Comparing Methods of Inferring Population Size from Incomplete Travel Data

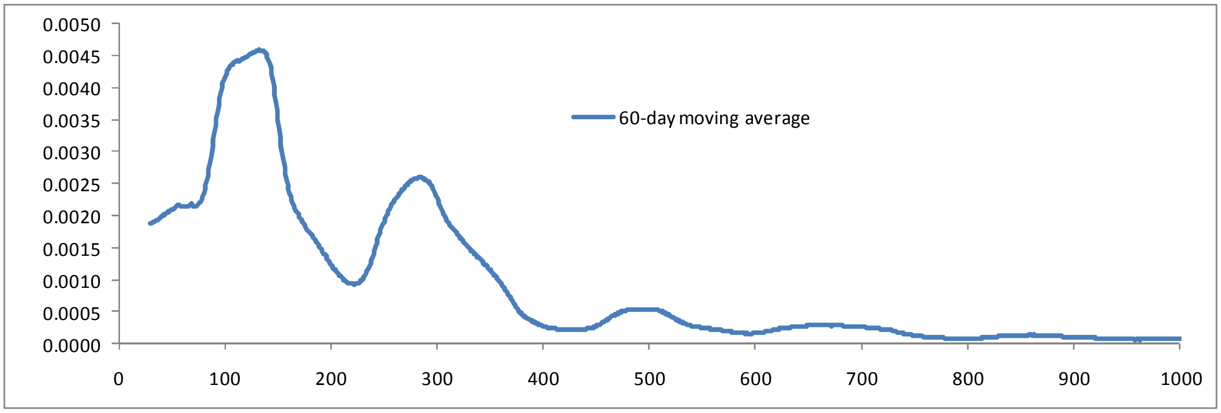
Introduction

Complete counts or direct estimates of the size of the resident nonimmigrant population in the United States are not readily available from surveys or other data sources. The U.S. Department of Homeland Security estimates the average population size during a given period of time indirectly using travel data, including arrival records and the average length of time nonimmigrants tend to stay in the U.S. before returning abroad. Assuming that the entire visit length probability mass function (PMF) is known, instead of simply the average visit length, the expected value of the population size during the given time period can be calculated exactly. Approximate methods may remain desirable, however, because calculating the exact expected value is computationally intensive (i.e., slow), requiring 1,000's of calculations, date comparisons, and lookup-table references for many of the more than 30 million arrivals from 2001-2010. This paper compares estimates and CPU times for several methods that use arrival dates and a known visit length distribution to estimate the average population size in 2010. The methods include calculating the exact expected value, methods previously used by DHS, approximating the PMF with a gamma model, and simulation. For illustrative purposes, results are presented separately for Chinese students.

Data

The user-level data consist of resident nonimmigrant arrival records from fiscal years 2001-2010 and a visit length PMF for each category of nonimmigrant (see Figure 1). Arrival records include the arrival date, class of admission (e.g., student or temporary worker), country of citizenship, and other demographic variables. Nonimmigrant categories are defined by class of admission and country of citizenship. The PMFs are generated from samples of base-level data, but are treated as error-free (see Appendix I) and are assumed to be determined entirely by the nonimmigrant category.

Figure 1.
Visit Length PMF for Chinese Students



Source: U.S. Department of Homeland Security.

Note: The distribution has been smoothed for illustrative purposes.

Methods

All of the methods begin by estimating the number of visit days that occur during the given time period separately for each arrival. The daily average population size is then estimated as the sum of the visit days across all arrivals, divided by the total number of days in the time period. Because CPU time will vary by computer, it is reported as a percentage of a baseline instead of in actual time units. The CPU time required for building the lookup tables (given that the PMF is known) was negligible and is excluded from the reported CPU times.

The Exact Method

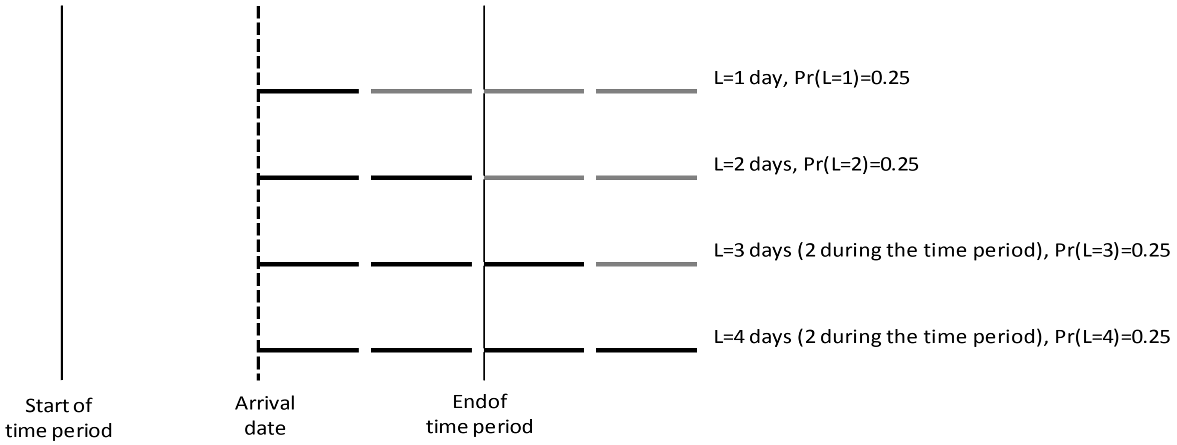
The Exact Method calculates the exact expected number of visit days for each arrival in 2001-2010, adds up the expected number of visit days across all arrivals, and divides by the number of days in 2010. The exact expected number of visit days for a given arrival is calculated in the usual way for a function of a random variable (see the inner sum in the equation below).

The estimate, S_{PMF}, is given by:

$$S_{PMF} = \frac{1}{\| T \|} \sum_{i} \sum_{l=1}^{\max_c} \left\| [A_i, A_i + l] \cap T \right\| \times \Pr\left(L = l \mid C = c\right),$$

where A_i is the arrival date for the ith arrival, l is visit length, $\max_c$ is the maximum visit length with nonzero probability for category c, T is the time period (fiscal year 2010), ||.|| is a metric denoting the number of days in the encapsulated time period, and the outer sum is taken over all arrivals from 2001 through 2010.

Figure 2:
Example: The Expected Number of Visit Days for a Given Arrival.
For simplicity, the visit length PMF is uniform with support = {1, 2, 3, 4}.


L=1 day, Pr(L=1)=0.25
L=2 days, Pr(L=2)=0.25
L=3 days (2 during the time period), Pr(L=3)=0.25
L=4 days (2 during the time period), Pr(L=4)=0.25
Start of time period
Arrival date
Endof time period

Expected number of visit days = 1*0.25 + 2*0.25 + 2*0.25 + 2*0.25 = 1.75 visit days

It was not necessary to calculate the inner sum for arrivals occurring early enough that $A_i + max_c$ occurred before the start of the time period, since the inner sum for all such arrivals is zero. The algorithm took this into account by checking $A_i + max_c$ for each arrival prior to initiating the inner summation loop.

The algorithm was further modified by using a cumulative distribution function (CDF) lookup table to take advantage of the fact that for any arrival, i, $\|[A_i, A_i + l] \cap T\|$ is the same value for all l such that $A_i + l$ occurs after the end date of time period T (see L=3 and L=4 in Figure 2). Because max_c is larger than 3,000 for some categories, this step eliminated many iterations of the inner summation loop and reduced the CPU time by a factor of 6. The time required to build the CDF lookup table was negligible.

The Exact Method yielded estimates of 1.88 million overall and 92,000 for Chinese students (see Table 1). The results of this method are the baselines against which the results of the other methods are compared. Approximate methods are considered "accurate" if they obtain similar estimates to the Exact Method, and are considered "fast" if they produced estimates much more quickly than the Exact Method.

Table 1.
Average Daily Population Size for Fiscal Year 2010: Estimates and CPU times.

	Overall			Chinese students		
	Population estimate			Population estimate		
Method	Number	Percent of baseline	CPU time	Number	Percent of baseline	CPU time
Exact method (baseline)	1,880,000	100%	100%	92,000	100%	100%
Product method	2,090,000	111%	0%	128,000	139%	2%
Forced average method	1,900,000	101%	1%	106,000	115%	1%
Gamma method	1,900,000	101%	57%	100,000	109%	49%
Simulation method (1 iteration)				92,000	100%	2%
Simulation method (10 iterations)				92,000	100%	6%
Simulation method (100 iterations)				92,000	100%	46%
Simulation method (1,000 iterations)				92,000	100%	448%

Source: U.S. Department of Homeland Security.

Product Method

The product method estimates the number of visit days by multiplying the average visit length by the number of arrivals in the given time period.

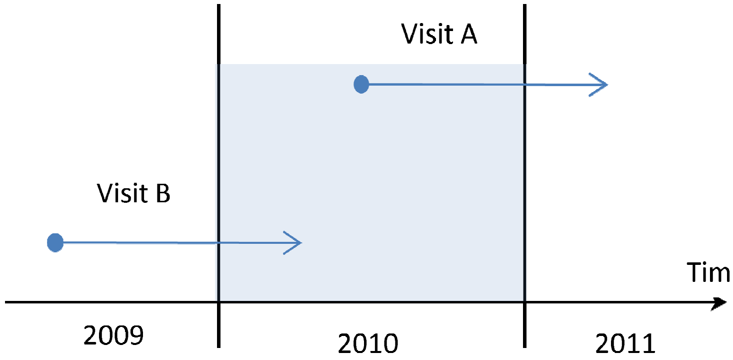
The estimated average daily population size, S_{PM}, during 2010 is given by:

$$S_{PM} = \left(\frac{1}{\|T\|}\right)\sum_{c \in C}\left(N_{c,2010} \times \mu_c\right),$$

where c is the nonimmigrant category, C is the set of all resident nonimmigrant categories, $N_{c,2010}$ is the number of arrivals in category c in 2010, μ_c is the average visit length for category c, and $\|T\|$ is the number of days in the time period (365 days in 2010).

Conceptually, visits originating in the time period (e.g., Visit A in Figure 3) contribute μ visit days to the total, regardless of whether or not all μ days occur during the time period. Visits that originate in other time periods (e.g., Visit B in Figure 3) do not contribute anything to the total.

Figure 3.



The Product Method yielded estimates of 2.09 million overall and 122,000 for Chinese students, respectively 11 and 39 percent larger than the estimates produced by the (baseline) Exact Method. The CPU time required was less than 1 percent of the baseline time.

The substantial overestimation compared to the baseline is partly explained by an overweighting of the portions of visits that extend beyond the end of the time period. Counting those portions at full value is akin to assuming that for each visit extending past the end of time period (e.g., Visit A in Figure X), there was exactly one arrival at the same point in the previous time period (e.g., Visit B in Figure X), and then counting the visit days that occur during the time period for both visits. However, there were fewer arrivals in the previous time period (see Table 2), meaning that there was, on average, less than one Visit B for every Visit A, thereby explaining some of the overestimation.

Table 2.
Resident Nonimmigrant Arrival Counts by Year of Arrival.

	Arrival counts (thousands)	
Arrival year	All categories	Chinese students
2001	2,890	31
2002	2,757	33
2003	2,894	27
2004	2,789	27
2005	2,446	28
2006	3,113	41
2007	3,449	59
2008	3,486	81
2009	3,191	115
2010	3,584	160

The overestimation described above tends to zero as average visit length decreases toward 1 day. Thus the Product Method should much more accurately estimate the average population size of *nonresident* nonimmigrants (e.g., tourists, business travelers, and alien crewmen of foreign airlines), who tend to make short visits relative to resident nonimmigrants. DHS used the product method to estimate the size of the nonimmigrant population in 2004 and the size of the resident nonimmigrant component of certain populations in 2005-08 (Grieco, 2006; Hoefer et al, 2006-09).

Forced Average Method

Like the product method, the forced average method treats all visits as being of average length. Unlike the product method, the visit length is used to calculate a departure date for all arrivals from 2001-2010, and only visit days that occur during 2010 count toward the total.

The population size estimate, S_{FAM}, is given by:

$$S_{FAM} = \left(\frac{1}{\| T \|} \right) \sum_i \left\| [A_i, A_i + \mu] \cap T \right\|,$$

where A is the arrival date, μ is the average visit length, T is the time period (2010), ||T|| is the number of dates in time period T, and the sum is taken over all arrivals (2001-2010). To reduce the number of unnecessary operations, the intersection and day counting metric are omitted whenever $A_i+\mu$ is a date occurring before the start of the time period.

The Forced Average Method yielded estimates of 1.90 million overall and 102,000 for Chinese students, about 1 percent and 15 percent larger than the respective baseline estimates. Processing took about 1 percent of the baseline time. DHS has used the forced average method to estimate the size of the resident nonimmigrant population in 2008 and the size of the resident nonimmigrant component of the lawfully resident foreign-born population in 2009-10 (Baker, 2010; Hoefer et al, 2010 and 2011).

Gamma Method

The gamma method produced probabilities for each possible visit length similarly to the Exact Method, except that the probabilities came from a gamma distribution with parameters chosen to fit the PMF instead of coming directly from the PMF itself. The parameters were chosen by the eyeball method, by varying the shape parameter and setting the scale parameter to the mean of the PMF divided by the shape parameter. The estimates were originally only produced for the Chinese student category, but were then repeated for the other categories by holding the shape parameter at the value chosen to fit the Chinese student PMF and calculating the scale parameter as above, thereby allowing a rough overall estimate.

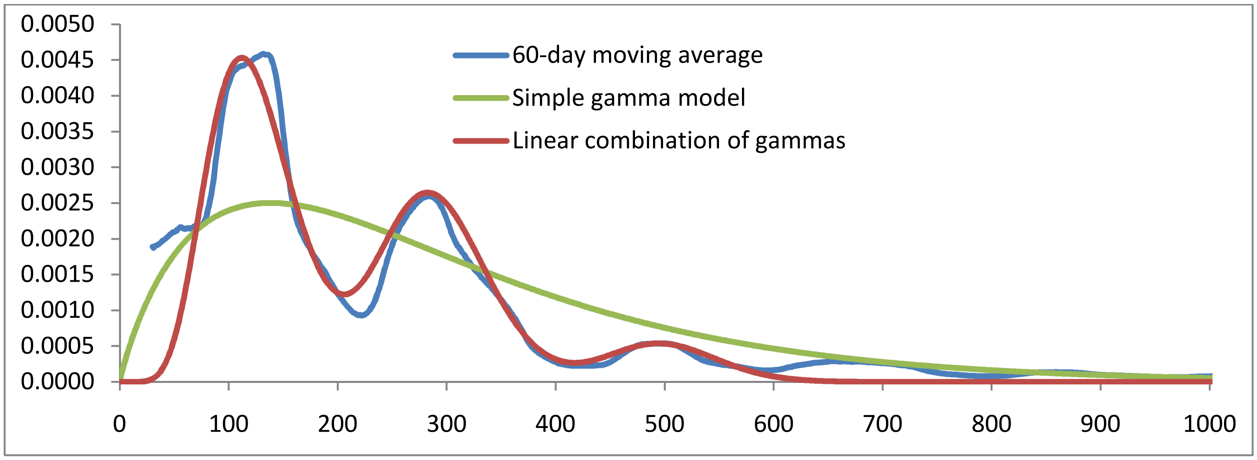
The estimate, S_{Gamma}, is given by:

$$S_{Gamma} = \frac{1}{\| T \|} \sum_{i} \sum_{l=1}^{\max_c} \left\| \left[A_i, A_i + l \right] \cap T \right\| \times \Pr\left(l - 1 < L \le l \mid C = c \right).$$

The Gamma Method produced estimates of 1.90 million overall and 100,000 for Chinese students, overestimating the baseline estimates by 1 and 9 percent, respectively. Processing by the CPU took about 50-60 percent as long as the baseline.

The estimates may have been improved by using numerical methods to estimate the shape parameter, which would also affect the CPU time. However, the fit would still have been crude for Chinese students, considering the multimodal nature of the PMF (see Figure 3). A more complex model, e.g., a linear combination of gamma distributions, could provide a better fit. Improving the model might also defeat the purpose of finding a fast way to produce estimates, given that the PMF is already thoroughly described by the observed data.

Figure 4.
Visit Length Probability Mass Function for Chinese Students:



Source: U.S. Department of Homeland Security.

Simulation Method

The Simulation Method produces estimates by taking the average of multiple simulations. In each simulation, a visit length is randomly drawn from the support of the PMF for each arrival, and the probability of selection is determined by the PMF. The Simulation Method was repeated for 1, 10, 100, and 1,000 simulations.

The estimated daily average population size in 2010 using K simulations is given by:

$$S_{Sim,K} = \left(\frac{1}{K \times \|T\|}\right) \sum_{k=1}^{K} \sum_{i} \left\| [A_i, A_i + l_c] \cap T \right\|,$$

where K is the number of iterations, A_i is the arrival date of the i^{th} arrival in 2001-2010, c is the category of the i^{th} arrival, l_c is the visit length randomly selected from the PMF for category c, and the inner sum is taken over all arrivals from 2001-2010.

The method produced an estimate of about 92,000 for Chinese students for each of the 4 runs, essentially identical to the estimate produced by the Exact Method. The CPU time was competitive with the product method when run with only one iteration (2 percent of the time required by the Exact Method). When run with 10, 100, and 1,000 simulations, CPU processing took 6, 46, and 448 percent as long as the Exact Method. An overall estimate was not produced.

Discussion

The Exact Method calculates the exact expected population size during the given period from arrival data and known PMFs. The CPU time required is high relative to the other methods, however, reflecting the relatively large number of operations. Even with modifications to reduce the number of unnecessary operations, the method must reference the PMF lookup table and calculate the intersection of the visit and the period of interest up to 366 times for each of the roughly 31 million arrivals before calculating the grand total.

The Product Method is computationally simple and more than 100 times faster than the Exact Method. It references only one lookup table value per arrival (the average visit length for the category), doesn't reference any lookup table values for 2001-2009 arrivals, and does not calculate the intersection of the visit with the time period. The method works by assuming that μ visit days are contributed to the total for each arrival during the time period. This is not an unreasonable assumption if visits tend to be short or if the arrival frequency distribution is identical from time period to time period. In fact, the Product Method would yield the same estimate as the Exact Method if the arrival distributions and counts were identical from one time period to the next, and the two methods are asymptotically equivalent as visit length approaches 1 day. The accuracy depends on the extent to which those conditions are met.

The Forced Average Method also references only one lookup table value per arrival, but must do so for all arrivals, and must also calculate the intersection of the visit with the time period for each arrival. It is more accurate than the Product Method, and, while not quite as fast, is still more than 100 times faster than the Exact Method. Like the Product Method, the Forced Average Method yields the same estimate as the Exact Method if the arrival distribution and count is identical from one time period to the next, and is asymptotically equivalent to the Exact Method as visit length approaches 1 day.

The Gamma Method is similar to the Exact Method, except that the visit length probabilities come from a gamma distribution with estimated parameters instead of from the PMF lookup table. The distribution offered only a crude fit for Chinese student visit lengths, and was not analyzed for fitness for the other categories. Considering that the method ran only twice as fast as the Exact Method, efforts to improve the method (linear combination of gammas, linear approximations of the parameter estimates, etc…) may not be justified.

The Simulation Method references only one lookup table value per arrival, but does so once for each iteration. Like the Forced Average Method, it calculates the intersection of the visit with the time period. Unlike the Forced Average Method, referencing the visit length requires the extra step of generating a random number. The Simulation Method runs about 50 times faster than the Exact Method with 1 iteration, and about twice as fast with 100 iterations. Because of the large number of arrivals, accurate estimates were obtained with only a single iteration. The estimates may be improved, within-category estimates in particular, by allowing the number of iterations to vary inversely with the number of arrivals from each category.

If the population is to be tabulated according to one or more auxiliary variables, then tracking which individuals contribute to the total is important. Since the PMFs are skewed right, the methods that rely on average visit lengths tend to overweight relatively recent arrivals at the expense of earlier arrivals, making the Product Method and Forced Average Method less appropriate than the Exact Method or the Simulation Method.

Conclusions

All methods require that the PMFs are known. Without further assumptions, the Exact Method is the only method that actually calculates the expected value of the population size. On the other hand, it is also slow relative to the approximate methods. If a reasonable approximation would be sufficient, simulation may be preferred. Both methods preserve the demographic characteristics of the individuals expected to contribute to the population estimate.

If the arrival distributions and counts are similar across time periods, or if visit lengths tend to be short, the Product Method and Forced Average Method should also provide accurate approximations. Neither method may be appropriate, however, if demographic analysis is desired, unless the latter of the two conditions is met.

The accuracy of all of the methods is dependent on the accuracy and perpetuity of the PMFs and on the explanatory power of the nonimmigrant category. For the present purposes of comparing estimates and CPU times, the PMFs are assumed to be correct, constant over time, and completely determined by the nonimmigrant category. Because of the increasing trend in arrival counts from year to year, the PMFs used here may actually be biased towards shorter visits, thereby biasing the population estimates downward. Furthermore, the PMFs may change over time (as with changes in immigration law, policy, or airline prices), and additional observed variables may (and sometimes do) influence visit length, thereby introducing additional error into the estimates. Actual attempts to estimate the resident nonimmigrant population size should take these concerns under consideration.

Appendix – Base Data and Construction of the PMFs

Base Data

The base data consist of DHS Form I-94 nonimmigrant arrival and departure records. The I-94 is a two-part form, consisting of an arrival stub and a departure stub. The arrival stub is submitted to customs officials by the nonimmigrant during the admission process, and the collection of arrival records is assumed to be complete. Departure stubs are requested and accepted by the airlines for departures by air, and collection is known to be incomplete[1]. Both the arrival and departure stubs include the nonimmigrant's name and date of birth and are pre-stamped with a unique identifier, making it possible to reassemble the stubs into a complete visit record, assuming both stubs are collected. In essence, the available datasets include a full set of arrival records and a partial set of complete visit records.

Construction of the Visit-length PMFs

Visit-length PMFs were estimated from the visit lengths in the partial sets of complete visit records with 2010 departures. The sets were made to be random subsets of all complete visits with 2010 departures by assuming that data collection failure occurs completely at random. The sets were made to be representative of all visits (not just those with 2010 departures) by further assuming that the PMFs are constant over time, that the PMFs are completely determined by the nonimmigrant category, and that the arrival date frequency distributions are identical from one year to the next. For the present purposes, the PMFs are assumed to be the "true" PMFs.

About 20-35 percent of departure records are never collected. Although there are logical reasons to suspect that the probability of data collection failure may be correlated with visit length, there are no data available to model the relationship. For lack of an empirical model, data collection failures were assumed to have occurred completely at random.

Visit-length distributions may fluctuate over time, as with changes in airline prices and other conditions, but there is no evidence of any major trends for the largest classes of admission. For the present purposes, it is assumed that visit-length PMFs are constant over time. It is further assumed that the PMFs are completely determined by the nonimmigrant category[2].

The arrival date probability distribution is consistent from one year to the next for some of the most important categories (e.g., students in general, and Chinese students in particular), but there are nearly 3,500 categories, and the year-to-year consistency has not been confirmed for all of them. For simplicity, it is assumed that arrival date probability distributions are identical from one year to the next for all categories.

Overall arrival frequencies increased by about 3 percent each year, increasing from about 2.9 million in 2001 to about 3.6 million in 2010. Because the criterion for the visit-length samples was a departure record in 2010, shorter visits may be slightly overrepresented, and estimates based on that sample may be biased downward.

A Cautionary Note

For categories with many matched records, the visit length PMF may be sufficiently described by the data. Categories with fewer matched records tend to also have few arrivals, and so have little impact on the overall estimate. So, estimates for categories based on PMFs with few data

[1] Typically 20-35% of departure stubs are never collected.

[2] In practice, other recorded variables may also influence visit length. For example, students who arrive in August are more likely to stay for 9 months or so (roughly 2 semesters) than students who arrive in January.

points should be treated with caution, even if the overall estimate and estimates for relatively large categories appear reasonable.

References

Baker, Bryan C., “Estimates of the Resident Nonimmigrant Population in the United States: 2008,” Office of Immigration Statistics, Policy Directorate, U.S. Department of Homeland Security, http://www.dhs.gov/xlibrary/assets/statistics/publications/ois_ni_pe_2008.pdf.

Grieco, Elizabeth M., 2006. “Estimates of the Nonimmigrant Population in the United States: 2004,” Office of Immigration Statistics, Policy Directorate, U.S. Department of Homeland Security, http://www.dhs.gov/xlibrary/assets/statistics/publications/NIM_2004.pdf

Hoefer, Michael, Nancy Rytina, and Christopher Campbell, 2006. “Estimates of the Unauthorized Immigrant Population Residing in the United States: January 2005,” Office of Immigration Statistics, Policy Directorate, U.S. Department of Homeland Security, http://www.dhs.gov/xlibrary/assets/statistics/publications/ILL_PE_2005.pdf. Updates for 2006-10 can also be found at http://www.dhs.gov/files/statistics/publications/.