ETHNOGRAPHIC EXPLORATORY RESEARCH
REPORT SERIES
(#2007-5)

Effect of Movers on Triple System Estimation

Sally W. Thurston
Alan M. Zaslavsky

Harvard University

Citation: Sally W. Thurston and Alan M. Zaslavsky. (1992) "Effect of Movers on Triple System Estimation." *Proceedings of the Survey Research Methods Section (American Statistical Association): 176-181.*

Report Issued: October 24, 2007

EFFECT OF MOVERS ON TRIPLE SYSTEM ESTIMATION

Sally W. Thurston and Alan M. Zaslavsky, Harvard University
Sally W. Thurston, Department of Statistics, 1 Oxford Street, Cambridge, MA 02138

Key Words: alternative list, ethnographers, mover rate

1 Introduction

The Census Bureau uses a 'capture-recapture' or dual system estimation (DSE) methodology to estimate total population including those missed by the census. The two 'systems' are the census and a Post Enumeration Survey (PES) (Hogan and Wolter 1988). One of the assumptions underlying use of the DSE to estimate population size is that within each poststratum (defined by some set of geographic and demographic variables), being in the census is independent of being in the PES. When these events are not independent, there is a 'correlation bias' which typically leads to underestimation of the number of people who are in neither the census nor the PES. Reasons for the possible failure of this assumption of independence have been discussed (Hogan and Wolter 1988). One method of checking this assumption, or indeed of measuring the correlation bias, makes use of a third source of names and addresses – an alternative list (Marks, Seltzer and Krotki 1974, chapter 7D; Zaslavsky and Wolfgang 1990). By using a third independent source of names and addresses, the 2×2 table underlying the DSE can be expanded into a $2 \times 2 \times 2$ table in which only one of the 8 cells is unknown. Estimates of the unknown cell and thus of correlation bias and total population may be calculated under suitable assumptions. Zaslavsky and Wolfgang (1990) discuss a number of methods to estimate this cell. In this paper we focus on two of these estimates, 'ratio r_1' and 'ratio r_2'.

One such alternative list is formed by combining several administrative lists. A list consisting of portions of lists from the Employment Security, Internal Revenue Service, Selective Service, Veteran's Administration, and driver's licence records was used in the 1988 Administrative List Supplement program conducted by the Census as part of the PES test in St. Louis, Missouri (Zaslavsky and Wolfgang 1990). For further discussion of the use of administrative lists, see also Alvey and Scheuren (1982) and Citro and Cohen (1985, chapter 4). Alternative lists may also be compiled by ethnographers (Vigil 1988; Brownrigg and De La Puente 1992). To date these have not been used for estimation purposes.

One of the challenges posed by triple system estimation is proper cross-classification of cases by inclusion/exclusion in each of three sources. Improper classification may bias the subsequent population estimates. In addition, movers and non-movers may have different coverage rates in each of the sources. Consequently, calculations based on considering movers separately from non-movers are likely to be more accurate than estimates in which movers are either dropped from the triple system estimates, or are combined with non-movers. In general, movers may either be over- or undercounted at a different rate than non-movers (Citro and Cohen 1985, chapter 5) and it is often harder to match movers than non-movers with census records (Schafer 1991).

In this paper, we discuss methods of estimating the number of movers and non-movers, cross-classified by inclusion in census, PES, and alternative list (administrative or ethnographer's lists). We also discuss how these estimates can be used to give total population estimates, and the relative merits of each estimate.

2 Methods of Estimation

We follow the notation of Zaslavsky and Wolfgang (1990), in which the number of people in a given cell is denoted by x_{epa}, where $e = 1$ for people in the census (in the PES block or elsewhere) or 0 otherwise, and p and a are likewise 1 for people in the PES or alternative list respectively, or 0 otherwise. Poststratification is implicit here, so all relationships are assumed to be within a single poststratum (see Diffendal (1988) for details about poststratification used in the PES). In order to distinguish between non-movers, people who move into PES blocks between census and PES days ('in-movers'), and people who move out of PES blocks between census and PES days ('out-movers'), when needed we add a fourth subscript, n, i, or o for non-movers, in-movers, and out-movers respectively.

Zaslavsky and Wolfgang propose five estimators using administrative list data. We restrict consideration to the 'ratio r_1' and 'ratio r_2' estimates because they are based on explicit assumptions about comparability of coverage rates in different subpopulations. Both of these estimators are DSEs in which the census and PES are treated as a single source. The 'ratio r_1' estimate is based on the as-

sumption that the event of being in neither the census nor PES is independent of the event of being in the alternative list. This gives an estimate of the unknown cell as

$\hat{x}_{000} = x_{001} \times (x_{110} + x_{100} + x_{010})/(x_{111} + x_{101} + x_{011})$.

The 'ratio r_2' estimate is based on the same assumption applied to the subpopulation of people who are not in both the census and PES. The rationale is that people captured in both of these sources are "easy to count" and therefore least comparable to those omitted in both. This gives the estimate

$\hat{x}_{000} = x_{001} \times (x_{100} + x_{010})/(x_{101} + x_{011})$.

Once this cell is estimated, the correlation bias between the census and the PES can be calculated, as can coverage rate and total population size estimates. Note that people omitted from both the census and the PES are more likely to be omitted from the alternative list than those included in the census and/or the PES. Thus both estimates of x_{000} are likely to be underestimates.

In making these estimates, we consider the sample of interest either to be PES-A (those residing in the sample blocks on Census day, i.e. the non-movers plus the out-movers), or PES-B (those residing in the sample blocks at PES time, i.e. the non-movers plus the in-movers). In principle PES-A and PES-B are both samples of the same population, and coverage rates for either are estimators of population coverage rates.

3 Administrative Lists

In the following calculations we assume that the administrative lists were last updated at census day and that followup is accurate enough so that movers can be distinguished from non-movers. Problems resulting from the failure of these assumptions will also be discussed.

We subdivide each cell (which has been cross-classified by inclusion in census, PES, and administrative list sources) into non-movers, in-movers, and out-movers (Figure 1). Since people who move into PES blocks between census day and the PES (in-movers) cannot be in the administrative lists for these blocks, $x_{111i} = x_{101i} = x_{011i} = x_{001i} = 0$. Similarly, people who move out of PES blocks after census day but before the PES (out-movers) cannot be in the PES, so $x_{111o} = x_{011o} = x_{110o} = x_{010o} = 0$. Not only do we have no direct information as to the number of people who are in none of the three sources (x_{000n}, x_{000i}, and x_{000o}) but we also do not know the number of in-movers who are in the census, but not in the PES or administrative lists (x_{100i}). The latter cell can not be observed because the only information about the addresses for these people is their census day address, which is not in the PES sample block. Under the stated assumptions, it is possible to count the number of people in all the remaining cells.

3.1 Administrative list estimates using PES-A

When PES-A is the sample of interest, the estimate $\hat{x}_{000}$ using the r_1 estimator is:

$\hat{x}_{000n} + \hat{x}_{000o} = (x_{001n} + x_{001o}) \times (x_{110n} + x_{100n} + x_{100o} + x_{010n})/(x_{111n} + x_{101n} + x_{101o} + x_{011n})$.

With the r_2 estimator,

$\hat{x}_{000} = \hat{x}_{000n} + \hat{x}_{000o} = (x_{001n} + x_{001o}) \times (x_{100n} + x_{100o} + x_{010n})/(x_{101n} + x_{101o} + x_{011n})$.

No other cells need be estimated to calculate $\hat{x}_{000}$.

3.2 Administrative list estimates using PES-B

The r_1 estimate $\hat{x}_{000}$ using PES-B is:

$\hat{x}_{000n} + \hat{x}_{000i} = x_{001n} \times (x_{110n} + x_{110i} + x_{100n} + \hat{x}_{100i} + x_{010n} + x_{010i})/(x_{111n} + x_{101n} + x_{011n})$.

The r_2 estimate is:

$\hat{x}_{000} = \hat{x}_{000n} + \hat{x}_{000i} = x_{001n} \times (x_{100n} + \hat{x}_{100i} + x_{010n} + x_{010i})/(x_{101n} + x_{011n})$.

In both cases, x_{100i} is the only cell which is not directly observable. One method of estimating this cell relies on two assumptions: (1) the number of people who move into the PES blocks is equal to the number of people who move out of the PES blocks in the period between census day and the PES, in each poststratum; and (2) census coverage of in-movers is equal to census coverage of out-movers. Both of these assumptions reflect a view that the size and characteristics of the poststratum are unchanging, i.e. that in-movers are numerically and qualitatively similar to out-movers. Under these assumptions, the number of in-movers in the census equals the number of out-movers in the census, so $x_{101o} + x_{100o} = x_{110i} + x_{100i}$. Then

$$\hat{x}_{100i} = x_{101o} + x_{100o} - x_{110i}. \qquad (1)$$

Another method of estimating this cell relies on the assumption that among people in the census, PES coverage for non-movers is the same as PES coverage for in-movers. Since in-movers cannot be on the administrative lists, the appropriate reference group for them is all non-movers regardless of whether or not they were on an administrative list. Under this assumption we have

$(x_{111n} + x_{110n})/(x_{101n} + x_{100n}) = x_{110i}/x_{100i}$

so

$$\hat{x}_{100i} = \frac{x_{110i}(x_{101n} + x_{100n})}{(x_{111n} + x_{110n})}. \qquad (2)$$

4 Ethnographer's Lists

Ethnographers work intensely in an area, and by getting to know individuals in the neighborhood, compile lists of names which may be more complete than the census or PES address list (Vigil 1988). In the 1990 evaluation programs using ethnographers, the ethnographers stayed in an area from before census day to the PES but typically collected data from May through July.

When using the ethnographer's data as a third source, we have the following situation (Figure 1): as with the administrative lists, out-movers cannot be in the PES, so $x_{111o} = x_{011o} = x_{110o} = x_{010o} = 0$. However, in contrast to the situation with the administrative lists, in-movers can be on an ethnographer's list. The only cells which are unobservable are those which correspond to people who are not on any of the three lists (x_{000n}, x_{000i}, and x_{000o}), and in-movers who are only in the census, x_{100i}.

The equations for the estimates of x_{000} under PES-A are identical whether the administrative lists or ethnographer's lists are used as the third source. Under PES-B however, the estimates are somewhat different.

The r_1 estimate for x_{000} using PES-B is:

$\hat{x}_{000n} + \hat{x}_{000i} = (x_{001n} + x_{001i}) \times (x_{110n} + x_{110i} + x_{100n} + \hat{x}_{100i} + x_{010n} + x_{010i})/(x_{111n} + x_{111i} + x_{101n} + x_{101i} + x_{011n} + x_{011i})$.

The r_2 estimate for x_{000} using PES-B is:

$\hat{x}_{000n} + \hat{x}_{000i} = (x_{001n} + x_{001i}) \times (x_{100n} + \hat{x}_{100i} + x_{010n} + x_{010i})/(x_{101n} + x_{101i} + x_{011n} + x_{011i})$

As with the administrative list, x_{100i} is unobservable, and could be estimated using either of the two methods previously described. Using the first method, assuming that the number of in-movers in the census equals the number of out-movers in the census, we can estimate x_{100i} by

$\hat{x}_{100i} = (x_{101o} + x_{100o}) - (x_{111i} + x_{101i} + x_{110i})$.

Using the second method, in which we assume that PES coverage for non-movers in the census is the same as PES coverage for in-movers in the census,

$(x_{111n} + x_{110n})/(x_{101n} + x_{100n}) = (x_{111i} + x_{110i})/(x_{101i} + x_{100i})$.

so

$\hat{x}_{100i} = (x_{111i} + x_{110i}) \times (x_{101n} + x_{100n})/(x_{111n} + x_{110n}) - x_{101i}$.

5 Population Estimates from one Ethnographer's Site

There were 29 sites used in the 1990 evaluation program using ethnographers. Four of these sites were put into the PES and PES data were collected, but the four sites were not actually used for PES evaluation. It is these four sites for which the potential exists for triple system estimation. It should be noted that the selection of the ethnographic sites was based on where the ethnographers lived at that time, rather than being randomly sampled.

Figure 3: Estimated cell counts based on one ethnographer's site, including movers

All mover categories:

On Ethnographer's List		
	In PES	Out of PES
In C	209	5
Out of C	43+4	8

Not on Ethnographer's List		
	In PES	Out of PES
In C	2	5 + 4
Out of C	0+1	2

Preliminary data from one ethnographer's site were used to demonstrate these methods, and to compare triple system estimates, when movers were included as well as when movers were excluded, with the DSE (Figure 2). In these calcuations, PES-B was considered the population of interest, and the r_2 estimator was used. All poststrata were combined in the following estimates.

Using ethnographers' lists and considering movers separately, x_{100i} must first be estimated. In this case, (1) gave an estimate of 4 people. The number of people missed by all 3 sources was estimated to be $8 \times (9+1)/(47+5) = 1.538 \approx 2$ people (Figure 3). Using this estimator, an estimated 58 people, or 20.49 percent of the population in this site was missed by the census.

When the r_2 estimator is used and movers are dropped from the roster, the number of people estimated to be missed by all 3 sources was $(8 \times 5)/48 = .833 \approx 1$ person. In this case, an estimated 52 people, or 19.05 percent of the population in this site was missed by the census.

In comparison, dual system estimation (ignoring movers) gives an estimate of $(43 \times 10)/211 = 2.27 \approx 2$ people missed by two sources (census and PES) (Figure 4). The estimated number of people missed by the census is then 45, or 16.92 percent of the population in the site.

6 Discussion

In the 1988 Test Census PES in St. Louis, administrative lists were used as a third source of cases.

Figure 4: Estimated cell counts based on one ethnographer's site using dual system estimation only

All mover categories:

	In PES	Out of PES
In C	209 + 2	5 + 5
Out of C	43	2

However the lists comprising the administrative lists were last updated before census day, and in most cases well before. The population of interest was PES-B. Follow-up was done after the PES, to determine whether people not in the PES were in the block at PES time; everyone who was not was dropped from the roster.

A general problem resulting from the outdated nature of the administrative lists concerns people who move into a PES block before census day but after the administrative lists were compiled, and who are on neither the census nor PES lists. Had the administrative lists been current at census day, these people would have contributed to x_{001}, but instead they became part of x_{000}, making $\hat{x}_{000}$ too small. This results in an estimate of the coverage rate which is too large, and underestimation of the true population size in the block.

Non-movers who are on an administrative list but not in the census or PES (x_{001n}) may be more difficult to locate than non-movers in the census and/or PES. The former group of people would have a greater chance of being misclassified as an outmover, and when PES-B is used, dropped from the roster. This too contributes to an overstated coverage rate.

In estimates based on PES-B, whether based on administrative or ethnographic lists, x_{100i} is unobservable and must be estimated. The assumption underlying estimator (2), is not likely to be accurate, as we would expect that the PES coverage rate for in-movers is smaller than that for non-movers. This too has the effect that our estimates of x_{100i} and ultimately of x_{000} are too small. One of the assumptions underlying estimator (1), that the number of in-movers equals the number of outmovers between census day and the PES, may be somewhat inaccurate, especially when the number of movers in an area is small. However, unless there are systematic population shifts between Census and PES time, (1) is likely to be less biased than (2).

PES-B estimates using ethnographer's lists may be more accurate than PES-B estimates using administrative lists. The former also rely on an estimate of x_{100i}, but since their lists are updated after census day, some in-movers are seen by the ethnographers, and consequently the estimate of x_{100i} is likely to be more accurate than when administrative lists are used. In-movers who are not in the PES are likely to be missed by the ethnographers more easily than in-movers in the PES since presumably they are 'harder to count', and consequently x_{101i} and x_{001i} are likely to be too small. Underestimation of x_{001i} leads to underestimation of x_{000}, while underestimation of x_{101i} leads to overestimation of x_{000}. Under the assumption that in-movers who are not in the census were more likely to be missed than in-movers who were in the census somewhere, $\hat{x}_{001i}$ is underestimated by a greater degree than $\hat{x}_{101}$. Thus on balance we might expect our estimate of x_{000} to be somewhat too small, leading again to an overstated coverage rate.

Estimates based on PES-A also lead to underestimates of x_{000}, particularly when administrative lists are used. It may be especially difficult to locate outmovers who were not in the census but were on an administrative list (x_{001o}). If not found at followup, the dated nature of the administrative list makes it impossible to determine whether such individuals were in the PES block at census day, so these individuals were dropped from the roster. In addition, non-movers who were on an administrative list but were in neither census nor PES (x_{001n}), and who could not be resolved at followup, may have been misclassified as out-movers and also dropped from the roster. By dropping these people, we underestimate x_{001}, which leads to an underestimate of x_{000}.

The problems with PES-A may be ameliorated when ethnographer's lists are used. If the ethnographer's lists are reliable, people who were on only this list at census day were quite certain to have been in the PES block. Furthermore, with these lists it may be possible to classify each person on the ethnographic list as a non-mover or an out-mover. In this case, if PES-A is used, the ethnographer estimates are preferable to those estimates based on the administrative lists. It should be noted, however, that PES-A data were not considered critical in the 1990 ethnographic program, and may be less reliable than PES-B data.

Defining movers has not been considered a part of ethnographic studies to date. Consequently, attempts to distinguish movers from non-residents has been somewhat problematic. However, defining mover status reliably should be possible when data are collected by ethnographers. In contrast, the outdated nature of administrative lists makes it unlikely that mover status could be defined accu-

rately with reference to these lists.

Preliminary data from the one ethnographer's site suggests that dual system estimation may underestimate the true population size. In this ethnographic site, triple system estimation led to a 4 percent increase in the estimated number of people who were missed by the census, when compared with a DSE. Further work with other sites would be needed to see whether this holds true in other areas.

Regardless of the source of names for the alternative list, it should be noted that when using PES-A, census coverage among people in the PES only applies to non-movers, since out-movers cannot be in the PES. Thus this coverage rate is not representative of the population as a whole.

Census experience shows that the non-match rate among movers is typically much greater than among non-movers (Schafer 1991). Although a large part of the reason for this high non-match rate is due to matching error, movers may be more prone to both over and undercounting (Citro and Cohen 1985, chapter 5). One way to improve the population estimates may be to consider movers separately from non-movers, drawing inferences about movers only from the mover population. In areas with a large number of movers, a separate triple system estimate for movers, combined with triple system estimates from non-movers, may lead to more accurate estimates. However, when the number of movers is small, we would expect a large sampling variability from estimates based on movers alone. In this case, some way to pool estimates of movers across similar poststrata would be desirable.

7 Conclusions

Proper consideration of movers when using triple system estimation may lead to both more accurate population estimates than are possible using a DSE, and to a way to measure the correlation bias in the DSE. Estimates based on PES-A and estimates based on PES-B are both likely to give underestimates of x_{000}. If the coverage rate among people in the PES is of interest, estimates should be based on PES-B so that movers are included in this estimate.

Estimates using ethnographer's lists have the potential to be more accurate than estimates using administrative lists, in part because the ethnographer's lists refer to a more relevant time period, and have more information as to the exact time each person resides in the block of interest. However if ethnographer's estimates using PES-B are desired, it is important to ensure that these lists are reasonably accurate at the time of the PES.

8 References

Alvey, W. and Scheuren, F. (1982), "Background for an Administrative Record Census", *American Statistical Association Proceedings of the Social Statistics Section* 137-146.

Brownrigg, L. and De La Puente, M. (1992) "An Analysis of the Census Underenumeration of Racial and Ethnic Minorities: Evidence from Small Area Studies", *American Statistical Association Proceedings of the Section on Survey Research Methods*, forthcoming.

Citro, C.F. and Cohen M.L., editors (1985), *The Bicentennial Census: New Directions for Methodology in 1990*, Washington, D.C.: National Academy Press

Diffendal, G. (1988), "The 1986 Test of Adjustment Related Operations in Central Los Angeles County", *Survey Methodology*, 14, 71-86.

Hogan, H. and Wolter, K. (1988), "Measuring Accuracy in a Post-Enumeration Survey", *Survey Methodology*, 14, 99-116.

Marks, E.S., Seltzer, W. and Krotki, K.J. (1974), *Population Growth Estimation: A Handbook of Vital Statistics Measurement.* New York: The Population Council.

Schafer, J.L. (1991), "A Comparison of the Missing-Data Treatments in the Post-Enumeration Program", *Journal of Official Statistics*, 7, 475-498.

Vigil, J.D. (1988), "Counting the Hard-to-Enumerate Population", *Proceedings, Bureau of the Census Annual Research Conference*, 25-27.

Zaslavsky, A.M. and Wolfgang, G.S. (1990), "Triple System Modeling of Census, Post-Enumeration Survey, and Administrative List Data". *American Statistical Association Proceedings of the Section on Survey Research Methods*, 668-673.

Figure 1: Estimates using Administrative (A) and Ethnographer's (E) lists

Non-movers:

On Alternative List		
	In PES	Out of PES
In C	x_{111n} √	x_{101n} √
Out of C	x_{011n} √	x_{001n} √

Not on Alternative List		
	In PES	Out of PES
In C	x_{110n} √	x_{100n} √
Out of C	x_{010n} √	x_{000n} x

In-movers:

On Alternative List		
	In PES	Out of PES
In C	x_{111i} A=o E=√	x_{101i} A=o E=√
Out of C	x_{011i} A=o E=√	x_{001i} A=o E=√

Not on Alternative List		
	In PES	Out of PES
In C	x_{110i} √	x_{100i} x
Out of C	x_{010i} √	x_{000i} x

Out-movers:

On Alternative List		
	In PES	Out of PES
In C	x_{111o} o	x_{101o} √
Out of C	x_{011o} o	x_{001o} √

Not on Alternative List		
	In PES	Out of PES
In C	x_{110o} o	x_{100o} √
Out of C	x_{010o} o	x_{000o} x

Key:
√ = seen, x = unseen, but exist, o = 0 by definition
For cells for which the status differs for administrative and ethnographer's lists, differences are indicated.

Figure 2: Preliminary data from one ethnographer's site

Non-movers:

On Ethnographer's List		
	In PES	Out of PES
In C	209	5
Out of C	43	8

Not on Ethnographer's List		
	In PES	Out of PES
In C	2	5
Out of C	0	?

In-movers:

On Ethnographer's List		
	In PES	Out of PES
In C	0	0
Out of C	4	0

Not on Ethnographer's List		
	In PES	Out of PES
In C	0	?
Out of C	1	?

Out-movers:

On Ethnographer's List		
	In PES	Out of PES
In C	0	0
Out of C	0	0

Not on Ethnographer's List		
	In PES	Out of PES
In C	0	4
Out of C	0	?