issue months and median issue months from our crosswalk are nearly identical: The January 13, 1977 parole (1) appears as an uptick in SSN issuance in February, followed by a peak in March and elevated issuance in April; the lag to SSN issuance is two to three months, consistent with INS processing times in Rome. The December 14, 1977 parole (2) appears as another uptick in February, rising to a plateau from March through May; the lag is two months from the end of the meetings. The June 14, 1978 parole (3), two to three times the size of the first two programs, has its first impact two months later in August, but the largest impact arrives in December, two months after new funds are authorized to support resettlement; with more arrivals than expected, the quota is exhausted early, so SSN issuance falls earlier than expected in April 1979. The April 30, 1979 parole (4), five to six times the size of the first two programs, has a substantial impact on SSN issuance two months later in June and remaining at an exceptionally high level through November, a two month lag from the backlog's planned clearance; after the program is extended, SSN issuance remains elevated through June 1980, three months after parole is replaced with the lower refugee quota. Finally, as the USSR restricts exit permits (5), SSN issuance declines.

A.2.2 Immigration Months—Limitations

Figure A6 suggests four limitations to the use of issue months as immigration months before 1972, revealed by the Hungarian case. First, some employment was not covered and some workers applied for an SSN only when they began covered employment.[22] Second, until 1962, spouses were rarely issued SSNs if they did not work. Third, until 1972, spouses did not need their own SSNs to claim benefits. Fourth, also until 1972, children were rarely issued SSNs until they began to work. Hungarian escapees entered the United States in 1956-57, and from 1958-72, a period of fifteen years, fifteen cohorts of children who arrived in 1956-57 obtained their SSNs as they reached age 14-16; since few children arrived after 1957, issuance after 1972 continues at progressively greater ages. Throughout this period, we see delayed issuance driven by transitions into covered employment, mixed with issuance to post-1957 arrivals. By the mid-1970s, most Hungarian escapees had been enumerated, some with issue months beyond the true immigration month.[23]

These limitations may be addressed by linking family members. When family members enter together, those entering the covered sector first are likely to have the SSN issue month that best represents the immigration month. Opportunities to link family members depend on the data source, and linkage may be aided by names and the birth dates and places if available, and the SSN issue states and months from our crosswalk.

These limitations do not affect the use of issue months as immigration months from 1972 onwards to the same degree, for example in the 1977, 1991, and 2001 cases above. Issue dates more reliably indicate immigration months in later years because the 1972 Social Security Amendments (Pub. L. 92-603, Section 137) required SSA to assign SSNs to all legally admitted non-citizens at entry, or later when they gained work authorization.[24]

A.3 Building the Probability Table

This appendix provides detail on the extract, exclusions, weights, and cells used to calculate the probabilities in our table.

[22]See Nelson (1985) for an overview of covered and non-covered employment. More workers had a SSN than were covered, but in early years the issue age varied according to the age at first covered employment.

[23]Rough calculations using the NARA Numident suggest that for male escapees aged 18-62 in 1957, who were enumerated by 1977, 60 percent were enumerated by 1957, 83 percent by 1964, and 98 percent by 1973. For female escapees the comparable figures are 20, 47, and 93 percent. For children of both genders aged 0-13 in 1957, the figures are 0, 30, and 89 percent. For seniors of both genders aged 65-99 in 1957, the figures are roughly 2, 61, and 96 percent. These figures illustrate the greater reliability of working-age mens' issue months as a proxy for immigration months, but otherwise are not definitive. The challenge is that the number of Hungarians that arrived in 1956-57 and was enumerated from 1958-77 is similar to the number that newly arrived from 1958-77 (INS, 1977, p. 67), each making up a third of the 1956-77 total: thus, the fraction of 1956-57 arrivals that was enumerated by 1957 is likely closer to one-half than one-third. Regardless, it is clear that some SSN issuance was immediate and some was delayed.

[24]Foreign-born children in non-immigrant statuses are not always eligible for an SSN (Kolker and Morton, 2023), and may still need to be linked to their parents to obtain reliable immigration months after 1972. It may be easier to link children and parents in more recent years: for example, in IRS records, children can be linked to parents through dependent claiming starting in tax year 1996 (birth cohort 1978).

Extract To calculate the probability of foreign birth conditional on sex and the issue state, year, and age, we need one observation for each SSN in the NARA Numident. Goldstein and Breen (2022) report that entries for a given SSN may contradict each other, or may be incomplete. So, we select the most credible non-missing value for each SSN record and each core variable—foreign birth, sex, and birth date—to build a composite entry.

We resolve discrepancies and edit variables as needed: in particular, we recover missing birth country and foreign birth indicators where possible, based on recorded birth cities and non-citizen codes; we recover missing sex from names, suffixes, titles, and the forms used to make entries; and we ensure consistency between birth dates across records, with issue, claim, and death dates, and with internal updates to birth dates.

We assume that observations are missing at random (MAR; see Little and Rubin, 1987) conditional on sex and the issue state, year, and age.

However, the the birth place is lost for some SSNs in the SSA Numident.[25] The birth place may not be MAR: foreign-born SSN holders in the NARA Numident might be more or less likely to have a recorded birth place. So, we impute foreign birth where it is missing in the NARA Numident, in four stages: First, we build a probability table for non-missing records by sex and the issue state, year, and age. Second, we build probability tables by first, middle, and last name, the ZIP in the death file, and the District Office in the claims file. Third, we code foreign birth based on each variable and combinations of the variables when the classification is near certain. Fourth, we predict the remainder of values using a logistic regression on the merged probability tables and other variables derived from claim and death records. The logistic model correctly classifies 88 percent of records with roughly 12 percent false positives and false negatives in the full NARA Numident, and 84 percent with roughly 16 percent false positives and false negatives in the NARA Numident claims file, which is the source of most of the missing data on foreign birth.

To minimize the influence of any MAR violation, the results presented in Figure 3 and in our probability table are based on our extract imputing foreign birth where it is missing. Using only non-missing foreign birth gives nearly-identical results.[26]

Exclusions To estimate the probability of foreign birth for the resident population, we exclude (most of) the non-resident population.

For birth cohorts from 1898 onward, it is not necessary to explicitly exclude the non-resident population. The NARA Numident consists of records that include a death entry, and deaths are recorded primarily for the resident population, so non-residents are implicitly

[25]Missing birth place minimally affects the current population, and does not affect the working population: birth place is missing for 45-85 percent of individuals in birth cohorts 1890-1914, 15 percent of those in birth cohorts 1915-25, 10 percent in 1926-49, 5 percent in 1950-55, falling to under 1 percent from 1956.

[26]The scope for imputation to change the results is limited. Foreign birth is missing primarily for SSN account records with a claim in the Numident claim file, which was used until 1975-76 (rarely 1977-78). Most claims are for retirement or death benefits by birth cohorts before 1911-14 (age 62-65 in 1976) or to a limited extent before 1957 (age 19 in 1976). Because we restrict the extract to SSN account records with a death in the Numident death file, both native- and foreign-born account holders are typically residents eligible for a life (disability or retirement) or death claim. This mitigates the most plausible MAR violation, which is that SSN holders with a claim may more likely be natives, due to eligibility. Thus, while it is prudent to impute missing values, it is relevant only for few birth cohorts, in a manner addressed by sample restrictions.

excluded from the NARA Numident.[27]

For birth cohorts up through 1897, it is necessary to exclude the non-resident population. These individuals are presumed deceased in 2007 and are included in the NARA Numident regardless of whether there is a death entry in their records. Because they are not screened on having a death entry, these individuals are much more likely to be non-residents.

We exclude the older non-resident population by excluding Numident records that do not have a death entry.

Because the NARA Numident provides population coverage for birth cohorts up through 1897, researchers with access to full SSNs should merge the SSNs to the cleaned data extract of the NARA Numident that we use to build our probability table (before records with no death entry are excluded), instead of imputing foreign birth from the probability table.

Weights We weight the NARA Numident to support analyses that are not conditioned on sex, age, and year. We weight Numident records based on sex-specific mortality tables: weights are inversely proportional to the risk of death between the issue year and 2007, conditional on survival to the issue year, modified to account for incomplete death recording in certain years in the original Numident death file. Death records appear to cover the entire eligible population from 1988, and most of the eligible population from 1977; thus, we treat 1977 as the first year at risk for SSNs issued before 1977.

We find that weights do not affect the probability table, because it is already conditioned on sex, age, and year.

Aggregation Our probability table includes a row for each sex, state, year, and age cell seen in the NARA Numident. We first calculate the probability of foreign birth within each cell that contains at least ten observations. We fill smaller cells by step-wise aggregating them until they contain at least ten observations: steps involve using multiple issue years, multiple issue ages, both sexes, and multiple similar states at once. In all cases we use weights to account for differences in mortality risk between cells. Aggregation steps preserve the sum of weights, and are sequenced to maximize similarity between cells grouped together.

Missing Cells We fill missing cells from non-missing cells, again following a step-wise aggregation approach. Missing cells are marked in the probability table, and do not have representative weights. Figures A9 and A10 show how missing cells are filled.

A.4 Using the Probability Table

Our probability table covers SSNs issued from 1936-2011.

Probability table users may impute foreign birth for all birth cohorts up through June 2011, in any year of SSN issuance. After SSN randomization in 2011, it is not possible to determine the date of issuance from the SSN, but it is possible to determine that the SSN was randomized, since our crosswalk does not provide an issue month for randomized SSNs.

[27] Deaths are recorded for small numbers of non-residents including citizens and non-citizens eligible for benefits: see SSA POMS GN 00304.005 *Preferred Evidence of Death* and GN 02402.230 *Death of an International Direct Deposit (IDD) Beneficiary.*

Individuals with randomized SSNs, but birth dates up through June 2011, are almost always foreign-born, because by 2011 nearly all native-born individuals obtained their SSNs under SSA's Enumeration at Birth (EAB) program.[28] Because the birth cohort of 2011 is only 13 years old in 2024, this extension covers the entire work-eligible population.

[28]Probability table users with access to the IRS DM1 file may further impute native and foreign birth for individuals born *after* SSN randomization, using the issue date recorded in that file for all SSNs issued under randomization, except SSNs issued from July to December 2011. SSNs issued from July to December 2011 can be identified by the fact that they are not included in our crosswalk, and that they do not have an issue date in the DM1 file; DM1 users may impute an issue date in those months, no earlier than the birth date, before merging the probability table as suggested for issue years 2008 and later.

A.5 Appendix References

136 Cong. Rec., Pt. 25. 1990

Acquisti, Alessandro and Ralph Gross. 2009. "Predicting Social Security Numbers from Public Data," *Proceedings of the National Academy of Sciences* 106:27, pp. 10975-10980

American Baptist Churches vs. Thornburgh, 760 F. Supp. 796 (N.D. Cal. 1991)

Barron, Erma and Felix Bamberger. 1982. "Meaning of the Social Security Number," *Social Security Bulletin* 45:11, pp. 29-30

Bruno, Andorra. 2020. "Immigration Parole," Congressional Research Service Report No. 46570, Oct. 15, p. 1-21

Carlin, James. 1989. "Appendix," *The Refugee Connection: A Lifetime of Running a Lifeline*, MacMillan: London, p. 164-197

Commission on Security and Cooperation in Europe. 1978. "The Belgrade Followup Meeting to the Conference on Security and Cooperation in Europe: A Report and Appraisal," H.R. Committee Print, Government Printing Office: Washington, D.C., p. i-105

Kolker, Abigail and William Morton. 2023. "Noncitizen Eligibility for Employment Authorization and Work-Authorized Social Security Numbers (SSNs)," Congressional Research Service Report No. R47483, Mar. 22, p. 1-21

Government Accountability Office. 2020. "Temporary Protected Status: Steps Taken to Inform and Communicate Secretary of Homeland Security's Decisions," *Report to Congressional Requesters* GAO-20-134, Apr. 3

H.R. Judiciary Comm., No. 15. 1991. Implementation of the Immigration Act of 1990: Hearing Before the Subcommittee on International Law, Immigration, and Refugees

Immigration and Naturalization Service. 1977. *1977 Annual Report of the Immigration and Naturalization Service*

Immigration and Naturalization Service. 1978. *1978 Annual Report of the Immigration and Naturalization Service*

Kaiser, Robert. 1979. "Trade Benefits for Russia, China Eyed: Rep. Vanik, Co-Author of Curbs, Sees Possibility of Removing Them This Year," *Washington Post*, January 5

Kestenbaum, Bert. 2012. "Consequences of Social Security Number Randomization," accessed online (26 May 2023) at: `https://nces.ed.gov/FCSM/pdf/Kestenbaum_2012FCSM_VII-C.pdf`

Kissinger, Henry. 1976. Letter to Attorney General Edward H. Levi, December 17, Washington, D.C. Accessed online: `https://www.fordlibrarymuseum.gov/library/document/0019/4520887.pdf`

Lazin, Fred. 2009. "'We Are Not One': American Jews, Israel, and the Struggle for Soviet Jewry," *David W. Belin Lecture in American Jewish Affairs* Vol. 16. University of Michigan Library: Ann Arbor

Martin, David. 1982, "The Refugee Act of 1980: Its Past and Future," *Michigan Journal of International Law* Vol. 3, No. 1, p. 91-123

Maydans, Seth. 1990. "Help for Salvadorans, With a Catch," *New York Times* Dec. 2, p. 30

Nash, Marian. 1983. "Refugees," *1979 Digest of United States Practice in International Law*, Department of State Publication No. 9374, p. 389-416

Nelson, Alan. 1987. "Control of Employment of Aliens," *Federal Register* Vol. 52, No. 84, p. 16216-16228

Pub. L. 92-603, 86 Stat. 1329. 1972. Social Security Amendments of 1972

Pub. L. 95-481, 92 Stat. 1591. 1979. Foreign Assistance and Related Programs Appropriations Act of 1979

Pub. L. 101-649, 104 Stat. 4978. 1990. Immigration Act of 1990

Pub. L. 102-65, 105 Stat. 322. 1991. To amend the Immigration Act of 1990 to extend for 4 months the application deadline for special temporary protected status for Salvadorans

Rosenburg, Victor. 2015. "Refugee Status for Soviet Jewish Immigrants to the United States," *Touro Law Review* Vol. 19, No. 2, p. 419-450

Rubin, Robert. 1991. "Ten Years After: Vindication for Salvadorans and New Promises for Safe Haven and Refugee Protection," 68 *Interpreter Releases* 97, Jan. 28

Rutland, Suzanne. 2017. "Conflicting visions: debates relating to Soviet Jewish emigration in the global arena," *East European Jewish Affairs*, Vol. 47, No. 2-3, p. 222-241

S. Judiciary Comm., No. 96-1. 1979. The Refugee Act of 1979, S. 643: Hearing Before the Committee on the Judiciary

S. Judiciary Comm., No. 96-55. 1980. U.S. Refugee Programs: Hearing Before the Committee on the Judiciary

S. Judiciary Comm., No. 96-79. 1980. U.S. Refugee Programs, 1981: Hearing Before the Committee on the Judiciary

Shulman, Marshall. 1979. "Memorandum of Conversation," Jun. 6, Washington, D.C., Doc. 198 in *Foreign Relations of the United States, 1977-1980, Volume VI, Soviet Union*, Department of State. Accessed online: `https://history.state.gov/historicaldocuments/frus1977-80v06/d198`

Snyder, Sarah. 2011. "Follow-up at Belgrade: The United States Transforms the Helsinki Process," *Human Rights Activism and the End of the Cold War: A Transnational History of the Helsinki Network*. Cambridge University Press: Cambridge, England, p. 81-114

Social Security Administration. 1974. "Enumeration Activities," *Annual Report of the Social Security Administration for Fiscal Year 1973*, p. 50-51

Social Security Administration. 2005. "Social Security Number Allocations." Accessed online: `https://web.archive.org/web/20050830025754/http://www.ssa.gov/foia/stateweb.html`

Social Security Administration. 2010. "RM 10201.040 - Void Social Security Numbers (SSNs)," *Program Operations Manual System (POMS)*. Accessed online: `https://secure.ssa.gov/poms.nsf/lnx/0110201040`

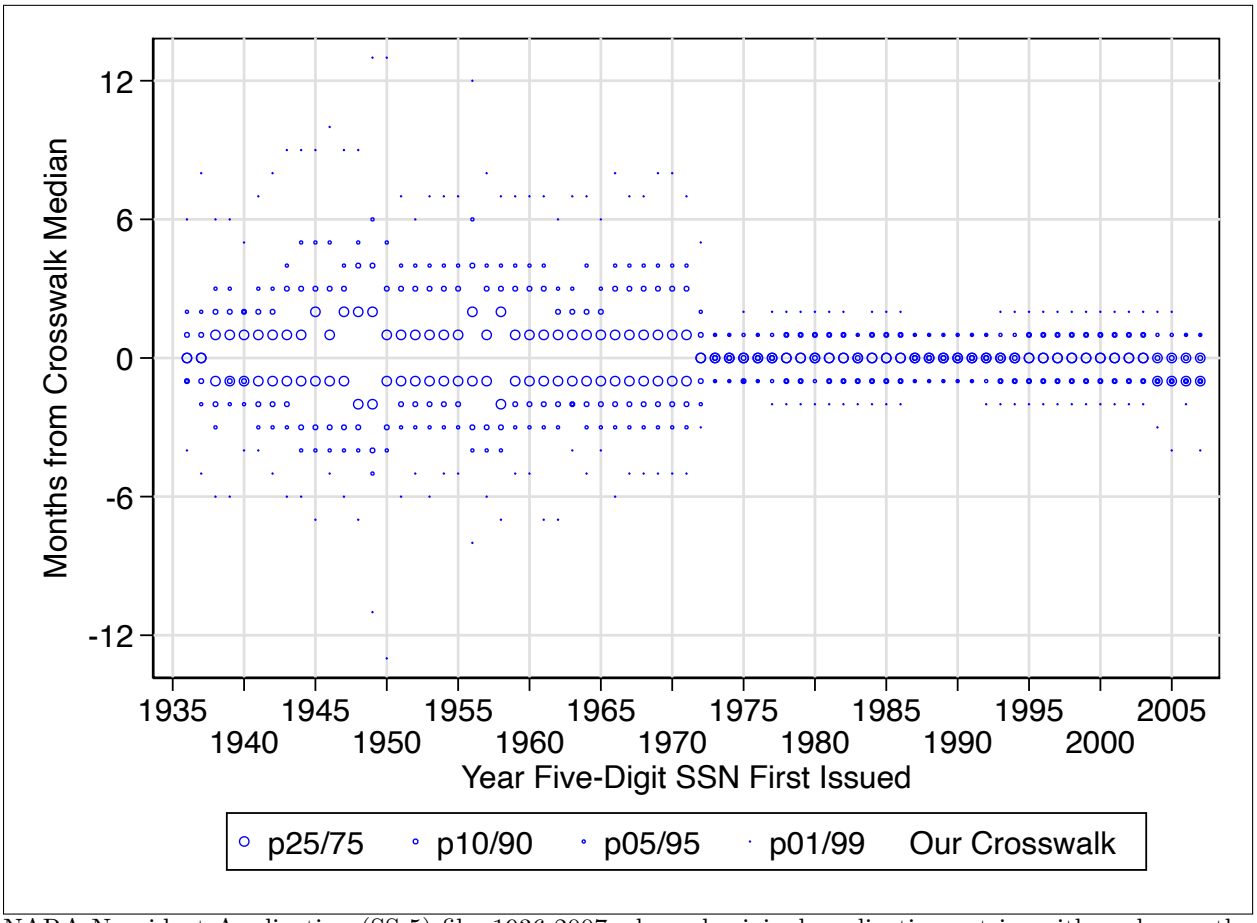
Social Security Administration. 2024a. "Highest Group Issued as of [Date]," for various dates starting from January 4, 1999, accessed online (29 September, 2024): `https://web.archive.org/web/19990203134611/http:/www.ssa.gov/foia/highgroup.htm`

Social Security Administration. 2024b. "High Group List And Other Ways To Determine If An SSN Is Valid," accessed online (March 7, 2024): `https://www.ssa.gov/employer/ssnvhighgroup.htm`

Staats, Elmer. 1977. "U.S. Assistance Provided for Resettling Soviet Refugees: Departments of State and Justice," *Comptroller General's Report to the Congress* ID-76-85, Jun. 20

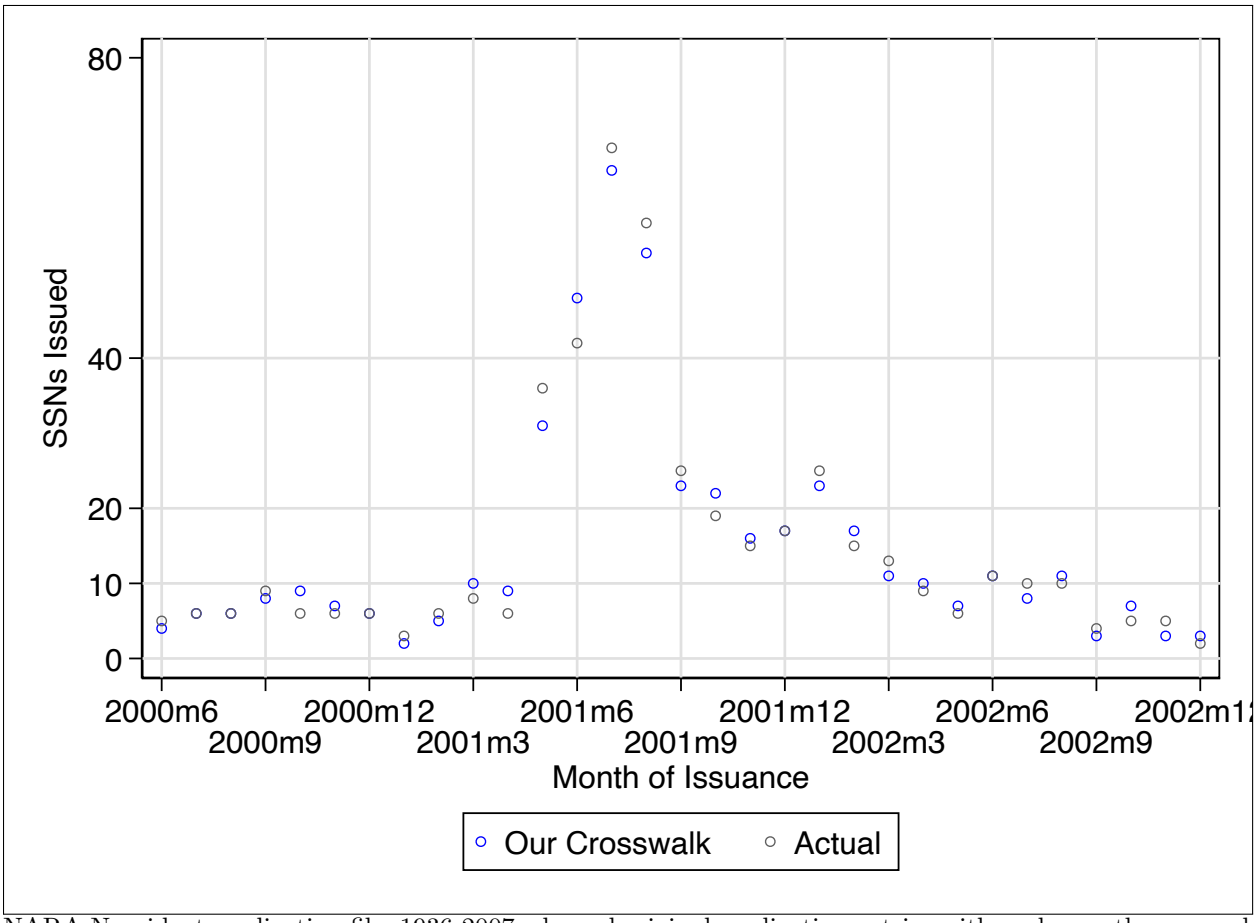
A.6 Appendix Figures

Figure A1: Distribution of Differences between Imputed and Actual Issue Months, Using Our Crosswalk



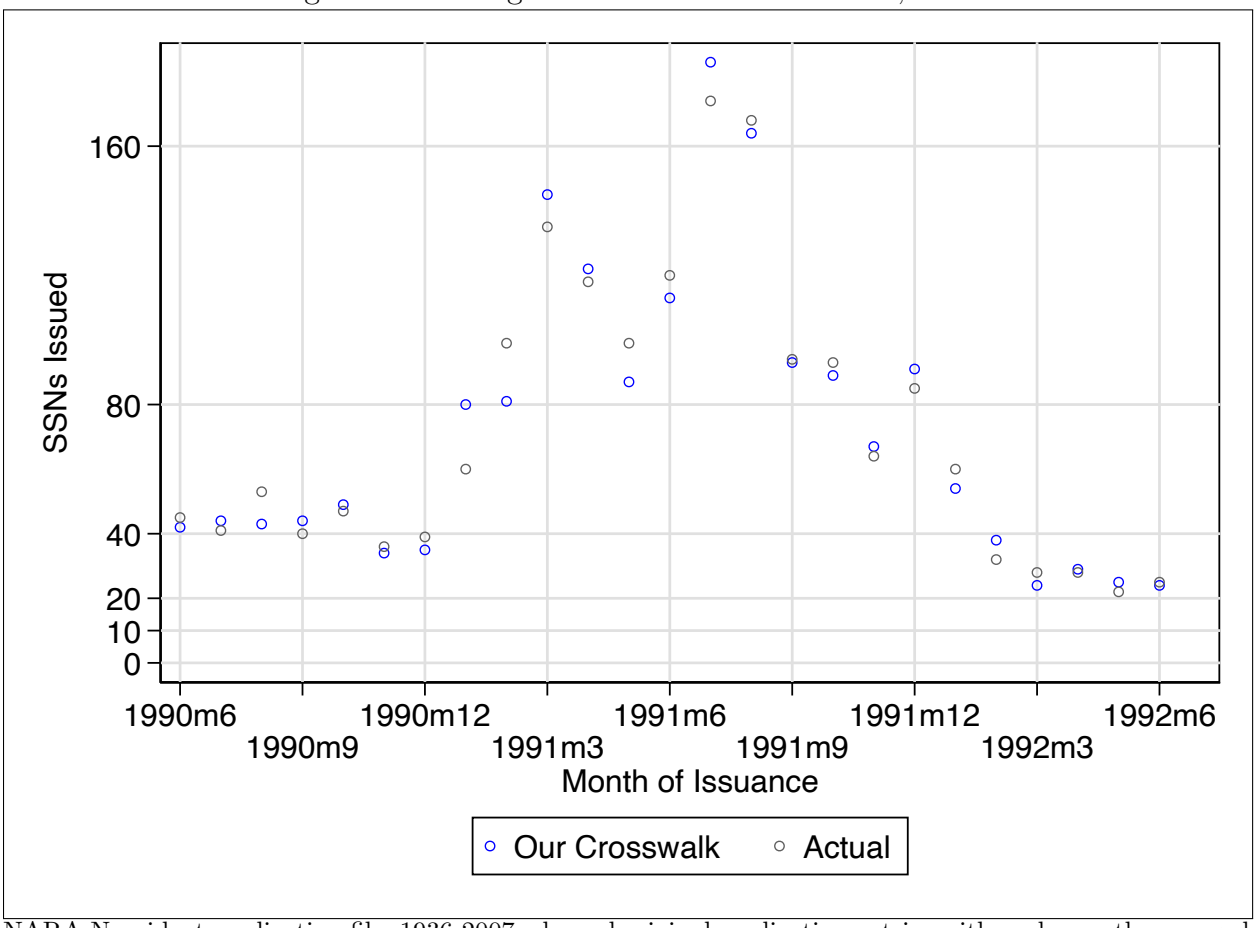
NARA Numident Application (SS-5) file, 1936-2007, cleaned original application entries with cycle months (N=35,580,763), merged to our crosswalk by five-digit SSN (e.g. 001-01). To show our crosswalk accurately represents true issue months, we calculate the difference in months between the cycle month and the crosswalk median month; for five-digit SSN 001-01, the cycle months are November and December 1936, and the crosswalk median is November 1936, for a difference of 0 or 1 months. We group entries by the five-digit SSN's year of first issuance and calculate the distribution; for 1936, the 1st percentile of differences is -4 months, the 5th and 10th percentiles are -1 month, the 25th and 75th percentiles are 0 months, the 90th percentile is 1 month, the 95th percentile is 2 months, and the 99th percentile is 6 months. The plotted points are percentiles of the differences in months, within the year first issued. Overall, this figure shows that cycle months are very near to the crosswalk median from 1936 to 2011, and that cycle months are nearer to the crosswalk median from 1972 onward.

Figure A2: Immigration Months: El Salvador, 2001



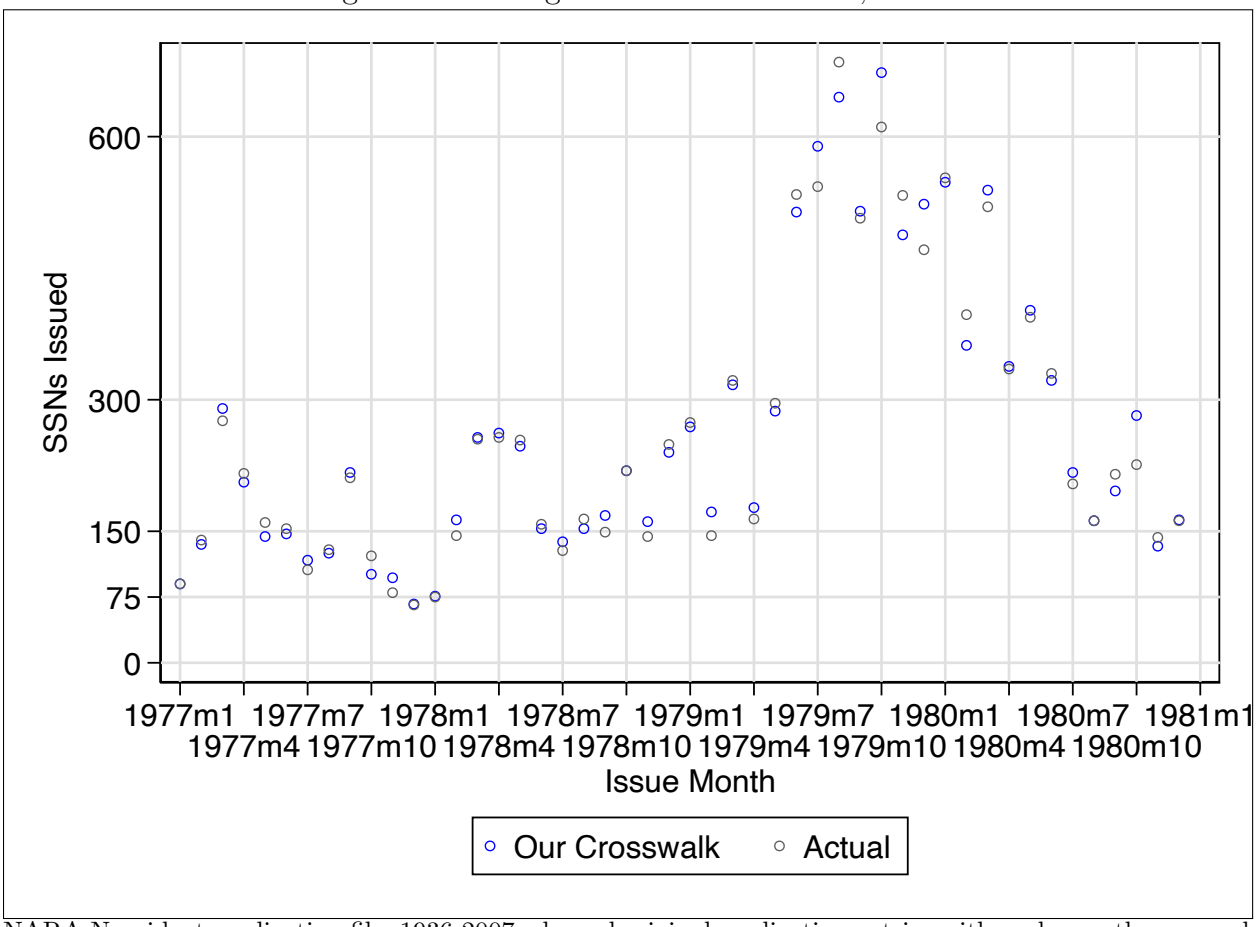
NARA Numident application file, 1936-2007, cleaned original application entries with cycle months; merged to best record of birth month, sex, and birth country code; and merged to our crosswalk for median cycle month. Restricted to number-holders born in El Salvador (N=17,320), issued SSNs from June 2000 through December 2002 according to NARA Numident (N=462) or our crosswalk (N=462). This figure shows that the issuance of SSNs to Salvadorans closely tracks their known dates of eligibility due to the 2001 Temporary Protected Status (TPS) designation of El Salvador, and that issue months imputed from our crosswalk reproduce the same time pattern. Salvadorans became eligible to apply for TPS and an Employment Authorization Document (EAD) from March 9, 2001 through September 9, 2002. EADs were issued within 90 days according to contemporary records, and could then be used to apply for an SSN. The figure shows that in May 2001, 2-3 months after Salvadorans became eligible to apply for an EAD, the first beneficiaries begin to obtain SSNs from the Social Security Administration (SSA), which suggests that the minimum application and processing time for the EAD and SSN together is slightly less than 90 days. Issuance peaks in July, 4-5 months after eligibility, suggesting that typical application and processing times for the EAD and SSN together were about 120 days. Most registration happens at the beginning of the eligible period, but SSN issuance does not return to baseline until September 2002, the end of the registration period.

Figure A3: Immigration Months: El Salvador, 1991



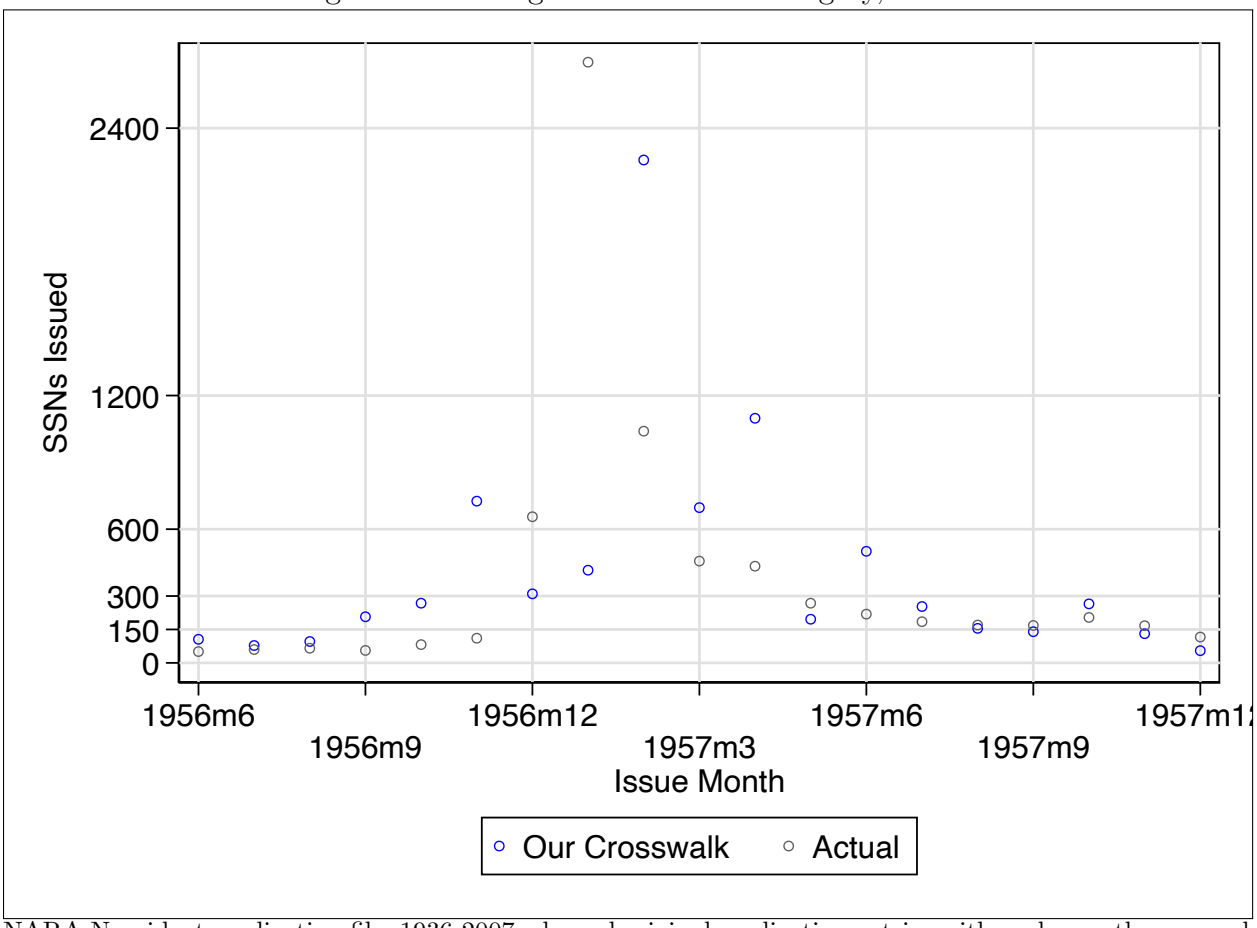
NARA Numident application file, 1936-2007, cleaned original application entries with cycle months; merged to best record of birth month, sex, and birth country code; and merged to our crosswalk for median cycle month. Restricted to number-holders born in El Salvador (N=17,320), issued SSNs from June 1990 through June 1992 according to NARA Numident (N=1,805) or our crosswalk (N=1,803). This figure shows that the issuance of SSNs to Salvadorans closely tracks their known dates of eligibility due to the 1990 Temporary Protected Status (TPS) designation of El Salvador, and that issue months imputed from our crosswalk reproduce the same time pattern. Salvadorans became eligible to apply for TPS and an Employment Authorization Document (EAD) from November 29, 1990 according to the law and from January 1, 1991 according to regulation, through June 30, later extended to October 31; in addition, Salvadorans were notified of protection from deportation under the American Baptist Churches vs. Thornburgh (ABC) judicial settlement, with a notice mailed on March 31 announcing a deadline of June 30 to claim ABC benefits; finally, the Immigration and Naturalization Service (INS) reduced application fees on May 15. EADs were issued within 60 days according to contemporary records, and could then be used to apply for an SSN. The figure shows that in January 1991, 1-2 months after Salvadorans become eligible to apply for an EAD under the law, the first beneficiaries begin to obtain SSNs from the Social Security Administration (SSA), which suggests that the minimum application and processing time for the EAD and SSN together is slightly less than 60 days. Issuance peaks in March and again in July, in the first case 2-3 months from eligibility under regulation and in the second case 3-4 months after the ABC mailing and 2-3 months after the fee reduction, suggesting a typical application and processing time of 90 days. Issuance remains near the peak in August, 1-2 months after the original TPS and ABC deadlines, and remains elevated through January 1992, 2-3 months after the final TPS deadline; these figures again suggest a minimum processing time just under 60 days and a maximum just over 90 days.

Figure A4: Immigration Months: USSR, 1977



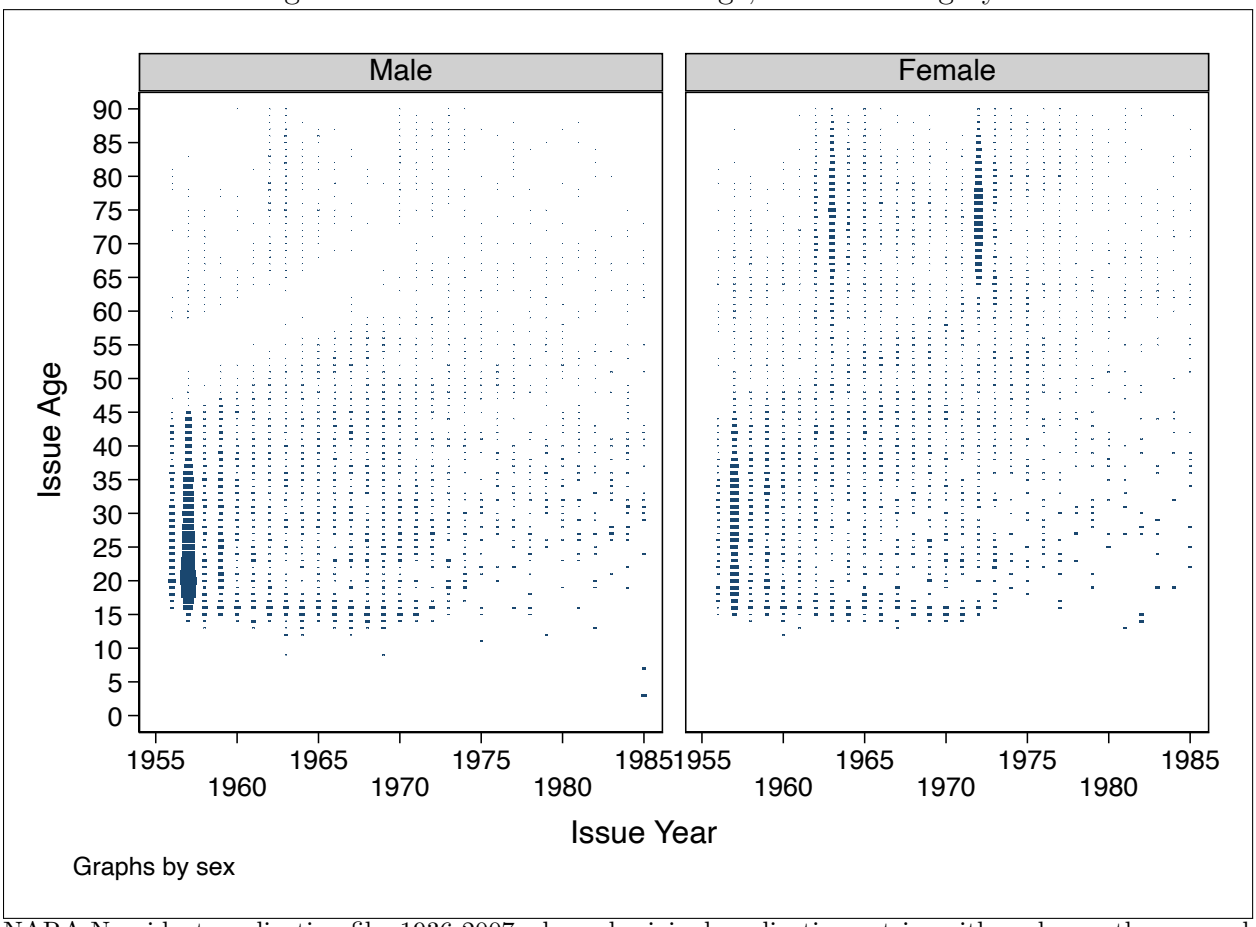
NARA Numident application file, 1936-2007, cleaned original application entries with cycle months; merged to best record of birth month, sex, and birth country code; and merged to our crosswalk for median cycle month. Restricted to number-holders born in the USSR (N=207,273), issued SSNs from January 1977 through December 1980 according to NARA Numident (N=12,672) or our crosswalk (N=12,650). This figure shows that the issuance of SSNs to Soviets (Soviet Jews) closely tracks their known exits from the USSR and admissions to the United States under various conditional entry and parole authorities, and that issue months imputed from our crosswalk reproduce the same time pattern. The flow of Soviet Jews to the United States was about 700 per month by the end of 1976, which serves as our baseline. The figure shows the following events: (1) On January 13, 1977, the Attorney General (AG) paroled 4,000 refugees into the United States; by the end of January one-third were paroled or issued visas and some entered the United States. Accordingly, SSN issuance rises in February within one month of the first arrivals, and peaks in March within one month of the remaining arrivals. (2) The AG paroled another 5,000 Soviet refugees on December 14, 1977, with the quota fully used by April 1978. Accordingly, SSN issuance rises just over one month later in February 1978 and reaches a high plateau in March through May, for a one-month lag. (3) The AG paroled another 12,000 Soviet refugees on June 14, 1978 through April 30, 1979, but they were not moved as quickly; on October 18, Congress authorized funds to accelerate resettlement. Accordingly, SSN issuance rises just over one month later in August but does not rise substantially until the period from October through May 1979, again for a one-month lag. (4) The AG paroled another 25,000 Soviet refugees on April 30, 1979 through September 30, and both the Immigration and Naturalization Service (INS) and Department of State hired and transferred additional staff to raise the rate of resettlement and resettle the full quota of 5,000 per month; the AG then paroled 3,000 per month through March 1980. Accordingly, SSN issuance rises one month after April 30, 1979, meaning June, to 7-8 times the baseline level, and remains near that level through March 1980, rising quickly through August 1979 and then falling slowly towards March 1980. (5) The USSR began to restrict exit permits in January 1980 and monthly arrivals at the INS processing center in Rome reached baseline levels by July 1980. Accordingly, SSN issuance declines to near baseline by August 1980.

Figure A5: Immigration Months: Hungary, 1956



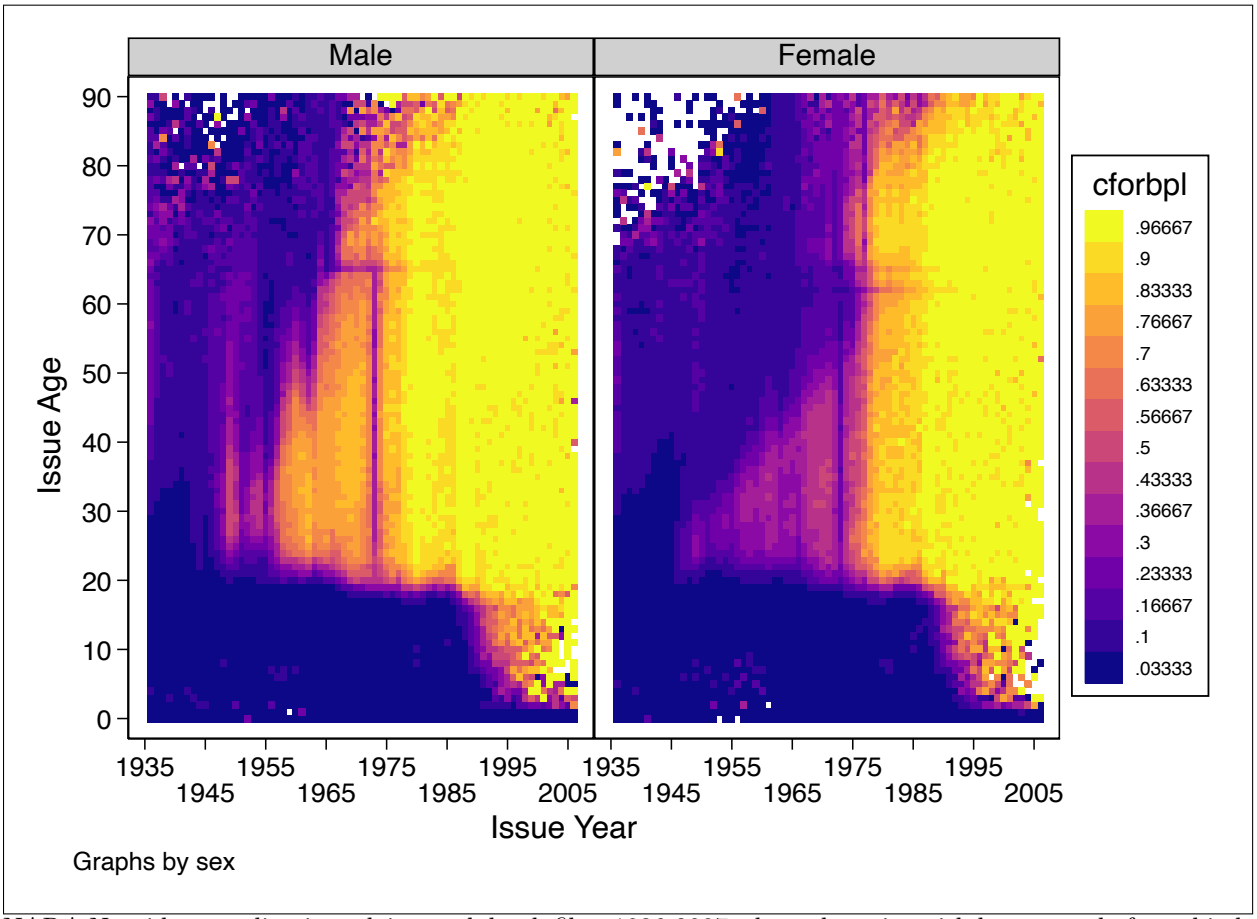
NARA Numident application file, 1936-2007, cleaned original application entries with cycle months; merged to best record of birth month, sex, and birth country code; and merged to our crosswalk for median cycle month. Restricted to number-holders born in Austria-Hungary (N=63,012), issued SSNs from June 1956 through December 1957 according to NARA Numident (N=7,206) or our crosswalk (N=7,140). This figure shows that the issuance of SSNs to Hungarian escapees closely tracks their known arrivals in the United States by airlift and sealift and their subsequent processing, and that issue months imputed from our crosswalk reproduce the same time pattern. The first Hungarians were airlifted to the United States arriving on November 21, 1956, with 5,000 arriving by December 1, 15,000 more arriving by January 2, 1957, 12,373 more by February 28, and a total of 32,000 by May 1; by December 31 a total of 38,000 had arrived. Processing generally took less than one month and included the issuance of a SSN. Accordingly, SSN issuance rises within the first month after arrival, in December 1956, peaks in January 1957 after the largest group of Hungarians arrives in the preceding month, falls by two-thirds in February after the rate of arrival falls to under half the peak rate, and returns to a stable but somewhat elevated level in May after emergency operations (including SSN issuance) conclude in April. Compared to our later examples, the relatively short lag between arrival and SSN issuance relates to three factors: first, emergency processing; second, the simpler procedure to apply for an SSN in the 1950s; and third, paper processing, which means that instead of a delayed cycle date for when the electronic record was first posted, the Numident records the actual application date. The crosswalk does not perform as well in this example, before 1972, as it does in our other examples, after 1972. It should not be a surprise that immigration months imputed from the crosswalk in earlier years are less precise than immigration months imputed from the crosswalk in later years, since we have seen from Figure A1 that issue months are generally less precise in earlier years. However, it is remarkable that immigration months imputed from the crosswalk are within 1-2 months of the true immigration months.

Figure A6: SSN Issue Year and Age, Born in Hungary



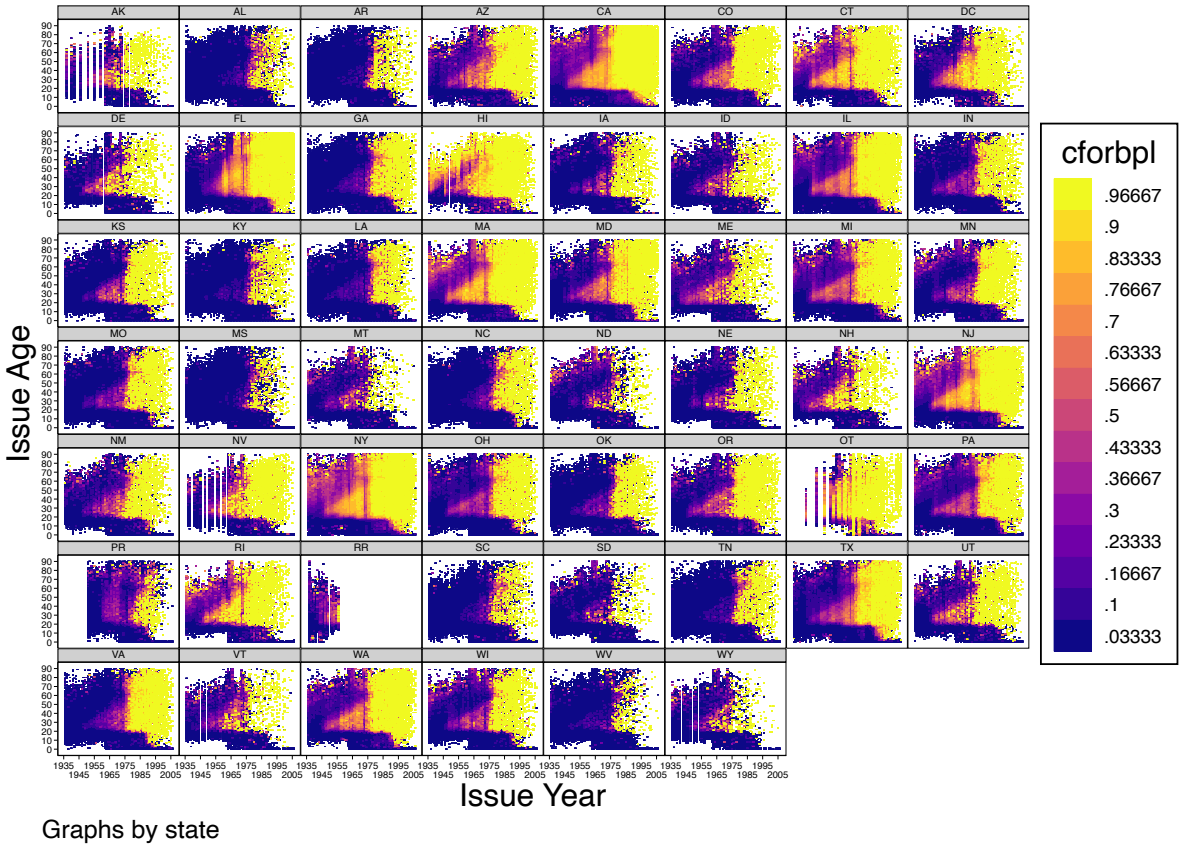
NARA Numident application file, 1936-2007, cleaned original application entries with cycle months; merged to best record of birth month, sex, and birth country code. Restricted to number-holders born in Hungary, issued SSNs from October 1956 through December 1985 (N=23,423). We define the issue age as the difference between the cycle month and birth month, divided by twelve and rounded down. We plot the weighted issue age by the issue year, in two panels by sex. The figure shows certain limitations affecting the use of the cycle month to indicate the immigration month prior to 1972. Reassuringly, we see that most adult Hungarian men were issued SSNs on arrival in late 1956 and in 1957. However, not all Hungarians were issued SSNs on arrival. Older Hungarian women and some older Hungarian men were issued SSNs in 1962-1964 by the IRS, when it began to use the SSN as a tax identification number and to require banks and other non-employer institutional payers to use the SSN to identify payees; most payees who did not already have an SSN did not have an SSN because they were retired, dependent, or working in a non-covered sector. In 1972, another group of older Hungarian women was issued SSNs by the SSA when it issued SSNs to all secondary beneficiaries. Hungarian children who were not working or work-eligible upon arrival also obtained their SSNs in later years as they began to work, leading to steady issuance at about age 16 for about 16 years after they arrived in the United States. Issuance at working ages in most years is to Hungarians who had worked in non-covered jobs and then switched to covered jobs, and to Hungarians who arrived after 1957. Finally, the light diagonal band in both panels, from about age 46 to 59 in 1957 up through age 62 to 75 in 1972, relates to data quality issues discussed in Appendix A.3.

Figure A7: Probability of Foreign Birth by Sex and Issue Year and Age



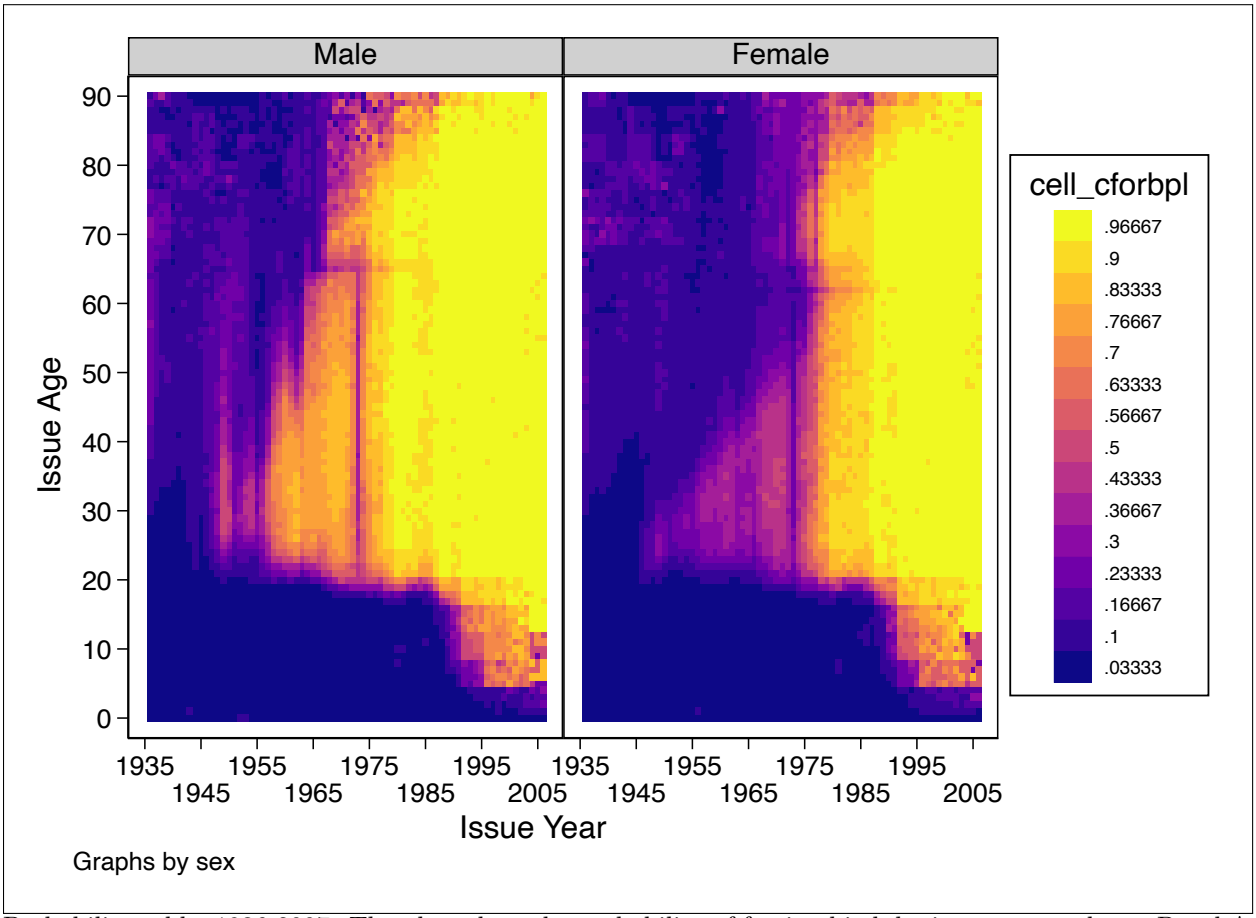
NARA Numident application, claim, and death files, 1936-2007, cleaned entries with best record of sex, birth date, and foreign birth; imputed where missing; merged to our crosswalk by five-digit SSN; restricted to SSN account records with known death date (N=49,379,843). The plots show the probability of foreign birth by issue year, and the issue age based on the birth month and crosswalk issue month; Panel A shows male applicants and Panel B shows female applicants. Issuance to foreign-born men is easier to distinguish by age of issuance for SSNs issued prior to 1986, and more immediately responds to changes in sectoral coverage, likely due to less-concentrated ages of entry into covered employment among native- and foreign-born women. Foreign-born women are less likely to obtain SSNs prior to the mid-1970s when SSNs became necessary for purposes other than work: since many native women worked before marriage, and many foreign-born women immigrated after marriage, foreign-born women are less likely to apply for an SSN for work alone.

Figure A8: Probability of Foreign Birth by Issue State, Year, and Age



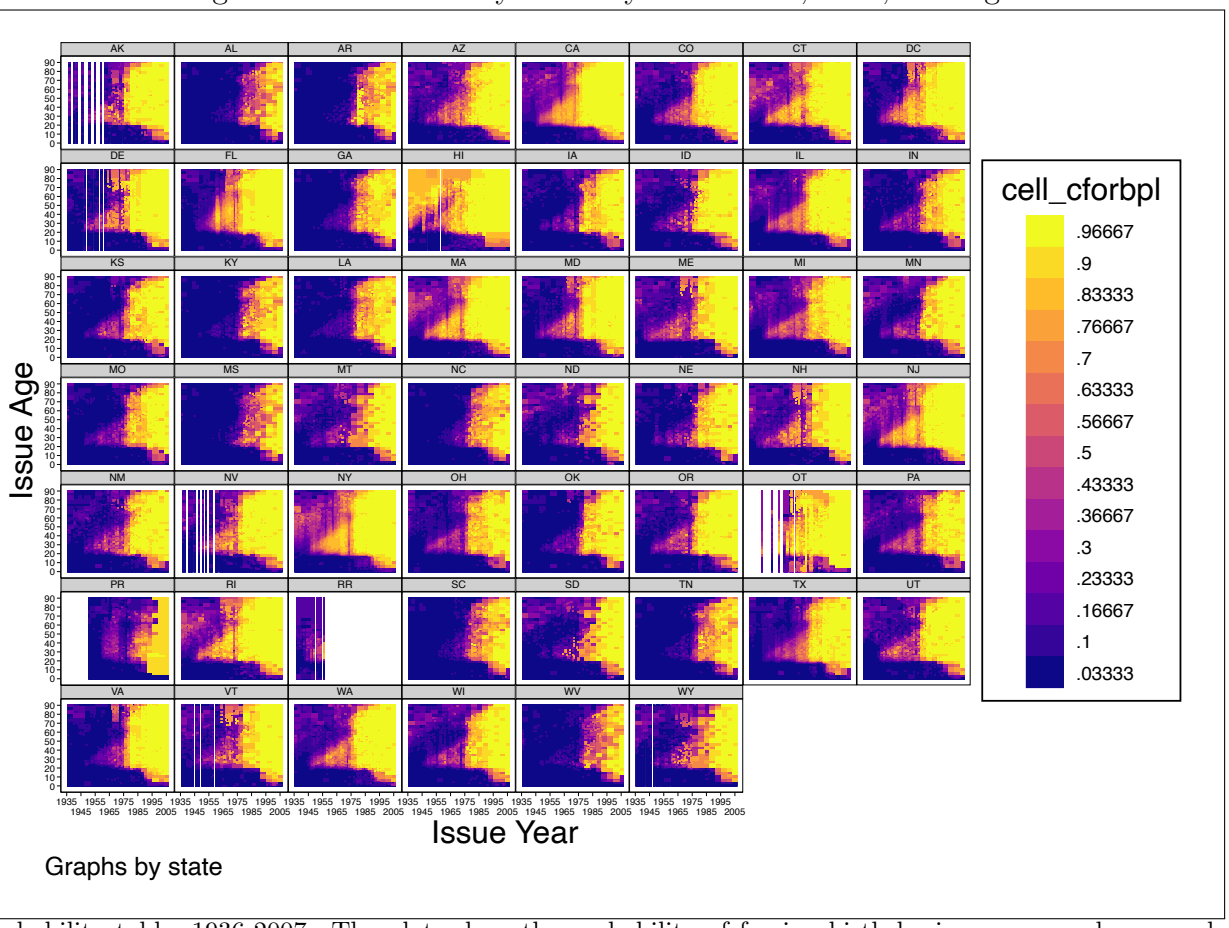
NARA Numident application, claim, and death files, 1936-2007, cleaned entries with best record of sex, birth date, and foreign birth; imputed where missing; merged to our crosswalk by five-digit SSN; restricted to SSN account records with known death date (N=49,379,843). The plots show the probability of foreign birth by issue year, and the issue age based on the birth month and crosswalk issue month; each panel shows applicants by state, Puerto Rico (PR), Railroad (RR), and possessions, territories, and overseas applicants (OT). Foreign birth is more common in some issue years, ages, and states than in others. This figure displays small and missing cells. Such cells are not quantitatively important for imputation, but in our probability table, are filled in as described in Appendix A.3. Further, some issue years are not assigned for certain issue states and places: Railroad issuance was discontinued; Puerto Rico issuance only began after Puerto Rico was covered by Social Security; small states including Alaska, Delaware, Hawaii, Nevada, Vermont, and Wyoming issued SSNs in all years but some five-digit SSNs spanned more than one year each, so that all SSNs issued in those five-digit ranges are assigned by our crosswalk to a subset of their actual issue years.

Figure A9: Probability Table by Sex and Issue Year and Age



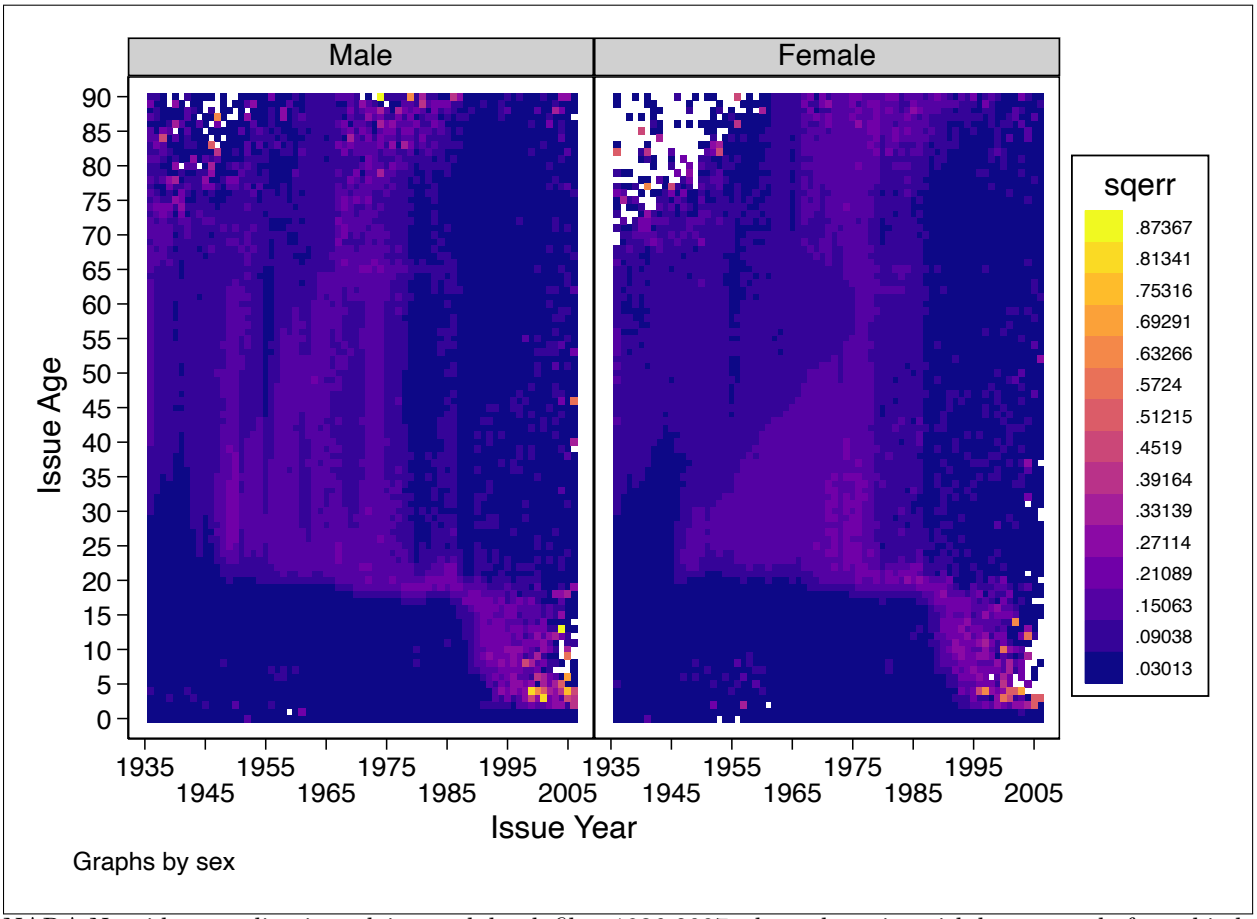
Probability table, 1936-2007. The plots show the probability of foreign birth by issue year and age; Panel A shows male applicants and Panel B shows female applicants.

Figure A10: Probability Table by Issue State, Year, and Age



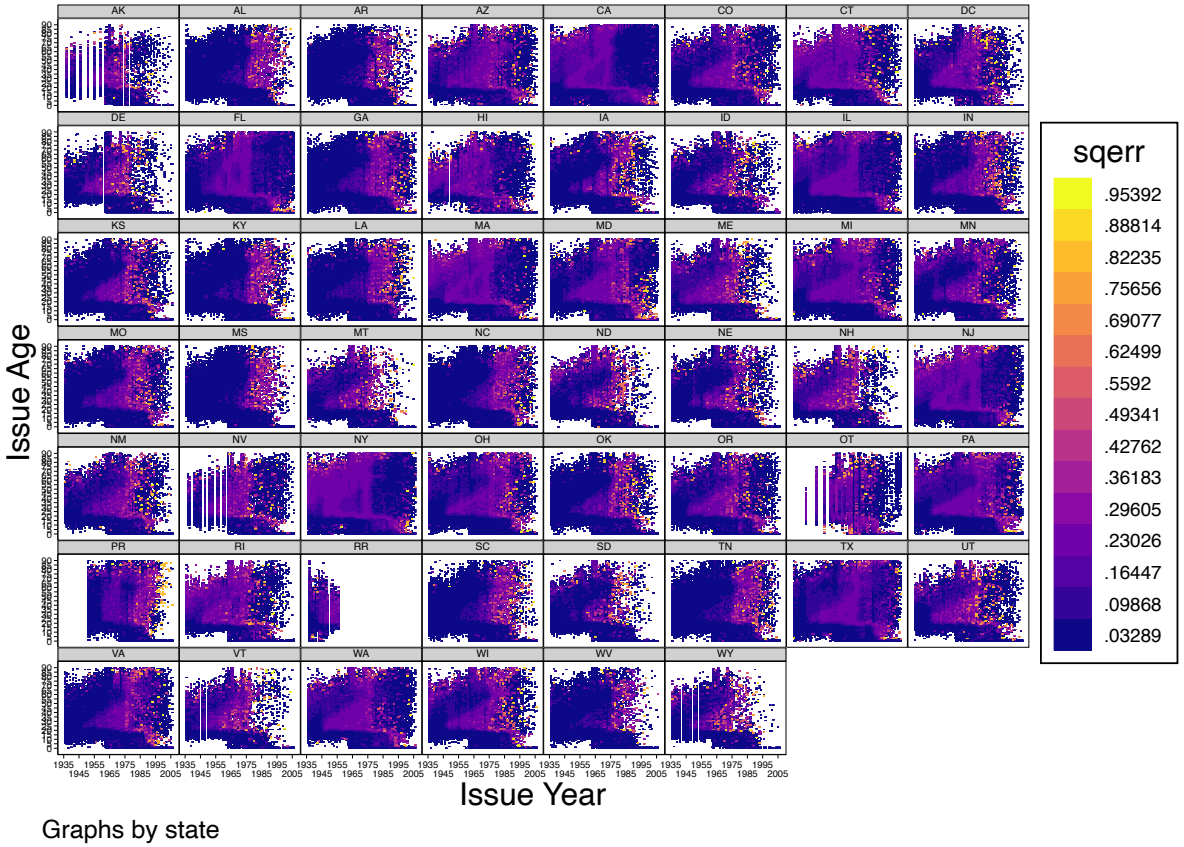
Probability table, 1936-2007. The plots show the probability of foreign birth by issue year and age; each panel shows applicants by state, Puerto Rico (PR), Railroad (RR), and possessions, territories, and overseas applicants (OT). Missing cells from the NARA Numident are filled in as described in Appendix A.3.

Figure A11: Mean Squared Error of Foreign Birth Imputation from Probability Table by Sex and Issue Year and Age



NARA Numident application, claim, and death files, 1936-2007, cleaned entries with best record of sex, birth date, and foreign birth; imputed where missing; merged to our crosswalk by five-digit SSN; restricted to SSN account records with known death date (N=49,379,843). The plots show the MSE of actual values of foreign birth minus imputed values of foreign birth by issue year and issue age, from our probability table: Panel A shows male applicants and Panel B shows female applicants. The side-by-side comparison shows that classification accuracy is similar for men and women, though modestly better for adult men before 1985.

Figure A12: Mean Squared Error of Foreign Birth Imputation from Probability Table by Issue State, Year, and Age



NARA Numident application, claim, and death files, 1936-2007, cleaned entries with best record of sex, birth date, and foreign birth; imputed where missing; merged to our crosswalk by five-digit SSN; restricted to SSN account records with known death date (N=49,379,843). The plots show the MSE of actual values of foreign birth minus imputed values of foreign birth by issue year and issue age, from our probability table; each panel shows applicants by state, Puerto Rico (PR), Railroad (RR), and possessions, territories, and overseas applicants (OT). The side-by-side comparison shows that classification accuracy is similar for all issue states, though low-immigration states have relatively higher (lower) classification accuracy for adults before (after) 1980, and high-immigration states show the opposite pattern. In all issue years, ages, and states, classification accuracy is better than both random chance and the issue-age cutoff imputation.