Projecting Trends in Undocumented and Legal Immigrant Populations in the United States*

Ryan Bhandari, University of Illinois, Chicago

Benjamin Feigenberg, University of Illinois, Chicago

Darren Lubotsky, University of Illinois, Chicago & NBER

Eduardo Medina-Cortina, University of Illinois, Urbana-Champaign

September 10, 2021

Abstract

We use administrative data on over 9 million Matrícula (identification) cards issued by the Mexican government between 2008 and 2017 to Mexican-born individuals living the United States to improve estimates of the undocumented foreign-born population. These cards are held by those who do not have legal status in the United States and therefore do not have other forms of valid identification. The key contribution of our work is to use this data to produce estimates of the undocumented population from Mexico and from other countries, carefully laying out the relevant issues, assumptions, and sources of uncertainty. The ability to use the Matrícula data to inform estimates of the undocumented population is particularly important because of the general lack of direct data on this group. Our preferred estimates indicate that there were on average 8.3-8.7 million undocumented Mexican individuals in the United States per year between 2008 and 2012 and 7.5-8.2 million between 2013 and 2017; both estimates are somewhat higher than the well-known estimates produced by the Pew Center. Our estimates of the undocumented immigrant population from other Latin American and Caribbean countries are more closely aligned with those from the Pew Center. Finally, we conclude that Matrícula data is unlikely to be useful in estimating the undocumented population from outside of the Latin America and Caribbean region.

*The research reported herein was performed pursuant to grant RDR18000003 from the US Social Security Administration (SSA) funded as part of the Retirement and Disability Research Consortium. The opinions and conclusions expressed are solely those of the author(s) and do not represent the opinions or policy of SSA, any agency of the Federal Government, or NBER. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of the contents of this report. Reference herein to any specific commercial product, process or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply endorsement, recommendation or favoring by the United States Government or any agency thereof.

1 Introduction

Basic facts about the size of the immigrant population, the fraction undocumented, and future trajectories are crucial for basic research on how immigration affects the economy and for the analysis of public policies, including the Social Security Administration's long-term projections and models. This work is hindered, however, by the lack of quality data on the number of undocumented immigrants that reside in the United States at a given point in time. Standard government surveys, such as the American Community Survey (ACS) and the Current Population Survey (CPS), and commonly used administrative records (such as tax records) do not indicate a person's legal status. Further, it clear that undocumented individuals are systematically under-counted in surveys (Passel and Cohn, 2019).

We use new data to improve the measurement of the undocumented population in the United States. Specifically, we use administrative data on over 9 million Matrícula (identification) cards issued by the Mexican government between 2008 and 2017 to its citizens who reside in the United States. Matrícula cards are issued by Mexican consulate offices to Mexican citizens residing abroad. Applicants for a card must document their Mexican nationality, their identity, and current place of residence. Cards are valid for five years and can be renewed. The card does not contain any information on an individual's legal status in the United States. The primary value of the card is as an identification card that can be used in the United States, including with many financial institutions, for the purpose of obtaining a driver's license in some states, and for the purpose of obtaining a taxpayer identification number through the federal government. Recent work has validated the quality and coverage of data on Matrícula cards using external data sources and concluded that it is reasonable to infer that all applicants are undocumented since there are no benefits to receiving an ID card for those in the U.S. legally (Caballero et al., 2018, Massey et al., 2010).

We use the Mexican Matrícula data to conduct three sets of analyses. First, we construct estimates of the Mexican-born undocumented population share across local areas and we estimate the total Mexican-born undocumented population at the national level. Then, in combination with data from the American Community Survey, we estimate the undocumented population that originates from countries other than Mexico. We show that undocumented population share estimates for the Mexican-born population, and for those from other Latin American and Caribbean countries, are highly correlated with estimates obtained from alternative methods.[1] Estimates of undocumented population shares for those from other parts of the world appear to be less reliable and thus we focus our analysis and discussion on estimates from Mexico and the rest of the Latin America and Caribbean region. Last, we leverage estimates of the undocumented population to project future population changes as a function of a range of economic, demographic, and policy parameters.

These Matrícula data have a number of potential advantages over existing data sources and methods to measure the undocumented population. As we discuss in more detail in Section 2, there are two broad alternative approaches to estimating the undocumented population. The first is a "residual"-based approach pioneered by Warren and Passel (1987) in which the undocumented population is inferred as the difference between measures of the total foreign-born population and the legally-resident population. The second method uses standard survey data, such as the American Community Survey, and imputes legal status to respondents based on their observable characteristics (Passel and Cohn, 2019; Borjas and Cassidy, 2019). Since Matrícula cards are issued almost exclusively to individuals who reside in the U.S. without legal authorization, they provide an external source of information on the Mexican-born undocumented population. We are therefore able to largely sidestep standard concerns about differential survey non-response (or sur-

[1] We follow the World Bank and define the Latin America and Caribbean (LAC) region to include South America, Central America, and the Caribbean.

vey mis-response) as a function of immigration status that may bias existing estimates. Since Matrícula cards must be renewed every five years, concerns related to unmeasured emigration, mortality, and status changes are mitigated to some degree. Finally, as an administrative data source, the Matrícula data are not subject to the small sample concerns that limit researchers' capacity to use existing methodologies to estimate undocumented populations at geographically disaggregated levels, such as by county or commuting zone. At the same time, the Matrícula data introduce new challenges. Most significantly, not all undocumented immigrants from Mexico acquire (or regularly renew) Matrícula cards, and so we must rely on estimated take-up and renewal rates to translate Matrícula-based counts into undocumented population counts.

An important contribution of our work is to use the Matrícula data to make estimates and future projections of the non-Mexican undocumented population. To do this, we start with new estimates of the Mexican-born undocumented population in U.S. counties and commuting zones in 2008 through 2012 formed by counting the number of valid cards in an area. Corresponding estimates of undocumented population shares are highly correlated with estimates derived from the American Community Survey in which we impute undocumented status using a method outlined in Borjas and Cassidy (2019). We then estimate a model that relates the Mexican-born undocumented population share in an area to the average characteristics of all Mexican-born respondents in the American Community Survey in that area. Using the parameter estimates from this model and the average characteristics of other groups in the ACS, we form estimates of the share of undocumented individuals from other parts of the world.

We validate this approach with out-of-sample predictions of local Mexican-born undocumented population shares in the period from 2013 to 2017, which are highly correlated with direct estimates of undocumented population shares based on Matrícula and ACS data from those same years. The performance of our predictive model for populations

born outside of Mexico is mixed. While undocumented population share predictions for those born elsewhere in the LAC region are highly correlated with imputation-based estimates, predicted shares of the undocumented population born outside of the LAC region are essentially uncorrelated with imputed values. To further validate our prediction model for those born in the LAC region, we take advantage of data from the Encuestas Sobre Migración en Las Fronteras de México (EMIF), which includes surveys of individuals deported from the United States to Mexico, El Salvador, Guatemala and Honduras. Specifically, we use EMIF data on interior deportees in conjunction with ACS files to verify that predictors of deportee status within the Mexican-born population are nearly identical to the predictors of deportee status within the population born in El Salvador, Guatemala, or Honduras.

We next turn to constructing national estimates of the undocumented population, separately for those born in Mexico and the remainder of the LAC region. One challenge in using our Matrícula card model to predict the undocumented population from the LAC region is that, even conditional on the observable area characteristics in our model, the propensity to be undocumented may differ for those born in Mexico versus elsewhere in the LAC region. We address this concern by constructing an alternative estimate of the undocumented population from the LAC region that uses the relative deportation rate of people from Mexico versus people from the rest of the LAC region. A second issue we confront is that not all undocumented migrants will obtain a Matrícula card and some fraction of the new cards that we observe in the data represent renewals among undocumented individuals who are already in the country. Our version of the Matrícula data do not allow us to estimate takeup and renewal rates directly, so we rely on estimates from Allen et al. (2019). Finally, we present estimates that adjust for changes in legal status.

Although additional work is certainly needed to improve the population adjustments

that we have made, our final estimate is that in the 2008 to 2012 period there were 8.3-8.7 million undocumented Mexican-born individuals residing in the U.S. each year, which is roughly 35 percent higher than the benchmark estimate by the Pew Center (Passel and Cohn, 2019). Consistent with Passel and Cohn (2019), we estimate about a ten percent decline in the undocumented population from Mexico between 2008-2012 and 2013-2017. In the 2008 to 2012 period, we estimate about 2.6 to 3.2 million undocumented LAC-born individuals (from outside of Mexico) residing in the U.S. each year; we estimate a 3.2 to 4.0 million undocumented LAC-born population for the 2013 to 2017 period. Both of these estimate are more similar to estimates from the Pew Center.

Our final objective is to leverage estimates of the undocumented population to project future population changes as a function of a range of economic, demographic, and policy parameters. We build on similar work by Hanson et al. (2017), who model and project the flow of low-skilled immigrants. A key innovation in our work is to use our new estimates of the undocumented population to project migration flows separately for legal and undocumented immigrants. We first project migration flows based only on the population and GDP of sending countries relative to the United States. This parsimonious model indicates a modest rise in undocumented migration from Mexico and the LAC region over the next 20 years, followed by a fall to its current level around 2060.

Finally, we have created an additional projection tool that allows a user to take our baseline projections described above and extend them to account for changes in policy variables, including Customs and Border Protection staffing levels, fence construction, Secure Communities enforcement, and a more general measure of "immigration policy tightness." We use existing estimates of their effects on migration from the recent literature. While a clearly very speculative exercise, this tool allows the user to combine the results from our work that quantifies undocumented migration with others' work on how key policies may affect it.

2 Measuring Undocumented Immigration

Because of the lack of direct data on the undocumented population, researchers have to rely on a variety of imperfect methods to estimate the undocumented population and factors that affect it. Broadly speaking, the most credible estimates of the undocumented population have relied on a residual approach in which the undocumented population is calculated based on the difference between the total foreign-born population (as measured in survey data) and the legal resident population of foreign-born individuals (as measured in administrative data). Alternatively, some researchers have employed an imputation approach to identify likely undocumented individuals based on survey data alone. This imputation approach assumes that all foreign-born people who meet certain criteria (such as receiving Medicaid or being a veteran) have legal status and then classifies all others as being undocumented. One key challenge is that alternative approaches generally cannot be benchmarked using external data on the undocumented population. In this section, we will summariaze the range of existing approaches to estimating the undocumented population. We will use this review of the literature to motivate our subsequent efforts to empirically investigate the extent to which administrative Mexican Matrícula (registration) data can be used to augment prior approaches to estimating the undocumented population.

Warren and Passel (1987) represents the earliest rigorous attempt to quantify the undocumented population using a "residual" approach that leverages a combination of survey and administrative data. Warren and Passel estimated the undocumented population in 1980 by constructing a measure of the total foreign-born population from the 1980 U.S. Census and then subtracting away the number of naturalized U.S. citizens as well as the estimated number of legally resident aliens in the U.S. (measured using data from Immigration and Naturalization Services).

In the decades since this seminal work, researchers have refined the "residual" approach

by incorporating additional administrative data sources that improve the accuracy of national estimates. Warren and Warren (2014), for example, use administrative data from the Department of Homeland Security (DHS) and the Department of Health and Human Services (HHS) to measure the legally resident foreign-born population based on the number of legal permanent residents, the non-immigrant population of foreign-born individuals, and counts of refugee arrivals. They then use data from the 2000 U.S. Census and 2001-2009 ACS surveys to calculate the number of foreign-born arrivals in the U.S. each year. These data sources, in combination with estimated Census-based emigration rates, DHS-based counts of unauthorized removals and status adjustments, and mortality rate measures, are used to construct estimates of the total undocumented population.[2]

While the "residual" approach is employed to construct aggregate undocumented population counts, its reliance on data sources that provide only total counts of legally resident foreign-born individuals means that it cannot be used in isolation to identify the likely legal status of surveyed individuals in common datasets, such as the Census, Current Population Survey, or the American Community Survey. In Passel and Cohn (2019), the authors employ the standard "residual" approach to construct undocumented population counts for age-gender groups in six individual states (California, Florida, Illinois, New Jersey, New York and Texas) and for the rest of the country combined. The authors then use a combination of survey-based individual attributes to impute legal status to respondents in a number of household surveys, while ensuring that the resultant number of individuals identified as likely undocumented equals the total estimated undocumented population within a given geography and when disaggregated by region of origin and age-gender group.

Specifically, to impute legal status in the ACS, the authors classify as legally resident those ACS respondents who satisfy any of the following criteria: (i) they work for the

[2] Warren and Warren (2014) also produce separate estimates of departures and arrivals of undocumented immigrants based on these data sources.

government or in certain occupations that require lawful status or government licensing (e.g., police officers and other law enforcement occupations), (ii) they are veterans or active-duty members of the armed forces, (iii) they are military Reserves or in the National Guard, (iv) they participate in government programs not open to unauthorized immigrants (Medicare, Medicaid, SNAP, etc.), (v) they arrived to the U.S. before 1980, (vi) they are children of citizens or lawful temporary migrants, (vii) they are immediate relatives of U.S. citizens, or (viii) they were born in Cuba.

A key challenge for researchers relying on survey data sources is that estimates must be adjusted to account for differential under-reporting (and mis-reporting) as a function of legal status and country of origin. For instance, while Passel and Cohn (2019) argue that most respondents who report being naturalized citizens are likely legal residents, the authors account for over-reporting of naturalized citizenship by allowing for individuals who self-report as naturalized but arrived in the past five years and are not married to U.S. citizens to be classified as likely undocumented.[3] To account for differential under-reporting, the authors draw on several recent papers on the topic (see, for instance, Van Hook et al., 2014) and ultimately inflate estimated undocumented population counts by eight to 13 percent for the 2000-2009 period and by five to seven percent for the 2010-2016 period.[4] The approach taken by U.S. government agencies to estimating the undocumented population residing in the U.S. has closely paralleled this methodology (Baker, 2017).

One key obstacle to extending existing methodologies, such as the Passel and Cohn (2019) approach described above, is that researchers do not typically provide sufficiently precise details to permit replication and extensions. Borjas and Cassidy (2019) represents

[3] Individuals who report naturalized citizenship but were born in Mexico or Central America may also be assigned likely undocumented status given higher rates of misreporting of naturalized citizenship for individuals from these countries of origin. The authors further refine their imputation approach based on refugee and asylee admissions, temporary visa issuances, etc.

[4] The overall immigrant population count is adjusted by two to five percent over the same period.

a notable exception. In Borjas and Cassidy (2019), the authors build on the Passel and Cohn (2019) imputation approach while providing sufficient details to facilitate replication.[5] [6] Specifically, after imputing undocumented status to individuals in the ACS that match the criteria from Passel and Cohn (2019) (as described above), the authors further refine the imputation method to assign likely legal status to highly-educated, foreign-born individuals who are likely to hold H-1B visas. These individuals are identified as those who work in common H-1B visa occupations, have resided in the U.S. for six years or fewer, and are college graduates.[7] [8] Given the close parallels between the Passel and Cohn (2019) and Borjas and Cassidy (2019) approaches, paired with the replicability of the Borjas and Cassidy (2019) methodology, we rely on the latter paper when constructing ACS-based undocumented population estimates to compare to the estimates we derive based on our alternative (Matrícula-based) methodology.

3 Data

In this section we describe the survey and administrative data sources that we rely on to construct estimates of the total foreign-born and undocumented populations and to produce future population projections. We begin with the administrative Mexican Matrícula

[5] The Borjas and Cassidy (2019) imputation method was developed by "reverse-engineering" the algorithm employed in Passel and Cohn (2019) based on files shared by the latter authors.

[6] In earlier work, Borjas (2017) first applied the Passel and Cohn (2019) approach to impute legal status in the Current Population Survey (CPS).

[7] Notably, Borjas and Cassidy (2019) do not employ any reweighting to account for survey undercounts.

[8] While most of the recent literature employs approaches similar to those detailed above and arrives at qualitatively similar population estimates, Fazel-Zarandi et al. (2018) represents a noteworthy exception. In that research, the authors start from a 1990 estimate of the undocumented population based on the "residual" approach but then use data on population inflows and outflows (as opposed to survey data) to construct an estimate of the growth in the undocumented population between 1990 and 2016. The authors ultimately conclude that "residual" approach-based estimates have significantly undercounted the undocumented population. A published response to Fazel-Zarandi et al. (2018), however, argues that this inflow/outflow-based approach has overestimated growth in the undocumented population by underestimating voluntary emigration rates during the 1990s (Capps et al., 2018).

data. We then discuss the American Community Survey, which is the key data used in much of the prior literature that employs "residual" and imputation approaches to estimate the U.S. undocumented population, and the Encuestas Sobre Migración en Las Fronteras de México (Survey of Migration Across Mexico's Borders), which includes surveys of deported individuals originally from Mexico, El Salvador, Honduras, and Guatemala.

3.1 Mexican Matrícula Data

The key innovation in our work is to use administrative data on Mexican Matrícula identification cards to count undocumented individuals in the United States. As described above, these cards are issued by Mexican consulate offices in the United States to Mexican citizens. The card serves as a valid form of photo identification for many purposes in the United States when another form of identification, such as an immigration document, is unavailable.

The number of Matrícula cards may over or understate the population of undocumented Mexican individuals. Not all migrants may apply for a card, especially if they anticipate a short stay in the United States. Some applicants may not be approved for a card if they fail to meet the consular identification requirements. Some undocumented individuals who have a Matrícula card may obtain U.S. identification, such as a driver's license, and then not renew their Matrícula card, even though they remain undocumented. Finally, of course some individuals who have a card eventually gain legal status in the United States. As we discuss in detail in Section 5, we count 4.7 million Matrícula cards issued between 2013 and 2017. By comparison, the Pew Research Center estimates that there were about 5.5 million undocumented Mexican migrants in the United States during that time. Later we return to a discussion of how we reconcile these estimates.

Our analysis uses a confidential version of the Mexican government's Matrícula Consular de Alta Seguridad (MCAS) database, a new set of microdata covering the universe

of identity cards produced under Mexico's MCAS program from 2002 to 2019. These data were provided by Mexico's Ministry of Foreign Affairs' Department of Consular Affairs. All personally identifying information, including individuals' exact addresses in Mexico and the United States, was removed. The data includes individuals' age, gender, educational attainment, occupation, state and Municipio of birth in Mexico (similar to a U.S. county), and state and county of residence in the United States.[9]

Our methodology below rests on the assumption that the number of Matrícula identification cards issued in a given geographical area over a particular time period can be used to produce an accurate estimate of the true resident undocumented population. Although the endogeneity of the application decision suggests that this assumption is imperfect, we argue that this data source significantly improves upon previously-available measures of the resident Mexican-born undocumented population and we rely on findings from related analyses to convert identification card application counts into measures of the stock of the Mexican-born undocumented population and the undocumented population born elsewhere. Specifically, we rely on the detailed analyses presented in Allen et al. (2019) on both the Matrícula card takeup and renewal rates to generate a plausible set of re-scaling factors that map Matrícula card counts to undocumented population estimates.[10]

[9] A key challenge is that the raw data on place of birth and current residence is provided by the applicant and not standardized. For example, current residence can be listed as Los Angeles County, Los Angeles, L.A., or LA; or the data clearly indicate the city of residence rather than the county. We hand correct these fields and assign FIPS state and county identifiers to all observations. A small number of observations with missing information on state or county are dropped. An analogous process is used to clean the place of birth field, though we do not use that information in this project. All replication and cleaning files are available upon request.

[10] Allen et al. (2019) are able to directly distinguish between new registrants and card renewals. Given the anonymized nature of the dataset to which we currently have access, we are unable to identify whether newly-issued cards represent renewals or new registrants at present. Acquiring data that allows for the identification of new cards issued versus renewals represents an important next step in this research project. One additional limitation of the Mexican Matrícula data is the lack of information available on the date of arrival to the U.S. A survey conducted by Pew indicates that Matrícula applicants have spent less time in the U.S., on average, than Mexican-born ACS respondents.

3.2 American Community Survey Data

The American Community Survey (ACS) is one of the key data sources used to impute legal status to the undocumented population. In our work, we use the ACS 1-year files from 2008 through 2017, which are 1-in-100 random samples of the population. The survey includes questions on household member demographic characteristics along with questions on various other topics including employment and earnings. The 1-year ACS files identify the Public Use Microdata Area (PUMA) for all respondents and identify the county of residence for 473 counties that have populations greater than 65,000 residents. 81 percent of Mexican-born individuals in the ACS live in these counties. In alternative models, we present estimates for the subset of identified counties, for all counties after probabilistically mapping the PUMA in which a respondent resides to a county, and at the commuting zone level (based on a PUMA to commuting zone mapping). After verifying that findings are consistent across these alternative geographies, we focus on commuting zone-level estimates in subsequent models. Since the ACS is designed to be representative at both national and sub-national levels, it can be used to estimate the total foreign-born population at each level of geographical aggregation.

3.3 Encuestas Sobre Migración en Las Fronteras de México (Survey of Migration Across Mexico's Borders)

We use the Encuestas Sobre Migración en Las Fronteras de México (EMIF) 1-year deportation files from 2008 through 2017. These files survey individuals born in Mexico, El Salvador, Honduras and Guatemala who have been deported by the U.S. Federal Government. Mexican respondents are surveyed across approximately 15 sites that include cities in the U.S.-Mexico border region as well as airports in the interior of Mexico. Surveys with respondents from El Salvador, Honduras and Guatemala are conducted at international

airports located in each country. The surveys are designed to be representative of the deported population from each home country and we use data on respondent education, gender, age, and time in U.S. prior to deportation to investigate whether the relevant predictors of undocumented status are common across origin countries. Across analyses, we restrict the sample to individuals who were detained by immigration authorities at least one month after arriving in the U.S. to avoid including individuals deported during the migration process (who may differ from individuals deported from the U.S. interior in various ways and would not be captured in U.S. survey data sources or in Matrícula files).

3.4 Supplementary Data Sources

We leverage a number of additional data sources for included analyses. To construct undocumented population estimates, we incorporate Department of Homeland Security data on persons obtaining lawful permanent resident status by broad class of admission and country of birth from 2008 to 2017. We also use Immigration and Customs Enforcement (ICE) data on individuals deported from the U.S. interior by month and country of origin for the 2008-2015 period to evaluate differences in undocumented population shares by origin country. To construct future immigrant population projections, we draw on population projections at the year by origin country by age level that are issued by the United Nations. We also make use of country-specific future gross domestic product predictions issued by the International Monetary Fund (IMF) and the Organisation for Economic Co-operation and Development (OECD). Lastly, we use a range of datasets, including data on annual ICE budgets and Customs and Border Protection staffing levels provided by the Department of Homeland Security, to investigate predictors of historical immigrant population estimates.

4 New Estimates of Local Undocumented Population Shares

In this section we first use Mexican Matrícula data to construct local measures of Mexican-born undocumented population shares. We demonstrate that Matrícula-based measures of local undocumented population shares are highly correlated with shares constructed using only the ACS and the Borjas and Cassidy (2019) imputation approach described in Section 2. We next undertake a two-step prediction exercise to provide proof of concept that the Matrícula data can be used to inform estimates of the undocumented population born outside of Mexico. To do so, we first present the results of descriptive analyses that identify key predictors of local Matrícula-based Mexican-born undocumented population shares from an extended set of ACS socio-demographic averages constructed at alternative levels of geographic aggregation (i.e., by county and commuting zone). Relying on data from 2013-2017, we then show that our resultant predictions of undocumented population shares for those born elsewhere in the LAC region (i.e., outside of Mexico) are highly correlated with corresponding undocumented population shares estimated using only the ACS and the Borjas and Cassidy (2019) imputation approach.

4.1 Estimates of Local Mexican-Born Undocumented Population Shares in 2008-2012

The lack of available administrative data on the true Mexican-born undocumented population means that it is not feasible to formally validate the Mexican Matrícula card data that we use to build on existing estimates. However, prior research has confirmed the quality and coverage of the Matrícula data (Caballero et al., 2018), and we next confirm that undocumented population shares measured using Mexican Matrícula card data are highly correlated with shares measured using only data from the American Community

Survey. To do so, we rely on the algorithm presented in Borjas and Cassidy (2019) (and described above) to identify likely undocumented Mexican-born respondents in the ACS and to construct estimates of the Mexican-born undocumented population share at alternative levels of geographical aggregation.[11] Using Matrícula data files, we calculate the total number of Matrícula cards issued in a given geography between 2008 and 2012. This five-year window ensures that we will not observe cards for the same individuals multiple times since cards remain valid for five years. We divide the number of cards issued by a population estimate from the ACS of the Mexican-born population over the same time period and residing in the same geographical area.[12] This ratio is a measure of the fraction of the Mexican-born population that is undocumented. As we discuss in Section 5, we will ultimately inflate Matrícula counts by re-scaling factors that correspond to alternative assumptions regarding card take-up and renewal rates.[13]

In Figure 1, we present a scatter plot characterizing the commuting zone-level relationship between Mexican-born undocumented population shares alternatively constructed using ACS and Matrícula data for the 2008-2012 period, where the ACS measure reflects the annual average undocumented share across these years. We weight each observation by the average Mexican-born population in the commuting zone over the same years. The scatter plot and associated regression estimates indicate that the Matrícula-based measure of Mexican-born undocumented population shares is highly predictive of the corresponding ACS-based estimates. At the same time, this figure is consistent with the Matrícula data providing additional information on the distribution of the undocumented population; in particular, the marginally higher variance of the Matrícula-based mea-

[11] Our approach makes use of ACS rather than CPS data given the various advantages of the ACS survey, including its larger sample size. The relative advantages of the ACS are discussed in more detail in Passel and Cohn (2019).

[12] Our population estimate is given by the sum of the population weights in an area.

[13] Since we are comparing undocumented population shares across areas during the same set of years, and since the re-scaling factors we estimate are common across geographies, correlations based on cross-sectional analyses are not influenced by our estimates of overall Matrícula card take-up and renewal rates.

sure is consistent with the coarseness of the ACS imputation algorithm, which relies on a relatively limited set of covariates to identify the likely undocumented population.[14] Appendix Figures A1 and A2 show qualitatively similar patterns when undocumented population shares are constructed at the county level, although smaller sample sizes at the county level introduce additional imprecision.[15]

4.2 Validation of a Predictive Model of Undocumented Population Shares

In this section we build a model that combines the Matrícula data and the ACS. These estimates then provide the foundation to use the ACS alone to predict local undocumented population shares for migrants from countries other than Mexico.

To do this, we first define m_c as the Matrícula-based undocumented share of the Mexican-born population in commuting zone (or county) c. As before, this measure is constructed as the total number of Matrícula identification cards issued between 2008 and 2012, divided by the estimated average number of Mexican-born residents in that commuting zone. We regress m_c on the average values of individual characteristics in the American Community Survey, $\bar{x}_c$. In particular, $\bar{x}_c$ represents averages taken over x_{ic} for the following vector of characteristics: gender, indicators for age groups, indicators for educational attainment, log income, occupational category, and year since arrival to

[14] No Matrícula cards are reported issued in approximately 8% of commuting zones for the 2008-2012 period. These commuting zones are notably smaller than those covered in the Matrícula files with an average population of 22,500 (as compared to 449,500 in covered commuting zones) and with only 75 Mexican-born residents, on average (as compared to 17,300 in covered commuting zones). For the 2.4% of commuting zones for which the number of Matrícula cards issued exceeds the average number of Mexican-born residents during the 2008-2012 period, we top code the undocumented population share at 1 (these commuting zones are also significantly less populated than those with estimated undocumented population shares between 0 and 1). In any case, since we weight observations by the average Mexican-born population during the relevant years in the given geography, estimates are not sensitive to this top coding or to the exclusion of geographies with extreme values.

[15] We present results separately for the subset of counties identified in the ACS, and for the universe of counties where county-level values are imputed based on a PUMA-county crosswalk.

the US for Mexican-born person i who lives in area c and was surveyed in the ACS between 2008 and 2012. We then run OLS regression models of the form $m_c = \alpha + \beta\bar{x}_c + \epsilon_c$. Observations are weighted by the average number of Mexican-born residents in area c. We use the estimated coefficients to form $\hat{\alpha} + \hat{\beta}\bar{x}_c$, the best linear prediction of the undocumented population share in commuting zone c for individuals from a given country/region of origin.

In Figure 2, we first present $\hat{\beta}$ estimates from models estimated at the commuting zone level, separately for subsets of included individual characteristics. In panel (a) we present results from a regression of the Mexican-born undocumented population share on the average fraction of the population that is aged 16 to 40, aged 41 to 65, and 65 or older. The omitted category is the fraction under age 16. The undocumented population share is largest in those geographies with the highest shares of Mexican-born residents aged 16 to 40 and smallest in areas with a higher fraction over 65. In particular, a one percentage point increase in the fraction of the population 16 to 40, relative to under 16, is associated with a 0.8 percent increase in the undocumented share.

Panel (b) shows a non-monotonic relationship between the undocumented population share and the educational attainment of the Mexican-born population. Commuting zones with higher shares of college-goers and college graduates have lower undocumented population shares than those with higher shares of Mexican-born respondents who have not completed high school (the omitted group), while the undocumented population share is highest where the share of Mexican-born respondents with high school degrees but no history of college enrollment is largest. Finally, panel (c) investigates the occupational composition of Mexican-born respondents. Here, the omitted category corresponds to the fraction of white-collar workers. We find that the Mexican-born undocumented population share is highest in commuting zones with high shares of workers in food and cleaning services, in construction, transportation, and production, and in agriculture (though the

latter coefficient is not statistically different from zero).[16]

Table 1 presents $\hat{\beta}$ estimates from regression models that include all potential predictors of local undocumented population shares in the same model. We present estimates separately for three alternative units of geography: commuting zones, counties identified in the ACS, and counties imputed using ACS PUMAs. Although patterns are broadly similar to those previously described, the estimates do change to some degree once we condition jointly on labor market and demographic characteristics. In addition to a general loss of precision associated with the inclusion of additional regressors, the attenuated (and in some cases, opposite-signed) coefficients corresponding to the male share of Mexican-born respondents are explained by the strong association between respondent gender and the other included covariates (which are themselves highly predictive of undocumented population share).[17]

We next assess how accurately we can predict Mexican-born undocumented population shares for the 2013 to 2017 period using $\hat{\beta}$ values derived from models estimated over the 2008 to 2012 period in conjunction with ACS-based averages of local Mexican-born population characteristics from 2013 to 2017. Figure 3 presents a scatter plot that summarizes our findings. As before, we weight observations by the average size of the Mexican-born population. We identify a strong positive relationship between Matrícula-based undocumented population shares for the 2013-2017 period and the undocumented population shares predicted using ACS covariates and $\hat{\beta}$ values. For comparison, the root

[16] See Appendix Figure A3 for estimates based on the subset of counties identified in the ACS and Appendix Figure A4 for imputed county-level results based on PUMA of residence.

[17] For the set of analyses that relate area-level undocumented population shares to average population characteristics, it is worth noting that the $\hat{\beta}$ parameters we estimate need not mirror the coefficients that would be derived from individual-level models. To provide an example, it may be the case that the likelihood that an individual is undocumented is decreasing in their educational attainment but that, conditional on education, undocumented individuals are more likely to reside in areas with more educated Mexican-born residents (perhaps because these individuals are likely to provide them with employment opportunities). Then, we may find that local undocumented population shares are rising in the average educational attainment of Mexican-born residents in spite of the fact that educational attainment is negatively correlated with an individual's own likelihood of being undocumented.

mean squared error (RMSE) of 0.109 is marginally lower than the corresponding RMSE of 0.120 from a regression of 2013-2017 Matrícula-based undocumented population shares on 2013-2017 undocumented population shares constructed using the ACS-based imputation procedure (Borjas and Cassidy, 2019).[18] In Appendix Figure A5, we present a complementary scatter plot based on a simple lasso-based covariate selection model to assess whether we can improve upon OLS-based predictions. The predictive power of our model is essentially unchanged when we employ the lasso-based approach, so we elect to focus primarily on the more transparent and parsimonious OLS-based prediction models in subsequent analyses.

By construction, there is no scope to externally benchmark the performance of our Matrícula-based models viz-a-viz standard approaches in the literature, such as imputation methods relying on ACS respondent characteristics. Indeed, the Matrícula data are useful precisely because there are no other more accurate measures of the undocumented population at present. That said, we find the correlational results described above to be encouraging. To the extent that the Matrícula data more accurately capture variation in Mexican-born undocumented population shares than prior methods, our estimates indicate that even the relatively coarse prediction model we have developed can improve upon existing imputation-based algorithms.[19]

[18] Correspondingly, the R-squared associated with the former regression (0.531) is marginally higher than the R-squared associated with the latter one (0.433).

[19] A distinct question is whether the Matrícula-based prediction model improves upon prior estimates constructed using the residual-based approach. The close alignment between results from that approach and results from the Borjas and Cassidy (2019) imputation method suggest that we should arrive at relatively similar conclusions when comparing our estimates to those based on the residual method. Ultimately, however, it is not feasible to reconstruct residual-based estimates given the precise algorithms that underlie these approaches are not made publicly available.

4.3 Using ACS Covariates to Estimate Local Undocumented Population Shares for Those Born Outside of Mexico

Having validated the out-of-sample predictive ability of the Matrícula-based prediction model, we use the estimates from the regression model above in conjunction with ACS data on foreign-born individuals to estimate undocumented population shares among non-Mexicans. We split the remaining foreign-born population into one subgroup that includes all individuals born in the LAC region (which includes Central America, South America, and the Caribbean) but outside of Mexico and a second subgroup that includes all individuals born outside of the LAC region.

We begin by briefly highlighting key differences in the foreign-born populations by country/region of origin. Table 2 summarizes commuting zone-level population averages for the 2013-2017 period, separately for those born in Mexico, elsewhere in the LAC region, and outside of the LAC region. During this period, the average commuting zone in our sample has 15,974 Mexican-born residents, 12,997 residents born elsewhere in the LAC region, and 31,205 residents born outside of the LAC region. Based on the Borjas and Cassidy (2019) methodology, we estimate that the average commuting zone-level undocumented population share is 46.4 percent for the Mexico-born population, 34.4 percent for the population born elsewhere in the LAC region, and 17.2 percent for the population born outside of the LAC region. The correlation across commuting zones between the Mexican-born undocumented population share and the undocumented share of those born elsewhere in the LAC region (0.44) is notably higher than the correlation between the Mexico-born undocumented population share and the undocumented share of those born outside of the LAC region (0.21). There are a number of indications based on the summary statistics in Table 2 that the foreign-born population from outside of the LAC region differs in important ways from the foreign-born population from the LAC region (Mexico or elsewhere). Most notably, foreign-born populations from outside of the

LAC region are less skewed towards males, have been in the U.S. for longer, and are more highly-educated. Those born outside of the LAC region are also the least likely to be aged 16-40 (the age range most positively associated with undocumented status based on the area-level Matrícula data).

To evaluate the performance of the Matrícula-based prediction model across the same three foreign-born subpopulations, Figures 4 through 6 show scatter plots of the predicted share of the relevant population that is undocumented based on the Borjas and Cassidy (2019) imputation-based approach (on the vertical axis) and based on our Matrícula-based prediction model (on the horizontal axis). Specifically, the Matrícula-based prediction is formed as $\hat{\alpha} + \hat{\beta}\bar{x}_c$, where $\bar{x}_c$ is the average characteristics of the relevant foreign-born population in commuting zone c. $\hat{\alpha}$ and $\hat{\beta}$ are the intercept and slope estimates from the model described in Section 4.2 and presented in Table A1.

While Matrícula prediction model-based estimates are strongly predictive of ACS imputation-based estimates of undocumented population shares for those born in Mexico and elsewhere in the LAC region, the prediction model seems to perform poorly for immigrants from the rest of the world. As shown in Figure 6, the alternative estimates of area-level undocumented population shares for those born outside of the LAC region are weakly negatively correlated. In Appendix Figures A6 through A8, we present corresponding scatter plots that rely on a lasso-based covariate selection approach to predict area-level undocumented population shares based on averages of ACS respondent characteristics. When predicting undocumented population shares for those born in Mexico or elsewhere in the LAC region, slopes are similar and RMSE values fall modestly (by less than 10%). The predictive power of the model for those born outside of the LAC region, however, remains poor (the associated regression slope remains statistically insignificant at conventional levels and changes sign).

The Matrícula-based prediction model appears to hold the most promise for measuring

changes in the undocumented population for those born in the LAC region. Our model relies on the assumption that the variation in area-level population characteristics that predicts undocumented population shares for those born in Mexico is also predictive of undocumented population shares for those born elsewhere in the LAC region, and the similar population averages for those born in Mexico and elsewhere in the LAC region (shown in Table 2) are consistent with this assumption. It is possible that the Borjas and Cassidy (2019) ACS imputation method itself performs poorly for immigrants born outside of the LAC region. In Appendix Figure A9, we compare Matrícula-based estimates to undocumented population shares constructed based on Migration Policy Institute (MPI) estimates of the undocumented population available at the state by country/region of origin level. These scatter plots similarly indicate that Matrícula card data accurately predict MPI-based counts and that our prediction model performs well within the LAC region but poorly outside of this region.

To provide additional support for the applicability of our Matrícula-based prediction model for the population born outside of Mexico, we next leverage data on a selected sample of undocumented immigrants from the EMIF. As described in Section 3.3, the EMIF includes survey responses from a representative sample of deportees from Mexico, El Salvador, Honduras and Guatemala, and we can thus use this data to assess whether the predictors of undocumented status vary by country of origin. Although survey coverage requires us to focus exclusively on those born in El Salvador, Honduras and Guatemala (as opposed to other countries in the LAC region), these three countries accounted for two-thirds of the undocumented population born in the LAC region and one-third of the total undocumented population born outside of Mexico as of 2016 (Passel and Cohn, 2019).

To make use of the EMIF, we append the 2008-2017 annual ACS files to the annual EMIF deportee files and create an indicator variable for whether a given observation

corresponds to an EMIF deportee.[20] We then estimate a regression of this indicator for a deportee on a set of socio-demographic characteristics that are available in both datasets (gender, age, years of schooling, and years spent living in the U.S.). Under the assumption that selection into deportation is as good as random within the undocumented population or does not vary as a function of country of origin, we can interpret resulting estimates as characterizing the degree to which the relative characteristics of the undocumented population differ by country of origin.

Table 3 presents the results of this regression analysis; each column presents regression results specific to the indicated country of origin. Although coefficient magnitudes vary across columns (in part due to differences in deportees per capita), the pattern of coefficients across variables is remarkably consistent. Across sample countries, deportees are younger, less educated, more likely to be male and have spent less time in the U.S. than the average survey respondent. As a summary measure, for each observation, we construct four predicted values based on the four sets of country of origin-specific coefficients (paired with respondent observable characteristics). We then correlate these predicted values. The results from this exercise indicate that the predictors of deportee status are indeed quite similar across countries of origin: pairwise correlations range from 0.95 to 0.98. These findings thus provide additional support for the applicability of our Matrícula-based prediction model to other origin countries in the LAC region.[21]

A logical next question is whether the EMIF data can be used to augment our prediction model for those born outside of Mexico. Unfortunately, this approach holds limited promise for two key reasons. First, while it is plausible that selection into deportation

[20] We restrict the EMIF sample to individuals who were deported from the U.S. interior after at least one month in the country to avoid including individuals who were apprehended while crossing the border but never established residency in the U.S. since these individuals would not be surveyed in the ACS.

[21] While we investigated the possibility of using alternative data sources, such as Latin America Migration Project data, to study the characteristics of the undocumented population from elsewhere in Latin America, limited sample sizes and temporal coverage made doing so infeasible.

does not vary by country of origin, it is unlikely that the deportation rate conditional on undocumented status is constant across geography given the myriad state and local policies that have been enacted to either facilitate or constrain the transfer of undocumented individuals from police to immigration authority custody. Indeed, if we correlate the state-level Matrícula-based undocumented population share with the number of state-level deportees per resident from Mexico, we identify a weak negative correlation consistent with the endogeneity of deportation policy. Given this, any adjustment to our prediction tool would have to rely on the estimated relationships between individual-level deportee status and socio-demographic characteristics. This gives rise to a second challenge: as noted previously, aggregate (area-level) estimates of the relationships between undocumented population shares and average population characteristics need not (and do not) align with estimates based on individual microdata. Thus, although we do not make further use of the EMIF data to improve our Matrícula-based predicted model, our analysis of EMIF data indicates that extrapolating the Matrícula-based model from the Mexican-born population to the wider Latin American-born population is informative.

5 National Estimates of the Undocumented Population

In this section we construct national-level estimates of the undocumented population. We leverage the Mexican Matrícula data and our Matrícula-based prediction model to construct estimates of the undocumented population born in Mexico and in the LAC region. Given concerns about the performance of our prediction model outside of LAC origin countries, we do not develop new estimates of the size of the undocumented population born outside of the LAC region. The national-level estimates we produce for those from Mexico and the rest of the LAC region will serve as an input in the models we employ

to make projections regarding future undocumented population trends, as discussed in detail in Section 6.

We begin by comparing raw Matrícula card counts and our unadjusted estimates of the undocumented population born in LAC countries outside of Mexico to estimates of the undocumented population in Passel and Cohn (2019), who rely on the residual method described in Section 2. Table 4 summarizes our findings. Column 1 is an estimate of the undocumented Mexican-born population. Column 2 is an estimate of the undocumented population that originated in other LAC countries. Rows 1 and 2 present Pew estimates, where we construct five year averages over the relevant time periods based on Passel and Cohn (2019) and preceding reports produced by the Pew Research Center. Rows 3 and 4 present raw Matrícula card counts in Column 1 and our unadjusted estimates of the undocumented population born in LAC countries outside of Mexico in Column 2. As expected, raw Matrícula card counts and unadjusted Matrícula-based predictions fall below the Pew estimates. Specifically, the raw count of Matrícula cards is about 30 percent lower than Pew estimates for the 2008-2012 period and 15 percent lower for the 2013-2017 period. The unadjusted estimates of the undocumented population for those born elsewhere in the LAC region are roughy one-third lower than Pew estimates.

In what follows we address a number of key challenges to using the Matrícula data to estimate the undocumented population. The first is that, conditional on area-level population characteristics measured in the ACS, the share of individuals who are undocumented may differ between those from Mexico and from the rest of the LAC region. An alternative method is to assume that the fraction of undocumented individuals from LAC relative to Mexico can be inferred from their relative rates of deportation. To implement this estimate, we make use of Immigration and Customs Enforcement (ICE) data on individuals deported from the U.S. interior. These data were harmonized and published by the Transactional Records Access Clearinghouse (TRAC, 2021), which received raw deporta-

tion records from ICE after successful court litigation. These data include monthly interior ICE immigrant removals for the 2008-2015 period by country of origin. We use these data in conjunction with ACS files to first estimate the per capita annual removal rate for those born in Mexico versus elsewhere in the LAC region. We estimate an annual removal rate that is 95 percent higher for Mexican-born individuals for the 2008-2012 period and 67 percent higher for the 2013-2015 period, relative to those from the rest of LAC. In rows 5 and 6 of Table 4, we present alternative estimates of the LAC-born undocumented population constructed by multiplying the average annual foreign-born population from this region (measured in the ACS) by the undocumented share of Mexican-born immigrants (constructed based on raw 5-year Matrícula card counts) and then by the inverse of the period-specific interior removal ratio. This adjustment reduces the estimated size of the undocumented population born in LAC countries outside of Mexico for the 2008-2012 period (when the Mexico: LAC removal rate ratio is higher) and increases the estimate for the 2013-2017 period.

We next adjust estimates to account for the selective take-up and renewal of Matrícula cards. Here, we rely on renewal and take-up estimates from Allen et al. (2019). Focusing first on the 2008-2012 period, Passel and Cohn (2019) estimate that approximately 30 percent of undocumented migrants were recent migrants (having arrived in the prior five years) during the pre-2012 period. Allen et al. (2019) estimate annual takeup rates of 26 percent and 12.5 percent for recent and established migrants, respectively, during this period. Aggregating over five years, these estimates indicate that 56 percent of new migrants and 49 percent of established migrants would be expected to acquire Matrícula cards during the 2008-2012 period under the assumption that arrival dates of new migrants are uniformly distributed.[22] Further assuming 30 percent of undocumented immigrants

[22] For established migrants, we calculate the five year takeup rate as $1-(1-.125)^5 = .49$ based on the 12.5 percent annual takeup rate and the five year period over which we construct our estimate. For recent migrants, under the assumption that year of arrival was uniformly distributed over the same five year period, we calculate a five year takeup rate of $.2 * (1-(1-.26)) + .2 * (1-(1-.26)^2) +$

during this period are recently arrived, we calculate a re-scaling factor of 1.96 (equal to 0.7*(1/0.49)+0.3*(1/0.56)).

Turning to the 2013-2017 period, Allen et al. (2019) estimate an 11.6 percent annual renewal rate beginning in 2012 for established migrants (renewal estimates are reliably available beginning only in 2011), a 56 percent annual takeup rate among new migrants during the post-2011 period, and a 4 percent annual takeup rate among established migrants. After 2011, the authors report 58 percent of cards issued are renewals. Again assuming a uniform distribution of new migrant arrivals, the authors' annual takeup rate implies that 85 percent of new migrants would be expected to acquire Matrícula cards during the 2013-2017 period and 46 percent of established migrants would be expected to renew Matrícula cards during this period.[23] Since the card takeup rate is lower for established than new migrants during this period, and since the share of new cards issued to established migrants is not observed, we conservatively apply the same 1.18 (1/0.85) re-scaling factor to all new issuances. Combining estimated takeup and renewal rates with the share of card issuances that are renewals, we calculate a total re-scaling factor of 1.75 (equal to 0.58*(1/0.46)+0.42*(1/0.85)) for the 2013-2017 period.

Rows 7 through 10 of Table 4 re-scale Row 3 through 6 estimates by the period-specific re-scaling factors constructed above. Comparing resultant counts of the Mexican-born undocumented population to those presented in Pew publications, we identify a baseline Mexican-born undocumented population that is roughly 40 percent larger, and we identify a decline in the Mexican-born undocumented population across periods that is somewhat smaller than that found by Pew. In Column 2, we present parallel estimates of the total undocumented population born in the LAC region outside Mexico based on the same re-scaling factors. Re-scaled estimates for this subpopulation in Rows 7 through 10 exceed

$.2 * (1 - (1 - .26)^3) + .2 * (1 - (1 - .26)^4) + .2 * (1 - (1 - .26)^5)) = .56$.

[23] We construct these estimates as above, now relying on the annual take-up and renewal rates for the post-2011 period.

Pew estimates by similar or lower percentages than the corresponding estimates for those born in Mexico. While Pew identifies a small increase in the LAC-born undocumented population across periods, we identify inconsistent estimated changes across periods in Rows 7-8 versus 9-10.

In the final rows of Table 4, we adjust Matrícula counts to account for changes of status within the undocumented population. Specifically, we calculate the total number of persons obtaining lawful permanent residence status by year, country of birth, and class of admission. Not all classes of admission are available to individuals illegally resident in the United States, and so we include only those classes of admission that could potentially apply to the undocumented population: immediate relatives of U.S. citizens, refugees and asylees, and Other (including U-visa admissions). Not all admissions within these broad classes will correspond to undocumented immigrants already resident in the U.S. In particular, a substantial share of immediate family admissions will correspond to admissions of Mexican-born individuals who are not yet resident in the U.S. Nonetheless, we present corresponding undocumented population estimates in Rows 11 to 14 of Table 4 that assume all admissions within these classes apply to undocumented immigrants in order to bound the contribution of status changes to overall population counts. Population estimates fall mechanically after accounting for changes in legal status during the relevant time periods. The inconsistency across estimation approaches in the sign of population changes over time for the LAC-born population is again consistent with the roughly constant estimates presented in Pew publications.

Our preferred range of undocumented population counts is based on the estimates derived in Rows 9 and 10, and 13 and 14, of Table 4. These ranges account for alternative assumptions regarding the share of newly admitted immigrants who were previously residing illegally in the U.S. We estimate an annual average of between 8.3 and 8.7 million Mexican-born undocumented residents in 2008 to 2012, and 7.5 to 8.2 million in 2013 to

2017. We estimate an additional 2.6 to 3.2 million undocumented residents from other LAC countries in 2008 to 2012, and between 3.2 to 4.0 million in 2013 to 2017. On net, our estimates of the Mexican-born undocumented population exceed Pew estimates. Our estimates of the undocumented population born in the remaining LAC region are more comparable to Pew estimates. It is important to note, however, that these estimates remain highly speculative for a number of reasons. First, the re-scaling we have undertaken is likely to overestimate the size of the Mexico-born undocumented population to the extent that we fail to capture return migration or mortality among those who have previously acquired Matrícula cards. Second, the re-scaling of estimates for those born elsewhere in the LAC region relies on the assumption that interior removal rates represent a useful proxy for relative undocumented population shares, and this assumption is ultimately untestable. Further examination of those factors that can explain estimated gaps in both levels and changes in undocumented population sizes (viz-a-viz Pew estimates) represent a worthwhile avenue for future research.

6 Future Projections

In this portion of the paper we develop a simple model that can be used to produce time series projections of the future legal and undocumented flow of migrants, separately by country of origin and immigrant legal status. Our method builds on that in Hanson et al. (2017), which is focused on low-skilled migration more generally. We leverage our estimates from Section 5 to extend their method to projections of flows of undocumented migrants. A key limitation of recent estimates from the literature that relate migration flows to economic and demographic factors is the lack of attention paid to the differential impacts of these factors on legal versus illegal migration. In supplementary models, we thus probe whether our historical data can help clarify the extent of heterogeneity

with respect to these Push-Pull factors, and we present projections based on alternative assumptions regarding the extent of such heterogeneity. As an addendum to this paper, we have produced a tool that allows users to construct their own projections based on alternative assumptions regarding the anticipated future levels of various factors that influence legal and undocumented immigration as well as the associations between these factors and immigration flows. Appendix B includes a README file describing this tool.

Our method departs from that of Hanson et al. (2017) in a number of ways. They develop a rich projection model that incorporates net migration measured at the level of country of origin, birth cohort, gender, and time period. With sufficiently rich data, we could estimate their model separately for undocumented migrants. In practice, however, our estimates of the recent undocumented population cover only two time periods (2008-2012 and 2013-2017) and ten countries of origin. We thus lack the disaggregated data needed to accurately model the relationships between predictors of interest and migration levels. As such, we use our estimates of the undocumented population to construct country of origin-by-legal status fixed effects that inform future projections. These future projections are based on demographic and economic projections produced by the United Nations, International Monetary Fund, and OECD, in combination with parameter estimates from the model in Hanson et al. (2017). We describe this in detail below. Finally, we note that there exists a much larger literature that examines bilateral migration flows and their determinants. It seems unlikely that estimates from this broader literature would be expected to generalize to our study sample given unobserved variation in the costs and benefits of migration as a function of origin and destination country. Given its focus on predicting future migration flows from the LAC region to the U.S. and its methodological rigor, we thus view Hanson et al. (2017) as the most appropriate benchmark for our projection exercise.

We begin by constructing baseline rates of undocumented migration by country of

origin for the ten Latin American countries with the largest undocumented populations in the United States (based on estimates in Pew reports).[24] Over the 2008 to 2017 period these ten countries account for approximately three-fourths of the total undocumented population. To formalize our approach, denote by m_{clt} the predicted number of migrants from county c, in year $t = 2008-2012, 2013-2017$, with legal status l. For undocumented migrants, we use the same approach to construct estimates as is used to produce Rows 9-10 of Table 4.[25] Since we use legal status change data in Table 4 only to bound estimates, we must select estimate values within the relevant ranges for the purposes of this prediction exercise. Specifically, we make the assumption that the fraction of immigrant adjustments that apply to the undocumented population is equal to the estimated fraction of the population from a given origin country that is undocumented, and we adjust estimates accordingly. For legal migrants we use the estimates produced by Pew, since the goal of our work is to focus on producing alternative undocumented population estimates.[26]

We next convert immigrant population counts, m_{clt}, to net migration rates, M_{clt}, by normalizing by the relevant contemporaneous origin country population. It is ex-ante unclear how to partition the origin country population into prospective legal and undocumented migrants. Given the strong negative association between age and undocumented status, we form the ratios of undocumented counts to the population under 40, and legal migrant counts to the population 40 and older. We then regress M_{clt}, the cell-specific net migration rate from origin country c at time period t and for those with legal status l, on

[24] In addition to Mexico, this set of countries includes Brazil, Colombia, Dominican Republic, Ecuador, El Salvador, Guatemala, Haiti, Honduras, and Peru.

[25] Interior removal rates vary widely across origin countries when we disaggregate the interior removals data beyond the Mexico/LAC dichotomy employed in Table 4. To generate plausible estimates, we top code the origin country-specific undocumented population share at 93%, corresponding to the 85th percentile of the imputed undocumented share distribution based on raw interior removal rate adjustments.

[26] We recover Pew estimates of the legal immigrant population from each origin country by taking the difference between the total immigrant population from that country residing in the U.S. (as measured in the ACS) and the corresponding Pew estimate of the undocumented population from that origin country.

country of origin-by-legal status fixed effects. These estimated fixed effects, denoted $\hat{\theta}_{cl}$, are our baseline rates of undocumented and legal migration, by country of origin.

The next step is to convert these baseline rates to estimated population counts based on projected trends in population and GDP. We use an adapted version of the model in Hanson et al. (2017). Broadly speaking, they estimate that increases in the population of a country relative to the United States increase low-skilled migration among those aged under 40, but not among those aged above 40. In contrast, the net migration rate is increasing in log relative GDP for those aged over 40 but not for those aged under 40. To formalize how we use their estimates, let $\log C_{cta}$ represent the log of the population ratio between country of origin c and the United States, in year t for age group a, relative to its mean value during the period 2008 to 2017. We use two age groups, those under 40 (group y) and those 40 or older (group o). Past population counts and projections through 2060 are from the United Nations. Following Hanson et al. (2017), we use "no-migration" estimates of future populations for all origin countries since alternative estimates rely on precisely the migration flows we are attempting to project.[27] We use estimates of the effect of population ratios on migration taken from Table 5 of Hanson et al. (2017), which are reproduced in our Table A2.[28] Specifically, we then form the weighted average

$$P_{ct} = \log C_{cty}\beta_y \times share_{yt} + \log C_{cto}\beta_o \times (1 - share_{yt}) \tag{1}$$

where $\beta_y = 4.7008$ and $\beta_0 = 0.7716$ are the estimated effects of population ratios on net migration; and $share_{yt}$ is the projected share of the sending country population in year t that is under 40 years old. We proceed analogously in incorporating the predicted effect

[27] For the U.S., we include "medium fertility" estimates that incorporate anticipated migratory flows since changes in migration from any one origin country will have a relatively more limited impact on aggregate migration flows to the U.S.

[28] While the positive association between origin country economic conditions and migration rates may initially appear puzzling, this finding is consistent with other existing research (see, for instance, Chort and de la Rupelle, 2016) and may be explained by credit constraints operating as a barrier to migration.

of the log of the ratio of origin country to U.S. GDP per capita, with heterogenous effects on younger and older individuals.

We then add our estimated benchmark migration from county c and legal status l, $\hat{\theta}_{cl}$, and the country-specific trend in migration, P_{ct}, to produce estimated levels of future migration. The results of this projection exercise are shown graphically in Figure 7 and associated estimates are presented separately by origin country in Table 5.[29] As noted, our projection model is more coarse than that included in Hanson et al. (2017). While we attempt to match estimates from that paper to our data structure, they do not align perfectly given that we ultimately construct outcomes by aggregating across cohorts. The application of prior estimates to our study sample is necessitated by the fact that we do not have sufficiently rich historical data to construct credible within-sample coefficient estimates. As such, these projections should be interpreted with particular caution. Nonetheless, our findings appear qualitatively similar to those from Hanson et al. (2017).

Focusing first on estimates of the Mexican-born population, we estimate a 30 percent larger population than Hanson et al. (2017) in the 2013-2017 period corresponding to their 2015 estimate (consistent with the difference in our undocumented population estimates as compared to those in Pew). Our projection of the Mexican immigrant population for 2040 exceeds the Hanson et al. (2017) projection by about 15 percent. More generally, our 2013-2017 total immigrant population estimates generally align with the estimates from Hanson et al. (2017) as do our 2040 projections totaled over legal and undocumented immigrant estimates.[30] Turning to immigrants' legal status, our projection model predicts consistent growth in the legal immigrant population between the 2008-2017 period and

[29] As in Hanson et al. (2017), we estimate negative net migration rates for a small number of cells. This reflects the limits of the linearity assumption that underpins our projection model.

[30] Two exceptions are our estimates for Ecuador and Peru, which exceed estimates from Hanson et al. (2017) by roughly 100%. In the case of Peru, this is partly explained by the negative estimate of the immigrant population aged under 40 that is presented in Hanson et al. (2017) and results from the linearity imposed in the projection model.

2060. In contrast, we predict that the total undocumented immigrant population from the 10 included LAC countries will peak in roughly a decade and decline gradually thereafter.

One key challenge in interpreting these findings, however, is that our benchmark projection model does not allow for demographic and economic variables to differentially alter patterns of legal versus undocumented migration. To the best our knowledge, there are no existing estimates from the recent literature that would allow us to do so. To probe how log population ratios and log GDP ratios may differentially affect legal versus undocumented immigration, we leverage the historical estimates of legal and undocumented immigrant populations by country of origin. In Appendix Table A3, we present regression results that alternatively rely on our own historical population estimates and those produced by Pew. These estimates are derived from a series of simple OLS models that regress net migration rates on country-by-legal status fixed effects while adding additional covariates of interest and time period-by-legal status fixed effects (in a subset of specifications).

Columns 1 and 6 of Appendix Table A3 highlight the imprecision associated with analyses relying on our five-year Matrícula-based estimates or alternatively on Pew annual estimates. While we identify a significant positive association between the log GDP ratio and the net migration rate in Column 6 (using Pew data), log birth cohort ratio estimates are inconsistent in sign and uniformly insignificant at conventional levels. Nonetheless, we do uncover suggestive evidence of heterogeneous associations in the remaining columns. In particular, increases in the log birth cohort ratio for those aged under 40 have a more positive effect on undocumented than legal migration across models, while increases in the log birth cohort ratio for those aged over 40 have a more positive (or, in some cases, less negative) effect on legal than undocumented migration across models. These patterns should be interpreted with caution. Coefficient magnitudes vary widely and estimates are not consistently statistically distinguishable from zero (or one another). As a result,

we would certainly not feel confident in applying coefficient estimates directly to our projection exercise (doing so results in implausible net migration rates).

With these caveats in mind, we consider how projected changes in net migration rates would differ if we imposed the heterogeneity we identify from the historical data in our projection model. Results are presented in Figure 8 and Table 6. While overall projected immigrant population counts are somewhat higher based on this alternative model, we also identify larger predicted future declines in the total undocumented population from Mexico and the rest of the LAC region (between 2040 and 2060). In the future, we plan to use Matrícula microdata to more rigorously develop projections that allow for regressors of interest to differentially impact legal versus undocumented immigration.[31]

While projections are available for economic and demographic variables of interest, we lack corresponding predictions regarding future immigration policy trajectories. As such, we have developed a tool that allows users to construct projections based on expected changes in immigration policy and enforcement in combination with estimates of the elasticity of immigration with respect to these measures.[32] Before discussing the existing parameter estimates we draw on to populate this projection tool, it is important to emphasize that the exercise of applying these existing estimates to future projections is a highly speculative one. One specific overarching limitation when applying existing esti-

[31] In Appendix Figure A10 and Appendix Table A5, we alternatively set the denominators for both the undocumented and legal net migration rates equal to one half of the total projected population for each origin country. The limited projected differential changes in overall legal versus undocumented net migration reflect the importance of changing demographics in driving the projected divergence seen in Figures 7 and 8 as well as the corresponding tables.

[32] In Appendix Table A4, we correlate net migration rates with two additional origin country measures (the homicide rate and the Gini coefficient of income inequality) as well as two border enforcement measures (log number of Customs and Border Protection agents and log annual ICE budget). Homicide rate estimates are positive for legal immigration and negative for undocumented immigration but are imprecise. While we identify a more robust negative association between the Gini coefficient and legal migration, we are hesitant to apply these parameter estimates to projections since it is unclear how log GDP ratio coefficients should be adjusted to accommodate this measure. Lastly, while we find that increases in the log ICE budget increase legal immigration and reduce illegal immigration, coefficients are exclusively derived from time series variation and so are particularly susceptible to omitted variables bias concerns.

mates to our model is that we are seeking to estimate future stocks of the undocumented and legal immigration populations, while previous research typically analyzes changes in flows. As such, the parameter values we incorporate are best viewed as coarse upper bounds (in terms of magnitudes) on underlying elasticities of immigrant stocks with respect to policy measures. Our projection tool allows the user to adjust the pre-populated estimated effects of policy/enforcement measures on immigration, and we encourage users to evaluate the sensitivity of projection results to alternative parameter values.

We first introduce two measures of border enforcement that characterize future staffing levels and barrier construction. Angelucci (2012) estimates a range of elasticities of illegal immigration with respect to Customs and Border Protection hours and we use the relatively conservative -0.41 parameter value to allow users to predict changes in net migration as a function of changes in CBP staffing.[33] Feigenberg (2020) examines how migration from Mexico to the U.S. responds to changes in border barrier construction and concludes that fence construction in an additional border municipality is associated with a roughly 35% decline in migration for Mexicans who had historically crossed through that municipality. Given 38 Mexican border municipalities, this would imply that border fence construction in a given municipality is expected to reduce migration from Mexico to the U.S. by 0.92 percent. If we further assume similar deterrence effects for those from other origin countries in the LAC region, the implied decline in the net migration rate associated with border fence construction in an additional municipality is 0.074%.[34] We

[33] To convert estimated percent changes to net migration rate changes, we rely on the baseline undocumented net migration rate of 8.1%. An elasticity of -0.41 implies that a one log point increase in agent line watch hours would be expected to reduce the undocumented net migration rate by 3.32%. As noted, an important caveat is that these estimates characterize the change in illegal immigration to the U.S. (as opposed to the change in the stock of undocumented immigrants). Estimates from Angelucci (2012) suggest that the change in the stock of undocumented immigrants associated with marginal increases in staffing is becoming more negative over time.

[34] Incorporating estimates from Feigenberg (2020) relies on a number of strong assumptions regarding the distribution of crossing locations, the remaining set of unfenced municipalities, the applicability of existing deterrence effect estimates to previously-unfenced municipalities, etc. Here as well, we rely on estimates of the change in migration to the U.S. in response to fence construction as opposed to estimates of the change in the stock of immigrants. While it is feasible to construct stock-based

impose the assumption that border fence construction, like changes in agent linewatch hours, impact illegal but not legal migration. Given that a substantial majority of undocumented immigrants from the Western Hemisphere arrive to the U.S. by illegally crossing the U.S.-Mexico border (Hanson, 2009), we apply these estimated deterrence effects to project undocumented immigrant populations for all origin countries in our sample.

We next consider interior immigration enforcement and we draw on Miles and Cox (2014), who estimate that Secure Communities enforcement (which required information on arrested individuals to be shared with the Department of Homeland Security) was associated with 1.13 percent of the non-citizen population being detained between 2008 and 2012.[35] Finally, to incorporate changes in legal immigration policies, we draw on Ortega and Peri (2013), which analyzes bilaterial immigration flows to OECD destinations and their responsiveness to a coarse measure of "immigration policy tightness" that reflects whether legislation passed in a given year increases or reduces restrictions on legal immigration.[36] The authors estimate that a one-unit increase in tightness in the sample of non-European destination countries is associated with a 6% decline in immigration, corresponding to a decline of 1.00 in the legal net migration rate given a base legal net migration rate of 16.69%.

estimates specific to undocumented immigration in this instance, the range of resultant estimates is wide and overlaps with our chosen parameter value.

[35] Although Secure Communities can result in the deportation of any non-citizen who is eligible for deportation, which includes both legal and undocumented immigrants, undocumented immigrants will be disproportionately eligible for deportation. Given uncertainty regarding the precise share of deportees without legal status, we make the simplifying assumption that all are undocumented immigrants. An important caveat underlying this estimate is that, while the vast majority of the population was subject to Secure Communities enforcement by the end of 2012, enforcement rolled out gradually across counties. Since there are other potential biases that may alternatively lead the 1.13 percent estimate to overstate the true effect of Secure Communities on the undocumented population (i.e., the fact that not all detained individuals are ultimately deported), we rely on this estimate in our benchmark projection model.

[36] Although the authors' estimates are based on OECD destination country immigration patterns, we apply them to the U.S. context given the importance of incorporating changes in legal immigration policy in addition to changes in immigration policy enforcement and given that the effects of U.S. visa policy changes cannot be rigorously identified in isolation since there is no cross-state policy variation that can be readily exploited. As for the border enforcement measures, a critical caveat is that the authors' estimates relate to immigration flows rather than stocks.

It is worth noting that we have excluded a large number of candidate factors affecting migration flows from our projection model. First, a number of existing studies have examined the role that various time-invariant factors, including cultural and linguistic similarity as well as geographic distance, play in determining cross-sectional variation in bilateral migration (Belot and Ederveen, 2012; Adserà and Pytliková, 2015; Mayda, 2010). Given our focus on projecting changes in migration levels from a fixed set of origin countries to the U.S., these estimates are not directly applicable. We have also excluded a range of studies that identify predictors of migration that are highly correlated with those already included (Chiquiar and Hanson, 2005; Simpson and Sparber, 2013). In other instances, we exclude factors for which associated estimates are unstable within or across publications.[37] We also abstract away from considering the general equilibrium effects of the policies we include in our analysis, such as the impact of border enforcement on legal immigration flows (see Chassamboulli and Peri, 2020, for a rich model that investigates these interlinkages).

7 Conclusions

Immigration to the United States is among the most powerful forces shaping the economy and society. Accurate data on the size and composition of this population is crucial for basic research, policy analysis, and forecasting future trends. Despite being a large share of the immigrant population, data on undocumented individuals is particularly poor. Traditional data sources on the immigrant population, such as the decennial Census and the American Community Survey, do not contain information on legal status. As such, researchers have estimated the undocumented population either as the difference between the total immigrant population and a measure of the legal immigrant population, or

[37] See, for instance, the varying estimates of the effect of E-Verify employment authorization mandates on the immigrant population as analyzed in Ayromloo et al. (2021) and Bohn et al. (2014).

by imputing legal status to respondents in household surveys based on their observable characteristics.

This project builds on this past work by using administrative data on over 9 million Mexican Matrícula cards that are issued to Mexican nationals in the United States to estimate the stock of undocumented migrants in the United States. Cardholders are almost exclusively in the United States without legal status. As such, the number of these cards and the geographic distribution of cardholders across the United States provides important, new information about the undocumented population.

We draw several key conclusions from our analysis. First, the geographic distribution of Matrícula cards across U.S. counties and commuting zones is highly correlated with the distribution of undocumented migrants obtained by imputing undocumented status to respondents in the American Community Survey. This gives us confidence in using the cards as a new source of information and, in future work, will allow us to improve the procedure for imputing legal status for respondents in household surveys.

Our second conclusion is that data on the Matrícula cards can be used to estimate the undocumented population from other Latin American and Caribbean countries. In particular, we model the cross-sectional relationship between Matrícula cards (normalized by population counts) and observable characteristics of the Mexican-born population and use the results to estimate undocumented population shares for migrants from other countries.

We estimate that there were about 8.3 to 8.7 million undocumented Mexican-born individuals in the United States annually between 2008 and 2012, larger than the 6.2 million estimate by the Pew Center (Passel and Cohn, 2019) using a different methodology. We also estimate about 7.5 to 8.2 million undocumented Mexican-born individuals per year in 2013 to 2017, or about nine percent fewer than in 2008 to 2012. We estimate an additional average of 2.6 to 3.2 and 3.2 to 4.0 million undocumented migrants from the

rest of Latin America and the Caribbean during these periods. These latter estimates are more closely aligned with estimates from the Pew Center.

Finally, we use our new estimates of the recent undocumented population in combination with projections of how population levels and relative GDP affect future migration flows to estimate future legal and undocumented migration through 2060. The key conclusions from this exercise are that net increases in immigration are likely to be among legal immigrants. We project a modest increase in undocumented immigration from Mexico and other Latin American countries over the next two decades, followed by a decline to the current level over the 2040-2060 period.

References

Adserà, A. and M. Pytliková (2015). The Role of Language in Shaping International Migration. *The Economic Journal 125*(586), F49–F81.

Allen, T., C. Dobbin, and M. Morten (2019). Border Walls. Working Paper.

Angelucci, M. (2012). U.S. Border Enforcement and the Net Flow of Mexican Illegal Migration. *Economic Development and Cultural Change 60*(2), 311–357.

Ayromloo, S., B. Feigenberg, and D. Lubotsky (2021). Employment Eligibility Verification Requirements and Local Labor Market Outcomes. Working Paper.

Baker, B. (2017). Population estimates. Department of Homeland Security: Office of Immigration Statistics.

Belot, M. and S. Ederveen (2012). Cultural Barriers in Migration between OECD Countries. *Journal of Population Economics 25*(3), 1077–1105.

Bohn, S., M. Lofstrom, and S. Raphael (2014). Did the 2007 Legal Arizona Workers Act reduce the state's unauthorized immigrant population? *Review of Economics and Statistics 96*(2), 258–269.

Borjas, G. J. (2017). The labor supply of undocumented immigrants. *Labour Economics 46*, 1–13.

Borjas, G. J. and H. Cassidy (2019). The wage penalty to undocumented immigration. *Labour Economics 61*, 101757.

Caballero, M. E., B. Cadena, and B. Kovak (2018). Measuring Geographic Migration Patterns Using Matrículas Consulares. *Demography 55*(3), 1119–1145.

Capps, R., J. Gelatt, J. V. Hook, and M. Fix (2018). Commentary on "The number of undocumented immigrants in the United States: Estimates based on demographic modeling with data from 1990 to 2016". *PLoS ONE 13*(9).

Chassamboulli, A. and G. Peri (2020). The economic effect of immigration policies: analyzing and simulating the U.S. case. *Journal of Economic Dynamics and Control 114*.

Chiquiar, D. and G. Hanson (2005). International Migration, Self-Selection and the Distribution of Wages: Evidence from Mexico and the United States. *Journal of Political Economy 113*(2), 239–281.

Chort, I. and M. de la Rupelle (2016). Determinants of Mexico-U.S. Outward and Return Migration Flows: A State-Level Panel Data Analysis. *Demography 53*(5), 1453–1476.

Fazel-Zarandi, M. M., J. Feinstein, and E. Kaplan (2018). The number of undocumented immigrants in the United States: Estimates based on demographic modeling with data from 1990 to 2016. *PLoS ONE 13*(9).

Feigenberg, B. (2020). Fenced out: The impact of border construction on U.S.-Mexico migration. *American Economic Journal: Applied Economics 12*(3), 106–139.

Hanson, G. H. (2009). The Economics and Policy of Illegal Immigration in the United States. Migration Policy Institute.

Hanson, G. H., C. Liu, and C. McIntosh (2017). The Rise and Fall of U.S. Low-Skilled Immigration. Brookings Papers on Economic Activity.

Massey, D., J. Rugh, and K. Pren (2010). The geography of undocumented Mexican migration. *Mexican Studies/Estudios Mexicanos 26*(1), 129–152.

Mayda, A. (2010). International Migration: A Panel Data Analysis of the Determinants of Bilateral Flows. *Journal of Population Economics 23*(4), 1249–1274.

Miles, T. and A. Cox (2014). Does immigration enforcement reduce crime: Evidence from secure communities. *The Journal of Law and Economics 57*(4).

Ortega, F. and G. Peri (2013). The Effect of Income and Immigration Policies on International Migration. *Migration Studies 1*, 1–28.

Passel, J. S. and D. Cohn (2019, June 25). U.S. unauthorized immigrant total dips to lowest level in a decade. https://www.pewresearch.org/hispanic/wp-content/uploads/sites/5/2019/03/Pew-Research−Center_2018−11−27_U−S−Unauthorized−Immigrants−Total−Dips_Updated−2019−06−25.pdf. Pew Research Center.

Simpson, N. and C. Sparber (2013). The short- and long-run determinants of less-educated immigrant flows into U.S. States. *Southern Economic Journal 80*(2), 414–438.

Transactional Records Access Clearinghouse (2021). Historical Data: Immigration and Customs Enforcement Removals. https://trac.syr.edu/phptools/immigration/removehistory/.

Van Hook, J., F. Bean, J. Bachmeier, and C. Tucker (2014). Recent Trends in Coverage of the Mexican-Born Population of the United States: Results From Applying Multiple Methods Across Time. *Demography 51*, 699–726.

Warren, R. and J. S. Passel (1987). A Count of the Uncountable: Estimates of Undocumented Aliens Counted in the 1980 United States Census. *Demography 24*(3), 375–393.

Warren, R. and J. R. Warren (2014). Unauthorized Immigration to the United States: Annual Estimates and Components of Change, by State, 1990 to 2010. *International Migration Reiew 47*(2), 296–329.

Table 1: Predictors of Matrícula-based Mexican-born Undocumented Population Share (Aggregated Age and Occupational Categories)

	(1) Commuting Zone	(2) County	(3) County (imputed)
Male	0.023	0.322	-0.004
	(0.232)	(0.206)	(0.139)
Log Income (Mexican-Born)	-0.230***	-0.135*	-0.079*
	(0.085)	(0.075)	(0.047)
Years in US	0.002	-0.008	-0.007
	(0.009)	(0.008)	(0.005)
Log Income (Natives)	0.167***	0.112***	0.114***
	(0.060)	(0.039)	(0.037)
Aged 16-40	0.903*	1.343***	0.820***
	(0.547)	(0.406)	(0.280)
Aged 41-65	0.447	0.999**	0.612*
	(0.676)	(0.468)	(0.321)
Aged 66+	1.086	1.681**	0.727
	(1.119)	(0.852)	(0.577)
High School Graduate	0.226	0.070	0.135
	(0.269)	(0.164)	(0.146)
Some College	-0.637*	-0.567**	-0.479***
	(0.367)	(0.224)	(0.165)
Completed College	-0.292	-0.133	-0.238
	(0.484)	(0.338)	(0.233)
Unemployed	-0.973*	-0.827**	-0.514*
	(0.512)	(0.364)	(0.288)
Healthcare and Personal Services	-0.538	-0.801	-0.808*
	(0.783)	(0.699)	(0.465)
Food and Cleaning Services	-0.129	-0.314	-0.024
	(0.360)	(0.251)	(0.179)
Construction, Transportation, and Production	0.001	-0.102	0.037
	(0.358)	(0.267)	(0.193)
Law Enforcement	-4.370**	-1.043	-2.254**
	(2.216)	(0.805)	(1.085)
Agriculture	-0.381	-0.510**	-0.324*
	(0.333)	(0.246)	(0.183)
Observations	741	472	3131

Each column presents coefficients and standard errors from a regression of the Matrícula-based Mexican-born Undocumented Population Share at the specified level of geography on the included covariates using data from 2008-2012. Observations are weighted by the size of the Mexican-born population. Heteroskedasticity-robust standard errors are presented in parentheses. In Column (3), we construct county-level estimates by probabilistically matching ACS PUMAs to counties. * significant at 10 percent level; ** significant at 5 percent level; *** significant at 1 percent level.

Table 2: Population Characteristics by Country/Region of Birth

	(1) Born in Mexico	(2) Born in LAC Region (Outside of Mexico)	(3) Born Outside of LAC Region
Population	15,974	12,997	31,205
	(101,814)	(94,462)	(152,654)
Share Undocumented	0.464	0.344	0.172
(Imputed)	(0.159)	(0.173)	(0.079)
Male	0.576	0.519	0.465
	(0.085)	(0.103)	(0.048)
Log Income	9.911	10.001	10.327
	(0.286)	(0.370)	(0.230)
Years in US	18.431	18.308	23.904
	(3.655)	(4.511)	(4.879)
Aged Under 16	0.054	0.081	0.087
	(0.038)	(0.066)	(0.036)
Aged 16-40	0.534	0.501	0.392
	(0.119)	(0.135)	(0.090)
Aged 41-65	0.366	0.343	0.377
	(0.103)	(0.121)	(0.066)
Aged 66+	0.046	0.075	0.145
	(0.045)	(0.064)	(0.063)
High School Dropout	0.522	0.302	0.137
	(0.119)	(0.158)	(0.063)
High School Graduate	0.272	0.226	0.225
	(0.088)	(0.097)	(0.055)
Some College	0.142	0.265	0.297
	(0.079)	(0.110)	(0.062)
Completed College	0.063	0.207	0.340
	(0.063)	(0.121)	(0.092)
Unemployed	0.212	0.221	0.269
	(0.096)	(0.120)	(0.071)
White-Collar Work	0.157	0.300	0.427
	(0.081)	(0.125)	(0.071)
Healthcare and Personal Services	0.024	0.039	0.053
	(0.029)	(0.040)	(0.023)
Food and Cleaning Services	0.179	0.139	0.080
	(0.093)	(0.093)	(0.034)
Construction, Transportation, and Production	0.331	0.254	0.153
	(0.113)	(0.131)	(0.078)
Law Enforcement	0.003	0.009	0.010
	(0.008)	(0.024)	(0.010)
Agriculture	0.094	0.038	0.007
	(0.097)	(0.063)	(0.011)
Observations	741	741	741

Share Undocumented (Imputed) is derived based on the imputation approach outlined in Borjas and Cassidy (2019). We present unweighted commuting zone-level mean values with standard deviations shown in parenthesis. Summary statistics are constructed for the 2013-2017 period. We exclude individuals aged below 17 when measuring educational attainment and we exclude individuals aged below 16 when constructing occupational variables.

Table 3: Predictors of Deportee Status By Country of Origin

	(1)	(2)	(3)	(4)
	Mexico	El Salvador	Guatemala	Honduras
Male	0.0103***	0.0138***	0.0183***	0.0275***
	(0.0002)	(0.0003)	(0.0003)	(0.0007)
Years in US	-0.0005***	-0.0009***	-0.0008***	-0.0017***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Aged 26-30	0.0009*	-0.0115***	-0.0099***	0.0008
	(0.0005)	(0.0009)	(0.0010)	(0.0018)
Aged 31-35	-0.0017***	-0.0144***	-0.0147***	-0.0083***
	(0.0005)	(0.0008)	(0.0010)	(0.0016)
Aged 36-40	-0.0015***	-0.0159***	-0.0163***	-0.0035**
	(0.0004)	(0.0008)	(0.0009)	(0.0016)
Aged 41-45	-0.0027***	-0.0156***	-0.0156***	-0.0130***
	(0.0004)	(0.0008)	(0.0010)	(0.0015)
Aged 46-50	-0.0024***	-0.0152***	-0.0161***	-0.0101***
	(0.0004)	(0.0008)	(0.0009)	(0.0015)
Aged 51-55	-0.0029***	-0.0137***	-0.0149***	-0.0066***
	(0.0004)	(0.0008)	(0.0009)	(0.0015)
Aged 56-60	-0.0024***	-0.0132***	-0.0126***	-0.0022
	(0.0004)	(0.0008)	(0.0010)	(0.0016)
Aged 61-65	-0.0016***	-0.0120***	-0.0106***	0.0032*
	(0.0004)	(0.0008)	(0.0010)	(0.0018)
Aged 66+	-0.0000	-0.0106***	-0.0086***	0.0105***
	(0.0004)	(0.0008)	(0.0010)	(0.0018)
High School Graduate	-0.0096***	-0.0082***	-0.0154***	-0.0171***
	(0.0002)	(0.0003)	(0.0005)	(0.0008)
Some College	-0.0096***	-0.0104***	-0.0168***	-0.0235***
	(0.0002)	(0.0003)	(0.0004)	(0.0007)
Completed College	-0.0099***	-0.0102***	-0.0156***	-0.0208***
	(0.0002)	(0.0003)	(0.0004)	(0.0008)
Observations	880,838	98,753	66,425	41,314

Observations reflect the raw number of survey respondents. Regression estimates are based on models that incorporate survey weights and append 2008-2017 annual ACS files to the annual EMIF deportee files. Each country-specific regression estimates the associations between the socio-demographic characteristics available in both datasets and an indicator variable for whether a given observation corresponds to an EMIF deportee. Heteroskedasticity-robust standard errors presented in parentheses. * significant at 10 percent level; ** significant at 5 percent level; *** significant at 1 percent level.

Table 4: Mexico-Born and LAC-Born Undocumented Population Estimates (in Thousands)

	(1) Mexico	(2) LAC Region (Excluding Mexico)
Panel A: Basic Estimates		
Row 1: Pew (2008-2012)	6,225	2,850
Row 2: Pew (2013-2017)	5,525	2,925
Row 3: Raw Matrícula (2008-2012)	4,425	2,000
Row 4: Raw Matrícula (2013-2017)	4,675	1,925
Row 5: Matrícula w/ LAC Adjustment (2008-2012)	4,425	1,650
Row 6: Matrícula w/ LAC Adjustment (2013-2017)	4,675	2,275
Panel B: Estimates Adjusted By Matrícula Takeup		
Row 7: Matrícula Takeup Correction (2008-2012)	8,675	3,925
Row 8: Matrícula Takeup Correction (2013-2017)	8,175	3,350
Row 9: Matrícula Takeup Correction w/ LAC Adjustment (2008-2012)	8,675	3,225
Row 10: Matrícula Takeup Correction w/ LAC Adjustment (2013-2017)	8,175	3,975
Panel C: Estimates Adjusted By Matrícula Takeup and Legal Status Changes		
Row 11: Matrícula Takeup Correction w/Legalizations (2008-2012)	8,275	3,325
Row 12: Matrícula Takeup Correction w/Legalizations (2013-2017)	7,525	2,525
Row 13: Matrícula Takeup Correction w/ LAC Adjustment and Legalizations (2008-2012)	8,275	2,625
Row 14: Matrícula Takeup Correction w/ LAC Adjustment and Legalizations (2013-2017)	7,525	3,150

All estimates are rounded to the nearest 25,000. Pew estimates for the population born in LAC countries outside of Mexico are based on averages across the subset of years with available data during the 2008-2012 and 2013-2017 periods. Raw Matrícula refers to raw card counts for the Mexico-born population and to predicted counts for those born elsewhere in the LAC region that do not apply any re-scaling. LAC Adjustment refers to the rescaling of estimates for those born outside of Mexico based on relative rates of interior removals per capita, constructed separately by time period. Matrícula Takeup Correction estimates re-scale raw estimates by 1.96 for the 2008-2012 period and by 1.75 for the 2013-2017 period.

Table 5: Projected Populations by Origin Country and Legal Status

	(1) 2008-2012 Undocumented	(2) 2008-2012 Legal	(3) 2013-2017 Undocumented	(4) 2013-2017 Legal	(5) 2040 Undocumented	(6) 2040 Legal	(7) 2060 Undocumented	(8) 2060 Legal
	Population Estimates (in 1000s)							
Mexico	8380	5560	7720	6310	8378	11239	7389	14106
Brazil	60	230	50	280	-138	239	-397	-147
Colombia	90	510	90	610	92	1060	12	1151
Dominican Republic	120	740	130	920	209	1615	196	2112
Ecuador	60	290	50	320	87	648	38	821
El Salvador	450	550	460	650	452	1075	392	1395
Guatemala	740	330	770	390	1155	1064	1160	1693
Haiti	30	490	30	570	60	1180	17	1667
Honduras	450	170	510	240	608	574	558	835
Peru	40	300	40	360	158	772	153	948
Rest of LAC (Combined)	2040	3610	2130	4340	2821	8226	2524	10623

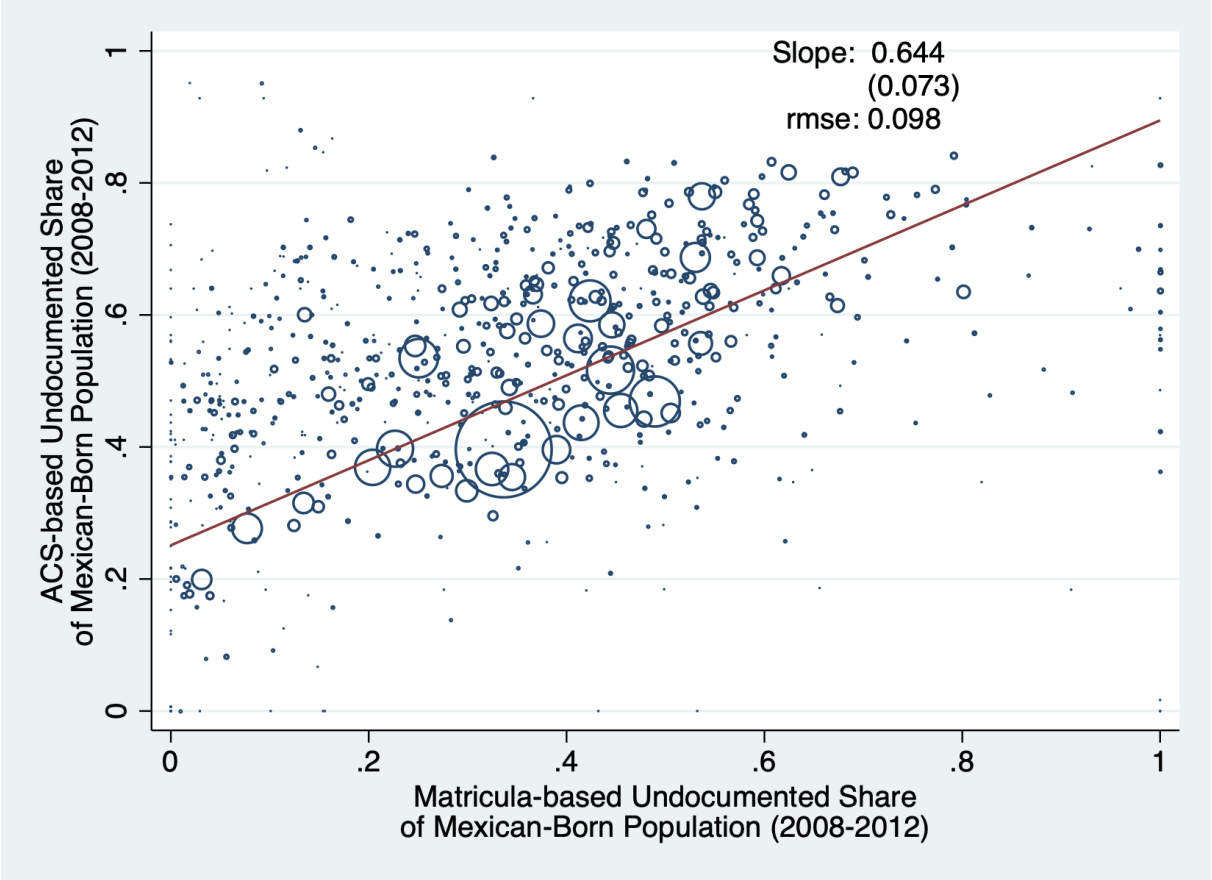
Each row presents the estimated population (for 2008-2012 and 2013-2017) and projected population (for 2040 and 2060) for a given country of origin, separately by immigrant legal status.

Table 6: Projected Populations by Origin Country and Legal Status w/ Heterogeneous Predictors by Legal Status

	(1) 2008-2012 Undocumented	(2) 2008-2012 Legal	(3) 2013-2017 Undocumented	(4) 2013-2017 Legal	(5) 2040 Undocumented	(6) 2040 Legal	(7) 2060 Undocumented	(8) 2060 Legal
	Population Estimates (in 1000s)							
Mexico	8380	5560	7720	6310	8245	11382	7050	14622
Brazil	60	230	50	280	-2788	868	-3952	974
Colombia	90	510	90	610	-513	1158	-989	1378
Dominican Republic	120	740	130	920	247	1610	114	2124
Ecuador	60	290	50	320	252	626	74	815
El Salvador	450	550	460	650	438	1083	361	1421
Guatemala	740	330	770	390	1178	1039	1142	1724
Haiti	30	490	30	570	763	1159	467	1647
Honduras	450	170	510	240	602	582	531	896
Peru	40	300	40	360	3	788	-258	1003
Rest of LAC (Combined)	2040	3610	2130	4340	3484	8913	2687	11982

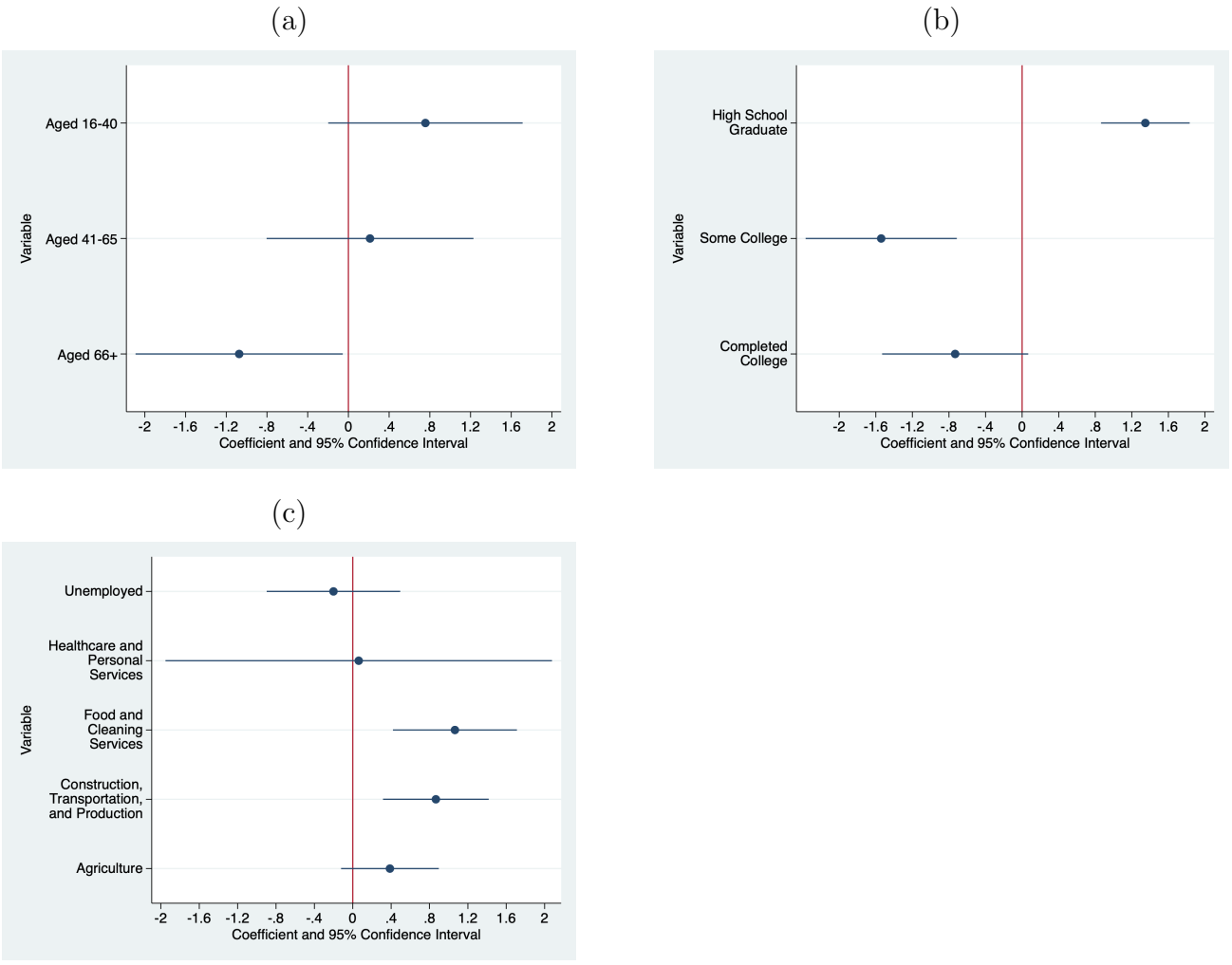
Each row presents the estimated population (for 2008-2012 and 2013-2017) and projected population (for 2040 and 2060) for a given country of origin, separately by immigrant legal status. Based on exploratory analyses using historical data, we impose the assumption that changes in the log cohort size ratio for those aged under 40 affect only undocumented immigrant population flows, while changes in the log cohort size ratio for those aged above 40 affect only legal immigrant population flows.

Figure 1: Commuting Zone-level Relationship Between ACS-based and Matrícula-based Measures



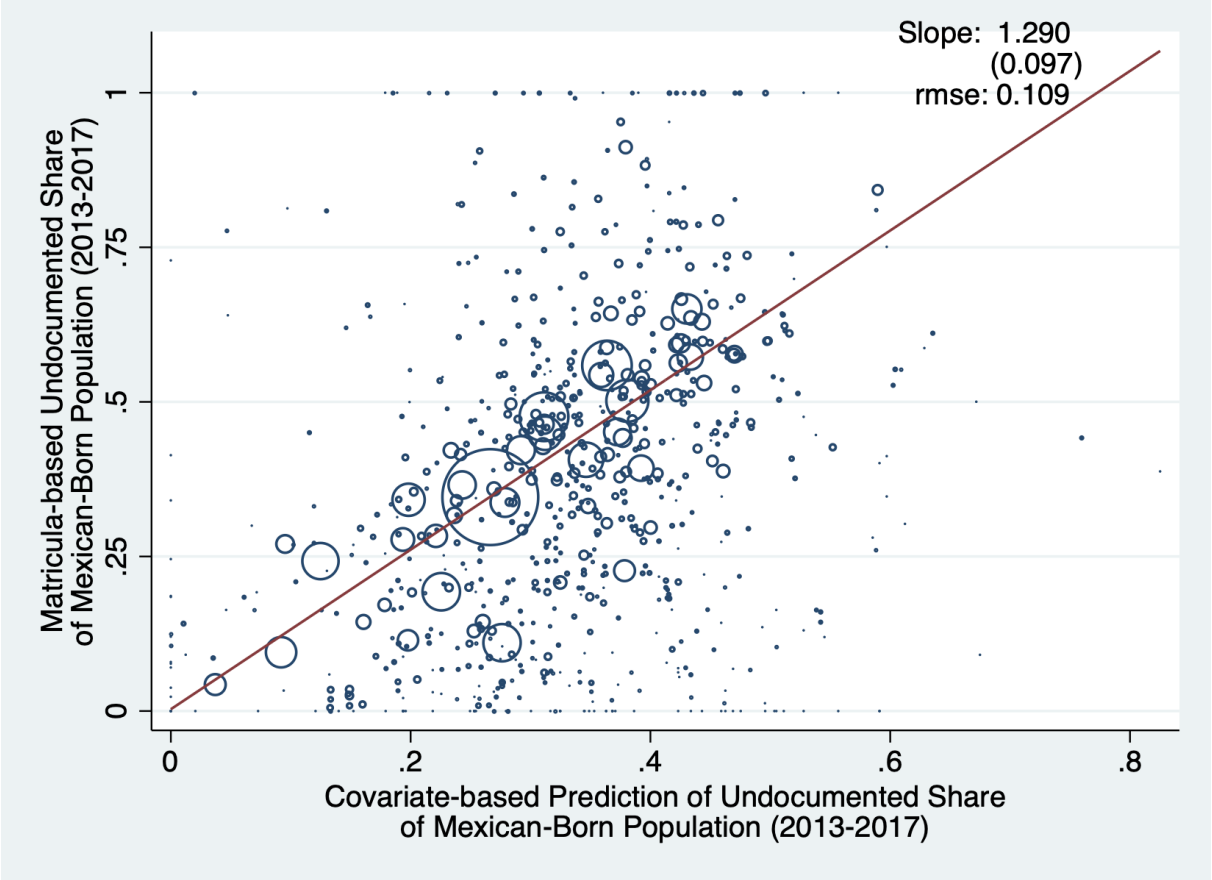
Notes: This figure presents a scatter plot comparing the predicted share of the relevant population that is undocumented based on the Borjas and Cassidy (2019) imputation-based approach (on the vertical axis) and based on Matrícula card counts (on the horizontal axis). Undocumented population shares are constructed using data from 2008-2012. Each commuting zone-level observation is weighted by the number of Mexican-born residents in that commuting zone (as measured in the ACS).

Figure 2: Commuting Zone-level Predictors of the Mexican-Born Undocumented Population Share



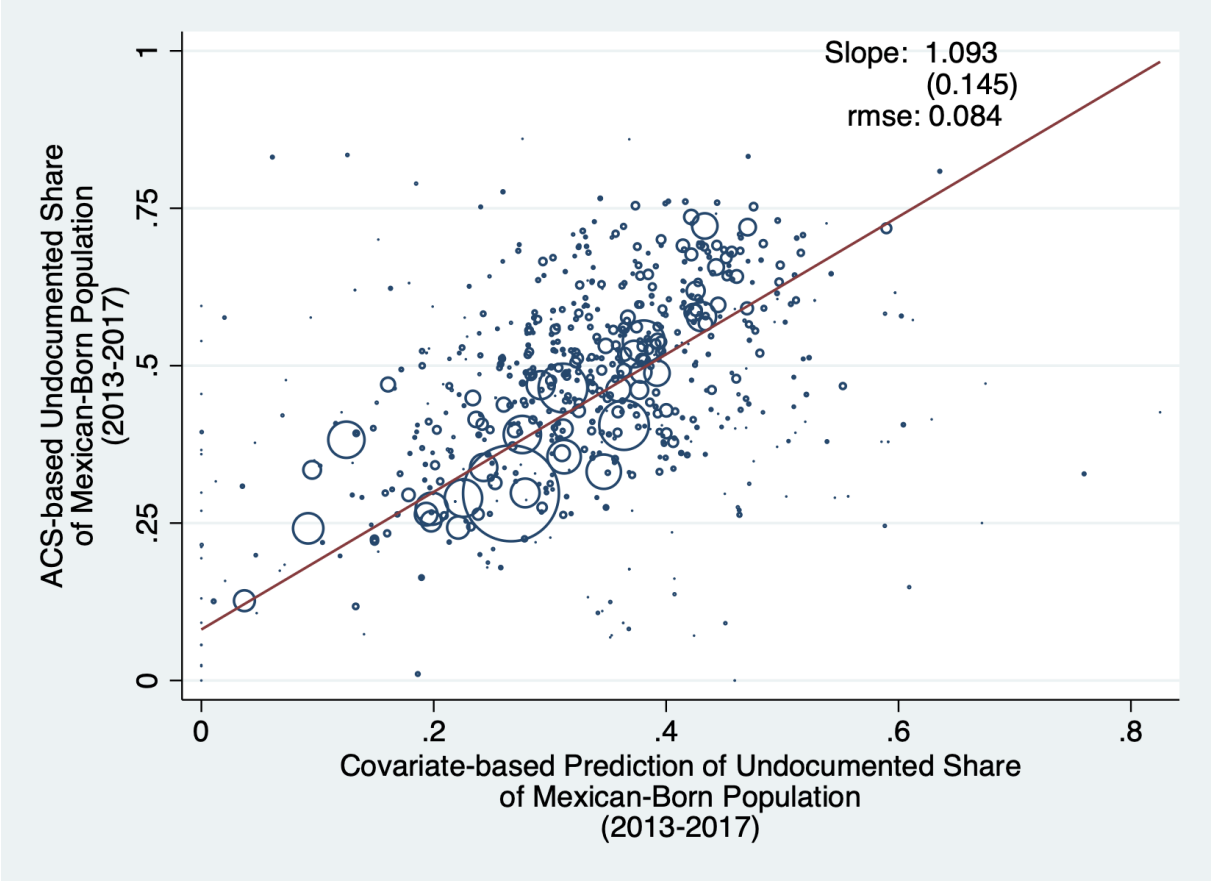
Notes: Each panel plots coefficient estimates and 95% confidence intervals from regressions of the Matrícula-based Mexican-Born Undocumented Population Share on the included covariates using data from 2008-2012. The omitted category from panel (a) is Aged Under 16, the omitted category from panel (b) is High School Dropout, and the omitted category from panel (c) is White-Collar Workers. Each commuting zone-level observation is weighted by the number of Mexican-born residents (as measured in the ACS). Confidence intervals are constructed from heteroskedasticity-robust standard errors.

Figure 3: OLS-Based Commuting Zone-level Prediction of Undocumented Mexican-Born Population (Matrícula)



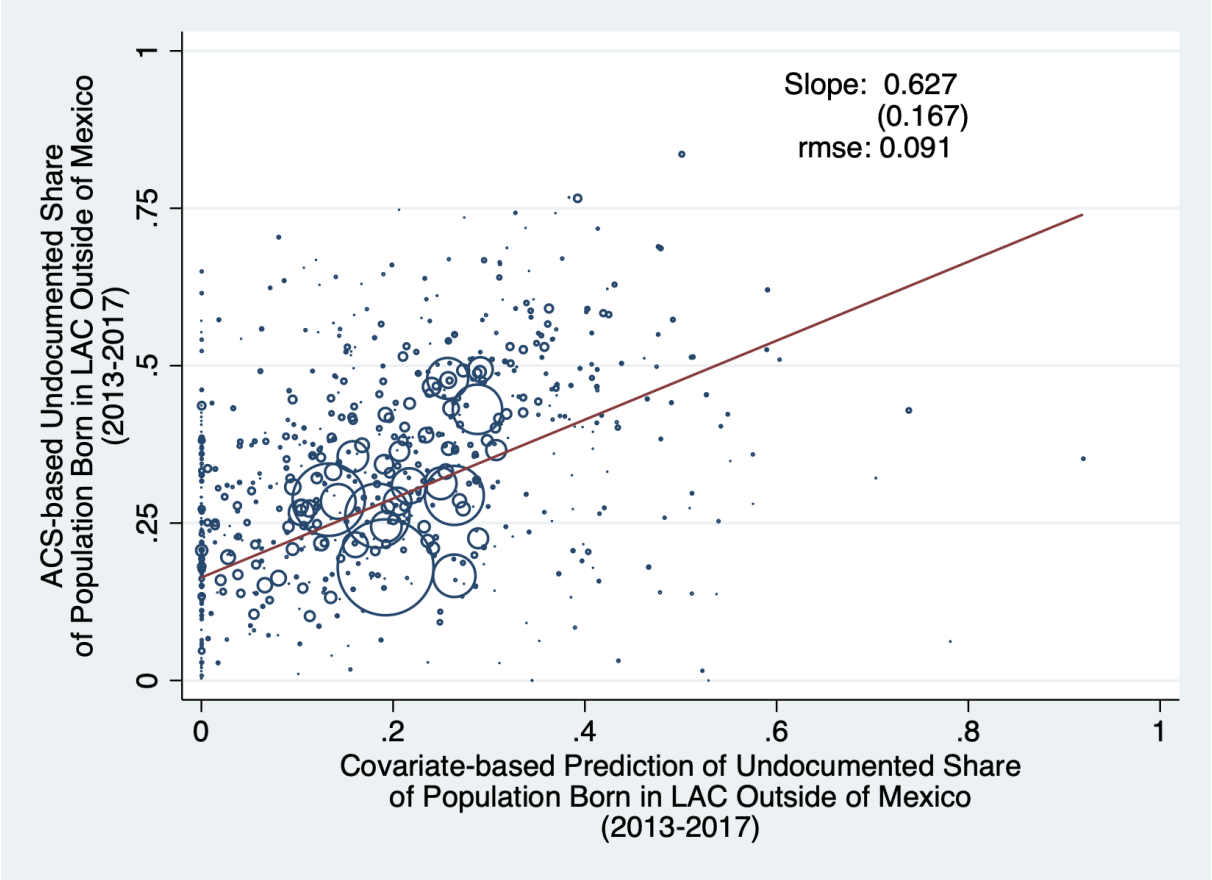
Notes: This figure presents a scatter plot comparing the predicted share of the relevant population that is undocumented based on the ratio of the number of valid Matrícula cards to the Mexican-born population (on the vertical axis) and based on our Matrícula-based prediction model (on the horizontal axis). Specifically, the Matrícula-based prediction is formed as $\hat{\alpha} + \hat{\beta}\bar{x}_c$, where $\bar{x}_c$ is the average characteristics of the relevant foreign-born population in commuting zone c. $\hat{\alpha}$ and $\hat{\beta}$ are the intercept and slope estimates from the model described in Section 4.2 and presented in Table A1 (estimated using data from the 2008-2012 period). Each commuting zone-level observation is weighted by the number of Mexican-born residents (as measured in the ACS).

Figure 4: OLS-Based Commuting Zone-level Prediction of Undocumented Mexican-Born Population (ACS)



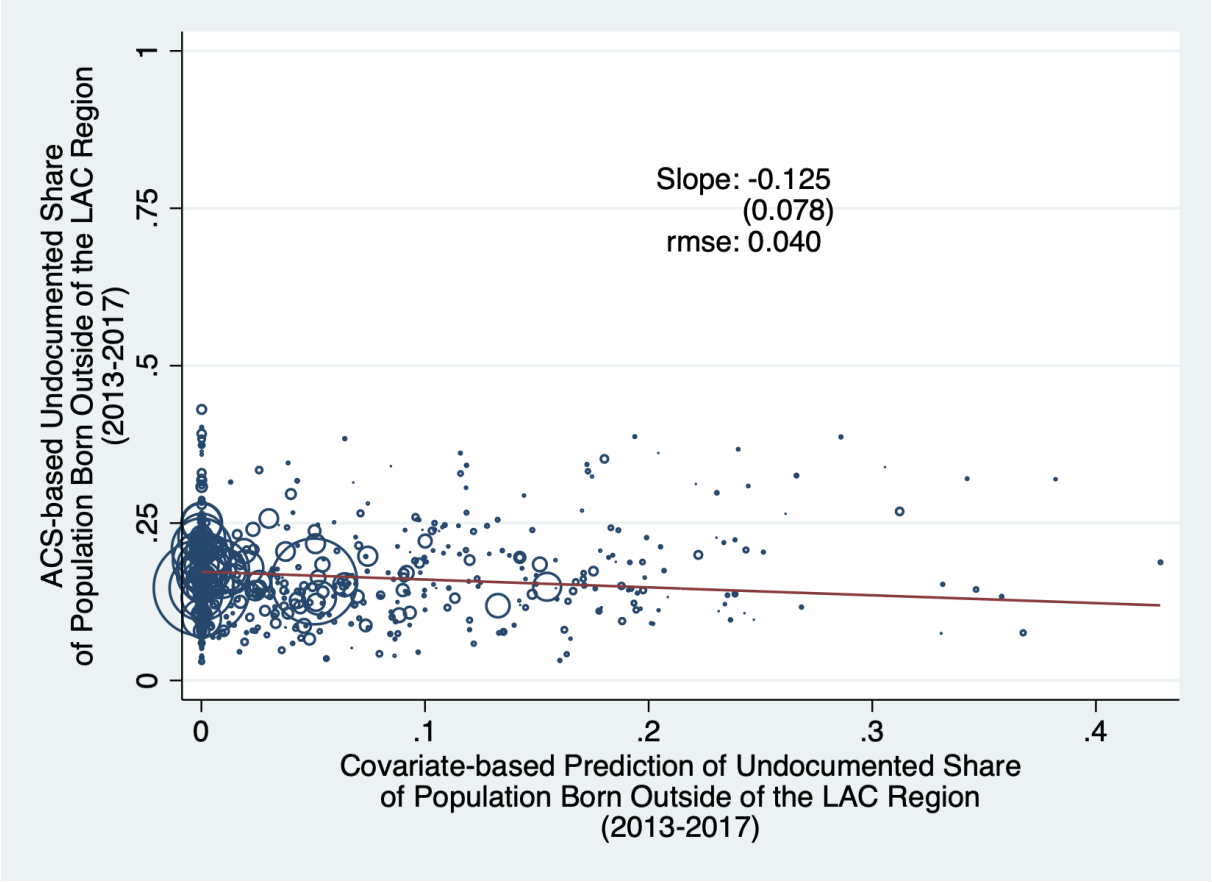
Notes: This figure presents a scatter plot comparing the predicted share of the relevant population that is undocumented based on the Borjas and Cassidy (2019) imputation-based approach (on the vertical axis) and based on our Matrícula-based prediction model (on the horizontal axis). Specifically, the Matrícula-based prediction is formed as $\hat{\alpha} + \hat{\beta}\bar{x}_c$, where $\bar{x}_c$ is the average characteristics of the relevant foreign-born population in commuting zone c. $\hat{\alpha}$ and $\hat{\beta}$ are the intercept and slope estimates from the model described in Section 4.2 and presented in Table A1 (estimated using data from the 2008-2012 period). Each commuting zone-level observation is weighted by the number of Mexican-born residents (as measured in the ACS).

Figure 5: OLS-Based Commuting Zone-level Prediction of Undocumented LAC-Born Population (ACS)



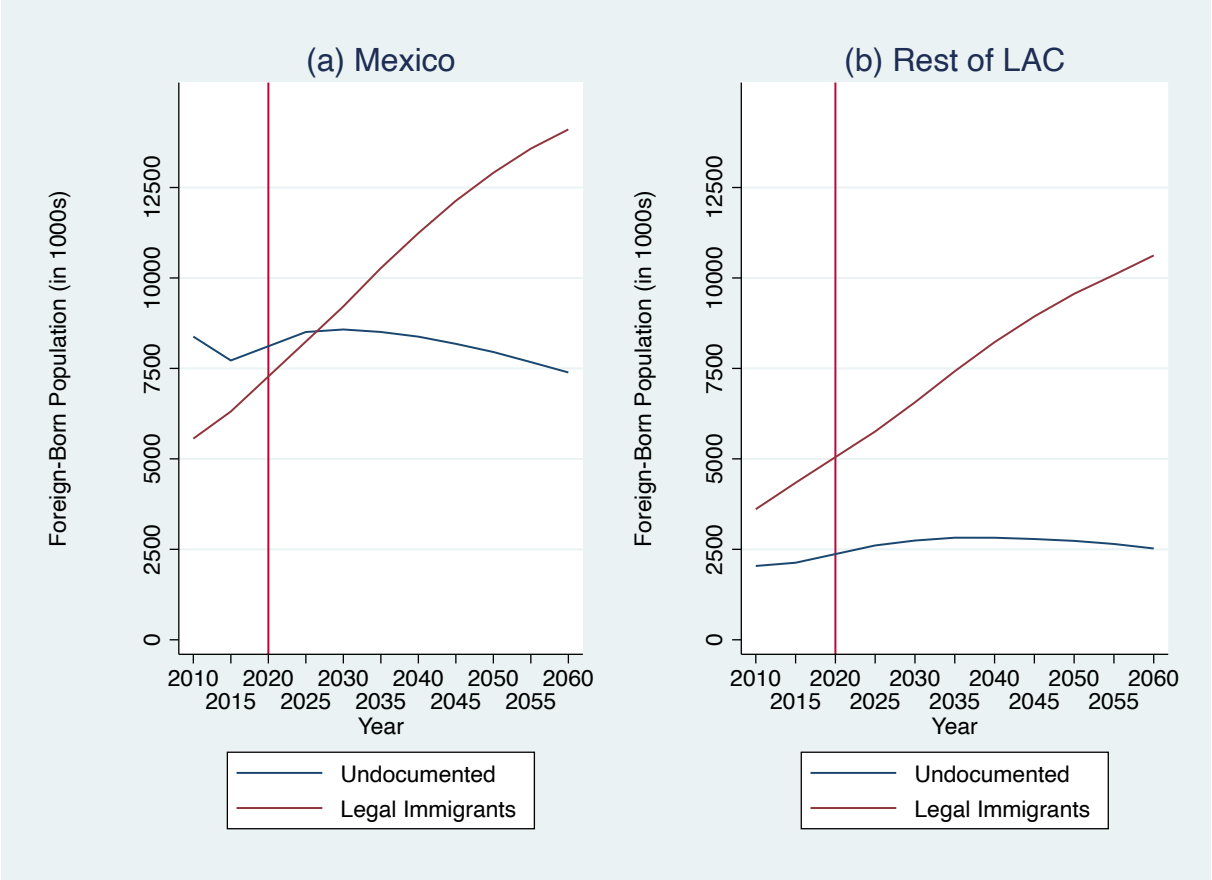
Notes: This figure presents a scatter plot comparing the predicted share of the relevant population (those born in the LAC region outside of Mexico) that is undocumented based on the Borjas and Cassidy (2019) imputation-based approach (on the vertical axis) and based on our Matrícula-based prediction model (on the horizontal axis). Specifically, the Matrícula-based prediction is formed as $\hat{\alpha} + \hat{\beta}\bar{x}_c$, where $\bar{x}_c$ is the average characteristics of the relevant foreign-born population in commuting zone c. $\hat{\alpha}$ and $\hat{\beta}$ are the intercept and slope estimates from the model described in Section 4.2 and presented in Table A1 (estimated using data from the 2008-2012 period). Each commuting zone-level observation is weighted by the number of LAC-born residents (as measured in the ACS and excluding Mexico).

Figure 6: OLS-Based Commuting Zone-level Prediction of Undocumented Population Born Outside of the LAC region (ACS)



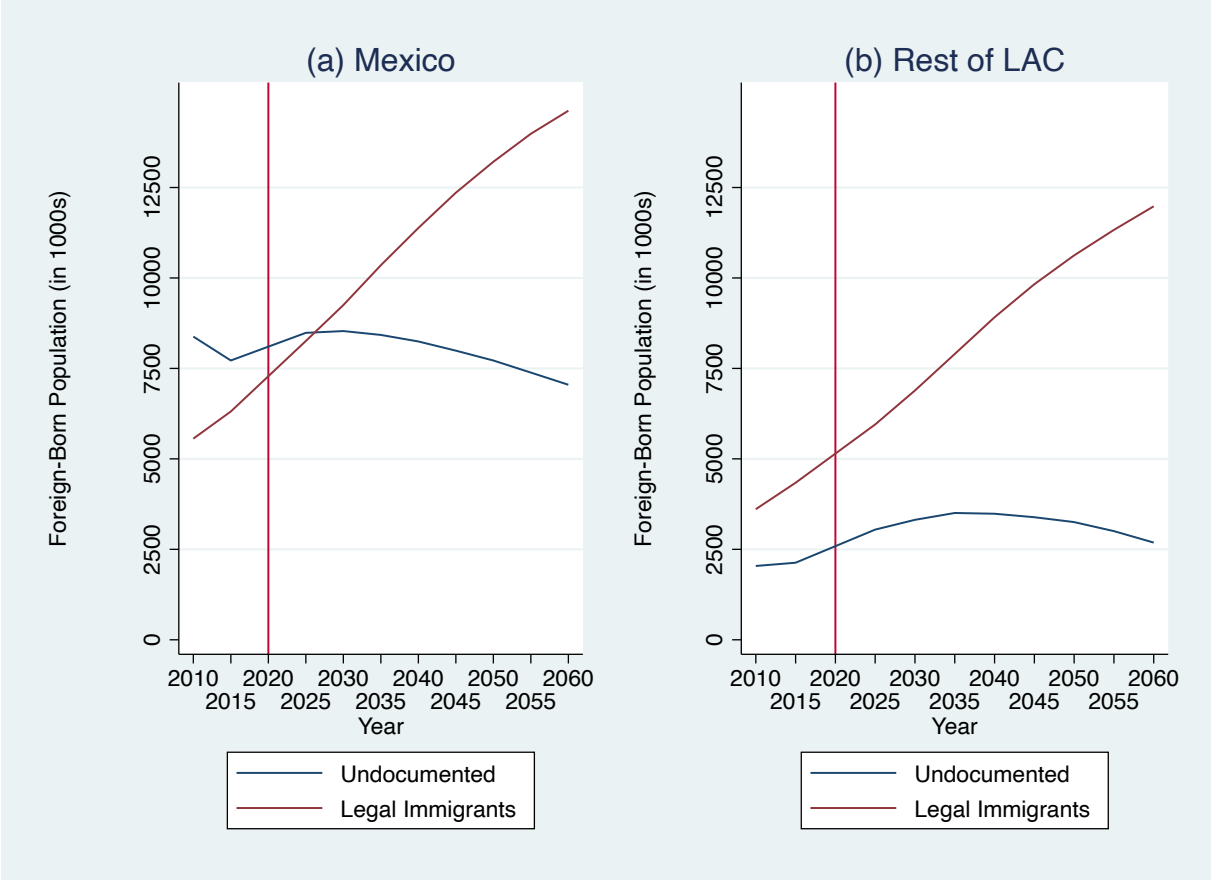
Notes: This figure presents a scatter plot comparing the predicted share of the relevant population (those born outside of the LAC region) that is undocumented based on the Borjas and Cassidy (2019) imputation-based approach (on the vertical axis) and based on our Matrícula-based prediction model (on the horizontal axis). Specifically, the Matrícula-based prediction is formed as $\hat{\alpha} + \hat{\beta}\bar{x}_c$, where $\bar{x}_c$ is the average characteristics of the relevant foreign-born population in commuting zone c. $\hat{\alpha}$ and $\hat{\beta}$ are the intercept and slope estimates from the model described in Section 4.2 and presented in Table A1 (estimated using data from the 2008-2012 period). Each commuting zone-level observation is weighted by the number of foreign-born residents from outside of the LAC region (as measured in the ACS).

Figure 7: Historical and Projected Population Estimates for Mexico and Rest of the LAC Region



Notes: This figure plots the estimated/projected number of legal and undocumented immigrants separately for Mexico and the rest of the LAC region.

Figure 8: Historical and Projected Population Estimates for Mexico and Rest of the LAC Region w/ Heterogeneous Predictors by Legal Status



Notes: This figure plots the estimated/projected number of legal and undocumented immigrants separately for Mexico and the rest of the LAC region. Based on exploratory analyses using historical data, we impose the assumption that changes in the log cohort size ratio for those aged under 40 affect only undocumented immigrant population flows, while changes in the log cohort size ratio for those aged above 40 affect only legal immigrant population flows.

Appendix A: Tables and Figures

Table A1: Predictors of Matrícula-based Mexican-born Undocumented Population Share (Full Model)

	(1) Commuting Zone	(2) County	(3) County (imputed)
Male	0.180	0.331	0.007
	(0.217)	(0.202)	(0.136)
Log Income (Mexican-Born)	-0.165*	-0.133*	-0.070
	(0.086)	(0.074)	(0.047)
Years in US	0.004	-0.000	-0.007
	(0.008)	(0.008)	(0.005)
Log Income (Natives)	0.164***	0.113***	0.121***
	(0.057)	(0.041)	(0.037)
Aged 16-25	-0.391	0.913	0.194
	(0.651)	(0.556)	(0.376)
Aged 26-30	-0.287	0.766	0.461
	(0.706)	(0.524)	(0.367)
Aged 31-35	-0.397	0.694	0.227
	(0.618)	(0.573)	(0.347)
Aged 36-40	-0.813	0.339	0.382
	(0.704)	(0.525)	(0.385)
Aged 41-45	-1.430*	0.459	0.095
	(0.742)	(0.573)	(0.405)
Aged 46-50	-1.277	0.263	0.102
	(0.805)	(0.624)	(0.409)
Aged 51-55	-0.488	0.463	-0.033
	(0.945)	(0.694)	(0.433)
Aged 56-60	-0.961	-0.118	0.366
	(0.945)	(0.771)	(0.520)
Aged 61-65	-0.171	-0.870	0.226
	(1.080)	(0.820)	(0.564)
Aged 66+	-0.551	1.163	0.190
	(1.039)	(0.806)	(0.575)
High School Graduate	0.124	0.037	0.128
	(0.232)	(0.152)	(0.138)
Some College	-0.571	-0.630***	-0.462***
	(0.354)	(0.221)	(0.161)
Completed College	0.044	0.260	-0.078
	(0.487)	(0.418)	(0.246)
Unemployed	0.061	-0.353	-0.094
	(0.502)	(0.428)	(0.308)
Healthcare Services	2.102*	0.228	0.237
	(1.171)	(1.295)	(0.678)
Personal Care Services	-0.639	-0.694	-0.690
	(0.932)	(0.822)	(0.594)
Law Enforcement	-4.092*	-0.432	-1.928*
	(2.278)	(0.784)	(1.152)
Sales	1.923***	0.709	0.738**
	(0.555)	(0.442)	(0.298)
Food Services	0.960**	0.354	0.507*
	(0.447)	(0.420)	(0.273)
Cleaning and Maintenance	0.793*	-0.052	0.208
	(0.472)	(0.398)	(0.254)
Agriculture	0.723*	-0.005	0.173
	(0.397)	(0.396)	(0.238)
Construction	0.981**	0.345	0.544**
	(0.395)	(0.410)	(0.244)
Transportation	0.811	0.045	0.361
	(0.512)	(0.472)	(0.335)
Production Occupations	1.085**	0.535	0.448
	(0.428)	(0.444)	(0.273)
Observations	741	472	3131

Each column presents coefficients and standard errors from a regression of the Matrícula-based Mexican-born Undocumented Population Share at the specified level of geography on the included covariates using data from 2008-2012. Observations are weighted by the size of the Mexican-born population. Heteroskedasticity-robust standard errors are presented in parentheses. In Column (3), we construct county-level estimates by probabilistically matching ACS PUMAs to counties. * significant at 10 percent level; ** significant at 5 percent level; *** significant at 1 percent level.

Table A2: Immigrant Population Prediction Parameters

	(1) Net Migration Rate (Legal)	(2) Net Migration Rate (Undocumented)
Log Birth Cohort Ratio for Origin Country to United States × Under 40 Years Old Reference: Hanson et al. (2017)	4.7008	4.7008
Log Birth Cohort Ratio for Origin Country to United States × Over 40 Years Old Reference: Hanson et al. (2017)	0.7716	0.7716
Log GDP Ratio for Origin Country to United States × Under 40 Years Old Reference: Hanson et al. (2017)	0.538	0.538
Log GDP Ratio for Origin Country to United States × Over 40 Years Old Reference: Hanson et al. (2017)	1.9581	1.9581
Log Agent Line Watch Hours Reference: Angelucci (2012)	–	-3.32
Number of Fenced Mexican Border Municipalities Reference: Feigenberg (2020)	–	-0.074
Share Secure Communities Counties Reference: Miles and Cox (2014)	–	-0.092
Immigration Policy Tightness Reference: Ortega and Peri (2013)	-1.00	–

Each row presents the estimated responsiveness of the legal (Column 1) and undocumented (Column 2) net migration rate to a one-unit change in the relevant regressor.

Table A3: Immigrant Population Prediction Model

	(1) Net Migration Rate	(2) Net Mig Rate	(3) Net Mig Rate	(4) Net Mig Rate	(5) Net Mig Rate	(6) Net Mig Rate	(7) Net Mig Rate	(8) Net Mig Rate	(9) Net Mig Rate	(10) Net Mig Rate
		Matricula-Based Estimates					Pew-Based Estimates			
Log Birth Cohort Ratio (Under 40)	8.54					-0.27				
	(16.64)					(6.59)				
Log Birth Cohort Ratio (Over 40)	-9.62					-5.15				
	(11.74)					(4.30)				
Log GDP Ratio	12.68					5.44**				
	(8.23)					(2.72)				
Log Birth Cohort Ratio (Under 40)		-13.94		-6.61	9.06**		-16.76**		-13.38**	5.74
		(12.53)		(9.83)	(3.34)		(6.88)		(6.24)	(3.61)
×*Legal*										
Log Birth Cohort Ratio (Under 40)		18.00		23.70	29.38*		9.73		12.84	29.12***
		(11.95)		(14.44)	(14.69)		(9.46)		(9.32)	(8.53)
×*Undocumented*										
Log Birth Cohort Ratio (Over 40)		9.20***		3.51	-27.44***		10.16***		7.46***	-28.87***
		(2.85)		(2.12)	(4.85)		(1.61)		(1.31)	(4.71)
×*Legal*										
Log Birth Cohort Ratio (Over 40)		-18.33***		-22.75***	-33.99**		-15.28***		-17.76***	-46.32***
		(4.65)		(6.32)	(12.45)		(3.15)		(3.34)	(9.54)
×*Undocumented*										
Log GDP Ratio			18.58***	14.27**	12.98***			10.87***	5.67***	6.20***
×*Legal*			(3.81)	(6.61)	(2.93)			(2.02)	(1.65)	(1.27)
Log GDP Ratio			-19.12	11.09	10.62			-8.35	5.21*	6.53**
×*Undocumented*			(17.12)	(8.67)	(9.05)			(5.31)	(2.69)	(2.75)
Observations	40	40	40	40	40	200	200	200	200	200
Origin Country× Legal Status FE	X	X	X	X	X	X	X	X	X	X
Time Period× Legal Status FE					X					X

Net Migration Rate (abbreviated as Net Mig Rate) refers to the legal/undocumented immigrant count with each normalized by one half of the origin country population from the same time period. Observations in Columns 1-5 rely on Matrícula-based estimates and are at the 5 year period (2008-2012 or 2013-2017) by country of origin by legal status (legal or undocumented) level. Observations in Columns 6-10 rely on Pew-based estimates and are at the year by country of origin by legal status (legal or undocumented) level. We present heteroskedasticity-robust standard errors and weight observations by the baseline immigrant population associated with the given country of origin and legal status.

Table A4: Additional Immigrant Population Predictors

	(1) Net Migration Rate	(2) Net Mig Rate	(3) Net Mig Rate	(4) Net Mig Rate	(5) Net Mig Rate	(6) Net Mig Rate	(7) Net Mig Rate	(8) Net Mig Rate
	Matricula-Based Estimates				Pew-Based Estimates			
Homicide Rate	-0.0002			0.002				
	(0.06)			(0.01)				
Gini Coefficient	0.001			-0.3***				
	(0.25)			(0.10)				
Homicide Rate		0.004	0.025		0.02	0.02**		0.01
×*Legal*		(0.07)	(0.07)		(0.01)	(0.01)		(0.01)
Homicide Rate		-0.004	-0.06		-0.02	-0.01		-0.01
×*Undocumented*		(0.07)	(0.09)		(0.03)	(0.02)		(0.01)
Gini Coefficient		-0.37*	-0.18		-0.35***	-0.22***		-0.19***
×*Legal*		(0.18)	(0.15)		(0.06)	(0.07)		(0.07)
Gini Coefficient		0.37	-0.10		0.07	-0.35***		-0.35***
×*Undocumented*		(0.42)	(0.22)		(0.16)	(0.11)		(0.11)
Log CBP Agents				0.40				
				(3.22)				
Log ICE Budget				-4.40				
				(3.42)				
Log CBP Agents							0.80	0.41
×*Legal*							(0.96)	(0.89)
Log CBP Agents							1.11	0.38
×*Undocumented*							(2.55)	(2.10)
Log ICE Budget							6.92***	4.91***
×*Legal*							(0.90)	(0.96)
Log ICE Budget							-9.99***	-13.72***
×*Undocumented*							(2.58)	(2.29)
Observations	40	40	40	200	200	200	200	200
Origin Country× Legal Status FE	X	X	X	X	X	X	X	X
Time Period× Legal Status FE			X			X		

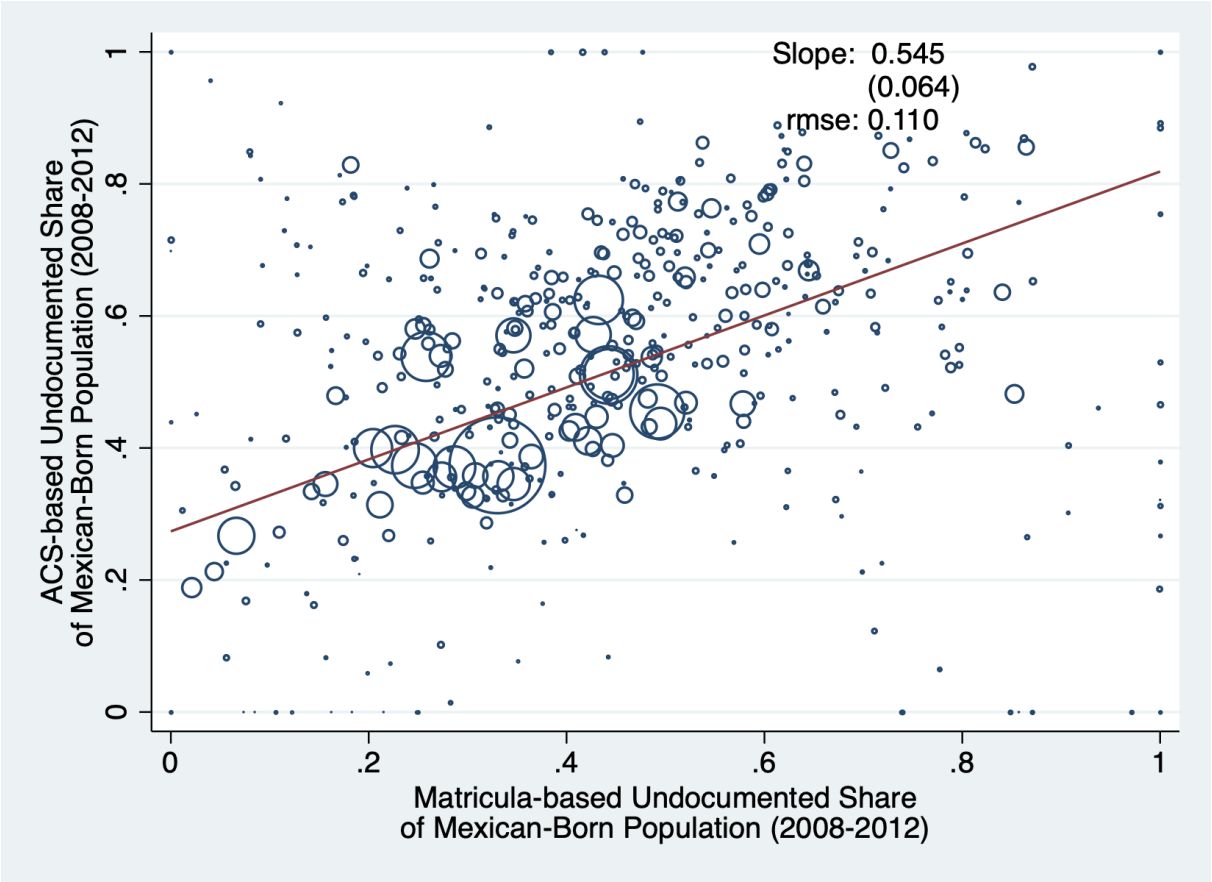
Net Migration Rate (abbreviated as Net Mig Rate) refers to the legal/undocumented immigrant count with each normalized by one half of the origin country population from the same time period. Observations in Columns 1-3 rely on Matrícula-based estimates and are at the 5 year period (2008-2012 or 2013-2017) by country of origin by legal status (legal or undocumented) level. Observations in Columns 4-8 rely on Pew-based estimates and are at the year by country of origin by legal status (legal or undocumented) level. We present heteroskedasticity-robust standard errors and weight observations by the baseline immigrant population associated with the given country of origin and legal status.

Table A5: Projected Populations by Origin Country and Legal Status w/o Differential Age-Based Exposure to Legal/Undocumented Immigration

	(1) 2008-2012 Undocumented	(2) 2008-2012 Legal	(3) 2013-2017 Undocumented	(4) 2013-2017 Legal	(5) 2040 Undocumented	(6) 2040 Legal	(7) 2060 Undocumented	(8) 2060 Legal
	Population Estimates (in 1000s)							
Mexico	8380	5560	7720	6310	10451	7712	10952	8052
Brazil	60	230	50	280	-130	99	-474	-251
Colombia	90	510	90	610	121	696	43	624
Dominican Republic	120	740	130	920	228	1149	248	1251
Ecuador	60	290	50	320	98	450	64	460
El Salvador	450	550	460	650	557	733	585	772
Guatemala	740	330	770	390	1327	697	1599	836
Haiti	30	490	30	570	61	767	32	842
Honduras	450	170	510	240	750	343	845	379
Peru	40	300	40	360	162	544	177	588
Rest of LAC (Combined)	2040	3610	2130	4340	3304	5478	3594	5753

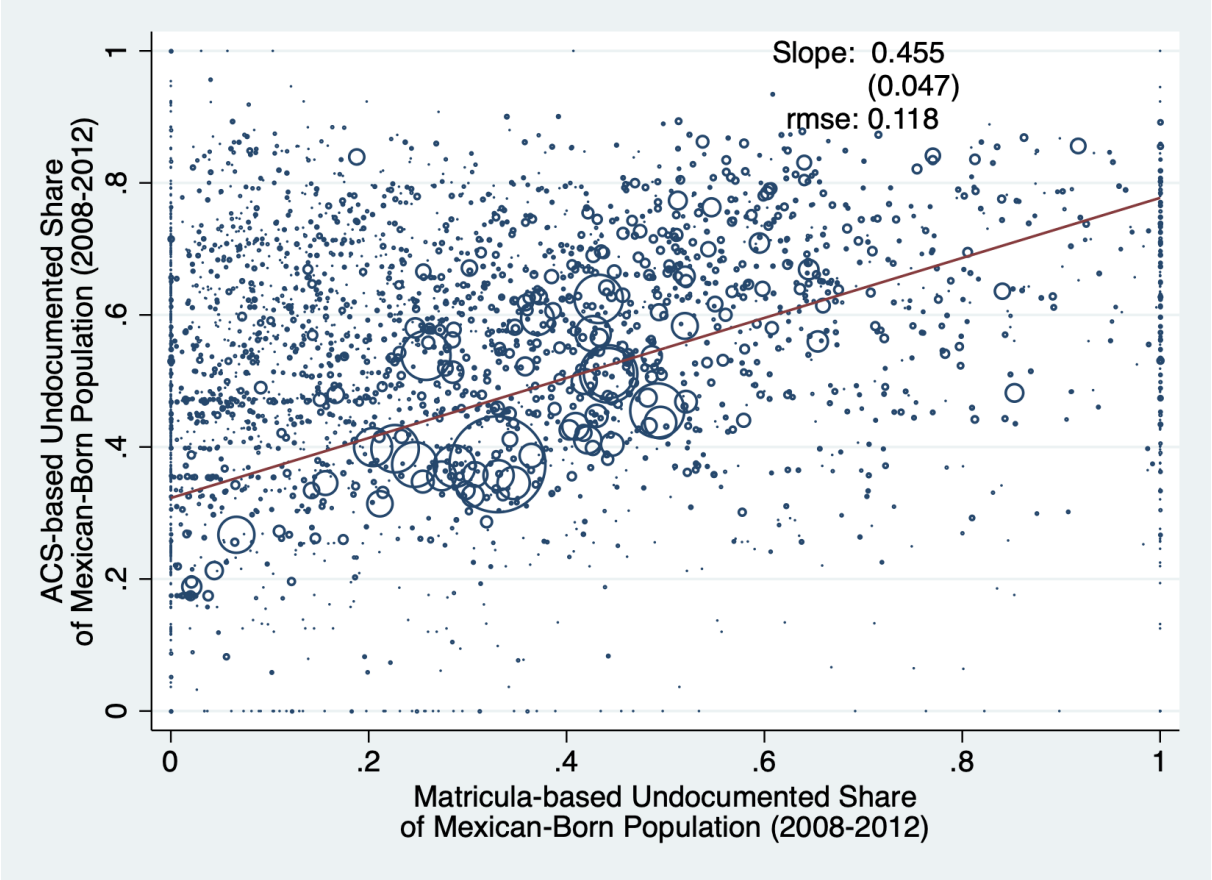
Each row presents the estimated population (for 2008-2012 and 2013-2017) and projected population (for 2040 and 2060) for a given country of origin, separately by immigrant legal status. For these projections, we set the denominators for both the undocumented and legal net migration rates equal to one half of the total projected population for each origin country.

Figure A1: County-level Relationship Between ACS-based and Matrícula-based Measures



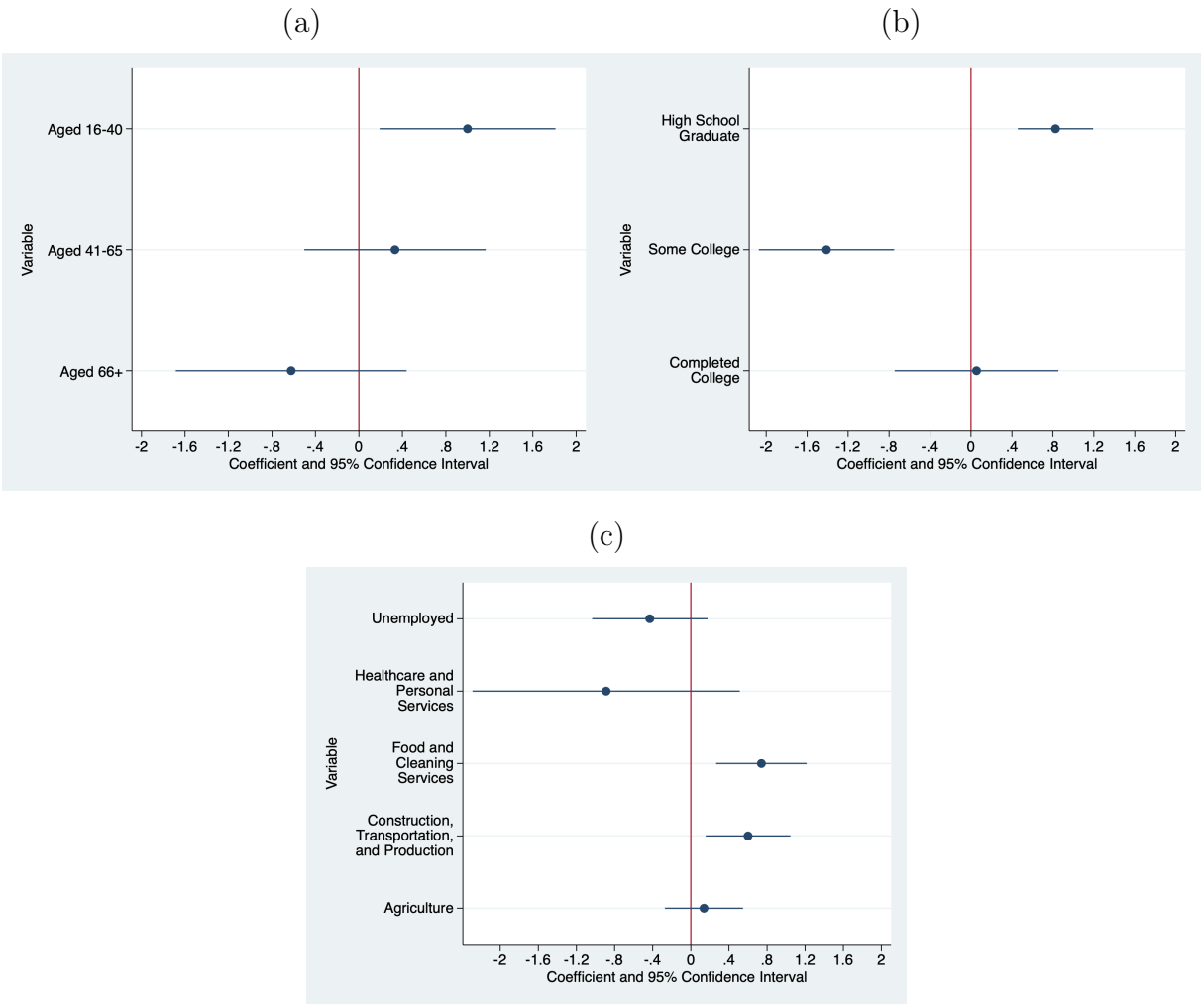
Notes: This figure presents a scatter plot comparing the predicted share of the relevant population that is undocumented based on the Borjas and Cassidy (2019) imputation-based approach (on the vertical axis) and based on Matrícula card counts (on the horizontal axis). Undocumented population shares are constructed using data from 2008-2012. Each county-level observation is weighted by the number of Mexican-born residents in that county (as measured in the ACS).

Figure A2: County-level Relationship Between ACS-based and Matrícula-based Measures (County Imputation)



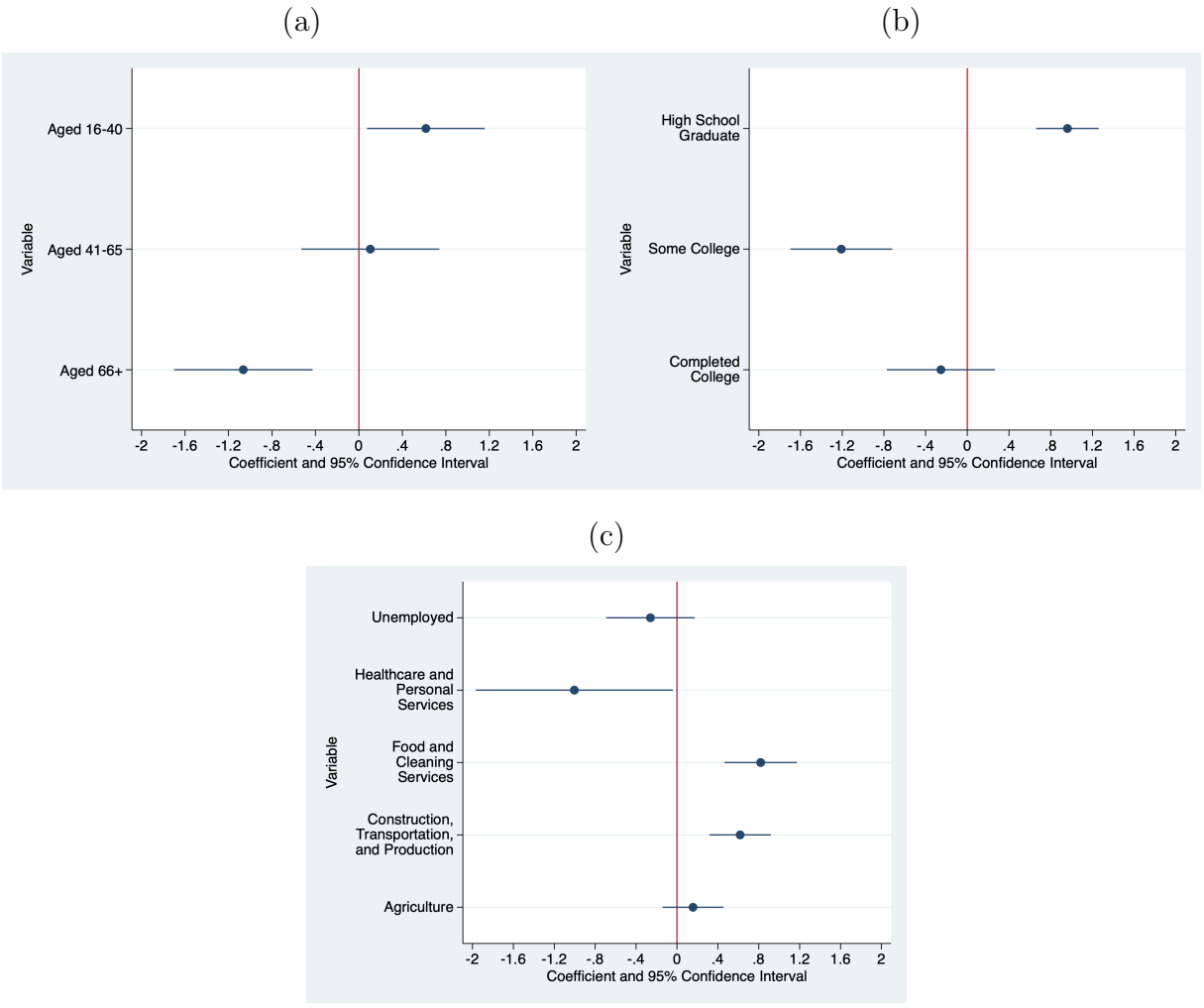
Notes: This figure presents a scatter plot comparing the predicted share of the relevant population that is undocumented based on the Borjas and Cassidy (2019) imputation-based approach (on the vertical axis) and based on Matrícula card counts (on the horizontal axis). Undocumented population shares are constructed using data from 2008-2012. Each county-level observation is weighted by the number of Mexican-born residents in that county (as measured in the ACS). We construct county-level estimates by probabilistically matching ACS PUMAs to counties.

Figure A3: County-level Predictors of the Mexican-Born Undocumented Population Share



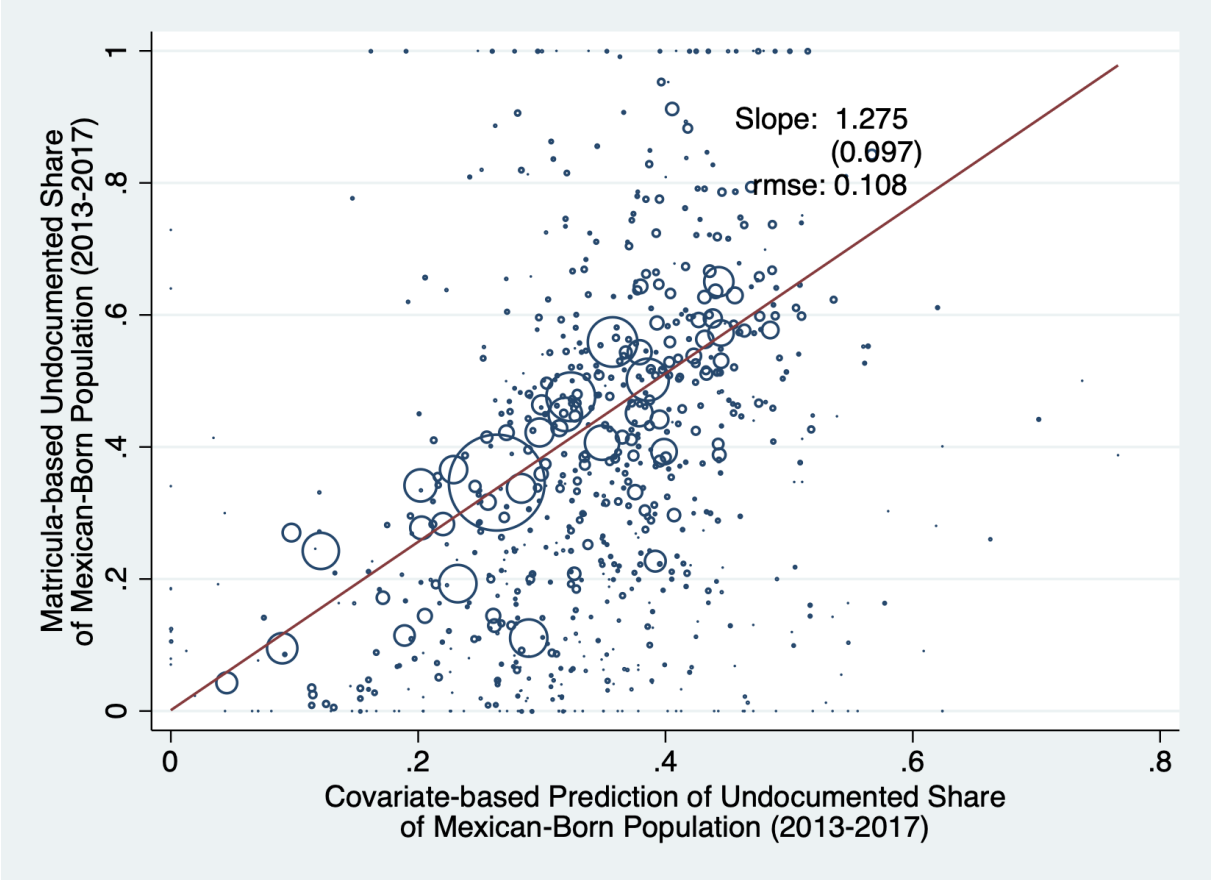
Notes: Each panel plots coefficient estimates and 95% confidence intervals from regressions of the Matrícula-based Mexican-Born Undocumented Population Share on the included covariates using data from 2008-2012. The omitted category from panel (a) is Aged Under 16, the omitted category from panel (b) is High School Dropout, and the omitted category from panel (c) is White-Collar Workers. Each county-level observation is weighted by the number of Mexican-born residents (as measured in the ACS). Confidence intervals are constructed from heteroskedasticity-robust standard errors.

Figure A4: County-level Predictors of the Mexican-Born Undocumented Population Share (County Imputation)



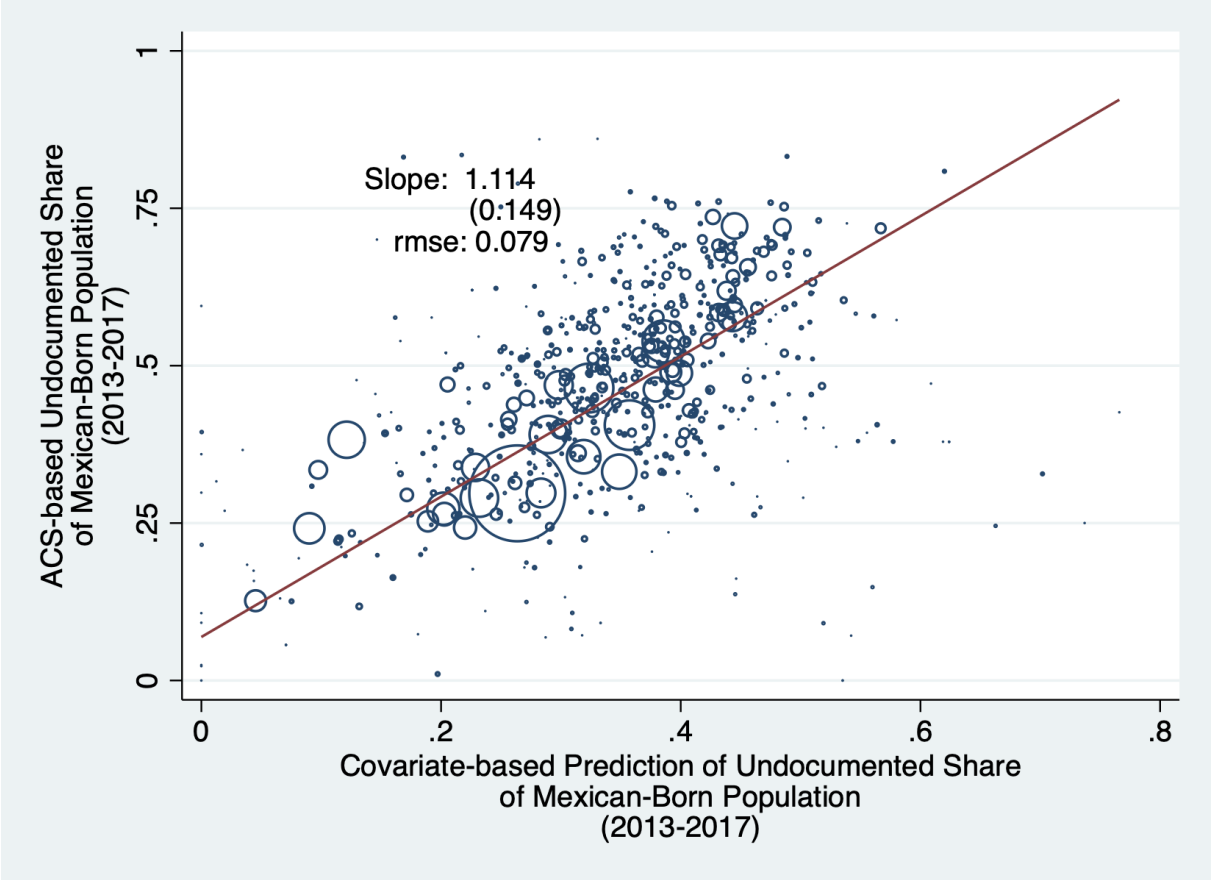
Notes: Each panel plots coefficient estimates and 95% confidence intervals from regressions of the Matrícula-based Mexican-Born Undocumented Population Share on the included covariates using data from 2008-2012. The omitted category from panel (a) is Aged Under 16, the omitted category from panel (b) is High School Dropout, and the omitted category from panel (c) is White-Collar Workers. Each county-level observation is weighted by the number of Mexican-born residents (as measured in the ACS). Confidence intervals are constructed from heteroskedasticity-robust standard errors. We construct county-level estimates by probabilistically matching ACS PUMAs to counties.

Figure A5: Lasso-Based Commuting Zone-level Prediction of Undocumented Mexican-Born Population (Matrícula)



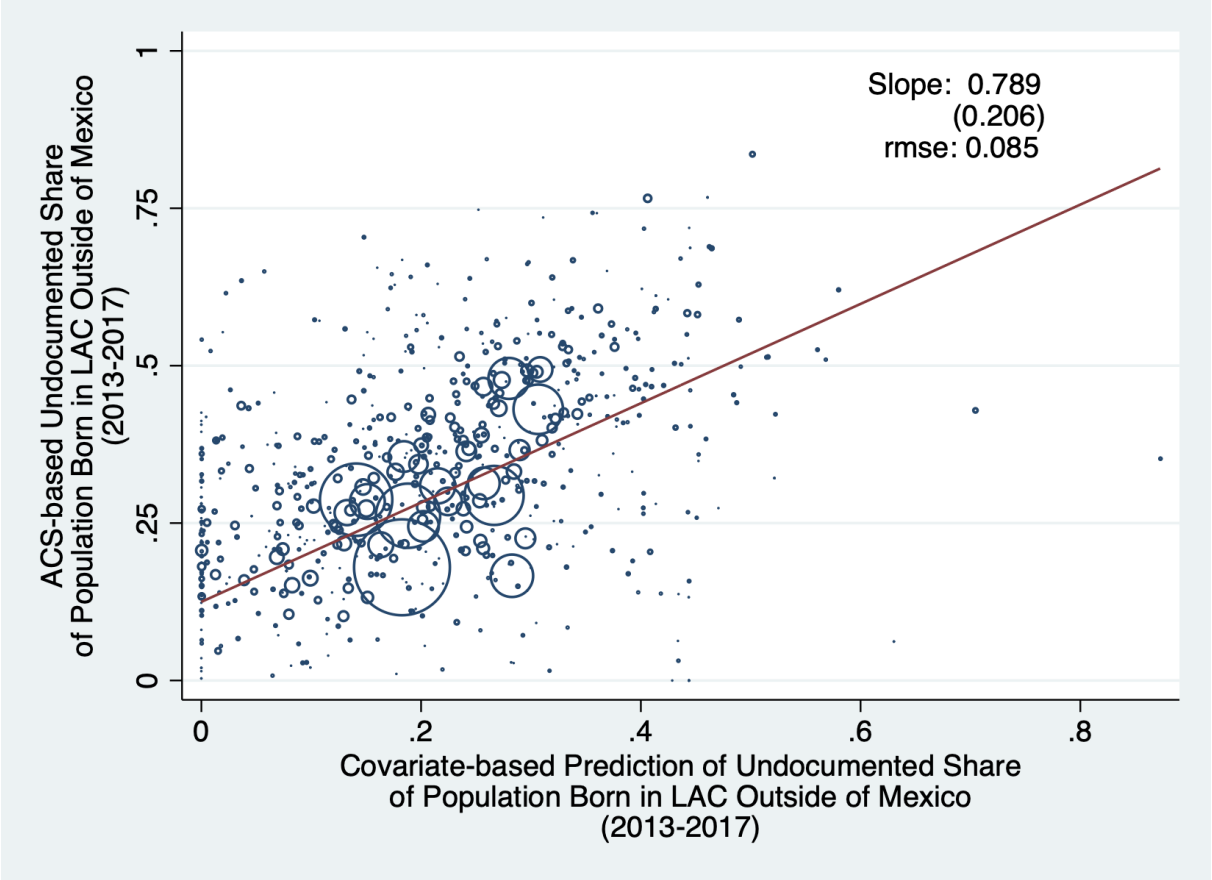
Notes: This figure presents a scatter plot comparing the ratio of the number of valid Matrícula cards to the Mexican-born population at the commuting zone level for the 2013-2017 period (on the vertical axis) to the ratio predicted based on ACS respondent characteristics during this same period and a lasso-based prediction model constructed using 2008-2012 Matrícula and ACS data (on the horizontal axis). Each commuting zone-level observation is weighted by the number of Mexican-born residents (as measured in the ACS).

Figure A6: Lasso-Based Commuting Zone-level Prediction of Undocumented Mexican-Born Population (ACS)



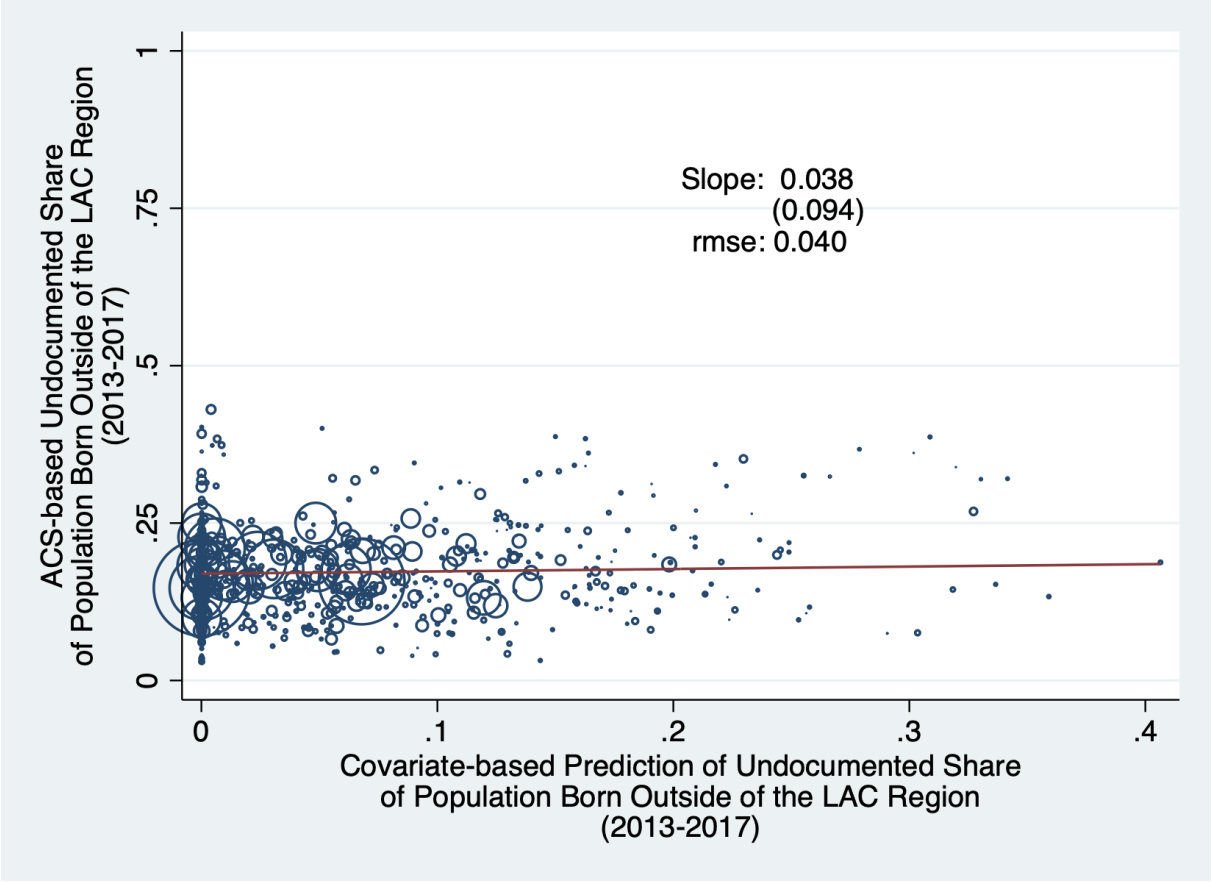
Notes: This figure presents a scatter plot comparing the predicted share of the relevant population that is undocumented based on the Borjas and Cassidy (2019) imputation-based approach (on the vertical axis) to the ratio predicted based on ACS respondent characteristics during this same period and a lasso-based prediction model constructed using 2008-2012 Matrícula and ACS data (on the horizontal axis). Each commuting zone-level observation is weighted by the number of Mexican-born residents (as measured in the ACS).

Figure A7: Lasso-Based Commuting Zone-level Prediction of Undocumented LAC-Born Population (ACS)



Notes: This figure presents a scatter plot comparing the predicted share of the relevant population (those born in the LAC region outside of Mexico) that is undocumented based on the Borjas and Cassidy (2019) imputation-based approach (on the vertical axis) to the ratio predicted based on ACS respondent characteristics during this same period and a lasso-based prediction model constructed using 2008-2012 Matrícula and ACS data (on the horizontal axis). Each commuting zone-level observation is weighted by the number of LAC-born residents (as measured in the ACS and excluding Mexico).

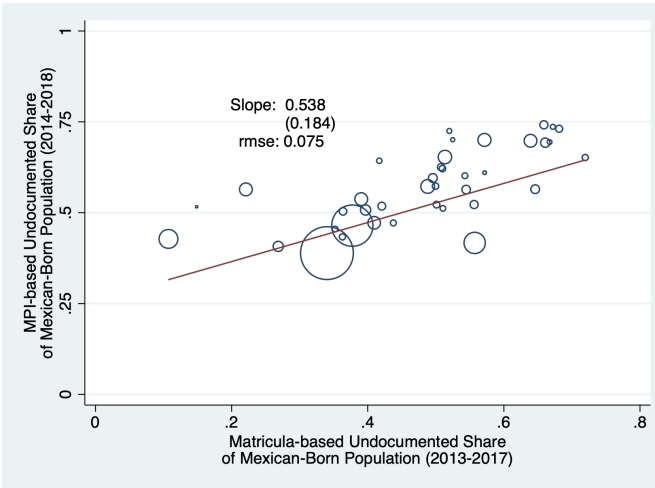
Figure A8: Lasso-Based Commuting Zone-level Prediction of Undocumented Population Born Outside of the LAC Region (ACS)



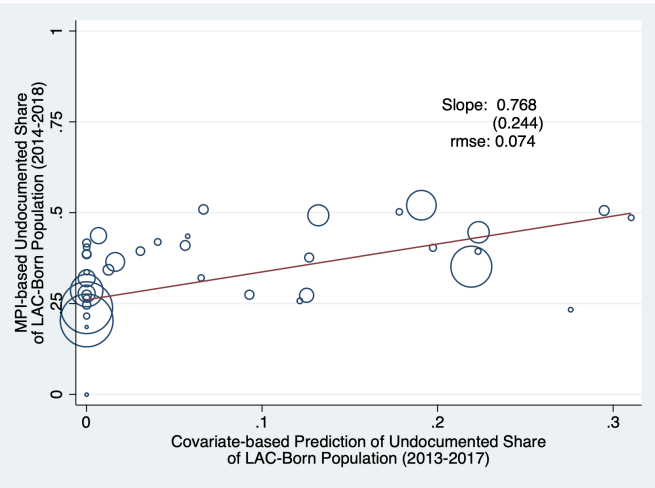
Notes: This figure presents a scatter plot comparing the predicted share of the relevant population (those born outside of the LAC region) that is undocumented based on the Borjas and Cassidy (2019) imputation-based approach (on the vertical axis) to the ratio predicted based on ACS respondent characteristics during this same period and a lasso-based prediction model constructed using 2008-2012 Matrícula and ACS data (on the horizontal axis). Each commuting zone-level observation is weighted by the number of foreign-born residents from outside of the LAC region (as measured in the ACS).

Figure A9: State-level Prediction of Undocumented Population Shares (MPI Comparison)

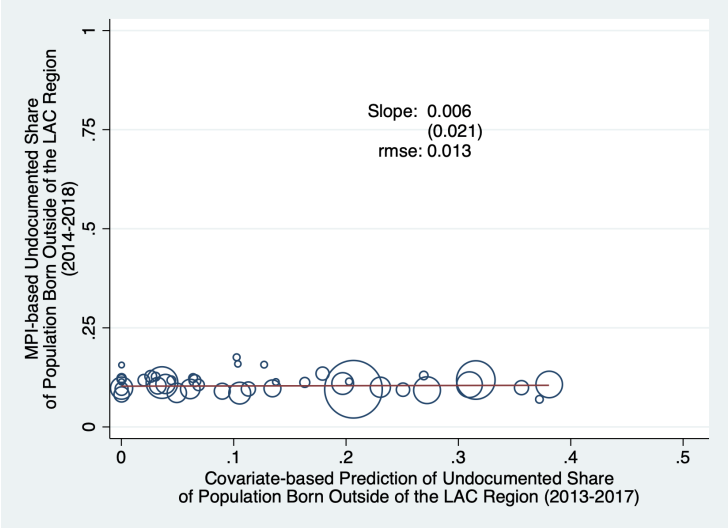
(a) Mexico



(b) LAC Region (Outside of Mexico)

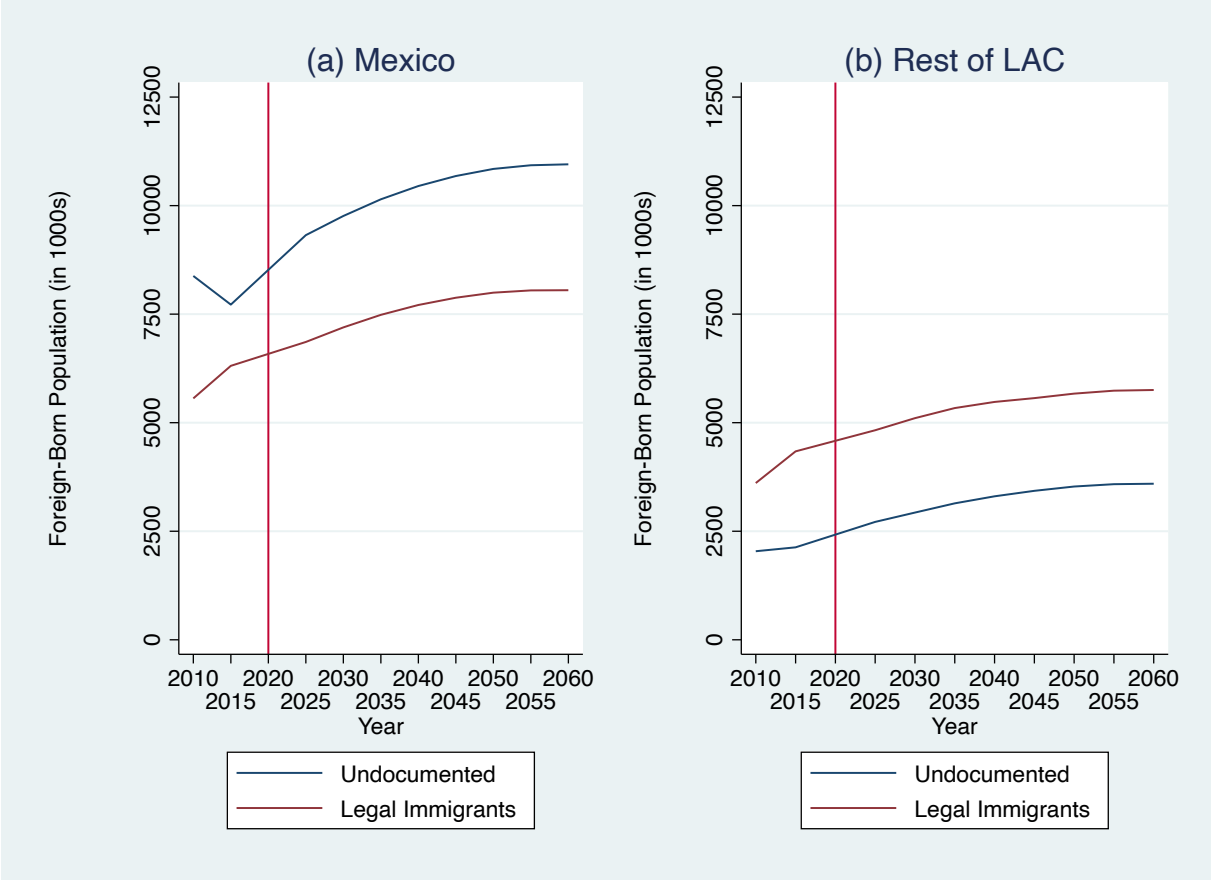


(c) Outside of LAC Region



Notes: This figure presents scatter plots comparing the Migration Policy Institute (MPI)-based measure of the undocumented share of the population born in each country/region at the U.S. state level for the 2014-2018 period to the ratio based on Matrícula card counts (for Mexico) and to the ratio predicted based on ACS respondent characteristics during this same period and a prediction model constructed using 2008-2012 Matrícula and ACS data (for the rest of LAC and non-LAC regions). Each state-level observation is weighted by the number of foreign-born residents from the given country/region (as measured in the ACS).

Figure A10: Historical and Projected Population Estimates for Mexico and Rest of the LAC Region w/o Differential Age-Based Exposure to Legal/Undocumented Immigration



Notes: This figure plots the estimated/projected number of legal and undocumented immigrants separately for Mexico and the rest of the LAC region. For these projections, we set the denominators for both the undocumented and legal net migration rates equal to one half of the total projected population for each origin country.

Appendix B: Projection Tool README File

"Projecting Trends in Undocumented and Legal Immigrant Populations in the United States"
Projection Tool Instructions

Ryan Bhandari
Benjamin Feigenberg
Darren Lubotsky
Eduardo Medina-Cortina

A. Overview

This README file describes the structure of the Projection Folder and explains how to use the "Projection_Tool.xlsx" and "Projection_Tool.do" files to produce figures and tables characterizing projected counts of legal and undocumented immigrant populations by year and country/region of origin. The projection tool requires access to Stata (a statistical software package) and Microsoft Excel.

B. Folder Contents

1. **Projection_Tool.xlsx** – In this Excel file, the user can enter projected future values for four covariates in Columns C-F: (1) *Annual Customs and Border Protection U.S.-Mexico Border Staffing (Number of Agents), (2) Projected Number of Mexican Municipalities with Border Barrier/Fence in Operation, (3) Number of Counties with Secure Communities Agreements in Place, and (4) Measure of Immigration Policy Tightness.* The Excel file provides additional details related to each measure as well as historical reference values. The user can enter anticipated future values, separately by year (for the 2025-2060 period). By default, each measure is set to its historical reference value. As such, by default, the projection tool will return the benchmark projection-based tables and figures included in the manuscript (see Table 5 and Figure 7). As described in Section 6 of the text, we impose the assumption that border staffing, border barrier/fence construction, and Secure Communities agreements affect only undocumented immigration, while immigration policy tightness affects only legal immigration. In Columns G-J, the user can specify additional determinants of legal immigration (in Columns G-H) and undocumented immigration (in Columns I-J). All entered values for the years 2025-2060 should be relative to average values for the 2008-2017 period. By default, these values are set to zero (indicating no anticipated changes in future immigration due to the optionally included measures in Column G-J).

2. **Projection_Tool.do** – After the user has entered preferred values in the Excel file (**Projection_Tool.xlsx**), they will need to run this dofile in order to produce outputted tables and figures. The only adjustment that the user **must** make to this dofile before running is to specify the directory location of the Projection Folder (see the instructions at the top of the dofile regarding the appropriate syntax). In addition, the user can adjust parameter values for the four pre-specified measures by changing covariate-specific coefficients (displayed in lines 11-14 of the dofile). These measures and the sources used to construct predicted associations with net migration flows are summarized in Appendix Table 2 of the text (and discussed in more detail in Section 6). If the user elects to specify additional determinants of legal immigration (in Columns G-H of

Projection_Tool.xlsx) and/or undocumented immigration (in Columns I-J of **Projection_Tool.xlsx**), the user must also specify corresponding coefficient values in lines 17-20 of **Projection_Tool.do.** Each coefficient value reflects the anticipated effect of a one-unit change in the included measure on the net migration rate. By default, coefficient values are set to zero so that changes in the additional measures included by the user in **Projection_Tool.xlsx** do not affect projected estimates of future legal and undocumented immigration levels.

3. **projections_prepped_tool.dta** – This is a Stata data file that is called in when the **Projection_Tool.do** dofile is run by the user and does not require any modification.

4. **country_estimates.xlsx** – Projection Tool results are exported to this pre-formatted Excel file (see below). This file should not be manually modified or deleted.

C. Output

1. **projected_trends.pdf** – This PDF file plots projected trends in legal and undocumented immigration, separately for Mexico and the nine other origin countries in the LAC region with the largest historical number of U.S.-based undocumented immigrants.
2. **country_estimates.xlsx** – This Excel file provides historical and projected legal and undocumented population estimates by LAC region origin country. Historical estimates are annual averages for two five-year periods (2008-2012 and 2013-2017) and projected estimates are provided for 2040 and 2060.