CAN WE IGNORE THE MIGRATION OF INCOME TAX NON-FILERS WHEN BENCHMARKING THE AMERICAN COMMUNITY SURVEY'S COUNTY ESTIMATES?

Yves Thibaudeau
U.S. Census Bureau

Introduction

When fully implemented the American Community Survey will provide population estimates at the census tract level. At the same time, demographic analysis provides independent population estimates that do not account for migration. The population division at the Census Bureau adjusts the counts obtained from demographic analysis for migration at the county level by matching administrative records, primarily income tax returns, for two successive years. By identifying the returns corresponding to filers who file only the first or the second year, for each county, the population division obtains estimates of domestic migration at the county level. Since demographic analysis is generally considered very reliable, and the matching operations involve all the tax returns representing a county, the population estimates corrected for domestic migration are likely to more reliable for the computation of the relative propensities of various subdomains. When deriving fresh estimates from the ACS, we can take advantage of the superior accuracy of the corrected population estimates by using them to benchmark the ACS estimates, thereby enhancing the quality of these estimates.

However, one problem with the correction of the population estimates for domestic migration is that the non-filers are not represented in the files containing the income tax returns corresponding to each county. At the same time, some research suggests (Huang, Kim, 2000) that non-filers could account for up to 8 % of the workforce. Therefore, if non-filers migrate at rates substantially different than filers, the estimates produced by adjusting the demographic analysis for county-level migration by matching the IRS files could be significantly biased, leading to biased benchmarks for some subdomains. In the paper we make a first step toward an evaluation of the benchmarks accuracy in order to make an educated use of these benchmarks in the future. Interestingly, it is data collected from the ACS itself that allows us to proceed with our investigation, because the ACS includes questions on migration that provide estimates on rates of in-migrations without excluding a priori the subpopulation of non-filers from the general workforce. Of course, the ACS presents its own difficulties when attempting to estimate migration rates based on filing status. For instance non-response biases, and measurement errors cause problems. The non-filers may be less likely to respond to the questions on migration than the rest of the population. The biggest problem is that the ACS does not have a question designed for the identification of the non-filers. That is filing status is unreported. Thus any study attempting to correlate filing status with behavioral traits, such as propensity to migrate, must be carried on the basis of proxy information, in lieu of direct information

on filing status. Despite these disadvantages, the ACS provides an independent source of knowledge that can be exploited in an investigation of the consequences of deriving adjustments for migration at the county level, when information on the non-filers is unavailable.

2. Adjusting the County Level Population Counts for Migration

The Census Bureau uses a comprehensive model (Sink, 2000) to create population estimates derived from demographic analysis. Four types of county-level migration are explicitly accounted for:

1. Out-of-State In-County Migration (OSIM)
2. Out-of-State Out-County Migration (OSOM)
3. Intra-State In-County Migration (ISIM)
4. Intra-State Out-County Migration (ISOM)

For each county, and for each age-sex-race origin category ,the model is

$$TM = OSIM - OSOM + ISIM - ISOM \qquad \textbf{(1)}$$

To estimate the terms on the right hand side (RHS) of equation (1) the Census Bureau matches together the income tax returns for two subsequent years at the county level. To help focus on biases associated with the content of the files of returns, we assume that other errors are under control. In particular we suppose that most individuals represented either by the filer of an income tax return, or by an exemption claimed on a return can be identified through the matching operation, and their between-county migration status can be determined. To produce estimates of domestic migration accounting for the non-filers and their exemptions, the Census Bureau takes advantage of available estimates of domestic migration at the state level that include the non-filing population, and the results of the matching operations at the county level. The Census bureau computes estimates of between-county migration at the county level with a raking procedure involving both state level estimates and county level matching results. However, the raking procedure assumes a uniform rate of migration computed from the matching results involving tax filers only. We are interested in the impact of excluding the non-filers from the raking procedure on the estimates of the between-county migration, for each county.

3. Differential Migratory Behavior of the Filers and Non-Filers

The assumptions of uniformity imbedded in the raking procedure are defined by two rules. First, the rate of between-county migration at the state level depends only on race and Hispanic origin. That is, total between-county migration for a state is obtained by multiplying the state population by an immigration rate proper to each race and Hispanic origin combination. In probabilistic terms we have the following rule.

Rule 1: Allocation of the Filing and Non-Filing Populations by Migration Status

At the state level, conditional on Race and Hispanic origin, between-county migration status (MS) is independent of age, sex, and filing status (FS). Let MS be either between-county migrant, or non-migrant. Given race and origin, the rule is

$$\text{Prob}(\text{MS}, \text{Age}, \text{Sex}) = \text{Prob}(\text{Age}, \text{Sex}) * \text{Prob}(\text{MS} \mid \text{FS} = \text{filer}) \qquad (2)$$

Summing up both sides of equation (2) over age and sex we obtain

$$\text{Prob}(\text{MS}) = \text{Prob}(\text{MS} \mid \text{FS} = \text{filer}) \qquad (3)$$

Therefore, migrant status is independent of filing status and

$$\text{Prob}(\text{MS} \mid \text{FS} = \text{filer}) = \text{Prob}(\text{MS} \mid \text{FS} = \text{non-filer}) \qquad (4)$$

In the paper, we question the assumption of independence implicit in (2). That is we ask the following questions: is the propensity to migrate dependent on FS (i.e. does (4) holds)? And does it matter, in terms of (1)?

After the state-level inter-county migration for each race-origin-age-sex category has been determined, it must be allocated between the counties. The raking procedure to do this is based on the following rule.

Rule 2: Allocation of the State Level Migrant Population by County of Residence

Let CR denote the county of residence. For migrants, given race and origin the rule is

$$\text{Prob}(\text{CR}, \text{Age}, \text{Sex}) = \text{Prob}(\text{Age}, \text{Sex}) * \text{Prob}(\text{CR} \mid \text{FS} = \text{filer}) \qquad (5)$$

Summing-up both sides of equation (5) over age and sex, we obtain that county of residence is independent of filing status. That is

$$\text{Prob}(\text{CR}) = \text{Prob}(\text{CR} \mid \text{FS} = \text{filer}) = \text{Prob}(\text{CR} \mid \text{FS} = \text{non-filer}) \qquad (6)$$

4. The American Community Survey

In 1999, the American community survey collected data for 36 counties, with various sampling rates for the different counties. One question on the survey requests the county and state of residence for the year previous to the survey. At this time, using this question, we can only investigate the hypothesis pertaining to ISIM in equation (1), and we can only evaluate rule 1, as data are not available yet for an entire state. Our objective is to assess the extent of any departures from rule 1 for between-county migration in the states covered by the ACS, and to evaluate the bias resulting from these departures (if any), in terms of the population estimates of the between-county migration. To achieve

this objective, we explore models with a general structure compatible with equation (3) only under specific conditions, and we assess if these conditions are violated when the model is applied to the data. For each race-Hispanic origin category we propose the following general model

$$\text{Prob}(\text{FS}, \text{MS}, \text{Age}) = \text{Prob}(\text{FS}) * \text{Prob}(\text{Age} \mid \text{FS}) * \text{Prob}(\text{MS} \mid \text{FS}) \qquad (7)$$

$$= \text{Prob}(\text{FS}, \text{Age}) * \text{Prob}(\text{MS} \mid \text{FS}) \qquad (8)$$

The model given in equations (7) and (8) stipulates that, conditional on filing status, migration status and age are independent. In other words, any dependence between migration status and age can be explained by their correlation with filng status. Then, the model in equation (8) agrees with rule 1 only if equations (3) and (4) hold. That is, only if the probability of between-county migration is the same for filers and non-filers. We use a model of the general form given in equations (7) and (8) to assess the extent of the departures from equations (3) and (4) in the data, and consequently, from rule 1. To provide a valid assessment, the model must fit the data well, and the estimates of the parameters involved in the model in equations (7) and (8) must be plausible. We conjecture that the non-filers are associated with a class of younger, less affluent householders, and are more prone to rent, rather than own, relative to the filers. Thus we incorporate income and tenure in the general structure given in equation (8)

5. A latent Class Model for Inter-County Migration

Because filing status is estimated implicitly in (8) and (9), additional covariates are necessary to discriminate between categories with different values of filing status. We choose tenure and income, in addition to age and in-migration, as our discriminatory variables. The specific model is:

$$L(\boldsymbol{N}; \boldsymbol{P}) = \tilde{\prod}_{i,j,k,l,m} \left(P(i,j,k,l,m)\right)^{N(i,j,k,l,m)} \quad ; \quad \boldsymbol{P} \in \Theta_{\boldsymbol{P}} \qquad (9)$$

In equation (9), i, j, k, l, m are binary indicators associated with filing status, age, tenure, salary, and between-county migration status, respectively. For age and salary, we choose specific cutoffs to define the value of the binary indicators. Tenure can only take the values “owner”, and “renter”.. $N(i, j, k, l, m)$ is the count of householders in the ACS sample for a county with values of filing status, age, tenure, salary, and between-county migration status according to i, j, k, l, m respectively. $\boldsymbol{P} = [P(i, j, k, l, m)]$ is the 16-dimension probability vecor representing the relative prevalence of each type of householder. Thus we have $P(i, j, k, l, m) > 0$, for $i, j, k, l, m = 0, 1$; and $\sum_{i,j,k,l,m} P(i, j, k, l, m) = 1$. $\Theta_{\boldsymbol{P}}$ is a hierarchical log-linear model (Agresti, 1990, Bishop, Fienberg, Holland, 1976) including all second order-interactions involving filing status and a covariate, and the second-order interactions

between tenure and migration. Thus, the model in (9) is not a conditional independence model. Our analysis of the data suggests that the bond connecting between-county migration status and tenure is simply too strong to be manifest only through the interim of the latent classes representing filing status. Since there are four observed categorical variables, the model support the estimation of all these interactions.

Equation (9) specifies a latent class model because, while the likelihood in (9) is expressed in terms of the $N(i, j, k, l, m)$'s, only counts aggregated on age, tenure, salary and migration, represented by $M(j, k, l, m) = \sum_i N(i, j, k, l, m)$, for $j, k, l, m = 0, 1$, are available from the ACS. Nevertheless, we can still construct ten independent likelihood equations involving ten parameters. If the maximum likelihood estimator (MLE) is not on the boundary, appropriate starting values allow us to find it with the EM algorithm.

Huang and Kim (2000) match a one percent sample of the IRS files to various administrative records, and estimate a rate of non-filers of approximately 8 %. This suggests using a constraint to ensure that the sizes of the latent classes are .92 and .08. Winkler shows how to use the EM algorithm under afine constraints. However, at a first trial we do not constrain the latent classes, and we pay particular attention to the sizes of the latent class in relation with the observations of Huang and Kim.

6. Results

We present results for 32 counties surveyed by the ACS in 24 states. We analyze data on inter-county migration for Non-Hispanic White householders in these counties and states. 30 of these counties reported inter-county in-migrants in that category. The categories defined by the cutoffs for age and salary are respectively less or equal to 24 years old, and less than 25K dollars a year. The householders with age under 18 are deleted from the analysis. Table 1 gives the resulting sample sizes, the estimated proportions of filers (class 1) and non-filers (class2), and the estimated rates of in-migration for filers (class 1) and non-filers (class 2), for 30 counties, in each case obtained with 2000 iterations of the EM algorithm implemented in-house with SAS. Table 1 also gives the proportion of unaccounted migration when (3) is assumed to hold. That is, when the rate of county in-migration for the non-filer is uniformly set to its value for the rate of county in-migration for the filers. Table 2 gives describes the distribution over the counties of statistical estimates for the covariates in the model over the two latent classes.

We see from table 1 that the two latent classes are clearly identifiable as the class of filers (class 1), which includes between 75 % and 100 % of the householders of each county and the class of non-filers, which include the remaining of the householders. The relative sizes of the classes naturally determine their identity. In addition, the statistical description of the distribution of the covariates over each class (table 2) is consistent with the identification. Indeed, home ownership is relatively higher in class 1 than in class 2, and so are the proportions of older and wealthier householders (based on the cut-offs in

table 2), and the proportion of between-county migrants. These higher proportions in class 1 than in class 2 make class 1 a natural candidate for the class of filers, and class 2 for the class of non-filers. Nevertheless, it is important to recall that a latent class analysis does not proceed from any direct information on the latent variable, which we associate with filing status here. Thus we can't make any formal inference on filers or non-filers. But we can make inference on the expected configuration of the files of income tax returns if they are to be representative, and on the ensuing bias, if they are not. In addition we can conjecture on the size of the bias if class2 does in fact correspond to the non-filers.

Based on the results of table 1 and 2, if they are to representative of all the householders in a county, the files of returns must represent householders in two clusters with the same relative proportions as in table 1. The first cluster is typically four times as large as the second, and exhibits comparatively low rates of renters, and of young or low-income householders. In addition, householders in this cluster are relatively unlikely to have migrated in the county recently. By contrast, the second cluster has a high proportion of renters, and of young, low-income householders. Furthermore, a relatively higher percentage of householders in the second cluster have recently migrated in the county. If the tow clusters are not proportionally represented in the files, then distributional bias will ensue. If the cluster corresponding to class 2 in underrepresented, then the estimate for the between-county rate of migration will underestimate the true rate. If indeed, the class of non-filers corresponds exactly to class 2, then the "unaccounted migration" in table 1 is quite substantial (up to, and above 90 %).

Table 1. Prevalence of Latent Classes, Rate Of Migration Per Class, and Unaccounted Migration when Using Only the Rate Of Class 1 (* Indicates Unstable Estimates after 2000 Iterations)

State	County	Sample Size	Prevalence Of Class 2	Rate of In-Migration Class 1	Rate of In-Migration Class 2	Un-Accounted Migration
004	019	3031	.091	.048	.082	.061
005	069	349	.091	.006	.188	.725
006	075	983	.050	.040	.255	.211
006	107	1102	.090	.016	.155	.445
012	011	4104	.047	.024	.093	.118
013	293	141	.194	.027	.145	.465
017	097	2521	.066	.037	.239	.267
018	103	337	.079	.010	.032	.144
019	013	1036	.092	.015	.022	.045
022	031	121	.254	.033	.000*	-.341*
024	009	459	.026	.024	.180	.145
025	013	2268	.063	.010	.029	.100
028	089	342	.086	.030	.496	.573
029	093	178	.056	.013	.188	.435
029	179	95	.067	.022	.792*	.696*
029	221	182	.074	.024	.145	.273
030	029	705	.020	.042	.069*	.012*
030	047	220	.080	.027	.261	.412
031	055	2719	.044	.016	.216	.355
035	035	220	.106	.036	.340	.475
036	005	591	.047	.008	.312	.656
036	087	1442	.047	.026	.026	.001
039	049	3263	.089	.021	.104	.254
041	051	3085	.142	.023	.132	.409

042	057	115	.065	.028	.137	.206
042	107	1263	.094	.000	.151*	1.00*
047	155	418	.036	.035	.649	.382
048	157	271	.014	.034	.769	.235
048	201	2134	.040	.029	.250	.231
048	505	8	.127	.010	.915	.915

Table 2 Distributions Of County Rates Ownership, and Off Age and Salary Above Or Equal Their Cutoffs

	Mean	**Median**	**Upper Quartile**	**Lower Quartile**
Ownership Class 1	.895	.930	.954	.900
Ownership Class 2	.281	.257	.363	.072
25 Years Or More Class1	.992	.996	.999	.987
25 Years Or More Class 2	.767	.781	.891	.691
$25K Or More Class 1	.404	.398	.546	.288
$25K Or More Class 2	.183	.101	.214	.039

7. Conclusion

Our analysis suggests that, because they do not account for householders who do not file income tax returns, the files of income tax returns may not yield an unbiased representation of between-county migration behavior, and thus may bias the population estimates at the county level. Our results indicate that the possible bias is potentially serious (table 1), at least relative to the size of the between-county migrant population, and should not be ignored. Thus the answer to the question in the title of the paper is no, at this time. To give a complete assessment of the possible bias of the population estimates of between-county migrants, it is sufficient to estimate the composition of the population of non-filing householders, in terms of the two clusters identified in the paper. This estimation could be done deductively, by comparing the distributions of the covariates used in the paper, income for example, for the ACS, and for the files of returns. Proceeding this way would avoid the difficulties associated with linking the ACS with the files of returns. If a substantial bias was found, the remedy needs not be onerous. The rates of between-county migration obtained from the files of returns could simply be adjusted on the basis of the estimates obtained from the ACS. Such an approach, if needed, would take full advantage of the data available from the files of income tax returns, and from the ACS.

References

Agresti, A. (1990). *Categorical Data Analysis*, Wiley-Intersience

Bishop, Y. M. M., Fienberg, S. E., Holland, P. W. (1975). *Discrete Multivariate Analysis*, MIT Press

Huang, E. T., Kim, J. (2000). "One-Percent Sample Study Report," U.S. Census Bureau, Statistical Research Division

Sink, L. (2000). "Producing County-Level Characteristics Estimates by the Cohort-Component Method," U.S. Census Bureau, Population Division