ETHNOGRAPHIC EXPLORATORY RESEARCH
REPORT SERIES
(#2007-4)

Using Ethnographic Data to Evaluate Dual System Estimates

Sally W. Thurston

Harvard University

Citation: Sally W. Thurston (1995) *Using Ethnographic Data to Evaluate Dual System Estimates.* Report prepared in partial fulfillment of Joint Statistical Agreement 91-31 with Harvard University. January 4, 1995.

Report Issued: October 24, 2007

USING ETHNOGRAPHIC DATA TO EVALUATE DUAL SYSTEM ESTIMATES *

Sally W. Thurston †

January 4, 1995

Abstract

Administrative or ethnographic lists provide a third source of names and addresses which can be used to expand the 2×2 table underlying the dual system estimate into a 3-way table ($2 \times 2 \times 2$) in which only one cell is unknown. Use of these lists makes it possible to check the independence assumption underlying the dual system estimate, and to estimate the correlation bias if this assumption is not met. I discuss ways in which knowledge of people's mover status can be incorporated into triple system estimates, and the relative merits of using administrative or ethnographic lists for population estimation. Population estimates and coverage rates for two ethnographic sites are compared with and without use of the ethnographers' lists. The statistical dependency of the census and PES in these two sites is evaluated.

Key Words: triple system estimation, movers, census undercount

*Submitted to the Department of Statistics, Harvard University as a qualifying paper, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

†Department of Statistics, Harvard University. This research was supported in part by Joint Statistical Agreement 91-31 between the Bureau of the Census and Harvard University. The author would like to thank Alan M. Zaslavsky, Elizabeth Martin, and Gregg Diffendal for their assistance in this work.

Contents

List of Figures

List of Tables

1 Introduction

In 1990, the U.S. Census Bureau used 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 estimating the statistical dependency between the census and PES, makes use of a third source of names and addresses – an alternative list (Marks, Seltzer and Krotki 1974, chapter 7D; Zaslavsky and Wolfgang 1993). 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 the count in only one of the 8 cells is unknown. Estimates of the number of people in this cell and of total population may also be calculated under suitable assumptions. Zaslavsky and Wolfgang (1993) discuss a number of methods to estimate the number in this cell. In this paper we focus on two of these estimates, 'DSE:$(E \bigcup P) \times A$' and 'DSE:$(E \Delta P) \times A$'.

One such alternative list may be 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 U.S. Bureau of the Census as part of the PES test in St. Louis, Missouri (Zaslavsky and Wolfgang 1993). 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, de la Puente 1993, Martin and de la Puente 1993). 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 had connections to the people in the area, either by having worked with members of the community, or in living nearby. The ethnographers typically collected data from May through July.

One of the challenges posed by triple system estimation is proper cross-classification of cases by inclusion/exclusion in each of three sources. 'Movers' are persons or households who change their residence between census day and the time of the PES. Movers are particularly difficult to classify because information from two different locations must be linked, and census day residence may be uncertain. It is often harder to match movers than non-movers with census records (Schafer 1991). In general, movers may either be over- or under-counted at a different rate than non-movers (Citro and Cohen 1985, chapter 5). 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 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 discuss how these estimators can be used to give total population estimates, and the relative merits of each estimator. These methods are applied to two sites in which data were collected by ethnographers – one site in rural North Carolina and the other in urban Florida. People who were erroneously enumerated are discussed in a separate section. Estimates of the number of people missed and of total population are calculated based only on people who were correctly enumerated in the site. We also discuss ways to estimate the statistical dependency of the census and PES, and apply these to the two sites.

2 Triple System Estimation When There are Movers

We follow the notation of Zaslavsky and Wolfgang (1993), 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 the census and the PES ('in-movers'), and people who move out of PES blocks between the census and the PES ('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 (1993) propose a number of estimators using administrative list data. We restrict consideration to the DSE:$(E \bigcup P) \times A$ and DSE:$(E \Delta P) \times A$ estimators 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 DSE:$(E \bigcup P) \times A$ estimator is based on the assumption 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 number of people missed by all three sources as:

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

The DSE:$(E \Delta P) \times A$ estimator is based on the same assumption applied to the subpopulation of people who are in either the census or PES, but not both. The rationale is that people captured in both census and PES are 'easy to count' and therefore least comparable to those omitted in both. This gives the estimate

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

Once this cell is estimated, the estimated correlation bias between the census and the PES can be calculated, as can estimates of coverage rate and

total population size. 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 estimate population coverage rates.

When movers are included in these estimates, we subdivide each cell (which has been cross-classified by inclusion or exclusion in census, PES, and administrative list sources) into non-movers, in-movers, and out-movers, creating a 4-way table (Figure 1). Estimates of x_{000} involve adding the number of in-movers or out-movers (depending on whether PES-A or PES-B is the sample of interest) to the number of non-movers, in each of the 8 cells of the 3-way table. People who move out of PES blocks after census day but before the PES (out-movers) are not directly seen by the PES, but through interviewing of current residents and neighbors, information about these people is collected by the PES.

2.1 Estimation Using Administrative Lists

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

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} √	x_{101o} √
Out of C	x_{011o} √	x_{001o} √

Not on Alternative List		
	In PES	Out of PES
In C	x_{110o} √	x_{100o} √
Out of C	x_{010o} √	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 1: Estimates using Administrative (A) and Ethnographer's (E) lists

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 out-mover, and when PES-B is used, dropped from the roster. This too contributes to an overstated coverage rate.

Not all cells in the 4-way table (incorporating movers) can be observed. Since people who move into PES blocks between census day and the PES (in-movers) cannot be in the administrative lists for these blocks, $x_{111i} = x_{101i} = x_{011i} = x_{001i} = 0$. 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.

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.

When PES-A is the sample of interest, the DSE:$(E \bigcup P) \times A$ estimator is:

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

With the DSE:$(E \Delta P) \times A$ estimator,

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

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

When PES-B is the sample of interest, under the DSE:$(E \bigcup P) \times A$ estimator:

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

With the DSE:$(E \Delta P) \times A$ estimator:

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

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 $x_{111o} + x_{101o} + x_{110o} + x_{100o} = x_{110i} + x_{100i}$. Then

$$\hat{x}_{100i} = x_{111o} + x_{101o} + x_{110o} + x_{100o} - x_{110i}. \quad (1)$$

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

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

2.2 Estimation Using Ethnographer's Lists

When using the ethnographer's data as a third source, the situation is somewhat different (Figure 1). In contrast to the situation with 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 (in another block), x_{100i}.

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

The DSE:$(E \bigcup P) \times A$ estimate for x_{000} using PES-B is:

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

The DSE:$(E \Delta P) \times A$ estimate for x_{000} using PES-B is:

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

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

$$\hat{x}_{100i} = (x_{111o} + x_{101o} + x_{110o} + x_{100o}) - (x_{111i} + x_{101i} + x_{110i}). \tag{3}$$

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

$$\hat{x}_{100i} = ((x_{111i} + x_{110i})(x_{101n} + x_{100n})/(x_{111n} + x_{110n})) - x_{101i}. \qquad (4)$$

2.3 Comparisons of Administrative and Ethnographers Lists as a Source for Triple System Estimation

PES-B estimates using ethnographers' lists are likely to be more accurate than PES-B estimates using administrative lists, assuming equal sample sizes. Both rely on an estimate of x_{100i}, but administrative lists do not include any in-movers, whereas some in-movers are seen by ethnographers. Consequently $\hat{x}_{100i}$ is likely to be more accurate when ethnographers' lists are used than when administrative lists are used. In addition, in-movers who were missed by both the census and PES may be on an ethnographer's list, but cannot be on an administrative list for the block of interest. When ethnographers' lists are used, these people are seen as part of the x_{001i} cell. When administrative lists are used, they become part of the x_{000} cell, leading to an estimate of x_{000} which is too small.

When ethnographers' lists are used for PES-B, we also expect $\hat{x}_{000}$ to be too small, but to a lesser extent 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 they are probably 'harder to count'. Consequently, people who would be part of the x_{101i} cell may become part of x_{100i}, while people who would be part of x_{001i} become part of x_{000i}. The former leads to an overestimation of x_{000}, while the latter leads to an underestimation 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, x_{001i} is underestimated by a greater degree than x_{101i}.

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}. Since the PES data collection was geared towards the PES-B population, information about census day residents who had moved out of the block was not considered as crucial as information about current residents. Consequently, people who would be categorized as x_{111o} may become part of x_{101o}, and similarly for x_{110o} and x_{100o}, x_{011o} and x_{001o}, and x_{010o} and x_{000o}. While the first two misclassifications have no effect on the estimate of x_{000}, a decrease in x_{011o} corresponding to an increase in x_{001o} leads to an overestimation of x_{000}, while a decrease in x_{010o} corresponding to an increase in x_{000o} leads to an underestimation of x_{000}.

When administrative lists are used, out-movers who are not in the census but are on an administrative list (x_{001o}) are followed up. If these individuals are not found, 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 the St. Louis study. 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.

3 Results from Two Ethnographers' Sites

There were 29 sites used in the 1990 ethnographic evaluation program, of which 28 were in the continental U.S. All sites were selected to be in areas which were difficult to enumerate, and which had a large concentration of minorities, including Blacks, Hispanics, Asians, and American Indians (de la Puente 1993, Martin and de la Puente 1993). Each site consisted of about 100 housing units, usually in one or two blocks. 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 should be noted that ethnographic sites were selected in areas where the ethnographers already had a relationship with people in the area. Since this was neither a random sample of the country nor a random sample of areas with high undercount, results from these sites may not be generalizeable to a population.

Processing ethnographic data for this study posed some difficulties. The 1990 ethnographic data were not collected for quantitative purposes, and defining mover and residency status were not part of the data which were collected. Consequently, attempts to categorize each person into a cell of the 3-way table, and to determine the mover status of each individual, were sometimes problematic.

Other aspects of the ethnographic study contributed to initial errors and uncertainties in the data. In particular, the coding required of the ethnographers was very meticulous, making room for errors, and errors were made. The error rate may be reduceable in future years if the coding can be simplified. Also, the ethnographic data was linked to the census and PES data later than when the census and PES data were linked, so the ethnographers were unable to comment on some uncertain PES cases. In addition, the ethnographers did the initial three-way matching, but they are not trained matchers and the results did not necessarily agree with the results that trained matchers would have obtained. It should be noted that after the initial matching by ethnographers, some of the data was clerically matched and the entire data set was then reviewed by Gregg Diffendal (a Census Bureau statistician with long and intimate experience with PES methods) and corrections were made.

In one of the four ethnographic sites which overlapped with the census and PES, the ethnographer did not proceed far enough with the coding to allow for quantitative estimates. In a second site, some missing coding has lead to uncertainties in the data, and we did not attempt to resolve these uncertainties. We have examined the data from the remaining two sites.

The first site is in rural North Carolina, and is part of the community of the Waccamaw Siouan Indian tribe (Lerch, 1992). Eighty-seven percent of the residents were Indian and the remaining 13% were white spouses and children. A household in this site does not consist of a stable set of people separate from people in other households. Rather, households form and then regroup as people move to other addresses (Lerch 1992). Adult children of residents in the site often live in mobile homes or newly-built houses close to the house of their parents. Mail is delivered to numbered mailboxes along the side of the road, and most of the mailboxes serve more than one household.

The second site is an urban site in downtown Fort Lauderdale, Florida. Haitians comprised about 70% of the residents in the sample area, Blacks (African Americans) about 25%, and 5% were of other races (Wingerd, 1992a). The site is one block away from a major drug dealing area, and drug transactions were common in the site. There were bullet holes and multiple deadbolts on doors, many drawn curtains, and people carried guns and knives (Wingerd, 1992b). The Haitians were Creole-speaking recent entrants, and often did not speak English. Although many of them were undocumented aliens, they were more approachable by the ethnographer than were the Blacks. There was a high rate of mobility among the Haitians, as some found better places to live and some returned temporarily to Haiti. The Black residents of the site were suspicious of anything relating to the government (Wingerd 1992a). Of the 4 census forms filled out from this site, none were from the Black community (Wingerd, 1992b).

In these two sites, as in other ethnographic sites, the ethnographers consistently found people who were missed by the census and/or PES. A common theme in the ethnographers' reports was that the ethnographers were able to enumerate hard-to-count people because they had gained the trust of the residents (Hamid 1992, Lerch 1992, Wingerd 1992a).

Non-movers:

On Ethnographer's List		
	In PES	Out of PES
In C	213	5
Out of C	46	8

Not on Ethnographer's List		
	In PES	Out of PES
In C	2	1
Out of C	0	?

In-movers:

On Ethnographer's List		
	In PES	Out of PES
In C	0	0
Out of C	3	0

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

Out-movers:

On Ethnographer's List		
	In PES	Out of PES
In C	5	0
Out of C	3	0

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

Table 1: Counts from Site 1 (Lerch)

3.1 Population Estimates

In the first site (Table 1), 275 non-movers, 3 in-movers, and 8 out-movers were correctly enumerated by the census, PES, and/or ethnographers. All but three of these people were seen by the ethnographer. In the second site (Table 2), 74 non-movers, 17 in-movers, and 1 out-mover were correctly enumerated by at least one source. In this site the ethnographer only missed one person who was found by another source.

In each site the DSE:$(E \bigcup P) \times A$ and DSE:$(E \Delta P) \times A$ estimators gave nearly identical population estimates. Less than one person was estimated to have been missed by all three sources, under all estimators and in both sites.

Non-movers:

On Ethnographer's List		
	In PES	Out of PES
In C	40	19
Out of C	6	8

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

In-movers:

On Ethnographer's List		
	In PES	Out of PES
In C	4	1
Out of C	7	5

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

Out-movers:

On Ethnographer's List		
	In PES	Out of PES
In C	0	0
Out of C	1	0

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

Table 2: Counts from Site 2 (Wingerd)

In both sites, a DSE would have underestimated the population and overstated the coverage rate, when compared with any of the triple system estimates. Ignoring movers in a DSE is one reason for these differences. A second reason was that the ethnographers found a greater number of people who were missed by both census and PES than would have been estimated by a DSE. In the first site, under the assumption of independence of census and PES, the DSE would have predicted that 1.28 people were missed by both census and PES, whereas the ethnographer found 8 such people, all of whom were nonmovers. In the second site, the DSE would have predicted that 3.00 people were missed by both census and PES. In this site the ethnographer found 8 non-movers as well as 5 in-movers who were missed by both census and PES.

In the first site, a DSE would have estimated the population to be 268, while a triple system estimate ignoring movers would have estimated a population of 275 (Table 3). Because 8 people moved out of the site, the PES-A population estimate was 283 people. Using PES-B, $\hat{x}_{100i} = 5$ under (3), and $\hat{x}_{100i} = 0$ under (4), giving population estimates of 284 for the DSE:$(E \bigcup P) \times A$ estimator (283 for the DSE$(E \Delta P) \times A$ estimator) under (3), and 278 under (4). A DSE would have estimated the census coverage rate to be 82%. The coverage rate was estimated to be about 80% under all the triple system estimators.

In the second site, a DSE would have estimated the population to be 69, while the triple system population estimate ignoring movers was 74. Only one person moved out of the site, so the PES-A population estimate was 75. Because so many people moved into the site and so few moved out, when PES-B is used, (3) estimated that $\hat{x}_{100i} = -5$, so we take $\hat{x}_{100i}$ to be 0. Under estimator (4), $\hat{x}_{100i} = 1$. These gave population estimates of between 91 and 93, depending on whether the DSE:$(E \bigcup P) \times A$ or DSE:$(E \Delta P) \times A$ estimators are used. In this site, the estimated coverage rate would be 87% under a DSE, 81% under a triple system estimator ignoring movers and 80% under the triple system estimator using PES-A. The triple system estimator using PES-B gave an estimated census coverage rate of 71 to 72%.

In each site, the ethnographers found 8 non-movers missed by both census and PES. In the first site, the 8 people consisted of 6 households, of which 2

	Site 1 (Lerch)					
	$\hat{x}_{000}$		Estimated population		Census coverage	
DSE:			268		82.46 %	
TSE:	$(E \bigcup P)$	$(E \Delta P)$	$(E \bigcup P)$	$(E \Delta P)$	$(E \bigcup P)$	$(E \Delta P)$
ignore movers:	0.091	0.157	275	275	80.36 %	80.36 %
include movers:						
PES-A:	0.088	0.148	283	283	79.86 %	79.86 %
PES-B						
w/ assumption (3):	0.240	0.889	283	284	79.86 %	79.58 %
w/ assumption (4):	0.090	0.148	278	278	79.50 %	79.50 %

	Site 2 (Wingerd)					
	$\hat{x}_{000}$		Estimated population		Census coverage	
DSE:			69		86.96 %	
TSE:	$(E \bigcup P)$	$(E \Delta P)$	$(E \bigcup P)$	$(E \Delta P)$	$E \bigcup P)$	$(E \Delta P)$
ignore movers:	0.123	0.320	74	74	81.08 %	81.08 %
include movers:						
PES-A:	0.121	0.308	75	75	80.00 %	80.00 %
PES-B						
w/ assumption (3):	0.169	0.394	91	91	71.43 %	71.43 %
w/ assumption (4):	0.338	0.788	92	93	71.74 %	70.97 %

Table 3: Estimates of population, coverage rates, and number of people missed by three sources

households (3 people) were missed completely by the census and PES, and 4 households (5 people) had other members who were found by the census and/or the PES. In the second site, the 8 people came from 6 households, of which 5 households (7 people) were missed completely by the census and PES, and 1 household (1 person) had other household members who were found by the census and/or PES.

3.2 Erroneous Enumeration

In the first site, the census found a total of 277 people, of whom 226 were correctly enumerated (non-movers plus out-movers), and 51 were erroneously enumerated. Of the 51 erroneously enumerations, 31 were duplicate census-only records, 13 were misgeocoded census-only records, 8 were either temporary residents, weekend-only residents or non-residents, and 3 were people for whom the ethnographer had no information. The erroneous enumeration rate was among census records was 51/(51+226) or 18%.

The second site contained 65 people who were on a census list. Sixty of these were correctly enumerated in the site, and five were correctly enumerated in-movers who were found by the census in another block. There were no erroneous enumerations among census records in this block.

3.3 Statistical Dependency of the Census and PES

The odds ratio in a 2×2 table is $(x_{00}x_{11})/(x_{01}x_{10})$, where the first subscript is 1 for inclusion in the first source and 0 otherwise, and likewise the second subscript indicates inclusion or exclusion in the second source. The subtable of non-movers who are on an ethnographer's list is a 2×2 table in which the counts in all four cells are observed. Under the assumption of independence of the census and PES, the odds ratio in this subtable should be about one. The observed odds ratio for this subtable was 7.41 in the first site, and 2.81 in the second site. Ninety-five percent confidence intervals for the odds ratio were generated under six methods (Table 4). Three methods modeled the individual as the unit, and did not use poststratification. In

Method	Site 1 (Lerch)	Site 2 (Wingerd)
parametric bootstrap	(2.19, 37.58)	(0.79, 12.30)
bootstrap	(2.16, 35.98)	(0.80, 11.96)
exact	(2.01, 29.84)	(0.72, 11.19)
exact w/ poststratification on age	(1.93, 28.35)	(0.60, 9.84)
exact w/ poststratification on age and race	(1.91, 28.54)	(0.62, 10.82)
exact based on households	(1.52, ∞)	(0.51, 24.96)

Table 4: Ninety-five percent confidence intervals for the odds ratios in the fully observed subtable of people on ethnographers' lists

the first method (a parametric bootstrap), the counts in each of the four cells were modeled as being Poisson random variables with the mean in each cell equal to the observed count. One observation from the relevant Poisson distribution was drawn for each of the four cells, the odds ratio for that table was calculated, and this was repeated 100,000 times. The resulting odds ratios were sorted, and 2.5 percent of the observations from each tail were removed. The remaining range of the observations gave a 95 percent confidence interval.

In the second method (a non-parametric bootstrap), a sample was drawn with replacement from the observed counts in the fully observed subtable, with each unit in the sample retaining identification as to the cell from which it came. The odds ratio was computed for the sample, and a 95 percent confidence interval was created in the manner just described. The third method was a Fisher exact confidence interval, calculated using StatXact (Mehta and Nitin 1991). The 95% exact confidence interval for the first site was (2.01, 29.84). The p-value for the null hypothesis that the odds ratio is one, versus the alternative that it is less than 1, is .0009. For the second site, the 95% exact confidence interval is (0.72, 11.19), with a p-value for the same hypotheses of .0779.

Within each site, these three intervals were very similar. In the first site, none of the intervals included one (and in fact none included two), whereas

in the second site, all three intervals included one.

There are several reasons why these first three methods might overestimate the statistical dependency of the census and PES. One reason is that the individuals within a site may come from a mixture of subpopulations, each with different probabilities of capture by the three sources. If this is the case in these sites, calculating the odds ratios separately for the different subpopulations and then calculating a common odds ratio for the subpopulations should result in more accurate estimates. We attempted to address this issue by poststratifying on first one, then two variables.

The fourth method we used to calculate a confidence interval used poststratification on age, using year of birth (before 1960, and 1960-1990). In each site, this separated the individuals into two nearly equal groups (with one missing year of birth in the second site). An exact confidence interval for the common odds ratio was calculated. The fifth method used poststratification on age and race. The 95 percent confidence intervals for the common odds ratio using one or two poststratifications were very similar to each other and to the interval when poststratification was not used, although intervals using poststratification had a smaller lower endpoint than intervals without poststratification.

A second reason that the odds ratio may have been overestimated using the methods just described is that individuals within a household may not be independent. If one individual is found by all three sources for example, it is likely that other individuals in the household were also found by all three sources. To address this issue, the sixth method of generating confidence intervals for the odds ratio used the household as the unit of analysis. A bootstrap was used to sample households with replacement from the households within the site, and all individuals within each sampled household were included in the sample. The odds ratio for the individuals was calculated. This was repeated 10,000 times, then the odds ratios were sorted and the range of the middle 95% of the odds ratios gave the 95% confidence interval. As expected, the clustering of individuals within households gave a wider interval. The interval for the first site was $(1.52, \infty)$, while for the second site it was (0.51, 24.96).

It should be noted that while the odds ratios in the subtable of people on the ethnographers lists may not be one in either site, the observed odds ratio in the other fully observed subtables are even further from one. This indicates that the ethnographers' lists are not at all independent from either the census or PES.

4 Discussion

In estimates based on PES-B, whether based on administrative or ethnographic lists, x_{100i} is unobservable and must be estimated. The assumption underlying estimators (2) and (4) – that among people in the census, PES coverage for non-movers equals PES coverage for in-movers – is not likely to be accurate. We would expect that the PES coverage rate for in-movers is smaller than that for non-movers. This has the effect that our estimates of x_{100i} and ultimately of x_{000} are too small. One of the assumptions underlying estimators (1) and (3), that the number of in-movers equals the number of out-movers between census day and the PES, may be inaccurate, especially when the number of movers in an area is small. This was the case for the second ethnographic site examined here. However, unless there are systematic population shifts between the census and the PES, (1) and (3) are likely to be less biased than (2) and (4) on average.

Determining who is a mover and who is a non-mover has not been considered a part of ethnographic studies to date. Consequently, distinguishing movers from non-residents in the sites was sometimes difficult. However, determining 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 accurately with reference to these lists.

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, separate triple system estimates for movers, combined with triple system estimates for non-movers, may lead to more accurate population 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.

5 Conclusions

Use of a triple system estimator is likely to lead to more accurate estimates of population and of census coverage rates than are possible using a DSE. In addition, an estimate of the statistical dependency of the census and PES is possible using triple system estimation. Proper consideration of movers may give more accurate estimates of population and coverage rates than when movers are dropped from the roster.

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 ethnographers' lists have the potential to be more accurate than estimates using administrative lists, in part because the ethnographers' 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. If ethnographers' estimates using PES-B are desired, it is important to ensure that these lists are reasonably accurate at the time of the PES.

Since estimates based on administrative lists tend to underestimate x_{000} to a greater extent than estimates based on ethnographers lists for both PES-A and PES-B, using ethnographers' lists is preferable to using administrative lists when possible. A larger sample size within each site would be desirable in order to obtain more precise estimates. Random selection of sites with

high undercount would be desireable if inference is to be made about sites with high undercount. Due to the difficulties in collecting and processing ethnographic data however, it may be necessary to use administrative lists or some other source of names and addresses when large sample sizes are needed. Using an updated administrative list containing changes to the earlier version for a follow-up at a later time should help to improve the accuracy of estimates based on administrative lists.

6 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) "Alternative Enumeration Methods and Results", *American Statistical Association Proceedings of the Section on Survey Research Methods*, 199-204.

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

de la Puente, M. (1993), "A Multivariate Analysis of the Census Omission of Hispanics and Non-Hispanic Whites, Blacks, Asians and American Indians: Evidence from Small Area Ethnographic Studies", *American Statistical Association Proceedings of the Section on Survey Research Methods*, 641-646.

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

Hamid, A. (1992), "Ethnographic Follow-up of a Predominantly African American Population in a Sample Area in Central Harlem, New York City: Behavioral Causes of the Undercount of the 1990 Census ", *Ethnographic Evaluation of the 1990 Decennial Census Report Series, Report 11*, Center for Survey Methods Research, Bureau of the Census, Washington D.C.

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

Lerch, P.B. (1992), "Coverage Differences in the Census of a Rural Minority Community in North Carolina: The Little Branch Area of the Waccamaw Sioux Tribe", *Ethnographic Evaluation of the 1990 Decennial Census Report Series, Report 20*, Center for Survey Methods Research, Bureau of the Census, Washington, D.C.

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

Martin, E. and de la Puente, M. (1993), "Research on Sources of Undercoverage Within Households", *American Statistical Association Proceedings of the Section on Survey Research Methods*, 1262-1267.

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.

Wingerd, J. (1992a), "Urban Haitians: Documented/Undocumented in a Mixed Neighborhood", *Ethnographic Evaluation of the 1990 Decennial Census, Report 7*, Center for Survey Methods Research, Bureau of the Census, Washington, D.C.

Wingerd, J. (1992b) "Triple Match Analysis: Mixed Urban Haitian", unpublished report, Center for Survey Methods Research, Bureau of the Census, Washington, D.C.

Zaslavsky, A.M. and Wolfgang, G.S. (1993), "Triple System Modeling of Census, Post-Enumeration Survey, and Administrative-List Data". *Journal of Business and Economic Statistics*, Vol. 11 No 3, 279-288.