COUNTING THE UNCOUNTABLE ILLEGALS: SOME INITIAL STATISTICAL SPECULATIONS EMPLOYING CAPTURE-RECAPTURE TECHNIQUES

Clarise Lancaster, Office of the Assistant Secretary for Planning and Evaluation, Department of Health, Education, and Welfare
Frederick J. Scheuren, Social Security Administration

This paper provides some initial statistical speculations on the number of illegal aliens residing in the United States. Our results come from the 1973 CPS-IRS-SSA Exact Match Study [1] which has been conducted jointly by the Census Bureau and the Social Security Administration, assisted by the Internal Revenue Service. Direct estimates are presented only for the age group 18 to 44 years old as of April 1973; however, there is some discussion of ways, using other sources, that one can extend these figures to all age groups and project them forward in time.

Organizationally, the paper is divided into five sections. Section 1 provides a brief introduction to what is known about the nature and magnitude of the illegal alien population. The approach we will take in obtaining estimates for 1973 is described in section 2. Some limitations on the data being used are set forth in section 3. Section 4 discusses the results of the exploratory analyses we have carried out so far. A few conclusions and possible implications for future study are given in section 5.

1. INTRODUCTION

Most of what we know about illegal aliens comes from data on apprehensions (about 800,000 in 1975) which suggest that Mexico is a major source of such individuals.1/ United States and Mexican authorities, however, have,on numerous occasions, cited the unreliability of the apprehension information as indicative of the nature of the total illegal alien population in the U.S. In particular, it is misleading to characterize the illegal alien population in the United States as predominantly male and Mexican based on these apprehension statistics: first, because we are dealing with those who are, in fact, caught, and there is no reason to believe that they are representative of those who are not caught; and, secondly, because Mexican illegal immigration may be substantially different from that of other source countries, mainly Jamaica, the Dominican Republic, Haiti, Korea, the Phillippines, Thailand, and China. It is suspected that both Mexicans and males are over-represented in apprehension data.

Not only is the composition of the illegal alien population unclear from official statistics, but the total number of illegals who are not apprehended is, of course, unknown and is a source of considerable speculation. To see how widely divergent some of the guesses are, it might be worth quoting from a recent article by Hobart Rowen [4] in the Washington Post--

> There are four million illegal aliens in the United States.
>
> There are eight million illegal aliens in the United States.
>
> There are twelve million illegal aliens in the United States.
>
> These are the estimates of [government] officials trying to evolve a policy to deal with illegal immigration. You can pick any one of them, or insert your own number and you will be--they confess--as accurate as they are."The truth is [an official says] that no one knows how many 'illegals' are in the country."

As will be seen later in this paper, our own preliminary investigations suggest that it is the smallest of these figures which is more nearly correct.

2. METHODOLOGY

2.1 General.--The approach we will use to estimate the number of illegal aliens makes use of two sources of information:

1. a sample of the total resident civilian noninstitutional population, including illegal aliens (who were not, however, identifiable as such); and

2. an independent estimate or "count" of the number of persons in the resident civilian noninstitutional population, excluding illegal aliens.

From the sample data, the Capture-Recapture procedure is used to estimate the total resident civilian noninstitutional population including illegal aliens. The independent population total, excluding illegal aliens, is then subtracted from this sample estimate to derive counts for "illegals."

The sample we are using to make estimates is the Census Bureau's March 1973 Current Population Survey (CPS). The capture-recapture technique can be applied to this sample because it has been matched to Internal Revenue Service (IRS) individual income tax records, and Social Security Administration (SSA) earnings and benefit data.

The independent population estimates on which we rely also come from the Census Bureau. They were obtained by adjusting the 1970 Census count for underenumeration and carrying forward the population totals taking account of subsequent aging of the population, births, deaths, and net legal migration [5, 6]. Also excluded from the population estimates were members of the Armed Forces in April 1973 and persons living in institutions [7].

2.2 Capture-Recapture techniques.--In order to explain how we employed the capture-recapture technique, let us examine table 1, which illustrates our approach for the total 18 to 44 year age group. Two observations should be made initially:

1. All the individual cell estimates, except for the lower right-hand corner total, were taken from a random half-sample selected from the 1973 CPS-IRS-SSA Exact Match Study. These were the data with which we started our exploratory analyses. 2/

2. The right-hand corner entry (shown in parenthesis) was obtained by subtracting the remaining cells from the April 1, 1973, Census Bureau estimate (73,893,000) for the total civilian noninstitutional population 18 to 44 (which excludes illegal aliens).

Now the capture-recapture [8], or multiple systems [9], estimation procedure that we used, essentially resolved itself into treating the cell entry in the parenthesis as missing and estimating it from the remainder of the table. Once this was done, the difference between the new entry for the "missing" cell and the original (parenthesized) entry provided our count of "illegals." 3/

To compute the capture-recapture estimate for the missing cell, we employed expression (6.4-15) from [8], that is:

$$m_{222} = \frac{m_{111}\, m_{221}\, m_{122}\, m_{212}}{m_{121}\, m_{211}\, m_{112}},$$

where the cell counts or entries $\{m_{ijk}\}$ are defined by letting i = 1 or 2, depending on whether there is a yes or no, respectively, on the IRS dimension (i.e., whether a person was in a unit with a taxfiler, "yes", or not, "no"); j=1 or 2, depending on whether there is a yes or no on the SSA covered employment dimension; and, finally, k=1 or 2, depending on whether there is a yes or no on the SSA beneficiary dimension.

The above formula for the missing entry m_{222} cannot be interpreted without making a number of (strong) assumptions. Two might be mentioned here:

1. To explain all the interrelationships which exist between the three "captures" (administrative systems), it is enough to look at just the pairwise associations between them. (More technically, the assumption is being made that there is no second-order interaction.)

2. The very same set of "capture" probabilities applies to each individual in the population. Such an assumption would only be tenable if the group we are dealing with were divided into very homogeneous subgroups--something we will discuss in section 4.

2.3 Definition of classifiers.--Some definitions are needed of exactly what we mean by the classifiers in table 1. These are provided in the following paragraphs:

1. SSA beneficiaries.--To be considered an SSA beneficiary, a person had to be receiving benefits in December 1972 (i.e., be in Current Pay Status for that month).

2. SSA covered employment.--To be considered as a covered worker, an individual had to have had taxable SSA wages or self-employment reported for calendar year 1972.

3. Federal income taxfiler.--To be considered a taxfiler, an individual had to have filed a tax return for 1972 on which he was designated as the primary taxpayer. 4/

4. STATS unit.--This is a nuclear family concept used at Social Security to designate individuals in CPS households who would generally be considered interdependent under social insurance programs [10]. The designation, STATS units, stands for "Simulated Tax and Transfer System" units. These units can consist of a single adult 22 years or older, an adult with children under 14, and married couples with or without children. Young adults (14 to 21 years old), depending on their living arrangements, are treated as separate units or as part of a unit containing their parent(s).

TABLE 1.--U.S. civilian noninstitutional population 18 to 44 years old as estimated from the 1973 Census-Social Security Exact Match Study and Census Bureau sources

(In thousands)

In STATS units with persons in SSA covered employment	Total	In STATS units with persons filing Federal income tax returns	
		Yes	No
Overall total....	76,893	67,289	9,604
IN STATS UNITS WITH SSA BENEFICIARIES			
Yes.....	1,321	1,142	179
No.....	509	79	430
NOT IN STATS UNITS WITH SSA BENEFICIARIES			
Yes.....	68,412	63,447	4,965
No.....	6,651	2,621	(4,030)

Note: For definitions of terms used, see section 2.3.

In table 1 above and in the tables used in our subsequent analyses, we do not classify an individual by whether or not he or she was "captured" by one of the administrative systems, but, rather, by whether or not anyone in his or her STATS unit had been so captured. Two (natural) questions arise in this connection: "Why didn't we classify individuals by their own characteristics?" and "How sensitive would our results be if we had done so?"

We didn't classify people just on the basis of their own characteristics for two reasons. First, the STATS unit, by construction, is conceptually more attractive as a classifier of an individual's relationship with regard to the beneficiary and tax systems. Second, by using the STATS unit as a classifier, we expected to increase the overlap

among all three systems, which, in turn, would reduce the probability of having zero cells and, perhaps,make more tenable our assumption of no second-order interaction.

When this paper was delivered in Chicago, we had not yet obtained an answer to the question of how sensitive our results would be if we did the analysis on a person, rather than a STATS unit, basis. The work we have done since then suggests that the results would be very sensitive indeed. The person-based estimates do not actually contradict the STATS unit ones, however. What seems to be happening is that the sampling error of the estimate of the missing cell has increased enormously, principally because much more of the sample was not "captured" by any system.

3. DATA LIMITATIONS

The assumptions which the method requires necessarily impose limitations on our estimates. In addition to these, however, there is also a second set of limitations which arises from the nature of the data on which we are using the method:

1. Survey and matching problems.--The starting point of the administrative record matches was the CPS and not the systems themselves. Problems of non-matches, mismatches, coverage, and non-interview nonresponse must necessarily be considered. (See [11], for example.) It is enough to say here that we believe that these data problems definitely raise interpretive issues, even though major efforts were made to adjust or "correct" for any impacts they might have had [7].

2. Administrative data problems.--The nature of the administrative systems we are using is such that illegal aliens might be less well-represented than their (other) socio-economic characteristics (income level, age, race, sex, etc.) might otherwise suggest. We do not know how serious this is, but it is a problem which we believe would (in the absence of other problems) lead to an underestimation of the total illegal population.

3. Independent population totals.--The Census Bureau population estimates needed for deriving "illegals" are themselves subject to error. Evidence from [12], for example, suggests that there may be a serious understatement in the allowance made for outmigration. For the 18 to 34 year olds this is likely to be the only important error. For the remainder of the 18 to 44 year age group, that is, persons 35 to 44, the undercount totals (Siegel's Preferred Series D) for 1970 are based on a combination of demographic techniques [5, p.6] and not, principally, on vital records, as is true of the younger ages (suggesting that there might be proportionately more error in the older age group).

4. EXPLORATORY ANALYSIS

When this paper was given at the meetings, we were still in the exploratory analysis phase of our research on illegals. In order to be able (at a later date) to do at least some confirmatory analysis, we restricted our attention to half the sample cases in the 1973 Exact Match Study.

4.1 Initial results.--To make more tenable the assumption that the capture probabilities were equal for every individual, we subdivided the age group 18 to 44 into four race-sex subgroups: white males, white females, males of other races, and females of other races. This also has the advantage, as Chandra Sekar and Deming have suggested [13], of tending to lower the overall variance.

Table 1 was repeated for each subgroup separately. The combined tabulation, consisting of 32 cells (four of which were to be treated as missing), was then subjected to "standard" log linear contingency table fitting procedures.5/ Our goal was, of course, the usual one: eliminating those parameters which the analysis showed were unnecessary. In other words, to create a model with fewer parameters which fits well enough to withstand statistical inspection while, at the same time, is sufficiently parsimonious to yield "sturdy" estimates.

Many models were considered before we settled on one to illustrate our results. The model chosen was fit by iterative proportional scaling to the following five sets of marginal totals:

1. Sex
2. Race and taxfiler status
3. Taxfiler status and covered worker status
4. Taxfiler status and beneficiary status
5. Covered worker status and beneficiary status.

Table 2.--Initial Exploratory Model Estimates for April 1973, of Total U.S. Civilian Noninstitutional Population 18 to 44 Years Old by Race and Sex

(Numbers in thousands)

Race and Sex	Total excluding illegals*	Total including illegals	Difference (illegals) Number	Difference (illegals) Percent
Total.....	76,893	79,951	3,058	100.0
Male.....	37,490	39,705	2,215	72.4
Female.....	39,403	40,246	843	27.6
White, total.....	66,673	68,603	1,930	63.1
Male.....	32,689	34,069	1,380	45.1
Female.....	33,984	34,534	550	18.0
Other races, total	10,220	11,348	1,128	36.9
Male.....	4,801	5,635	834	27.3
Female.....	5,419	5,712	293	9.6

(*)Population totals not adjusted for understatement of 1960-73 outmigration.

Once we had obtained our fitted model, we then used the estimates it provided in each of the four race-sex subtables to obtain new entries for the "missing" cells. From the "before" and "after" totals for each race-sex group we then constructed table 2.

4.2 Further results.--We brought a computer terminal with us to the meetings and invited anyone interested in the results in table 2 to try his own hand at still other models. Our basic data set had literally hundreds of dimensions we had not yet looked at. Two we thought most promising were age and income; and we had come prepared to fit models involving these variables if anyone suggested them. As luck would have it, the interactive APL computer service we use was down most of the day of the meeting, and no one was able to take us up on our offer. Matters did not rest at this point, however.

A number of discussions have been held, since the paper was delivered, with various individuals interested in and knowledgeable about illegal alien immigration. From these conversations, we concluded three things. First, we had to provide at least one model which split up the rather broad age group 18 to 44. Second, we had to adjust our initial estimates for the rather serious understatement (over 500,000) in the outmigration estimates used to obtain population totals that excluded illegal aliens. Third, since our initial and improved results had a certain amount of plausibility, they were likely to be believed and used. Therefore, as "responsible" researchers, we had to provide at least some rough idea about the magnitude of the uncertainty surrounding our figures.

In accord with these excellent suggestions, we returned to our exploratory work with the same half sample that was used to obtain table 2. This time we added age as a dimension (18 to 34 and 35 to 44) and looked at models for the 6-way table involving sex, race, age, and the three administrative systems. The model we finally settled on was obtained by fitting the following marginal totals:

1. Sex
2. Race and taxfiler status
3. Age and taxfiler status
4. Taxfiler status and covered employment status
5. Taxfiler status and beneficiary status
6. Covered employment status and beneficiary status.

To test this model, we fit it on the second half of our sample. While the fit (as expected) was not nearly as good on the second half, it still could be accepted at the $\alpha = .05$ level of significance.

Our next step was to combine the two half samples and refit the model on all the data. The estimates obtained in this way are shown in table 3, column (2). The final step we took was to revise the population estimates not including illegals (column (1) of table 3) to account for the understatement of outmigration. The Warren-Peck paper [12], set B estimates were our basic source. These were aged to 1973, the effect of additional outmigrant underestimation between 1970 and 1973 was imputed, and a rough adjustment was made to take account of changes in the foreign student population not originally reflected in [12].6/ The result of these steps is shown below.

Age Group	Understatement of Outmigrants (in thousands)		
	Total	Male	Female
Total.....	568	244	324
18 to 34 years..	440	180	260
35 to 44 years..	128	64	64

Since virtually all of the outmigrants involved were believed to be white, we made the entire adjustment in that racial group.

Table 3.--Overall Revised Model Estimates for April 1973 of Total U.S. Civilian Noninstitutional Population 18 to 44 Years Old by Race and Sex

(Numbers in thousands)

Race and Sex	18 to 34 Years of Age			35 to 44 Years of Age		
	Total excluding illegals*	Total including illegals	Difference (illegals)**	Total excluding illegals*	Total including illegals	Difference (illegals)**
Total.....	53,401	56,583	3,182	22,924	23,627	703
Male.....	25,973	27,974	2,001	11,273	11,681	408
Female.....	27,428	28,609	1,181	11,651	11,946	295
White, total.....	46,198	48,379	2,181	19,910	20,304	394
Male.....	22,613	23,918	1,305	9,834	10,038	204
Female.....	23,585	24,461	876	10,076	10,266	190
Other races, total.	7,203	8,204	1,001	3,014	3,323	309
Male.....	3,360	4,056	696	1,439	1,643	204
Female.....	3,843	4,148	305	1,575	1,680	105

(*)Adjusted for outmigration as explained in the text.
(**)This estimate differs from that in table 2 due to the adjustment for outmigrants discussed in the text, to the fact that the whole sample is being used, not just half, and to the fact that the models fit in the two cases are different.

4.3 Crude measures of uncertainty.--It is a formidable, perhaps impossible, task to do a "good" job of assigning measures of uncertainty to the entries for "illegals" in table 3. We have to obtain the approximate sampling errors of the estimates, quantify the impact of the nonsampling errors, and assess the robustness of the figures to possible failures in the assumptions underlying our application of the capture-recapture method.

Time considerations precluded our making more than a crude attempt to quantify the uncertainty surrounding the estimates in table 3. Perhaps we should not even have tried, since subjective judgments play such an important role in our assessments and, undoubtedly, other researchers may reach quite different conclusions.

Table 4 provides the rough confidence bounds we constructed.7/ Notice that they are not symmetric, reflecting our belief that the counts of "illegals" in table 3 may be downwardly biased. The bounds also are quite far apart. This is in keeping with the early stage at which our analysis stands. Further research probably would lead to estimates with narrower bounds of uncertainty.

Table 4.--Subjective 68 percent Confidence Intervals for the Overall Revised Model Estimate of the Number of Illegal Aliens 18 to 44 Years of Age in April 1973 by Age, Race and Sex

(In thousands)

Race and Sex	18 to 44 Years of Age		18 to 34 Years of Age		35 to 44 Years of Age	
	Lower	Upper	Lower	Upper	Lower	Upper
Total.....	2,904	5,722	2,438	4,574	466	1,148
Male.....	2,046	3,318	1,726	2,689	320	629
Female.....	858	2,404	712	1,885	146	519
White, total.....	1,961	3,724	1,715	3,052	246	672
Male.....	1,282	2,077	1,133	1,735	149	342
Female.....	679	1,647	582	1,317	97	330
Other races,total..	943	1,998	723	1,522	220	476
Male.....	764	1,241	593	954	171	287
Female.....	179	757	130	568	49	189

5. SOME CONCLUSIONS AND IMPLICATIONS

According to the overall model shown in table 3, there were some 3.9 million resident "illegals" 18 to 44 years of age in April 1973. Rough, subjective, 68 percent confidence bounds on this estimate (from table 4) suggest that the actual value could be anything from 2.9 million to 5.7 million. Generally speaking, such widely (wildly?) varying speculations would cause most people to make no further demands on the present results. We certainly would not wish to do so were it not for the fact that the questions of most interest are--

> "How many illegals were there, altogether, in 1973?"
> "How much has the total increased since 1973?"

We cannot offer any statistical speculations of our own on these questions, but it might be worth mentioning how others have answered them. First, David North, in [14], cites various studies which ..."suggest that the 18-44 age range would cover most, but not all, of the illegal aliens; a 10% upward adjustment would appear appropriate. ..." On the second question, we turn to some conclusions of Alex Korns [15], who has examined the relationship between the BLS establishment and CPS employment series for nonagricultural wage and salary jobs. He notes that while there may have been a sharp rise in illegal alien employment during the business expansion of 1964-1969, there appears to be no sustained increase since then.

With these two outside sources in mind, we feel reasonably comfortable in restating the assertions about the number of "illegals" that Rowen quoted:

> There are probably not twelve million illegal aliens in the United States.
>
> There are probably not eight million illegal aliens in the United States.
>
> There may, however, be about four million illegal aliens in the United States.

AN AFTERWORD

We debated whether or not to submit this paper to the Proceedings. The subject is, after all, important and controversial; hence, it deserves a careful, studied treatment. Unfortunately, time and resource constraints intervened. Our results, therefore, are quite preliminary and could be misleading if taken too seriously.

Ultimately, what persuaded us to give the paper and, then, have it published was an expectation that other statisticians interested in "illegals" would learn about the 1973 Exact Match Study data base and use it in their own research. The public-use files from the study are now available and may provide the means to do the complete, thorough job that the subject deserves. We would be more than happy to assist in any such effort.

ACKNOWLEDGEMENTS AND FOOTNOTES

The authors would like to thank several individuals for sharing their expertise on illegal aliens: David North, Alex Korns, Muffie Houstoun, and especially Robert Warren. We also benefitted considerably from discussions with Jeff Passel and Jacob Siegel at the Census Bureau after the paper was delivered at the Chicago meetings. Editorial and other assistance was provided by Ben Bridges, H. Lock Oh, Linda DelBene and, especially, Wendy Alvey. The typing was done by Joan Reynolds and Helen Kearney.

We would also like to take this opportunity to mention two points about the title of our paper. First, "Counting the Uncountables" is apparently an irresistable phrase. The Illegal Alien Study Design report [3], for example, uses the expression, something we were not aware of when we chose it ourselves. The Design report also suggests that the well-known "Capture-Recapture" technique be employed to estimate the number of illegal aliens. In doing so, the authors of that report add a graceful apology, with which we concur, for the necessity of using such (customary) terminology with respect to this population.

1/ The authors have relied primarily on [2] and [3] for the brief overview of the illegal alien immigration situation in this section.

2/ The estimates were obtained by using twice the "Final" administratively weighted [7] sample figures from rotation panels entering the survey in March for the first, third, sixth or eighth time.

3/ In the more general settings later in section 4, the "count of illegals" is obtained by calculating the difference between the model estimated total population derived from the sample (which includes "illegals") and the Census supplied population (where "illegals" are excluded). It might be mentioned also that just because we sometimes calculate our estimates from the "missing" cell does not imply that this is where all the illegals will be found. Quite the contrary. If none of the "illegals"

were ever "captured" by the administrative systems, then our procedure simply would not work.

4/ For nonjoint returns, there was considered to be only one taxpayer; for joint returns filed by married couples, there were two. In such cases, the husband was designated as the primary taxpayer.

5/ Actually, standard log linear procedures require simple random sampling. The CPS sample design and estimation procedures were such that we had to modify the ordinary minimum discrimination information (maximum likelihood χ^2) test statistic by dividing by the product of the base weight for the half sample (3,200) times a preliminary estimate of the design effect (taken to be quite large, about 3). The data for both half samples is available upon request.

6/ The updating and adjustments were prepared with the help of Robert Warren.

7/ The actual steps we went through to obtain these crude bounds are available upon request.

REFERENCES

[1] U.S. Social Security Administration, Studies from Interagency Data Linkages, Reports Nos. 4 and following.

[2] Domestic Council Committee on Illegal Aliens, Preliminary Report, December 1976.

[3] U.S. Immigration and Naturalization Service, Illegal Alien Study Design, Vol. 1-Final Report, May 1975.

[4] Rowen, H., "Illegal Alien Dilemma," Washington Post, Section A, p. 19, July 21, 1977.

[5] Siegel, J., Estimates of Coverage of Population by Sex, Race and Age: Demographic Analysis, 1970 Census of Population and Housing: Evaluation and Research Program, PHE(E)-4, 1974.

[6] U.S. Bureau of the Census, "Population Estimates and Projections," Current Population Reports, Series P-25, No. 614.

[7] Scheuren, F., "Methods of Estimation for the 1973 Exact Match Study" (unpublished working paper to appear in the series Studies from Interagency Data Linkages).

[8] Bishop, Y., Fienberg, S., and Holland, P., Discrete Multivariate Analysis: Theory and Practice, Cambridge: MIT Press, 1975, Chapter 6.

[9] Marks, E., Seltzer, W., and Krotki, K., Population Growth Estimation: A Handbook of Vital Statistics Measurement. New York: The Population Council, 1974.

[10] Projector, D., Millea, M., and Dymond, K., "Projection of March Current Population Survey: Population Earnings, and Property Income, March 1972 to March 1976," Studies in Income Distribution, Report No. 1, Social Security Administration, 1975.

[11] Yuskavage, R., Hirschberg, D., and Scheuren, F., "The Impact on Personal and Family Income of Adjusting the Current Population Survey for Undercoverage," 1977 American Statistical Association Proceedings, Social Statistics Section.

[12] Warren, R. and Peck, J., "Emigration from the United States: 1960 to 1970," paper presented at the Population Association's annual meeting in Seattle, Washington, July 17-19, 1975.

[13] Chandra Sekar, C. and Deming, W., "One Method of Estimating Birth and Death Rates and the Extent of Registration," Journal American Statistical Association, Vol. 44, 1949, pp. 101-115.

[14] North, D., "Manpower Policy and Immigration Policy in the United States: An Analysis of a Nonrelationship," Chapter IV and Appendix D. (This report is to be published shortly by the National Commission for Manpower Policy.)

[15] Korns, A., "Coverage Issues Raised by Comparisons between CPS and Establishment Employment," 1977 American Statistical Association Proceedings, Social Statistics Section.