Local Post-Stratification in Dual System Accuracy and Coverage Evaluation for the U.S. Census

Song Xi CHEN, Cheng Yong TANG, and Vincent T. MULE, JR.

We consider a local post-stratification approach to analyze the capture–recapture dual system Accuracy and Coverage Evaluation (A.C.E.) data associated with the 2000 U.S. Census. The local post-stratification is carried out via a nonparametric regression estimation of the census enumeration and the correct enumeration functions. We propose a nonparametric population size estimator that is designed to accommodate some key aspects of the A.C.E.: missing values, erroneous enumerations, and extra covariates affecting the missingness and correct enumeration. The resulting estimates are compared with estimates from a conventional post-stratification and a logistic regression approach in an analysis on the 2000 Census A.C.E. data.

KEY WORDS: Capture–recapture; Erroneous enumeration; Kernel smoothing; Missing values; Population size estimation.

1. INTRODUCTION

Dual system capture–recapture surveys have been conducted in 1980, 1990, and 2000 in conjunction with the last three decennial U.S. Censuses. Their main objective was to obtain information on the population net coverage and enumeration errors for both the whole population and subpopulations defined by race, geographic region, age, and other demographic/socioeconomic variables. The 2000 Census Accuracy and Coverage Evaluation (A.C.E.) consisted of two surveys. The first survey verified the census enumerations on selected sample block clusters of the census. The data collected was the enumeration (E) sample. The aim of the E sample was to identify erroneous enumerations, such as fictitious records and people who were born after or who died before the census date. The second survey was independent of the first one and was conducted soon after the census in the same sample block clusters of the E sample. The data collected was the population (P) sample; the purposes were to identify "matches" (recaptures) to the census records and to facilitate estimation of the E-sample enumeration probability; see Hogan (1992, 1993, 2000a, 2000b), Haberman, Jiang, and Spencer (1998), and Bell (1999) for comprehensive discussions on the dual system surveys. The US is not the only country that conducts the dual system surveys to gain information on the accuracy of the census counts. Australia, New Zealand, Turkey, Switzerland, and the UK also carry out similar surveys to evaluate their national censuses; see Census Customer Service (2002), Dunstan et al. (2001), Ayhan and Ekni (2003), and Rhind (2003).

It is known that different individuals may have different probabilities of being enumerated in the census (Hogan 1993 and 2000b); this is called heterogeneity in enumeration. A group of variables called ROAST is known to differentiate the heterogeneity in the census enumeration (Hogan 1993). Here, RO stands for race/(Hispanic) origin, A for age, S for sex, and T for housing tenure (owner or renter).

Let $\mathbf{X} = (X_1, \ldots, X_d)$ be a vector of covariates that influences the enumeration of individuals. We set Y to be a binary indicator that takes a value 1 for enumeration and 0 otherwise for each E-sample person. The enumeration probability of an individual with covariates $\mathbf{X} = \mathbf{x}$ is $p(\mathbf{x}) = P(Y = 1|\mathbf{X} = \mathbf{x})$. Without the P sample, only individuals with $Y = 1$ would be observed and that is insufficient for the estimation of $p(\mathbf{x})$. The P sample makes estimation feasible by providing enumerations with both Y outcomes. As $Y = 1$ indicates a match (recapture) to a census record, $p(\mathbf{x})$ is also called the match rate function.

Figure 1 displays the kernel estimates of the match rates as a function of age for selected values of region and the other ROAST variables. The estimates are based on the 2000 A.C.E. data using nonparametric imputations for missing values based on the kernel estimator approach shown in Section 5. The estimated $p(\mathbf{x})$ indicates strong heterogeneity with respect to the age, region, and ROAST variables. Each panel in Figure 1 displays a V-shape within the age range of 18 to 29. This age interval is known for having fluctuating enumeration properties. However, the specific details of this V-shape vary substantially for different regions and the discrete ROAST combinations.

Post-stratification (Sekar and Deming 1949) has been the method used in the U.S. Census to counter heterogeneous enumerations by subdividing the covariate space. Although it reduces the heterogeneity, there is still a substantial amount that is unaccounted for, as shown in Figure 1. One limitation of the post-stratification is that continuously valued covariates such as age are grouped into discrete categories. This implies that $p(\mathbf{x})$ is piecewise constant with respect to the age strata. Any remaining heterogeneity may result in "correlation bias" (Wolter 1986; Chen and Lloyd 2000) in the population size estimates. The other limitation of the post-stratification approach is that some strata have small sample sizes. To control the variance, small

Song Xi Chen is Professor of Statistics, Department of Statistics, Iowa State University, Ames, IA 50010 and Department of Business Statistics and Econometrics, Guanghua School of Management, Peking University, China (E-mail: *songchen@iastate.edu*). Cheng Yong Tang is Assistant Professor, Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546 (E-mail: *statc@nus.edu.sg*). Vincent T. Mule, Jr. is from the U.S. Census Bureau, Washington, DC 20233 (E-mail: *vincent.t.mule.jr@census.gov*). We thank the U.S. Census Bureau for supporting our research. We are grateful to Bill Bell for insightful comments, suggestions, and logistic arrangements, Doug Olson for providing versions of the A.C.E. data, and Philip Gbur for overseeing the administrative matters. Thanks also go to Richard Griffin, Donna Kostanich, Don Malec, Mary Mulry, and Tommy Wright of the Census Bureau for insightful comments and discussions. The project is supported by a National Science Foundation grant SES-0518904. Tang is also supported by a National University of Singapore startup grant. The views expressed in this paper are not necessarily those of the U.S. Census Bureau. We thank the Professor David Banks for careful reading of the paper, the AE, and two referees for constructive comments that led to improving the presentation of the paper.

Journal of the American Statistical Association
March 2010, Vol. 105, No. 489, Applications and Case Studies
DOI: 10.1198/jasa.2009.ap08404

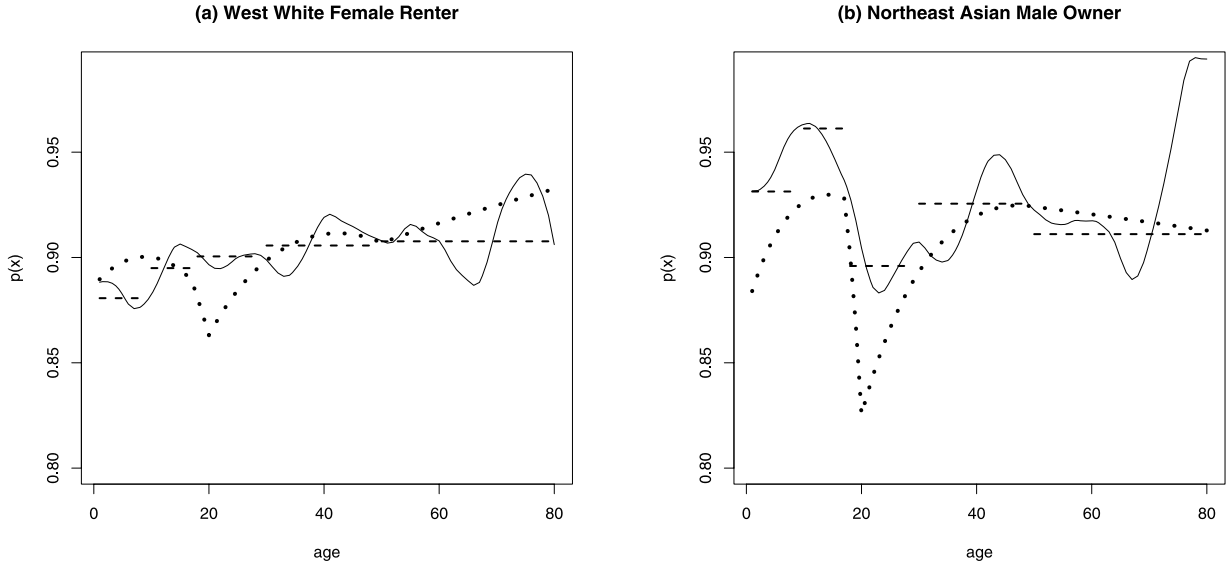

Figure 1. Estimates of the enumeration function $p(\mathbf{x})$ by the local post-stratification (solid line), post-stratification (dashed line), and logistic regression (dotted line).

strata are usually combined but doing so increases the heterogeneity.

There have been numerous studies that use additional data sources in conjunction with the census dual system surveys. Wolter (1990) and Bell (1993) considered correcting the correlation bias within each post-stratum by utilizing demographic analysis (DA) information. Elliott and Little (2000, 2005) proposed Bayesian hierarchical modeling approaches that can implement constraints on the DA cell counts. Another source is administrative records. This provides information on sections of the U.S. population based on Medicare, Social Security, and the Internal Revenue Service files. The administrative records may be considered as a third system in addition to the dual system E and P samples. The triple system approach and its modeling for the 1990 post-enumeration survey have been discussed in Darroch et al. (1993), Zaslavsky and Wolfgang (1993), and Chao and Tsay (1998).

In this paper we propose a local post-stratification approach for the dual system estimation for the 2000 Census A.C.E. study. The local stratification is made through nonparametric kernel estimation of the match and correct enumeration probability functions. The kernel estimation produces a local stratum around each value of the covariate $\mathbf{X}$. The size of the local stratum is allowed to shrink as the number of observations increases. This leads to the removal of the "correlation bias."

The proposed local post-stratification approach can accommodate categorical covariates. This suits the census well since large numbers of covariates are of this type. The local post-stratification approach constructs strata with respect to the categorical covariates by combining data within a ring of the neighboring strata. This potentially leads to variance reduction compared with the traditional post-stratification estimates that only utilize information from within each stratum.

The paper is organized as follows. Section 2 outlines the dual system capture recapture estimation. Section 3 discusses the missing values and erroneous enumerations in the A.C.E. Section 4 introduces the nonparametric kernel estimators. Section 5 discusses the missing value imputation and the proposed population size estimator. Section 6 discusses the smoothing bandwidth selection for the local post-stratification. Section 7 describes the post-stratification and logistic regression approaches. Section 8 analyzes the A.C.E. data and provides population undercount estimates for the U.S. population and various demographic and geographic subpopulations. Results from a simulation study are reported in Section 9. Section 10 discusses our conclusions.

2. DUAL SYSTEM ESTIMATION

Let $\mathcal{E}$ and $\mathcal{P}$ be the sets of individuals enumerated by the E and P samples, respectively, n_e and n_p be the respective sample sizes, and $\mathcal{E}^+$ be the set that includes the E sample and other census enumerations in one ring of blocks surrounding $\mathcal{E}$. Here $\mathcal{E}^+$ reflects the Target Extended Search (TES) operation in the A.C.E. that was designed to reduce the variance on the dual system estimates associated with census geocoding error. This also was designed to reduce bias from potential A.C.E. listing errors.

Let $\mathbf{X}_i$ be the covariate of the ith individual in $\mathcal{P}$ and $Y_i = 1$ if $i \in \mathcal{P} \cap \mathcal{E}^+$ (i.e., if $i \in \mathcal{P}$ but not in $\mathcal{E}^+$); otherwise, $Y_i = 0$. As $E(Y_i|\mathbf{X}_i = \mathbf{x}) = p(\mathbf{x})$ due to the capture–recapture design, the E-sample enumeration probability $p(\mathbf{x})$ can be estimated based on $\{(\mathbf{X}_i, Y_i)\}_{i=1}^{n_p}$ via binary regression. A heteroscedastic nonparametric regression model is $E(Y_i|\mathbf{X}_i) = p(\mathbf{X}_i)$ and $\operatorname{var}(Y_i|\mathbf{X}_i) = \sigma_p^2(\mathbf{X}_i)$ with the forms of $p(\mathbf{x})$ and $\sigma_p^2(\mathbf{x})$ both unknown.

Suppose that there are no erroneous enumerations, there are no missing values, and that each individual has a nonzero probability to be enumerated in the census. Let $\hat{p}(\mathbf{x})$ be a consistent estimator of $p(\mathbf{x})$. A Horvitz–Thompson type estimator of the population size N is

$$\hat{N} = \sum_{i \in \mathcal{C}} \frac{1}{\hat{p}(\mathbf{X}_i)}, \tag{2.1}$$

where $\mathcal{C}$ denotes the set of census records.

The above estimator is not the most efficient in dual system capture–recapture experiments. Let $q(\mathbf{x}) = P(\text{enumeration by the P sample}|\mathbf{X} = \mathbf{x})$ be the P-sample enumeration probability and $g(\mathbf{x}) = p(\mathbf{x}) + q(\mathbf{x}) - p(\mathbf{x})q(\mathbf{x})$ be the probability of being enumerated by either sample. Chen and Lloyd (2002) proposed a more efficient estimator than (2.1), with

$$\tilde{N} = \sum_{i \in \mathcal{C} \cup \mathcal{P}} \frac{1}{\hat{g}(\mathbf{X}_i)}, \tag{2.2}$$

where $\hat{g}(\mathbf{x})$ is a consistent estimator of $g(\mathbf{x})$. This is a two-way approach as the E-sample is used to calibrate the P-sample enumeration $q(\mathbf{x})$ as well. The estimators proposed by Huggins (1989) and Alho (1990) are of this type. However, the dual system A.C.E. is insufficient to do the estimation of the P-sample enumeration $q(\mathbf{x})$. Hence, the A.C.E. is a one-way approach instead. As noted in Alho et al. (1993) the one-way approach prevents one from obtaining the so-called triples that are needed to use the two-way estimator (2.2) based on logistic regression methods.

The one-way estimator (2.1) cannot be applied to the census dual system estimation without some modifications. This is due to the presence of missing values and erroneous enumerations. Modifications are also needed to include some extra variables that are related to the correct enumeration function and the missing value mechanism. An aim of this paper is to extend (2.1) so that it can be used for the A.C.E. estimation.

3. ISSUES IN A.C.E.

Missing values are significantly present in both samples of A.C.E. The match indicator, Y_i, may be missing if i is an unresolved case. Erroneous enumerations are invalid records and lead to overestimation of the population size. Hogan (1993) and Haberman, Jiang, and Spencer (1998) described two main sources of missing values. One source are those who should not have been enumerated. These include duplicates, fictitious records, and people born after or who died before the census. The other source are enumerations included in the wrong geographical location. The E sample is designed to identify the erroneous enumerations based on information collected during the initial or follow-up interview. Let n_e be the total number of enumerations in the E sample, which consists of both correct and erroneous enumerations. For $i = 1, \ldots, n_e$, let $e_i = 1$ if i is a correct enumeration, $e_i = 0$ if i is an erroneous enumeration. The indicator e_i is also subject to missing like Y_i.

Research on the U.S. Census reveals that in addition to the $\mathbf{X}$ (ROAST plus region), the groupings of match codes were another covariate related to enumeration status (U.S. Census Bureau 2004). These match code groupings (MCG) are mutually exclusive and exhaustive groups based on the match codes before the follow-up operation. Some information from the follow-up operation was coded in time to be used as well to form these groups. As illustrated in the analysis by Belin et al. (1993) of the 1990 post-enumeration survey, individuals with certain MCG status had a lower probability of being correctly enumerated. In general, let $\mathbf{S}_i$ be any extra variables that affect the correct enumeration status e_i. We assume a nonparametric regression model: $E(e_i|\mathbf{X}_i, \mathbf{S}_i) = e(\mathbf{X}_i, \mathbf{S}_i)$ and $\text{var}(e_i|\mathbf{X}_i, \mathbf{S}_i) = \sigma_e^2(\mathbf{X}_i, \mathbf{S}_i)$ where both $e(\mathbf{x}, \mathbf{s})$ and $\sigma_e^2(\mathbf{x}, \mathbf{s})$ are unknown. Here $e(\mathbf{x}, \mathbf{s})$ is the correct enumeration function.

Unlike $\mathbf{X}$, the extra covariates $\mathbf{S}$ are not available for the census enumerations outside the A.C.E. We overcome this issue by defining $e(\mathbf{x}) = E\{e(\mathbf{X}, \mathbf{S})|\mathbf{X} = \mathbf{x}\}$, which is the marginal correct enumeration function, by conditioning on the observable $\mathbf{X} = \mathbf{x}$. Then, the population size estimator (2.1) is modified to be

$$\hat{N} = \sum_{i \in \mathcal{C}} \frac{\hat{e}(\mathbf{X}_i)}{\hat{p}(\mathbf{X}_i)}, \tag{3.1}$$

where $\hat{e}(\mathbf{x})$ and $\hat{p}(\mathbf{x})$ are consistent estimators of $e(\mathbf{x})$ and $p(\mathbf{x})$, respectively.

Missing values and erroneous enumerations are likely to be present in other capture–recapture experiments for both wildlife and human populations. However, these issues, as far as we are aware, have not been well studied in the literature. This is partly due to the fact that the ability to identify erroneous enumerations may require substantial resources. It also appears that there is a lack of a general estimation theory about how to handle both erroneous enumerations and missing values in capture–recapture surveys.

4. LOCAL STRATIFICATION VIA KERNEL ESTIMATION

The key components in the population size estimator (3.1) are the estimates of $p(\mathbf{x})$ and $e(\mathbf{x})$. The proposed local post-stratification approach estimates these two functions based on nonparametric kernel smoothing. We consider in this section the case where there are no missing values in Y_i and e_i. The incorporation of missing values will be discussed in the next section.

The local post-stratification is an alternative to the post-stratification typically used in dual system estimation. Pollock (1976, 1991) proposed a parametric approach that can be used for the estimation of $p(\mathbf{x})$ and $e(\mathbf{x})$. The logistic regression model of Huggins (1989), Alho (1990), and Alho et al. (1993) may be modified to suit the one-way matching of the A.C.E. The main limitation of the parametric approach is the risk of model misspecification. When that happens, systematic model bias is present in the estimates. An aspect that encourages our proposed nonparametric estimation is that there is a large amount of data collected—the A.C.E. has more than 700,000 records in both samples. With this amount of data, the need to specify a parametric model is reduced. We illustrate local post-stratification in the estimation of $p(\mathbf{x})$. The estimate of $e(\mathbf{x}, \mathbf{s})$ is formulated in the same way.

The covariate $\mathbf{X}$ consists of both continuous and categorical variables. We treat age as a continuous variable as smoothing of an ordered categorical variable is essentially the same as the smoothing of a continuous variable (Simonnoff 1995). Most variables in the census are unordered categorical. Write $\mathbf{X}_i = (\mathbf{X}_i^c, \mathbf{X}_i^u)$ where $\mathbf{X}_i^c$ is d_c-dimensional and continuous and $\mathbf{X}_i^u$ is d_u-dimensional and unordered categorical with $d_c + d_u = d$.

To smooth the continuous covariates, we propose to use a d_c-dimensional kernel K, which is a radially symmetric probability density function in R^{d_c}. Let $K_h(\mathbf{x}) = h^{-d_c}K(\mathbf{x}/h)$ where h is a smoothing bandwidth than controls the amount of smoothness of the kernel estimate (Härdle 1990; Fan and Gijbels 1996).

Without loss of generality, we choose $K(\mathbf{u}) = \prod_{i=1}^{d_c} K_1(u_i)$, where $K_1(\cdot)$ is a symmetric univariate density function. We use $K_1(x) = 15/16(1-x^2)^2 I(|x| \le 1)$ throughout the paper.

To smooth unordered categorical covariates, we use the discrete kernel proposed by Aitchison and Aitken (1976); see also Hall (1981), Racine and Li (2004), and Hall, Racine, and Li (2004). Smoothing categorical variables is designed to utilize data from the neighboring strata to help improve the estimation efficiency. The neighboring strata may have similar characteristics as the target stratum. One way to define neighboring strata for a target cell $\mathbf{x}^u$ is based on the number of different components from $\mathbf{x}^u$. The nearest neighbors consist of cells that have only one different component, the second-nearest neighbors have two different components and so on. We believe the strata that are more similar to $\mathbf{x}^u$ should lend more information to the estimation at $\mathbf{x}^u$ than the strata with more differences. This data borrowing is ideally suited for the census since small post-strata sizes are often encountered.

Suppose X_{ij}^u, the jth component of $\mathbf{X}_i^u$, takes c_j values in $\{0, 1, \ldots, c_j - 1\}$. The bandwidth for smoothing X_{ij}^u is λ_j and the kernel weight at x_j^u is

$$\lambda_j I(X_{ij}^u = x_j^u) + \frac{1-\lambda_j}{c_j - 1} I(X_{ij}^u \ne x_j^u),$$

where $I(\cdot)$ is the indicator function and λ_j takes values in $[c_j^{-1}, 1]$. Assigning $\lambda_j = c_j^{-1}$ leads to a uniform weight irrespective of the difference between X_{ij}^u and x_j^u, whereas $\lambda_j = 1$ gives a weight of 1 if $X_{ij}^u = x_j^u$ and zero otherwise, which coincides with the standard frequency weight. The other λ_j values between c_j^{-1} and 1 offer a range of choices for efficiency improvement. The kernel used to smooth the entire $\mathbf{X}_i^u = (X_{i1}^u, \ldots, X_{id_u}^u)$ at $\mathbf{x}^u = (x_1^u, \ldots, x_{d_u}^u)$ is

$$L(\mathbf{x}^u, \mathbf{X}_i^u; \boldsymbol{\lambda}) = \prod_{j=1}^{d_u} \left\{ \lambda_j I(X_{ij}^u = x_j^u) + \frac{1-\lambda_j}{c_j - 1} I(X_{ij}^u \ne x_j^u) \right\}, \quad (4.1)$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{d_u})$ is the bandwidth vector. The overall kernel weight drawn from $\mathbf{X}_i = (\mathbf{X}_i^c, \mathbf{X}_i^u)$ for local estimation at $\mathbf{x} = (\mathbf{x}^c, \mathbf{x}^u)$ is $K_h(\mathbf{x}^c - \mathbf{X}_i^c) L(\mathbf{x}^u, \mathbf{X}_i^u; \boldsymbol{\lambda})$.

The kernel estimator of $p(\mathbf{x})$ with no missing values is

$$\hat{p}_0(\mathbf{x}) = \frac{\sum_{i=1}^{n_p} K_h(\mathbf{x}^c - \mathbf{X}_i^c) L(\mathbf{x}^u, \mathbf{X}_i^u; \boldsymbol{\lambda}) Y_i}{\sum_{i=1}^{n_p} K_h(\mathbf{x}^c - \mathbf{X}_i^c) L(\mathbf{x}^u, \mathbf{X}_i^u; \boldsymbol{\lambda})}. \quad (4.2)$$

The above estimator belongs to the class of Nadaraya–Watson type kernel regression estimators (Härdle 1990 and Signorini and Jones 2004). It has similar properties as the estimator considered in Racine and Li (2004) that accommodates both continuous and discrete covariates. The kernel estimator (4.2) carries out a weighted average of the binary responses Y_i in such a way that the "closer" a $\mathbf{X}_i$ is to $\mathbf{x}$, the larger the weight assigned by the a kernel. An estimator for $e(\mathbf{x}, \mathbf{s})$ based on the E-sample records is obtained by changing Y_i, n_p, and X_i in (4.2) to e_i, n_e, and $(\mathbf{X}_i, \mathbf{S}_i)$, respectively.

5. INCORPORATING MISSING VALUES

As illustrated earlier, both the match and correct enumeration indicators Y_i and e_i can have missing values. Let δ_i and η_i be the missing value indicators of Y_i and e_i, respectively, namely $\delta_i, \eta_i = 0$ (1) for missing (observed) Y_i, e_i. In this paper, we account only for missing values of Y_i and e_i. An extension to account for the missing values of covariates $\mathbf{X}_i$ will be considered in a separate study. Extensive effort has been made to carry out follow-up interviews to collect more information for subsequent matching. Reasons for sending cases out for a follow-up interview are related to the missingness in Y and e. We take this into account by incorporating extra variables $\mathbf{Z}$ in the estimation.

Missing at random (MAR) (Rosenbaum and Rubin 1983) is an important assumption in missing data analysis. This assumption means that $P(\delta_i = 1 | Y_i, \mathbf{X}_i) = P(\delta_i = 1 | \mathbf{X}_i) =: w_p(\mathbf{X}_i)$. Here, w_p is the missing propensity of Y_i.

The MAR in the above form may not be a realistic assumption for the census. Analyses on the census data indicate that extra variables in addition to $\mathbf{X}_i$ may be associated with the missingness. For instance, the match code groups (MCG) and the mover status (nonmover or in/out-mover) in the P sample have been shown to be influential in Belin et al. (1993) and Cantwell et al. (2001). To reflect these prior results, we shall also assume that in addition to $\mathbf{X}_i$, there are extra P-sample covariates $\mathbf{Z}_i^p$ that are related to the missingness of Y_i. Specifically, we assume that Y_i is missing at random given $(\mathbf{X}_i, \mathbf{Z}_i^p)$

$$P(\delta_i = 1 | Y_i, \mathbf{X}_i, \mathbf{Z}_i^p) = P(\delta_i = 1 | \mathbf{X}_i, \mathbf{Z}_i^p) =: w_p(\mathbf{X}_i, \mathbf{Z}_i^p), \quad (5.1)$$

where $w_p(\mathbf{x}, \mathbf{z})$ is the unknown P-sample missing propensity. At the same time, we assume $E(Y|\mathbf{X}, \mathbf{Z}^p) = p(\mathbf{X})$ and $\mathrm{var}(Y|\mathbf{X}, \mathbf{Z}^p) = \sigma^2(\mathbf{X})$ which means that $\mathbf{Z}^p$ does not have any predictive power for the enumeration of individuals.

For the missingness of e_i, we assume that e_i is MAR given $(\mathbf{X}_i, \mathbf{S}_i, \mathbf{Z}_i^e)$ where $\mathbf{Z}_i^e$ are extra E-sample variables that affect the missingness of e_i, namely

$$P(\eta_i = 1 | e_i, \mathbf{X}_i, \mathbf{S}_i, \mathbf{Z}_i^e) = P(\eta_i = 1 | \mathbf{X}_i, \mathbf{S}_i, \mathbf{Z}_i^e) =: w_e(\mathbf{X}_i, \mathbf{S}_i, \mathbf{Z}_i^e). \quad (5.2)$$

Like $\mathbf{Z}_i^p$, $\mathbf{Z}_i^e$ does not affect the correct enumeration, namely $E(e|\mathbf{X}, \mathbf{S}, \mathbf{Z}^e) = e(\mathbf{X}, \mathbf{S})$ and $\mathrm{var}(e|\mathbf{X}, \mathbf{S}, \mathbf{Z}^e) = \sigma^2(\mathbf{X}, \mathbf{S})$.

We now discuss estimation of $p(\mathbf{x})$ and $e(\mathbf{x}, \mathbf{s})$ in the presence of missing values. We first introduce the so-called complete case estimator

$$\hat{p}_c(\mathbf{x}) = \frac{\sum_{i=1}^{n_p} K_h(\mathbf{x}^c - \mathbf{X}_i^c) L(\mathbf{x}^u, \mathbf{X}_i^u; \boldsymbol{\lambda}) \delta_i Y_i}{\sum_{i=1}^{n_p} K_h(\mathbf{x}^c - \mathbf{X}_i^c) L(\mathbf{x}^u, \mathbf{X}_i^u; \boldsymbol{\lambda}) \delta_i}. \quad (5.3)$$

Although this estimator is consistent, its efficiency can be improved (Chen and Tang 2008a) by imputing each missing Y_i by $\hat{p}_c(\mathbf{X}_i)$. This leads to the imputation based estimator

$$\hat{p}(\mathbf{x}) = \frac{\sum_{i=1}^{n_p} K_h(\mathbf{x}^c - \mathbf{X}_i^c) L(\mathbf{x}^u, \mathbf{X}_i^u; \boldsymbol{\lambda}) \{\delta_i Y_i + (1-\delta_i) \hat{p}_c(\mathbf{X}_i)\}}{\sum_{i=1}^{n_p} K_h(\mathbf{x}^c - \mathbf{X}_i^c) L(\mathbf{x}^u, \mathbf{X}_i^u; \boldsymbol{\lambda})}. \quad (5.4)$$

The complete-case estimator for $e(\mathbf{x}, \mathbf{s})$, say $\hat{e}_c(\mathbf{x}, \mathbf{s})$, can be formulated in a same way by using the discrete kernel smoothing on $\mathbf{S}$. As $\mathbf{S}$ is not observable outside the A.C.E., we estimate the marginal correct enumeration $e(\mathbf{x}) = E\{e(\mathbf{X}, \mathbf{S})|\mathbf{X} = \mathbf{x}\}$ by

$$\hat{e}(\mathbf{x}) = \sum_{i=1}^{n_e} K_h(\mathbf{x}^c - \mathbf{X}_i^c) L(\mathbf{x}^u, \mathbf{X}_i^u; \boldsymbol{\lambda}) \times \{\eta_i e_i + (1 - \eta_i)\hat{e}_c(\mathbf{X}_i, \mathbf{S}_i)\} \Big/ \sum_{i=1}^{n_e} K_h(\mathbf{x}^c - \mathbf{X}_i^c) L(\mathbf{x}^u, \mathbf{X}_i^u; \boldsymbol{\lambda}). \tag{5.5}$$

With the kernel estimators $\hat{e}(\mathbf{x})$ and $\hat{p}(\mathbf{x})$, the local post-stratification estimator of the population size that takes into account both erroneous enumerations and missing values is

$$\hat{N} = \sum_{i \in \mathcal{C}} \frac{\hat{e}(\mathbf{X}_i)}{\hat{p}(\mathbf{X}_i)} \tag{5.6}$$

which is of the same type as those in Haberman, Jiang, and Spencer (1998) and Griffin (2005). Conditions under which (5.6) and $\hat{e}(\mathbf{x})$ are consistent are available in Chen, Tang, and Mule (2009). See also Chen and Tang (2008b) for more technical details.

The population size of a small geographic area or demographic domain, say $\mathcal{S}$, can be estimated by

$$\hat{N}_{\mathcal{S}} = \sum_{i \in \mathcal{C} \cap \mathcal{S}} \frac{\hat{e}(\mathbf{X}_i)}{\hat{p}(\mathbf{X}_i)}. \tag{5.7}$$

6. BANDWIDTH SELECTION

In implementing the nonparametric estimation, the smoothing bandwidths $(h, \boldsymbol{\lambda})$ need to be determined. We propose two-stage bandwidth selection based on a cross-validation. The first stage selects the bandwidth for smoothing the continuous variable, and the second stage selects the ones for smoothing the discrete variables.

For the estimation of $p(\mathbf{x})$, the discrete variables $\mathbf{x}^u$ are $\{x_1^u(\text{Race Domain}), x_2^u(\text{Region}), x_3^u(\text{Sex}), x_4^u(\text{Tenure})\}$ which define 112 cross-classification categories $\{\mathcal{C}(\mathbf{x}^u)\}$. While the three large race domains (White, Hispanic, and Black) have a large amount of data, the other domains experience different levels of sparsity. The same is seen for renters and owners. To reflect the varying levels of sparsity in the data, we select 112 h-bandwidths for each category. Let

$$\hat{p}_{h,\boldsymbol{\lambda}}^{-i}(\mathbf{x}) = \frac{\sum_{j \in \mathcal{P}, j \neq i} K_h(\mathbf{x}^c - \mathbf{X}_i^c) L(\mathbf{x}^u, \mathbf{X}_i^u; \boldsymbol{\lambda}) \delta_i Y_i}{\sum_{j \in \mathcal{P}, j \neq i} K_h(\mathbf{x}^c - \mathbf{X}_i^c) L(\mathbf{x}^u, \mathbf{X}_i^u; \boldsymbol{\lambda}) \delta_i}$$

be the complete case estimator formed by leaving out the ith individual, and let $h(\mathbf{x}^u)$ be the h-bandwidth used in the $\mathbf{x}^u$-category $\mathcal{C}(\mathbf{x}^u)$. The first-stage cross-validation (CV) score function for selecting $h(\mathbf{x}^u)$ is

$$CV_c(h) = \sum_{i \in \mathcal{C}(\mathbf{x}^u)} \{Y_i - \hat{p}_{h,\mathbf{1}}^{-i}(\mathbf{X}_i)\}^2 \delta_i,$$

found by setting the discrete bandwidths $\boldsymbol{\lambda}$ to $\mathbf{1}$, the vector of 1s. Let $h^{cv}(\mathbf{x}^u)$ be the bandwidth that minimizes $CV_c(h)$. We repeat this procedure for all 112 categories by changing $\mathbf{x}^u$ values.

In the second stage, we define the CV score for $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_4)$ as

$$CV_u(\boldsymbol{\lambda}) = \sum_{i \in \mathcal{P}} \left\{Y_i - \hat{p}_{h^{cv}(\mathbf{X}_i^u), \boldsymbol{\lambda}(\mathbf{X}_i^u)}^{-i}\right\}^2 \delta_i,$$

which is found by plugging in the $h^{cv}(\mathbf{x}^u)$ obtained in the first stage. We choose $\boldsymbol{\lambda}^{cv}(\mathbf{x}^u)$ by minimizing the above CV score. These procedures can be used to select bandwidths for $e(\mathbf{x}, \mathbf{s})$.

7. POST–STRATIFICATION AND LOGISTIC REGRESSION

When we analyze the census data, two other approaches are compared with the proposed local post-stratification estimator (5.6). The first is post-stratification. This approach has been used in the past three census dual system estimations. The second uses a logistic regression approach to estimate the match and correct enumeration probabilities. This is the approach that is planned to be implemented for the 2010 Census Coverage Measurement.

7.1 Post-Stratification

Post-stratification is designed to form homogeneous post-strata with respect to the census capture probabilities (Hogan 2003). In addition to reducing heterogeneity, the post-strata are used for synthetic estimation of subpopulations below the post-stratum level. To allow consistency and comparability between the local post-stratification and the post-stratification, both approaches use region, race/Hispanic origin domain, age, sex, and tenure as the covariates. Specifically, we use race/Hispanic origin (7 levels: American Indian or Alaska Natives on Reservation, Off-Reservation American Indian or Alaska Native, Hispanic, Non-Hispanic Black, Native Hawaiian or Pacific Islander, Non-Hispanic Asian, and Non-Hispanic white or other races), Region (4 levels: Northeast, Midwest, South, and West), housing tenure (2 levels: owner and renter), sex (2 levels), and age. For the three large race domains (Hispanic, Black, and White), the age and sex are combined to form 8 groups: (0–9, 10–17, 18–29 male, 18–29 female, 30–49 male, 30–49 female, 50+ male, and 50+ female). Each of the three larger domains has 64 post-strata, formed by crossing the 4 regions with the 2 tenures and the 8 age/sex groups. For the two smaller race domains, American Indian living on reservation and Non-Hispanic Asian, we combine the Northeast, Midwest, and South regions into one region. Thus, each of these two domains has 32 post-strata by crossing 2 regions with 2 tenures and 8 age/sex groups. For American Indians living off reservation, there are only 16 strata after we combine all regions into one. For the Native Hawaii and Pacific Islander domain, there are only 8 strata, formed by crossing 2 housing tenures with 4 age/sex groups: (0–9, 10–17, 18+ males, 18+ females). This collapsing led to a total of 280 post-strata that are different from the post-strata used for the March 2001 A.C.E. or the A.C.E. Revision II.

This post-stratification was based on research in Schindler (2008). It was able to utilize region more effectively when forming post-strata for several of the minority domains. The age/sex groupings are the same as those used in the A.C.E. Revision II estimation. The post-stratification for the A.C.E.'s

original estimates (Haines 2001) or the A.C.E. Revision II estimates (Kostanich 2003) used region only for the Non-Hispanic White domain.

Let $\{S_l\}_{l=1}^{280}$ be the 280 post-strata defined by the covariates in ROAST plus region. Let n_{pl} and n_{el} be respectively the numbers of P-sample and E-sample observations in the stratum S_l, and $\tilde{n}_{pel}$ be the number of definite matches in S_l. The underlying assumption of post-stratification is that both $p(\mathbf{x})$ and $e(\mathbf{x})$ functions are piecewise constant within each stratum. Under this assumption, the equivalent form of the imputation-based local post-stratification estimator (5.4) is

$$\hat{p}_{sl}(\mathbf{x}) = \frac{\tilde{n}_{pel} + n^{\star}_{pel}}{n_{pl}} \quad \text{for } \mathbf{x} \in S_l, \tag{7.1}$$

where $n^{\star}_{pel}$ is the number of imputed matches assigned to the unresolved cases. The post-stratification estimator $\hat{e}_{sl}(\mathbf{x})$ for $e(\mathbf{x})$ is similarly defined. The post-stratification based population size estimator is

$$\hat{N}_s = \sum_{l=1}^{280} \sum_{i \in \mathcal{S}_l} \frac{\hat{e}_{sl}(\mathbf{X}_i)}{\hat{p}_{sl}(\mathbf{X}_i)}. \tag{7.2}$$

We have not incorporated the sampling weights when formulating these estimators in order to keep the notation simple. Analysis of the A.C.E. data in Section 8 incorporates the weights.

7.2 Logistic Regression

The following logistic model with 86 main effects and interactions was considered for $p(\mathbf{x})$:

$$\begin{aligned}
&\text{logit}\{p_{log}(\mathbf{x})\} \\
&\quad = \text{Intercept} + \text{Age} + (\text{Age}^2 - (\text{Age} - 17)_+^2) \\
&\qquad + (\text{Age} - 17)_+ + (\text{Age} - 20)_+ + \{(\text{Age} - 20)_+^2 \\
&\qquad - (\text{Age} - 50)_+^2\} + (\text{Age} - 50)_+ + \text{Race} + \text{Sex} \\
&\qquad + \text{Tenure} + \text{Region} + \text{Race} * (\text{Tenure} + \text{Sex}) \\
&\qquad + \text{Tenure} * \text{Sex} + (\text{Tenure} + \text{Sex} + \text{Black} + \text{Asian} \\
&\qquad + \text{Indian on Reservation}) * \text{Age Splines} \\
&\qquad + \text{Regions} * (\text{Tenure} + \text{Hispanic} + \text{Hispanic} * \text{Tenure} \\
&\qquad + \text{Black} + \text{Black} * \text{Tenure} + \text{Asian} \\
&\qquad + \text{Asian} * \text{Tenure}),
\end{aligned} \tag{7.3}$$

where $(a)_+ = aI(a > 0)$ is the positive truncation function, race corresponds to the seven racial domains, and Hispanic, Black, Asian, and American Indian on Reservations are binary indicators of being in that domain. A similar model is used to model $e(\mathbf{x}, \mathbf{s})$.

In addition to the main effects and their interactions based on the four discrete variables race, sex, tenure and region, the logistic regression model uses six truncated polynomial pieces to form age splines. The age splines specify quadratic forms from age 0 to 17 and from age 20 to 50. It also specifies a linear form from 17 to 20 and from age 50 on; see Mule et al. (2007) for more details. The logistic model also includes interactions between the four discrete covariates and the six polynomial splines. This is an attempt to model the age effect continuously instead of keeping it fixed over post-strata. Smith (1979) presented a spline regression based on truncated polynomial functions that allows standard multiple regression procedures to be used. Malec et al. (1997) used piecewise linear splines to model the age effect in their small area inference in the National Health Interview Survey. Nandram and Choi (2005) used a similar join-point regression model of age in hierarchical Bayesian nonignorable nonresponse regression models for small area estimation.

A similar logistic regression model to (7.3) is used to model the correct enumeration function $e(\mathbf{x}, \mathbf{s})$. Griffin (2005) provides details on applying the logistic regression approach for missing values and extra covariate in estimation of $e(\mathbf{x}, \mathbf{s})$. Let $\hat{p}_{log}(\mathbf{x})$ and $\hat{e}_{log}(\mathbf{x})$ be the logistic regression estimators of $p(\mathbf{x})$ and $e(\mathbf{x})$. The logistic regression based population size estimator is

$$\hat{N}_{log} = \sum_{i \in \mathcal{C}} \frac{\hat{e}_{log}(\mathbf{X}_i)}{\hat{p}_{log}(\mathbf{X}_i)}. \tag{7.4}$$

The logistic regression approach represents a step forward from the post-stratification in actively modeling heterogeneity in the enumeration and correct enumeration due to covariates. The model of the heterogeneity is described fully parametrically by (7.3). Like any parametric approach, it is subject to the risk of model misspecification. The proposed local post-stratification is similar to the logistic regression in that both try to incorporate the age effect and the interactions among the covariates. The difference is that the local post-stratification approach does not impose the effects of the covariates explicitly and allows the data determine it.

8. ANALYZING CENSUS DATA

We analyzed the 2000 A.C.E. data using the three population size estimators: the local post-stratification, the post-stratification, and the logistic regression. All three estimators used the same set of covariates, ROAST plus region, in the estimation of $p(\mathbf{x})$ and ROAST plus region plus MCG in the estimation of $e(\mathbf{x}, \mathbf{s})$.

We used the March 2001 A.C.E. data to estimate $p(\mathbf{x})$ and $e(\mathbf{x})$ first. We then constructed population size estimates for each category in $\{\mathcal{C}(\mathbf{x}^u)\}$ defined by the four discrete covariates. Each record in the samples is properly weighted reflecting the multiphase sampling procedure in the census.

We are aware that there are some measurement and enumeration errors in the original A.C.E. estimates. Results reported here are for research purposes only to show the performance of different estimation procedures. We could carry out our estimation based on the A.C.E. Revision II data that corrected some errors in the A.C.E. These corrections were based on the results of additional computer matching and a further evaluation of a 10% subsample of the full A.C.E. data (U.S. Census Bureau 2004). Our choice of the original A.C.E. data is to show how different estimation approaches work for the usual dual system estimation situation without having to deal with the special A.C.E. issues. Also, since the full A.C.E. has 10 times more data than the evaluation sample, using the full data will largely reduce the variability of the final population size estimates.

In the 2000 census, a person with two or more recorded characteristics is called a data-defined person. Let $\mathcal{D}$ be the set of data-defined census persons, which is a subset of all census

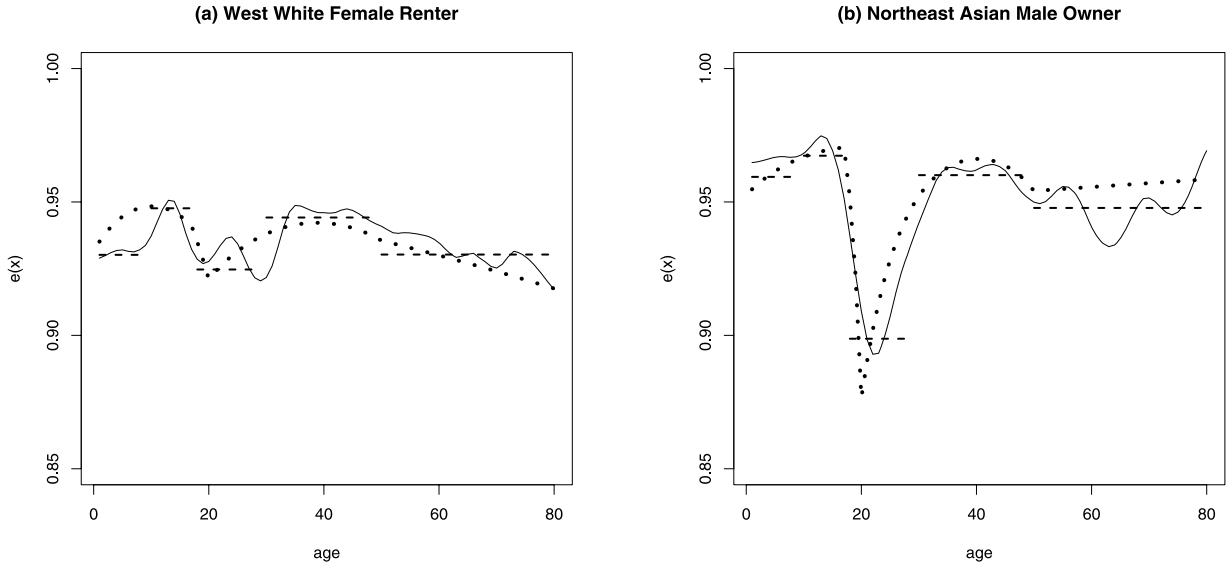

Figure 2. Estimation of the correct enumeration function $e(\mathbf{x})$ by the local post-stratification (solid line), post-stratification (dashed line), and logistic regression (dotted line).

persons $\mathcal{C}$. People with fewer than two characteristics recorded (non-data-defined persons) were removed from the E sample and the correct enumeration determination. These cases were also excluded from matching of the P sample to the census. The number of removed persons was about 8,000,000. For the estimators shown in (5.6), (7.2), and (7.4), we restrict $\mathcal{C}$ to $\mathcal{D}$. For the P sample, we used only the nonmovers and out-movers, namely Procedure A for handling movers. As expected, our estimates are lower than the A.C.E. estimates released in March 2001 that used the PES-C method (which used the in-movers to estimate the total mover population and out-movers to determine the match rate of the mover population) for most of the post-strata (U.S. Census Bureau 2004).

In addition to Figure 1 shown earlier, we display in Figure 2 the kernel estimates of $e(\mathbf{x})$ for the same categories of $\{\mathcal{C}(\mathbf{x}^u)\}$ shown in Figure 1. Both Figures 1 and 2 contain estimates using the post-stratification and the logistic regression approaches so comparisons can be made among the three. It is observed from these figures that the region and ROAST variables contribute significantly to the heterogeneity seen in both the enumeration and correct enumeration functions. The age effect was the most noticeable, as seen by the dip in enumeration probability between age 17 and 25. The kernel estimates also change significantly with respect to the other ROAST variables. Although the post-stratification accounted for much of the age heterogeneity, some still remained. The local post-stratification revealed more details on the heterogeneity than seen by the post-stratification. The logistic regression estimates were able to account for more heterogeneity than the post-stratification. However, its use of linear or quadratic polynomials over the predetermined age groups made its results less responsive to local data features, as is evident from the local post-stratification approach.

To gain information on the missingness, we show in Figure 3 the kernel estimates of the missing propensity score $w_e(\mathbf{x}, \mathbf{z}^e)$ with $\mathbf{z}^e$ being the matching code group (MCG) variable. These estimates can be constructed in the form of (4.2) by replacing Y_i with the missing indicator η_i. Figure 3 shows that, like the enumeration and correct enumeration functions, there is a lot of heterogeneity in the missing propensity function. Panel (a) shows that MCG is associated with different levels of missing propensities. The heterogeneity in the missingness also differs among the different race/domain groups, as seen in panels (b), (c), and (d). These confirm the association of MCG with missingness in the E-sample correct enumeration.

For the post-stratification and logistic regression estimates, we implemented similar missing data procedures as those used in the A.C.E. (Cantwell and Ikeda 2003). When the correct enumeration or match status was missing, mean cell imputation was used. For the correct enumeration status, initial cells were formed based on the match code groupings. As in the A.C.E., additional cells were formed within some of the match code groupings. Three of the groups were split into (i) Non-Hispanic whites and (ii) all others. The partial household nonmatches group was split into (i) whether the case was an 18 to 29 year old child of the reference person or (ii) all other cases. For match status, nonmovers and out-movers were the two cells formed. The estimates from these two approaches did not utilize certain covariates, such as the amount of characteristic imputation or the results of housing unit matching, that were used in the A.C.E. This was done so that these results could be more comparable to the results based on the initial $\mathbf{X}$ and extra $\mathbf{Z}$ covariates used in the local post-stratification approach.

As can be seen from the estimator shown in (5.6), the functional ratio of $e(\mathbf{x})/p(\mathbf{x})$ contributes significantly to the population size estimates. We call it the correction ratio function. If, in an ideal situation, $e(\mathbf{x})/p(\mathbf{x}) \equiv 1$ for all $\mathbf{x}$ or is piecewise constant with respect to the post-stratification, then there would be no bias in the population size estimates despite the heterogeneity in both $p(\mathbf{x})$ and $e(\mathbf{x})$ seen in Figures 1 and 2. This would mean that the bias due to the heterogeneity in $p(\mathbf{x})$ and $e(\mathbf{x})$ are canceling each other. To check on this possibility, we plot in Figure 4 estimates of $e(\mathbf{x})/p(\mathbf{x})$ based on the three estimation

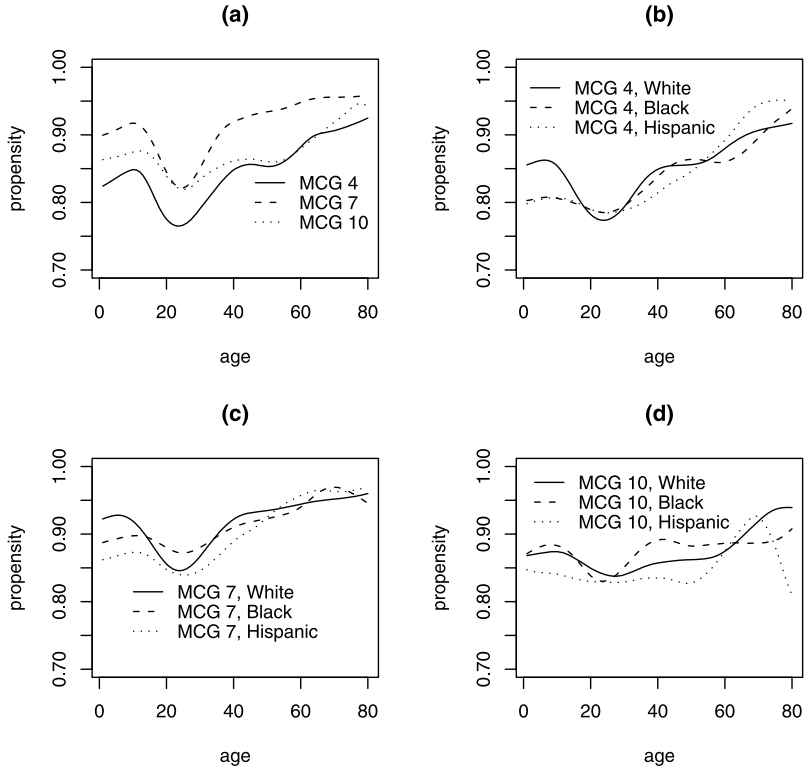

Figure 3. Estimates of the E-sample missing propensity function by the local post-stratification for selected MCG groups and the racial domains; MCG 4—Whole household nonmatches where the housing units matched, MCG 7—Household nonmatches where the housing units did not match in housing units matching, and MCG 10—Target Extended Search People.

approaches. As a reference, we also show the 95% confidence bands for the local post-stratification estimates based on a jackknife method whose details will be provided later.

It is observed from Figure 4 that there is indeed some cancellation of the heterogeneity between $p(\mathbf{x})$ and $e(\mathbf{x})$. However, the heterogeneity is still apparent in the correction ratios although it is much less than the heterogeneity observed for $p(\mathbf{x})$ and $e(\mathbf{x})$. The reduction is expected because of a strong positive correlation between the $p(\mathbf{x})$ and $e(\mathbf{x})$ estimates due to overlapping E and P samples. We note from Figure 4 that the logistic regression estimates of the ratio were quite different than the estimates from the local post-stratification and post-stratification approaches over a large interval of ages. It suggests an increasing and then decreasing trend in the ratio that is probably due to

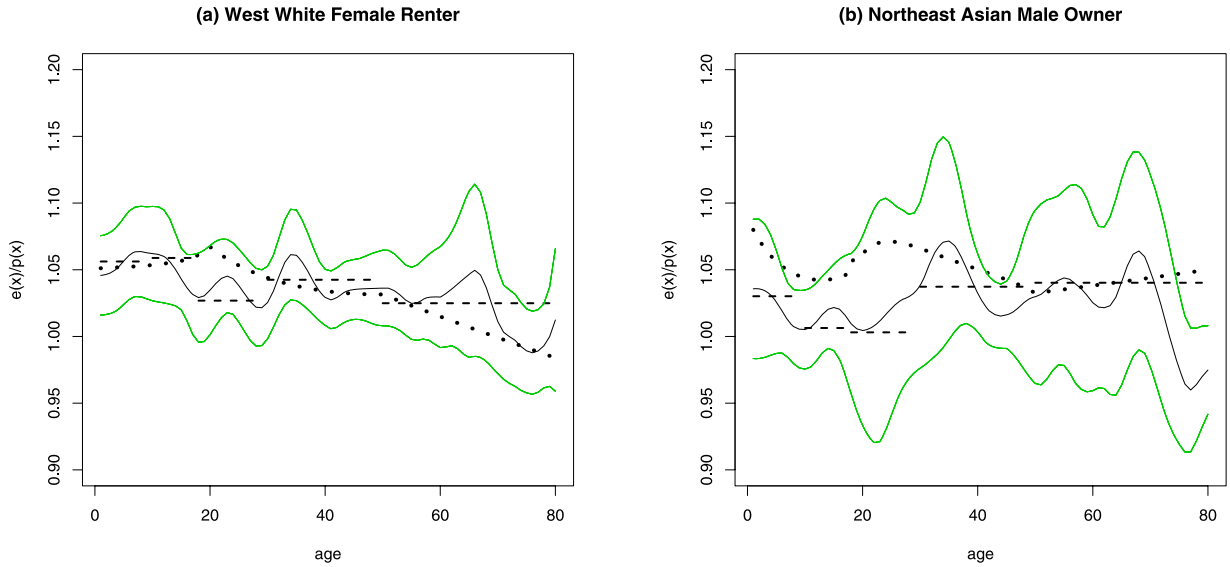

Figure 4. Estimation of the ratio of $e(\mathbf{x})/p(\mathbf{x})$ by local post-stratification with 95% estimated confidence bound (solid line), post-stratification (dashed line), and logistic regression (dotted line). A color version of this figure is available in the electronic version of this article.

the forms of the age splines. Overall, we see that the local post-stratification varies from the post-stratification estimates with local post-stratification showing more local features. It is reasonable to say that the correction ratios are heterogeneous with respect to the covariate $\mathbf{X}$. Hence, the correlation bias will be present in the dual system estimates when the post-stratification is applied. This is especially a concern for any synthetic estimates from the post-stratification approach as the estimates would be subject to synthetic error.

We then calculated population estimates for each of the 112 categories $\{\mathcal{C}(\mathbf{x}^u)\}$ defined by the discrete covariates. Let $N(\mathbf{x}^u)$ be the unknown population size of $\mathcal{C}(\mathbf{x}^u)$. We used the census counts for these cells as the baseline and report the percentages of undercounts based on the population estimate

$$\hat{N}(\mathbf{x}^u) = \sum_{i \in \mathcal{C}(\mathbf{x}^u)} \frac{\hat{e}(\mathbf{X}_i)}{\hat{p}(\mathbf{X}_i)}. \tag{8.1}$$

Let $|\mathcal{C}(\mathbf{x}^u)|$ be the census count for $\mathcal{C}(\mathbf{x}^u)$, so the total census count is $|\mathcal{C}| = \sum_{\mathbf{x}^u} |\mathcal{C}(\mathbf{x}^u)|$. The percentages of undercount in $\mathcal{C}(\mathbf{x}^u)$ and in the entire census are, respectively,

$$u(\mathbf{x}^u) = \frac{\hat{N}(\mathbf{x}^u) - |\mathcal{C}(\mathbf{x}^u)|}{\hat{N}(\mathbf{x}^u)} \quad \text{and} \quad u = \frac{\hat{N} - |\mathcal{C}|}{\hat{N}}. \tag{8.2}$$

To obtain the standard errors, we used a jackknife variance estimation approach (Shao and Wu 1989), which is commonly used in estimating variance in survey samples (Shao and Tu 1995 and Wolter 2007). The conventional delete-one jackknife variance estimation approach is not applicable since the primary sampling units (PSU) in the A.C.E. are clusters of collection blocks. Deleting one individual or even one household is inadequate in reflecting the multiphase sampling procedure used in the A.C.E. and would result in biased variance estimates. To respect the multiphase sampling procedure in the A.C.E., we considered the delete-one-PSU method for the jackknife variance estimation. Because the delete-one PSU approach results in a large number of replicates (around 30,000), we formed 100 random groups according to the last two digits of the cluster identification numbers. We then performed the jackknife variance estimation based on the 100 random groups. For the proposed local post-stratification estimation, let $\hat{p}^{-j}(\mathbf{x})$ and $\hat{e}^{-j}(\mathbf{x})$ be the estimates of $p(\mathbf{x})$ and $e(\mathbf{x})$ made by leaving out the jth group. The leave-out-one-group correction ratio estimate is $\hat{r}^{-j}(\mathbf{x}) = \hat{e}^{-j}(\mathbf{x})/\hat{p}^{-j}(\mathbf{x})$. Based on the 100 jackknife replicates, the variance of $\hat{r}(\mathbf{x})$ is estimated by

$$v_r^2(\mathbf{x}) = \frac{99}{100} \sum_{j=1}^{100} \{\hat{r}_k^{-j}(\mathbf{x}) - \hat{r}_k(\mathbf{x})\}^2, \tag{8.3}$$

which is used to form the 95% confidence bands for the correction ratio as shown in Figure 4. The jackknife replicates for the populations size are, for $j = 1, \ldots, 100$,

$$\hat{N}^{-j} = \sum_{i \in \mathcal{C}} \frac{\hat{e}^{-j}(\mathbf{X}_i)}{\hat{p}^{-j}(\mathbf{X}_i)},$$

and the jackknife variance estimate of $\hat{N}$ is $v_N^2 = \frac{99}{100} \times \sum_{j=1}^{100} (\hat{N}^{-j} - \hat{N})^2$. The jackknife variance estimation of the population size estimates and percentages of undercounts for other subpopulations can be carried out in a similar fashion.

Table 1. The estimates of percentage undercounts and their standard errors (in parentheses) for the local post-stratification (Local P-S), the post-stratification (P-S), and the logistic regression (Logistic) estimates for selected demographic aggregates. AIANR means American Indian or Alaska Native on reservations while AIANO means American Indian or Alaska Native off reservations

Domain	Local P-S	P-S	Logistic
U.S. National	0.94 (0.15)	0.92 (0.15)	0.92 (0.15)
AIANR	3.47 (1.66)	4.48 (1.19)	4.45 (1.14)
AIANO	2.45 (1.19)	2.81 (1.14)	2.57 (1.11)
Hispanic	2.46 (0.37)	2.45 (0.36)	2.44 (0.36)
Non-Hisp Black	1.87 (0.31)	1.82 (0.30)	1.83 (0.31)
Hawaiian & PI	5.10 (3.11)	4.59 (2.87)	4.55 (2.91)
Non-Hisp Asian	0.63 (0.66)	0.56 (0.66)	0.51 (0.65)
White & Other	0.48 (0.15)	0.47 (0.16)	0.46 (0.16)
Northeast	0.21 (0.27)	0.19 (0.27)	0.27 (0.28)
Midwest	0.30 (0.18)	0.30 (0.18)	0.28 (0.19)
South	1.63 (0.24)	1.60 (0.24)	1.56 (0.25)
West	1.09 (0.30)	1.08 (0.30)	1.09 (0.30)
Owner	0.22 (0.17)	0.23 (0.17)	0.22 (0.17)
Renter	2.48 (0.25)	2.40 (0.26)	2.40 (0.25)

Table 1 reports the percentages of undercounts and their standard errors for the entire U.S., the seven race/origin domains, the four regions, and housing tenure using the three estimation approaches. Figure 5 shows the undercounts for the 80 age groups and Figure 6 shows the correction ratios for four of the 112 demographic categories. The undercount estimates among the three estimation methods are not that different for broader aggregates of demographic categories including the entire U.S., the two types of housing tenure, and the four regions. This is satisfying and shows that the effect of the correlation bias with the post-stratification estimate is not very severe for larger population aggregates. This is not due to a lack of the correlation bias in the dual system estimation, but rather the can-

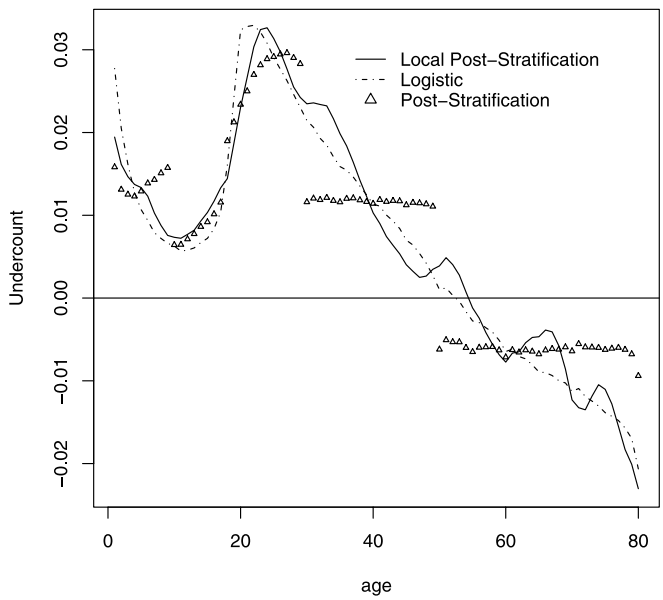

Figure 5. Estimated percentage undercount for single age groups by local post-stratification, post-stratification, and logistic regression.

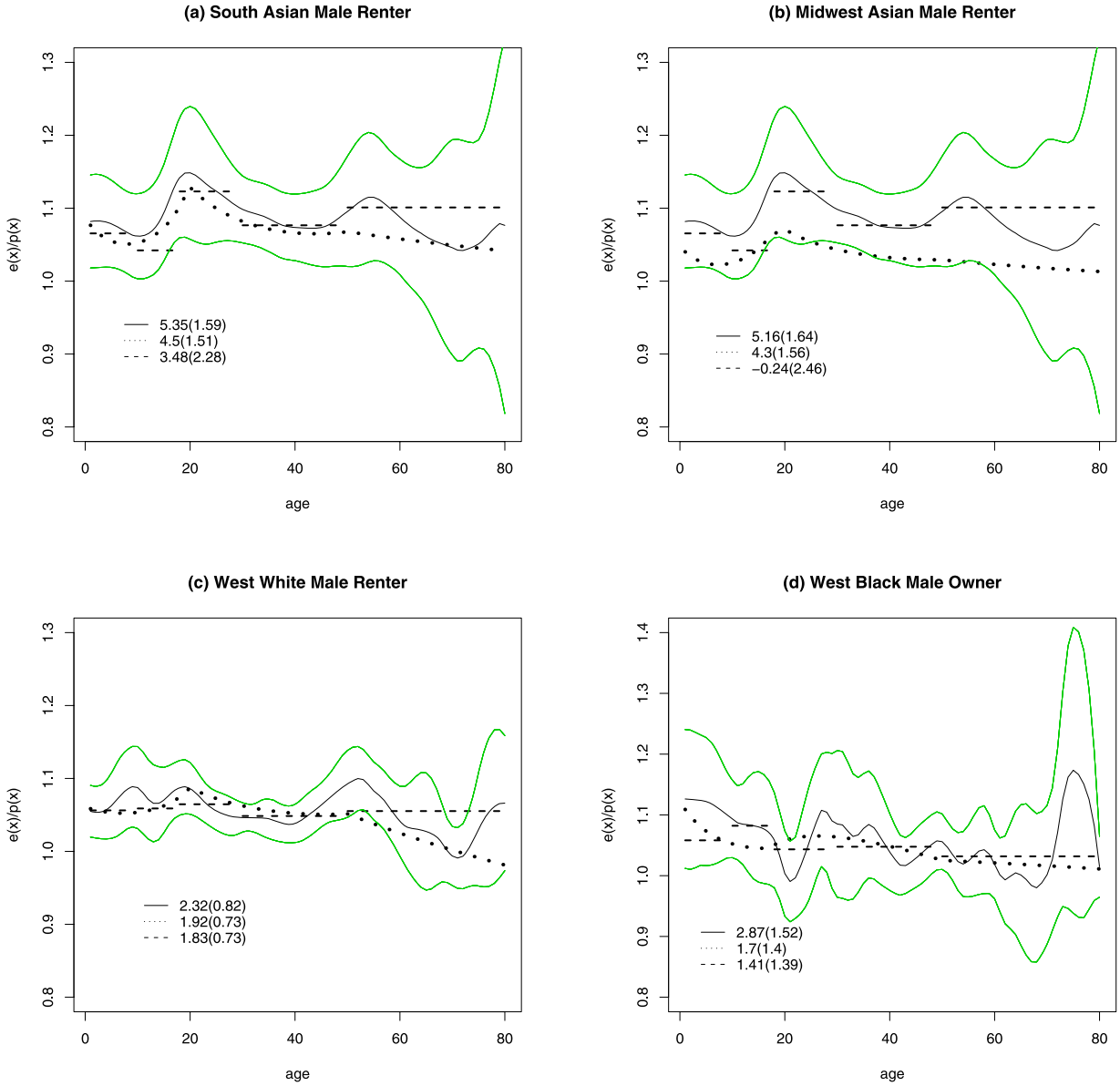

Figure 6. Estimated correct ratio $e(\mathbf{x})/p(\mathbf{x})$ and undercounts (given in the legends) for some demographic categories by the local post-stratification (solid line), post-stratification (dashed line), and logistic regression (dotted line). The legends report the percentages of undercount and their standard errors in parentheses. A color version of this figure is available in the electronic version of this article.

cellation of biases among sections of the population. Indeed, when we evaluate estimates for smaller subpopulations, larger differences emerge. These differences are seen in the estimates for smaller racial domains in Table 1 and for single years of age in Figure 5. The difference in undercount estimates are as large as 1% in some places, which is substantial considering the size of these subpopulations. Figure 6 shows that for even smaller demographic categories there are large discrepancies among these estimates.

The largest discrepancy (almost 1%) among the Table 1 estimates is for the American Indians living on reservations, which has the smallest sample size in A.C.E. As the data is sparse in certain areas of the covariate space (in particularly for some of the age intervals), the local post-stratification employed larger bandwidths to smooth for age. Plots of the correction ratios show large differences for the three regions other than "West" between the local post-stratification and the other two approaches. This led to a lower population size estimate by the local post-stratification approach. We note the undercount estimates had a larger standard error for another small domain, the Native Hawaiian and Pacific Islander. This is likely caused by the sparsity of data among the jackknife replications in the variance estimation.

The local post-stratification and the logistic regression produced similar single age undercount estimates. The post-stratification significantly underestimates the undercount for ages between 30 and 40, and 50 to 60, and overestimates the results in some other age intervals as well. The local post-stratification detects age heaping at ages 50 and 65, which reflects a known aspect of the census data. At the same time, the logistic model produced a smoother trend of the undercount es-

Table 2. State level research estimates of undercount percentage and their standard errors (in parentheses) for the local post-stratification (Local P-S), the post-stratification (P-S), and logistic regression (Logistic)

			Local P-S		P-S		Logistic	
State	r	$\hat{u}$	$\hat{u}_1$	$\hat{u}_2$	$\hat{u}_1$	$\hat{u}_2$	$\hat{u}_1$	$\hat{u}_2$
				Northeast region				
MA	97.5	0.27	0.96 (0.27)	0.25 (0.27)	0.93 (0.27)	0.23 (0.27)	1.01 (0.28)	0.31 (0.29)
NY	95.1	0.4	−0.73 (0.29)	0.37 (0.29)	−0.77 (0.29)	0.33 (0.29)	−0.68 (0.31)	0.43 (0.3)
VT	96.1	0.18	−1.05 (0.29)	0.18 (0.29)	−1.08 (0.3)	0.17 (0.29)	−1.05 (0.3)	0.19 (0.3)
				Midwest region				
IL	96.6	0.55	−0.23 (0.24)	0.59 (0.24)	−0.24 (0.24)	0.57 (0.24)	−0.29 (0.24)	0.52 (0.24)
IA	98.4	0.15	0.52 (0.17)	0.16 (0.17)	0.52 (0.17)	0.16 (0.17)	0.49 (0.17)	0.12 (0.17)
NE	98.4	0.33	0.93 (0.18)	0.34 (0.18)	0.93 (0.18)	0.34 (0.18)	0.89 (0.19)	0.31 (0.19)
				South region				
DC	96.1	2.44	2.14 (0.34)	2.51 (0.34)	2.1 (0.34)	2.46 (0.34)	2.03 (0.35)	2.4 (0.35)
LA	97.1	1.59	1.74 (0.23)	1.64 (0.23)	1.69 (0.23)	1.6 (0.23)	1.65 (0.24)	1.56 (0.24)
TX	96.5	1.89	1.72 (0.32)	1.95 (0.32)	1.68 (0.32)	1.91 (0.32)	1.65 (0.33)	1.87 (0.33)
				West region				
AK	97.4	1.12	1.49 (0.33)	1.15 (0.33)	1.42 (0.32)	1.08 (0.32)	1.39 (0.35)	1.05 (0.35)
NV	96.1	1.05	0.2 (0.3)	1.08 (0.3)	0.16 (0.29)	1.05 (0.29)	0.17 (0.3)	1.06 (0.3)
NM	95.5	1.28	0.32 (0.28)	1.22 (0.28)	0.38 (0.28)	1.27 (0.28)	0.39 (0.29)	1.29 (0.29)

timates at the older ages as the age-splines did not allow any local variation.

We then calculated population size estimates for the 50 states and the District of Columbia (DC). This allowed us to evaluate the results that the three approaches produced for smaller local geographic areas. The states and DC are nested within the four regions which have been used in the analysis so far. Brown and Zhao (2008) has shown that estimates of smaller geographic/demographic areas are sensitive to synthetic assumptions, the non-data-defined cases and imputations. We consider two sets of state estimates derived from a generic form in line with (5.7):

$$\hat{N}_{\mathcal{S}} = \sum_{i \in \mathcal{C} \cap \mathcal{S}} \{1 - \Pi(\mathbf{X}_i)\} \frac{\hat{e}(\mathbf{X}_i)}{\hat{p}(\mathbf{X}_i)}, \tag{8.4}$$

where $\Pi(\mathbf{X}_i)$ denotes a rate function for an individual with covariate $\mathbf{X}_i$ for being non-data-defined, and $\mathcal{S}$ denotes the state. To examine the sensitivity to the non-data-defined cases, two forms of $\Pi(\mathbf{x})$ were considered that led to two sets of state estimates.

Results for selected states are shown in Table 2. The first set of state estimates ($\hat{u}_1$ in Table 2) were obtained by assigning $\Pi(\mathbf{X}_i)$ to be the indicator of whether the person record was non-data-defined. This gave the same estimates as those shown earlier that used (8.1). This estimator is equivalent to the first alternative method in Brown and Zhao (2008). The second set of state estimates ($\hat{u}_2$) was obtained by modeling the rate function $\Pi(\mathbf{X}_i)$ based on the census records. We used region, tenure, and basic demographic variables to create 8,960 cross classifications of covariates for the local post-stratification and 9,072 for the post-stratification and logistic regression in the estimation of $\Pi(\mathbf{X}_i)$. This is similar to the Census Bureau's Census Coverage Correction Factor approach (Hogan 2003; Brown and Zhao 2008). The difference in the numbers of cross classifications was due to using different minimum age between the local post-stratification and the other two methods. The local post-stratification estimation started from age 1 by merging ages 0 and 1 together, while the other two methods started from age 0. We also show estimates ($\hat{u}$) where we estimate the $\Pi(\mathbf{X}_i)$ rate, correct enumeration rate and match rate for the 280 post-strata. This is the same synthetic approach used by the bureau as reported in Hogan (2003). This approach allocated the population in each post-stratum proportionally to subgroup census counts (including imputations).

Table 2 reports the results of the two sets of undercount estimates for selected states grouped by the four census regions together with their estimated standard errors. The results show that the three estimation methods produced similar results within each set of $\hat{u}_1$ or $\hat{u}_2$ estimates. There are some important differences between the two sets of estimates for each of the three estimating methods. This is similar to the findings in Brown and Zhao (2008). It is seen that states with lower data-defined rates like New York tended to have overcounts in the $\hat{u}_1$-estimates but have undercounts in the $\hat{u}_2$-estimates. The reverse was seen in the Midwest states like Nebraska and Iowa. Results from the second set of estimates were very comparable to the synthetic estimates based on 280 post-strata using the bureau's A.C.E. synthetic approach.

9. SIMULATION STUDY

To provide a finite sample evaluation on the performance of the three estimation methods, we conducted a simulation study on a simulated population with a structure that mimics the U.S. population as described by the 2000 census. The size of the simulated population was 281,421,906, which was the same as the 2000 census count. The population was created by generating each person's state according to the census state population distribution. Each person after being assigned a state was given the ROAST values according to the census marginal frequency distributions for the state. We also tried to match the race by

housing tenure distribution for each state. Except for the race-tenure dependence, the ROAST variables within each state were assumed to be mutually independent.

We designed $p(\mathbf{x})$ and $e(\mathbf{x})$ (no extra covariate $\mathbf{Z}$) as well as the missing propensity $w_p(\mathbf{x})$ and $w_e(\mathbf{x})$ in order to create the simulated P and E samples, respectively. We tried to reflect the heterogeneities in these functions as was revealed in the empirical local post-stratification estimation for the A.C.E data. The population heterogeneity was dependent on region (X_1), race (X_2), (housing) tenure (X_3), sex (X_4), and age (X_5), with X_0 being the latent state variable. The four discrete covariates (X_0, X_2, X_3, X_4) constituted 1428 nonoverlapping cells to which each person's characteristics (without age) belong uniquely.

Let

$$l(X_0, X_2, X_3, X_4) = 28(X_0 - 1) + 4(X_2 - 1) + 2(X_3 - 1) + X_4, \quad (9.1)$$

be a one-to-one mapping from the 1428 cells, and $u(t; \boldsymbol{\beta}) = [1 + \exp\{-b(t; \boldsymbol{\beta})\}]^{-1}$ where $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{10})$,

$$b(t; \boldsymbol{\beta}) = \beta_0 + \beta_1 t + \beta_2 \phi\left(\frac{t - \beta_3}{\beta_4}\right) + \beta_5 \phi\left(\frac{t - \beta_6}{\beta_7}\right) + \beta_8 \phi\left(\frac{t - \beta_9}{\beta_{10}}\right),$$

$\phi(\cdot)$ is the standard normal density, and $(\beta_2, \beta_5, \beta_8)$, $(\beta_3, \beta_6, \beta_9)$, and $(\beta_4, \beta_7, \beta_{10})$ represent, respectively, the location, dispersion, and the magnitude parameters. Having three $\phi(\cdot)$s in $b(t; \boldsymbol{\beta})$ mimics the fluctuations in $p(\mathbf{x})$ as observed in the empirical study.

At each cell determined by (X_0, X_2, X_3, X_4), we assigned $p(\mathbf{x}) = u(x_5; \boldsymbol{\beta}_l^{(p)})$ and $g(\mathbf{x}) = u(x_5, \boldsymbol{\beta}_l^{(g)})$ where the subscript $l = l(X_0, X_2, X_3, X_4)$ as given in (9.1) and x_5 denotes the age. Here $g(\mathbf{x})$ is the enumeration function that generated the P sample in the simulation. To create a difference in heterogeneity among cells and to make the state variability within one region smaller than the variability observed between two regions, we randomly generated the parameters $\boldsymbol{\beta}_l^{(p)}$ and $\boldsymbol{\beta}_l^{(g)}$ from the following additive model:

$$\boldsymbol{\beta}_l = \boldsymbol{\beta}_{0,X_0} + \boldsymbol{\beta}_{1,X_1} + \boldsymbol{\beta}_{2,X_2} + \boldsymbol{\beta}_{3,X_3} + \boldsymbol{\beta}_{4,X_4},$$

where for any X_a, $a \in \{0, \ldots, 4\}$, $\boldsymbol{\beta}_{a,X_a}$ is a 11-dimensional normal random vector with mean $\boldsymbol{\mu}_a$ and covariance matrix $\boldsymbol{\Sigma}_a$. We set $\boldsymbol{\mu}_2^{(p)}$ and $\boldsymbol{\Sigma}_2^{(p)}$ as those given in Table 3, $\boldsymbol{\mu}_0^{(p)} = \boldsymbol{\mu}_1^{(p)} = \boldsymbol{\mu}_3^{(p)} = \boldsymbol{\mu}_4^{(p)} = 0$, $\boldsymbol{\Sigma}_0^{(p)} = 0.04\boldsymbol{\Sigma}_2^{(p)}$, $\boldsymbol{\Sigma}_1^{(p)} = 0.25\boldsymbol{\Sigma}_2^{(p)}$, $\boldsymbol{\Sigma}_3^{(p)} = 1.25\boldsymbol{\Sigma}_2^{(p)}$, and $\boldsymbol{\Sigma}_4^{(p)} = 0.09\boldsymbol{\Sigma}_2^{(p)}$. We recreated some of the empirical features of $p(\mathbf{x})$, for instance the White had the highest $p(\mathbf{x})$ among the seven racial domains, and owners had higher $p(\mathbf{x})$ than renters. Similarly, we set $\boldsymbol{\mu}_2^{(g)}$ to the values in Table 2, $\boldsymbol{\mu}_a^{(g)} = 0$ for $a \in \{0, 1, 3, 4\}$ and $\boldsymbol{\Sigma}_a^{(g)} = \boldsymbol{\Sigma}_a^{(p)}$ for $a \in \{0, \ldots, 4\}$.

The correct enumeration function $e(\mathbf{x})$ was

$$e(\mathbf{x}) = \left[1 + \exp\left\{-b\big(x_5; \boldsymbol{\beta}_l^{(e)}\big) + \phi\left(\frac{x_5 - 65}{2}\right) - \phi\left(\frac{x_5 - 70}{2}\right) + \phi\left(\frac{x_5 - 75}{2}\right)\right\}\right]^{-1}$$

which has three more $\phi(\cdot)$-terms than $p(\mathbf{x})$ to allow more fluctuation with respect to age, and $\boldsymbol{\beta}_l^{(e)}$ was specified in a same way as $\boldsymbol{\beta}_l^{(p)}$. The missing propensities $w_p(\mathbf{x})$ and $w_e(\mathbf{x})$ were established according to $w_p(\mathbf{x}) = u(x_5; \boldsymbol{\beta}_l^{(w_p)})$ and $w_e(\mathbf{x}) = u(x_5; \boldsymbol{\beta}_l^{(w_e)})$, where $\boldsymbol{\beta}_l^{(w_p)}$ and $\boldsymbol{\beta}_l^{(w_e)}$ were similarly specified as $\boldsymbol{\beta}_l^{(p)}$ with $\boldsymbol{\mu}_2^{(e)}$, $\boldsymbol{\mu}_2^{(w_p)}$, and $\boldsymbol{\mu}_2^{(w_e)}$ given in Table 3, and the other means were 0 and all the other covariate matrices were the same as those of the $\boldsymbol{\beta}^{(p)}$s.

In each simulation, we applied the above $p(\mathbf{x})$ and $g(\mathbf{x})$ on each person in the simulated super-population, which gave rise to two independently simulated censuses: $\mathcal{C}$ by applying $p(\mathbf{x})$ that mimics the U.S. Census and $\mathcal{P}$ that leads to the P-sample. By comparing the simulated $\mathcal{C}$ with the simulated super-population, the population undercounts for each subpopulation as categorized in Table 1 were obtained. We generated the E and P samples by carrying out a stratified simple random sampling over each state according to the census state population frequency distribution. Both E and P sample sizes were 1 million which was about 1/3 more than the A.C.E. Applying $e(\mathbf{x})$ and $w_e(\mathbf{x})$ on the E sample created both erroneous enumerations and unresolved correct enumerations, and a similar application of $w_p(\mathbf{x})$ led to unresolved matches between the E and P samples.

For each simulated E and P samples, we estimate the entire population and subpopulations using the three estimation methods. The jackknife variance estimation was implemented to estimate the variance of the undercount. The bandwidths of the local post-stratification were selected by the cross-validation procedure. All the simulation results were obtained based on 2000 simulations.

Table 3. Parameters used in the simulation

Parameter	Parameter value
$\boldsymbol{\mu}_2^{(p)}$	$(2.2, 0.0083, -10.7, 24, 4.3, -2.0, 55, 2.0, 2.0, 60, 2.0)^T$
$\boldsymbol{\mu}_2^{(g)}$	$(2.4, 0.0073, -8.7, 28, 5.3, -1.0, 55, 2.0, 1.0, 60, 2.0)^T$
$\boldsymbol{\mu}_2^{(e)}$	$(3.2, -0.0063, -8.0, 21, 5.3, -1.0, 55, 2.0, 1.0, 60, 2.0)^T$
$\boldsymbol{\mu}_2^{(w_p)}$	$(4.0, -0.0063, -8.0, 23, 3.3, -1.0, 55, 2.0, 0, 0, 0)^T$
$\boldsymbol{\mu}_2^{(w_p)}$	$(3.5, -0.0063, -8.0, 25, 5, -1.0, 55, 2.0, 0, 0, 0)^T$
$(\boldsymbol{\Sigma}_2^{(p)})^{1/2}$	$\mathrm{diag}(0.5, 0.005, 1.5, 2, 1, 0.1, 1, 0.1, 0.1, 1, 0.1)$

Table 4. Simulation based Bias, Standard Deviation (SD), and Root Mean Squared Error (RMSE) in the estimation of the true undercount percentages (U) using the local post-stratification (Local P-S), post-stratification (P-S), and logistic regression (Logistic), together with the jackknife standard deviation estimation ($\hat{SD}$)

		Local P-S		P-S		Logistic	
Domain	U	Bias SD	RMSE ($\hat{SD}$)	Bias SD	RMSE ($\hat{SD}$)	Bias SD	RMSE ($\hat{SD}$)
Overall	3.42	0.002 0.039	0.039 (0.038)	−0.016 0.040	0.043 (0.038)	−0.006 0.040	0.040 (0.038)
Race 1	6.22	0.040 0.353	0.354 (0.352)	−0.042 0.353	0.355 (0.340)	−0.035 0.343	0.345 (0.333)
Race 2	7.38	0.013 0.157	0.159 (0.162)	−0.042 0.156	0.161 (0.161)	−0.079 0.155	0.173 (0.169)
Race 3	3.16	0.030 0.240	0.242 (0.262)	0.040 0.240	0.242 (0.268)	−0.020 0.240	0.241 (0.264)
Race 4	3.51	−0.020 0.050	0.055 (0.050)	−0.021 0.052	0.056 (0.052)	−0.016 0.051	0.054 (0.525)
Race 5	7.66	0.071 0.244	0.254 (0.285)	−0.10 0.237	0.256 (0.287)	0.213 0.234	0.316 (0.284)
Race 6	2.62	0.002 0.125	0.125 (0.139)	−0.016 0.127	0.128 (0.140)	0.002 0.131	0.131 (0.139)
Race 7	0.10	0.011 0.063	0.064 (0.063)	0.013 0.064	0.064 (0.062)	0.011 0.063	0.064 (0.062)
Region 1	3.37	0.002 0.069	0.069 (0.070)	−0.034 0.065	0.073 (0.071)	−0.011 0.070	0.071 (0.077)
Region 2	2.84	0.006 0.079	0.079 (0.074)	−0.043 0.074	0.086 (0.075)	0.011 0.082	0.082 (0.084)
Region 3	4.02	−0.010 0.065	0.066 (0.064)	0.012 0.067	0.068 (0.062)	−0.014 0.066	0.068 (0.065)
Region 4	3.18	0.005 0.083	0.083 (0.082)	−0.012 0.074	0.075 (0.076)	−0.006 0.083	0.084 (0.081)
Tenure 1	1.91	−0.011 0.047	0.048 (0.047)	−0.024 0.048	0.054 (0.047)	−0.012 0.049	0.050 (0.048)
Tenure 2	3.68	0.008 0.061	0.062 (0.063)	0.008 0.064	0.065 (0.064)	0.007 0.063	0.064 (0.062)
Sex 1	3.46	−0.006 0.052	0.053 (0.052)	−0.022 0.047	0.052 (0.050)	−0.006 0.053	0.053 (0.053)
Sex 2	3.38	0.001 0.056	0.056 (0.054)	−0.009 0.056	0.056 (0.050)	−0.006 0.056	0.056 (0.054)

Table 4 reports the simulated undercount estimates for the same population aggregates as those shown in Table 1 for the real census analysis. The true undercounts of the simulation are reported under the U-column. It is observed that all three approaches produced comparable undercount estimates. This is similar to the results seen in Table 1. A closer examination reveals that the local post-stratification consistently had smaller bias than the post-stratification, and outperformed the logistic regression with smaller root mean squared errors (RMSE). The quality of the jackknife variance estimation is demonstrated by comparing the average jackknife standard deviation estimates ($\hat{SD}$) with the average standard deviation from the simulations (SD). We see the two quantities agreed with each other at the various levels of population aggregates. This confirms the jackknife variance estimation procedure.

For the three estimation approaches, Figure 7 shows the average undercount estimates and their average RMSE for each age cohort. These figures are counterparts of Figure 5 in the census data analysis. Both figures display substantial differences among the three estimation methods. The local post-stratification reflected the pattern of the true undercounts quite well, whereas the logistic regression captured the overall trend, especially at the left end, but failed to take into account the fluctuations around ages 20 and 50. The post-stratification preformed quite poorly in the comparisons. Judging by the RMSE in Figure 7, the local post-stratification had the most stable performance across all ages.

We also collected information on the estimated undercounts at the state level, whose details are available in a technical report (Chen, Tang, and Mule 2009). Among the 50 states and DC, the local post-stratification had the smallest RMSE on 33 of them. In the South region, which had about about $1/3$ of the U.S. population and had the most undercounts among the four regions, the local post-stratification performed the best in most of the states. If we define the cumulated root mean squared error (CRMSE) to be the raw summation of the RMSE of the estimated undercounts over the states, then the CRMSE of the local post-stratification, post-stratification, and logistic regression

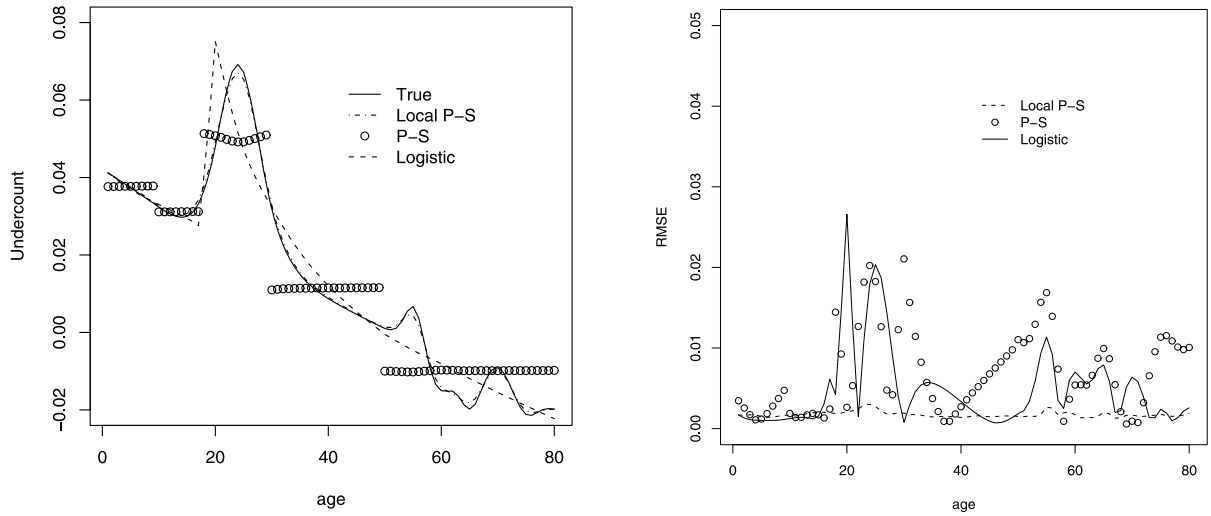

Figure 7. Simulation comparison of estimated percentage undercount for age by the local post-stratification (dashed line), post-stratification (circle points), and logistic regression (dotted line) with the true undercounts (solid line). Left panel: comparison of undercount estimates. Right panel: root mean square error (RMSE) comparison.

were 9.122, 9.904, and 10.195, respectively. This shows the robust and promising performance of the proposed approach.

10. DISCUSSION

In this paper, we propose a local post-stratification approach for the census dual system estimation based on the nonparametric kernel estimation for the enumeration and correct enumeration functions. The local post-stratification can capture the underlying data characteristics objectively and is free of model misspecification. Compared with the post-stratification, it avoids construction of post-strata explicitly and accounts for the correlation bias in the estimation automatically. An major advantage of our proposal is the ability to smooth the discrete variables, an aspect which is relevant to the census. This leads to more efficient estimation of the enumeration and correct enumeration function as compared to single-cell based estimation (Chen and Tang 2008a, 2008b).

The A.C.E. provides fresh and challenging research issues for capture–recapture surveys. The erroneous enumerations and missing values treated in this paper are only some of the new aspects. The proposed local stratification, although having been described in close connection to the census, is applicable for other capture–recapture surveys after some modifications. The inclusion of extra covariates $\mathbf{Z}^e$ and $\mathbf{Z}^p$ in the missing propensities for the E and P samples was done to make the MAR assumption more realistic.

The issue of extra covariate $\mathbf{S}$ in the correct enumeration function $e(\mathbf{x}, \mathbf{s})$ introduces a new issue regarding how to treat variables that are only observed in the A.C.E. but not available for the general census. These variables are related to both the enumeration and correct enumeration functions. We have assumed that $\mathbf{S}$ does not contribute to the enumeration function p. If $\mathbf{S}$ affects both e and p functions, then the population size estimator (5.6) may not be consistent, even if we define a marginalized version $p(\mathbf{x})$ like $e(\mathbf{x})$ by integrating out the $\mathbf{S}$ variable by conditioning. We may have to use

$$\hat{N} = \sum_{i \in \mathcal{E}} \frac{\hat{e}(\mathbf{X}_i, \mathbf{S}_i)}{\pi_i \hat{p}(\mathbf{X}_i, \mathbf{S}_i)}, \tag{10.1}$$

where π_i is the known survey weight for $i \in \mathcal{E}$. Although this estimator is consistent, the estimates will show a large sampling variance as it excludes data collected outside the A.C.E. The estimator (10.1) is similar to the N2 estimator in Griffin (2005). He was also concerned about the variability and proposed a ratio adjustment.

The issue of data-defined persons is critically important for the dual system estimation. Our estimates for larger population aggregates, as reported in the first part of the Section 8, were obtained by substituting $\mathcal{C}$ with $\mathcal{D}$, the set of data-defined enumerations in the estimation of $e(\mathbf{x})$ and $p(\mathbf{x})$. The same replacement was made for the population size estimates in (5.6), (7.2), and (7.4). As discussed in Brown and Zhao (2008), different assumptions regarding the heterogeneity of an individual being data-defined can produce quite different population estimates. This is confirmed by our state estimates reported in Table 2. As there is most likely differential heterogeneity between data-defined and non-data-defined cases, the data-defined rate function needs to be modeled (Griffin 2005). The Census Bureau is considering this for the 2010 Census Coverage Measurement Study.

[Received July 2008. Revised May 2009.]

REFERENCES

Aitchison, J., and Aitken, C. (1976), "Multivariate Binary Discrimination by the Kernel Method," *Biometrika*, 63, 413–420. [108]

Alho, J. M. (1990), "Logistic Regression in Capture–Recapture Models," *Biometrics*, 46, 623–635. [107]

Alho, J. M., Mury, M. H., 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. [107]

Ayhan, Ö. H., and Ekni, S. (2003), "Coverage Error in Population Censuses: The Case of Turkey," *Survey Methodology*, 29, 155–165. [105]

Belin, T. R., Diffendal, G. J., Mack, S., Rubin, D. B., Schafer, J. L., and Zaslavsky, A. (1993), "Hierarchical Logistic Regression Models for Imputation of Unresolved Enumeration Status in Undercount Estimation" (with discussion), *Journal of the American Statistical Association*, 88, 1149–1166. [107,108]

Bell, W. R. (1993), "Using Information From Demographic Analysis in Post-Enumeration Survey Estimation," *Journal of the American Statistical Association*, 88, 1106–1118. [106]

——— (1999), "Accuracy and Coverage Evaluation Survey: Ratio Adjusting Logistic Regression DSEs (Target Model) Using 1990 Census Counts," in *DSSD Census 2000 Procedures and Operations Memorandum Series*, Vol. Q-11, Washington, DC: U.S. Census Bureau. [105]

Brown, L., and Zhao, Z. (2008), "Alternative Formulas for Synthetic Dual System Estimation in 2000 Census," *IMS Collections. Probability and Statistics: Essays in Honor of David A. Freedman*, 2, 90–113. [115,118]

Cantwell, P., and Ikeda, M. (2003), "Handling Missing Data in the 2000 Accuracy and Coverage Evaluation Survey," *Survey Methodology*, 29, 139–153. [111]

Cantwell, P., McGrath, D., Nguyen, N., and Zelenak, M. F. (2001), "Accuracy and Coverage Evaluation: Missing Data Results," in *DSSD Census 2000 Procedures and Operations Memorandum Series*, Vol. B-7. [108]

Census Customer Service (2002), *Census Coverage Survey: Evaluation Report*, London, U.K.: Office for National Statistics. [105]

Chao, A., and Tsay, P. K. (1998), "A Sample Coverage Approach to Multiple-System Estimation With Application to Census Undercounts," *Journal of the American Statistical Association*, 93, 283–293. [106]

Chen, S. X., and Lloyd, C. J. (2000), "A Non-Parametric Approach to the Analysis of Two Stage Mark-Recapture Experiments," *Biometrika*, 87, 633–649. [105]

——— (2002), "Estimation of Population Size Based on Biased Samples Using Nonparametric Binary Regression," *Statistica Sinica*, 12, 505–518. [107]

Chen, S. X., and Tang, C. Y. (2008a), "Nonparametric Regression With Discrete Covariates and Missing Values," technical report, Iowa State University, Dept. of Statistics. [108,118]

——— (2008b), "Nonparametric Estimation of Population Size From Dual System Enumeration Surveys," technical report, Iowa State University, Dept. of Statistics. [109,118]

Chen, S. X., Tang, C. Y., and Mule, V. T. (2009), "Local Post-Stratification in Dual System Accuracy and Coverage Evaluation for US Census," technical report, Iowa State University. [109,117]

Darroch, J., Fienberg, S. E., Glonek, G. F., and Junker, B. W. (1993), "A Three-Sample Multiple-Recapture Approach to Census Population Estimation With Heterogeneous Catchability," *Journal of the American Statistical Association*, 88, 1137–1148. [106]

Dunstan, K., Heyen, G., and Paice, J. (2001), "Measuring Census Undercount in Australia and New Zealand," Demography Working Paper 99/4, Australian Bureau of Statistics. [105]

Elliott, M. R., and Little, R. J. A. (2000), "A Bayesian Approach to Combining Information From a Census, a Coverage Measurement Survey, and Demographic Analysis," *Journal of the American Statistical Association*, 95, 351–362. [106]

——— (2005), "A Bayesian Approach to 2000 Census Evaluation Using ACE Survey Data and Demographic Analysis," *Journal of the American Statistical Association*, 100, 380–388. [106]

Fan, J., and Gijbels, I. (1996), *Local Polynomial Modeling and Its Applications*, London: Chapman & Hall. [107]

Griffin, R. (2005), "Net Error Estimation for the 2010 Census," in *2010 Census Coverage Measurement Memorandum*, Vol. 2010-E-01. [109,110,118]

Haberman, S., Jiang, W., and Spencer, B. (1998), "Activity 7: Develop Methodology for Evaluating Model-Based Estimates of the Population Size for States. Final Reports," technical report, U.S. Census Bureau. [105,107,109]

Haines, D. (2001), "Accuracy and Coverage Evaluation Survey: Computer Specifications for Person Dual System Estimation (U.S.)—Re-Issue of Q-37," in *DSSD Census 2000 Procedures and Operations Memorandum Series*, Vol. Q-48. [110]

Hall, P. (1981), "On Nonparametric Multivariate Binary Discrimination," *Biometrika*, 68, 287–294. [108]

Hall, P., Racine, J., and Li, Q. (2004), "Cross-Validation and the Estimation of Conditional Probability Densities," *Journal of the American Statistical Association*, 99, 1015–1026. [108]

Härdle, W. (1990), *Applied Nonparametric Regression*, Cambridge: Cambridge University Press. [107,108]

Hogan, H. (1992), "The 1990 Post-Enumeration Survey: An Overview," *The American Statistician*, 46, 261–269. [105]

——— (1993), "The 1990 Post-Enumeration Survey: Operations and Results," *Journal of the American Statistical Association*, 88, 1047–1060. [105,107]

——— (2000a), "Accuracy and Coverage Evaluation 2000: Decomposition of Dual System Estimate Components," in *DSSD Census 2000 Procedures and Operation Memorandum Series*, Vol. B-8. [105]

——— (2000b), "Accuracy and Coverage Evaluation 2000: Dual System Estimate Results," in *DSSD Census 2000 Procedures and Operation Memorandum Series*, Vol. B-9. [105]

——— (2003), "The Accuracy and Coverage Evaluation: Theory and Design," *Survey Methodology*, 29, 129–138. [109,115]

Huggins, R. M. (1989), "On the Statistical Analysis of Capture Experiments," *Biometrika*, 76, 133–140. [107]

Kostanich, D. (2003), "A.C.E. Revision II: Design and Methodology," in *U.S. Census Bureau DSSD A.C.E. Revision II Memorandum Series*, Vol. PP-30. [110]

Malec, D., Sedransky, J., Moriarity, C., and LeClere, F. (1997), "Small Area Inference for Binary Variables in the National Health Interview Survey," *Journal of the American Statistical Association*, 92, 815–826. [110]

Mule, T., Schellhamer, T., Malec, D., and Maples, J. (2007), "Using Continuous Variables as Modeling Covariates for Net Coverage Estimation," in *DSSD 2010 Census Coverage Measurement Memorandum Series*, Vol. 2010-E-09-R1. [110]

Nandram, B., and Choi, J. (2005), "Hierarchcial Bayesian Nonignorable Regression Models for Small Areas: An Application to the NHANES Data," *Survey Methodology*, 31, 73–84. [110]

Pollock, K. H. (1976), "Building Models of Capture–Recapture Experiments," *The Statistician*, 25, 253–260. [107]

——— (1991), "Modeling Capture–Recapture, and Removal Statistics for Estimation of Demographic Parameters for Fish and Wildlife Populations: Past, Present, and Future," *Journal of the American Statistical Association*, 86, 225–238. [107]

Racine, J. S., and Li, Q. (2004), "Nonparametric Estimation of Regression Functions With Both Categorical and Continuous Data," *Journal of Econometrics*, 119, 99–130. [108]

Rhind, D. (2003), *The 2001 Census in Westminister*, London, U.K.: Statistics Commission. [105]

Rosenbaum, P. R., and Rubin, D. B. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55. [108]

Schindler, E. (2008), "Post-Stratification by Age for Small Intervals," in *Census 2000 Procedures and Operations Memorandum Series*, Vol. Q-94. [109]

Sekar, C. C., and Deming, W. E. (1949), "On a Method of Estimating Birth and Death Rates and the Extent of Registration," *Journal of American Statistical Association*, 40, 101–115. [105]

Shao, J., and Tu, D. (1995), *The Jackknife and Bootstrap*, New York: Springer-Verlag. [113]

Shao, J., and Wu, C. F. J. (1989), "A General Theory for Jackknife Variance Estimation," *The Annals of Statistics*, 17, 1176–1197. [113]

Signorini, D. F., and Jones, M. C. (2004), "Kernel Estimators for Univariate Binary Regression," *Journal of the American Statistical Association*, 99, 119–126. [108]

Simonnoff, J. S. (1995), "Smoothing Categorical Data," *Journal of Statistical Planning and Inference*, 47, 41–69. [107]

Smith, P. (1979), "Splines as a Useful and Convenient Statistical Tool," *The American Statistician*, 33, 57–62. [110]

U.S. Census Bureau (2004), *Accuracy and Coverage Evaluation of Census 2000: Design and Methodology*, Washington, DC: U.S. Census Bureau. [107,110,111]

Wolter, K. M. (1986), "Some Coverage Error Models for Census Data," *Journal of the American Statistical Association*, 81, 338–346. [105]

——— (1990), "Capture–Recapture Estimation in the Presence of a Known Sex Ratio," *Biometrics*, 46, 157–162. [106]

——— (2007), *Introduction to Variance Estimation*, New York: Springer-Verlag. [113]

Zaslavsky, A. M., and Wolfgang, G. S. (1993), "Triple-System Modeling of Census, Post-Enumeration Survey, and Administrative-List Data," *Journal of Business & Economic Statistics*, 11, 279–288. [106]