A Model for Developing Estimates of U.S. Citizens Abroad: Final Technical Report

Fors Marsh Group LLC

TABLE OF CONTENTS

Fors Marsh Group, LLC

Submitted to the Federal Voting Assistance Program

July 23, 2013

Introduction

The *Uniformed and Overseas Citizens Absentee Voting Act* (*UOCAVA*) requires that States allow certain eligible citizens—including members of the uniformed services who are absent from their voting jurisdiction due to their service, their family members and dependents, and other U.S. citizens residing outside the United States—to apply to register to vote and vote by absentee ballot in Federal elections. Although the Act applies only to Federal elections, most States and territories also allow *UOCAVA* citizens to use the process for State and local elections. This legal protection is critical in ensuring that *UOCAVA* citizens are able to exercise their right to vote. The Federal Voting Assistance Program (FVAP), under the authority of the Secretary of Defense, is the agency charged with administering *UOCAVA*. More specifically, FVAP's purpose is to

- inform and educate U.S. citizens worldwide of their right to vote
- report on participation rates for uniformed service members and overseas civilians after each Presidential election
- promote efficiency and effectiveness in administering *UOCAVA*

FVAP works closely with the Military Services, the U.S. State Department, and other Federal agencies to accomplish these goals. FVAP also provides information and voting assistance directly to *UOCAVA* voters.

To most effectively support military and overseas voters as they exercise the right to vote, FVAP must know the size and distribution of the *UOCAVA* population. Although the U.S. Department of Defense has up-to-date information on the number and location of military members and their dependents, estimating the number of all U.S. citizens living outside the United States is much more difficult; no official census of this population exists. Developing a method for estimating the population of U.S. citizens abroad will allow FVAP to more effectively and efficiently allocate resources, target its voter assistance outreach to the greatest number of *UOCAVA* voters, identify the degree of success *UOCAVA* citizens experience when voting, and more broadly serve as a foundation for its policy decisions moving forward. Moreover, it is critical that the process of estimating this population be conducted in a scientifically grounded fashion—theoretically driven, transparent, reproducible, and well documented.

Other Federal agencies might also find these estimates valuable in achieving their policy goals, and U.S. business interests may use the estimates to identify overseas U.S. populations as potential markets for their exports. In an increasingly integrated global economy, it is likely, too, that the population of U.S. citizens abroad will become increasingly important in a number of policy areas, and having some information on its size and geographic distribution will facilitate the analysis and implementation of these policies.

However, analyzing the population of U.S. citizens abroad is complicated by the issues that arise from the lack of available data for many (particularly developing) countries, the diverse motivations for U.S. citizens traveling and living overseas, and the economic and institutional environments of many of the countries in which U.S. citizens reside. Consequently, any attempt to estimate this population will face necessary trade-offs between breadth (i.e., the number of countries that can be estimated) and depth (i.e., the accuracy and detail of the estimates that can be made). Currently there are several estimates (varying from 1 million to 7 million) that academics, government organizations, nongovernmental organizations, and private industry use to plan and implement programs targeted to the overseas U.S. citizen population. Unfortunately, some of these estimates have often been accompanied by little documentation, have used varying definitions of the population of U.S. citizens abroad, and seem to have suffered from problematic or unclear methods.

This report describes a research effort that expands upon previous overseas citizen demographic research conducted by FVAP. The result of this effort is a method of estimating the population of U.S. citizens abroad, by country, from 2000 to 2010. This report will (1) review existing estimates and discuss their shortcomings for FVAP's purposes, (2) describe several possible approaches that were considered during this research, (3) detail the methodology ultimately selected for this effort, (4) present the estimates developed from this methodology and discuss trends and noteworthy results, and (5) identify limitations of this effort and next steps for future researchers. It is important to note that this method of estimating the population of U.S. citizens abroad is not intended to replace or supersede other estimates; rather, this model presents an alternative method for estimating the number of U.S. citizens abroad and provides FVAP with additional information about this population.

Past Efforts to Estimate the Overseas U.S. Citizen Population

In the past, a variety of organizations have attempted to develop estimates of either the population of U.S. citizens abroad, specifically, or of migrants worldwide. Different organizations have used different methods to develop their estimates, but these efforts have been hampered by data issues that limit their utility. These limitations fall into two general categories: those resulting from having a limited quantity of data and those resulting from having poor-quality data. Documentation differences are also very prevalent; however, these differences affect the assessment of the various estimates more so than the estimates themselves.

There have been five significant attempts to develop similar estimates of the overseas U.S. citizen population. These efforts provided a substantial starting point for the current work because each effort relied on different data sources and estimation procedures. Unfortunately, none of the past efforts provided an estimate useful for FVAP's purpose and mission. The goal of the current research project was to extend these past efforts and produce a more accurate method for estimating the size and geographic distribution of FVAP's target population. These efforts are described below and include data sources, a brief description of the methodology, and the limitations of these various sources for FVAP's use.

U.S. Census Bureau Estimate

The Census Bureau considered attempting a full enumeration of the population of U.S. citizens abroad for the 2010 Census. In 2004, the Census Bureau conducted a test to determine the feasibility of conducting an overseas census. Several test countries (France, Kuwait, and Mexico) were selected, and questionnaires were distributed through overseas organizations that were thought to have substantial contact with overseas U.S. citizens in those countries. In addition, a marketing firm was employed to promote the questionnaire to overseas U.S. citizens. Despite these efforts, the U.S. Government Accountability Office (GAO; 2004) reported that response rates were low due to the voluntary nature of the survey and difficulty in monitoring overseas partners. The GAO also concluded that the survey would be difficult to scale up across all countries due to country-specific factors such as privacy laws, the lack of address lists for overseas U.S. citizens, the inability to do follow-up interviews, and the lack of Census Bureau overseas offices, which could deal with localized problems in implementation. As a result of this pilot effort, the Census Bureau did not attempt to count overseas U.S. citizens in 2010.

U.S. State Department Estimate

The U.S. State Department produces annual estimates of the number of Americans located overseas. Based on information that has been released publically about these estimates (GAO, 2007), country-level estimates are based on a combination of consulate registrations and an estimate for the U.S. population living in the country who are nonregistrants using country-specific information. Country-level estimates are developed primarily to facilitate preparation for evacuations of U.S. citizens. A more detailed methodology for developing these estimates has not been released publically. According to the GAO (2007), consulates vary in their procedures for estimating the number of U.S. citizens. Given that consulate registrations are likely to represent only a fraction of the U.S. population residing in a country and that the proportion of the U.S. population that registers at the consulate is likely to vary by country, as discussed in the GAO report, the methodology used to estimate the nonregistered part of the population is likely to have a significant effect on the final estimates.

Several factors limit the usefulness of the State Department's estimates to FVAP. Only regional-level data is released publically, so country-level estimates are not generally available. For FVAP, the lack of available country-level data hinders the ability to use State Department estimates for the purpose of determining how resources should be geographically allocated. In addition, the use of the estimates by the State Department to plan for evacuations can result in their estimates including subpopulations that may not be considered long-term residents, or individuals who may not be eligible U.S. voters. This creates challenges for FVAP when it considers using these estimates to establish a measure of voter success because the number of *UOCAVA* citizens who cast their ballots would be compared with a potentially inflated estimate of the number of potential voters.

World Bank Estimate

The World Bank has developed estimates of bilateral migration stocks for all origin-host country pairs for the years 1960, 1970, 1980, 1990, 2000 (Ozden, Parsons, Schiff, & Walmsley, 2011). Data is primarily based on approximately 1,000 decennial censuses and registries, referred to throughout this report as foreign government estimates (FGEs), developed by host country governments. The researchers discuss many of the complications involved in harmonizing reports from different governments with respect to definitions of migrants and origin regions. For the large number of missing values, data is either imputed using a linear trend or is extrapolated using a prior or future decade's migrant composition, in the case that the country missing an observation has data available for other years. When a country has two or fewer observed decades, aggregate migration stocks are taken from the United Nation's Trends in Migration Stocks (total migration stocks by country every five years), and the average of bilateral migration shares from the decades that are available are used to assign portions of the migrant stocks to different origin countries. For countries lacking bilateral data, the total migrant stock is divided among countries using bilateral data from other countries in the region.

More specifically, when a country's number of U.S. residents was missing, that number was imputed based on the share of the total number of immigrants in the country composed of individuals born in the United States in earlier decades. As a result, the World Bank estimates may underestimate the U.S. population in a country if the propensity of U.S. citizens to migrate to that country increased relative to other countries since the last estimate. Further, the World Bank uses estimates based on both registries and censuses and makes no adjustment for the fact that different FGEs can represent either counts of individuals born in the United States or U.S. citizens, and citizen counts do not necessarily include dual citizens. Consequently, estimates may not be comparable across countries. All of these factors limit the usefulness of the World Bank results to FVAP because they may lead to a misallocation of resources across countries and regions, allocating more resources where there is more data, but not necessarily more need.

United Nations Estimate

The United Nations (UN) produces estimates using a methodology similar to that used by the World Bank (UN, 2011), relying on FGEs for countries when available and imputing missing values for missing years. Like the World Bank approach, this methodology could result in estimates lower than the "true" number of U.S. born and U.S. citizens if the propensity for U.S. citizens to migrate to different countries changes over time. The imputation methodology used by the United Nations and World Bank could also result in overestimates of a country's U.S. population if the size of the U.S. population relative to other immigrant communities has declined over time.

As a result of this methodology, the utility of the UN data to FVAP is likely to be limited by the inaccurate picture given of the distribution of the population of U.S. citizens abroad across countries and regions. Also, like the World Bank estimates, the UN estimates are primarily of U.S.-born individuals, rather than U.S. citizens, and therefore these counts may not capture dual citizens very well.

FVAP 2011 OCC

In 2010, FVAP commissioned a research team to conduct exploratory research into developing a method for estimating the population of U.S. citizens abroad. This report, produced in 2011, contained extensive information on potential sources of data on overseas U.S. citizens and effectively catalogues some of the challenges inherent in such an endeavor. The 2011 OCC Report (FVAP, 2011) did produce estimates of U.S. citizens abroad, but the researchers faced challenges in data collection, data analysis, and data interpretation:

Data Collection

The 2011 OCC Report initially identified 35 countries (later expanded to 47) of interest for which data on the resident U.S. citizen population was collected. The selection of the country sample was largely motivated by the size of the U.S. citizen population (minimum 25,000) and the proportion of the total overseas population residing within the countries of interest (approximately 80%). These criteria were derived from prior estimates based on State Department consular births abroad and the number of hits to the Overseas Vote Foundation (OVF) website as reported by Google Analytics, though the exact procedure was not described in detail.

Additional countries were added to the countries of interest based on individual data sources, including data on students abroad. Each of these sources likely only accurately reflects a subset of the overseas U.S. population, and thus may lead to some countries being inaccurately categorized as being “of interest.” Further, the choice to focus only on a specific set of 47 countries prior to data collection makes predicting U.S. citizen populations for countries outside the 47-country sample, based on the subsequent statistical analysis, untenable. Finally, while the report detailed extensive information on various potential data sources, the documentation provided on these sources was limited. This presents challenges in assessing the quality of the data and in replicating these data collection procedures.

Data Analysis

The basic methodology of the 2011 OCC Report was to use FGEs as a proxy for the overseas U.S. citizen population, and to then use country-level variables (population, GDP, etc.) to construct a model of the FGEs. This model was then used to produce an estimate of the U.S. citizen population of countries without an FGE. These estimates were then compared with counts of U.S. citizens based on administrative records, which were taken as the minimum estimate. The highest of the FGEs, the imputations, and the administrative-based minimum count was taken as the final estimate.

The analysis does not account for and discuss biases or the size of the confidence intervals for the country estimates. As a result, interpreting the resulting predictions' accuracy or testing the predictions of the empirical model on out-of-sample countries is challenging. This makes it uncertain whether the model would be as effective in explaining variation in the number of U.S. citizens for countries outside the sample used to initially generate the model. Finally, taking the highest of the administrative-based minimum count, the foreign government estimate, and the imputed foreign government estimate likely introduced bias in the final estimate.

Data Interpretation

While the report provided an estimate of the number of U.S. citizens per country, as an initial exploratory effort it did not include a time frame for these numbers, an estimate of the confidence interval for the estimates, or criteria for when people move in and out of the population. The data used to produce the estimates comes from various years, making it difficult to assign the final country-level and global estimates to any particular point in time. In addition, the estimate from the 35 countries was multiplied to provide an estimated range of all overseas U.S. citizens. The authors noted that they "assume that these (35) countries represent 75% to 90% of the total U.S. citizens living abroad," proportions derived from the sources used to choose the countries of interests. However, no attempt to assess to the degree to which the final estimates correlated with the estimates from the sources used to choose the countries was reported. This is particularly problematic with respect to the imputation of the population outside the focus countries, which assumes that the final estimates vary proportionally with these sources in order to arrive at the final global estimate.

These last issues—the lack of information on the time frame for the estimates and the lack of estimates for out-of-sample countries—especially limit the utility of the estimates to FVAP. The lack of information on the time frame for estimates makes it difficult for FVAP to use the results to determine either the rate of participation of eligible voters or the change in that rate because the estimated size of the *UOCAVA* population cannot be matched with a specific election year. In addition, the lack of estimates for out-of-sample countries means that the relative access to voting resources of a potentially large and growing segment of the overseas population cannot be assessed.

Methodology

After reviewing the existing work in this field (see previous section and References), the research team began investigating various approaches to creating estimates of the population of U.S. citizens abroad. The aim was to refine and enhance the work that had been done in the past, building on previous estimates and techniques, while adding elements that could result in more precise method of estimating U.S. citizens living abroad.

Initial Explorations

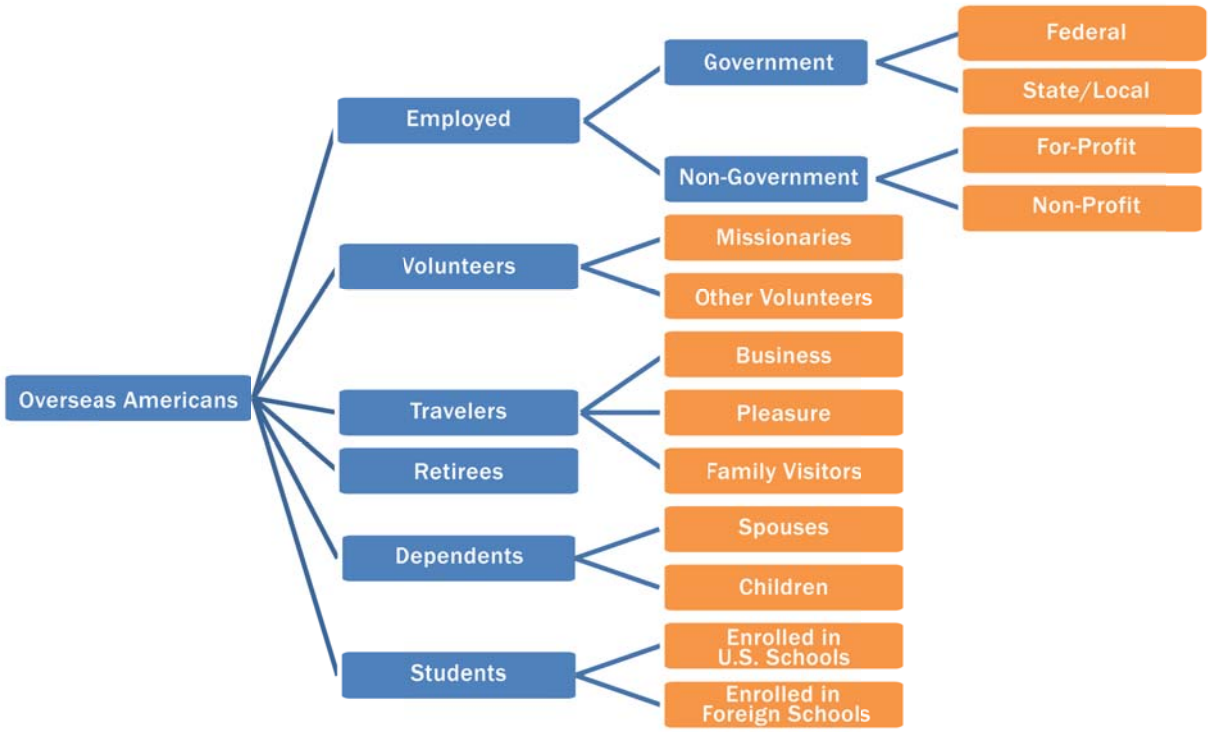
Initially, the research team investigated two overarching methods for addressing the population of U.S. citizens abroad: one that modeled the U.S. citizen population at the level of subpopulations (i.e., number of overseas workers, students, etc.), and one that modeled the population at the country-level aggregate. The team conducted a review of the technical and theoretical strengths, as well as the feasibility, of each method. Estimates produced this way could then be interpreted in light of the assumptions and limitations inherent in the framework. What follows are high-level summaries of the subpopulation method and the country-level aggregate modeling method.

Estimating Subpopulations

As documented by Klekowski von Koppenfels (2013), Americans migrate overseas for a wide variety of various reasons. Many, for example, initially go overseas to temporarily work in either a civilian or military capacity, with some choosing to stay in the host country as a result of marrying a local or to retire. Another example would be children born to immigrants in the United States whose parents then subsequently return to their home country. These are just two examples of situations in which a U.S. citizen could also have citizenship in the foreign country, and these individuals might not be represented in counts of resident U.S. citizens made by foreign governments. Given the diversity in the motives of U.S. citizens, one potentially fruitful approach would be to create separate models for each type of overseas U.S. citizen.

In a subpopulation approach, researchers would develop a taxonomy of the overseas U.S. citizen population such as that illustrated in Figure 1, with every hypothetical overseas U.S. citizen assigned to one of a number of mutually exclusive categories. Estimates of these subpopulations would then be developed for every country through multiple sources (government agencies, surveys of businesses, international organizations, etc.). For countries in which an estimate of a given subpopulation is not available, the value can be imputed through the creation of a model for that specific subpopulation. The subpopulation estimates can then be summed and the total for each country taken as the estimate of the U.S. citizen population of that country.

Figure 1. An Initial Taxonomy of the Overseas U.S. Citizen Population



This method has some significant advantages with regard to data transparency and specificity. One advantage is that there could be transparency in the origin, consistency, and biases in the subpopulation estimates because the source of each one would be identified and described. In addition, because the data is disaggregated, it can more easily be used to make population-specific policy and resource decisions, such as tailoring voter assistance efforts to given subpopulations. For instance, development/aid workers may tend to be located in rural areas, while U.S. government and company employees may tend to be located in larger cities.

However, there are several shortcomings to a subpopulation count approach. First, as a practical matter, the collection of the primary data would be an extremely resource- and time-intensive process. Further, if the subpopulations identified do not encompass the population of U.S. citizens abroad or are correlated with FGEs as a result of being estimated using the same, potentially flawed counting method, then problems in data collection could lead to problems with the analysis and the final estimate. Second, while the subpopulation-based estimates have the benefit of being transparent, unless the taxonomy of the U.S. citizen population is comprehensive and has a minimum of overlap, there is likely to be error in these estimates where certain individuals and groups are counted twice. This was a problem that hampered the 2011 OCC Report (FVAP, 2011), where counts of registered voters likely included substantial numbers of individuals in other subpopulations (such as those employed by Fortune 50 companies).

Capture-Recapture

If the degree of overlap between two subpopulations is known, then a capture-recapture methodology (Bishop, Fienberg, & Holland, 1975) may be used to estimate the remainder of the total population. Rather than attempting to create a complete enumeration of all individuals abroad, administrative record sources can be treated as samples in a capture-recapture approach and used to provide estimates of the population of overseas Americans. These estimates have the advantage of being independent of any FGEs; thus, using this technique can also avoid many of the limitations of FGEs.

Capture-recapture statistical methods are used to estimate the size of a population given samples from that population. In its original form, this technique was used to estimate hard-to-access wildlife populations. Two samples are obtained: a first sample of wildlife is captured, tagged, released, and then a second sample is taken sometime later. The overlap between the two samples is determined and allows for estimation of the size of the population. This two-sample approach requires independence between the two samples. With a larger number of samples, there is more flexibility in the assumption of independence. See Appendix B for the mathematical details.

To estimate the number of overseas U.S. citizens, existing administrative record sources can be used as the input lists or groups. One potential list would be of overseas Social Security recipients, maintained by the Social Security Administration (SSA). Another potential list would be the filers of foreign income (Form 2555) maintained by the IRS. State records of U.S. citizens requesting absentee ballots would be another potential source. Private entity sources of data also exist, maintained by professional organizations, universities, religious groups, etc. By determining the size of these lists and the overlap (through record linkage), population totals can be estimated.

A major limitation of the capture-recapture methodology is that it requires information on the overlap between samples. This would likely require access to detailed, sensitive microdata for record linkage. Absent such data, estimates can be generated under multiple scenarios. However, given the wide range of potential overlap and our current limited information on the overlaps, there will be a large degree of uncertainty in the country-level estimates. These points are discussed in more detail in Appendix B.

Modeling Using a Cross-country Regression

Another method of estimating the overseas U.S. citizen population would be to use a regression-based methodology where the estimate of a country's U.S. citizen population is based on FGEs of the population and how they interact with multiple predictor variables. These variables include those that the research literature on migration has identified as antecedents of bilateral migration stocks (i.e., population size) and flows (i.e., change in population size) as well as counts of particular subpopulations within the country derived from administrative records of U.S. agencies and organizations.

Because it is uncertain how well each variable predicts the size of the overseas U.S. citizen population, and because extraneous variables increase the danger of overfitting to the data, a weighted average of multiple models can be taken. Averaging the estimates from different models mitigates the potential for any individual "wrong" model introducing error in the final estimates, and this approach has been effectively applied to political forecasting, specifically the prediction of violent conflict and election outcomes (Montgomery, Hollenbach, & Ward, 2012). As noted, this approach relies heavily on FGEs, a strategy that has both benefits and drawbacks.

The benefits of using foreign government–produced counts include:

- FGEs are largely representative of the population of interest (U.S. citizens) by the desired unit of analysis (country).
- FGEs are easily acquired from foreign government statistical agencies and are updated on a routine basis.
- Prior studies (World Bank and OECD) have relied on FGEs, establishing precedent, albeit limited, in the research literature.

However, there are also drawbacks to using FGEs. These include:

- FGEs use different instruments by country (census versus registry) that may differ in accuracy.
- Not all censuses and registries are created equal; the quality of the data is directly dependent on the methodology and implementation of the data collection by the individual country. Different countries are likely to have different capacities with respect to data collection (the number and quality of census field workers) as well as the ability of the central statistical office to compile and analyze the collected data.

- The definition of a long-term or permanent resident is likely to vary by country based on individual immigration statutes.
- FGEs will have definitional differences (U.S.-born individuals versus U.S. citizens) and so are not strictly comparable.
- Foreign governments may not include dual citizens in their counts of U.S. citizens, leading to underestimates of the count of U.S. citizens (Ozden et al., 2011; United Nations, 2011).
- Because not all countries develop FGEs, using FGEs to create an estimate model could result in the possibility of having a potentially unrepresentative sample of countries, even after weighting procedures.

Overall, a cross-country statistical analysis using FGEs represents an efficient means of obtaining estimates for all countries for which common cross-country variables are available, thus facilitating a global estimate of overseas U.S. citizens. However, an approach using FGEs to model the overseas U.S. population will need to address the conceptual and methodological shortcomings listed above.

Our Chosen Method

A cross-country-based modeling approach was selected because it utilizes information on the size of overseas U.S. citizen populations already generated by foreign governments, and thus is likely to provide more reliable and accurate way of estimating the overseas U.S. citizen population than the subpopulation approach. This methodology involved developing estimates for the overseas U.S. citizen population following roughly from the 2011 OCC Report (FVAP, 2011) by using FGEs as the best estimate of the "true" U.S. citizen population within a given country. A model was developed using countries that publish estimates of the U.S. citizen population. This model is used to predict the number of U.S. citizens for every country that lacks an FGE as well as adjust the estimates for countries that use alternative definitions of in their estimates of overseas U.S. citizens.

Our basic methodology consists of three steps:

1) Estimate the relationship between counts of U.S. citizens and country characteristics for all countries and years for which FGEs are available.

2) Generate many different models (combinations of predictors) to estimate FGE with the final estimate being an average of these models, weighted by their fit (better-fitting models given a greater weight).

    – Although every predictor is considered in the final estimate, the impact of less-effective predictors (i.e., worse fit) is mitigated by giving those models a smaller weight.

3) Calculate confidence intervals that reflect the range of values that have a high probability of containing the "true" estimate.

   – The confidence intervals measure the uncertainty that results from the limited sample size and the possibility that the results would differ if the sample included different countries.

Our strategy builds upon previous work in several significant ways. First, it uses variation in the size of the U.S. population between countries and differences between countries on relevant characteristics to produce estimates for all countries. Second, it accounts for differences in the FGEs based on how U.S. residents were counted and who was considered a U.S. resident. And finally, it provides confidence intervals, which reflect at least some of the uncertainty in the estimates. The subsequent sections describe the process for (1) identifying and collecting the FGEs, (2) specifying the predictor set, and (3) calibrating the resulting models and weighting their predictions using ensemble Bayesian model averaging (EBMA).

Identifying and Collecting Foreign Government Estimates (FGEs)

FGEs were identified using several different sources of data. The initial estimates were obtained from the OECD International Migration Database, which provided data on the number of U.S. citizens during the years 2000 to 2010 for most OECD countries. Second, estimates were obtained from each of the individual countries or directly from their national statistical agencies. Links for foreign government statistical agencies websites were identified using the U.S. Census Bureau webpage titled "International Collection of the U.S. Census Bureau Library."[1] Estimates obtained from countries' websites were usually from their most recent census. In other cases, estimates were obtained from specific reports on migration commissioned by the national government. These estimates were obtained from foreign government censuses and immigrant registries Third, data were supplemented with an additional set of FGEs available in a U.S. Census Bureau internal document titled "Estimating native emigration from the United States," (Schachter, 2008), which was compiled as part of a project to estimate U.S. net emigration. Although this document included estimates for a period that roughly covered the years 1990 to 2008, only estimates from post-1999 were included (to avoid complexity introduced by the large number of border changes that occurred in the 1990s). In cases where a country has an estimate available for more than one year in the 2000–2010 period of study, each estimate is included in the sample, but the country is weighted

[1]Links to foreign government statistical office websites were retrieved from http://www.census.gov/population/international/links/stat_int.html

based on the inverse of the number of years of data. For example, for countries that have estimates available for two years, each estimate is given half the weight. This should result in a more representative sample and lead to more accurate estimates. Finally, unmodified FGEs for several countries were found in the 2011 OCC Report (FVAP, 2011). For countries without 2010 estimates, but with estimates in 2011, the 2011 estimate was used in place of the 2010 estimate. The following table lists the countries with an FGE by source. Table 1 lists countries for which an FGE was located, and Table 2 lists countries for which an FGE was unable to be identified/collected.

Table 1. Countries with FGEs by Source

2011 OCC Report	Schachter (2008)	OECD	Foreign Government Statistics Offices
Colombia, 2005	Argentina, 2001	Australia, 2000-2010	Albania, 2010
Dominican Republic, 2002	Bahamas, 2000	Austria, 2001-2010	Antigua & Barbuda, 2001
Panama, 2010	Barbados, 2000	Belgium, 2000-2009	Armenia, 2001
Russia, 2002	Belize, 2000	Canada, 2001, 2006	Belarus, 2009
United Kingdom, 2010	Bolivia, 2001	Czech Republic, 2000-2010	Bermuda, 2000
	Brazil, 2000	Denmark, 2000-2006, 2008-2010	Cyprus, 2001
	Chile, 2002	Finland, 2000-2010	Latvia, 2000 and 2010
	Costa Rica, 2000	France, 2006-2008	Lithuania, 2004-2010
	Croatia, 2001	Germany, 2000 – 2010	Mauritius, 2000 and 2010
	Ecuador, 2000	Greece, 2001 and 2010	Micronesia, 2000
	Guatemala, 2002	Hungary, 2000-2010	Peru, 2007
	Guyana, 2002	Italy, 2000-2010	Romania, 2002
	Honduras, 2001	Japan, 2000-2010	Sierra Leone, 2004
	Hong Kong, 2006	Korea, 2000-2010	St. Vincent and the Grenadines, 2001
	Iceland, 2008	Luxembourg, 2001	Tanzania, 2002
	India, 2001	Mexico, 2000 and 2010	Thailand, 2010
	Israel, 2006	Netherlands, 2000-2010	Uruguay, 2010
	Jamaica, 2001	New Zealand, 2001 and 2006	
	Jordan, 2004	Norway, 2000-2010	
	Kiribati, 2005	Poland, 2002 and 2006-2009	
	Malta, 2005	Portugal, 2000-2010	
	Nicaragua, 2005	Slovak Republic, 2001 and 2004-2010	
	Panama, 2000	Spain, 2000-2010	
	Palau, 2000	Sweden, 2000-2010	
	Paraguay, 2002	Switzerland, 2000-2008	
	Philippines, 2000	Turkey, 2000	
	Samoa, 2001		
	Slovenia, 2002		
	South Africa, 2001		
	St. Kitts and Nevis, 2001		
	St. Lucia, 2001		
	Trinidad and Tobago, 2000		
	United Kingdom, 2006		
	Venezuela, 2001		
	Zambia, 2000		

Table 2. Countries without an FGE

Afghanistan	Ghana	Papua New Guinea
Algeria	Grenada	Qatar
Angola	Guinea	Rwanda
Azerbaijan	Guinea-Bissau	Sao Tome and Principe
Bahrain	Haiti	Saudi Arabia
Bangladesh	Indonesia	Senegal
Benin	Iran	Serbia
Bhutan	Iraq	Seychelles
Bosnia and Herzegovina	Kazakhstan	Singapore
Botswana	Kenya	Solomon Islands
Brunei	Kuwait	Somalia
Bulgaria	Kyrgyzstan	Sri Lanka
Burkina Faso	Laos	Sudan
Burundi	Lebanon	Suriname
Cambodia	Lesotho	Swaziland
Cameroon	Liberia	Syria
Cape Verde	Libya	Taiwan
Central African Republic	Macao	Tajikistan
Chad	Macedonia	Timor-Leste
China	Madagascar	Togo
Comoros	Malawi	Tonga
Congo, Dem. Rep.	Malaysia	Tunisia
Congo, Republic of	Maldives	Turkmenistan
Cote d`Ivoire	Mali	Uganda
Cuba	Marshall Islands	Ukraine
Djibouti	Mauritania	United Arab Emirates
Dominica	Moldova	Uzbekistan
Egypt	Mongolia	Vanuatu
El Salvador	Montenegro	Vietnam
Equatorial Guinea	Morocco	Yemen
Eritrea	Mozambique	Zimbabwe
Estonia	Namibia	
Ethiopia	Nepal	
Fiji	Niger	
Gabon	Nigeria	

Specifying the Set of Predictor Variables

One of the primary ways that this method builds upon prior work is by having an explicit justification for the selection of explanatory variables. When variables are selected without this justification, but rather selected purely based on empirical results from a single sample source, it can result in overfitting, especially when working with a small sample size. The model introduced in this report includes a number of theoretically established interaction variables, including distance (Lewer & Van den Berg, 2008), the difference in income per capita (Grogger & Hanson, 2011), and immigrant stocks from the foreign country residing in the United States (Artuc, Docquier, Ozden, & Parsons, 2013). Much of these data are publicly available from sources such as the World Bank (The World Bank Group, 2012) and the Penn World Table Version 7.1 (Heston, Summers, & Aten, 2012).

There are a number of theoretical frameworks for modeling and predicting estimates of the aggregate overseas U.S. citizen population by country that were examined separately as well as in combination. Two specific models of the interaction between the United States and countries that host U.S. citizen populations are (1) a "gravity model" and (2) an immigration–emigration model.

> Gravity Model: Assumes that the flow of U.S. migrants to other countries and the resulting stocks of U.S. citizens in those countries is a function of (a) the size of the country, usually measured in GDP, with countries with larger economies attracting more U.S. migrants; and (b) the distance of the country from the United States, with countries closer in distance attracting more U.S. migrants. This modeling approach has recently been used to impute migration stocks for all country pairs (Artuc, Docquier, Ozden, & Parsons, 2013).

> Immigration–Emigration Model (Warren & Peck, 1980): Assumes that the number of U.S. citizens residing in another country is a function of the number of immigrants residing in the United States from that country, whereby countries that send more immigrants to the United States receive more emigrants from the United States in turn.

These models are not mutually exclusive and can be combined in both a single theoretical and statistical framework.

Each of the variables used to predict the FGE can be placed into one of three categories:

> (1) Administrative: Administrative records–based counts of the number of particular subpopulations of U.S. citizens living in a given country ("count" variables). Variables derived from administrative records directly reflect the size of a subset of the overseas U.S. citizen population of a country. Consequently, an increase in an administrative records–based variable would be expected, on average, to be reflected in an increase in the aggregate FGE.

> (2) Theoretical: Noncount-based variables that have a theoretical relationship with bilateral migration. Theoretical variables have been theoretically and empirically identified as

correlates of bilateral migration stocks and flows for samples including all origin countries for which data is available; however, it is unclear to what degree they are associated with migration by U.S. citizens.

(3) Measurement: Capture differences in how foreign governments estimated or counted their U.S. citizen population. Measurement variables are used to adjust the predictions of the model such that they reflect the size of the population of interest, specifically U.S. citizens. These adjustments require that they be included in every model.

In deriving the estimates, multiple models were tested using a variety of combinations of the three types of variables. Descriptive statistics for the FGEs and predictor variables for observations used to generate the estimates are listed in Table 3.

Table 3. Descriptive Statistics, In-Sample Country-Years

Variable	N	Mean	Standard Deviation	Minimum	Maximum
FGE	272	25151.02	59013.2	41	738103
Measurement Variables					
Citizenship	272	.75	.43	0	1
Dualcitizenship	272	.35	.48	0	1
Dualcitizenship * Citizenship	272	.19	.40	0	1
Registry	272	.69	.46	0	1
Administrative Records Variables					
Social Security Beneficiaries	272	7898.19	13278.69	14.72	102123
IRS Form 2555s	272	4488.27	6305.65	16.64	34213.93
Students	272	4000.73	7031.67	0	34024
Federal Government Employees	272	1290.12	3436.97	0	18232
Theoretical Variables					
Ln(Difference in GDP per capita)	272	-.66	.74	-4.11	.51
Population	272	29630.8	68996.54	46.19	1023295
Distance	272	3696.60	1153.73	3.45	9093.53
Mean (World Governance Indicators)	272	1.05	.68	-1	1.88
Trade	272	3.29E+04	6.25E+04	3.52	5.33E+05
Immigrants in U.S.	272	2.10E+05	5.75E+05	1240	6.40E+06
Military Aid	272	6.26E+09	1.20E+10	0	1.29E+11
English	272	.54	.50	0	1
Spanish	272	.41	.49	0	1
Year of Estimate	272	2004.66	3.29	2000	2010

Administrative Records Variables

Administrative records variables serve as potential indicators of the number of U.S. citizens in a particular subpopulation within a country. Because they can estimate a subset of the population of interest, there is reason to believe that they will help predict the size of the FGE because individuals included in these administrative records should also be counted in the FGE. Consequently, they are included in every model.

- *Number of Social Security Beneficiaries, 2000–2010:* The number of overseas Social Security beneficiaries published by the SSA. Counts were available for each year between

2000 and 2010, aggregated for all regions, but provided individually only for some countries. To create estimates for countries missing individual counts, a Poisson regression imputation model of the number of beneficiaries was developed using the (logged) number of foreign exchange students, the (logged) number of U.S. Federal Government civilian employees, and the additional theoretical variables (see the first column of Table 4) to generate predicted Social Security beneficiaries. As opposed to using the predicted values themselves as an estimate of Social Security beneficiaries for countries without counts, unassigned beneficiaries in a region (those in countries with fewer than 500 beneficiaries) as reported by the SSA were assigned to a country in the region missing a count proportional to the predicted number of beneficiaries.[2]

- *Number of Foreign Earned Income Returns, 2000–2010*: The estimated number of IRS Form 2555 returns, used to declare foreign income, filed by U.S. citizens living in the country in a given year (Hollenbeck & Kahr, 2009). Each form represents at least one U.S. citizen residing in the country. Data was not available for some countries, and for the subset of countries with estimates, they were only available for 1996, 2001, and 2006. To obtain estimates for missing countries and years, the number of returns was first estimated using a Poisson regression imputation model with the theoretical variables discussed below as predictors of the (logged) number of returns. The total number of Form 2555s filed for countries without an estimated number of returns was available by region. Unassigned Form 2555s in each region were assigned to countries without an estimate proportional to their predicted number of returns based on the imputation model. These were used to create estimates for 1996, 2001, and 2006 for all countries. Using these imputed estimates of the number of tax returns, estimates for 2000 and 2002 through -2005 were imputed using linear interpolation. To create estimates for the years 2007 through 2010, an imputation model of (logged) growth in tax returns between 2001 and 2006 was estimated using tax return growth between 1996 and 2001, (logged) number of tax returns in 2001, imputed values for Social Security beneficiaries, students, government employees, and the theoretical variables

[2]The number of Social Security beneficiaries is subject to a natural log transformation for the purpose of regression. Other variables that are logged include the number of foreign earned income returns, the number of U.S. exchange students, the number of civilian government employees, the ratio of GDP per capita of the foreign country to the GDP per capita of the United States (logged difference), foreign country population, distance, trade, the number of immigrants originating in the foreign country in the United States, and military aid. This transformation reduces the leverage of countries with extreme values on these predictors. Generally, when a country has a 0 value on a given predictor, the variable is increased by 1 for each country. This ensures that these predictors remain defined for all countries after the log transformation, and can thus be included in the regression.

for 2001. Using this model, data for 2006 (i.e., growth between 2001 and 2006, initial number of returns in 2006, etc.) was used to predict growth between 2006 and 2011. Using this predicted five-year growth, an estimate of the number of returns in 2011 was created for each country. Linear interpolation was then used between the 2006 estimate and the 2011 estimate to create estimates for 2007 through 2010. See Table 4 for model results.

- *Number of U.S. Exchange Students, 2000–2010*: The total number of U.S. exchange students attending foreign universities for each year in the period 2000–2010 (Institute of International Education, 2012). Countries without an estimate for any year were assigned a value of 0. Estimates for countries with at least two estimates but with missing years were generated using linear interpolation and/or extrapolation.
- *Number of Civilian U.S. Federal Government Employees, 2000–2010*: The number of civilian U.S. Federal Government employees residing in a country in a given year, as reported in data provided to FVAP by the Office of Personnel Management on April 3, 2013.

While additional administrative records such as State Department consulate registrations and Department of Defense counts of the number of military personnel and their dependents could have been included, these data were not publically available due to security considerations. As a result, including this data in the analysis would have precluded outside researchers from reproducing the results and thus undermined the transparency of the analysis. Therefore, these variables were not included in the analysis.

Table 4. IRS and Social Security Imputations

	# SS Beneficiaries	# IRS 2555 Returns (Est.)	Growth in 2555s (2001–2006)
Ln(IRS Returns, 2001)			-.80*** (.30)
Ln(Growth in IRS Returns, 1996-2001)			-.77** (.33)
DUALCITIZENSHIP	.47* (.27)	.54*** (.20)	-.43 (.46)
Ln(# of SS Beneficiaries)			-.30 (.20)
Ln(STUDENTS)	-.14 (.09)		.19 (.12)
Ln(US Government Employment)	.07 (.08)		-.88*** (.15)
Ln(Difference in GDP per capita)	.73 (.52)	.53*** (.20)	.66*** (.23)
Ln(Population)	.32* (.16)	.16 (.14)	.42** (.18)
Ln(Distance)	-.10 (.11)	.22*** (.07)	-.19 (.22)
Mean(World Governance Indicators)	.44 (.38)	-.21 (.20)	-.19 (.39)
Ln(Trade)	-.08 (.18)	.58*** (.10)	.31 (.21)
Ln(Immigrants in US)	.43** (.17)	.04 (.09)	.32* (.19)
Ln(Military Aid)	.05** (.02)	.00 (.01)	.23* (.13)
ENGLISH	.78*** (.30)	.49*** (.16)	-.05 (.33)
SPANISH	.31 (.21)	.07 (.18)	.21 (.45)
Year Effects	YES	YES	N/A
Countries	60	60	182
N	584	164	182
Pseudo R^2	.82	.74	.91

*p < .10. **p < .05. ***p < .01. Model coefficients are estimated using Poisson regression. Robust standard errors clustered by country in parentheses.

Theoretical Variables

What are referred to as "theoretical variables" are those that have been found in the research literature to be associated with higher levels of migration between countries. These studies have typically used comparisons between pairs of many different origin and destination countries to empirically test the effects of these variables on bilateral migration. There may be differences between what drives emigration from the United States and what drives emigration from other countries (as has been found in the empirical literature on international migration) due to the failure of many of these empirical studies to account for the changing propensity of residents of particular origin countries to migrate, or multilateral resistance (Bertoli & Fernandez-Huertas Moraga, 2013). Consequently, these variables may be poor predictors of the number of U.S. citizens in foreign countries and lead to inaccurate final estimates if included in the regression. For this reason, these variables were only included in some regressions, to ascertain whether the inclusion of these variables enhanced or detracted from the ability of the model to predict the FGEs. The weight given to the individual models was adjusted such that models that produced more accurate predictions were given larger weights. Consequently, the influence of these variables on the final estimate was based partly on the degree to which they were actually able to predict the FGEs.

- *The Difference Between Foreign Country GDP Per Capita and U.S. GDP Per Capita*: The difference between the purchasing power parity (PPP)–converted[3] GDP per capita of the foreign country in a given year in constant 2005 prices and the GDP per capita of the United States in the same year, as reported by Penn World Table Version 7.1 (Heston, Summers, & Aten, 2012). The empirical literature on international migration identifies differences in wages between origin and host countries as a primary determinant of bilateral migration flows (i.e., travel and resettling between two countries; Grogger & Hanson, 2011; Mayda, 2010). Consequently, countries that are highly developed relative to the United States, as determined by the difference in GDP per capita, would be expected to be more attractive to U.S. citizens and thus have larger U.S. citizen populations.
- *Population*: The population (in thousands) of the foreign country, as reported in the Penn World Table Version 7.1 (Heston et al., 2012). The empirical literature on international migration has typically found that countries with larger populations/economies tend to attract more migrants (Lewer & Van den Berg, 2008). Consequently, countries with larger populations would be expected to have larger numbers of U.S. citizens.
- *Distance from the United States*: The distance between the closest foreign country–U.S. pair of cities with populations over 750,000. For countries that do not have a city with a population over 750,000, the distance between the capital city of the foreign country and the closest U.S. city with a population of at least 750,000 was used. The latitude and longitude coordinates used to generate the distance measures were obtained from the United Nations' *World Urbanization Prospects, the 2011 Revision*. Distance has typically been found to be associated with lower levels of migration between two countries (Lewer & Van den Berg, 2008), likely because of the fact that more distance is related to higher costs of migration (e.g., owing to travel and moving expenses). Consequently, countries farther away from the United States would be expected to have smaller numbers of U.S. citizens.

[3]The U.S. dollar value of GDP per capita without a PPP adjustment is a problematic proxy for a country's level of development because it does not reflect differences in prices across countries. By contrast, PPP-converted GDP attempts to represent the actual amount of goods and services that the country's residents can obtain given their income.

- *Trade with the United States*: The mean end-of-year product trade (imports + exports) between the United States and the foreign country for the years in which data are available during the years 2000–2013 as reported by the Census Bureau.[4] Trade has been both theoretically and empirically linked to migration between trading countries (Felbermayr & Toubal, 2012; Sangita, 2013). Consequently, countries with higher levels of trade with the United States would be expected to have larger numbers of U.S. citizens.
- *Institutional Quality*: The average of the six World Bank's World Governance Indicators (Voice and Accountability, Political Stability and Absence of Violence, Government Effectiveness, Regulatory Quality, Rule of Law, and Control of Corruption) averaged across the years 1996–2011. Institutional quality, and particularly the degree of political stability, has been found to be a determinant of net migration to countries (Ziesemer, 2010). Consequently, countries with good institutional quality would be expected to have higher numbers of U.S. citizens.
- *Number of Immigrants in the United States*: The number of immigrants from the foreign country ages 25 and up in the United States in the year 2000 as reported by Artuc et al. (2013). One type of potential out-migrant from the United States is an immigrant from a foreign country (or his or her offspring) who then decides to return to his or her country of origin (Scheuren, 2012). A more general justification for the inclusion of this variable is that it may proxy for factors that promote or inhibit migration both to and from the United States, such as transportation costs. Consequently, countries with larger numbers of immigrants in the United States would be expected to have larger numbers of U.S. citizens. On the other hand, the number of immigrants in the United States from the country may also be negatively associated with the number of U.S. citizens in that country, if factors that affect migration flows asymmetrically (such as political instability) are salient. It is worth noting that the uncertainty regarding relationship direction is not a limitation for this predictor because the estimation strategy does not require an assumption of a positive or negative relationship.
- *U.S. Military Aid*: The total amount of military assistance in constant dollars made by the United States to the foreign country between 1946 and 2011 as reported by USAID. Aid to foreign countries by the U.S. Government, and the associated interaction between those governments, may promote migration from the United States to the foreign beneficiary countries by facilitating the transfer of information about the foreign country to potential U.S. migrants (Berthelemy, Beuran, & Maurel, 2009). In addition, aid may be a proxy for general diplomatic ties (Alesina & Dollar, 2000) that may be associated with foreign government

[4]Census Bureau trade data was retrieved from http://www.census.gov/foreign-trade/balance/

policies that are advantageous to U.S. migrants, leading to increased U.S. migration to the country. Since development aid is likely to be inversely correlated to the level of development, the effect of such aid on the number of U.S. migrants is ambiguous and may not be predictive of migration and the U.S. population overseas (Fleck & Kilby, 2010). Consequently, military aid, which should be a stronger proxy for strategic interests and diplomatic ties, is used here (Fleck & Kilby, 2010).

- *English or Spanish:* These variables indicate whether English or Spanish is spoken in the foreign country, respectively. The information is taken from *Ethnologue: Languages of the World* (Lewis, Grimes, Simons, & Huttar, 2009). These variables may proxy for cultural distance between the United States and the foreign country as well as the ability to succeed in the host country's labor market (Adsera & Pytlikova, 2012). Given that English and Spanish are the two most widely spoken languages in the United States, countries where these languages are commonly spoken would be expected to attract more U.S. citizens.
- *The Year to Which the FGE Applies*: This variable is included to control for unobserved trends in the size of the overseas U.S. citizen population common to all countries. These factors may include population growth through births of U.S. citizens, whether overseas or within the United States, which would be expected to affect the total number of overseas U.S. citizens. In addition, this variable may also capture changes in transportation costs over the 2000–2010 period of study, which would also be expected to affect the tendency of U.S. citizens to migrate.

Measurement Variables

One issue with using the FGEs as a proxy for the true overseas U.S. citizen population is that the specific population of overseas U.S. citizens being counted by each country is likely to vary (Artuc et al., 2013; Ozden et al., 2011). These differences may be due to an intentional decision on the part of the foreign government to only count a specific part of the U.S. population, such as U.S. citizens versus those who are U.S. born, or single citizenship versus dual citizenship. Alternatively, the differences could represent unintentional error resulting from the method used to count the U.S. citizen population, such as a registry versus census estimates (Ozden et al., 2011). Consequently, it is difficult to interpret what an estimate for a specific country represents, other than whom the government is willing or able to count. If the policy/methods applied by a significant number of foreign governments result in systematic differences in estimates, overall overseas U.S. citizen population estimates could be consistently biased.

Any approach that uses FGEs as part of its model will need to address the error that is inevitably present in these estimates. The potential for measurement error can be addressed in two ways. The first way involves splitting the sample of countries with FGEs based on whether the estimate counts U.S. citizens versus non-U.S. citizens and uses a registry versus a census. If, for instance, estimates derived from a registry that counts the number of U.S. citizens (including dual citizens) most accurately represents the population of interest, the sample used can be restricted to build the model to those countries that meet these criteria. Such an approach suffers, however, from the problem of small sample size. Only four countries (i.e., Austria, Germany, the Netherlands, and Norway) meet the above criteria—too few to construct the models and likely even less representative of the global sample of countries.

A second way of addressing this issue would be to explicitly model the differences in the country measurement instruments. This approach is common to meta-analysis (e.g., Card & Krueger, 1995) and can be incorporated in the regression-based gravity and immigration models.

For instance, in the following model:

$$Ln(USPOP) = \beta X + \gamma M + e$$

Where *USPOP* is the foreign government estimate of the U.S. citizen population, *X* is a vector of structural variables that explain variations in the "true" U.S. citizen population of the country (gravity, immigration to the United States, etc.) and *M* is a series of variables that capture differences in the definition of the U.S. citizen population and the methods used to estimate it. Three variables could be used to estimate the conditional difference in *USPOP*: (1) whether a country uses a census or a registry, (2) whether a country counts citizens versus U.S. born, and (3) whether a country allows or does not allow dual citizenship with the United States. These variables are not thought to have an effect on the "true" number of U.S. citizens in the country, but only affect the FGE. Including these variables in the regressions provides an estimate of the differences between the population as estimated by the FGE and the population of interest. Explicitly including these confounding variables in the prediction models of FGEs will ultimately allow for generation of estimates that mitigate these biasing effects and are thus more accurate representations of the "true" count of U.S. citizens living in foreign countries per FVAP's objectives.

- *FGE Based on a Registry*: A variable indicating if the FGE was generated using the government's administrative-based records. The primary difference between census and registry is that census data is drawn from a single source whereas registry data is drawn from a number of sources (e.g., tax forms, visas, school records, etc.; Ewing, 1998; Punch, 2001). Utilizing data from multiple sources is beneficial in that it may allow for more complete coverage of overseas U.S. citizens (because a citizen is unlikely to be “missed” by several different sources). However, one major disadvantage of registries is that data quality is completely dependent upon the quality of the administrative records on which the data are based (United Nations, 1969), and when attempting to enumerate overseas citizens, registries can be particularly problematic. One of the major problems is that migrants who have registered with a host country often do not de-register upon leaving—thus resulting in an overcount of overseas citizens (Dumont & Lemaître, 2005). A census conducted in a country may have a longer tradition, broader usage, and may be able to capture more data elements by asking multiple questions about citizenship, birth country, dual citizenship, and employment.

Relatively few nations currently use a population registry. Although a number of countries are transitioning to a population registry (Singapore Department of Statistics, 1999; Statistics Netherlands, 2012) or are considering transitioning to a register-based system, most countries, including those in the sample, still use the traditional census. See Table 5 for information on which countries use a census.

Table 5. Countries with FGEs Based on Census

Albania	Croatia	Kiribati	Sierra Leone
Antigua and Barbuda	Cyprus	Lithuania	Slovak Republic
Argentina	Czech Republic	Luxembourg	Slovenia
Armenia	Ecuador	Malta	South Africa
Australia	Finland	Mauritius	South Korea
Bahamas	France	Mexico	St. Kitts & Nevis
Barbados	Greece	Micronesia	St. Lucia
Belarus	Guatemala	New Zealand	St. Vincent & Grenadines
Belgium	Guyana	Nicaragua	Taiwan
Belize	Honduras	Panama	Tanzania
Bermuda	Hong Kong	Paraguay	Thailand
Bolivia	Hungary	Peru	Trinidad and Tobago
Brazil	India	Philippines	Turkey
Canada	Ireland	Poland	Uganda
Chile	Italy	Portugal	United Kingdom
China	Jamaica	Romania	Uruguay
Colombia	Japan	Russia	Venezuela
Costa Rica	Jordan	Samoa	Zambia

Countries with registries (i.e., Austria, Germany, the Netherlands, and Norway) and those without appear to differ with respect to factors that influence the size of their U.S. citizen populations. For example, nations with registries tend to be well-developed and European, both of which are characteristics that attract U.S. citizens (Wennersten, 2008). Consequently, any simple calculation of the mean difference in the FGE between registry and nonregistry countries cannot be interpreted as systematic "measurement" difference between a census and a registry, but may be due to real differences in the size of the U.S. citizen population. This indicator variable is therefore included to account for this possibility and to adjust the predictions so they represent what the model would generate if the FGE had been constructed using a government census, while controlling for the other country characteristics. Data on whether a government used a registry or census was obtained from the 2011 OCC Report, the U.S. Census Bureau internal document titled "Estimating native emigration from the United States," (Schachter, 2008), and websites of individual foreign government statistical agencies or through phone calls to those agencies.

- *FGE Counts of U.S. Citizens*: A variable indicating if the FGE was a count of U.S. citizens as opposed to U.S.-born individuals was included to focus on the number of overseas U.S.

citizens who can potentially vote. Therefore, the estimate should exclude U.S.-born individuals who migrated overseas and who, for whatever reason, are no longer U.S. citizens with the right to vote in U.S. elections. Including this variable also accounts for the potential underestimation that could result from children born to overseas U.S. citizens being excluded from an FGE that only includes U.S.-born individuals. Data on whether a government counted only U.S. citizens (rather than U.S.-born individuals) was obtained from the 2011 OCC Report, the Census Bureau data set (Schachter, 2008), and websites of individual foreign government statistical agencies.

- *Country Allows Dual Citizenship with the United States*: a variable indicating whether a foreign country generally allows its citizens to also have U.S. citizenship after they have migrated to the United States.[5] This variable acts as a proxy for a foreign government's attitude toward dual citizens. FGEs taken from countries that allow dual citizenship may undercount the number of resident U.S. citizens because dual citizens may be treated as citizens of their host country rather than as U.S. citizens. Including an indicator of whether a country allows dual citizenship with the United States allows for the potential mitigation of this source of error (see Appendix C for more information).

The definition of the U.S. citizen population also remains an issue in this study. For the purposes of this project, individual host country governments define what constitutes a resident U.S. population, using the number of long-term residents rather than the total number of U.S. born/citizens when such a subpopulation is enumerated. It should be noted that even what constitutes a resident typically varies by country. These definitional issues should be kept in mind in interpreting the final results of the analysis.

Calibrating and Weighting Models Using Ensemble Model Averaging (EMA)

Estimating the overseas U.S. citizen population is complicated by uncertainty about which predictors should be used to model this population. To address this uncertainty, a variant of a method called ensemble Bayesian model averaging (EBMA) was used, which has been found to yield more accurate out-of-sample predictions than using a single model in applications such as armed conflict prediction and forecasting the outcome of presidential campaigns (Montgomery et al., 2012). The general approach of EBMA is to take predictions from multiple models (i.e., ensembles) and create an

[5]Information on whether a country allows dual citizenship with the United States was obtained from immihelp, a website that provides information to recent immigrants to the United States concerning green cards, visas, and other necessary documents. Retrieved from http://www.immihelp.com/citizenship/dual-citizenship-recognize-countries.html

average of all the estimates weighted by the model's fit to the data in combination with each model's correlation or redundancy with predictions derived from other models. The resulting estimate is designed to be more accurate than the estimates derived from any single model by minimizing the effects of overfitting the data resulting from individual model specifications. At the same time, this method allows the final estimate to incorporate as much information as possible from the predictor variables. The model space from which this average prediction is derived takes the form of all possible combinations of predictor variables. For *k* predictors, the number of models, *N*, equals 2^ (*k*) (including the model with no theoretical predictors, as described above). As applied to the estimation of overseas U.S. citizens, the approach is not likelihood-based (instead, it is based on root mean square error; see below) and, therefore, is not Bayesian (See Appendix C for an analysis of merits and drawbacks of using likelihood-based weights). Consequently, the modeling approach is simply ensemble model averaging (EMA).

The *N* models take the form:

$$\mathrm{FGE}_{it}^{m} = \beta \mathrm{C}_{it} + \beta \mathrm{X}_{it}^{m} + \gamma 1\mathrm{REGISTRY}_{it} + \gamma 2\mathrm{CITIZEN}_{it} + \gamma 3\mathrm{DUAL}_{it} + \gamma 4(\mathrm{DUAL}_{it} * \mathrm{CITIZEN}_{it}) + \mathrm{e}_{it}^{m}$$

Where *FGE* is the foreign government estimate of the size of the U.S. citizen population in country *i* in year *t*; *C* is a vector of variables common to every model that are believed to determine the size of the U.S. citizen population; *X* is a vector of predictor variables that are likely to explain variations in the U.S. citizen population of country *i* included in model *m* (and thus will vary from model to model); *REGISTRY* is a dummy variable that takes a value of 1 if the country's *FGE* is based on a registry count; *CITIZEN* is a dummy variable that takes a value of 1 if the *FGE* pertains to the number of U.S. citizens in the country, and 0 otherwise; *DUAL* is a dummy variable that takes a value of 1 if the country allows dual citizenship with the United States; *DUAL * CITIZEN* is an interaction variable that takes a value of 1 if the country both allows dual citizenship and has an *FGE* that counts U.S citizens, and 0 otherwise; and *e* is an error term. Because the FGE is bounded at 0, each model was estimated using the Poisson Pseudo-Maximum Likelihood Estimator, following Santos Silva and Tenreyro (2006).

The measurement variables (i.e., those not included in vectors *C* or *X*) are included to control for differences in how FGEs estimated their U.S. population and whom they decided to count. For the purposes of generating predictions, *REGISTRY* is assumed to equal 0*, CITIZEN* is assumed to be equal to 1, and (*DUAL * CITIZEN*) is assumed to be equal to 0 for all countries. The constraints applied to *REGISTRY*, *CITIZEN*, and the *DUAL*CITIZEN* product were applied to make the final predictions more comparable with respect to the population they represent. To be specific, a count of

U.S. citizens (i.e., *CITIZEN* = 1) is enumerated using a census (*REGISTRY* = 0). However, this count should also seek to include individuals whom foreign governments of countries that allow dual citizenship might count as their own citizens. Consequently, the goal is to estimate the difference in the count of overseas U.S. citizens between countries that both allow dual citizenship and count the number of U.S. citizens and countries that do not meet one or both of these conditions. Specifically, predictions are generated under the assumption that no country meets both of these conditions (i.e., *DUAL***CITIZEN* = 0) as it is under such circumstances one is most likely to encounter citizenship misclassification and thus inaccurate citizen counts. In other words, citizenship-based FGEs for countries that allow dual citizenship are adjusted such that the prediction incorporates dual citizens.

Although this adjustment incorporates dual citizens in citizenship-based counts, and predictions between countries that allow dual citizenship with the United States and those that do not may still differ, the size of the difference does not depend on whether the FGE counts citizens or U.S. born. Allowing predictions to vary with *DUAL* is important in the present circumstance because whether a country allows dual citizenship with the United States may have an effect on the size of the U.S. citizen population given that the prospect of gaining citizenship in the host country while retaining U.S. citizenship may encourage immigration to that country. In addition, *DUAL* may proxy for unobserved policies that encourage U.S. citizen migration as well as historical connections with the United States. Many countries encourage dual citizenship as a way to promote continued engagement with their expatriate populations (Lafleur, 2012). These policies may therefore promote return migration, reflected in a larger FGE.

Mitigating Selection Bias

To account for the selection bias that may result from countries with FGEs being different in ways that may also affect the size of their overseas U.S. population, each country is given a weight for the purpose of model estimation:

$$\alpha_i = \frac{1}{\mathrm{Pr(FGE)}_i * n_i}$$

Where *Pr(FGE)* is the predicted probability that a country has an FGE during the years 2000 through 2010 based on its observable characteristics and *n* is the number of years for which country *i* has an FGE. The predicted probability of having an FGE is generated using a logit regression where the sample is all countries for which predictions are made. Predictor variables include all variables in vectors *C* and *X* in the estimation equation along with U.S. State Department region dummy variables. Data for the predictor variables for this selection equation were obtained for the year 2000. The results of the logit regression are displayed in Table 6. The result of the weighting is that countries with FGEs that have a low probability of having an estimate (based on the selection bias equation) will have more weight when generating model parameters and predictions, resulting in more accurate EMA predictions for countries without estimates and more accurate parameter estimates than those that would be generated in an unweighted model. This mitigates selection bias when there is not an unobserved factor (i.e., one not included in the model) that affects both the size of the FGE and whether a country has an FGE (Wooldridge, 2002). Including the *n* in the denominator of the weight accounts for the overrepresentation of some countries in the sample because of their having FGEs for multiple years.

Table 6. Determinants of a Country having at least one FGE for the period 2000-2010.

	Pr (1 = Country has estimate, 0 = Country does not have estimates)
DUALCITIZENSHIP	.16** (.15)
Ln(# of Social Security Beneficiaries)	1.79 (.66)
Ln(# of IRS Returns)	3.33** (1.86)
Ln(STUDENTS)	.91 (.22)
Ln(US Government Employment)	.95 (.32)
Ln(Difference in GDP per capita)	.15*** (.10)
Ln(Population)	.79 (.28)
Ln(Distance)	1.67 (.77)
Mean(World Governance Indicators)	18.89*** (17.14)
Ln(Trade)	.64 (.21)
Ln(Immigrants in US)	1.40 (.41)
Ln(Military Aid)	.91 (.06)
ENGLISH	1.75 (1.37)
SPANISH	11.30** (12.46)
Western Hemisphere	20.96** (28.01)
South/Central Asia	.61 (.69)
Near East	1.24 (2.04)
Europe	16.26** (18.31)
East Asia/Pacific	.74 (.83)
N	182
Adj. R^2	.63

**p* < .10. ***p* < .05. ****p* < .01. Model estimated using a logit regression. Odds ratios reported. Robust standard errors reported in parentheses. All predictors are from the year 2000. The reference region is Africa.

The final estimate of the overseas U.S. citizen population for country *i* in year *t* is:

$$\exp(\mathrm{P}_{\mathrm{it}}) = \exp(\sum_{\mathrm{m}=1}^{\mathrm{N}} \mathrm{w}^{\mathrm{m}} \mathrm{P}_{\mathrm{it}}^{\mathrm{m}})$$

Or the average of all predictions for the country across *N* models, weighted by model validation metric *w*. The sampling variance of P_{it} (i.e., the square of the standard error of the population estimate) is estimated by:

$$\mathrm{Var}(\mathrm{P}_{\mathrm{it}}) = \sum_{\mathrm{m}=1}^{\mathrm{N}} (\mathrm{w}^{\mathrm{m}})^2 \mathrm{Var}(\mathrm{P}_{\mathrm{it}}^{\mathrm{m}}) + 2 \sum_{\mathrm{m}=1}^{\mathrm{N}} \sum_{\mathrm{j}=1}^{\mathrm{N}-1} \mathrm{w}^{\mathrm{m}} \mathrm{w}^{\mathrm{j}} \mathrm{Cov}(\mathrm{P}_{\mathrm{i}}^{\mathrm{m}}, \mathrm{P}_{\mathrm{i}}^{\mathrm{j}})$$

Thus, to obtain 95% confidence intervals[6] for country *i* in year *t*, take:

$$\exp(\mathrm{P_{it}} \pm \left(1.96 * \sqrt{\mathrm{Var(P_{it})}}\right))$$

The model validation metric *w* can be expressed in reduced form as:

$$\mathrm{w^m} = \frac{\mathrm{f^m} * \mathrm{c^m}}{\sum_{\mathrm{m}=1}^{\mathrm{N}} \mathrm{f^m} * \mathrm{c^m}}$$

Where f^m is the component of the metric that indicates how well model *m* fit the data. f^m can be written as:

$$\mathrm{f^m} = \frac{(\frac{1}{\mathrm{MSE^m}})}{\sum_{\mathrm{m}=1}^{\mathrm{N}} (\frac{1}{\mathrm{MSE^m}})}$$

Where the MSE is the mean squared error. The MSE is determined through *K*-fold cross-validation (Stone, 1977), where each observation in the sample is randomly assigned to one of *K* subsamples, the model is estimated using the *K* – 1 subsamples, predictions are estimated for the excluded validation sample, and the MSE (weighted by the selection bias weight α_i, from above) is generated for that subsample. The cross-validation procedure is repeated *K* times, with each subsample acting as the validation sample in turn. The cross-validation step is then repeated *S* times, with the average of the *S* * *K* MSEs used as the model MSE. In this application, it set *K* = 5 and *S* = 10. Each model's contribution to the final estimate is therefore determined by its out-of-sample predictive ability, minimizing overfitting that could result from determining model performance based on in-sample fit only. Testing the model using countries that were not used to build the model allows for a more robust test of the model as its predictive power is more likely due to variation in the U.S. citizen populations in these countries and not random measurement error (Hawkins, 2004; Ward, Greenhill, & Bakke, 2010).

The other component of the model validation metric, $\mathrm{c^m}$, captures the degree to which the predictions generated by a model are correlated with predictions generated by other models. Specifically:

---

[6] It should be noted that these confidence intervals only incorporate uncertainty related to sampling variability, and not uncertainty related to issues of data quality, particularly for imputed variables, as well as assumptions related to the "ideal" set of measurement variables values, specifically the relative accuracy and registry versus census. Consequently, the "true" confidence intervals are likely to be wider. One objective of future research would be to obtain some sense of the reliability of different FGEs.

$$c^m = \frac{1/\sum_{j=1}^{N-1} \mathrm{Corr}\left(P^m, P^j\right)}{\sum_{m=1}^{N}(1/\sum_{j=1}^{N-1} \mathrm{Corr}(P^m, P^j))}$$

Corr is the correlation coefficient between models *m* and *j*. In other words, c^m is larger when a model is relatively uncorrelated with other models. The model validation metric w^m is larger when models simultaneously (1) make relatively accurate out-of-sample predictions, and (2) are uncorrelated or not redundant with predictions made from other models. The validation metric therefore focuses on the models that are best at prediction, while also being sure to include a diverse set of model specifications rather than just minor variations of the same model. The proposed validation metric thus rewards accuracy and penalizes redundancy.

Results

Estimates Resulting from this Model

The results of using this model to develop estimates of the population of U.S. citizens living abroad by State Department regions[7] are displayed in Table 7; individual country estimates for 2010 are provided in Appendix A.

The estimates show that the number of U.S. citizens living overseas has grown steadily from 2000 to 2013, increasing 60% overall during that period. These estimates also show that the majority of the population of U.S. citizens abroad is located in the Western Hemisphere and Europe, and this remained the case throughout the 2000–2010 period.

Table 7. Estimate of the Population of U.S. Citizens Abroad by Global Region, 2000–2010

Year	Africa	East Asia & Pacific	Europe & Eurasia	Near East	South & Central Asia	Western Hemisphere	Global Total
2000	52,763	370,009	923,066	119,414	33,259	1,203,359	**2,701,869**
2001	54,852	380,651	948,868	119,358	33,112	1,223,450	**2,760,291**
2002	54,298	392,833	969,335	112,028	39,512	1,261,526	**2,829,533**
2003	58,033	416,567	1,002,806	127,111	45,102	1,317,421	**2,967,039**
2004	62,538	438,368	1,048,491	149,712	53,070	1,383,127	**3,135,305**
2005	69,460	462,839	1,089,428	162,078	61,763	1,455,999	**3,301,566**
2006	67,516	518,835	1,123,249	169,325	65,897	1,507,595	**3,452,418**
2007	77,297	578,090	1,176,333	189,119	78,893	1,781,450	**3,881,182**
2008	89,888	603,188	1,179,756	203,939	85,259	1,953,433	**4,115,463**
2009	91,470	601,856	1,109,921	211,874	95,017	2,018,579	**4,128,716**
2010	100,052	626,189	1,071,890	234,552	107,732	2,189,973	**4,330,387**
% Change, 2000-2010	90%	69%	16%	96%	224%	82%	**60%**
Average Annual Growth Rate	6.61%	5.40%	1.51%	6.98%	12.47%	6.17%	**4.83%**

Note: Totals are rounded to the nearest person. The sum of the region totals will consequently not equal the global totals.

However, the data also show that Europe displayed by far the slowest rate of growth, while the U.S. populations in Africa, the Near East, and South and Central Asia grew at much higher rates. Among the other regions, East Asia and the Pacific dominate, with a U.S. population that exceeds that of

[7]State Department Region definitions were retrieved from http://www.state.gov/countries/

Africa, the Near East, and South and Central Asia combined. In the Western Hemisphere, the majority of the estimated population is accounted for by Mexico. Within Europe, the largest U.S. populations are in the United Kingdom, France, and Germany. Countries with the largest estimates tend to be those with the largest number of reported Social Security beneficiaries, individuals/households filing tax returns, and exchange students. In addition, countries with the greatest degree of economic (trade), demographic (immigration to the United States), and diplomatic (military aid) interaction with the United States also tend have the largest estimated populations of U.S. citizens.

Country-level estimates for 2000 through 2010 are provided in a separate Excel document, but Table 8 displays the countries with the 10 highest and 10 lowest U.S. citizen population estimates for 2010.

Table 8. Largest and Smallest Estimated Populations of U.S. Citizens Abroad, 2010

10 Largest Estimates		10 Smallest Estimates	
Country	Estimate	Country	Estimate
Mexico	1,109,974	East Timor	18
Canada	365,514	Bhutan	25
United Kingdom	221,118	Solomon Islands	41
France	175,994	Guinea-Bissau	54
Israel	134,647	Sao Tome and Principe	54
Germany	102,894	Comoros	73
Australia	102,176	Vanuatu	81
Japan	94,709	Maldives	96
Taiwan	82,598	Kiribati	111
India	79,562	Djibouti	135

Table 9 shows the countries with the fastest growth and slowest average annual growth rates in U.S. citizen populations over the 2000 to 2010 period. Countries with the fastest growth rates in their estimated number of U.S. citizen residents tended to have an initially small estimated population of U.S. citizens in 2000 and to have traditionally experienced internal and external conflict. Many countries with the highest growth rates in estimated U.S. citizen populations have had historic conflict with the United States. By contrast, countries with the slowest growth rates in estimated U.S.

citizen populations are countries with relatively large U.S. populations at the beginning of the period of interest, and small island states.

Table 9. Largest and Smallest Annual Average Percent Change in Estimated Populations of U.S. Citizens Abroad, 2000–2010

10 Fastest-Growing Countries		10 Slowest-Growing Countries	
Country	Growth Rate	Country	Growth Rate
Afghanistan	41%	Samoa	-3.83%
Jordan	24%	Zimbabwe	-3.75%
Vietnam	24%	United Kingdom	-3.61%
Chad	22%	Hong Kong	-3.32%
Libya	22%	Kiribati	-2.95%
Algeria	21%	Solomon Islands	-2.73%
Iran	22%	Germany	-2.71%
Laos	21%	Macao	-2.62%
Lithuania	21%	Micronesia	-2.54%
Lebanon	21%	Marshall Islands	-2.53%

The tendency for countries with initially small estimated U.S. citizen populations to see greater growth is consistent with trends at the regional level. While the estimated population of U.S. citizens in Europe is relatively high, that region also saw the lowest rates of growth over the 2000–2010 period. By contrast, Africa, the Middle East, and Southern Asia, while having the lowest totals throughout the period, saw the fastest growth. This is consistent with a change in the geographic distribution of the population of U.S. citizens abroad, with U.S. citizens becoming less concentrated over time, and the population of lagging regions beginning to converge with the higher population regions. This is also consistent with trends in the World Bank's estimates of the size of overseas U.S. born/citizen populations by country for the period 1990–2000, where countries with relatively small U.S. populations in 1990 saw faster growth over the subsequent decade than countries with relatively large populations (Ozden, et al., 2011). Figures 2, 3, and 4[8] identify the location of the

[8]In all maps, China, Hong Kong, and Macao are treated as a single observation.

countries with large, but slow-growing overseas U.S citizen populations and those with small, but fast growing populations.

Figure 2. Total Number of Estimated Overseas U.S. Citizens by Country, 2000

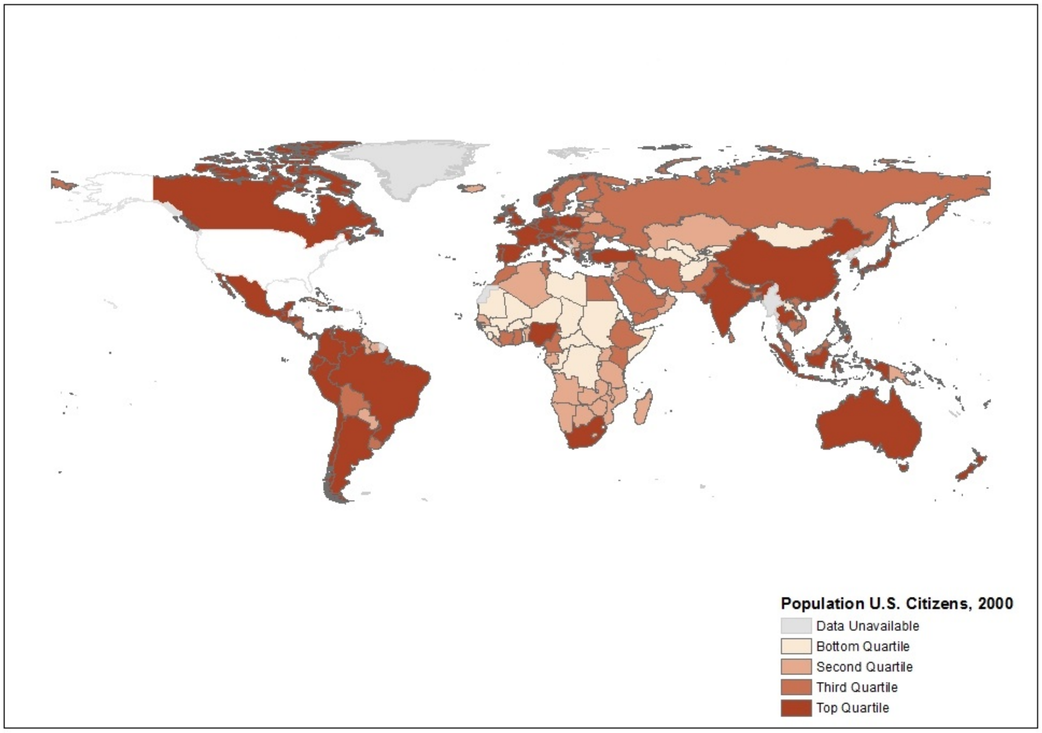
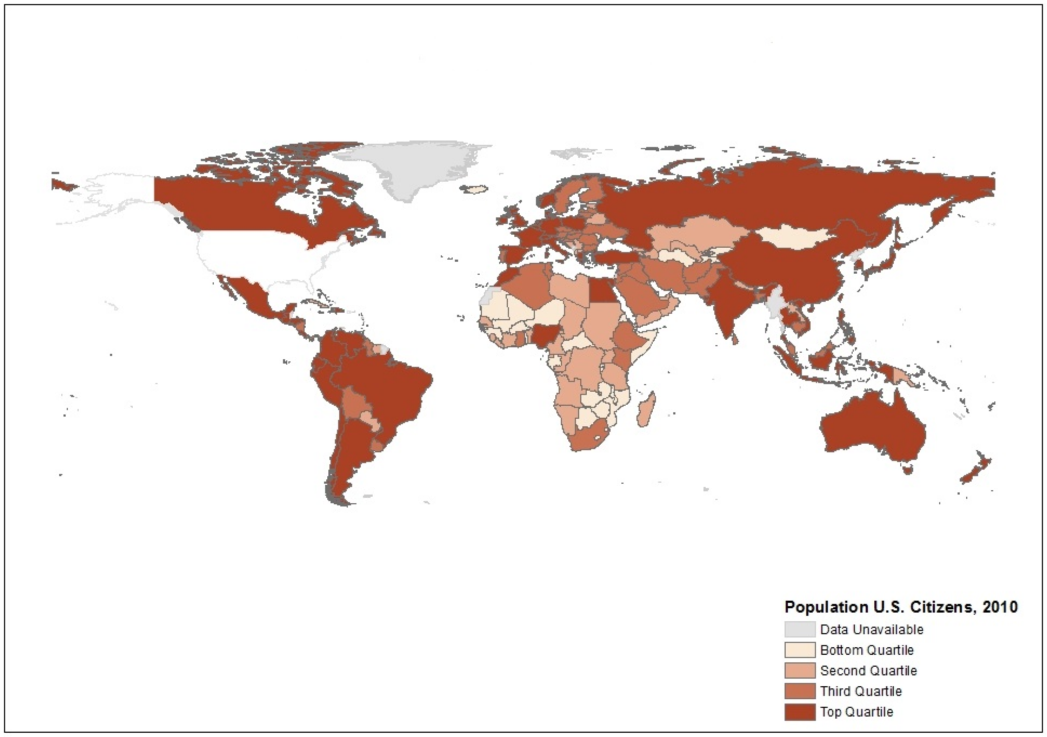
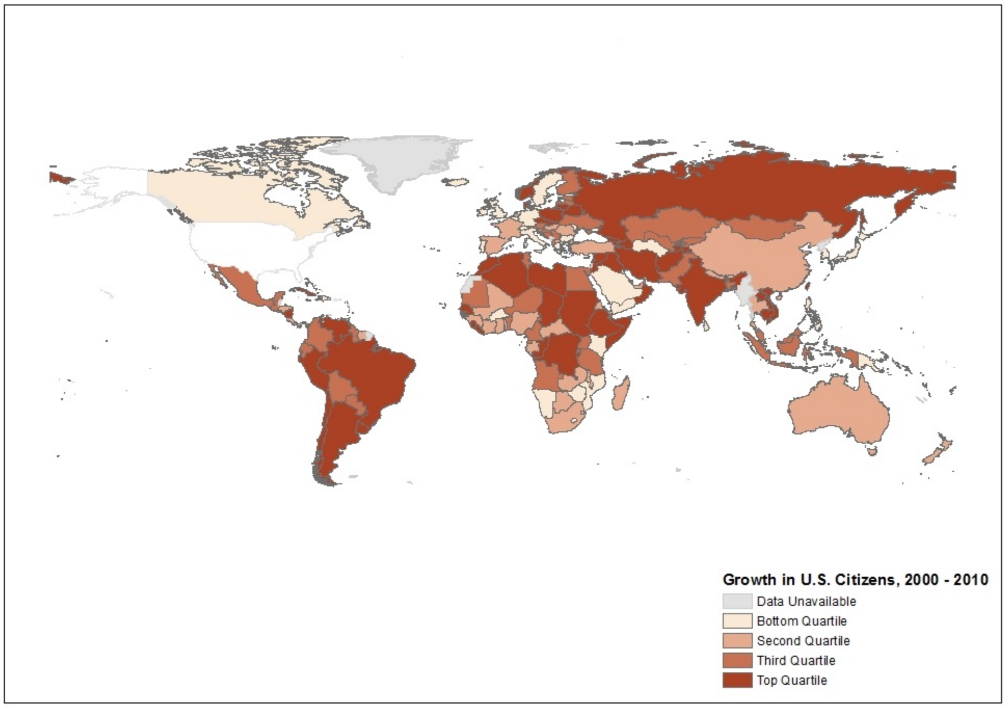


Figure 3. Total Number of Estimated Overseas U.S. Citizens by Country, 2010



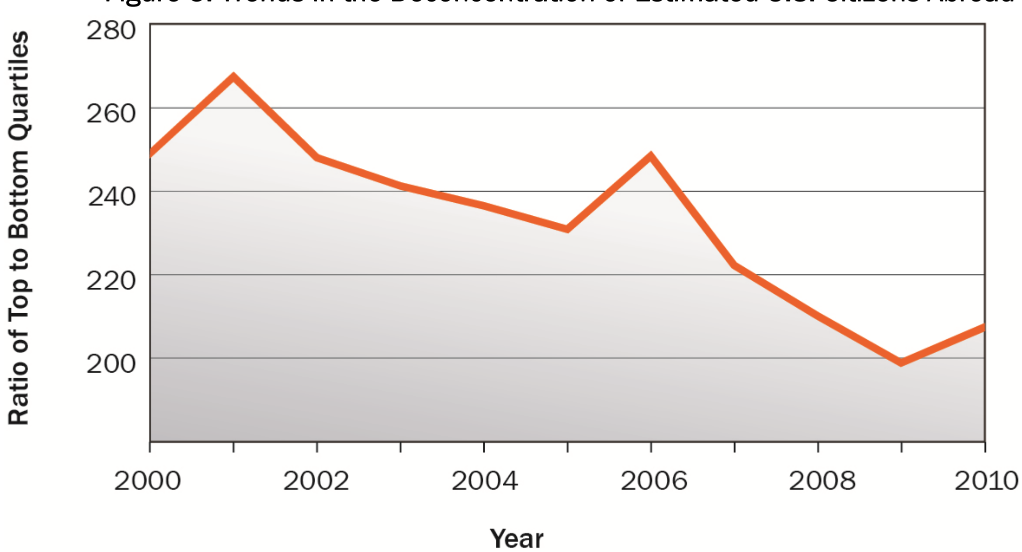
While Western European countries and former British dominions outside Southeast Asia (e.g., Canada, Australia) are in the top quartile of countries with respect to the size of their population in both 2000 and 2010, they are in the lower two quartiles with respect to the growth of their estimated overseas U.S. citizen populations. By contrast, many countries in Africa are in the lower two quartiles in the size of their estimated U.S citizen population in 2000 and 2010, but are in the upper quartile of countries in terms of the growth in that population. It should be noted, however, that several countries in Latin America such as Brazil, Argentina, and Chile are in the top quartiles both in terms of the size of their overseas U.S citizen population at the beginning and end of the 2000–2010 period and are among the top countries with respect to growth. This is consistent with the Western Hemisphere already having the highest estimated number of overseas U.S. citizens in 2000 while still seeing significant estimated growth for the 2000–2010 period.

Figure 4. Growth in the Number of Estimated Overseas U.S. Citizens by Country, 2000–2010



This trend can be seen in Figure 5, where the ratio of estimated overseas U.S. citizens in the top 25% versus bottom 25% of countries is plotted across time. If overseas U.S. citizens were equally distributed across the world, this ratio would be expected to take a value of 1, with higher values representing greater departure from equal distribution.

Figure 5. Trends in the Deconcentration of Estimated U.S. Citizens Abroad



Note: The vertical access is the ratio of the total number of estimated U.S. citizens abroad in countries in the top quartile to the total number of estimated U.S. citizens abroad in the bottom quartile.

In Figure 5, there is an apparent downward trend in the concentration of U.S. citizens abroad. Specifically, in 2000 there were approximately 249 estimated overseas U.S. citizens in the top 25% of countries for every one U.S. citizen in the bottom 25% of countries, but by 2010 there were only 207 estimated U.S. citizens in the top 25% of countries for every U.S. citizen in the bottom 25% of countries. When Mexico is excluded, this trend becomes even more prominent, with 203 estimated U.S. citizens in the top quartile for every U.S. citizen in the bottom quartile in 2000 declining to approximately 149 estimated U.S. citizens in the top quartile of countries for every U.S. citizen in the bottom quartile in 2010.

Any estimate of the population of U.S. citizens living abroad will have some level of uncertainty because of data and sample issues; this uncertainty is reflected in the confidence interval. A confidence interval reflects the range of estimates that has a high probability (95%) of containing the true population count. Table 10 shows the countries whose 2010 estimates displayed the largest and smallest confidence intervals, relative to their mean estimate.

Table 10. Largest and Smallest Confidence Intervals of Estimated Populations of U.S. Citizens Abroad, 2010

10 Largest Confidence Intervals				10 Smallest Confidence Intervals			
Country	Lower Bound	Mean	Upper Bound	Country	Lower Bound	Mean	Upper Bound
Afghanistan	249	3,619	52,488	Belgium	21,611	23,811	26,236
Libya	409	2,143	11,238	Barbados	4,085	4,607	5,196
Laos	234	1,152	5,668	Iceland	670	782	913
Iran	2,030	9,059	40,425	Philippines	57,931	68,449	80,876
Vietnam	5,358	23,420	102,362	Singapore	6,625	7,840	9,278
Lithuania	1,368	5,645	23,292	Namibia	976	1,173	1,409
Algeria	907	3,738	15,402	Netherlands	20,219	24,312	29,234
Lebanon	2,383	9,325	36,490	Maldives	79	96	116
Iraq	1,400	5,264	19,792	Canada	297,742	365,514	448,713
Azerbaijan	382	1,407	5,179	Kenya	5,004	6,194	7,667

Countries with large confidence intervals tend to be those with a high growth in the estimated size of their U.S. citizen populations from 2000 to 2010. This growth appears to be driven to a large degree by high values along country characteristics such as administrative records variables and/or trade. In these countries with large confidence intervals, predictions made by the different models also tend to be similar; this increases the uncertainty for these countries' estimates. By contrast, those countries that have characteristics that result in different models producing less-similar estimates of the number of U.S. citizens tend to have smaller confidence intervals. These less-similar estimates produced by the different models likely result in a “cancelling out” of the error introduced in the different models by limited sample size, resulting in a smaller range that likely contains the true value.

The Consistency of the Results of the Model with Theory

The validity of the analysis in the prior section is dependent upon the validity of the models used to generate estimates of the overseas U.S. citizen populations. This in turn is dependent upon the validity of the predictors. One way to test this validity is to examine the relationship between the final estimates and the country-level predictors and test if the direction of that relationship is consistent with expectations set by the theory used to choose the predictors in the first place. If the predictors are unrelated to the final estimates or the relationship is in the “wrong” direction, this potentially calls into question the model(s) and resulting final estimates because it would indicate a failure to capture the factors that explain the relative sizes of overseas U.S. citizen populations. Descriptive statistics for the FGEs and predictor variables for all country-years for which an estimate was made are listed in Table 11.

Table 11. Descriptive Statistics, All Estimated Country-Years

Variable	N	Mean	Standard Deviation	Minimum	Maximum
FVAP Estimate	2012	18689.75	67605.28	6.34	1109974
World Bank Estimate	182	10097.4	36821.78	0	350626
United Nations Estimates	274	15224.82	51797.21	3	563315
Dualcitizenship	2012	.31	.46	0	1
Administrative Records Variables					
Social Security Beneficiaries	2012	2503.59	9591.34	.04	108194
IRS Form 2555s	2012	1872.06	4490.44	1.20	48644.31
Students	2012	1125.22	3816.23	0	34024
Federal Government Employees	2012	234.38	1340.96	0	18232
Theoretical Variables					
Ln(Difference in GDP per capita)	2012	-2.02	1.34	-5.40	1.19
Population	2012	33283.31	127604.6	45.66	1330141
Distance	2012	4593.24	2014.07	3.45	9093.53
Mean (World Governance Indicators)	2012	-.07	.89	-2.24	1.88
Trade	2012	13745.44	50332.71	.2	600641.2
Immigrants in U.S.	2012	132461.5	503087.5	0	6400000
Military Aid	2012	3.77E+09	1.34E+10	0	1.29E+11
English	2012	.50	.50	0	1
Spanish	2012	.19	.39	0	1
Year of Estimate	2012	2005.00	3.16	2000	2010

In order to examine the relationships between the predictors and estimates, in the first three columns of Table 12 the final estimate is regressed on the administrative records and theoretical variables. In the first column, both the administrative records variables and theoretical variables are included to examine the association between each variable and the final estimate, conditional on the other variables. In the second column, the administrative records variables are dropped because it is expected that the effect of the theoretical variables on the final estimates would be mediated by the size of the different subgroups reflected in the administrative records variables, and thus controlling for them would attenuate the expected relationship of the theoretical variables with the final estimate. Finally, in the third column, the theoretical variables that directly measure the interaction between the United States and the host country (trade, immigration to the United States, and military aid) are dropped so that the effects of the structural variables (level of economic and institutional development, population, distance, and language) can be identified.

Table 12. Determinants of Final Estimates

Variable	(1)	(2)	(3)
Ln(# of Social Security Beneficiaries)	.28*** (.03)		
Ln(# of IRS Returns)	.45*** (.03)		
Ln(STUDENTS)	.18*** (.02)		
Ln(U.S. Government Employment)	-.04 (.02)		
Ln(Difference in GDP per capita)	-.40*** (.08)	.15 (.10)	.58*** (.11)
Ln(Population)	-.18*** (.04)	.07 (.08)	.61*** (.05)
Ln(Distance)	-.09*** (.02)	-.09** (.04)	-.30*** (.04)
Mean(World Governance Indicators)	.03 (.07)	.33** (.15)	.30* (.16)
Ln(Trade)	.17*** (.03)	.29*** (.07)	
Ln(Immigrants in U.S.)	.10*** (.03)	.35*** (.08)	
Ln(Military Aid)	.01** (.01)	.04** (.02)	
ENGLISH	.11 (.07)	.36** (.14)	.60*** (.18)
SPANISH	.07 (.10)	.43* (.23)	.59* (.31)
Year	-.00 (.00)	.03** (.01)	.04*** (.01)
Countries	183	183	183
N	2012	2012	2012
Pseudo R^2	.99	.94	.89

*$p < .10$. **$p < .05$. ***$p < .01$. Model estimated using a Poisson regression. Robust standard errors clustered by country in parentheses.

As indicated in Column 1, the number of Social Security beneficiaries, tax returns filed by U.S. citizens, and students abroad are all positively and significantly associated with the final estimate, consistent with expectation. By contrast, the coefficient on the number of U.S. civilian government employees is statistically insignificant and has a negative sign. This may be due to the fact that government employees may be more likely to be posted to countries subject to external and internal security threats and political instability, which may discourage migration (Ziesemer, 2010). Consequently, this variable could be capturing unobserved conditions in a country that makes it less attractive as a destination to many U.S. migrants.

Among the theoretical variables that capture interactions between the United States and the host country, trade, migration, and military aid are each, as expected, positively and statistically significantly associated with the final estimate, both when controlling for the administrative records variables and after dropping them. When the administrative records variables are dropped, the coefficient on each theoretical variable becomes larger. This indicates that while the administrative records variables might be capturing some of the effect of these interaction variables on the final estimates, the interaction variables may be proxying for the existence of populations not directly captured in the administrative records variables. Finally, the coefficients for the "structural" variables, with the exception of distance, are either statistically insignificant (English and Spanish dummies, institutional quality), or have the wrong sign (population and difference in GDP per capita) when controlling for both the administrative records and interaction variables. Once the administrative records variables are dropped in the second column, none of the structural variables have the wrong sign, and some (the language dummies, institutions) gain statistical significance. Once the interaction variables are dropped in the third column, each structural variable has both the expected sign and is statistically significant. This indicates that while the estimates have the theoretically expected relationship with the predictor variables, the structural variables added relatively little additional explanatory power to the model set.

Differences between the Estimates from this Methodology and Prior Estimates

In Table 13, the impacts of the administrative records and theoretical variables on the size of the estimates relative to the World Bank and United Nations estimates are analyzed by regressing the logged ratio of the World Bank and United Nations estimates to the FVAP estimates. In columns 1 and 2, the dependent variable is the ratio the World Bank estimate to the FVAP estimate. In columns 3 and 4, the dependent variable is the United Nations estimate to the FVAP estimate. Positive coefficients indicate that countries with high values on a given predictor have FVAP estimates that are small relative to their World Bank/United Nations estimates, and countries with negative coefficients have FVAP estimates that are relatively large.

Table 13. Correlates of Deviations from Prior Estimates

Variable	World Bank/FVAP Estimates, 2000		United Nations/FVAP Estimates, 2000 and 2010	
DUALCITIZENSHIP	-.93*** (.28)	-.95*** (.24)	-.63** (.25)	-.78*** (.22)
Ln(# of Social Security Beneficiaries)	.25* (.14)		-.07 (.11)	
Ln(# of IRS Returns)	-.03 (.21)		-.35*** (.09)	
Ln(STUDENTS)	-.01 (.06)		.03 (.07)	
Ln(U.S. Government Employment)	.15** (.07)		-.05 (.09)	
Ln(Difference in GDP per capita)	.41** (.19)	.44*** (.14)	.10 (.15)	-.06 (.15)
Ln(Population)	.38** (.10)	.45*** (.11)	.48*** (.12)	.45*** (.12)
Ln(Distance)	.27 (.31)	.14 (.29)	-.22* (.13)	-.28** (.12)
Mean(World Governance Indicators)	-.24 (.22)	-.04 (.20)	.44** (.22)	.54** (.23)
Ln(Trade)	-.39** (.17)	-.35*** (.10)	-.14 (.10)	-.37*** (.08)
Ln(Immigrants in U.S.)	-.22*** (.07)	-.12** (.05)	-.05 (.08)	-.10 (.07)
Ln(Military Aid)	-.10*** (.02)	-.08*** (.02)	-.07*** (.02)	-.10*** (.01)
ENGLISH	.21 (.25)	.23 (.20)	.87*** (.32)	.74*** (.26)
SPANISH	-.66** (.33)	-.46 (.29)	-.03 (.36)	-.05 (.34)
Western Hemisphere	1.37** (.61)	1.50** (1.03)	.23 (.56)	.42 (.50)
South/Central Asia	-.46 (.41)	-.77** (.35)	.70 (.45)	.95** (.43)
Near East	1.51*** (.37)	1.60*** (.29)	1.72*** (.43)	1.87*** (.42)
Europe	.62 (.50)	.91 (.71)	-.03 (.43)	.00 (.38)
East Asia/Pacific	.03 (.34)	.05 (.36)	-.03 (.38)	-.15 (.40)
2010			-.36* (.20)	-.16** (.07)
Countries	182	182	137	137
N	182	182	274	274
Pseudo R^2	.42	.40	.61	.59

**p* <.10. ***p* < .05. ****p* < .01. Model estimated using a Poisson regression. Robust standard errors clustered by country in parentheses. The reference region is Africa.

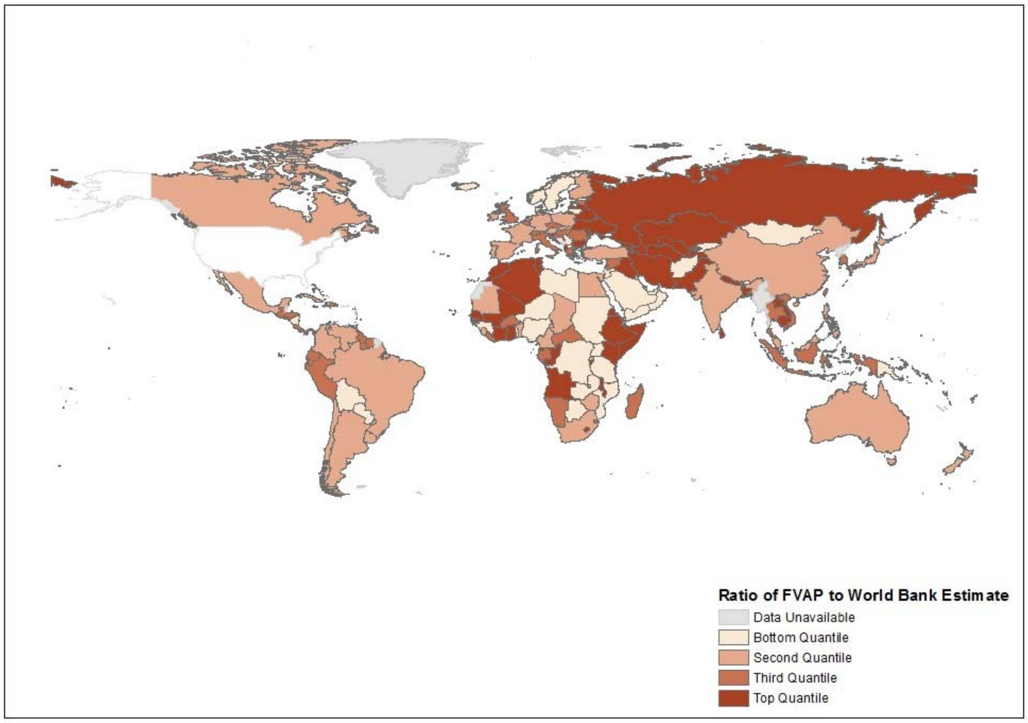
There are several variables for which the sign is consistent for both the World Bank and United Nations regressions.[9] The results indicate that countries that allow dual citizenship with the United States have FVAP estimates which are large relative to the World Bank and United Nations

[9]Although there are some variables that have opposite signs for the World Bank and United Nations regressions, the documentation on the generation of the United Nations is relatively light, and does not offer a basis for explaining differences between the two alternate sets of estimates.

estimates. This can be explained by the adjustments made to the FVAP estimates to ensure that countries that provide a count of citizens also count dual citizens. Regressions also indicate that countries with large populations tend to have World Bank and UN estimates that are large relative to their FVAP estimates. This may be an artifact of how the World Bank (and potentially the United Nations) imputed values for missing later decades. Specifically, they would assume that the share of total migrants in a country composed of individuals originating in the United States remained fixed relative to some prior decade or else took on a regional average. Consequently, if countries with large populations also had large numbers of migrants (from any country), then the number of U.S. born/citizens in the country would rise with population. By contrast, the FVAP estimates are derived using the empirical association between population and the size of the overseas U.S. citizen population.

The other consistent difference between the FVAP estimates and the World Bank and United Nations estimates is that the countries with high values on trade and military aid have FVAP estimates that are large relative to the World Bank and United Nations estimates. Given that each of these variables also was positively associated with the absolute size of the FVAP estimate, this may simply reflect the fact that these variables do not have an association with the data used to generate the World Bank and United Nations estimates. This might reflect the fact that the World Bank and UN estimates were imputed based on past estimates and/or regional averages. If there have been significant changes in the patterns of trade, perhaps because of the end of the Cold War and other factors that are leading to a more integrated global economy, then countries that have significant trade with the United States today would not necessarily have had significant trade with the United States in the past. With respect to military aid, if military aid is assigned based on need, then countries that are receiving military aid may not be attractive destinations for migrants. However, if that aid leads to better relations with the United States, then over the long run the number of U.S. migrants in the recipient country of the aid might increase. Thus, military aid might have an insignificant or even negative association with past migration, but a positive relationship with contemporary migration.

Figure 6. FVAP Estimate Relative to the World Bank Estimate, 2000



Further evidence for the importance of lagged data in explaining the difference between the FVAP estimates and the World Bank estimates is presented in Figure 6, which depicts the ratio of the 2000 FVAP estimate to the World Bank estimate in quartiles. Note that countries in the top quartile (i.e., those where the FVAP estimate is particularly high relative to the World Bank estimates) are heavily clustered in the former Soviet Union.[10] This likely reflects a situation that dominated in the Cold War, where there was limited migration between the United States and the former Soviet Union, but is less true now. Consequently, the interpretation of the differences between the size of the World Bank and United Nations estimates and the FVAP estimates is that the latter are produced using contemporary cross-country relationships between predictors and FGEs. By contrast, the World Bank estimates are imputed based on lagged data, resulting in the World Bank and United Nations estimates having relatively higher estimates in countries to which U.S. citizens have traditionally migrated, while the FVAP estimates are relatively high for countries with which the United States currently has strong links with respect to trade and migration.

[10]The United Nations does not provide estimates for many countries, and specifically many countries in the former Soviet Union. Consequently, a comparison based upon quartiles would not provide much information.

Discussion

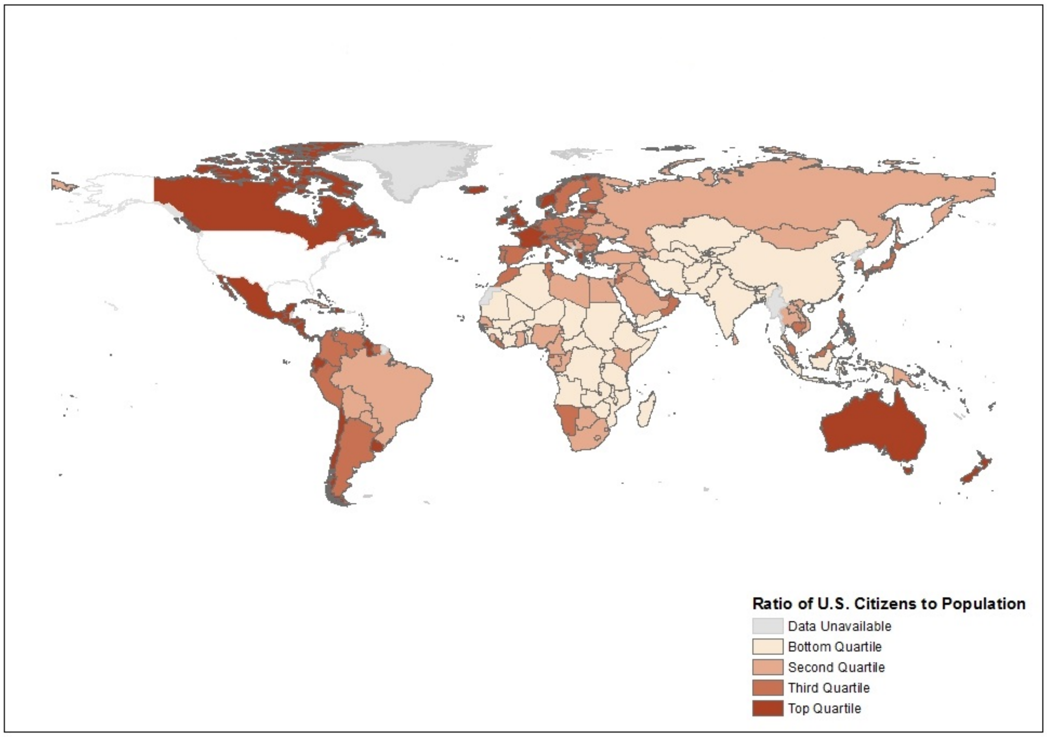
The goal of this effort was to develop a method of estimating the population of U.S. citizens abroad that was efficient, transparent, reproducible, created in a scientifically grounded fashion, and which would allow FVAP to more effectively and efficiently allocate resources, target its voter assistance outreach to the greatest number of *UOCAVA* voters, and identify the levels of success *UOCAVA* citizens are experiencing when voting. Although no estimate should ever be considered “final,” the methodology described in this report produced country-level estimates for each year from 2000 to 2010, using a method that can be reproduced and refined by future researchers. In contrast to the full enumeration methodology considered by the Census Bureau, this method did not require extensive field collection work to produce data, but rather utilizes data already produced by foreign governments, which presumably have greater capacity to estimate U.S. citizens in their own territories. By developing a model of these estimates, the size of the population of U.S. citizens abroad can be estimated more efficiently than by using a full enumeration approach. Unlike the World Bank and United Nations data sets, these estimates are made using relatively contemporary (2000–2010) FGEs and related predictors of the size of the overseas U.S. population. Consequently, this method of estimating should better reflect the current geographic distribution and dating of this population. This approach is broadly similar to that used by the 2011 OCC Report, insofar as it is based on using foreign sources to estimate the overseas U.S citizen population by country. But because it uses contemporary predictors of migration, rather than lagged migration data, the set of estimates provided in this report are likely to suffer less from the shortcomings described above. In addition, these estimates were generated using predictors that are theoretically justified, and the estimation procedure mitigates issues related to sample selection by weighting observations and predictions from different models such that the estimates are more likely to be valid for countries for which FGEs are unavailable. Finally, this methodology is has been subject to a variety of robustness checks discussed in Appendixes B and C.

The estimates provided in this report help to provide a picture of the size and geographic distribution of the population of U.S. citizens abroad as well as its change over time, and the changing geographic distribution of the overseas U.S. citizen population revealed could have strong implications for how FVAP allocates resources in the future. Specifically, while the estimates indicate the U.S. citizen population is to a large extent concentrated in Europe and the Western Hemisphere and has remained so throughout the 2000–2010 period, there are substantial differences in the estimated rate of growth between countries and regions that suggest an increase in the geographic dispersion of U.S. citizens. Though there is a large degree of uncertainty in the numbers of U.S.

citizens located in the countries seeing the fastest growth, FVAP may wish to consider how it will adapt to a potential rise in the number of U.S. citizens in Africa, Asia, and the Near East.

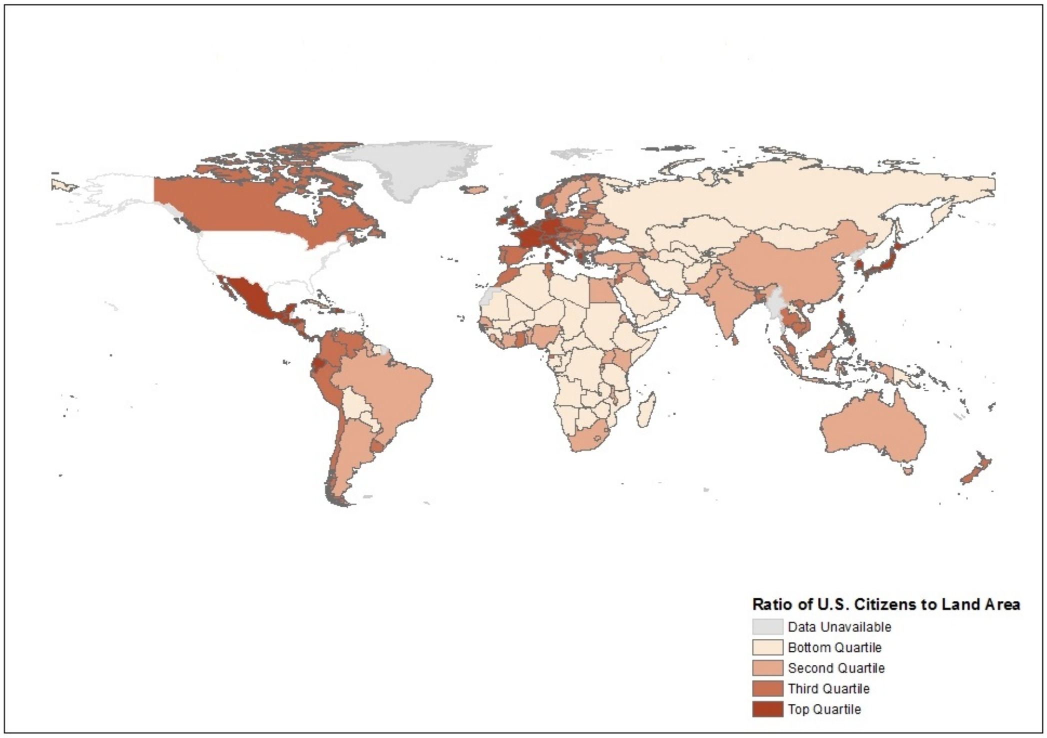
While the total number of overseas U.S. citizens within a country gives some indication of the benefit to FVAP and other organizations interested in engaging with overseas U.S. citizens in investing resources in the country, another relevant factor is the cost of reaching out to these citizens, which is likely to vary by country. Two proxies for these costs are used: population and land area.

Figure 7. Ratio of Estimated U.S. Citizens Abroad to Country Population, 2010



The necessity to identify a country's residents as either U.S. citizens or non-U.S. citizens might be greater in countries with large total populations, holding the number of U.S. citizens constant, as the probability that any given resident of the country is a U.S. citizen will be lower. If distinguishing between U.S. citizens and noncitizens is costly, then investing resources in a country with a large U.S. citizen population but where U.S. citizens make up a small percentage of the total population may be inefficient. Figure 7 displays countries coded by the ratios of the estimated number of U.S. citizens in 2010 to the country's total population. U.S. citizens comprise a relatively large (top two quartiles) percentage of the total population in Europe, North America, and Latin America as well as in some East Asia and Pacific countries. By contrast, countries with a relatively low fraction of their total population composed of U.S. citizens are largely concentrated in Africa, the former Soviet Union, and the Asian mainland.

Figure 8. Ratio of Estimated U.S. Citizens Abroad to Country Land Area, 2010



It is also expected to be costly to identify and engage with U.S. citizens in geographically large countries because the transportation costs involved in reaching these populations may be large in these countries. As seen in Figure 8, geographic patterns in the percentage of a country's total population composed of U.S. citizens largely holds when the ratio of U.S. citizens to land area is used instead, though in this case the former Soviet Union is relatively worse off with respect to the density of U.S. citizens than the Asian mainland. It should be noted that while population and land area may influence the costs of engaging in face-to-face outreach, they may be less relevant in countries where social media and other forms of online communication are viable. One the other hand, many countries in which there is the greatest density (per capita or per unit land area) of U.S. citizens are also likely to have the most developed Internet infrastructure, as indicated by the high density of U.S. citizens in Western Europe, former European colonies, and Japan.

With all of that being said, it is also key to remember that a handful of countries—Mexico, Canada, the United Kingdom, France, Israel, Germany, and Australia—continue to represent slightly over half (approximately 52%) of the population of U.S. citizens abroad. Any outreach and voter support resources that address those countries will continue to target most potential *UOCAVA* voters.

Limitations

Within any study of this nature, there are inherent limitations. Most have been covered within the discussion. It makes sense, however, to summarize them as a way to frame expectations and look for improvements in the future by:

- Refining methods;
- Improving availability; and
- Increasing "actionability."

To begin with an overview of the limitations in what has been done, in any exercise of this sort, with neither the budget nor the time to collect new data, reliance is placed on existing, largely official statistical sources. These were censuses and registries, drawn from U.S. and other national statistical offices around the world. This meant that the results were subject to differences in approach, usually driven by the individual country or administrative source. Commonly enough, these sources were originally available for a purpose other than the use intended here. Timing and definitional differences were major challenges, not always surmountable. Fortunately, because of the European Union (EU) there was somewhat greater uniformity of reporting in that part of the world. Still, much of the problem is model- or adjustment-driven.

Efforts were made to align the foreign country-by-country results provided here to make the exercise as consistent as possible. However, one can assume that the observed association between having a registry or census and the estimate of the size of the overseas U.S. population reflects differences in how the foreign government estimated the population rather than differences in the "true" U.S. population. If there are systematic, unobserved differences between countries that produced an estimate with a registry or census, and these differences affect the size of a country's U.S. citizen population, then bias may be introduced in the final estimates. This approach thus relies on the assumption that the administrative records–based variables and theoretical variables captured these systematic differences. However, with this assumption, the incomparability between census- and registry-based estimates has been more or less satisfactorily addressed.

Another limitation to the statistical methodology relates to how the possibility that the sample of countries was not representative was addressed. Inverse-probability weighting corrects for nonresponse bias to the degree that there are not unobserved factors that affect both the size of the FGE and probability that a country has an FGE. If the logit model did not capture all relevant nonignorable factors, then the results will still suffer from selection bias, and MSE and other measures of fit will not reliably indicate the quality of a model with respect to its ability to create an accurate prediction for countries without an FGE. This selection bias is potentially exacerbated by the fact that for many countries outside the sample, the administrative records variables had to be imputed, and are thus likely of lower quality. This adds additional uncertainty to the FVAP estimates for these countries that is not incorporated into the confidence intervals.

Because of these issues, a second, entirely separate, U.S.-based approach using a generalization of the Sekar–Deming capture-recapture model was developed. This generalized Sekar–Deming effort is described in Appendix B. To make it workable, tabulations were required from Federal Government agencies. Typically these would have been affordable and obtainable. However, the relevant agencies reported that they were unable to provide the research team with the data in time for use in this effort because of the limitations that sequestration put on their resources. As the necessary data was unavailable, a microsimulation that was unsatisfactory as a fully developed alternative to the regression effort employed here had to be used. This effort, however, did show that both model approaches were getting at the same thing and had a degree of reassuring interchangeability.

Next Steps/Future Research

There are a number of different options that could be pursued in future research efforts. Some efforts could focus on further refining this method of estimation, and other efforts could focus on using estimates produced through this model to answer research questions relevant to the FVAP mission. Two possibilities are detailed below.

Producing Subnational Estimates of the Population of U.S. Citizens Abroad

To obtain estimates of the number of U.S. citizens by subnational region consistent with the current country aggregates, researchers could estimate the population of all subnational regions by applying a multilevel modeling framework to a sample of regions with a current estimate/proxy of the number of U.S. citizens. The spatial distribution of a given country's population as predicted by this model would be used to apportion country-level estimates.

The data required to create these subnational estimates would include:

1) *Region Definition:* A definition of regions that is inclusive of all territory in a country will be needed. This requirement derives from the fact that the 2013 estimates are intended to represent the entire country. Consequently, a definition based on metropolitan area would not be useful because it would exclude the rural population and the apportionment of the country aggregate to the individual regions would have an upward bias. A definition based on state/province/prefecture would be more useful because the sum of all regional populations could be equated with a rescaled national population. Although it may be reasonable to

assume that in most countries U.S. citizens are concentrated in cities, this would be better captured by including urbanization as a predictor in the model.

2) *Regional Predictors:* Creating the model will require data on salient characteristics of the region (i.e., factors that explain variation in the number of U.S. citizens). Access to equivalent data at the region level as at the count level may not be available (particularly GDP per capita, trade dependence, and administrative count data from U.S. sources). However, population, population density, urbanization, and the availability of certain types of relevant infrastructure (airports, seaports, distance from the coast, capital city dummies, etc.) are likely heavily correlated with the above. Country-level estimates can also be incorporated to account for heterogeneity in the effect of the region-level factors. For instance, if distance from the nearest airport was used as predictor, one would expect to find a larger negative effect in developing countries because of weaker land transportation infrastructure in such countries. This heterogeneity could be incorporated into the model using an interaction between GDP per capita and distance.
3) *U.S. Citizen Estimates by Region:* To construct a dependent variable, an estimate/proxy of the number of U.S. citizens by subnational region would be needed. This could be provided by the governments either at the preferred level of territorial aggregation or, if the data were available at a lower, nested level, such as city, the researchers could aggregate it. There is also the issue of country representation. If U.S. citizens are heavily concentrated in developed countries, then there may be a rural bias in the estimate because expats may be more willing to locate in rural regions in developed countries as a result of there being a lower urban–rural gap in the provision of public goods and services in such countries. Although this heterogeneity can be modeled in a multilevel model, this will only work if there is variation in the country-level data. This would require estimates for at least some developing countries. If there are not observed U.S. citizens in any developing countries, then that makes creating valid estimates for those countries impossible.

Examining the Voting Success Rate of the Population of U.S. Citizens Abroad

Researchers could compare the percentage of the registered and/or potential *UOCAVA* population who successfully submit a ballot to that of similar domestic voters. This could be done by developing a demographic profile of *UOCAVA* voters by country and state and comparing the percentage of registered/total potential voters in that population who were able to vote to the voting success rate of a sample of U.S. citizens (registered and nonregistered) residing in the United States with similar demographics.

The data required to examine the voting success rate would include:

1) A demographic profile of the *UOCAVA* population by country and state. It is unlikely that demographic data for the civilian overseas U.S. citizen population will be available from most countries' statistical agencies. However, microlevel data may be available for OECD countries with respect to education, sex, and occupation, as well as through microlevel data taken directly from foreign government statistical agencies. For developing countries, the demographics may have to be imputed based on various covariates (level of development, distance/contiguity with the United States, etc.). For registered voters, data may be available from U.S. state governments.
2) Demographic profile of nonimmigrated U.S. population by state. Similar demographic variables will be needed in order to create a sample that matches the overseas population. This data may be available from the Census Bureau and/or state government voting agencies.

References

Abrahantes, J. C., Sotto, C., Molenberghs, G., Vromman, G., & Bierinckx, B. (2011). A comparison of various software tools for dealing with missing data via imputation. *Journal of Statistical Computation and Simulation*, 81(11), 1653-1675.

Adsera, A., & Pytlikova, M. (2012). *The role of language in shaping international migration* [IZA Discussion Paper No. 6333]. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2003666

Alesina, A., & Dollar, D. (2000). Who gives foreign aid to whom and why? *Journal of Economic Growth, 5*(1), 33–66.

Artuc, E., Docquier F., Ozden C., & Parsons, C. (2013). *A Global Assessment of Human Capital Mobility: the Role of non-OECD Destinations.* Working Paper retrieved from http://perso.uclouvain.be/frederic.docquier/filePDF/DOPA_V2.pdf

Berthelemy, J., Beuran, M., & Maurel, M. (2009). Aid and migration: Substitutes or complements? *World Development, 37*(10), 1589–1599.

Bertoli, S., & Fernandez-Huertas Moraga, J. (2013). Multilateral resistance to migration. *Journal of Development Economics, 102*, 79–100.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice* (chapter 6). Cambridge, Massachusetts Institute of Technology.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

Buja, A., Hastie, T., & Tibshirani, R. (1989). Linear smoothers and additive models. *The Annals of Statistics*, 453-510.

Burnham, K.P. and Anderson D.R. (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods Research* 33: 261 - 304.

Card, D., & Krueger, A. B. (1995). Time-series minimum-wage studies: A meta-analysis. *The American Economic Review 85*(2), 238–243.

Dumont, J. C., & Lemaître, G. (2005, July). *Counting immigrants and expatriates in OECD countries: A new perspective*. Working paper presented at the United Nations Expert Group Meeting on International Migration and Development, New York, NY. Retrieved from http://www.un.org/esa/population/meetings/ittmigdev2005/P09_Dumont%26Lemaitre.pdf

Ewing, I. (1998, September). Technological change and the use of administrative data in economic statistics: What does it mean for the Business Register? Paper presented at the 12th International Roundtable on Business Survey Frames. Helinski, Finland.

Federal Voting Assistance Program. (2011). *Overseas citizens count (OCC) 2011.*

Felbermayr, G. J., & Toubal, F. (2012). Revisiting the trade-migration nexus: Evidence from new OECD data. *World Development, 40*(5), 928–937.

Fleck, R. K., & Kilby, C. (2010). Changing aid regimes? U.S. foreign aid from the Cold War to the War on Terror. *Journal of Development Economics, 91*(2), 185–197.

Government Accountability Office. (2004). 2010 Census Counting Americans Overseas as Part of the Census Would not be Feasible. Washington, DC: U.S. Government Accountability Office. Retrieved from http://www.gao.gov/new.items/d041077t.pdf

Government Accountability Office. (2007). Evacuation Planning and Preparations for Overseas Posts Can be Improved. Washington, DC: U.S. Government Accountability Office. Retrieved from http://www.gao.gov/assets/270/268208.pdf

Grogger, J., & Hanson, G. H. (2011). Income maximization and the selection and sorting of international migrants. *Journal of Development Economics, 95*(1), 42–57.

Hawkins, D. M. (2004). The problem of overfitting. *Journal of Chemical Information and Computer Sciences, 44*(1), 1–12.

Heston, A., Summers, R., & Aten, B. (2012, November). Penn World Table Version 7.1. Center for International Comparisons of Production, Income and Prices at the University of Pennsylvania.

Hollenbeck, S., & Kahr, M. K. (2009). Individual foreign-earned income and foreign tax credit, 2006. *IRS Statistics of Income Bulletin*. Retrieved from http://www.irs.gov/pub/irs-soi/09sprbulinforincometc.pdf

Institute of International Education (2012). Host regions and destinations of U.S. study abroad students, 2009/10–2010/11. *Open Doors Report on International Educational Exchange*.

Internal Revenue Service. Individual Foreign-Earned Income and Foreign Tax Credit. Retrieved from: http://www.irs.gov/uac/SOI-Tax-Stats-Individual-Foreign-Earned-Income-Foreign-Tax-Credit (Accessed 7/11/2013)

Klekowski von Koppenfels, A. and Costanzo J. (2013). Counting the Uncountable: Overseas Americans. Migration Policy Institute.

Kulish N. and Cottrell C. (2013, May 31). Germany Counts Heads and Finds 1.5 Million Fewer Residents Than It Expected. *The New York Times.* Retrieved from: http://www.nytimes.com/2013/06/01/world/europe/census-shows-new-drop-in-germanys-population.html?_r=0

Lafleur, J.M. (2012). *Transnational politics and the state: The external voting rights of diasporas.* New York, NY: Routledge.

Lewer, J., & Van den Berg, H. (2008). A gravity model of immigration. *Economic Letters, 99*(1), 164–167.

Lewis, M. P., Grimes, B. F., Simons G. F., & Huttar, G. (2009). *Ethnologue: Languages of the world* (Vol. 9). Dallas, TX: SIL International.

Markham, T., Falk E., & Scheuren, F. (2013). Nonresponse Modeling in Repeated Independent Surveys in a Closed Stable Population--Did the Local Election Officials (LEOs) Roar in 2012? Proc. of Washington Statistical Society Seminar.

Mayda, A. M. (2010). International migration: a panel data analysis of the determinants of bilateral flows. *Journal of Population Economics, 23*(4), 1249–1274.

Montgomery J.M. and Nyhan B. (2010). Bayesian Model Averaging: Theoretical Developments and Practical Applications. Political Analysis 18(2): 245-270.

Montgomery, J. M., Hollenbach, F. M., & Ward, M. D. (2012). Improving predictions using ensemble Bayesian model averaging. *Political Analysis, 20*(3), 271–291.

Ozden, C., Parsons, C. R., Schiff, M., & Walmsley, T.L. (2011). Where on earth is everybody? The evolution of global bilateral migration, 1960–2000. *The World Bank Economic Review, 25*(1), 12–56.

Punch, A. (2001). Traditional census versus alternatives. Paper presented at the Symposium on Global Review of 2000 Round of Population and Housing Censuses: Mid-Decade Assessment and Future Prospects. New York, New York.

Sangita, S. (2013). The effect of diasporic business networks on international trade flows. *Review of International Economics, 21*(2), 266–280.

Santos Silva, J. M. C., & Tenreyro, S. (2006). The log of gravity. *The Review of Economics and Statistics, 88*(4), 641–658.

Schachter, J. (2008, December 24). Estimating native emigration from the United States. Memorandum developed during contract work for the U.S. Census Bureau.

Scheuren, F. (2012). *Overseas citizens count OCC report review: Review and recommendations*.

Sekar, C. C., & Deming, W. E. (1949). On a method of estimating birth and death rates and the extent of registration. Journal of the American Statistical Association, 44(245), 101-115.

Singapore Department of Statistics (1999). Singapore Census of Population, 2000: The First Register-Based Census. Paper presented at the ESCAP Working Group of Statistical Experts, 11th Session. Bangkok, Thailand.

Statistics Netherlands (2012, May). Methodology used for estimating Census tables based on incomplete information. Paper presented at the Economic Commission for Europe Conference of European Statistics, Geneva, Switzerland.

Stekhoven, D.J., & Buehlmann, P. (2012). MissForest - nonparametric missing value imputation for mixed-type data. Bioinformatics, 28(1), 112-118. doi: 10.1093/bioinformatics/btr597

Stone, C. J. (1985). Additive regression and other nonparametric models. The Annals of Statistics, 689-705.

Stone, M. (1977). Asymptotics for and against cross-validation. *Biometrika, 64*(1), 29–35.

The World Bank Group. (2012). *The world development indicators 2012.*

U.S. Census Bureau. (2001, September 27). *Issues of Counting Americans Overseas in Future Censuses*.

United Nations, Department of Economic and Social Affairs, Statistical Office (1969). *Methodology and evaluation of population registers and similar systems.* Retrieved from http://unstats.un.org/unsd/demographic/sources/popreg/Series_F15.pdf

United Nations, Department of Economic and Social Affairs, Population Division. (2011). *Trends in international migrant stock: migrants by age and sex*. Retrieved from http://esa.un.org/MigOrigin/

Ward, M. D., Greenhill, B. D., & Bakke, K. M. (2010). The perils of policy by p-value: Predicting civil conflicts. *Journal of Peace Research 47*(4), 363–375.

Warren, R., & Peck, J. M. (1980). Foreign born emigration from the United States: 1960 to 1970. *Demography 17*(1), 71–84.

Wennersten, J. (2008). Leaving America: The New Expatriate Generation. Westport, CT: Praeger Publishers.

Wooldridge J.M. (2002). Inverse probability weighted M-estimators for sample selection, attrition, and stratification. *Portuguese Economic Journal 1*(2), 117-139.

Ziesemer, T. H. W. (2010). Developing countries' net-migration: The impact of economic opportunities, disasters, conflicts, and political instability. *International Economic Journal, 25*(3), 373–386.

Appendix A: Estimates of the Population of U.S. Citizens Abroad, by Country, 2000 and 2010

	2000			2010			Growth in Overseas Citizens Populations, 2000-2010	
Country	95% CI Lower Bound	Mean Estimate	95% CI Upper Bound	95% CI Lower Bound	Mean Estimate	95% CI Upper Bound	% Change in Mean Estimate, 2000–2010	Average Annual Growth Rate
Global Totals	1,832,636	2,701,869	4,210,347	2,622,359	4,330,387	7,790,496	60%	4.83%
Afghanistan	8	113	1,642	249	3,619	52,488	3097%	41.41%
Albania	373	643	1,110	886	1,527	2,633	137%	9.03%
Algeria	130	537	2,213	907	3,738	15,402	596%	21.41%
Angola	373	627	1,056	816	1,373	2,311	119%	8.15%
Antigua and Barbuda	1,015	1,474	2,142	1,108	1,610	2,339	9%	.88%
Argentina	4,944	13,989	39,581	17,246	48,798	138,078	249%	13.31%
Armenia	148	346	809	463	1,083	2,533	213%	12.09%
Australia	54,067	69,101	88,315	79,950	102,176	130,580	48%	3.99%
Austria	4,980	8,371	14,071	10,045	16,884	28,382	102%	7.27%
Azerbaijan	72	266	978	382	1,407	5,179	429%	18.13%
Bahamas	1,723	3,045	5,380	2,557	4,517	7,980	48%	4.02%
Bahrain	352	616	1,079	505	884	1,548	44%	3.68%
Bangladesh	1,611	2,998	5,577	3,336	6,206	11,548	107%	7.55%
Barbados	3,801	4,282	4,824	4,085	4,607	5,196	8%	.73%
Belarus	213	512	1,230	827	1,986	4,769	288%	14.51%
Belgium	23,016	25,335	27,887	21,611	23,811	26,236	-6%	-.62%
*Belize**								
Benin	241	456	864	509	963	1,823	111%	7.76%
*Bermuda**								
Bhutan	8	15	27	13	25	46	71%	5.51%
Bolivia	871	1,550	2,758	1,901	3,384	6,023	118%	8.12%
Bosnia and Herzegovina	309	628	1,279	726	1,478	3,008	135%	8.93%
Botswana	294	488	811	478	794	1,319	63%	4.98%
Brazil	9,525	21,513	48,589	29,937	67,623	152,751	214%	12.13%
Brunei	154	213	294	147	203	281	-5%	-.46%
Bulgaria	2,348	2,987	3,800	3,186	4,052	5,155	36%	3.10%
Burkina Faso	171	259	391	241	364	551	41%	3.49%
Burundi	26	68	176	86	222	572	226%	12.53%
Cambodia	1,655	4,143	10,375	5,960	14,924	37,367	260%	13.67%
Cameroon	694	1,344	2,603	1,388	2,690	5,211	100%	7.19%

Canada	338,523	415,642	510,330	297,742	365,514	448,713	-12%	-1.28%
Cape Verde	233	338	490	358	519	752	54%	4.38%
Central African Republic	70	136	264	141	272	528	100%	7.18%
Chad	40	142	509	301	1,075	3,845	656%	22.42%
Chile	5,253	12,893	31,649	17,198	42,217	103,634	227%	12.59%
China	6,277	18,414	54,018	25,376	74,429	218,307	304%	14.99%
Colombia	9,421	17,523	32,596	22,541	41,922	77,966	139%	9.11%
Comoros	19	37	72	37	73	144	100%	7.19%
Congo, Dem. Rep.	154	406	1,069	528	1,390	3,661	242%	13.10%
Congo, Republic of	145	398	1,091	543	1,489	4,079	274%	14.09%
Costa Rica	14,841	23,581	37,467	25,882	41,141	65,394	74%	5.72%
Cote d`Ivoire	1,099	1,714	2,672	1,801	2,809	4,380	64%	5.06%
Croatia	2,237	3,891	6,768	5,023	8,737	15,197	125%	8.43%
Cuba	328	812	2,008	991	2,452	6,063	202%	11.69%
Cyprus	1,307	2,149	3,533	1,368	2,248	3,697	5%	.45%
Czech Republic	1,190	3,770	11,949	5,358	16,984	53,835	351%	16.24%
Denmark	3,113	7,182	16,572	10,385	23,963	55,292	234%	12.80%
Djibouti	32	57	100	77	135	239	138%	9.08%
Dominica	439	877	1,751	1,059	2,114	4,220	141%	9.19%
Dominican Republic	41,859	54,406	70,714	61,201	79,530	103,350	46%	3.87%
Ecuador	20,431	35,608	62,061	44,289	77,226	134,658	117%	8.05%
Egypt	3,587	7,495	15,662	9,840	20,563	42,972	174%	10.62%
El Salvador	7,100	12,654	22,550	17,072	30,422	54,209	140%	9.17%
Equatorial Guinea	221	440	877	563	1,122	2,236	155%	9.81%
Eritrea	181	378	791	344	720	1,508	91%	6.66%
Estonia	349	773	1,712	909	2,013	4,460	160%	10.04%
Ethiopia	556	1,386	3,456	2,035	5,074	12,655	266%	13.86%
Fiji	1,141	1,479	1,916	1,850	2,397	3,106	62%	4.95%
Finland	1,710	3,989	9,307	4,337	10,120	23,615	154%	9.76%
France	67,133	99,365	147,073	118,906	175,994	260,489	77%	5.88%
Gabon	247	436	771	435	770	1,361	76%	5.84%
Gambia, The	127	206	335	201	327	531	58%	4.71%
Georgia	121	282	658	442	1,034	2,415	267%	13.88%
Germany	101,631	135,483	180,613	77,176	102,894	137,181	-24%	-2.71%
Ghana	5,155	7,206	10,072	8,280	11,570	16,167	61%	4.85%
Greece	23,723	30,712	39,759	30,841	39,904	51,630	30%	2.65%
Grenada	805	1,434	2,554	1,619	2,883	5,135	101%	7.23%
Guatemala	10,334	18,445	32,921	23,390	41,746	74,509	126%	8.51%
Guinea	191	282	415	314	464	684	65%	5.11%

Guinea-Bissau	10	20	41	27	54	108	164%	10.19%
Guyana	613	1,075	1,885	1,615	2,832	4,965	163%	10.17%
Haiti	946	1,918	3,892	2,375	4,819	9,779	151%	9.65%
Honduras	7,725	12,455	20,081	14,911	24,042	38,763	93%	6.80%
Hong Kong	24,316	31,598	41,063	17,360	22,550	29,292	-29%	-3.32%
Hungary	6,509	9,771	14,667	12,039	18,067	27,114	85%	6.34%
Iceland	721	842	984	670	782	913	-7%	-.74%
India	7,318	19,366	51,249	30,066	79,562	210,542	311%	15.18%
Indonesia	4,014	8,646	18,625	9,072	19,543	42,100	126%	8.50%
Iran	293	1,306	5,829	2,030	9,059	40,425	593%	21.37%
Iraq	339	1,274	4,792	1,400	5,264	19,792	313%	15.24%
Ireland	25,034	31,969	40,825	29,161	37,240	47,556	16%	1.54%
Israel	61,089	86,797	123,322	94,778	134,647	191,287	55%	4.49%
Italy	53,364	66,443	82,728	50,121	62,408	77,707	-6%	-.62%
Jamaica	16,645	22,520	30,468	24,557	33,223	44,948	48%	3.97%
Japan	57,994	82,049	116,082	66,943	94,709	133,991	15%	1.45%
Jordan	221	809	2,962	1,951	7,144	26,161	783%	24.34%
Kazakhstan	188	421	944	475	1,065	2,387	153%	9.72%
Kenya	4,036	4,999	6,191	5,004	6,194	7,667	24%	2.17%
Kiribati	105	149	212	78	111	157	-26%	-2.95%
Korea, Republic of	17,559	23,807	32,278	25,294	34,287	46,477	44%	3.72%
Kuwait	339	597	1,050	494	868	1,527	45%	3.81%
Kyrgyzstan	42	82	161	97	191	377	133%	8.84%
Laos	34	169	832	234	1,152	5,668	581%	21.15%
Latvia	1,140	2,253	4,450	2,487	4,913	9,708	118%	8.11%
Lebanon	364	1,424	5,571	2,383	9,325	36,490	555%	20.68%
Lesotho	356	585	962	347	571	939	-2%	-.25%
Liberia	179	538	1,613	821	2,462	7,389	358%	16.44%
Libya	57	300	1,572	409	2,143	11,238	615%	21.74%
Lithuania	202	833	3,437	1,368	5,645	23,292	577%	21.09%
Luxembourg	314	434	600	303	419	579	-3%	-.35%
Macao	863	1,268	1,863	662	972	1,428	-23%	-2.62%
Macedonia	305	579	1,098	524	994	1,884	72%	5.55%
Madagascar	550	789	1,131	953	1,366	1,959	73%	5.65%
Malawi	326	484	719	513	761	1,130	57%	4.63%
Malaysia	2,496	5,617	12,639	6,653	14,971	33,688	167%	10.30%
Maldives	90	109	132	79	96	116	-12%	-1.28%
Mali	129	206	329	230	368	588	79%	5.97%
Malta	1,646	2,537	3,909	2,480	3,823	5,891	51%	4.19%
Marshall Islands	384	503	660	297	390	510	-23%	-2.53%

Mauritania	55	111	223	157	318	641	187%	11.11%
Mauritius	403	615	939	670	1,023	1,562	66%	5.22%
Mexico	250,509	467,880	873,870	594,335	1,109,974	2,072,977	137%	9.02%
Micronesia, Fed. Sts.	290	446	686	224	345	530	-23%	-2.54%
Moldova	250	368	542	315	464	684	26%	2.35%
Mongolia	124	207	344	272	452	752	119%	8.14%
*Montenegro**								
Morocco	2,655	6,304	14,965	8,181	19,421	46,105	208%	11.91%
Mozambique	390	484	601	310	384	476	-21%	-2.29%
Namibia	879	1,057	1,270	976	1,173	1,409	11%	1.05%
Nepal	590	885	1,329	891	1,337	2,006	51%	4.21%
Netherlands	15,655	18,825	22,636	20,219	24,312	29,234	29%	2.59%
New Zealand	11,615	16,034	22,134	19,867	27,422	37,849	71%	5.51%
Nicaragua	685	2,478	8,961	3,966	14,340	51,853	479%	19.19%
Niger	83	166	333	167	335	670	101%	7.26%
Nigeria	7,204	11,519	18,416	13,791	22,045	35,242	91%	6.71%
Norway	4,097	10,108	24,939	13,388	33,035	81,515	227%	12.57%
Oman	170	439	1,137	631	1,632	4,221	271%	14.02%
Pakistan	1,815	3,320	6,073	4,579	8,378	15,325	152%	9.70%
Palau								
Panama	7,715	11,771	17,960	12,159	18,551	28,306	58%	4.65%
Papua New Guinea	522	701	940	665	893	1,198	27%	2.45%
Paraguay	509	972	1,854	1,130	2,156	4,113	122%	8.29%
Peru	6,839	15,916	37,040	22,293	51,878	120,727	226%	12.54%
Philippines	48,384	57,181	67,577	57,931	68,449	80,876	20%	1.81%
Poland	6,083	15,944	41,792	21,710	56,909	149,178	257%	13.57%
Portugal	5,284	8,016	12,160	6,482	9,834	14,920	23%	2.07%
Qatar	70	150	321	142	302	646	101%	7.25%
Romania	3,876	6,069	9,501	7,294	11,418	17,875	88%	6.52%
Russia	2,801	6,823	16,622	8,186	19,943	48,586	192%	11.32%
Rwanda	115	237	489	292	603	1,246	155%	9.80%
Samoa	513	737	1,058	347	498	715	-32%	-3.84%
Sao Tome and Principe	17	28	45	34	54	87	93%	6.79%
Saudi Arabia	3,970	5,094	6,535	3,452	4,428	5,679	-13%	-1.39%
Senegal	266	575	1,242	786	1,698	3,667	195%	11.43%
Serbia	568	863	1,311	1,147	1,743	2,650	102%	7.29%
Seychelles	142	225	356	147	232	367	3%	.31%
Sierra Leone	115	330	947	520	1,491	4,272	351%	16.27%
Singapore	5,791	6,854	8,114	6,625	7,840	9,278	14%	1.35%

Slovak Republic	321	1,063	3,521	1,588	5,260	17,421	395%	17.34%
Slovenia	1,335	2,281	3,897	2,161	3,693	6,312	62%	4.94%
Solomon Islands	44	54	68	33	41	51	-24%	-2.73%
Somalia	60	211	740	247	866	3,038	311%	15.17%
South Africa	5,649	8,491	12,761	10,505	15,787	23,724	86%	6.40%
Spain	21,485	27,807	35,989	31,650	40,960	53,010	47%	3.95%
Sri Lanka	4,033	5,410	7,257	4,296	5,761	7,726	6%	.63%
St. Kitts & Nevis	707	1,178	1,963	1,291	2,152	3,587	83%	6.21%
St. Lucia	690	1,395	2,821	1,594	3,222	6,514	131%	8.73%
St. Vincent & Grenadines	205	354	611	440	760	1,313	115%	7.95%
Sudan	151	387	993	499	1,279	3,281	230%	12.70%
Suriname	352	529	793	616	924	1,386	75%	5.74%
Swaziland	256	402	629	405	635	994	58%	4.68%
Sweden	5,665	7,586	10,158	5,320	7,126	9,545	-6%	-.62%
Switzerland	28,667	38,680	52,191	23,735	32,035	43,238	-17%	-1.87%
Syria	488	1,175	2,830	1,840	4,432	10,674	277%	14.19%
Taiwan	8,355	21,713	56,427	31,788	82,598	214,623	280%	14.29%
Tajikistan	27	80	239	131	387	1,149	382%	17.02%
Tanzania	589	1,002	1,704	1,263	2,149	3,657	115%	7.93%
Thailand	9,922	15,657	24,707	19,338	30,516	48,155	95%	6.90%
Timor-Leste				8	18	40		
Togo	214	363	614	388	657	1,113	81%	6.12%
Tonga	263	411	640	516	805	1,256	96%	6.96%
Trinidad &Tobago	4,238	6,286	9,323	7,315	10,850	16,094	73%	5.61%
Tunisia	880	2,273	5,870	2,437	6,294	16,260	177%	10.72%
Turkey	8,780	13,900	22,005	15,741	24,933	39,495	79%	6.02%
Turkmenistan	90	135	202	130	195	293	44%	3.75%
Uganda	348	676	1,314	926	1,801	3,502	167%	10.30%
Ukraine	1,439	3,169	6,980	3,947	8,693	19,149	174%	10.62%
United Arab Emirates	650	1,381	2,932	1,195	2,539	5,392	84%	6.28%
United Kingdom	243,778	319,218	418,005	168,937	221,118	289,417	-31%	-3.61%
Uruguay	598	1,783	5,320	2,390	7,130	21,270	300%	14.86%
Uzbekistan	156	327	681	436	910	1,900	179%	10.80%
Vanuatu	43	69	112	50	81	130	16%	1.51%
Venezuela	5,291	15,121	43,218	15,886	45,415	129,831	200%	11.62%
Vietnam	638	2,788	12,186	5,358	23,420	102,362	740%	23.72%
Yemen	838	1,444	2,487	1,085	1,869	3,221	30%	2.62%
Zambia	276	451	737	515	842	1,377	87%	6.45%
Zimbabwe	769	1,011	1,328	525	689	906	-32%	-3.75%

Appendix B: Using Capture-Recapture Techniques to Estimate the Number of Overseas U.S. Citizens

Motivation of the Capture-Recapture Approach

In its report, “Issues of Counting Americans Overseas in Future Censuses,” the Census Bureau explored how best to count the number of overseas citizens.[11] It concluded:

> At this time, the Census Bureau cannot estimate accurately the size of the universe of the overseas population or the specified components other than the federally affiliated groups. No acceptable tested methodology for providing an independent measure of the coverage of that population is available. We need to conduct extensive research and development work to see if we can develop an estimate that would meet our quality standards.

Direct enumeration of U.S. citizens was deemed too expensive. The Census Bureau had considered the use of foreign government and U.S. administrative records, and indicated there were several limitations even with those approaches as well:

- Each potential administrative records source has coverage, accuracy, and access issues.
- None of these sources by themselves would give a complete, reliable estimate of the size of the Americans overseas universe. Some sources would provide information only for specific components of this population (military, college students, missionaries, those residing in specific countries, and so forth), while others may have broader coverage.
- The likelihood of some degree of duplication between these sources is great.

Many of these limitations, however, are based on an “enumeration-minded” approach: that the records, once linked and de-duped, would represent a complete listing of Americans abroad.

Overview of Capture-Recapture

Here the use of a statistical technique known as capture-recapture is considered as a method to estimate the number of overseas Americans. Rather than attempting to create a complete enumeration of all individuals abroad, if the administrative records sources are treated as samples in a capture-recapture approach, estimates of the population of U.S. citizens abroad can be made. These estimates have the advantage of being independent of any FGEs. Using this technique could also avoid many of the limitations mentioned by the Census Bureau.

Capture-recapture statistical methods are used to estimate the size of a population given samples from that population. In its simplest form, this technique is commonly used to estimate the wildlife

[11]U.S. Census Bureau, September 27, 2001.

population in a given environment using two samples: a first sample of wildlife is captured, tagged, released, and then a second sample is taken weeks later. The overlap between the two samples is determined and allows for estimation of the size of the population. This two-sample approach requires independence between the two samples.

With a larger number of samples, there is more flexibility in the assumption of independence. The technique can be applied to estimating population sizes in many applications. Several practical examples follow, taken from “Discrete Multivariate Analysis: Theory and Practice” [Bishop, Fienberg and Holland, 1975]:[12]

- Estimating the number of children in Massachusetts possessing a specific congenital anomaly. Five sources or lists of names of such children were available, and there were some clear relationships (or dependencies) among the lists.
- Estimating the number of volunteer organizations in small cities and towns in Massachusetts. In each city or town there were three techniques [“samples”] used to identify individual volunteer organizations [the samples were linked to determine the overlap and estimate the population].
- Estimating the number of drug addicts in the United States. There are five different Federal agencies which have registries of drug addicts. Most of the individuals whose names appear in these registries have had their names recorded because of crime-related activities stemming from their involvement with narcotics. The five-sample version of the techniques described here can yield an estimate of the size of the drug addict population which will probably exclude those individuals who have an extremely small probability of being apprehended, either for a narcotics offense or for a criminal activity necessitated by the monetary demands of addiction.
- Estimating the number of crimes committed in a given area. Crime reports are collected by the local, state, and Federal police groups. Not all crimes reported are recorded by any one police group. In addition, several local community agencies receive information about neighborhood crimes.

In the above examples, populations are estimated using lists, or groups, as the samples, and typically the groups are not completely independent. The lists share the characteristic that they are a

[12]See also Markham, Falk, and Scheuren (2013) for a recent application of the methodology for estimating nonresponse bias in surveys.

subset of the same population, but they may overlap in ways that suggest relationships (nonindependence). For example, the group of U.S. Social Security beneficiaries in a given foreign country may be expected to overlap significantly with the group of U.S. taxpayers, but minimally with the group of U.S. students studying in that country.

To estimate the number of overseas U.S. citizens, existing administrative records sources can be used as the input lists or groups. One list would be of overseas Social Security recipients, maintained by the Social Security Administration. Another potential list would be the filers of foreign income (Form 2555) maintained by the IRS. State records of Americans requesting absentee ballots through SF-76 forms constitute another potential administrative records source. Private entity sources of data also exist, maintained by professional organizations, universities, religious groups, etc. By determining the size of these lists and the overlap (through record linkage), population totals can be estimated.

The remainder of this supplemental report briefly outlines the capture-recapture methodology as applied to the task of estimating overseas U.S. citizens. Using the limited data available, some regional and country-level estimates are provided, but with limitations, and with wide credibility intervals,[13] representing a high level of uncertainty. This report serves primarily as the proof of concept of a promising approach. Successfully applying this approach only requires access to currently available administrative records. Not enough data could be obtained for this effort to fully apply the capture-recapture approach. Although some group totals were available, access to the actual administrative records data sources was not available, and any material to provide an idea of the overlap between the different sources was not available. The U.S. Government should be capable of coming up with a good estimate (at least better than any derived so far), but it requires multiple data sources being placed on the same server for record linkage.

Three-Sample Illustration: Estimating the Total Number of U.S. Citizens Abroad

The illustration of capture-recapture methodology begins with an overall estimate of overseas U.S. citizens using three sources as input:

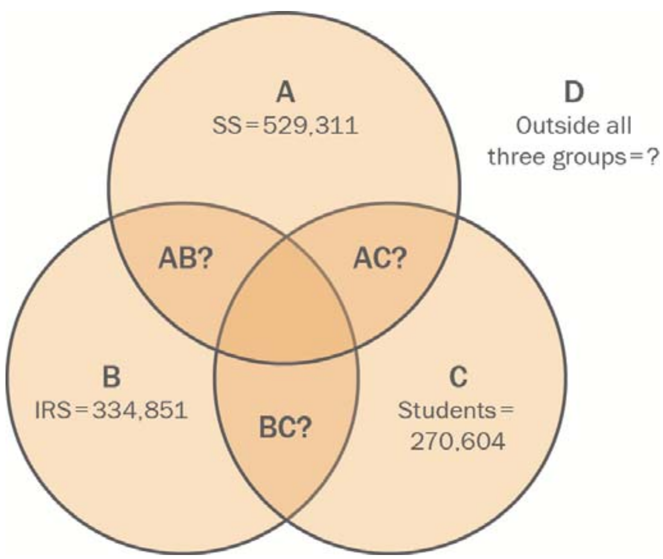
529,311: Total number of Social Security Beneficiaries overseas in 2010, published by the Social Security Administration. This data is available online with regional totals and limited by-country data (countries with over 500 beneficiaries).

[13]Given the reservations about some of the input data, the term "credibility interval" is used here in place of "confidence interval."

334,851: Total estimated number of tax returns with an accompanying Form 2555 (overseas income) for 2006, published by the Statistics of Income Division at the IRS. This number suffers from a few limitations: it is four years older than the other sources, it represents an estimate derived from a random sample of returns, and this is a count of returns, not individuals. This data, like that from the SSA, is available by region and country—but the list of countries for which data is published is a little different here than for the Social Security data. And the definitions of the regions differ as well.

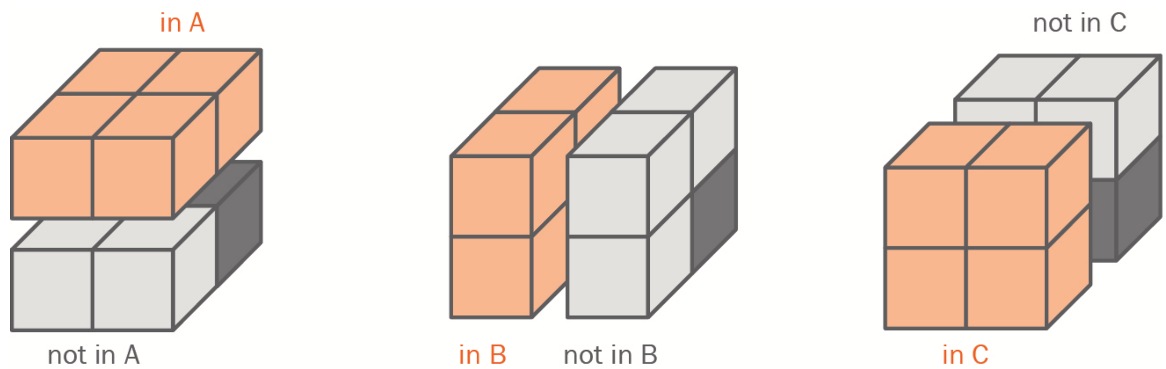
270,604: Total number of students studying abroad in 2009/2010 according to the Institute of International Education. The data is available by country and by region. Again, the regional definitions here differ from those above, but data are available for all countries, not just a selection.

The following diagram shows the data in a Venn diagram, illustrating the additional values that could be obtained through record linkage.



The three groups do not have to be independent samples. The degree of dependence or independence between the different groups in the capture-recapture approach will simply affect the variance in the final estimates. What is needed to estimate the total population is information on the overlap between the groups (the values ab, ac, and bc). Then the exclusion (Group D: the population outside the three groups) can be estimated using capture-recapture techniques, appropriately adjusting for any dependence between groups.

Another representation of the three-sample approach is given by a 2 × 2 × 2 table, with each dimension identified through a dichotomous variable indicating existence of an individual in a cell:



Each diagram shows how the dimension is represented by existence or nonexistence in a group. The shaded cell is the same as Group D in the Venn diagram—this is the missing cell—the population not in any group/sample. This tabular representation lends itself to a log-linear modeling, where the partial independence of the groups can be tested. The expected values under the degree of independence found and accompanying models of interaction can be calculated.

Assuming access to complete databases of individuals and good linking variables are available, the overlap can be determined quite accurately through record linkage. Alternatively, through a sample of data or a dedicated study, an estimate of the overlap with some degree of uncertainty can be developed. With less information, more uncertainty in the overlap is modeled, leading to greater uncertainty in the totals. What follows is an illustration of possible models of uncertainty for the overlap:

Full-Range Uncertainty	Simulates the greatest degree of uncertainty in the overlap between groups. Here, nothing is known about the overlap between the groups. The range of values for the overlap is from 0 (no overlap) to the size of the smaller group (100% overlap—the larger group subsumes the smaller group).[14]	ab: 0–335K ac: 0–271K bc: 0–271K
Modeled Uncertainty	Simulates some knowledge of the overlap, which could be a confidence interval obtained through a sample or specific study. Alternatively, the uncertainty could be more qualitative than quantitative, represented through a rather broad range of values. Here, the overlap ab is simulated by arbitrarily assuming that 25%–50% of Form 2555 filers are SS beneficiaries. The overlap ac is simulated by assuming 5%–15% of students studying abroad are	ab: 25%–50% ac: 5%–15% bc: 10%–20%

[14]For technical reasons, exactly 0% or 100% is not allowed—this would lead to undefined values in the equations. Instead, small values approaching 0% and 100% are used (for these simulations a minimum of 5 and a maximum of 5 below the group totals).

	SS beneficiaries. And the overlap bc is simulated by assuming 10%–20% of students are Form 2555 filers. Note that this does not represent any truth; these are arbitrary values chosen to illustrate the approach.	
No Uncertainty	This would represent the case where the groups A, B, and C have complete enumerated lists that can be linked. The exact overlap would be known if the administrative records sources can be linked to determine the records shared between multiple sources. There could be linkage error, but this is ignored for now. Again, the values chosen to the right are arbitrary and do not represent what the actual overlaps among the groups might be.	ab: 100K ac: 20K bc: 20K

Detailed Methodology

No matter how many dimensions are in the problem (how many input groups are included), only the two-group interactions will be modeled—here ab, ac, and bc. This assumes that there are no higher-order interactions. This approach can lead to a more stable estimate of the variance and is a common assumption that can be controlled through intelligent selection of the groups to be used.[15] The capture-recapture methodology is analogous to modeling, where each group count and overlap count is a data point and the goal is to estimate one data point: the count of individuals not in *any* group. As with most modeling problems, interaction terms are avoided where possible. With the capture-recapture methodology, two-order interactions are allowed if the model fit would be biased without them, but all three-order and higher terms are set to their maximum likelihood equivalent. If there is a choice of input samples (or groups, lists), then samples can be chosen with the greatest independence so as to improve the precision of the model.

The capture-recapture simulation methodology can be described at a high level here in a series of steps:

1) Obtain the overlap values of ab, ac, and bc. This is done through record linkage, or through sampling from a defined probability distribution. For example, under the "modeled uncertainty" scenario, ab is drawn from a uniform distribution between 25% and 50% of 334,851 (IRS Form 2555 filers). Under the "full-range uncertainty" scenario, ab is drawn from a uniform distribution between 0 and 334,851. The value abc is derived under the assumption there are no high-order interaction terms.
2) There are seven possible relationships in a three-sample scheme, all of which are calculated:

[15]e.g., Sekar and Deming (*Journal of the American Statistical Association,* 1949)

I. All groups are independent. The maximum likelihood estimates of the 7 nonmissing cells in the 2 × 2 × 2 table are determined from the A, B, C, ab, ac, and bc values through raking (described in Bishop, Fienberg, & Holland as iterative proportionate fitting). From these numbers, the estimate of the population and the variance are derived, and the goodness-of-fit statistics chi-square and *G*-square. The number of degrees of freedom is 3.

II. Groups A and B are related, but Group C is independent of the first two. The maximum likelihood estimates are derived from a mix of observed values and estimated values. The variance is larger than in *I*, but there is still a good deal of independence in this case. Again, the estimate of the population and the variance are derived, and the goodness-of-fit statistics chi-square and *G*-square. The number of degrees of freedom is 2.

III. Groups A and C are related. This is the same as *II* with groups B and C reversed.

IV. Groups B and C are related. This is the same as *II* with groups A and C reversed.

V. The only independence is between groups A and C. There is a relationship between A and B, and a relationship between B and C. This case, like *II*, has a closed-form solution, but there is less independence in this model and the variance is increased. The number of degrees of freedom is 1.

VI. Same as *V*, where the only independence is between groups A and B.

VII. Same as *V*, where the only independence is between groups B and C.

VIII. No independence. All groups are related. The maximum likelihood estimates of the expected cell counts are all equal to the observed values. The only independence is the assumption that there are no higher-order interaction terms. This leads to the highest variance. There is no goodness-of-fit test, since the model completely describes the cells in the 2 ×2 × 2 table. There are 0 degrees of freedom.

3) For each of the seven modeled cases, the estimate of the population, the standard error, the chi-square and *G*-square goodness-of-fit statistic and degrees of freedom are compared. A selection among the seven models is made according to the fit statistics, favoring greater independence if the model fits reasonably. The selection according to this criterion is against a .1 significance.
4) Using the chosen model, the distribution of population estimates for the given simulation is given by the point estimate and the variance estimate.
5) Steps 1 through 4 are repeated many times for each of the scenarios requiring simulation of the overlap uncertainty (if exact values for the overlap are known, through record linkage, the simulation isn't necessary). The simulation returns a final distribution of estimates—taking into

account the sampling variance due to the capture-recapture approach plus the variance due to the uncertainty in the overlap.

The methodology above describes a microsimulation model that contains capture-recapture statistical techniques. The capture-recapture theory itself is a fairly straightforward application of equations using as input the group sizes and all pairwise overlap totals. The “microsimulation” part of the methodology is necessary only so that the uncertainty in any input values can be modeled. If all group sizes and overlap totals are known, the capture-recapture equations can be applied directly, without the microsimulation modeling steps. Currently, this microsimulation is necessary because the overlap totals are unknown.

Results

With the data currently available, that is, totals from three subgroups and no information on the overlap, the full-range uncertainty model is used to determine a very wide confidence interval of the total number of U.S. citizens overseas:

SS Beneficiaries:	529,311	
IRS Form 2555 Filers:	334,851	Overlap: No information
Students:	270,604	

➔ Estimated Population: 90% Confidence interval: [678K–6,613K]

This represents a wide range of uncertainty. The uncertainty is due almost entirely to not knowing the value of the overlaps among the three groups. To illustrate this point, below are simulated values for potential values of the overlap:

Simulation assuming limited information on the overlap:

SS Beneficiaries:	529,311	Overlap:	ab: 25%–50%
IRS Form 2555 Filers:	334,851		ac: 5%–15%
Students:	270,604		bc: 10%–20%

➔ Estimated Population: 90% Confidence interval: [2,143K–4,464K]

Simulation assuming exact information on the overlap:

SS Beneficiaries:	529,311
IRS Form 2555 Filers:	334,851
Students:	270,604

Overlap: ab: 100K
ac: 20K
bc: 20K

➔ Estimated Population: 90% Confidence interval: [4,106K–4,193K]

As can be expected, with greater uncertainty in the overlap comes a greater variance in the final estimate. A greater variance is also suffered with less independence between the groups, but this effect is far less important than lacking better estimates for the overlap between groups.

The results of applying the above methodology to all available country data follows (using the three groups: Social Security beneficiaries, IRS Form 2555 filers, and students—and with no knowledge of the overlap). The 90% confidence intervals are provided without point estimates (the intervals are too wide given current data to give any credence to exact point estimates).

Overall	SS Beneficiaries	IRS Form 2555 Returns	Students	90% Credibility Interval	
				Lower	Upper
Total	529,311	334,851	270,604	678,450	6,613,198

This is the Estimate for the Total, as provided above.

North America	SS Beneficiaries	IRS Form 2555 Returns	Students	90% Credibility Interval	
				Lower	Upper
North America Total	157,129	36,179	9,046	169,621	1,526,665
Canada	107,074	30,067	1,750	115,514	962,326
Mexico	49,802	6,112	7,157	53,935	434,473

The estimates for North America as a whole and individually for Canada and Mexico are above. These are believed to be the countries with the largest number of U.S. citizens. Here, the range of estimates for Canada appears to be higher than that for Mexico. This runs contrary to what is believed to be the case. Limitations in the input data can explain some of the discrepancy (this method models a large range of possible overlaps, and it does not know the actual overlap). Note also that the confidence intervals overlap considerably. Because the individual values ab, ac, and bc for each country are not known, the intervals are wide. Also, the ab, ac, and bc values are likely to be different between Canada and Mexico.[16]

[16]If, for example, a U.S. citizen in Mexico was less likely to file an IRS Form 2555 than a U.S. citizen in Canada, there would be small overlaps here. The estimate for Mexico would fall near the high-end (or even over) the 90% credibility interval while the estimate of Canada would stay near the middle.

Europe	SS Beneficiaries	IRS Form 2555 Returns	Students	90% Credibility Interval	
				Lower	Upper
Germany	37,725	21,513	8,551	44,803	435,757
United Kingdom	32,359	28,409	32,683	48,040	435,191
Italy	32,721	5,199	27,940	40,473	421,680
Spain	10,838	2,453	25,411	28,639	255,870
France	12,829	9,653	17,161	22,546	209,474
Greece	23,561	1,484	3,700	25,177	187,475
Ireland	9,376	1,896	6,798	11,391	104,464
Switzerland	7,438	7,093	1,863	9,358	87,592
Portugal	12,451	387	198	12,900	62,467
Netherlands	5,087	3,263	2,369	6,473	54,124
Poland	8,152	735	437	8,606	48,472
Norway	6,940	1,215	440	7,467	45,403
Sweden	4,594	1,399	1,002	5,266	36,560
Austria	2,765	1,361	2,701	3,656	30,581
Czech Republic	749	1,091	3,409	3,922	25,602
Denmark	1,091	1,754	2,228	2,901	21,891
Belgium	1,959	1,881	1,244	2,648	20,718
Hungary	1,997	604	920	2,404	15,693
Finland	922	354	211	1,085	5,478

There is not a capture-recapture estimate for Europe as a whole. This is due to the lack of a consistent definition of what defines "Europe" between the three data sets. And the two Federal data sets, from SSA and IRS, do not include enough individual country data to make one consistent definition between them. But the above table does include several countries from Europe for which all three data sets provide data. The top three countries, Germany, the United Kingdom, and Italy, appear to rival Mexico for the number of Americans abroad. But again, note that the intervals are large, and that the values for overlap totals ab, ac, and bc are likely to differ between Mexico and European countries.

Asia and the Middle East	SS Beneficiaries	IRS Form 2555 Returns	Students	90% Credibility Interval	
				Lower	Upper
Japan	41,874	23,529	6,166	48,825	480,552
China	959	12,430	13,910	17,028	150,793
Philippines	23,044	2,313	238	24,098	134,403
Israel	9,798	8,986	3,146	12,453	119,497

Hong Kong	1,327	10,792	1,196	11,632	78,185
South Korea	2,019	6,668	2,137	7,803	59,571
India	1,550	4,214	3,884	5,419	47,671
Taiwan	665	6,588	850	7,095	44,725
Thailand	3,069	3,643	1,231	4,611	38,547
Turkey	811	1,199	1,522	2,004	14,055
Japan	41,874	23,529	6,166	48,825	480,552
China	959	12,430	13,910	17,028	150,793
Philippines	23,044	2,313	238	24,098	134,403
Israel	9,798	8,986	3,146	12,453	119,497
Hong Kong	1,327	10,792	1,196	11,632	78,185
South Korea	2,019	6,668	2,137	7,803	59,571
India	1,550	4,214	3,884	5,419	47,671
Taiwan	665	6,588	850	7,095	44,725
Thailand	3,069	3,643	1,231	4,611	38,547

Estimates for countries in Asia and the Middle East are combined in the table above. There are several countries missing from the list. The table captures only those countries that exist in all three data sources. The regional total for Asia is not in the table because the definition is not the same for all data sources.

Oceania	SS Beneficiaries	IRS Form 2555 Returns	Students	90% Credibility Interval	
				Lower	Upper
Oceania Total	10,657	9,724	13,566	18,225	165,316
Australia	9,047	6,420	9,962	13,366	123,549
New Zealand	1,328	2,518	3,113	4,000	32,517
Other Oceania	282	787	491	987	5,616

The three data sources appear to agree for the regional definition of Oceania. Therefore, regional and “other” totals are provided here.

Central and South America and the Caribbean	SS Beneficiaries	IRS Form 2555 Returns	Students	90% Credibility Interval	
				Lower	Upper
Costa Rica	4,681	1,662	6,262	7,742	67,247
Dominican Republic	7,322	1,093	1,324	8,056	53,899

Argentina	3,750	751	4,835	5,898	46,568
Brazil	2,132	2,696	3,099	4,188	33,837
Colombia	4,516	1,870	180	5,014	28,603
Chile and Easter Island	1,788	902	3,115	3,783	27,904
Peru	1,351	419	2,316	2,744	18,351
Panama	1,826	1,032	691	2,265	15,176

The above table represents estimates for countries in Central and South America as well as the Caribbean. Several countries are missing because they could not be found in one or more of the data sources. The regional definitions also differ between the three data sources, so they are also not provided.

Africa	SS Beneficiaries	IRS Form 2555 Returns	Students	90% Credibility Interval	
				Lower	Upper
Africa Total	2,693	9,697	14,769	17,696	166,731

The two Federal data sources have no information on individual African countries, so only the regional total is provided above.

Limitations and Further Notes

The estimates provided have several limitations. The data sources do not agree on regional definitions, so not all regions have estimates. Not all data sources have data for all countries (often contained in the catch-all, undefined "other countries" category). And most limiting, no data on the overlap between groups are available, so the confidence intervals are wide.

The IRS Form 2555 return estimates are outdated by four years, they represent returns, not individuals, and the numbers have not been properly represented with a confidence interval (the IRS provides estimates, not the true number of actual returns).

While the capture-recapture idea together with administrative records outlines a promising approach to obtaining better estimates, it would still suffer from some undercount bias. Specifically, there will be an undercount of individuals seeking to live "under the radar" (this is also a limitation suffered by FGEs and any derived models). In capture-recapture terms, this phenomenon is called *trap avoidance*. By the same token, however, it is unlikely that these individuals would take advantage of any voter outreach programs.

Expanding upon the number of groups (the number of administrative records sources) will decrease the undercount bias and increase the possible number of degrees of freedom, leading to even smaller variance. But it will also increase the number of pairwise overlaps that need to be obtained or modeled and could increase the number of linkage errors.

The capture-recapture simulation program as it stands is a tool that provides a range of estimates that reflects the current knowledge of the overseas U.S. population. This range of estimates is very wide, given the current limited information. A huge advantage of the tool, though, is that if better numbers for the overlap between samples are collected, it could improve future estimates.

Specifically, if an analyst were granted access to administrative records for currently available records from SSA and IRS, other Federal agencies, and other data from private sources, determining the overlap among groups could lead to much sharper estimates. Each improved input value will lead to a more accurate output estimate.

The capture-recapture approach is sound, promising, and provides calculable confidence intervals. It also has the advantage of providing these estimates independent of FGEs. U.S. administrative records data currently exist to follow the approach outlined in this supplement. But obtaining access to these data (simultaneously for the same year, for record linkage) is a major obstacle to be overcome. Some of this lack of data can be compensated for through microsimulation modeling, which adds the appropriate “uncertainty penalty” to the confidence intervals. But the current amount of unreliability is large, and the resulting confidence intervals are so wide as to be not very useful.

The alternative estimation methodology used in the main report (employing models based on FGEs of U.S. citizen counts) has broader direct coverage given the limited data currently available. However, as stated there, the U.S. citizen counts are not comparable from country to country and have their own definitional and timing issues. A combined approach might afford more robustness than either alone, and warrants further study.

Appendix C: Attempted Modeling Strategies

In addition to model averaging using cross-validated based weights, three other estimation methodologies were considered. One model estimation approach considered was Bayesian model averaging (BMA), a model averaging routine very similar to the preferred method, but one that uses an alternative model weighting scheme based on the Bayesian information criterion (BIC). The other two model estimation approaches considered were random forests and additive regression imputation. Random forests is a machine learning algorithm that uses heuristic rules to search the model space in a manner that is potentially more efficient than the model averaging methods. Additive regression, by contrast, is similar to a generalized linear model, save for it fits some function of each predictor to the data in predicting the outcome. This Appendix briefly describes these three alternative methods, and explains the procedure used to settle on a final methodology.

- *Bayesian Model Averaging* is a method of deriving parameter estimates by creating a weighted average of parameters and/or predictions from a set of possible models, where the weight is typically a function of the probability of observing the dependent variable given a model, or model likelihood (Montgomery and Nyhan, 2010). This measure of model likelihood reflects how well the model fits data. A critical difference between this report's methodology and BMA is that the measure of fitness in BMA is typically based on in-sample fit, rather than explicitly testing how well the models predict observations that were not used to calibrate the model. A traditionally popular choice of metrics used to generate model weights in BMA is the BIC, where the BIC can be written as:

    - $BIC_m = -2\ln(L_m) + k_m * \ln(p)$

  where *L* is the likelihood, or fitness of model *m*, *k* is the number of parameters in model *m*, and *p* is the number of observations. Higher values of the BIC correspond to a lower model fitness, and BIC-based weights are inversely related to the value of the BIC. Note that as the number of parameters increases, the BIC increases, and the model weight declines. Given that additional parameters that do not increase model fitness may lead to overfitting, the BIC in theory mitigates problems related to overfitting. In addition, models that have many parameters may be expected to produce predictions highly correlated with predictions from models that use some subset parameters. Consequently, a BIC-based weight may also punish model redundancy, similar to the correlation-based component of the model weights in the preferred method.

  The implementation of BMA considered here uses weights based on BIC that take the following form:

$$w_k = \frac{1/BIC_k}{\sum_{j=1}^{N} 1/BIC_j}$$

Unlike Burnham and Anderson (2004) and Montgomery and Nyhan (2010), the anti-log of the BIC was not taken because, in practice, the resulting numerators and denominators were too small for the software to process. In practice, the variant of the BIC weight would be expected to lead to greater equalities in weights across all models than would be the case if the BIC were subject to an anti-log transformation. To account for nonindependence in the observations, the number of countries (79) is used to calculate BIC rather than the number of country-years.

- *Random Forests Imputation* is a nonparametric, regression-tree based ensemble method that imputes missing values for all missing data. The random forest imputation procedure is sequential, imputing values for each variable with missing values in turn as the algorithm can have only one dependent variable at a time. The random forest imputation procedure then proceeds by imputing plausible values into all missing data points (often the mean [continuous variables] or mode [categorical variables]). A random forest algorithm (Breiman, 2001) is then run on the observed values of each variable in the data set with missing data. Random forests are recursive partitioning algorithms in which the data are divided into subsets based on splits defined by predictor variables that optimally predict the outcome. The result of a single recursive partitioning estimation run is a "tree" of splits or "decision points" that define the subgroups that optimally predict the outcome. The random forest algorithm computes many (i.e., hundreds or thousands) of individual trees with very low predictive quality standards (hence, "grows a random forest"). However, predicted values are derived as a weighted average (or modal category) of all the trees and, perhaps counterintuitively, usually constitutes a better prediction than a predictive algorithm with more stringent predictive quality standards. Based on the random forest results, predicted values are imputed for the missing values for each variable. The difference between the newly imputed and previously imputed values is assessed. If a stopping criterion is met based on the difference between the new and old imputed results, the algorithm stops; otherwise, the random forests imputation procedure continues (Stekhoven & Buehlmann, 2012).

- *Additive Regression with Observation Matching* is a nonparametric methodology imputing values into variables based on nonlinear functions of all other observed variables.

Specifically, the additive regression with observation matching proceeds in two steps. First, for each variable with missing values in turn, the observed values on the focal dependent variable are used in an additive regression onto the observed values for all other variables in the data set; the process is also bootstrapped—obtaining subsets of observations with replacement from the data to fit additive regression functions. Additive regression is a method whereby each input variable is allowed to vary in its functional form and is fit using regression splines yet still producing functions for each variable that are independent of the other input variables (i.e., no interactions "built into" estimates; e.g., Stone, 1985) using a process known as "backfitting" (see Buja, Hastie, & Tibshirani, 1989). The ideal functional form obtained through a series of cross-validations and bootstrap samples. Second, values for missing data are then "donated" or imputed from the most similar observed values' predicted value based on the additive regression or from a weighted combination of several predicted values (e.g., Abrahantes, Sotto, Molenberghs, Vromman, & Bierinckx, 2011).

In determining which of these approaches to take, the primary interest was in how the resulting models, or model averages, predicted country-years for which FGEs were unavailable. Consequently, each method was subjected to five-fold cross-validation, where each country-year in the full sample of observations used to calibrate the core model was randomly assigned to one of five groups. Each method is then executed five times, with the models calibrated using all observations from four of the groups and none from one of the five groups. The fitness metric for each of these runs is based on the Root Mean Squared Errors (RMSE) and squared correlation coefficient (R^2) for the excluded group. A higher RMSE corresponds to a worse fit. A higher R^2 corresponds to a better fit. The observations are randomly assigned to five mutually exclusive groups ten times, for a total of fifty different groups and runs. The mean of the RMSE across all fifty runs is used to assess model performance. Only a random 10% of the model space is used in the EMA and BMA methods in order to conserve computational resources. Trial runs revealed that there was little difference in the point estimates when using a random subset of models versus using the entire model space. The predictor set for each method includes all the measurement, administrative records, and theoretical variables, with the EMA and BMA methods including the measurement and administrative records variables in all models. All observations are given the same weight.

Method Validation

Method	Mean RMSE	Mean Pseudo R^2
EMA	22,976.15	.89
BMA	23,048.9	.89

Random Forests	30,156.82	.83
Additive Regression	37,467.75	.70

The EMA and BMA estimates both have a substantially better out-of-sample fit, as measured by the mean RMSE and R^2 across all folds, than the two nonparametric methodologies. This implies that the random forests and additive regression were both overfitting the data. Note, however, that with the exception of the additive regression, all methods have respectable out-of-sample performance as indicated by the R^2. This potentially speaks to the quality of the predictor variables with respect to their ability to predict the FGEs. Between the model averaging methodologies, the EMA performs slightly better on the mean MSE metric than the BMA, but has approximately equal performance on the R^2 metric. However, they are both quite similar with respect to both metrics. Despite their similar performance, given the difficulties with specifying the "correct" anti-logged BIC weight discussed above, and the fact that the cross-validated model weights used by EMA directly test for model overfitting and correlation, the EMA approach was preferred.

Feedback on the preferred Ensemble Model Averaging was provided by several external reviewers. We would like to thank Jason Schachter of the United Nations Economic Commission for Europe; Melissa Scopilitti of the United States Census Bureau's Net International Migration Branch; Kirsten West of the United States Census Bureau's Methodology, Research, and Development Branch; and Dr. Jacob Montgomery of the Department of Political Science at Washington University in St. Louis.