Estimation in Multiple Groups in the Presence of External Constraints that Prohibit Explicit Data Pooling

Jerome P. Reiter, Harvard University Statistics Dept
Dept of Statistics, Science Center, Cambridge, MA 02138

Key Words: Census 2000, Data Pooling, Hierarchical Models, Information Pooling, Model Selection, Regression.

1. Introduction

If the Census Bureau uses sampling for integrated coverage measurement (ICM), it will need to estimate population size adjustment factors at state and sub-state levels. In many demographic groups and geographic locales, sample sizes will not be large enough to provide direct estimates with tolerable variances. In such small area problems, statisticians can improve estimation accuracy by smoothing the direct estimates across areas. For example, the adjustment factors can be smoothed with a hierarchical regression model that pools data across states.

Experience from Census 1990 suggests that the Census Bureau's clients view models that pool data across states with suspicion. Thus, to avoid controversy in Census 2000, the Census Bureau has expressed the desire to avoid explicitly pooling data across states [1, 2]. Nonetheless, there may be across-state information that, if somehow tapped, could improve the accuracy of the within-state estimates. This paper presents several ways of teasing out this across-state information without estimating adjustment factors by explicit data pooling.

External constraints that prohibit explicit data pooling potentially exist in many settings outside of Census 2000. For example, when sampling is used to audit or assess several groups, the groups may reject explicit data pooling if they fear it will shrink their direct estimates in a way that makes them look worse. Constraints could also exist when: (1) clients do not allow the explicit use of prior years' data; (2) statisticians cannot release data from one group to another; and, (3) statisticians want to avoid explaining hierarchical models to their clients. The techniques in this paper may be useful in addressing these constrained estimation problems.

This research was funded through a contract from the United States Census Bureau. The author thanks Donald Rubin, David van Dyk, Alan Zaslavsky, John Barnard, and Ann Vacca.

2. Potential Solutions: Information Pooling

By explicit data pooling, I mean using a model in which multiple groups' (e.g., states) data enter directly into the formulas used to estimate any of the parameters ultimately included in each group's model. Thus, a hierarchical model is an example of explicit data pooling.

Explicit data pooling is one technique in a more general class of approaches to improving prediction accuracy, namely *information pooling*. I define information pooling to be using both a group's data and knowledge not contained in that group's data to make estimates in that group. This is a broad definition, and nearly every statistical analysis employs some form of information pooling. For example, an essential form of information pooling is relying on past experience to design the data collection mechanism and to build the statistical models for an estimation task. Two ongoing applications of this information pooling strategy that will improve the estimates in Census 2000 are: 1) using the collective knowledge of the Census Bureau to design the ICM sampling scheme; and, 2) performing simulation studies with previous census data from many states to determine which statistical techniques best predict population sizes. Another form of information pooling– and one that is a focus of this paper– is to use the estimates of parameters in multiple groups to help identify the predictors that should be included in each group's model. This form of information pooling is not explicit data pooling if, once the statistician specifies the model in each group, multiple groups' data are not used to estimate the parameters that are ultimately included in each group's model.

Information pooling techniques exist conceptually on a continuum ordered by how directly the techniques rely on multiple groups' data to make estimates in each group. At one extreme of the continuum is solely using past experience, which I call *minimal information pooling*. At the other extreme of the continuum is the use of explicit data pooling, which I call *maximal information pooling*. In

between these two extremes is a host of information pooling techniques that use multiple groups' data somewhat indirectly, such as the model selection technique mentioned in the above paragraph. Since this model selection technique relies on estimates of parameters from multiple groups to specify each group's model, it uses multiple groups' data more directly than does solely using past experience. Since it ultimately estimates the included parameters in each group from just the data in that group, it does not use multiple groups' data as directly as explicit data pooling. I call such techniques *medial information pooling.*

It seems clear that a legislated or self-imposed prohibition of *all* information pooling strategies would severely restrict, if not eliminate, the ability of the Census Bureau to provide accurate estimates. However, just because one form of information pooling is unacceptable, namely explicit data pooling, it does not follow that all forms of information pooling are unacceptable. Thus, the pertinent question the Census Bureau should consider is not *whether* information pooling is permissible; rather, it is, "*how much* information pooling is permissible?" In other words, how far along the information pooling continuum is the Census Bureau willing to travel? Undoubtedly, as the techniques move towards maximal information pooling, they are more likely to be perceived by Census clients as similar to explicit data pooling, and hence less likely to be acceptable to clients that disapprove of explicit data pooling.

If there is little reduction in estimation errors when a potentially controversial form of information pooling is employed, then it is not worthwhile to argue for that strategy. Thus, a second pertinent question the Census Bureau should consider is, "does using an information pooling strategy reduce estimation errors by a sufficient amount?" Implicit in the answers to this question is a tradeoff between accuracy and acceptability. We expect the strategies that rely more directly on multiple groups' data to produce estimates with smaller mean-squared errors, yet these strategies will be more controversial. Those strategies that look less like explicit data pooling will be easier to justify, but they will not give as large a payoff in estimation accuracy. Therefore, to find a strategy with a satisfactory balance between accuracy and acceptability, it is necessary to consider strategies at many locations of the information pooling continuum.

3. Some medial information pooling strategies for regression models

In this paper, we assume that the statistician will model the data with multiple regressions:

$$y_{ij} \sim N(x_{ij}\beta_i, \sigma^2_{ij}),$$

where x_{ij} represents a $1 \times p$ row vector of predictors for the jth observation in the ith group. Assuming known sampling variances σ^2_{ij} (and covariances), ICM smoothing fits into this framework as follows. Let i index a state, j index a demographic or geographic post-stratum, and y_{ij} be the direct estimate of the ijth adjustment factor. Let x_{ij} be a vector of dummy variables corresponding to main effects and interactions that define the post-strata. When we drop interactions or main effects, we smooth the factors. For example, consider a universe with only 4 post-strata in each state: black/white crossed with renter/owner. Then, an intercept, a main effect for black post-strata, a main effect for renter post-strata, and an interaction between black and renter post-strata would account for all post-strata. Dropping the interaction or the main effects would smooth the adjustment factors.

What kind of medial strategies are located on the information pooling continuum? Or, putting it a different way, what useful information might be teased out of the multiple groups' data that cannot be found in individual groups' data? To answer this question, it is helpful to construct a wish-list of information that, if known, might aid modelers to improve predictions. The goal of a medial information pooling strategy is to use multiple groups' data to make one of these wishes come true, or at least approximately true, without violating the constraints.

The first thing we might wish for is knowing if estimates of regression coefficients are close to the true values of the coefficients. If they are not, we would be better off removing the predictors from the model instead of estimating their coefficients. How might multiple groups data help grant this wish? Consider the following anecdote: say the coefficient for predictor X_1 in one group is estimated as $\hat{\beta}_1 = 5$ with large variance, but in all of the other groups the coefficient of X_1 is close to two with small variance. We might believe that the one large $\hat{\beta}_1$ resulted from sampling variability, and in reality the coefficient is similar to those in other groups. Eliminating X_1 from that group will yield more accurate predictions since $\hat{\beta}_1 = 0$ is closer to two than $\hat{\beta}_1 = 5$. This example can be translated into a medial information pooling strategy: use multiple groups' data to help determine which predictors have coefficients that are

poorly estimated, and then estimate the coefficients ultimately included in each group's model using just that group's data. I call this a *Matching* strategy.

A second item on the wish-list, if the first is unavailable, is knowing whether a predictor is unimportant; that is, does dropping a predictor reduce the mean squared errors of predictions. Satisfying this wish is typically the goal of traditional model selection strategies, such as choosing the model in each group that minimizes the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC). Using across-group information might allow us to more effectively achieve this goal. Consider the following example: the coefficient for predictor X_2 is estimated as unimportant in all of the groups' regressions except for one. We might believe that the estimated coefficient in the one group is a result of sampling variability, and in reality the coefficient is similar to those in other groups. We may want to eliminate X_2 from all the models to improve accuracy. This example leads to the following information pooling strategy: use multiple groups' data to help identify unimportant predictors that can be excluded from the models, and then estimate the coefficients ultimately included in each group's model using just that group's data. I call this an *Importance* strategy.

If knowledge from *Matching* or *Importance* strategies is not available, it would helpful to know whether a traditional model selection procedure should be performed in a group. Traditional model selection procedures can improve accuracy when there are many unimportant predictors, but can also lead to excluding important predictors. If multiple groups' data can be used to decide if traditional selection strategies should be employed, we may be able to take advantage of the benefits of these procedures while avoiding the drawbacks. I call this a *Selection* strategy.

A next item is knowing the number of predictors in each group that make an important contribution to the predictive ability of the model. With knowledge of the number of important predictors, we can use standard techniques, like selecting the model that maximizes R^2, in hopes of finding those important predictors and, in the process, excluding unimportant predictors. Since multiple groups' data can help identify important predictors, they also can help identify the number of important predictors. This leads to the final information pooling strategy examined in this paper: use multiple groups' data to determine the number of predictors to include in each group's model, and then determine those predictors separately in each group. I call this a *Dimension* strategy.

How can we implement these four strategies? To create viable procedures, it is necessary to mine the extra information that exists in multiple groups' data. Hierarchical models easily and effectively take advantage of across-group information. It makes sense, then, to use these hierarchical models as tools for extracting this information. Effectively, this means fitting a hierarchical model that explains the data as well as possible and using the results to assist in model specification. Importantly, this form of information pooling may not be a violation of the constraints since it uses explicit data pooling only as a tool to extract across-group information which might otherwise be difficult to tease out. In other words, even though we may not be permitted to estimate parameters via explicit data pooling, we may be permitted the to improve model construction by using the *results* of explicit data pooling.

Below are some procedures that attempt to accomplish the goal of each strategy, assuming a good-fitting hierarchical normal regression model (HNRM) has been found. Each procedure uses a stepwise model selection in each group that stops when a criterion is satisfied. These procedures are constructed assuming that the variances $\sigma_{ij}^2 = \sigma_i^2$ are unknown and that each observation is independent (i.e., the usual OLS set-up). This does not correspond to ICM smoothing, but leads to easier demonstrations of the potential of medial information pooling. The procedures can be adjusted to account for known variances and/or weighted least squares. I have also created procedures that select a set of predictors that must be included in every group, but I do not comment on these in this paper.

1. *Matching* strategy, *M1*: In each group, choose the set S_i of predictors that produces OLS fitted values with the smallest squared distance from the fitted values of the HNRM:

$$\min_{S_i} (\hat{Y}_{hnrm}^{F_i} - \hat{Y}_{sep}^{S_i})^t (\hat{Y}_{hnrm}^{F_i} - \hat{Y}_{sep}^{S_i}),$$

where $\hat{Y}_{hnrm}^{F_i}$ is the vector of fitted values in the ith group using the HNRM with all predictors included, and $\hat{Y}_{sep}^{S_i}$ is the vector of fitted values in the ith group using an OLS model with only the predictors in set S_i included. Since the HNRM yields estimates that are on average closer to the truth than OLS estimates, matching to the HNRM's estimates should lead to non-hierarchical models that yield better predictions.

2. *Importance* strategy, *I1*: In each group, choose the set S_i of predictors that leads to the smallest

prediction mean squared error:

$$\min_{S_i} (\hat{Y}^{F_i}_{hnrm} - E(\hat{Y}^{S_i}_{sep}))^t (\hat{Y}^{F_i}_{hnrm} - E(\hat{Y}^{S_i}_{sep})) + p^{S_i} \hat{\sigma}^2_{seg}),$$

where $\hat{\sigma}^2_{seg}$ is the estimate of the conditional variance from fitting the full OLS model, and $E(\hat{Y}^{S_i}_{sep}) = (X^t_{S_i} X_{S_i})^{-1} X^t_{S_i} \hat{Y}^{F_i}_{hnrm}$. Effectively, this treats the fitted values from the HNRM as the true mean response of the full OLS model. I have also created procedures that minimize AIC-like and BIC-like criteria, but I do not describe these in this paper.

3. *Selection* strategy, *S1*: In each group, use the predictors in the set S_i from the model that minimizes the AIC if:

$$(\hat{Y}^{F_i}_{hnrm} - \hat{Y}^{S_i}_{AIC})^t (\hat{Y}^{F_i}_{hnrm} - \hat{Y}^{F_i}_{AIC}) \leq (\hat{Y}^{F_i}_{hnrm} - \hat{Y}^{F_i}_{sep})^t (\hat{Y}^{F_i}_{hnrm} - \hat{Y}^{F_i}_{sep}),$$

or else use the full OLS model. Here, $\hat{Y}^{F_i}_{sep}$ is the vector of fitted values in the ith group using the OLS regression with all predictors included. This is similar to *M1*, but only two models are compared to the HNRM.

4. *Dimension* strategy, *D1*: In each group, determine the number of predictors to include, and then do best-subsets regression to select the model with the largest coefficient of determination, R^2, for that number of predictors. The number of predictors in each group is the number of predictors in the group's model after application of *I1*.

Determining which, if any, of these procedures might be acceptable to Census clients is, of course, the responsibility of the Census Bureau. Roughly, I anticipate the ordering of acceptability to be similar to the order in which the procedures are presented.

4. Assessing accuracy via a simulation study

To investigate the effectiveness of these procedures, I perform a simulation study. Data sets are simulated from the following data generation model:

$$y_{ij} = x_{ij}\beta_i + \epsilon_{ij}, \qquad \epsilon_{ij} \sim N(0, \sigma^2_i).$$

Here y_{ij} represents the dependent variable for the jth unit in the ith group, x_{ij} is a row vector of predictors for the ijth unit, β_i is a column vector of regression coefficients in the ith group, and $\sigma^2_i = 40$ is the conditional variance in the ith group. There are 50 groups with 40 observations in each group, and there are 14 predictors considered for inclusion in each group. The predictors are drawn independently from a $N(0, I)$, and the intercept term is 0. This set-up does not correspond to ICM estimation, but it does facilitate investigation of the potential of medial information pooling strategies.

With this simulation design, the variance matrix of the estimates of the OLS coefficients in the full model is on average the identity matrix. This makes it easy to define values of the regression coefficients that have meaningful interpretations: the value of a coefficient equals the number of standard deviations it is from zero. To examine the performance of the procedures at different distances from zero, I simulate six scenarios:

1. **mu=4**: $\beta_{ik} = 4$ for $k = 1 \ldots 14$, $i = 1 \ldots 50$.
2. **mu=3**: $\beta_{ik} = 3$ for $k = 1 \ldots 14$, $i = 1 \ldots 50$.
3. **mu=2**: $\beta_{ik} = 2$ for $k = 1 \ldots 14$, $i = 1 \ldots 50$.
4. **mu=1**: $\beta_{ik} = 1$ for $k = 1 \ldots 14$, $i = 1 \ldots 50$.
5. **mu=0**: $\beta_{ik} = 0$ for $k = 1 \ldots 14$, $i = 1 \ldots 50$.
6. **mu=0/4**: $\beta_{ik} = 4$ for $k = 1 \ldots 7$, $i = 1 \ldots 50$ $\beta_{ik} = 0$ for $k = 8 \ldots 14$, $i = 1 \ldots 50$.

Thus, as we move from scenario **mu=4** to scenario **mu=0**, the predictors become less important. The scenario **mu=0/4** represents the realistic scenario of a mixture of unimportant and important predictors.

The procedures depend on finding a good-fitting hierarchical model. The hierarchical model I fit to this data is the random coefficients model:

$$y_{ij} \sim N(x_{ij}\beta_i, \sigma^2_i), \qquad \beta_i \sim N(\mu, \Sigma).$$

I estimate parameters by maximum likelihood. In these simulations, the random coefficients model is very effective because of the large shrinkage in the direct estimates. Thus, this simulation study contains best-case scenarios for the procedures: if they cannot perform well when a HNRM is really helpful, they are unlikely to perform well in general settings.

5. Conclusions from the simulations

For each scenario, I simulate 4 data sets containing 80 observations. I run the procedures on the first 40 observations and use the resultant models to predict the true dependent variables from the second set of 40 observations. I also make predictions using

the model that minimizes the AIC criterion (procedure *MAIC*) and the full OLS regression (procedure *SEG*). The results from all procedures are summarized on the graph at the end of this paper. Each symbol on the graph represents the average squared prediction error for a procedure divided by the average squared prediction error for fitting a full OLS model. Thus, any procedure whose symbol is to the left of 1.00 is an improvement on full OLS regression, and any procedure to the right of 1.00 is worse than OLS regression. The standard errors for the symbols on the graph are around 1%.

Regardless of where the coefficients are, the HNRM yields about a 33% improvement in prediction accuracy. Of course, we cannot use the hierarchical model because it is explicit data pooling.

When the coefficients are four standard deviations from zero, all the medial information pooling procedures select the full OLS model. Thus, these procedures are appropriately identifying and including the important predictors. On the other hand, *MAIC* performs 20% worse than *SEG*. *MAIC* drops predictors when their coefficients' estimates are sufficiently small, even though they are truly important.

As the coefficients approach two standard deviations from zero, we see some separation in the procedures: *M1* jumps out to a 6% improvement over *SEG*. *M1* drops predictors whose coefficients' estimates are far from their true values, even if the estimates are far from zero. Other medial information pooling procedures perform similarly to *SEG*, while *MAIC* continues to perform poorly.

When coefficients are one standard deviation away from zero, *M1* increases to a 15% improvement in accuracy, and *I1* yields about an 8% improvement in accuracy. *I1* is beginning to identify and drop unimportant predictors, and it is doing so more effectively than *MAIC*, which yields about the same accuracy as *SEG*. *S1* shows a 3% gain in accuracy, indicating that in some groups it is appropriately choosing the *MAIC* model and in other groups appropriately choosing the full OLS model. As in previous scenarios *D1* hovers conservatively around 1.00.

When all the coefficients equal zero, *M1*, *I1* and *D1* yield about a 28% improvement over *SEG*. All three procedures are doing an excellent job of picking off unimportant predictors. In fact, in this scenario the gains in accuracy are close to gains in accuracy from using the HNRM. In contrast, *MAIC* improves accuracy by only 16%. As expected, *S1* mirrors the performance of *MAIC*.

In the more realistic scenario with half unimportant and half important predictors, *M1* and *I1* yield a 20% improvement, *D1* yields a 9% improvement, *S1* yields a 7% improvement, and *MAIC* yields a 5% improvement. The medial information pooling procedures are identifying the important predictors while excluding the unimportant predictors, and they are doing so more effectively than *MAIC*.

Although this is a limited simulation study, it seems clear that, by using medial information pooling procedures, we can substantially improve estimates without explicit data pooling.

6. Applying these ideas in Census 2000

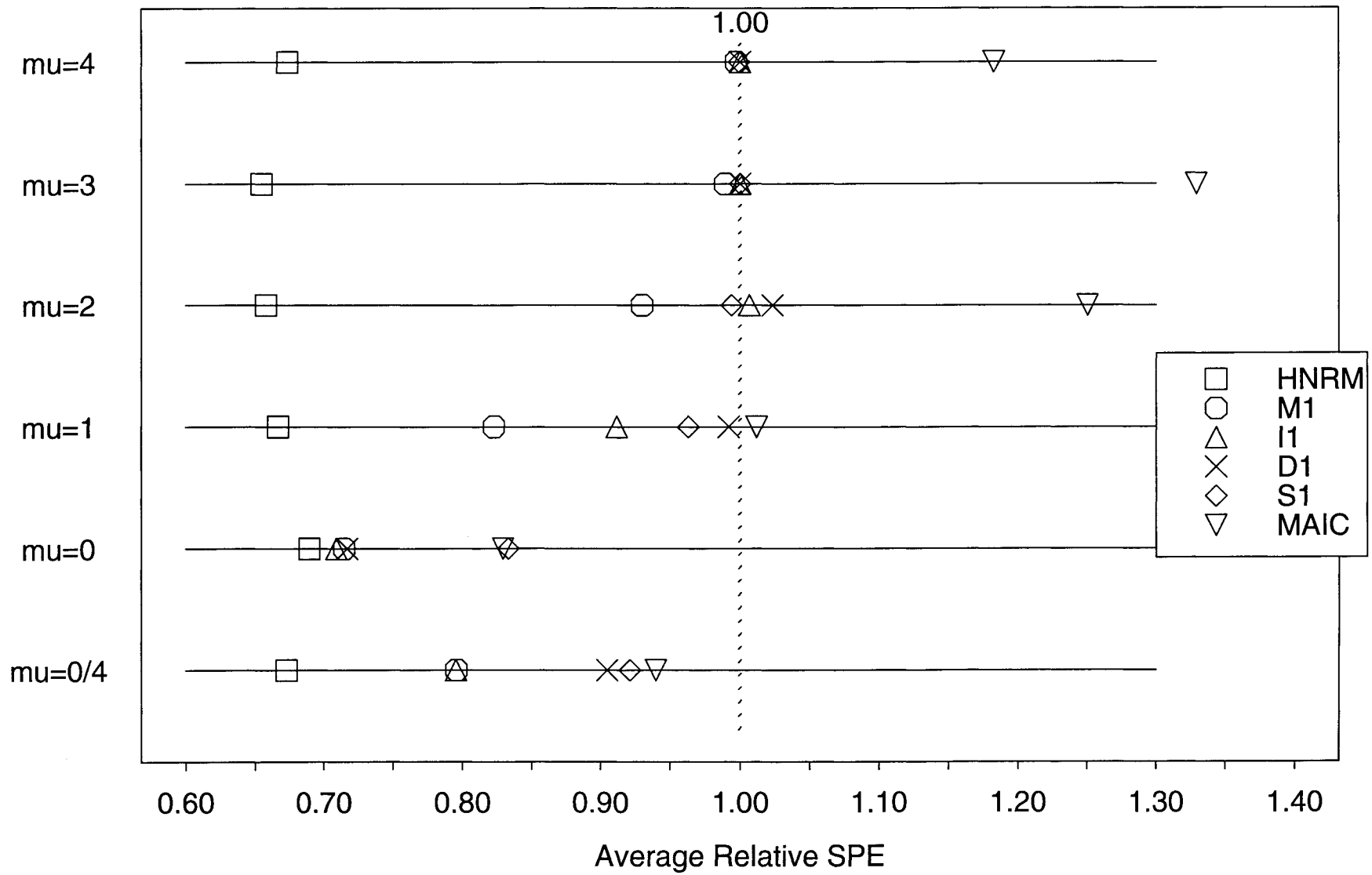
A possible application of these techniques in Census 2000 is to use multi-state information to help determine the adjustment factors for the census counts in sub-state demographic and geographic post-strata. Using dummy variables to represent main effects and interactions for post-strata, we can build a hierarchical model that smooths adjustment factors, and then use medial information pooling to construct a smoothing model in each state. If raking is to be used, predictors from the state-specific smoothing regressions can define the margins used in the raking matrix for each state. With modifications, these procedures could also be based on log-linear models.

For ICM, the medial information pooling procedures will have to be modified to account for known sampling variances and covariances of the estimates of adjustment factors. In future work, I plan to explain these modifications.

References

[1] Robert E. Fay and John Thompson. The 1990 post enumeration survey: Statistical lessons, in hindsight (with discussion). In *Proceedings of the Bureau of the Census Annual Research Conference*, volume 9, pages 71–96, 1993.

[2] Mary H. Mulry. Comment on a paper by Kadane. *Journal of Official Statistics*, 12:101–104, 1996.


Average Relative SPEs for Procedures
1.00
mu=4
mu=3
mu=2
mu=1
mu=0
mu=0/4
HNRM
M1
I1
D1
S1
MAIC
0.60
0.70
0.80
0.90
1.00
1.10
1.20
1.30
1.40
Average Relative SPE