Triple System Estimation with Erroneous Enumerations

Paul Biemer, RTI, G. Gordon Brown, RTI, D.H. Judson, Bureau of the Census, and Christopher Wiesen, Odum Institute

Abstract. A central assumption in standard population coverage error estimation approaches is that all non-residential units (so-called erroneous enumerations) have been removed from the data so that the only errors remaining are missed units. This assumption is violated in many situations, notably in the U.S. Census 2000, where undetected erroneous enumerations were a primary reason that the Post-Enumeration Survey (PES) results could not be used in census undercount adjustments. This paper develops a latent class modeling approach that allows for varying levels of undetected erroneous enumerations in one of the population lists. Our approach requires three population lists which may be the Census, the PES, and a list derived from merging records from administrative systems. The resulting data take the form of an incomplete 2^3 contingency table which can be represented by a latent class model where the latent variable is an individual's true status (i.e., resident or non-resident of the population). Latent class analysis is used to estimate the expected values of the observed cells of this table and then to project these estimate onto the unobserved cells in order to estimate the total number of population members. Using artificial populations, we evaluate the improvement in mean squared error using this approach compared with other log-linear estimation approaches from the census undercount and capture-recapture literature. The paper also discusses extensions of this approach for modeling erroneous enumerations in all three lists.

Keywords: Census undercount; conditional multinomial model; finite mixture models; latent class model; capture-recapture.

1. Background

An important issue in estimating the number of persons residing in an area from census data is the evaluation of census coverage error. Various techniques using multiple input sources have been developed for estimating the error in the census count. One widely used method is dual-system estimation (Sekar and Deming, 1949). With this approach, a post-enumeration survey (PES) of the population is conducted and the persons in the PES are matched to persons in the census enumeration. For the 2000 U.S. Census, the PES involved enumeration of the occupants of 300,000 households in a random national sample of 12,500 housing blocks (Hogan, et al.,

2002).

For the dual system estimation (DSE) approach, data from the enumeration process and the PES are combined in a 2×2 table of counts cross-classifying the presence or absence of persons in the census enumeration with their presence or absence in the PES. The DSE approach provides an estimator of the number of persons in the fourth cell of this table which corresponds to persons missed by both the census and the PES. The sum of the three observed and one estimated cells of the Census-PES cross-classification table provides an estimate of the total population count.

Three key assumptions are made for the DSE approach:

1. *Independence*: The probability of inclusion of an individual on the second list (the PES) does not depend upon inclusion or exclusion from the first list (the census). Failure of this assumption will induce correlations between the errors in the two lists, sometimes referred to as behavioral correlation (Wolter, 1986). If a third list is available, the independence assumption can be tested (see, for example, Bishop, Fienberg, and Holland, 1975, Chapter 6). Zaslavsky and Wofgang (1993) provide models for dealing with the behavioral correlation in three systems.
2. *Homogeneity*: The probability of inclusion on a list does not vary from individual to individual. Although this assumption is known not to hold for the population as a whole, various strategies have been used to address the problem of heterogeneous enumeration probabilities, including post-stratification (Sekar and Deming, 1949) and logistic regression (Alho, Mulry, Wurdeman, and Kim, 1993). Methods involving three systems have been explored by, Darroch, Fienberg, Glonek, and Junker (1993), Fienberg, Johnson, and Junker (1999), and Chao and Tsay (1998). This is the correlation bias problem (see, e.g., Wolter, 1986).
3. *Perfect enumeration and matching*. Individuals in both lists are all population members that can be accurately matched between the two lists and any nonresidents who have been

erroneously enumerated can be identified and eliminated. Matching errors can be fairly substantial (Biemer and Davis, 1991a) and methods for dealing with these can be found in Biemer (1988) and Ding and Fienberg (1992). Biemer and Davis (1991b) show how undetected erroneous enumerations can seriously bias the estimates of census coverage error. Usually, the bias is positive, resulting in overcorrecting the census counts for the undercount.

The models considered in this paper seek to address failures of all three assumptions to some extent, but particularly assumptions 1 and 3. The assumption of independence may be relaxed if a third counting system is introduced; for example, an administrative records list (ARL) of persons in the population. Although erroneous enumerations can occur in all three systems, the problem is much greater for administrative lists as will be discussed subsequently. Therefore, the focus in present paper is on erroneous enumerations in the ARL. A subsequent paper (in progress) will extend the ideas of the present paper to erroneous enumerations in all three lists.

Erroneous enumerations (EEs) occur when individuals who are not residents of the target population are erroneously counted as residents. Erroneous enumerations may be persons who were deceased prior to Census Day, born after Census Day, or nonresidents of the target area on Census Day. EEs also include geocoding (or location) errors, duplicated persons, and fictitious or nonexistent persons. In the following, we will refer to all of these entities as nonresidents regardless of their source. Further, any nonresident who is classified as a resident will be called an EE.

In this paper, a statistical framework for dealing with undetected EEs using a latent class modeling (LCM) approach is presented. Latent class models are essentially log-linear models where one or more of the variables are latent or unobservable. Since traditional capture-recapture models can also be written as log-linear models, LCMs are straightforward extensions of the traditional capture-recapture models. LCMs provide a convenient statistical framework

for specifying capture-recapture models with undetected EEs as well as missed residents in all three systems. Unfortunately, the identifiability of LCM for population size estimation has never been explored in the literature. Further, little is known about the statistical properties of the LCM estimators in census coverage error evaluation applications.

To simplify the exposition of the general ideas and the theory, the paper is confined to the situation where undetected EEs are present only in the ARL. That is, we assume the census process is successful in identifying and removing EEs in the census and the PES. This somewhat restricted class of models represents an important generalization over the traditional assumption of no EEs and provides a useful alternative to other dual and triple system models for applications where the numbers of EEs in the census and PES are small compared to the number in the ARL. Further, study of this restricted case will provide important insights regarding the much more complex case of EEs in all three systems, the ultimate goal of this research.

The problem of EEs in the census process has long been recognized and adjustments of the DSE of N for EEs is an essential component of the estimation process. A special survey, referred to as the E-sample (see, for example, Hogan, 1993), is conducted simultaneously with the PES in order to estimate the number of EEs in the census and adjust the DSE for them. Despite these efforts, some EEs are not identified and are included in the dual system thereby inducing bias in the estimates of N.

In 2000, the U.S. Census Bureau Evaluation Followup (EFU) estimated that about 1.8 million enumerations in the PES were actually EEs (ESCAP, 2001). Further, 365,000 persons classified as EEs were in fact correct enumerations. Based on these results, the Census Bureau concluded that the net undercount was overstated by three to four million persons and, thus, adjustments to the census count on the basis of the PES would substantially overcorrect the population counts in many areas. For the 1990 Census, Biemer and Davis (1991b) reported that the level of misclassified EEs in the 1990 PES exceeded 5 percent of the PES count for many areas of the country. In the worst areas, the Northeast urban and the Midwest non-central city

areas, the EE rate exceeded 20 percent.

Although the availability of an ARL as the third list provides the means for modeling the correlation between the census and the PES, the risk of including EEs in the estimation process is is substantially increased since an ARL may contain many non-population members and duplicate persons that are difficult to accurately identify and remove from the process. An example of an ARL that is being considered for census undercount evaluation purposes is the Census Bureau's Statistical Administrative Records System (StARS; see Judson, 2000). The StARS consists of seven merged databases including IRS returns, selective service files, Medicare enrollment database, Indian Health Service patient file, and the HUD tenant resident certification system. Since individuals may be on two or more of these lists, the potential for duplicate persons on StARS is quite high. The address information on the files may be incomplete or erroneous, thus increasing the opportunity for geocoding errors. The files may not be completely current, which can cause the application of the Census residency rules to the StARS to be problematic. Although many EEs can be identified through intensive field followup, such evaluations are costly and quite time consuming, considering the schedule for producing the census counts. In addition, for a large-scale implementation, the error rate of such a field verification process is likely to be unacceptable for census adjustment purposes.

In the next section, we introduce the notation and describe the models that will be used in our study. This includes both models with and without erroneous enumerations. Section 3 describes the estimators of the total population size that can be obtained from the models and provides an illustration of the ideas using real data. Section 4 reports on an extensive simulation study using artificial populations which compared the estimators under various adverse population conditions. Finally, in Section 5, we summarize the results and discuss their implications for using the estimators in census coverage error evaluation studies.

2. Models

This section briefly describes a few of the basic models in the capture-recapture literature, elaborating on two models that will be used extensively in our work. In addition, a new class of capture-recapture models based upon latent class analysis is proposed and the identifiability and utility of these models are explored.

Let U denote the persons in a target area (i.e., the area to be enumerated by the census) who the union of persons included on at least one of the three lists as well as all residents in the area who are not included on any of the lists. Thus, U denotes all actual residents of the target area as well as nonresidents and fictitious persons that are erroneously included on the lists. Let $\wp \subset U$ denote all persons who are true residents of the area and should be counted. Let $\mathscr{E}$ denote the complement of $\wp$, (i.e. $\mathscr{E}=U\sim\wp$); i.e., persons who are not residents of the target area and should not be counted. The number of persons in U will be denoted by M and the number of persons in $\wp$ by N. The objective of this research work is to obtain a robust estimate of N based upon data from the census, PES, and ARL when the members of $\mathscr{E}$ are classified as in $\wp$ and the members of $\wp$ are either missed or classified as in $\mathscr{E}$. As previously mentioned, in this paper we assume that EEs enter the estimation process solely as the result of a misclassification by the ARL.

Let $\boldsymbol{X}_i$ denote a dichotomous variable defined for the i^{th} person in U, where $\boldsymbol{X}_i = 1$ if person $i \in \wp$ and $\boldsymbol{X}_i = 0$ if person $i \in \mathscr{E}$. We assume that $\boldsymbol{X}_i$ is an unknown and unobservable (latent) variable for all $i \in U$. For triple system estimation there are three indicators of X_i corresponding to the census denoted by A_i, the PES denoted by B_i, and the ARL denoted by C_i. Like $\boldsymbol{X}_i$, each indicator variable takes on the value 1 if person i is classified as in $\wp$ and 0 if classified as in $\mathscr{E}$. Note that the definitions of $\boldsymbol{X}_i$ and its indicators depend upon the definition of the target area.

For notational convenience, in the following we will drop the subscript i when it is clear we are referring to an individual in the universe.

2.1 Model Assumptions and Notation

Let π_x denote P(X=1), $\pi_{A=a|X=x} = \pi_{a|x}$ =P(A =a |X =x) with analogous definitions for $\pi_{b|x}$ and $\pi_{c|x}$ where x, a, b, and c can be either 1 or 0. The probability the census correctly enumerates a resident in U is $\pi_{A=1|X=1}$, referred to as the correct enumeration probability. An EE occurs when a person in $\mathscr{E}$ is classified as in $\mathscr{P}$. Thus, the probability of an EE in the census is $\pi_{A=1|X=0}$.

Let XABC denote the (unobservable) cross-classification table for the variables X, A, B, and C for all $i \in U$ and let (x, a, b, c) denote the cell associated with $X = x$, $A = a$, $B = b$, and $C = c$ in this table. Define π_{xabc} = P(x,a,b,c) as the expected proportion in cell (x,a,b,c) and note that π_{xabc} can be expressed as

$$\pi_{xabc} = \pi_x \pi_{a|x} \pi_{b|ax} \pi_{c|abx}. \tag{1}$$

Although the XABC table is not observable, (1) is still useful to specify the cell probability for the observable ABC table, i.e.,

$$\pi_{abc} = \sum_x \pi_x \pi_{a|x} \pi_{b|ax} \pi_{c|abx}. \tag{2}$$

When all parameters in this likelihood are identifiable, they can be estimated using maximum likelihood estimation techniques. Since the unrestricted model (2) contains 95 parameters and only 47 degrees of freedom are available in the ABC table, the model is substantially over-parameterized and not identifiable. Restrictions on the probabilities will be introduced to reduce the number of parameters associated with the model and obtain an identifiable model. The plausibility of these restrictions and other model assumptions for census coverage error

evaluation applications is a key issue in our research.

In the next section, we consider models that assume no EEs in the census estimation process, i.e., $\pi_{A=1|X=0} = \pi_{B=1|X=0} = \pi_{C=1|X=0} = 0$. These are traditional capture-recapture models that are appropriate when the probability of undetected EEs in the three systems is negligibly small. In that case, we can ignore the latent variable *X* in the analysis and consider models for π_{abc} rather than π_{xabc}.

2.2 Models with No Erroneous Enumerations

In this section, we present a few classic closed population models as defined in Pollock, Nichols, Brownie, and Hines (1990) and discuss their utility for coverage error estimation. In order to remain consistent with census terminology, we will use the term "enumeration probability" instead of the traditional term "capture probability" used in the capture-recapture literature. In addition, the models are written using the notation introduced in Section 2.1.

1. Model M_0 - Equal Catchability Model.

Model M_0 assumes that every individual in the population has the same probability of being enumerated on each sampling occasion, i.e., $\pi_A = \pi_B = \pi_C = \pi_1$, and enumerations at future time points are independent of previous enumerations. For this model, $\pi_{abc} = \pi_1^{a+b+c}(1 - \pi_1)^{3-a-b-c}$. Although model M_0 is very unlikely to hold in practice, it is still important as the basis for all closed population models. Instances where M_0 has been used in practice are quite rare; however, it should be noted that, when the enumeration probabilities are in fact equal, inference obtained from model M_0 is nearly identical to inference obtained from the next model we will consider: M_t or Schnabel's model.

2. Model M_t - Schnabel's Model.

Schnabel (1938) originally developed the M_t model for situations where it may be assumed that every individual in the population has the same enumeration probability for a particular list, but enumeration probabilities vary across the lists. As with the M_0 model, future enumerations are assumed to be independent of previous enumerations. This model does not allow heterogeneity of enumeration rates within a list or a behavioral response to trapping. For this model, $\pi_{abc} = \pi_a \pi_b \pi_c$. The next model we consider allows enumeration probabilities for subsequent enumerations to depend upon previous enumerations.

3. Model M_b - The Trap Response Model.

For the M_b model, previously enumerated individuals can have future enumeration probabilities that differ from previously unenumerated individuals. Consequently, enumeration outcomes across the three lists may be correlated. The model specifies that P(*A*=1) = P(*B*=1|*A*=0) = P(*C*=1|*A*=0,*B*=0) = π_U. Note that π_U is the probability of enumeration for any individual not previously enumerated. The corresponding probability for individuals previously enumerated by the census, the PES or both is P(*B*=1|*A*=1) = P(*C*=1|*A*=1 or *B*=1) = π_M. Thus, the cell probabilities for this model can be written as products of π_U, (1-π_U), π_M, and (1-π_M). As an example, P(*A*=1, *B*=1, *C*=1) = $\pi_U \pi_M^2$; P(*A*=0, *B*=1, *C*=0) = $(1 - \pi_U)\pi_U(1 - \pi_M)$; and so on.

In a census context, correlations may be introduced between the census and the PES due to the reactions of individuals to the census enumeration process. For example, individuals who were enumerated in the census may have enjoyed the experience or may determine that any fears they may have had about the process were unfounded. This reaction might cause their probabilities of enumeration in the PES to be higher than for individuals missed by the census - referred to as "trap happy" behavior. Conversely, individuals whose experience with the census enumeration process was less than favorable, might engage in avoidance or "trap shy" behavior in

the PES.

In general, trap shy behavior causes enumeration rates for previously enumerated individuals to decrease, leading to overestimation of the population size. Trap happy behavior causes enumeration rates of previously enumerated individuals to increase, leading to underestimation of the population size.

4. Model M_{tAB} - Non-Stationary Behavioral Response Models.

A natural extension of the M_b is the M_{tb} model which has both time variation and behavioral response to the enumeration process. Although the standard form of the M_{tb} model is not identifiable, a very useful and identifiable model, the M_{tAB} model, can be obtained by imposing a plausible restriction on the M_{tb} model. The index AB on the M_{tAB} model is used to indicate that the model contains one interaction term representing behavioral correlation between the *A*- and *B*-lists and that the *C*-list is assumed to be independent of the other two lists. Under this model $\pi_{abc} = \pi_a \pi_{b|a} \pi_c$.

The motivation for this model stems from the realization that behavioral correlation is likely to be much greater between the census and the PES than for either of these enumerations and the ARL. This is because enumeration by the census or PES depend largely on an individual's attitude towards being interviewed and census participation in general, whereas, the listing of an individual on an administrative record usually depends upon factors that provide more direct benefits to the individual; examples are Medicare benefits, unemployment compensation, automobile ownership, payment of taxes, and so on. Therefore, whether an individual appears on the ARL should not be greatly influenced by the individual's choice or ability to participate in the census or PES.

It seems reasonable to assume that being listed on the ARL is uncorrelated with enumeration by the either census or the PES and, consequently, the interaction terms AC and BC

are negligible. This assumption also greatly reduces the complexity of the models. When both the AC and BC interactions are included in the model, identifiability problems result that can only be remedied by adding parameter constraints which are implausible for census applications. It is possible, however, to fit models that allow for correlations between enumerations in all three lists (see, for example, Zaslavsky and Wolfgang, 1993) and one such model will be considered later in Section 4. In this paper, the M_t and M_{tAB} models are examined in some detail since they appear to be the most likely of the traditional closed population models to mirror triple systems data.

2.4 Models with Erroneous Enumerations in the ARL

For the models presented in this section, several assumptions made for the traditional closed population models are relaxed. We still assume that $\pi_{A=1|X=0}=\pi_{B=1|X=0}=0$, but now we allow $\pi_{C=1|X=0}>0$, i.e., EEs are allowed to enter into the census estimation process through the ARL, or the *C*-list. Thus, we define a new class of population size estimation models which we refer to as L-models. The L-model assumptions essentially parallel those made for the M-models discussed previously except now we introduce a latent "true enumeration status" variable to account for the possibility that nonresidents may be misclassified as residents by their inclusion on the *C*-list.

The assumption of no undetected EEs in the census and PES is consistent with traditional assumptions made for these two systems, but, as previously discussed, these assumptions are unlikely to hold in some enumeration areas. In this regard, the L-models we consider in this paper suffer from the same limitations as the M-models with respect to EEs in the census and PES. The L-models should be preferred when the majority of EEs in the estimation process are introduced through the *C*-list, as is likely when the *C*-list is the ARL. It is possible to extend the general

ideas described here for modeling EEs in the *C*-list to the case where non-negligible EEs occur in the *A*- and *B*-lists. This research is currently underway and will be reported in a subsequent paper.

Another issue for the L-models is the value of P(C=1|X=0), which is the probability that a non-resident in the population is an EE in the ARL. Since all models considered in this paper assume that EEs only enter into the estimation process through the ARL, the only EEs that are of any consequence in the analysis are those that are brought in through the ARL. This implies that, given that we observe an EE in the data, the probability that it was introduced through the ARL is 1. Thus, we know that P(C=1|X=0)=1 and P(A=1|X=0) = P(B=1|X=0) = 0 for all L-models considered.

1. Model L_0 - Equal Catchability LCM.

The L_0 model extends the M_0 model to include EEs in the *C*-list. Like the M_0 model, the L_0 assumes that every individual in the target population has the same probability of enumeration on all three lists. An additional parameter is included to account for the EEs in the *C*-list. Let π_x denote P(X=1), $\pi_{A=1|X=1} = \pi_{B=1|X=1}=\pi_{C=1|X=1} =\pi_1$, and $\pi_{C=1|X=0}=\pi_2$. Then

$$\pi_{xabc} = \pi_x \pi_1^{a+b+c}(1-\pi_1)^{3-a-b-c} + (1-\pi_x)\pi_2^{c}(1-\pi_2)^{(1-c)}(1-a)(1-b). \qquad (3)$$

Note that the second term on the right simplifies to $(1-\pi_x)(1-a)(1-b)$ noting that $\pi_2=1$. This model parallels the M_0 model and is the least complex of the L-models. Like the M_0 model, it unlikely to hold in practice, it will not be considered further in this paper.

2. Model L_t - Non-Stationary, Independent LCM.

The L_t model extends model M_t to reflect EEs in the *C*-list. Thus, we have

$$\pi_{xabc} = \pi_x \pi_{a|x} \pi_{b|x} \pi_{c|x} \qquad (4)$$

which corresponds closely to the classical latent class model for the ABC table except for the

structural zero in the 000 cell. If there is no correlation between the *A*- and *B*-lists, then the L_t model should provide a good estimate of *N*. If there is correlation between the *A*- and *B*-lists then the L_t model is not appropriate, and the quality of inference will decline as the magnitude of the AB interaction increases. The L_t model is the least complex of the L-models that is likely to hold in practice and is investigated in this paper.

3. Model L_{tAB} - Non-Stationary, Behavioral Response Latent Class Model.

Among the models considered in this paper, the L_{tAB} is the most complex LCM and the most likely to accurately represent triple systems data. The L_{tAB} model accounts for correlation between the *A*- and *B*-lists and for list-dependent enumeration probabilities. Under this model, $\pi_{xabc} = \pi_x \pi_{a|x} \pi_{b|xa} \pi_{c|x}$. Unfortunately, the L_{tAB} model has several parameters that are not estimable when using the full or unconditional version of the likelihood. Additionally, the conditional version of the likelihood, which conditions on the seven observable cells and is used in latent class analysis, is not identifiable. Lack of identifiability implies that additional information must be provided in order to obtain meaningful inference from the L_{tAB} model.

Our solution to the identifiability problem is to specify a value for $\gamma = P(X=0|C=1)$. Knowledge of γ makes the conditional likelihood fully identifiable and allows all parameters to be estimated in the unconditional likelihood. For the majority of our study, we will assume that γ is known or can be estimated with negligible bias and sampling error. We also present an example of a potential method for estimating γ. A generalization of this model that allows error in the estimate of γ is currently being developed and is beyond the scope of this paper.

3. Estimation

3.1 Estimating *N*

Both the method of moments and maximum likelihood estimation methods have been used for parameter estimation in the literature for capture-recapture models. Method of moments estimates are often easy to calculate but can have undesirable properties such as large variance or large bias. Consequently, maximum likelihood estimates are often preferred. The standard MLE method consists of using the conditional likelihood of the model being considered (see White and Burnham 1999). In this method, the enumeration probabilities are estimated and an estimate of population size is derived using a Horvitz-Thompson estimator.

For the M-models, the estimator of N has the general form:

$$\hat{N}_M = \frac{n}{(1 - \hat{\pi}_{000})} \tag{5}$$

where n is the number of persons enumerated (all assumed to be in $\mathcal{P}$) and $\hat{\pi}_{000}$ is the estimate of the proportion of persons in $\mathcal{P}$ in the 000 cell of the ABC table.

For the L-models, we use two methods for estimating N. One method involves estimating $\hat{\pi}_{000}$ and $\hat{\pi}_{x}$ using the conditional likelihood for the ABC table (see, for example, Section 6.3 in Bishop, Fienberg, and Holland, 1975). This leads to the estimator

$$\hat{N}_L = \frac{m}{(1 - \hat{\pi}_{000})} \hat{\pi}_{x} \tag{6}$$

where m is the number of persons enumerated (including EEs) and $\hat{\pi}_{x}$ is an estimate of $\pi_x = P(X=1)$.

The other method uses the full likelihood by performing a search over likely values of M. For this search method, the initial value of M is set to its minimum value, $M=m$. Then, the likelihood is maximized over the other parameters conditional on $M=m$. The process is then repeated for $M = m+k$, for $k = 1, 2, 3, \ldots$, and so on until the global maximum is found for all of the parameters. The multi-modal nature of the likelihood necessitated the use of this simple

search algorithm. Let M_{opt} and π_{opt} denote the estimate of M and π_x from this process. Then the estimator of N from the search method is

$$\hat{N}_L = M_{opt}/\pi_{opt}. \tag{7}$$

Bishop, et al. (1975) and Cormack (1989) show how the M-models can be fit using traditional log linear analysis. Haberman (1979) provides a similar structure for estimating LCMs using log-linear analysis with latent variables. For example, the L_t model is equivalent to the following hierarchical log-linear model

$$\log m_{xabc} = u + u_x^X + u_a^A + u_b^B + u_c^C + u_{xa}^{XA} + u_{xb}^{XB} + u_{xc}^{XC} \tag{8}$$

where $m_{xabc} = m\pi_{xabc}$ and m is the number of enumerated individuals. This model is represented in shorthand notation by including the highest order terms in braces; viz., {AX, BX, CX}. Likewise, the L_{tAB} model is represented as {AX, BX, CX, AB} with constraints as noted above.

In Section 4, we illustrate an application of two of the more complex models described in Section 2: the M_{tAB} and L_{tAB} models. These models are applied to data from a study conducted by Zaslavsky and Wolfgang (1993). Estimates from the M_{tAB} and L_{tAB} models are compared to the corresponding estimates from a similar model considered in their paper.

3.2 Illustration Using Real Data

In this section, we illustrate the properties of the M_{tAB} and L_{tAB} estimators using the triple system data reported in Zaslavsky and Wolfgang (1993), hereafter referred to as ZW. For

comparison purposes, we also compare these two estimators with a similar estimator proposed by ZW which is based upon method of moments estimation principles.

ZW propose several models for estimating population size using triple system data. Three sources of data were considered in their study from the 1988 Dress Rehearsal: the census, the PES, and the ARL. These sources were labeled E, P, and A, respectively, in their study but are re-labeled as the *A*-, *B*- and *C*-list, respectively, to be consistent with our current notation.

Among the models used by ZW, the one that most closely resembles our M_{tAB} model is their $\alpha_{EP|A}$ model. The primary difference is that $\alpha_{EP|A}$ allows for behavioral correlations among all three lists while the M_{tAB} model allows correlation only between the *A*- and *B*-lists. Specifically, the $\alpha_{EP|A}$ model forces equality between the *AB* odds ratio conditioned on *C*=1 and the marginal *AB* odds ratio as follows

$$\alpha_{EP|a=1} = \frac{n_{001}n_{111}}{n_{011}n_{101}} = \frac{n_{00+}n_{11+}}{n_{01+}n_{10+}} \tag{9}$$

while M_{tAB} forces three-way equality between the conditional given *C*=0, conditional given *C*=1 and marginal odds ratios as follows:

$$\frac{n_{000}n_{110}}{n_{010}n_{100}} = \frac{n_{001}n_{111}}{n_{011}n_{101}} = \frac{n_{00+}n_{11+}}{n_{01+}n_{10+}} \tag{10}$$

where '+' indicates summation over the index.

The ZW model uses the observed value of the *AB* odds ratio, conditioned on *C*=1 (i.e., denoted by $\alpha_{EP|a=1}$) as an estimate of the unconditioned odds ratio. The estimate of n_{000} is the value that results in this equality, so

$$\hat{n}_{000} = \alpha_{EP|a=1}\frac{n_{01+}n_{10+}}{n_{11+}} - n_{001}. \tag{11}$$

Note that in this formulation, the *AB* odds ratio conditioned on *C*=0 is not restricted and *C*-list is assumed to be dependent on the *A*- and *B*-list.

The estimator of n_{000} from the M_{tAB} model is

$$\hat{n}_{000} = n\left(\frac{\hat{\pi}_{000}}{1-\hat{\pi}_{000}}\right) \tag{12}$$

where $\hat{\pi}_{000}$ is the MLE of π_{000} under the M_{tAB} model. An approximate expression for $\hat{n}_{000}$ which can be compared to (11) is derived in the Appendix A.

Variances of the estimators were estimated using traditional capture-recapture variance estimation techniques for population size such as those described in Seber (1982). The methodology typically used depends on the Taylor series expansion of the Horvitz-Thompson estimate of population size. Program MARK is a software package that calculates parameter estimates and their variances for a wide variety of capture-recapture models (see White and Burnham, 1999) and was used to calculate the traditional variance estimates that are presented the Table 3.2 below. Similar procedures were used for estimates derived from the log-linear models which were fit using the latent class analysis software, ℓEM (Vermunt, 1992).

Although the ZW model is theoretically similar to the M_{tAB} model, the two models can yield very different estimates of population size as shown below. Further, estimates of *N* from the L_{tAB} model exhibit even greater differences from the $\alpha_{EP|A}$ model depending upon the size of γ. To illustrate this, we fit ZW's $\alpha_{EP|A}$, the M_{tAB} model, and the L_{tAB} model for two data sets in Table 3.1 reproduced from ZW's Table 1. The first three columns of Table 3.1denote the eight cells of the ABC table with the cell counts displayed in columns 4 and 5 for two groups: home owners aged 20-29 years and home renters aged 30-44 years.

[INSERT TABLE 3.1 ABOUT HERE]

For owners, aged 20-29 years , the *AB* odds ratio estimate conditioned on *C*=1 is about 12.9, as is the marginal *AB* odds ratio. When *C*=0, the *AB* odds ratio is about 6.8. Under the M_{tAB} model, all *AB* odds ratios are about 5.8.

Table 3.2 provides the estimates of n_{000}, EE, and N for three estimators: $\alpha_{EP|A}$, M_{tAB}, and L_{tAB} shown in column 1 for L_{tAB} computed at two values of γ, γ=0.05 and γ=0.10. The 'EE' column gives the number of EEs detected by the model in the 001 cell and the estimated number of residents given in the 000 cell is given in the column labeled n_{000} column. Thus, the estimate of N is $m+n_{000}$-EE given in the column labeled N.

[INSERT TABLE 3.2 ABOUT HERE]

Three different standard errors are given for these point estimates of N which are shown in the last three columns. The first, expressed by 'SE', represents the standard error generated by LEM[1]. The second, expressed by 'Sim', is the standard error derived by the simulation experiments described in Section 4. The third, expressed by 'Mk', is the standard error given by program MARK. For the Owners, 20-29 years data, m=228; for the Renters, 30-44 years data, m=260.

For owners 20-29 years (top half of Table 3.2), the estimates of N for the ZW estimator is quite discrepant from the MLE estimators; however, the large standard error for ZW estimate (s.e. = 64) suggest that the discrepancies are due more to model instability rather than bias. Also note that changing from γ=0.05 and γ=0.10 for the L_{tAB} estimator has a small effect on the estimates of N suggesting that the L_{tAB} estimates of N are fairly robust to error in estimates of γ for these data.

The bottom half of Table 3.2 corresponds to renters, 30-44 years. For these data, the marginal *AB* odds ratio and the *AB* odds ratio conditional on C=1 are both approximately 34.0, whereas the *AB* odds ratio conditional on C=0 is approximately 27.4. Under the M_{tAB} model, the estimates for all odds ratios are 9.9. The large discrepancy in the odds ratio estimates is reflected

[1]LEM is a software package for fitting log-linear models with latent variables written by Jeroen Vermunt, Tilburg University, Tilburg, The Netherlands (see Vermunt, 1997).

in the difference between the estimates for n_{000} from the two models (the difference is 247). Again, this difference is small relative to the standard error of the ZW estimate (s.e. = 432).

As we did for owners, the L_{tAB} model was fit twice using 0.05 and 0.10 as plausible values for γ. The estimate for n_{000} from the L_{tAB} model decreased by 15% and 30%, respectively, as compared to the M_{tAB} model. This decrease is expected, as removing EEs from the data will lower the estimate of the population size. Note, however, that the change in $\hat{N}$ is relatively small.

In this example, the L_{tAB} and M_{tAB} models yielded substantially lower estimates for n_{000} than did ZW's $\alpha_{EP|A}$ model. These smaller estimates of n_{000} appear to be more plausible as they imply census enumeration rates which are more consistent with prior experience for these areas (see, for example, Hogan, 1993). Our studies of artificial populations such as those described in the next section suggest that in populations where either the $\alpha_{EP|A}$ or the M_{tAB} assumptions maintain, estimates of n_{000} based on $\alpha_{EP|A}$ and M_{tAB} are very close. The standard errors of the M_{tAB} estimate are much smaller in these populations, however, suggesting the M_{tAB} estimate is preferable to $\alpha_{EP|A}$ in populations where the $\alpha_{EP|A}$ is appropriate.

4. Assessing Estimation Accuracy Using Artificial Data

One key objective of our research is to investigate the bias and variance of our triple system estimators of N. In particular, we are interested in examining the properties of the M_t, M_{tAB}, L_t, and L_{tAB} models' estimates of N with varying levels of EEs in the estimation process. In addition, we wish to investigate the consequences of misspecifying the estimation model when behavioral interactions between the indicators are present in the data.

Analytical methods for assessing the bias and variance of the estimates from capture-recapture models are quite complex and are often only available for method of moments estimators. Even in that case, the formulas for the mean square error components are often

asymptotic expressions (see Seber 1982). To circumvent these difficulties, current research has focused on numerical methods of parameter estimation. Since numerical estimation methods lack analytical equations for the parameters, estimates of bias are usually obtained by performing simulation experiments.

In one analysis, we generated data deterministically to simulate a situation where the entire population is sampled. Thus, the parameters specified for the population also hold true exactly in the analysis data set. Since the population parameter values are known and pre-specified exactly for the analysis data set, examination of bias and variance components without the effects of sampling variance is possible. The primary goal of this type of analysis is to study model bias when underlying model assumptions have been violated. We refer to this type of simulation as artificial population analysis without sampling.

Section 4.2 summarizes the results of the artificial population analysis without sampling. The four models of interest were compared using the deterministically generated artificial data. The formulas for generating these data are given in the next section. Variance estimates were not calculated for this analysis since there was no meaningful method of testing their validity.

In a second type of analysis, also described in Section 4.2, numerous samples were randomly selected from an artificial population. The models or formulas under study are then applied to each sample in order to estimate the population parameters of interest. Since the true parameter values are known, the bias of the parameter estimators can be accurately determined provided a sufficiently large number of samples of a given size are generated. In addition to the estimation of bias, the sampling distributions and the variance of the estimators can be determined so that the coverage properties of interval estimates can also be assessed. We refer to this type of simulation experiment as artificial population analysis with sampling.

The primary purpose of our simulation experiments is to determine the bias for point estimates and validity of variance estimates derived from the four selected models presented in Section 3. Additionally, once point and variance estimates have been obtained, the mean square

error can be computed to determine which model produces the estimate of N have smallest total error. An extensive simulation experiment for the four models listed above was conducted and results are given in this section.

4.1 Simulation Methodology

Generating the Artificial Data without Sampling. The data consist of the number of individuals in each of the seven observable cells in an ABC table; i.e., all cells except the 000 cell whose count was set to 0. Although the true number of residents in the 000 cell is known for the artificial populations, this information was suppressed in estimation process since it is unobserved in ABC table.

As stated, all values for the ABC table are generated using the deterministic equation for the number of observations in cell (a,b,c) given by:

$$m_{abc} = N \pi_x \pi_{a|x} \pi_{b|a,x} \pi_{c|x} \quad (13)$$

In order to narrow the focus of the study, several of the population parameters were held constant over all of the artificial data sets. We chose a population size of N = 8,000 corresponding to roughly the size of a census tract in a central city area. We set the enumeration probability for the Census at $\pi_{A=1|X=1} = 0.70$ corresponding to a difficult census enumeration area (see, for example, the estimates for renters in Table 4.2 or Hogan, et al., 2002). The probability of enumeration for the PES given enumeration by the Census was set at $\pi_{B=1|A=1|X=1} = 0.90$ which corresponds to a moderately high behavioral dependence. Finally, the probability of being listed on the administrative list was set at $\pi_{C=1|X=1} = 0.50$ which corresponds to a list with fairly poor coverage properties. Some exploration of other values for these parameters has been undertaken within a 25 percentage-point range of these values and, in general, the results are consistent with

those reported here. We make no claim, however, that our results will hold beyond the range investigated.

The remaining two parameters were varied over a fairly wide range of plausible values as determined by previous census experience. The parameter $\pi_{B=1|A=0,X=1}$, which specifies the level of behavioral correlation between *A* and *B*, was varied over the values 0.40 through 0.90 by increments of 0.10. The parameter $\gamma = \pi_{X=0|C=1}$, which determines the number of EEs in the *C*-list, was varied over the values 0.0, 0.02, 0.05, 0.10, and 0.15. All possible combinations of parameters are considered with each possible combination yielding one artificial data set.

Generating the Artificial Data with Sampling. The data for the artificial data analysis with sampling were derived using the same set of parameters as described for the case without sampling. For each parameter combination, 1000 artificial data sets were generated. Each data set was randomly generated by the following five-step algorithm: (1) calculate the probabilities associated with the eight cells (π_i, i = 1,...,8, say) using the true parameter values of the artificial population; (2) compute the quantities $s_0 = 0$, $s_k = \sum_1^k \pi_i$ for k = 1,...,8, (3) for residents, draw a Uniform(0,1) random number, r; and increment the count in cell k by 1 if $s_{k-1} \le r < s_k$, k = 1, ..., 8; (4) repeat step 3 for 8000 residents; (5) add EEs to the 001 cell such that exactly γ percent of the enumerations on the *C*-list were erroneous.

Fitting the Models. For each artificial data set, all four models were fit in order to obtain an estimate of *N*. The M_t, M_{tAB}, and L_t models can be fit using only the data from the ABC tables. As stated in Section 3, the L_{tAB} model requires an estimate for γ in order to obtain meaningful inference about *N*. This shifts the focus of inference for the L_{tAB} model from bias due to violation of model assumptions to bias in the estimate of *N* due to misspecification of γ.

The parameter estimates were obtained using the unconditional likelihood of the models of interest. The results were compared to estimates obtained using LEM which uses a conditional

likelihood estimation approach as described in Bishop et al. (1975). The two estimation methods provided similar inference.

The results from the analysis of the artificial data sets are given in two parts. The first part described in Section 4.2, contains the results for the models that do not require knowledge of γ to produce meaningful inference; viz., M_t, $M_{tAB,}$ and L_t. Section 4.3 is devoted solely to the study of robustness of the L_{tAB} model estimates of N to failures of the model assumptions and misspecification of γ.

4.2 Results for the M_t, M_{tAB} and L_t Models

In the following tables, γ is the proportion of EEs in the *C*-list for the artificial population, δ is the error in the value of γ specified in the model, and $\rho_{AB|X=1}$ is the degree of *AB* interaction in the artificial population, which corresponds to the correlation between the *A*- and *B*-list given $X = 1$. In Tables 4.1, 4.2, 4.3, and 4.4, the models being compared are listed across the top row of the tables. The model estimate of N, the standard error of $\hat{N}$ (SE column), the mean square error (MSE column) and the percent bias of that estimate (%Bias column) are shown for each model. All variances, biases, and MSEs were estimated directly from the simulation results. The tables only report the results from the simulations with sampling since the bias results for the simulations without sampling were essentially the same. As stated previously, for all cases the resident population size being estimated is 8000.

[INSERT TABLE 4.1 ABOUT HERE.]

Table 4.1 explores the level of bias in the M_t, M_{tAB}, and L_t models when there are EEs, but no *AB* interaction. As expected, the L_t model is capable of producing an estimate of N that is virtually unbiased when EEs are present in the *C*-list. Both the M_t and M_{tAB} model yield biased estimates of N; however, the bias of the M_{tAB} estimate is greater than the bias of the M_t estimate. The point estimates for N from the with sampling and without sampling data are similar. The

MSEs clearly show that the L_t model performs better than the M_t or M_{tAB} when EEs are present in the data.

Table 4.2 shows the MSE components for the M_t, M_{tAB}, and L_t models when there are no EEs in any list, but there is an *AB* interaction. The values of $\rho_{AB|X=1}$ correspond to the changing levels of $\pi_{B=1|A=0,X=1}$. For example, when $\pi_{B=1|A=0,X=1}$=0.80, then $\rho_{AB|X=1}$=0.14. The M_{tAB} model accurately estimates *N* in the presence of an *AB* interaction. The other two models show significant bias due to the *AB* interaction. The L_t model shows considerably more bias and has a larger MSE than the M_t model. It should be noted that the M_t model tends to have the smallest standard error of the three models and the smallest MSE when there is no interaction or EEs present in the data and it behaves poorly when either of these assumptions are violated.

[INSERT TABLE 4.2 ABOUT HERE.]

Tables 4.3 reports the MSE components for the M_t, M_{tAB}, and L_t models when there are EEs and an *AB* interaction. The *AB* interaction is set at the highest level explored in this study, $\rho_{AB|X=1}$=0.53, which corresponds to $\pi_{B=1|X=1,A=0}$=0.40. All three of the models are substantially biased and produce a large MSE when an *AB* interaction and EEs are present in the data. It appears the L_t exhibits the largest MSE of the three models, suggesting that, despite the fact the L_t model can account for EEs, this advantage is negated in the presence of behavioral correlation. These tables highlight the need for a model (e.g., the L_{tAB} model) that is capable of fitting this type of data.

[INSERT TABLE 4.3 ABOUT HERE.]

Results for the L_{tAB} Model. Tables 4.4. and 4.5 show the key results for the L_{tAB} model. As mentioned previously, identifiability of L_{tAB} can be achieved if information on the number of EEs in the *C*-list is entered into the model. Therefore, we fit the L_{tAB} model using a known value for γ and consider situations where γ is not known exactly. For example, γ may be estimated from a study where a random sample of the persons on the *C*-list is selected and sent to the field in

order to verify their residential statuses. In that case, our estimate of γ would be subject to non-sampling and sampling errors and would not be known exactly (see the next section). In Tables 4.4 and 4.5 we consider the effect on the model estimate of N when γ is subject to error equal to δ.

[INSERT TABLE 4.4 ABOUT HERE.]

In tables 4.4 and 4.5, the value of γ is listed in the first row of the table and the amount of error in γ, denoted by δ, is given in the first column of the table. For example, if γ=0.10 and $\delta = -0.20$, then the value of γ used to fit the model is γ=0.08. For a given error percentage, the estimate for N along with the percent bias is given in the two columns below the error percentage. The tables consider both positive and negative values of δ.

Table 4.4 explores the level of bias in the L_{tAB} model when there are EEs but no *AB* interaction for different values of γ. Table 4.5 explores the level of bias in the L_{tAB} model when there are EEs and an *AB* interaction for different values of γ. For this table, $\rho_{AB|X=1}$=0.53 in all cases.

Both tables illustrate that the L_{tAB} model produces a virtually unbiased estimate of N when γ is correctly specified. In addition, the estimate of N appears to be robust to mis-specification of γ. For example, even with as much as 50 percent error, the estimates of N are still within 10 percent of the true value.

There are differences in the value of $\hat{N}$ between Tables 4.4 and 4.5. These differences occur primarily when γ is large (10%, 15%) and the amount of error in γ is positive and large (i.e., δ in the range of 0.30 to 0.50). This is likely due to the fact that $\pi_{B=1|A=0,X=1}$ is equal to 0.40 for Table 4.4 and 0.90 for Table 4.5. When $\pi_{B=1|A=0,X=1}$=0.40, fewer individuals tend to be included in the observable cells of the ABC table. Thus, the estimate of N can take on lower values. This is true since the lower bound of the estimate of N is equal to the number of individuals enumerated minus the number of recognized EEs in the data.

[INSERT TABLE 4.5 ABOUT HERE.]

5. Summary and Discussion

All four models we considered (M_t, M_{tAB}, L_t, and L_{tAB}) produce virtually unbiased estimates of *N* when a given model's assumptions are valid. For example, when *N*=8000, $\rho_{AB|X=1} = 0$ and $\gamma = 0$, the M_t model produces an estimate for *N* of 7999. As the assumptions are violated, all models begin to show biased results. The results of each model will be summarized separately.

The M_t model is the least complex of the four models that we studied in detail. This model's inability to account for either EEs or a behavioral correlation was evident from tables 4.2 and 4.3. Erroneous enumerations induce a positive bias in the estimate of *N*, while an *AB* interaction induces a negative bias. When both an *AB* interaction and EEs are present in the data, the biases due to these conditions tend to offset each other. The result is that the M_t model shows less bias in the estimate of *N* as compared to the M_{tAB} and L_t models when the effect of EEs in the data is approximately equivalent to the behavior correlation effect. Of course, this is in no way a desirable property of the model since balancing these two errors is not under the control of the experimenter. When correlation bias and EEs are not off-setting, the bias in the M_t estimator can be substantial.

The L_t model is designed to estimate *N* when there are EEs present in the *C*-list. As seen from Table 4.1, the L_t model produces estimates of *N* that are virtually unbiased when EEs are present in the data and there is no correlation bias. As demonstrated by Table 4.2, an *AB* interaction induces a severe negative bias in the estimate of *N*. This is similar to the negative bias associated with the correlation induced by population heterogeneity discussed in other work (see, e.g., Alho, et al, 1993). In addition, the *AB* interaction induces a much larger bias and MSE for the L_t model than for the M_t model. For the values of $\rho_{AB|X=1} > 0$ presented in Table 4.2, the MSE for the L_t model is approximately six times larger than that for the M_t model. As illustrated in Table 4.3, when both EEs and an *AB* interaction are present in the data, the M_t model will likely have a lower MSE than the L_t model. The exceptions occur when the *AB* interaction is small and γ is large. In general, it appears as if the L_t model is not very robust to the presence of an *AB* interaction.

It is interesting to compare the estimates from the M_t and L_t models. As stated above, the L_t model's estimates of *N* tend to have more bias and a larger MSE than the M_t estimates when an *AB* interaction is present. By comparison, the estimate of *N* from the M_t model appears to be relatively robust when the proportion of EEs on the *C*-list is small. Therefore, if information on γ is not available and the choice is between M_t and L_t, we recommend using the M_t model over of the L_t model, particularly if a sizeable *AB* interaction is expected. The L_t model is preferred when there is a large proportion of EEs in the *C*-list and γ is unknown. If γ is known, it might be possible to improve the inference obtained by the L_t model by incorporating an estimate of γ into the likelihood.

The M_{tAB} model is designed to estimate *N* when an *AB* interaction, but no EEs, are present in the data. As shown from Table 4.2, the M_{tAB} model produces virtually unbiased estimates for *N* for a range of values for $\rho_{AB|X=1}$. As compared to the M_t model, the presence of EEs induce a large positive bias in the M_{tAB} model. As seen from Table 4.1, when EEs are present in data, the MSE for the M_{tAB} model is approximately 2.5 times larger than that of the M_t model.

The reason for this increase can be explained by the additional parameter in the M_{tAB} model. The M_{tAB} model has two parameters for enumeration probabilities for the *B*-list, $\pi_{B=1|A=1,X=1}$ and $\pi_{B=1|A=0,X=1}$. For the M_t model, these two probabilities are equal, $\pi_{B=1|X=1}=\pi_{B=1|A=1,X=1}=\pi_{B=1|A=0,X=1,}$ and both the unenumerated and the previously enumerated individuals in the *B*-list are used to estimate $\pi_{B=1|X=1}$. Thus, more information is used to estimate $\pi_{B=1|X=1}$, and hence *N*, in the M_t model which, consequently, improves its robustness to EEs.

One nice property of the M_{tAB} model is that the degree of the *AB* interaction does not effect the bias in *N* due to EEs. This concept can be seen by comparing the results for the M_{tAB} model given in Tables 4.1 and 4.3. Additionally, from Table 4.3, it appear that the M_{tAB} model outperforms the M_t model when there few to moderate amount of EEs in the data, $\gamma < 0.05$. Thus, it appears to be preferable to use the M_{tAB} model over the M_t model when a moderate number of EEs are expected in the data. Unfortunately, as illustrated in Table 4.3, the M_t, L_t and

M_{tAB} models exhibit large biases and MSEs when both EEs and an *AB* interaction are present in the data. The results from Table 4.3 highlight the need for the L_{tAB} model.

Of the four models given notable consideration in this study, the L_{tAB} model is the most likely to accurately represent the triple system data. This model can account for both EEs in the *C*-list and for an *AB* interaction. Unfortunately, given only the ABC table, the L_{tAB} model is unidentifiable and requires the inclusion of additional information to provide meaningful inferences for *N*. Our solution to the lack of identifiability is to provide a value for the proportion of EEs in the *C*-list, γ. By specifying γ, the L_{tAB} model becomes fully identifiable and produces virtually unbiased estimates for *N* as seen in Table 4.4. Moreover, as seen in Table 4.5, the inclusion of an *AB* interaction does not affect the inference that is obtained from the L_{tAB} model when γ is known. Thus, our results indicate that the L_{tAB} model can accurately represent data with an *AB* interaction without affecting the nature of the inference.

Another concern for this model is the robustness of the estimate of *N* from the L_{tAB} model to the misspecification of γ. Both Tables 4.4 and 4.5 explore the levels of bias induced in the estimate of *N* when γ is misspecified. In general, the bias tends to be low, implying that the estimates of *N* are fairly robust. For example, consider the case when γ=0.10 and an $\rho_{AB|X=1}$=0.53. For the different values of δ presented in Table 4.5, the MSE for the L_{tAB} model ranges from 1,651 when δ=0 to 195,767 when δ=0.50. Similarly, the bias ranges from 0.0% when δ=0 to 5.5% when δ=0.50. By comparison, under this scenario, the M_t, L_t, and M_{tAB} models have MSEs of 1,016, 1,986,191 and 1,782,739, respectively. Even when γ is badly misspecified, the L_{tAB} model appears to outperform the M_{tAB} and L_t model.

In order to fully utilize the L_{tAB} model, a reasonable value for γ must be obtained from a separate data source. One possible method for obtaining an estimate of γ is to conduct a field study by drawing a random sample from the observations in cell 001 of the ABC table. In this situation, the MSE formulas in the present paper can be expanded to include variation in the estimate of *N* due to estimating γ. Our preliminary investigations of this method suggest that

even in situations where there is considerable sampling variability in the estimate of γ, the L_{tAB} model MSE of the L_{tAB} model estimate is still considerably smaller than that of the M_{tAB} model when γ is in the range of 0.05 to 0.15.

In general, when undetected EEs appear in the *C*-list and a reasonable estimate of γ is available, the L_{tAB} model performed best for estimating *N*. If an estimate of γ cannot be obtained, then the selection of an appropriate model to use for inference about *N* is less clear. It appears, however, that the M_{tAB} and M_t models outperform the L_t model. The selection of which model to use would depend on the nature of data, specifically on the strength of the *AB* interaction and the number of EEs in the *C*-list.

An alternative to using the L_{tAB} model for dealing with EEs in the ARL is to proceed with the M_{tAB} model and use an estimate of γ in *post hoc* correction of $\hat{N}_{M_{tAB}}$ for EEs. Our empirical studies suggest that such corrections can produce unbiased results in populations that are also ideal for the L_{tAB} model. One such *post hoc* estimator (derived in Appendix B) appears to produce very good results is

$$\hat{N}_{M_{tAB}}(\gamma) = (1-\gamma)\hat{N}_{M_{tAB}} \tag{14}$$

In practice, if an unbiased estimate, $\hat{\gamma}$ of γ is available, using (14) after substituting $\hat{\gamma}$ for γ will generally reduce the bias in $\hat{N}_{M_{tAB}}$; however, the resulting MSE increase depending upon the size of the bias relative to the standard error of $\hat{\gamma}$.

An important advantage of using L-models to explicit account for EEs rather than using post hoc corrections of M-models is the ease with which L-models can be extended to more complex situations. When EE's appear in more than one list, post hoc corrections for EEs are impractical due to their complexity. Latent class analysis provides an integrated structure for modeling much more complicate scenarios than were described in this paper. Thus, L_{tAB} model should be viewed

as a foundation for more complex models that involve list by list interactions, EEs in all three lists, and four or more lists. The current paper lays the groundwork for dealing with these more complex situations.

References

Agresti, A. 1994. Simple capture-recapture models permitting unequal catchability and variable sampling effort. *Biometrics* 50:494-500.

Alho, J., Mulry, M., Wurdeman, K, and Kim, J. 1993, Estimating heterogeneity in the probabilities of enumeration for dual-system estimation, *Journal of the American Statistical Association*, 88, 1130-1136.

Biemer, P. and Davis, M., 1991a. "Estimates of P-sample Clerical Matching Error from a Rematching Evaluation, Evaluation Project P7 Internal Bureau of the Census Report, July 1991.

Biemer, P. and M. Davis. 1991b. "Measurement of Census Erroneous Enumerations - Clerical Error Made in the Assignment of Enumeration Status." Evaluation Project P10, Internal Bureau of the Census Report, July 199

Biemer, P., H. Woltman, D. Raglin, and J, Hill, 2001. "Enumeration accuracy in a population census: an evaluation using latent class analysis." *Journal of Official Statistics*, Vol. 17, No. 1, pp. 129-148.

Bishop, Y. M. M., S. E. Fienberg, and P. W. Holland. 1975. *Discrete Multivariate Analysis: Theory and Practice*. Cambridge: MIT Press.

Brown, G. G. 2001 *Comparing Bayesian, Maximum Likelihood and Classical Estimates for the Jolly-Seber Model*. Ph.D. Dissertation. North Carolina State University, Raleigh, NC.

Chao, A. and Tsay, P.K. 1998. A sample coverage approach to multiple-system estimation with application to census undercount. *Journal of the American Statistical Association*, 93, 283-293.

Cormack, R. M. 1989. Log-linear models for capture-recapture. *Biometrics* 48:201-216.

Darroch, J.N., Fienberg, S.E., Glonek, G.F.V., and Junker, B.W. 1993. A three-sample multiple-recapture appro census population estimation with heterogeneous catchability, *Journal of the American Statistical Association*, 88, 1137-1148.

Ding, Y. and Fienberg, S. 1992. Estimating population and census undercount in the presence of matching error, unpublished manuscript.

ESCAP. 2001. ESCAP II Report No. 1 Recommendation and Report of the Executive Steering Committee for A Policy (ESCAP II), Bureau of the Census, Washington, D.C.

Fienberg, S.E., Johnson, M.S., and Junker, B.W. 1999. Classical multilevel and Bayesian approaches to populati estimation using multiple lists. *Journal of the Royal Statistical Society, A*, 162, 383-405.

Goodman, L. A. 1974. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* 61:215-231.

Goodman, L. A. 1978. *Analyzing Qualitative/Categorical Data; Loglinear models and Latent Structure Analysis*. (Jay Magidson ed.) Abt Books.

Haberman, S. 1979. *Analysis of Qualitative Data, Vol. 2, New Developments*. New York, Academic Press.

Hogan, H. 1993. "The 1990 post-enumeration survey: operations and results," *Journal of the American Statistical Association*, Vol. 88, No. 423, 1047-1071.

Hogan, H., Kostanich, D., Whitford, D., and Singh, R. 2002. "Research Findings of the Accuracy and Coverage Evaluation and Census 2000 Accuracy," American Statistical Association Joint Statistical Meetings, *Proceedings of the Section on Survey Research Methods*.

Judson, D. 2000. "The Statistical Administrative Records System: System Design, Successes, and Challenges." Internal Census Bureau Report, Nov. 11, 2000.

Lehmann, E. L. 1983. *Theory of Point Estimation.* New York: Wiley.

Norris, J. L., and K. H. Pollock. 1996. Nonparametric MLE under two closed capture-recapture models with heterogeneity. *Biometrics* 52:639-649

Pollock, K. H., J. D. Nichols, C. Brownie, and J. E. Hines. 1990. Statistical inference for capture-recapture experiments. *Wildlife Monographs* 107.

Pledger, S. 2000. Unified maximum likelihood estimates for closed capture-recapture models using mixtures. *Biometrics* 56:434-442.

Schnabel, Z. E. 1938. The estimation of the total fish population of a lake. *Am. Math. Mon.* 45:348-352.

Sekar, C.C. and Deming, W.E. 1949. On a method of estimating birth and death rates and extent of registration, *Journal of the American Statistical Association*, 44, 101-115.

Seber, G. A. F. 1982. *The Estimation of Animal Abundance and Related Parameters*, 2nd ed. London: Griffin.

U.S. Census Bureau, Executive Steering Committee for Accuracy and Coverage Evaluation Policy. 2001. "Report of the Executive Steering Committee for Accuracy and Coverage Evaluation Policy on Adjustment for Non-Redistricting Uses." Online: http://www.census.gov/dmd/www/EscapRep2.html, retrieved October 17, 2001.

Vermunt, J.K. 1996. *Log-Linear Event History Analysis*, Tilberg Press, Tilburg, The Netherlands.

Vermunt, J. 1997. *LEM: A General Program for the Analysis of Categorical Data*, Tilburg, University.

White, G. C., and K. P. Burnham. 1999. Program MARK: Survival estimation from populations of marked animals. *Bird Study* 46 Supplement: 120-138.

Wolter, K.M. 1986. Some coverage error models for census data, *Journal of the American Statistical Association*, 81, 157-162.

Zaslavsky, A., and G. Wolfgang. 1993, "Triple-system modeling of census, post-enumeration survey, and administrative-list data." *Journal of Business and Economic Statistics*, 11:279-288.

Appendix A. Derivation of M_{tAB} and L_{tAB} Estimators of n_{000}

The likelihood for the M_{tAB} model, denoted by $L_{M_{tAB}} = L(N, \pi_a, \pi_{b|a=1}, \pi_{b|a=2}, \pi_c \mid n_{ijk})$ can be written as:

$$L_{M_{tAB}} = \frac{N!}{\prod_{ijk} n_{ijk}!\,(N - n_{+++})!} \pi_a^{n_{1++}} (1-\pi_a)^{N-n_{1++}} \pi_{b|a=1}^{n_{11+}} (1-\pi_{b|a=1})^{n_{1++}-n_{11+}} \times \pi_{b|a=2}^{n_{01+}} (1-\pi_{b|a=2})^{N-n_{1++}-n_{01+}} \pi_c^{n_{++1}} (1-\pi_c)^{N-n_{++1}} \tag{A.1}$$

where n_{ijk} denotes the cell count in cell (i,j,k) of the ABC table, "+" indicates summation over the corresponding index, and the other notation is as defined in Section 2. To find the value of the parameters that maximizes (A.1), we take the logarithm of this function and differentiating with respect to the paramaters, set these partial derivatives equal to 0, and solve for the parameters. Holding N constant and maximizing with respect to the other parameters produces the following MLEs for π_a, $\pi_{b|a=2}$ and π_c conditional on N:

$$\hat{\pi}_{a|N} = \frac{n_{1++}}{N}$$
$$\hat{\pi}_{b|a=2,N} = \frac{n_{01+}}{N - n_{1++}} \tag{A.2}$$
$$\hat{\pi}_{c|N} = \frac{n_{++1}}{N}$$

Replacing these parameters in (A.1) by their conditional MLEs, the likelihood can be written as a function of N only. Simplifying and removing factors that do not contain N, (A.1) simplifies to

$$\frac{N!}{(N - n_{+++})!} N^{-2N} (N - n_{1++} - n_{01+})^{N-n_{1++}-n_{01+}} (N - n_{++1})^{N-n_{++1}} \tag{A.3}$$

Several approximations will be used in determining an MLE for *N*. First, *N* will be treated as a continuous variable and second, in order to take a derivative of *N*!, Stirling's approximation to the factorial will be used. This yields the following approximation to (A.2):

$$\frac{N^{N+0.5}e^{-N}}{(N-n_{...})^{N-n_{...}+0.5}e^{-N+n_{...}}}N^{-2N}(N-n_{1..}-n_{01.})^{N-n_{1..}-n_{01.}}(N-n_{..1})^{N-n_{..1}}$$

Again, eliminating factors that do not involve *N* yields

$$(N-n_{...})^{-(N-n_{...}+0.5)}N^{N+0.5}(N-n_{1}-n_{01.})^{N-n_{1}-n_{01.}}(N-n_{..1})^{N-n_{..1}}$$

Taking the natural log of the above expression gives:

$$\begin{aligned}&-(N-n_{...}+0.5)\log(N-n_{...})-(N+0.5)\log(N)\\&+(N-n_{..1}-n_{01.})\log(N-n_{1..}-n_{01.})+(N-n_{..1})\log(N-n_{..1})\end{aligned} \tag{A.3}$$

Now we can take the derivative of (A.3) with respect to *N* and set the resulting expression to 0. To obtain the following expression, we use a third approximation, viz., $\log(1+\alpha)=\alpha$ where α is a small positive constant. Upon simplifying, this yields

$$\log\left[\frac{(N-n_{1..}-n_{01.})(N-n_{..1})}{(N-n_{...}+0.5)(N+0.5)}\right]=0.$$

Finally, solving for *N* and further simplifying yields

$$\hat{N}_{MLE}=\frac{n_{..1}(n_{1..}+n_{01.}+0.5)+0.5n_{...0}-0.25}{n_{1..}+n_{01.}+n_{..1}-n_{...}+1} \tag{A.4}$$

Subtracting n_{+++} from the above expression produces estimate of n_{000} that can be compared with (11).

MLEs for the L_{tAB} model can be derived in a similar fashion. Since for the L_{tAB} model, EEs only appear on the *C* list, specifying γ is equivalent to specifying the number of EEs, say n_{EE}

that occur on the *C* list since $\gamma = \frac{n_{EE}}{n_{*\,*1}}$. Repeating the above steps and approximations for this model yields the following approximate MLE for the L_{tAB} model when γ is known:

$$\hat{N}_{L_{tAB}} = \frac{(n_{*\,*1} - n_{EE})(n_{1\,*\,*} + n_{01*} + 0.5) + 0.5n_{*\,*0} - 0.25}{n_{1\,*\,*} + n_{01*} + n_{*\,*1} - n_{*\,*\,*} + 1}. \qquad \text{(A.5)}$$

Appendix B. Derivation of the Estimator $\hat{N}_{M_{t+1}}(\gamma)$

Using the results of Appendix A, the ratio of $\hat{N}_{M_{t+1}}$ to $\hat{N}_{L_{t+1}}$ can be written as

$$\frac{\hat{N}_{L_{t+1}}}{\hat{N}_{M_{t+1}}} = \frac{(n_{\cdot\cdot 1} - n_{EE})}{n_{\cdot\cdot 1}} \times \frac{(n_{1\cdot\cdot} + n_{01\cdot} + 0.5) + \dfrac{0.5n_{\cdot\cdot 0} - 0.25}{(n_{\cdot\cdot 1} - n_{EE})}}{(n_{1\cdot\cdot} + n_{01\cdot} + 0.5) + \dfrac{0.5n_{\cdot\cdot 0} - 0.25}{n_{\cdot\cdot 1}}} \tag{B.1}$$

The remainder of this proof will show that the second factor on the right hand side of (B.1), denoted by F can be approximated by 1 for values of $\gamma < 0.5$. In that case, $\hat{N}_{L_{t+1}} = (1-\gamma)\hat{N}_{M_{t+1}}$

To show that $F \approx 1$ for small γ, we multiply and divide F by $(n_{1\cdot\cdot} + n_{01\cdot} + 0.5)$. Ignoring the term -0.25 which is negligible compared with $0.5n_{\cdot\cdot 1}$, we obtain

$$F = \frac{1 + \dfrac{0.5n_{\cdot\cdot 0}}{(n_{1\cdot\cdot} + n_{01\cdot} + 0.5)(n_{\cdot\cdot 1} - n_{EE})}}{1 + \dfrac{0.5n_{\cdot\cdot 0}}{(n_{1\cdot\cdot} + n_{01\cdot} + 0.5)n_{\cdot\cdot 1}}} \tag{B.2}$$

Note that $(n_{1\cdot\cdot} + n_{01\cdot} + 0.5) > n_{\cdot\cdot 0}$, which implies that $\dfrac{0.5n_{\cdot\cdot 0}}{(n_1 + n_{01\cdot} + 0.5)} = c_1$, for some constant $c_1 < 0.5$. Thus, $F = \dfrac{1 + \dfrac{c_1}{(n_{\cdot\cdot 1} - n_{EE})}}{1 + \dfrac{c_1}{n_{\cdot\cdot 1}}}$. Replacing $n_{\cdot\cdot 1} - n_{EE}$ with $n_{\cdot\cdot 1}(1-\gamma)$ and simplifying yields

$$F = \frac{\frac{n_{\bullet\bullet 1}(1-\gamma)+c_1}{(1-\gamma)}}{n_{\bullet\bullet 1}+c_1} = \frac{n_{\bullet\bullet 1}}{n_{\bullet\bullet 1}+c_1} + \frac{c_1}{(n_{\bullet\bullet 1}+c_1)(1-\gamma)}$$

When $\gamma=0$ then this expression is exactly 1. When $\gamma=0.5$, F reduces to $\frac{n_{\bullet\bullet 1}+2c_1}{n_{\bullet\bullet 1}+c_1}$. Since $c_1 <$ 0.5, $F \approx 1$ when $n_{\bullet\bullet 1}$ is reasonably large or, in general, for any value of γ between 0 and 0.5.

Table 3.1. Triple System Data from the 1988 Dress Rehearsal Census in Louis, Missouri

A	B	C	Owners, 20-29 years	Renters, 30-44 years
0	**0**	**0**	–	–
0	**0**	**1**	59	43
0	**1**	**0**	8	04
0	**1**	**1**	19	13
1	**0**	**0**	31	30
1	**0**	**1**	19	7
1	**1**	**0**	13	69
1	**1**	**1**	79	72

Table 3.2. Estimates of $\hat{N}$ for $\alpha_{EP|A}$, M_{tAB}, L_{tAB} (γ=0.05), and L_{tAB} (γ=0.10)

Standard Error Estimates

Post-stratum	EE	$\hat{N}$	N	SE	Sim	Mk
Owners, 20-29 years (ZW)	0	130	358	64	17.6	NA
Owners, 20-29 years(M_{tAB})	0	26	254	4.7	7.6	7.5
Owners, 20-29 years (L_{tAB}) γ=0.05	9	22	241	4.3	6.8	6.4
Owners, 20-29 years (L_{tAB}) γ=0.10	18	18	228	3.9	5.8	5.4
Renters, 30-44 years (ZW)	0	305	565	432	38.8	NA
Renters, 30-44 years (M_{tAB})	0	58	318	9.4	13.8	14.1
Renters, 30-44 years (L_{tAB}) γ=0.05	7	49	302	10.0	13.2	11.1
Renters, 30-44 years (L_{tAB}) γ=0.10	14	40	286	8.5	10.9	8.8

Table 4.1. No *AB* Interaction and EEs Given by γ with Sampling

γ	M_t				M_{tAB}				L_t			
%	N	SE	MSE	%B	N	SE	MSE	%B	N	SE	MSE	%B
0	7999	12.28	151	0.0	7999	15.98	257	0.0	7993	15.21	280	-0.1
2	8098	12.41	9773	1.1	8162	16.20	26432	2.0	7999	20.54	425	0.0
5	8257	12.27	66112	3.2	8420	15.99	176588	5.3	8000	18.95	402	0.0
10	8546	14.50	298697	6.8	8887	17.78	786553	11.1	8000	20.20	402	0.0
15	8875	16.31	766748	10.9	9409	19.00	2922785	17.6	8000	19.95	398	0.0

Table 4.2 *AB* Interaction Given by $\rho_{AB|X=1}$ and No EEs with Sampling

ρ	M_t				M_{tAB}				L_t			
	N	SE	MSE	%B	N	SE	MSE	%B	N	SE	MSE	%B
0.0	7999	12.29	151	0.0	7999	15.99	257	0.0	7993	15.22	280	-0.1
0.14	7893	16.22	11678	-1.3	7999	22.04	486	0.0	7731	24.10	72652	-3.4
0.25	7785	20.45	46519	-2.7	8000	29.50	870	0.0	7466	28.58	286122	-6.7
0.35	7674	21.81	106732	-4.1	8000	31.94	1020	0.0	7199	30.38	643181	-10.0
0.44	7561	24.54	192972	-5.5	8000	37.34	1394	0.0	6933	33.45	1139608	-13.3
0.53	7444	27.47	309890	-6.9	7998	43.90	1931	0.0	6665	34.42	1785386	-16.7

Table 4.3 *AB* Interaction with $\rho_{AB|X=1}$=0.53 and EEs Given by γ with Sampling

γ	M_t				M_{tAB}				L_t			
	N	SE	MSE	%B	N	SE	MSE	%B	N	SE	MSE	%B
0	7444	27.74	309890	-7.0	7998	43.90	1931	0.0	6664	34.42	1783596	0.0
2	7542	26.84	209834	-5.7	8160	41.98	27513	2.0	6666	34.43	1776659	0.0
5	7700	27.93	90684	-3.8	8420	43.49	178879	5.3	6666	35.76	1790793	0.0
10	7987	29.41	1016	-0.2	8886	45.50	787811	11.1	6666	35.78	1790783	0.0
15	8312	30.53	97989	3.9	9408	49.60	1986191	17.6	6665	34.40	1782739	0.0

Table 4.4 L_{tAB} Model: no *AB* Interaction and EEs Given by γ with Sampling

γ	0%				2%				5%				10%				15%			
δ	*N*	SE	MSE	Bias	*N*	SE	MSE	Bias	*N*	SE	MSE	Bias	*N*	SE	MSE	Bias	*N*	SE	MSE	Bias
-50	n/a	n/a	n/a	n/a	8081	15.89	6795	1.0	8207	16.16	43092	2.6	8441	16.82	194764	5.5	8703	17.47	494514	8.8
-40	n/a	n/a	n/a	n/a	8063	16.92	4203	0.8	8167	15.89	28132	2.1	8354	16.74	125596	4.4	8561	17.20	315017	7.0
-30	n/a	n/a	n/a	n/a	8045	15.55	2259	0.6	8124	15.24	15619	1.5	8263	16.37	69437	3.3	8421	16.46	177512	5.3
-20	n/a	n/a	n/a	n/a	8030	16.12	1160	0.4	8080	16.56	6643	1.0	8173	16.53	30202	2.2	8278	16.44	77554	3.5
-10	n/a	n/a	n/a	n/a	8013	15.59	412	0.2	8037	17.21	1612	0.5	8085	16.68	7503	1.1	8136	15.67	18742	1.7
0	7999	16.34	267	0.0	7997	15.42	252	0.0	7996	16.06	273	0.0	7996	15.52	257	0.0	7994	15.36	272	0.1
10	7960	15.55	1841	-0.5	7981	15.77	604	-0.2	7955	15.66	2268	-0.6	7902	16.40	9873	-1.2	7807	43.46	39138	2.4
20	7919	15.30	6475	-1.0	7965	15.47	1468	-0.4	7909	15.87	8533	-1.1	7773	24.86	52147	-2.8	7762	14.76	56862	3.0
30	7878	16.44	15118	-1.5	7949	15.81	2844	-0.6	7866	17.27	18254	-1.7	7761	14.63	57335	-3.0	7760	14.94	57823	3.0
40	7837	15.62	26803	-2.0	7931	15.96	5004	-0.9	7810	29.30	36958	-2.4	7761	15.31	57355	-3.0	7762	15.01	56869	3.0
50	7789	19.59	44755	-2.6	7916	15.62	7299	-1.1	7765	16.41	55494	-2.9	7762	15.07	56871	-3.0	7761	15.23	57353	3.0

Table 4.5. L_{tAB} Model: *AB* Interaction with $\rho_{AB|X=1}$=0.53 and EEs Given by γ with Sampling

γ	0%				2%				5%				10%				15%			
δ	*N*	SE	MSE	Bias	*N*	SE	MSE	Bias	*N*	SE	MSE	Bias	*N*	SE	MSE	Bias	*N*	SE	MSE	Bias
-50	n/a	n/a	n/a	n/a	8078	42.32	7875	1.0	8210	42.58	45913	2.6	8440	45.55	195767	5.5	8705	47.92	499321	8.8
-40	n/a	n/a	n/a	n/a	8064	43.24	5966	0.8	8165	43.82	29145	2.1	8356	43.66	128642	4.4	8563	45.71	319058	7.0
-30	n/a	n/a	n/a	n/a	8047	40.11	3818	0.6	8125	39.63	17196	1.5	8265	41.74	71967	3.3	8424	45.11	181811	5.3
-20	n/a	n/a	n/a	n/a	8031	41.94	2720	0.4	8085	42.15	9002	1.0	8173	41.24	31630	2.2	8281	43.00	80810	3.5
-10	n/a	n/a	n/a	n/a	8015	45.37	2283	0.2	8039	42.04	3288	0.5	8089	41.35	9631	1.1	8140	42.70	21423	1.8
0	7999	41.54	1727	0.0	7997	43.21	1876	0.0	7997	43.52	1903	0.0	7996	40.43	1651	0.1	7998	42.97	1850	0.0
10	7958	42.66	3584	-0.5	7980	41.66	2136	-0.2	7956	41.84	3687	-0.6	7912	40.84	9412	-1.1	7857	41.70	22188	-1.8
20	7917	41.20	8586	-1.0	7965	40.33	2852	-0.4	7916	39.60	8624	-1.1	7822	41.14	33376	-2.2	7716	40.47	82294	-3.6
30	7877	40.38	16760	-1.5	7950	38.72	3999	-0.6	7872	39.40	17936	-1.7	7730	39.21	74437	-3.4	7573	42.30	184118	-5.3
40	7841	40.17	26895	-2.0	7933	41.28	6193	-0.9	7827	42.35	31723	-2.4	7639	38.13	131775	-4.5	7430	39.87	326490	-7.1
50	7799	38.46	41880	-2.5	7916	40.31	8681	-1.1	7785	42.73	48051	-2.9	7555	39.66	199598	-5.6	7292	39.05	502789	-8.9