Journal of Official Statistics, Vol. 19, No. 3, 2003, pp. 237–252

A Post-stratified Raking-ratio Estimator Linking National and State Survey Data for Estimating Drug Use

Trent D. Buskirk[1] *and Jane L. Meza*[2]

Estimation of statewide and county-specific drug use can be improved by combining national and state survey data. Obtaining county estimates of drug use from state-level data is difficult due to the rarity of drug use and the small sample sizes within counties. In this article, we propose using post-stratified raking-ratio estimators to link state data with national surveys stratified by state. We also propose a raking estimator that links state census data to national survey data using auxiliary variables associated with drug use. Specifically, the adjusted raking estimator produces estimates of drug use by region and age-class to be used in a small-area model that will produce empirical Bayes estimates of county drug use. The methods are presented using data from the Nebraska Adult Household Survey and the National Household Survey on Drug abuse. Two choices of small-area models are discussed.

Key words: Post-stratified estimator; sampling weights; raking algorithm; small-area models; empirical Bayes estimates; borrow strength.

1. Introduction

Tracking drug use in the U.S. has become an important social and economic issue. Drug addiction programs and legal agencies benefit from accurate estimates of the prevalence of drug use. State health agencies requesting federal funds for drug rehabilitation and prevention programs are better served with accurate estimates of county-level drug use. Using this information, states may more appropriately allocate funds to county health agencies for the development of these programs.

Many states conduct studies to estimate the need for substance abuse treatment services (Bray, Kroutil, and Wheeless 1999). However, relatively few studies investigate drug use for small areas such as counties: ‘‘these smaller areas are responsible for many of the policy decisions regarding substance abuse services and as such, require sound data upon which policy makers can base their decisions’’ (Marsden and Bray 1999, p. 56). Estimating drug use for rural sectors of a state may be difficult as there may be no reported cases of drug use among the sampled respondents. By linking state and national data, one may be able to obtain more accurate estimates of state-wide drug use.

Wright et al. (1997) estimated the prevalence of hard-core drug use based on a post-stratified ratio estimator. The post-stratum population sizes were obtained from marginal population counts along with an estimate of one post-stratum population size from an

[1] Graduate Program in Public Health and Biostatistics and Epidemiology Core, Eastern Virginia Medical School, Norfolk, VA 23501-1980, U.S.A. Email: buskirtd@evms.edu
[2] Department of Preventive and Societal Medicine, University of Nebraska Medical Center, Omaha, NE 68198-4350 U.S.A. Email: jmeza@unmc.edu

independent national survey. Such a procedure requires knowledge of at least one cell population size and the marginal population sizes. In this article we propose applying a raking ratio estimator to state survey data that has been post-stratified using auxiliary variables associated with drug use. By using a raking algorithm to calibrate cell population size estimates to known marginal totals, we eliminate the need for the additional estimate of one of the post-stratum population sizes. This is advantageous for estimating state drug use since additional independent state surveys may not be available.

Chattopadhyay et al. (1999) used an empirical Bayes method for estimating drug prevalences for small areas based on data stratified by planning region. To improve these estimates we propose a design-based method that links state data to national data using auxiliary variables associated with drug use. We then propose empirical Bayes estimates of county prevalence based on the new estimates.

1.1. Nebraska adult household survey

The Nebraska Adult Household Survey (NAHS) is funded by the Center for Substance Abuse Treatment and conducted periodically by public and private agencies contracted by Nebraska Health and Human Services, Division of Alcoholism, Drug Abuse, and Addiction Services. The respondent universe for the 1995 NAHS included adult Nebraska residents (ages 19 or older) living in a household with a telephone. Residents without a land telephone or who were homeless, institutionalized or who were students living in dormitories were excluded.

The 1995 NAHSs was a multistage stratified random digit dialing telephone survey with 4,324 completed interviews. Hispanics and African Americans were oversampled to provide adequate precision of estimates for these minority groups. Respondents were questioned about use of drugs, treatment and past arrest history. The overall response rate for the 1995 NAHS was 75 per cent.

1.2. National household survey on drug abuse

The National Household Survey on Drug Abuse (NHSDA), administered by The Substance Abuse and Mental Health Services Administration, is a primary source of information about drug use in the U.S. The 1995 NHSDA employed a multistage area probability sample of 17,747 persons with oversampling of Hispanics, African Americans and cigarette smokers to ensure specified precision constraints within these populations.

The NHSDA collects survey data via personal interviews and solicits information about drug use. To increase the honesty of responses questions are read by an interviewer and responses are placed in a sealed envelope to which the interviewer has no direct access. The overall response rate was 80.4 per cent.

2. Raking Ratio Estimator

2.1. Background and motivation

Raking estimators are typically used to adjust sampling weights so that their sums match known marginal population totals (see Deming and Stephan 1940 or Brackstone and Rao

1979 for a discussion of raking estimators). In the context of survey data, Deville et al. (1993) considered several distance functions for evaluating the calibration equations used in raking cell totals to known marginal totals. Here we use the classical raking algorithm proposed by Deming and Stephan (1940) sometimes referred to as iterative proportional fitting. Of those functions considered by Deville et al. (1993), the classical raking algorithm results in calibration equations that produce positive calibrated weights. However, the raking algorithm may not converge if one or more of the cell estimates are zero (Lohr 1999).

The classical raking procedure usually requires two or more variables for which a two-way (or multi-way) cross-classification table is formed and assumes marginal totals are known. To employ the raking algorithm, we post-stratify the sample according to the cells in the classification table. The resulting estimates for the population size within each cell may be more precise than naïve cell estimates from the original survey weights since these new estimates are based on known marginal population totals (Lohr 1999). The raked estimate of the population total will have the same variance (asymptotically) as the traditional pooled ratio estimator provided the interaction effects are negligible (Deville and Särndal 1992). Unlike the traditional post-stratified ratio estimator, the raking estimator does not require prior knowledge of the joint distribution of all cell population sizes, but only requires the marginal distribution of the population sizes for each of the variables used in forming the classification table. For a numeric example of the raking algorithm, see Lohr (1999).

Wright et al. (1997) used the traditional post-stratified estimator to estimate the number of heroin users in the U.S. based on a post-stratification using two auxiliary variables, ARREST and TREATMENT, selected because of their association with drug use. Wright et al. (1997) used known marginal counts (or estimated marginal counts assumed to have negligible error) for the number arrested and number treated along with the overall population size to generate the ''true'' marginal population sizes for each variable in the two-way classification table. Since knowledge of these marginal totals creates one degree of freedom in the cross-classification table, Wright used data on the percentage of those in treatment who were ever arrested from the 1990 Drug Services Research Survey (DSRS) and arrest and treatment data from the 1992 NHSDA to estimate the number of people who were arrested and treated in 1992. This estimate along with the marginal totals created a complete table of population counts for the number of people who were arrested and treated, arrested and not treated, treated and not arrested, and neither arrested nor treated. Assuming negligible error for the cell population estimates and by applying the estimated proportion of heroin users within each of the cells to the population counts described above, the estimator of Wright et al. (1997) can be considered a standard post-stratified estimator.

We propose a raking ratio estimator similar to Wright et al. (1997); however, we replace the ''known'' cell population sizes with those obtained by raking naïve estimates of the cell population sizes from the state data to known marginal totals. No knowledge about the population sizes for any of the cells within the classification table is assumed and thus no additional information from the 1990 DSRS is required. This method has an advantage over the traditional post-stratified estimator in that it produces an estimate with approximately the same variance (Deville and Särndal 1992) (provided the interaction effect of treatment and arrest is negligible) and uses only cell estimates and marginal totals (i.e., population totals for all cells in the cross-classification table are not necessary). This method may be more desirable for estimating drug use within states as independent state surveys for estimating the population count within any given cell in the table may not be available.

2.2. *Raking ratio estimate of heroin use in the U.S.*

We now illustrate the use of a raking ratio estimator using the 1992 public-use NHSDA data file to estimate the number of heroin users within the U.S. in 1992. Throughout this illustration we will make use of the following notation:

- $\hat{N}^R_{11}$ = estimated number of people who were arrested and treated.
- $\hat{N}^R_{12}$ = estimated number of people who were arrested but not treated.
- $\hat{N}^R_{21}$ = estimated number of people who were not arrested but treated.
- $\hat{N}^R_{22}$ = estimated number of people who were neither arrested nor treated.
- $N_{.1}$ = number of people treated.
- $N_{.2}$ = number of people who were not treated.
- $N_{1.}$ = number of people who were arrested.
- $N_{2.}$ = number of people who were not arrested.

Using the values provided in Wright et al. (1997): $N_{.1} = 1{,}789{,}000$, $N_{.2} = 203{,}924{,}000$, $N_{1.} = 9{,}722{,}671$ and $N_{2.} = 195{,}990{,}329$, we rake the estimated number of persons within each combination of the levels of TREATMENT and ARREST obtained from the 1992 public-use NHSDA data file. Table 1 displays the resulting cell population estimates.

Estimates of the proportion of heroin users for each cell of the classification table, $\hat{p}_{ij}$ for $i = 1, 2; j = 1, 2$, were available from the 1992 public-use NHSDA data. Using these estimated proportions along with the estimated cell population sizes from the raking algorithm ($\hat{N}^R_{ij}$) (Table 1), the estimated number of heroin users $\hat{D}(A, T)$ within the U.S. in 1992 is:

$$\hat{D}(A, T) = \sum_{i=1}^{2} \sum_{j=1}^{2} \hat{p}_{ij} \hat{N}^R_{ij} = 675{,}974$$

Applying the post-stratified estimator proposed in Wright et al. (1997) to the 1992 public-use NHSDA data, we estimated that there were 623,469 heroin users in the U.S. in 1992. This raking estimate is 52,505 higher than the estimate obtained using the post-stratified estimator proposed by Wright et al. (1997), primarily due to the estimated number of persons arrested *and* treated using the raking algorithm. The raking estimate for this cell is 227,102 persons more than the 1990 DSRS used by Wright et al. (1997). This cell also has the highest estimated proportion of heroin users. The raking ratio estimate may be compensating for underreporting and/or undercoverage of the naïve estimate

Table 1. Raked treatment and arrest population estimates in the U.S. in 1992

	Treated	Not treated	Total
Arrested	943,417	8,779,254	9,722,671
Not arrested	845,583	195,144,746	195,990,329
Total	1,789,000	203,924,000	205,713,000

Table 2. Raked treatment and arrest population estimates for Nebraska in 1995

	Treated	Not treated	Total
Arrested	1,046.102	25,745.898	26,792
Not arrested	4,336.898	1,139,303.102	1,143,640
Total	5,383	1,165,049	1,170,432

obtained from the 1990 DSRS. Both estimates produce an overall U.S. heroin drug use rate of about .3 per cent.

2.3. *Raking ratio estimate of lifetime drug use in Nebraska*

As a second illustration, the 1995 NAHS along with statewide marginal totals from national surveys stratified by state are used to estimate the number of lifetime drug users in the state of Nebraska in 1995. A respondent is classified as a lifetime drug user if he or she has used an illicit drug (marijuana, cocaine, hallucinogens, opiates, amphetamines, or any other drug) at some point in his or her life.

From the Uniform Crime Report (U.S. Department of Justice 1995) there were a total of 58,914 arrests in Nebraska in 1995 for persons age 19 years or older (not accounting for the number of repeat offenders). To allow for multiple arrests per person, the 1995 public-use NHSDA data were used to estimate an average of 2.199 arrests per person and the number of arrests of persons age 19 or older in Nebraska was modified to 26,792. An estimated 5,383 Nebraskans age 19 or older were enrolled in some form of drug treatment during 1995[3] using the Uniform Facility Data Set.

According to the 1995 update to the 1990 decennial census, there were 1,170,432 adults age 19 or older living in Nebraska. Using this figure along with the number of arrests and number treated we have: $N_{1.} = 26{,}792$; $N_{2.} = 1{,}170{,}432 - N_{1.} = 1{,}143{,}640$; $N_{.1} = 5{,}383$, $N_{.2} = 1{,}170{,}432 - N_{.1} = 1{,}165{,}049$. These marginal totals along with the original weighted cell population totals for the cells in our classification table were used to produce the calibrated cell population estimates based on the classical raking algorithm ($\hat{N}^R_{ij}$) displayed in Table 2.

Horvitz-Thompson (1952) estimates of the proportion of lifetime drug users within each cell of the table ($\hat{p}_{ij}$ for $i = 1, 2; j = 1, 2$) were obtained from the NAHS. Using these estimated proportions along with the estimated cell population sizes ($\hat{N}^R_{ij}$) (Table 2), the post-stratified raking ratio estimate of the number of lifetime drug users $\hat{D}(A, T)$ in Nebraska in 1995 is:

$$\hat{D}(A, T) = \sum_{i=1}^{2} \sum_{j=1}^{2} \hat{p}_{ij} \hat{N}^R_{ij} = 339{,}914$$

This estimate results in an overall Nebraska lifetime drug use rate of about 30 per cent.

[3] This may overestimate the number of patients admitted for drug treatment in Nebraska during 1995 since information on the number of episodes per individual in treatment was not available.

3. Adjusted Raking Estimator

3.1. Background and motivation

Sampling for many state surveys is executed at the state or regional planning levels without regard to estimation at the county level. Traditional direct estimates of drug rates for some counties may be unreliable due to the small sample size in the county. If none of the respondents in the county are drug users, the direct estimates may be misleading since the direct estimate and its estimated standard error are both zero. To improve the direct estimates we borrow strength from the estimated drug rates for post-strata formed using related auxiliary variables. In this section, we propose an adjusted raking estimator of the proportion of lifetime drug users for post-strata to be used as parameter estimates in the small-area model in Section 5.

Associations between drug use, age, and population density have been documented in the literature (see for example Maxwell 2000). Hard-core drug use tends to be more prevalent among members of younger age-classes across regions (i.e., metropolitan, small metropolitan, and rural) and hard-core drug use rates tend to be higher in larger metropolitan regions across age-classes. Therefore we expect that these variables may be good predictors (with possibly additive effects) for estimating drug use. However, lifetime drug-use rates reflect a wider spectrum of time of use, so the effects of age-class and region may have a nonnegligible interaction when used to predict lifetime drug use. In the case of nonnegligible interaction, the complete classification table based on the combinations of all levels of region and age-class may improve the precision of our estimators.

In the methodology that follows, we treat REGION as a categorical variable with three levels (Large Metropolitan Region, Small Metropolitan Region, and Rural Region) and AGECLASS as a categorical variable with three levels (19 to 25, 26 to 34, and 35 or older). In Nebraska three counties were identified as large metropolitan regions, 14 as small regions and 76 as rural regions.

3.2. Adjusted raking estimator of Nebraska drug use by region and age-class

Treating county as a small area embedded within region and post-stratified by age-class, we propose a method for obtaining potentially more efficient post-stratified estimates of the proportion of lifetime drug users within each region by age-class. This method borrows information from the 1995 public-use NHSDA data, also post-stratified by region and age-class.

Assuming the joint distribution of the cell population sizes for the nine cells of the REGION and AGECLASS cross-classification table is known, we combine these proportions with the NHSDA direct estimates of the proportion of lifetime drug users within each of the cells to obtain the proportion of drug users within the levels of AGECLASS and REGION. Since we want consistent estimators of lifetime drug use and have reliable information about the population distribution of the state[4], we only borrow information pertaining to drug use from the larger U.S. population. Multiplying these marginal proportions by the number of adult citizens (age 19 or older) in Nebraska in 1995 (1,170,432) gives the

[4] If only the marginal totals are available for these two variables, we modify our method using the raking ratio estimator proposed earlier.

expected number of drug users within each of the age-classes and regions. Using these marginal totals, we then apply an iterative proportional fitting algorithm to the weighted cell estimates (e.g., Horvitz-Thompson (1952) estimates) of the total number of lifetime drug users within each of the nine post-strata from the 1995 NAHS. Combining these raked estimates with the known population distribution of Nebraska, we estimate the proportion of lifetime drug use within each combination of REGION and AGECLASS. We call this estimator the Adjusted Raking Estimator (ARE). To illustrate this method formally, we define the following quantities:

- d_{ij} = proportion of Nebraskans in age-class i and region j ($i, j = 1, 2, 3$) in 1995.
- $\hat{p}_{ij}$ = proportion of lifetime drug users in age-class i and region j ($i, j = 1, 2, 3$) estimated using the 1995 NHSDA.[5]
- The proportion of drug users in each age class ($i = 1, 2, 3$) in Nebraska is estimated by $\hat{p}^*_{i\cdot} = \sum_{j=1}^{3} d_{ij}\,\hat{p}_{ij}$ and similarly, for each region ($j = 1, 2, 3$) by $\hat{p}^*_{\cdot j} = \sum_{i=1}^{3} d_{ij}\,\hat{p}_{ij}$.
- Because there were 1,170,432 Nebraskans age 19 or older in 1995, $(\hat{p}^*_{1\cdot}, \hat{p}^*_{2\cdot}, \hat{p}^*_{3\cdot}) \times 1{,}170{,}432$ is the marginal distribution for the number of drug users within each level of AGECLASS and similarly $(\hat{p}^*_{\cdot 1}, \hat{p}^*_{\cdot 2}, \hat{p}^*_{\cdot 3}) \times 1{,}170{,}432$ is the marginal distribution for the number of drug users within each level of REGION.

These marginal distributions were used to apply an iterative proportional fitting algorithm to the naïve estimates of the number of drug users within each combination of AGECLASS and REGION to obtain:

- $\hat{N}^R_{ij}$ = estimated number of Nebraska drug users in age-class i and region j using the classical raking algorithm.
- $\hat{p}^{NE}_{ij} = (1{,}170{,}432)^{-1} \times \hat{N}^R_{ij}/d_{ij}$ = ARE of the proportion of lifetime drug users within age-class i and region j in Nebraska.

The 1995 U.S. Census for Nebraska was used to estimate the population proportions (d_{ij}, $i, j = 1, 2, 3$). From the 1995 NHSDA data we obtained the Horvitz-Thompson (1952) estimates of the proportion of lifetime drug users within each level of AGECLASS and REGION ($\hat{p}_{ij}$, $i, j = 1, 2, 3$). Combining these two estimates, we estimate the proportion of lifetime drug users between 19 and 25 years old within Nebraska to be $\hat{p}^*_{1\cdot} = \sum_{j=1}^{3} \hat{p}_{1j}\,d_{1j} = .06295151$, which implies there were an estimated 73,680 Nebraskan lifetime drug users between the ages of 19 and 25.

Performing similar computations yields the following vectors for the estimated number of drug users within Nebraska for the variables AGECLASS and REGION, respectively: (73,680, 109,705, 214,330) and (185,186, 123,663, 88,866). Using these marginal totals we raked the weighted cell estimates of the total number of lifetime drug users within each of the levels of AGECLASS and REGION to obtain the cell estimates given in Table 3.

Finally, using these estimates along with the population distribution for Nebraska, we compute the ARE of the proportion of lifetime drug users within each combination of AGECLASS and REGION ($\hat{p}^{NE}_{ij}$, $i, j = 1, 2, 3$) as shown in Table 4.

These ARE proportions borrow strength from the estimated distribution of lifetime drug users within the U.S. from the 1995 NHSDA public-use data file while constraining the age distribution for the U.S. to that of Nebraska (across levels of REGION and AGECLASS

[5] These estimates were weighted estimates computed using survey weights (ANALWT) and the SUMFLAG variable.

Table 3. Raked estimates of total lifetime drug users in Nebraska by region and age-class

	Lg. Metro	Sm. Metro	Rural	Total
19–25	42,831	21,764	9,085	73,680
26–34	49,398	37,589	22,718	109,705
35 +	92,957	64,310	57,063	214,330
Total	185,186	123,663	88,866	397,715

Table 4. ARE proportion of lifetime drug users in Nebraska by region and age-class

	Lg. Metro	Sm. Metro	Rural
19–25	.52398	.46328	.31046
26–34	.51758	.59299	.48921
35 +	.29342	.26160	.23335

using the 1995 NHSDA public-use data file produced significant main effects and a significant interaction term (all p-values $< .0001$). The Hosmer and Lemeshow Goodness-of-Fit test was not significant (p-value ≈ 1) indicating that these two variables and the interaction term are strong predictors of lifetime drug-use. Since the ARE estimates contain one component from the state and another from the nation and since these two components are linked via the cross-classification (i.e., the interaction term is significant) of two variables that are strong predictors for lifetime drug use, we expect that these adjusted raking estimates will be more stable and more precise than the direct estimates.

If this assumption holds across all regions within each age-class, the raked cell estimates of the proportion of lifetime drug users should be more stable and precise than the direct estimates.

4. Small-Area Estimates

4.1. Background and motivation

The term small-area refers to a geographic area or demographic group in which a limited number of observations are available. Here, small sample sizes for the counties (small-areas) are the result of a sample design which provides adequate sample size for the state but not for individual counties. Information from the state survey is not adequate to produce reliable estimates of lifetime drug use at the county level, so small-area methods borrow information from the region, entire state, and/or auxiliary data. We propose to borrow strength from the adjusted raking estimate for the region (decomposed by age-class).

The small-area model must be appropriate for estimating a rare event like drug use. Folsom and Liu (1994) and Folsom et al. (1999) discuss state-level small-area estimation in the NHSDA. Chattopadahyay et al. (1999) considered small-area estimation for a rare event and proposed a composite estimate that takes into account the survey weights and combined information for counties within the same region. Here, we propose an estimate that is a composite of the direct estimate of the county drug rate for a particular age group and the adjusted raking estimate of the corresponding regional and age-class rate, rather than using the regional/age-class direct estimate as in Chattopadahyay et al. (1999). The regions used here are based on the REGION variable defined earlier.

4.2. *Composite estimate*

Since the sample size in many counties is small, some age-classes may not be represented by sample respondents. Let S_{jk} be the set of age-classes in the kth county ($k = 1, \ldots, 93$) of region j ($j = 1, 2, 3$) that are represented by respondents in the sample. If age-class $i \in S_{jk}$ the lifetime drug rate for county k is estimated by the direct estimate:

$$\hat{p}_{ijk}^{NAHS} = \frac{\sum_{l=1}^{n_{ijk}} w_{ijkl} y_{ijkl}}{\sum_{l=1}^{n_{ijk}} w_{ijkl}} \tag{1}$$

where y_{ijkl} is an indicator of lifetime drug use for individual l ($l = 1, \ldots, n_{ijk}$), w_{ijkl} is the survey weight for individual l and n_{ijk} is the sample size for age-class i ($i = 1, 2, 3$) in county k ($k = 1, \ldots, 93$) of region j ($j = 1, 2, 3$). If $i \notin S_{jk}$, the lifetime drug rate for county k is estimated by the ARE estimate ($\hat{p}_{ij}^{NE}$) for the region. Each estimated rate is weighted by a_{ijk}, the proportion of people in age-class i in county k of region j. The composite estimate for county k is found by summing across the age classes and is given by:

$$\hat{p}_{jk}^{C} = \sum_{i \in S_{jk}} a_{ijk} \hat{p}_{ijk}^{NAHS} + \sum_{i \notin S_{jk}} a_{ijk} \hat{p}_{ij}^{NE} \tag{2}$$

The a_{ijk} were estimated using the 1995 Census data and thus are not the true proportions, but rather estimates based on census data. Chattopadahyay et al. (1999) caution that the estimators will be biased if unreliable or outdated estimates of a_{ijk} are used. The direct estimates $\hat{p}_{ijk}^{NAHS}$ in (2) may be unreliable if estimated from a small sample. Next we adapt the method proposed by Chattopadahyay et al. (1999) to improve the direct estimate (1) by borrowing strength from the ARE estimate ($\hat{p}_{ij}^{NE}$).

4.3. *Empirical Bayes method*

Let p_{ijk} denote the true lifetime drug use rate for age-class i in county k of region j ($i, j = 1, 2, 3$; $k = 1, \ldots, 93$) and assume that:

1. For each individual l, given p_{ijk}, the y_{ijkl} are uncorrelated with $E(y_{ijkl} | p_{ijk}) = p_{ijk}$ and $V(y_{ijkl} | p_{ijk}) = p_{ijk}(1 - p_{ijk})$.
2. For each county, p_{ijk} are uncorrelated with $E(p_{ijk}) = \mu_{ij}$ and $V(p_{ijk}) = h\mu_{ij}^2$.

The first assumption produces conditionally uncorrelated direct estimates $\hat{p}_{ijk}^{NAHS}$ by county, given the p_{ijk} values. Assumption 2 allows for variability in the lifetime drug-use rates for counties within the same region by age-class. Assuming squared error loss, the linear Bayes estimator of p_{jk} under this framework is

$$\hat{p}_{jk}^{B} = \sum_{i \in S_{jk}} a_{ijk} [B_{ijk} \hat{p}_{ijk}^{NAHS} + (1 - B_{ijk}) \mu_{ij}] + \sum_{i \notin S_{jk}} a_{ijk} \mu_{ij} \tag{3}$$

where:

$$B_{ijk} = \frac{h\mu_{ij}^2}{h\mu_{ij}^2 + c_{ijk}[\mu_{ij} - (h+1)\mu_{ij}^2]} \tag{4}$$

$$c_{ijk} = \frac{\sum_{l=1}^{n_{ijk}} w_{ijkl}^2}{\left(\sum_{l=1}^{n_{ijk}} w_{ijkl}\right)^2} \tag{5}$$

The empirical Bayes estimate is found by replacing B_{ijk} (4) and μ_{ij} by their estimates $\hat{B}_{ijk}$ and $\hat{\mu}_{ij} = \hat{p}_{ij}^{NE}$, respectively, where

$$\hat{B}_{ijk} = \frac{h(\hat{p}_{ij}^{NE})^2}{h(\hat{p}_{ij}^{NE})^2 + c_{ijk}[\hat{p}_{ij}^{NE} - (h+1)(\hat{p}_{ij}^{NE})^2]} \tag{6}$$

Using the proposed framework we construct two models for lifetime drug use within the county. In Model 1, we assume a uniform (0, μ_{ij}) prior on the p_{ijk}'s ($h = \frac{1}{12}$). For Model 2, we assume the prior distribution of the p_{ijk}'s to be uniform (0, $2\mu_{ij}$) ($h = \frac{1}{3}$). Applications of these two models to the 1995 NAHS data will be illustrated in the next section.

4.4. *Estimates of county drug use*

The direct estimates of lifetime drug use by county are displayed in Figure 1. There were 18 counties with a direct estimate of zero. As mentioned earlier, such an estimate results when none of the survey respondents reported lifetime drug use and often underestimates the true rate. We considered a county to have a moderately elevated drug use rate if the estimate is up to 110% of the state direct estimate and a highly elevated drug use rate if the estimate is 110% to 177% of the state direct estimate. Using the direct estimates, four counties had moderately elevated rates and ten counties had highly elevated rates.

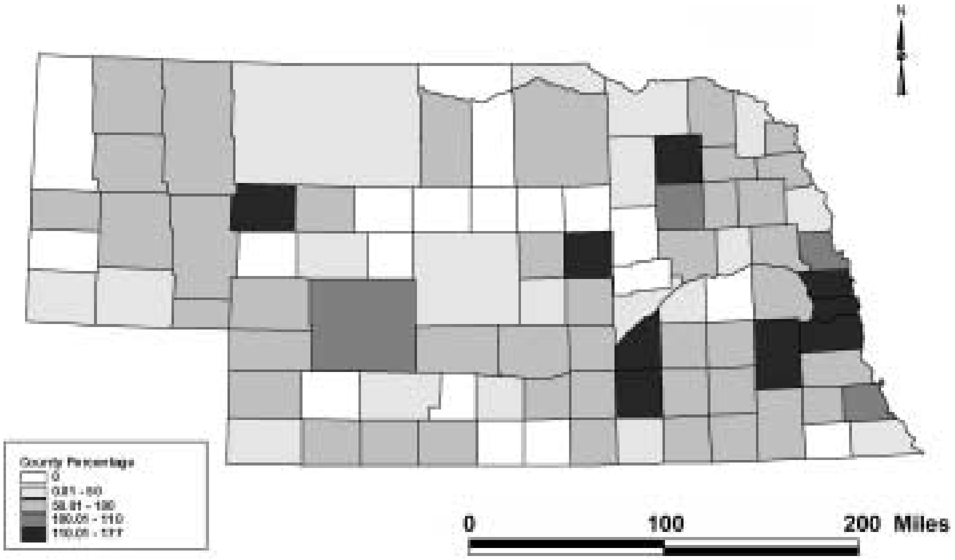

Fig. 1. Direct estimates of lifetime drug use by Nebraska county

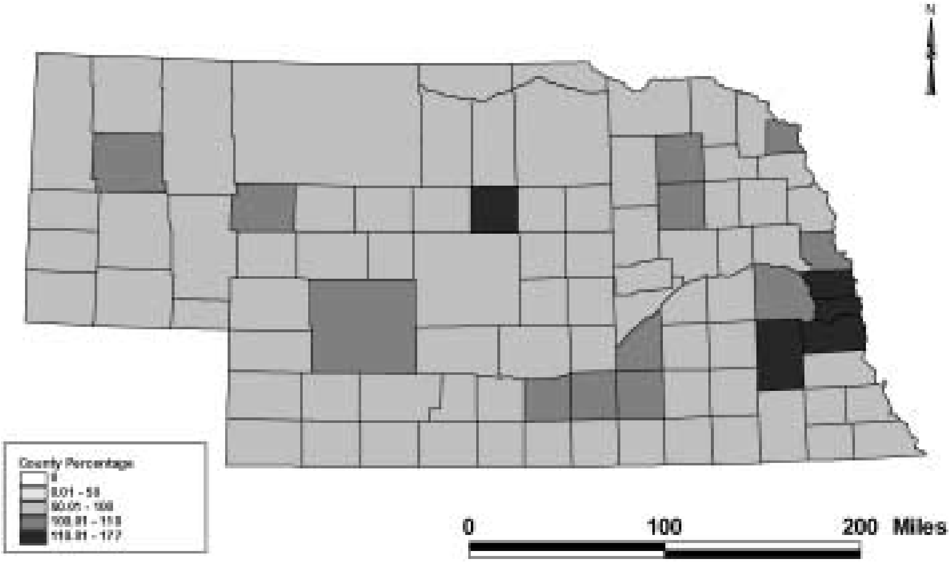

Fig. 2. Empirical Bayes estimates of lifetime drug use by Nebraska county, Model 1 (h = 1/12)

To improve the direct estimates we used the empirical Bayes estimators given in (3) in conjunction with (6) under Model 1 to produce the county drug use rates displayed in Figure 2. Here, all 93 Nebraska county lifetime drug use rate estimates are nonzero. There were 12 counties with moderately elevated rates and five counties with highly elevated rates. In Figure 3, estimates derived under Model 2 are displayed. Using this model, there were seven counties with moderately elevated rates and seven counties with highly elevated rates.

Clearly, the choice of prior is an important component in estimating drug use rates. All

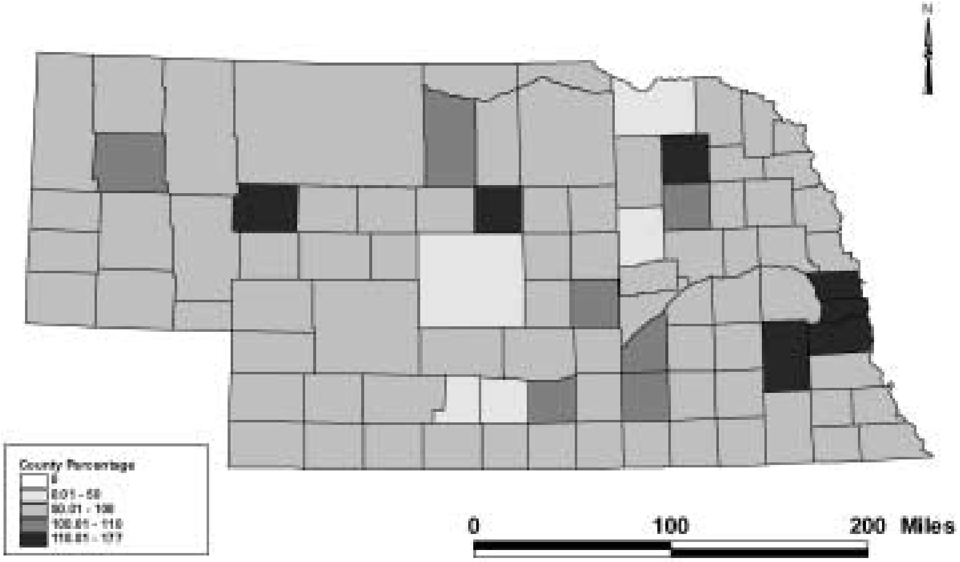

Fig. 3. Empirical Bayes estimates of lifetime drug use by Nebraska county, Model 2 (h = 1/3)

five counties with highly elevated drug use rates under Model 1 were also identified as having highly elevated ones under Model 2. However, Model 2 identified two additional counties with highly elevated rates that were only considered to have moderately elevated ones under Model 1. On the other hand, there were five counties with moderately elevated drug rates under Model 1 that had estimates *below* the state direct survey estimate under Model 2.

5. Precision of Estimates

Since the estimator in Section 4.3 involves both a raking ratio estimator and an empirical Bayes estimator, we investigate the precision of these two estimators in Sections 5.1 and 5.2, respectively. The overall estimator for the ARE will then have these two components of variation.

5.1. Raking ratio estimators

To estimate the precision of the raking ratio estimators that use marginal totals for two auxiliary variables each having two levels (e.g., ARREST and TREATMENT), we refer to the framework of Deville and Särndal (1993), who derive the variance and variance estimator for the raking ratio estimator assuming an additive effects model. Under such a model the asymptotic variance of the raking ratio estimator, $\hat{D}(A, T)$ is given by:

$$AV\{\hat{D}(A, T)\} = \sum \sum_{U} \Delta_{kl}(w_k E_k)(w_l E_l) \tag{7}$$

where $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$, $E_k = y_k - (A_i + B_j)$, A_i $(i = 1, 2)$ represents the row effect (e.g., ARREST) and B_j $(j = 1, 2)$ represents the column effect (e.g., TREATMENT). These quantities satisfy the finite population normal equations

$$\sum_{j=1}^{2} N_{ij}(A_i + B_j) = \sum_{j=1}^{2} N_{ij} p_{ij} \quad \text{for } i = 1, 2 \tag{8}$$

$$\sum_{i=1}^{2} N_{ij}(A_i + B_j) = \sum_{i=1}^{2} N_{ij} p_{ij} \quad \text{for } j = 1, 2 \tag{9}$$

where N_{ij} is the number of population units within cell ij and p_{ij} is the proportion of drug users within cell ij for $i, j = 1, 2$. For each k, $y_k = 1$ if unit k is a drug user and 0 otherwise. The probability that both units k and l are included in the sample is given by π_{kl} and the sample inclusion probability for unit k is π_k. The sample weight for unit k is then $w_k = 1/\pi_k$.

In the case of using marginal totals to rake sample cell estimates, the N_{ij} values are unknown. Using the notation defined in Section 3, the sample counterparts to the population normal equations (8) and (9) are (Deville and Särndal 1993):

$$\sum_{j=1}^{2} \hat{N}_{ij}^{R}(\hat{A}_i + \hat{B}_j) = \sum_{j=1}^{2} \hat{N}_{ij}^{R} \hat{p}_{ij} \quad \text{for } i = 1, 2 \tag{10}$$

$$\sum_{i=1}^{2} \hat{N}_{ij}^{R}(\hat{A}_i + \hat{B}_j) = \sum_{i=1}^{2} \hat{N}_{ij}^{R} \hat{p}_{ij} \quad \text{for } j = 1, 2 \tag{11}$$

Because we have a two-way classification table in which both marginals have the same population total, one of the equations in the system (10) and (11) is redundant. Upon removing one of these equations (e.g., set $B_2 = 0$), a computer algorithm (such as GREGWT from the Australian Bureau of Statistics, 2000) can be employed to generate the estimates $\hat{A}_i$, $\hat{B}_j$ for $i, j = 1, 2$.

Using these estimates, an estimator of the asymptotic variance of the raking ratio estimator (7) can be derived. From Deville and Särndal (1993) this estimate is given by:

$$\widehat{AV}\{\hat{D}(A,T)\} = \sum \sum_{S} (\Delta_{kl}/\pi_{kl})(w_k^* e_k)(w_l^* e_l) \tag{12}$$

where Δ_{kl} and π_{kl} are defined as in (7) and $e_k = y_k - (\hat{A}_i + \hat{B}_j)$ for k in cell ij. Finally, $w_k^* = w_k \hat{N}_{ij}^R / \hat{N}_{ij}$ where $\hat{N}_{ij}$ is the weighted sample estimate of the population size of cell ij in the two-way classification table ($i, j = 1, 2$).

While the computer algorithm can be used in conjunction with the survey data to obtain estimates of the A_i's and B_j's, calculations of the joint inclusion probabilities (i.e., π_{kl} and consequently Δ_{kl}) were not provided in the public-use files of the NAHS. It is possible to compute these inclusion probabilities using detailed descriptions of the sampling plan including unit level information. However, due to confidentiality issues, the publicly available NAHS survey documentation does not provide enough of these details. Using replicate and pseudo-psu information available from the 1992 public-use NHSDA data file along with the SAS macro GREGWT developed by the Australian Bureau of Statistics (2000), we were able to estimate the number of heroin users in the U.S. in the past year (1992) along with standard error estimates. The post-stratified estimate (Wright et al. 1997) was 623,469, with an estimated standard error of 93,859. The post-stratified raking ratio estimate of the number of heroin users in the past year was computed to be 675,974, with an estimated standard error of 118,156. In the case of additive effects for the chosen auxiliary variables, ''the raking ratio estimator has a variance that is slightly larger than that of the post-stratified estimator'' (Deville and Särndal 1993, p. 1016). A logistic regression model predicting the proportion of heroin users from the variables arrest and treatment was constructed using the 1992 public-use NHSDA data. All factors were significant, including the interaction term (all p-values < .0001). In the case of significant interaction effects the equivalence of the precision of the two estimators is not necessarily implied. The variance of the raking ratio estimator in the presence of interaction effects is currently under investigation.

5.2. *Empirical Bayes estimates*

The uncertainty of the empirical Bayes estimates is estimated by the mean squared error, defined as $MSE(\hat{p}_{jk}^B) = E(\hat{p}_{jk}^B - p_{jk})^2$, where the expectation is with respect to the model. Here, $MSE(\hat{p}_{jk}^B)$ can be shown to be:

$$MSE(\hat{p}_{jk}^B) = h\left(\sum_{i \in S_{jk}} a_{ijk}^2 (1 - B_{ijk}) \mu_{ij}^2 + \sum_{i \notin S_{jk}} a_{ijk}^2 \mu_{ij}^2\right)$$

A naïve estimate of the MSE of the empirical Bayes estimator can be found by replacing B_{ijk} (4) and μ_{ij} by their estimates $\hat{B}_{ijk}$ (6) and $\hat{\mu}_{ij} = \hat{p}_{ij}^{NE}$.

The naïve estimator, however, underestimates the uncertainty since the variance estimation of $\hat{p}_{ij}^{NE}$ is not considered (see Prasad and Rao 1990). These ARE's represent the cell estimates for a subdomain of interest within the cell – namely, lifetime drug users. Considering only this subdomain and raking only naïve estimates of the number of lifetime drug users within each cell using marginal totals for the number of users categorized by variables assumed to have additive effects on drug use, we would expect the raked estimates of the number of users within each cell to be fairly close to the actual number, provided one assumes that the marginal totals of the number of lifetime drug users in the state obtained from the NHSDA have negligible error. Since the denominators of each of the cell ARE's are known, a lower variation in these cell counts represents a lower variation in the overall ARE's ($\hat{p}_{ij}^{NE}$).

Chattopadahyay et al. (1999) used a jackknife estimator of the MSE. In this setting, a jackknife estimator of the MSE requires repeatedly computing the adjusted ratio estimator (ARE) after deleting an individual observation from the data set and is computationally intensive. Further consideration of estimating the MSE (along with the variation of the ARE's) will be discussed in a forthcoming paper.

6. Discussion

Estimating drug use at the state level can be more complex than estimation at the national level since the availability of independent state surveys may be limited. In addition, the survey instrument may not be as extensive or exhaustive as national counterparts. Using the post-stratified ratio estimator to estimate the number of drug users within a state based on two or more auxiliary variables may be accomplished if all cell sizes are known, either from the census or as in Wright et al. (1997) through an independent survey. However, as the number of levels of the auxiliary variables or the number of auxiliary variables increases, the classification table used to form the post-strata increases in dimension. If marginal totals are known, then $(r-1)\times(c-1)$ cells would need to be estimated from independent sources in order to apply the post-stratified estimator.

The raking ratio estimator eliminates the burden of obtaining independent sources to estimate the number of drug users. By using only marginal totals along with the survey data, the number of drug users may be estimated to nearly the same precision offered by the post-stratified estimator assuming no interaction effects of the auxiliary variables (Deville and Särndal 1993). ''The need to calibrate on marginal counts rather than on cell counts would be more strongly felt for a table with three or more dimensions'' (Deville and Särndal 1992, p. 380). In the case of known cell counts, the raking ratio estimator is equivalent to the post-stratified ratio estimator regardless of the distance function used in the raking algorithm (Deville and Särndal 1993).

One precaution for the raking ratio estimator arises when there are no sampled respondents within one of the cells in the classification table. In this case, the iterative proportional fitting algorithm may not converge (Lohr 1999). However, if an independent survey yielded an estimate of zero respondents in a particular cell, the post-stratified ratio estimator would not use any survey data from this cell. Thus, the empty cell problem affects both estimators. Additional research is needed for the evaluation of the raking ratio estimator in the presence of empty cells and in the case of nonnegligible interaction effects between the auxiliary variables.

State estimates may not be as accurate as estimates from national data since many state surveys collect data via telephone whereas many national surveys employ a personal interview. There has been some work on the comparability of estimates constructed using surveys with differing modes of collection (see Gfroerer and Huges 1991 or Biemer 2001). By linking state and national data it may be possible to improve the state-level estimates of drug use within the cells of a cross-classification table based on variables associated with drug use. These new estimates can provide better parameter estimates for small-area models used to produce county estimates. However, linking state and national survey data requires compatible auxiliary variables. Adjusted ratio estimates link cell estimates from the national data set to state population quantities through the cells of the classification table formed using these compatible variables. In order to improve accuracy of the cell counts, it is imperative that these auxiliary variables describe the same quantities for the same population (e.g., noninstitutionalized adult household population). Even when auxiliary variables are analogous, levels of these variables may need to be adjusted to provide compatible cross-classification tables for both the state and national survey data.

Once adjusted cell estimates are obtained, small-area models can be constructed to estimate county drug use rates. One may construct several models (such as the models considered in Section 4.3.1) for estimating small-area drug use rates. Additional models may also provide alternative variance structures for the region effects. One such alternative is to specify a correlation structure that allows for counties with close geographical proximity to have stronger correlation than counties with less geographical proximity. In any model-based application, it is important to check the model assumptions to select a model that fits the data. There is a some literature on model diagnostics in this setting which is reviewed briefly in Ghosh and Rao (1994). Additional research on small area models and model selection for estimating drug use is needed.

7. References

Australian Bureau of Statistics (2000). GREGWT User's Guide.

Biemer, P.P. (2001). Nonresponse Bias and Measurement Bias in a Comparison of Face to Face and Telephone Interviewing. Journal of Official Statistics, 17, 295–320.

Brackstone, G.J. and Rao, J.N.K. (1979). An Investigation of Raking Ratio Estimators. Sankhyā: The Indian Journal of Statistics, 41, 97–114.

Bray, R.M., Kroutil, L.A., and Wheeless, S.C. (1999). Comparing and Integrating Findings Across Populations. In Drug Use in Metropolitan America, R.M. Bray and M.E. Marsden (eds.), 267–295, Sage Publications, Inc., Thousand Oaks, CA.

Bureau of Sociological Research (1995). Nebraska State Demand and Needs Assessment Studies: Alcohol and Drugs – Adult Household Survey Results [and dataset]. University of Nebraska–Lincoln.

Chattopadhyay, M., Lahiri, P., Larsen, M., and Reimnitz, J. (1999). Composite Estimation of Drug Prevalences for Sub-state Areas. Survey Methodology, 25, 81–86.

Deming, W.E. and Stephan, F.F. (1940). On a Least Squares Adjustment of a Sampled Frequency Table When Expected Marginal Totals Are Known. Annals of Mathematical Statistics, 11, 427–444.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration Estimators in Survey Sampling. Journal of the American Statistical Association, 87, 376–382.

Deville, J.-C., Särndal, C.-E., and Sautory, O. (1993). Generalized Raking Procedures in Survey Sampling. Journal of the American Statistical Association, 88, 1013–1020.

Folsom, R.E. and Liu, J. (1994). Small-area Estimation for the National Household Survey on Drug Abuse. Proceedings of the American Statistical Association, Section on Survey Research Methods, 565–570.

Folsom, R.E., Shah, B., and Vaish, A. (1999). Substance Abuse in States: A Methodological Report on Model Based Estimates from the 1994–1996 National Household Survey on Drug Abuse. Proceedings of the American Statistical Association, Section on Survey Research Methods, 371–375.

Gfroerer, J.C. and Hughes, A.L. (1991). The Feasibility of Collecting Drug Abuse Data on Telephone. Public Health Reports, 106, 384–393.

Ghosh, M. and Rao, J.N.K. (1994). Small Area Estimation: An Appraisal. Statistical Science, 9, 55–93.

Horvitz, D.G. and Thompson, D.J. (1952). A Generalization of Sampling Without Replacement From a Finite Universe. Journal of the American Statistical Association, 47, 663–685.

Lohr, S.L. (1999). Sampling: Design and Analysis. Brooks/Cole Publishing Company; Pacific Grove, CA.

Maxwell, J.C. (2000). Methods for Estimating the Number of ''Hard-Core'' Drug Users. Substance Use and Misuse, 35, 399–420.

Prasad, N.G.N. and Rao, J.N.K. (1990). The Estimation of Mean Squared Errors of Small-area Estimators. Journal of the American Statistical Association, 85, 163–171.

U.S. Department of Commerce, Economics and Statistics Administration (1994). Geographic Areas Reference Manual, U.S. Bureau of the Census.

U.S. Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, Office of Applied Studies. NATIONAL HOUSEHOLD SURVEY ON DRUG ABUSE, 1995 [Computer file and Codebook]. ICPSR Version. Research Triangle Park, NC: Research Triangle Institute/Chicago, IL: National Opinion Research Center [producers], 1997. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 1997.

U.S. Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, Office of Applied Studies. UNIFORM FACILITY DATA SET, 1995: [UNITED STATES] [Computer file]. ICPSR version. Arlington, VA: Synectics for Management Decisions, Inc. [producer], 2000. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2000.

U.S. Department of Justice, Federal Bureau of Investigation. UNIFORM CRIME REPORTING PROGRAM DATA [UNITED STATES]: STATE-LEVEL DETAILED ARREST AND OFFENSE DATA (for Nebraska), 1995 [Computer file]. 2nd ICPSR ed. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [producer and distributor], 2001.

Wright, D., Gfroerer, J., and Epstein, J. (1997). Ratio Estimation of Hard-Core Drug Use. Journal of Official Statistics, 13, 401–416.

Received January 2002
Revised January 2003